National Technical University of Athens
School of Rural, Surveying and Geoinformatics Engineering
Laboratory of Photogrammetry

Integrating scene priors in Multiple View Stereo (MVS)

Ph.D. Dissertation

Elisavet Konstantina Stathopoulou

Supervisors:
Prof. Andreas Georgopoulos
Dr. Fabio Remondino
Prof. Konstantinos Karantzalos

Athens, September 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών - Μηχανικών Γεωπληροφορικής
Εργαστήριο Φωτογραμμετρίας

Ενσωμάτωση δεσμεύσεων στην πολυεικονική ανακατασκευή

Διδακτορική Διατριβή

Ελισάβετ-Κωνσταντίνα Σταθοπούλου
Α.Μ. 65140002

Επιβλέποντες:
Καθ. Ανδρέας Γεωργόπουλος
Dr. Fabio Remondino
Καθ. Κωνσταντίνος Καράντζαλος

Αθήνα, Σεπτέμβριος 2022

Advisory Committee:
Andreas Georgopoulos
Professor, National Technical University of Athens

Fabio Remondino
Head of Unit, 3D Optical Metrology, Fondazione Bruno Kessler

Konstantinos Karantzalos
Associate Professor, National Technical University of Athens

Examination Committee:
Andreas Georgopoulos
Professor, National Technical University of Athens

Fabio Remondino
Head of Unit, 3D Optical Metrology, Fondazione Bruno Kessler

Konstantinos Karantzalos
Associate Professor, National Technical University of Athens

Charalambos Ioannidis
Professor, National Technical University of Athens

Andrea Fusiello
Associate Professor, University of Udine

Anastasios Doulamis
Associate Professor, National Technical University of Athens

Maria Pateraki
Assistant Professor, National Technical University of Athens

From October 2018 and till the end of this dissertation the author has been a fully-funded researcher in 3D Optical Metrology group (3DOM) at Fondazione Bruno Kessler (FBK), Italy.

Since the beginning of this dissertation and till October 2016 the author received funding from the EU Marie Sklodowska Curie Actions under the project ITN-DCH: Initial Training Network for Cultural Heritage, FP7 PEOPLE research framework (Grant Agreement no. 608013), hosted by National Technical University of Athens (ELKE 63/200).

The custom data used in the experiments section of Chapters 4 and 6 were acquired during the June 2021 campaign in Limassol, Cyprus, conducted for the needs of the project "Portal for heritage buildings integration into the contemporary built environment" PERIsCOPE INTEGRATED/0918/0034 that is co-financed by the European Regional Development Fund and the Republic of Cyprus through the Research Innovation Foundation, and is coordinated by The Cyprus Institute. The author has been involved in the project within FBK-3DOM (January 2020-September 2021).

Cite this thesis as:
Stathopoulou, E.K. (2022). Integrating scene priors in multiple view stereo, PhD Thesis, National Technical University of Athens.

Η έγκριση της διδακτορικής διατριβής από την Ανώτατη Σχολή Αγρονόμων και Τοπογράφων Μηχανικών - Μηχανικών Γεωπληροφορικής του Εθνικού Μετσοβίου Πολυτεχνείου, δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα (Ν. 5343/1932, άρθρο 202).

Acknowledgements

This dissertation is the outcome of several years of research and involvement in various projects on an international scale. I started this journey in NTUA as a Marie Curie fellow when my general research objectives were set, yet the final shape of this thesis was carved during the last years that I have been a member of the FBK-3DOM group. In these few lines, I would like to express my gratitude to those who stood by me during this time.

First, I would like to thank my advisors; I deeply thank Prof. Andreas Georgopoulos for his unconditional support all these years. Words are not enough to describe my gratitude and admiration as Andreas has been a mentor, a parent, and a friend since my undergraduate years. With his inspirational presence as a professor, he motivated me to pursue doctoral studies and introduced me to academia. Since then, he has been guiding me in research paths, advising me as an educator, supporting me through difficult times, and always believing in me. But mostly, with his distinctive ethos and humbleness, he has been a great life example to me, and I am sincerely grateful for this. Then, I would like to express my gratitude to my co-supervisor, Dr. Fabio Remondino; this thesis could not have been realized without his support. I want to thank him for believing in me and encouraging me since the beginning of my research career. Fabio welcomed me into his group and involved me in many interesting research projects while motivating me to finalize my thesis and giving me guidance; I am grateful for being given this opportunity. I would also like to thank my third supervisor, Prof. Konstantinos Karantzalos, for the valuable discussions, constructive comments, and support, especially during the last months of this journey.

I sincerely thank Dr. Dan Cernea for his time and guidance in practical matters regarding the implementation, as well as theoretical ones. A big part of this work builds upon his OpenMVS library, which is an excellent example of open-source code in our community, manifesting the importance of transparency and solidarity in research. It was an honor to collaborate with him in extending the framework and publishing articles together. Moreover, I would like to say a huge thanks to my good friend and colleague Roberto Battisti for his support in the implementation part of the last articles leading to this thesis. Working side to side with such a skilled developer has been of irreplaceable value. I would also like to thank my colleague Salim Malek for his kind advice on semantic segmentation topics during these last months. I also thank my other colleagues, former and current, in FBK-3DOM, especially: Eleonora, Fabio, Isabella, Daniele, Michele, Ali, Emre, Simone, and Alessandro; our fruitful discussions definitely helped me in this research path, but I also thank you for being my friends and family in Trento.

My special thanks to Erica Nocerino for the collaboration in research articles, but also for being my friend, supporting, and advising me.

I would like to acknowledge the other committee members, Prof. Charalambos Ioannidis, Prof. Anastasios Doulamis, Prof. Maria Pateraki, and Prof. Andrea Fusiello, for their interest and willingness to participate in the examination process. I shall also express my gratitute to the professors and personnel of the Photogrammetry Laboratory for collaborating on several projects during my years at NTUA. Also, many thanks to Dr. Pol Kolokousis from the Remote Sensing Laboratory for showing me the way toward open-source, both as an undergrad student and in the beginning of this research path. Moreover, I thank Andrea Romanoni and Qingshan Xu who kindly run experiments on the two custom datasets used in Chapter 5 with TAPA-MVS and ACMM/ACMP respectively, back in November 2021 when their code was not publicly available. Acknowledgements also to Matteo Rappana for his contribution in the manual annotation of the 3DOM Semantic Facade dataset.

I shall also mention my co-fellows and supervisors in the Marie Curie ITN-DCH project, during which I had the opportunity to be a visitor to several research groups and companies across Europe, work on interesting projects and participate in several conferences. The years I spent with these wonderful people were full of unforgettable experiences around the world. Most of them will be friends for life to me; special mention to Anais, Diego, Gina, Matevz, Matthew, Magda, and Rosella for all the moments we shared. A very special thanks to Nicola for the constant moral support all these years.

I must dedicate some words to all my good, life-long friends who have been by my side in recent years, in sunny and dark moments; the name list is long, and I consider myself blessed for having them. With most of them spread around the world, we constantly prove that *geographic gaps can be bridged.* To some of them, great engineers and researchers, I owe an additional acknowledgment because apart from the support, they enjoyed research-related discussions with me: thank you, Katerina, Lena, Lida, Valia, and Panos.

Finally, I would like to thank my family; my beloved parents, Eleni and Fotis, for raising me with love, inheriting me moral principles, and supporting my education, as well as my dear siblings, Katerina and Panagiotis, for always being there for me.

«Εύρον πράσινην πέτραν ωραιοτάτην. Ελθέ αμέσως, Ζορμπάς.»
Νίκος Καζαντζάκης, Βίος και Πολιτεία του Αλέξη Ζορμπά, (1946).

Abstract

Image-based 3D reconstruction addresses the problem of generating 3D representations of 3D scenes given overlapping 2D images as observations. It is one of the fundamental topics in photogrammetry and computer vision, counting many decades of research. In recent decades, many robust algorithms have been introduced for pixel depth estimation and 3D reconstruction, achieving great results in various applications. However, there are still several open challenges and space for improvement toward efficient, complete, and accurate 3D reconstruction in real-world scenarios. Geometric 3D reconstruction is closely related to scene understanding, another hot topic in computer vision research that has seen tremendous growth due to the recently developed deep learning algorithms. Indeed, advanced scene prior cues can potentially support efficient 3D reconstruction and vice versa. However, semantic reasoning directly in the 3D space is non-trivial mainly due to the limited availability of training data and the computational complexity; on the contrary, algorithms for 2D semantic segmentation are mature enough to obtain robust results, and the existence of large-scale datasets facilitates the generalization of the trained models. In this dissertation, both 3D reconstruction and semantic segmentation are comprehensively studied and interlinked; the main open challenges and limitations are identified, while innovative and easy-to-implement solutions in real-world scenarios are proposed.

In the field of semantic segmentation, a new benchmark with GT semantic maps of pixel-level accuracy for historic building facades is introduced, *3DOM Semantic Facade*, acknowledging the lack of existing, high-resolution benchmarks for similar purposes. Using this benchmark, a straightforward pipeline for model training based on state-of-the-art learning algorithms is proposed, and the inference results are experimentally evaluated on unseen data. Moreover, a new functionality is built upon the open-source and broadly-used MVS pipeline OpenMVS to enable label transfer from 2D to 3D, yielding semantically enriched dense point clouds. At the same time, selective (class-specific) reconstruction is made possible based on the semantic label of each scene pixel; in this way, the reconstruction of only the areas of interest is enabled according to the needs of each application. These functionalities are domain-independent and can, thus, be generalized in every MVS scenario for which semantic segmentation maps are available.

Regarding depth estimation and reconstruction, this thesis focuses on the multi-view stereo (MVS) part of the 3D reconstruction; it proposes methods to integrate advanced scene priors in the process in order to obtain high-quality and complete 3D point clouds. Depth estimation typically relies on correspondence search based on visual appearance between image pixels and is commonly measured using photometric consistency metrics. A variety of robust algorithms exist for efficient correspondence search and subsequent depth reconstruction for both stereo (two-view) and multi-view scenarios. Yet, certain limitations regarding the geometry of the scene (slanted surfaces, occlusions), material properties (repetitive patterns, textureless, reflective, or transparent surfaces), and acquisition conditions remain challenging. The main goal and objectives of the thesis refer to the development of novel practical approaches toward confronting the inevitable matching ambiguities in large, non-Lambertian surfaces due to the nature of the photometric consistency costs.

The first proposed method exploits semantic priors to indicate important cues for the 3D scene structure. Thus, a novel strategy is proposed to guide the depth propagation in such challenging surfaces under a PatchMatch-based scenario using RANSAC-based plane hypotheses in the 3D space. Then, a novel, adaptive cost function is introduced to leverage the prior hypotheses with the standard photometric cost and adaptively promote more reliable depth estimates across the image. During the experimental evaluation on the *ETH3D* benchmark as well as on custom scenes, the proposed algorithm achieved constantly better results than the baseline method in point cloud completeness while not sacrificing accuracy. Given the growing availability of semantically segmented data, this approach can be implemented in a variety of scenarios, indoor and outdoor. However, in real-world applications, it is not always trivial to obtain such semantic cues for every scene; a large amount of additional GT data may be required, and model training or fine-tuning is often a laborious task. Thus, an alternative, generic and domain-independent solution is also proposed, guided only by local structure and textureness cues. Based on quadtree decomposition on the image, groups of pixels with similar color attributes are grouped together. Similar to the previous method, planar hypotheses are extracted in 3D and guided by the quadtree blocks. The adaptive cost function is also used here to support PatchMatch depth propagation. Results on the entire training and test set of the *ETH3D* dataset demonstrate the effectiveness of the proposed approach and show a clear improvement in performance scores with respect to the baseline method while being competitive with other state-of-the-art algorithms. To further prove the applicability of the new method under varying scenarios, two more custom datasets were considered, on which similar improvements were achieved. The proposed methodologies are integrated into the well-established, open-source framework OpenMVS to promote usability and reproducibility.

Εκτεταμένη Περίληψη

Η τρισδιάστατη (3Δ) αποτύπωση του χώρου αποτελεί ένα σημαντικό ερευνητικό αντικείμενο τόσο στην επιστήμη της φωτογραμμετρίας όσο και στην όραση υπολογιστών (computer vision), που βρίσκει πληθώρα εφαρμογών όπως η τεκμηρίωση μνημείων, η δημιουργία 3Δ χαρτών και ψηφιακών διδύμων digital twins, η αυτόματη πλοήγηση, ο εντοπισμός, η εικονική και επαυξημένη πραγματικότητα, μεταξύ άλλων. Τις τελευταίες δεκαετίες, έχουν αναπτυχθεί διάφοροι αισθητήρες για το σκοπό αυτό, όπως για παράδειγμα οι σαρωτές laser (laser scanners), ωστόσο, η 3Δ ανακατασκευή από ψηφιακές εικόνες (image-based 3D reconstruction) παραμένει μια αποτελεσματική, ευρέως διαδεδομένη και χαμηλού κόστους μέθοδος. Για το λόγο αυτό, τις τελευταίες δεκαετίες έχουν προταθεί πολλοί αλγόριθμοι που στοχεύουν στη δημιουργία 3Δ μοντέλων από επικαλυπτόμενες εικόνες, είτε με τη μορφή νεφών σημείων είτε ως όγκοι ή επιφάνειες. Παρ'όλα αυτά, η πλήρης και ακριβής ανακατασκευή παραμένει ένα μερικώς άλυτο πρόβλημα, ειδικά στην περίπτωση που τα προς αποτύπωση αντικείμενα περιλαμβάνουν επιφάνειες με ιδιαίτερα γεωμετρικά, ραδιομετρικά ή φυσικά χαρακτηριστικά και οι συνθήκες λήψης (λ.χ. φωτισμός, γωνίες λήψης) δεν είναι ιδανικές. Στην 3Δ ανακατασκευή διακρίνονται δύο βασικές διαδικασίες, η Δομή από Κίνηση (Structure from Motion, SfM) και η Πολυεικονική Συνταύτιση (Multiple View Stereo, MVS). Οι αλγόριθμοι SfM προτείνουν λύσεις για την αυτόματη εξαγωγή και συνταύτιση χαρακτηριστικών σημείων μεταξύ των εικόνων με ταυτόχρονη ανακατασκευή τους στον 3Δ χώρο, με σκοπό τον υπολογισμό των προσανατολισμών των εικόνων, που σε αυτή την περίπτωση μπορεί να ειπωθεί ότι ισοδυναμεί με τον αγγλικό όρο camera poses. Η διαδικασία MVS, από την άλλη, αφορά στην πυκνή ανακατασκευή του 3Δ χώρου και συνεπώς στη δημιουργία 3Δ μοντέλων, συνταυτίζοντας, αν είναι εφικτό, κάθε εικονοστοιχείο (pixel) και λαμβάνοντας υπ'όψιν όλες τις επικαλυπτόμενες εικόνες.

Ταυτόχρονα με την 3Δ ανακατασκευή, το ενδιαφέρον της επιστημονικής κοινότητας έχει πρόσφατα προσελκύσει και η «κατανόηση» του 3Δ χώρου (3D scene understanding) είτε μέσω εικόνων είτε απευθείας σε 3Δ μοντέλα. Μάλιστα, τα τελευταία χρόνια, η 3Δ ανακατασκευή συχνά συνδέεται στην βιβλιογραφία με το scene understanding, και πιο συγκεκριμένα με τη σημασιολογική κατάτμηση (semantic segmentation) και ανίχνευση (object detection) των αντικειμένων και αντιμετωπίζονται σε κοινό πλαίσιο. Μελέτες δείχνουν ότι η σημασιολογική συσχέτιση των αντικειμένων του χώρου μπορεί να βοηθήσει την 3Δ ανακατασκευή σε περιπτώσεις που η γεωμετρική και ραδιομετρική πληροφορία δεν είναι αρκετή. Ακόμα, μια τέτοια συσχέτιση, μπορεί να δημιουργήσει, τελικά, 3Δ μοντέλα που θα εμπεριέχουν αυτή τη σημασιολογική πληροφορία και θα ανοίγει το δρόμο, με αυτό τον τρόπο, προς το 3D scene understanding. Κατά τη διάρκεια της τελευταίας δεκαετίας, έχουν γίνει σημαντικά βήματα σε αυτό τον ερευνητικό τομέα, καθώς η τεχνολογική πρόοδος και η ολοένα αυξανόμενη υπολογιστική ισχύς έχει επιφέρει σημαντικές εξελίξεις στην μηχανική μάθηση (machine learning), και πιο συγκεκριμένα στη δημιουργία μοντέλων βαθιάς μηχανικής μάθησης (deep learning). Ωστόσο, η εύρεση ενός αλγορίθμου που θα είναι εξίσου αποτελεσματικός σε διαφορετικά πεδία εφαρμογών (domains) είναι ακόμα μια ανοικτή πρόκληση.

Η παρούσα διατριβή μελετά διεξοδικά το πρόβλημα της πολυεικονικής (multi-view) συνταύτισης και κατ΄ επέκταση, του υπολογισμού του βάθους και της ανακατασκευής (depth estimation and reconstruction) με μεθόδους MVS. Παράλληλα, πραγματεύεται τη σημασιολογική κατάτμηση εικόνων και προτείνει πρωτότυπες μεθόδους για την αξιοποίηση αυτής της πληροφορίας κατά τη διάρκεια της πολυεικονεικής ανακατασκευής. Όλες οι μέθοδοι που περιγράφονται στη διατριβή στηρίζονται σε open-source αλγορίθμους, διευρύνοντας τις λειτουργίες τους και επεκτείνοντας τις δυνατότητές τους. Το εισαγωγικό κεφάλαιο (Κεφάλαιο 1) περιγράφει συνοπτικά το πλαίσιο της διατριβής και τα κίνητρα για την συγκεκριμένη έρευνα και παρουσιάζει τους επιμέρους στόχους και τις πρωτοτυπίες της.

Σχετικά με τον υπολογισμό του βάθους, αρχικά, γίνεται εκτενής βιβλιογραφική έρευνα στο πεδίο της συνταύτισης εικόνων, ξεκινώντας από τη διεικονική (two-view) περίπτωση, δηλαδή ενός και μόνο στερεοζεύγους, και επεκτείνοντας το πρόβλημα στην πολυεικονική (Κεφάλαιο 2). Έτσι, περιγράφονται βασικές θεωρητικές έννοιες και παρουσιάζονται συνοπτικά οι πιο διαδεδομένοι αλγόριθμοι, και ταυτόχρονα εντοπίζονται τα βασικά τους όρια και οι περιορισμοί. Αυτή η, όσο το δυνατόν πληρέστερη, βιβλιογραφική έρευνα αποτελεί ταυτόχρονα και έναν από τους στόχους της διατριβής, με σκοπό την κατανόηση του προβλήματος, των ιδιαιτεροτήτων και των διαφορών μεταξύ των αλγορίθμων, τόσο των παλαιοτέρων όσο και των σύγχρονων, που βασίζονται είτε σε συμβατικές μεθόδους είτε σε μεθόδους βαθιάς μηχανικής μάθησης. Στην πραγματικότητα, το πρόβλημα της πολυεικονικής συνταύτισης στηρίζεται στις ίδιες θεμελιώδεις αρχές με το αντίστοιχο της διεικονικής. Ένα βασικό κοινό χαρακτηριστικό και στις δύο περιπτώσεις είναι αναμφίβολα η καθαυτή συνταύτιση των pixel, που συνήθως πραγματοποιείται με βάση την οπτική εμφάνιση (visual appearance) ενός παραθύρου συγκεκριμένου μεγέθους γύρω από το προς εξέταση pixel στην εικόνα αναφοράς και του αντιστοίχου στην εικόνα αναζήτησης. Για τον υπολογισμό της ομοιότητας μεταξύ των δύο παραθύρων χρησιμοποιούνται παραμετρικά ή μη παραμετρικά μέτρα ομοιότητας με βάση κυρίως τη ραδιομετρία (photometric consistency), όπως το άθροισμα των απόλυτων διαφορών (Sum of Absolute Differences - SAD), το άθροισμα των τετραγώνων των απολύτων διαφορών (Sum of Squared Differences - SSD), ο συντελεστής συσχέτισης (Normalized Cross Correlation - NCC), μεταξύ άλλων. Αυτά τα μέτρα ομοιότητας συνήθως είναι αρκετά αποτελεσματικά σε επιφάνειες με υφή και γενικά near-Lambertian περιοχές, αλλά αδυνατούν να βρουν αξιόπιστες λύσεις σε περιοχές με ομοιογενή υφή ή ανακλαστικές και διαφανείς επιφάνειες λόγω των πολλαπλών τοπικών ελαχίστων που οδηγούν σε ασάφειες συνταύτισης (matching ambiguities). Τις τελευταίες δεκαετίες, για τη διαδικασία συνταύτισης με βάση αυτά τα μέτρα, έχουν αναπτυχθεί πολλοί αλγόριθμοι τοπικής συνταύτισης (local algorithms), καθολικής συνταύτισης (global algorithms) όπως και ημί-καθολικής συνταύτισης (semi-global). Η κάθε μέθοδος έχει πλεονεκτήματα και μειονεκτήματα, όμως όλες κάνουν συγκεκριμένες παραδοχές για τοπική ομαλότητα, και έτσι, συνήθως αδυνατούν να υπολογίσουν σωστά το βάθος σε περιπτώσεις κεκλιμένων επιφανειών (fronto-parallel surfaces). Παράλληλα, οι περισσότερες μέθοδοι στηρίζονται στη δημιουργία της επονομαζόμενης «εικόνας του χώρου των ψηφιακών παραλλάξεων» (Disparity Space Image, DSI), για τον έλεγχο όλων των πιθανών τιμών παράλλαξης (βάθους) για κάθε pixel, μια διαδικασία που απαιτεί μεγάλη υπολογιστική μνήμη.

Όμως, η πολυεικονική ανακατασκευή δεν μπορεί να θεωρηθεί απλή γενίκευση της διεικονικής διαδικασίας, καθώς υπάρχουν κάποιες σημαντικές διαφορές. Η περίπτωση ενός και μόνο ζεύγους εικόνων, συνήθως περιλαμβάνει γεωμετρίες λήψης με μικρό μήκος βάσης, ενώ η πολυεικονική περίπτωση στις σύγχρονες εφαρμογές συχνά αφορά σε πιο άτακτη γεωμετρία με μεγάλες διαφορές στη γωνία λήψης ή ακόμα και στην κλίμακα. Οι αλγόριθμοι MVS είναι λοιπόν σχεδιασμένοι να αντιμετωπίζουν τέτοιες περιπτώσεις και απαιτούν την προσεκτική επιλογή των καλύτερων δυνατών πιθανών ζευγών, μια διαδικασία που στη διεθνή βιβλιογραφία αναφέρεται και ως visibility reasoning. Ακόμη, η ύπαρξη πλεοναζουσών παρατηρήσεων, καθώς το ίδιο σημείο του χώρου συνήθως προβάλλεται σε περισσότερες από δύο εικόνες, τείνει να βελτιώνει την ακρίβεια υπολογισμού του βάθους. Ταυτόχρονα, οι σύγχρονοι αλγόριθμοι MVS, έχουν τη δυνατότητα να υπολογίζουν το βάθος χωρίς να έχει προηγηθεί επιπολική επανασύσταση των εικόνων, μια διαδικασία που συνήθως αποτελεί αναγκαία προϋπόθεση στη διεικονική περίπτωση.

Μέσω αυτής της μελέτης, αφού εξετάστηκαν διεξοδικά οι διαθέσιμοι αλγόριθμοι και αξιολογήθηκαν τα πλεονεκτήματα και τα μειονεκτήματά τους, επιλέχθηκε να χρησιμοποιηθεί ο αλγόριθμος PatchMatch, ως μια σύγχρονη και αποτελεσματική λύση που στη διεθνή βιβλιογραφία θεωρείται state-of-the-art και εφαρμόζεται σε πολλές βιβλιοθήκες ελεύθερου λογισμικού (open-source libraries) (Κεφάλαιο 3). Ο αλγόριθμος PatchMatch βασίζεται στην πολύ απλή παραδοχή της τοπικής συνοχής (local coherency) της εικόνας και έχει αποδειχθεί ιδιαίτερα αποτελεσματικός τόσο σε ακρίβεια όσο και σε χρόνο υπολογισμού τα τελευταία χρόνια. Ξεκινώντας από τυχαίες τιμές, υπολογίζει το βάθος και τον προσανατολισμό (normal) του κάθε σημείου στο χώρο χρησιμοποιώντας τοπικά εφαπτόμενα παράθυρα (local tangent windows). Σύμφωνα με την παραδοχή της τοπικής συνοχής, γειτονικά pixel θα τείνουν να διέπονται από ομαλές μεταβολές βάθους, και έτσι, με βάση κάποιο μοτίβο μετάδοσης (propagation scheme), αξιόπιστες τιμές, τόσο βάθους όσο και προσανατολισμού (normal) διαδίδονται στα γειτονικά pixel. Αυτή η διαδικασία, αν και απλή, έχει αποδειχθεί ότι λειτουργεί αποτελεσματικά και σε πρακτικές εφαρμογές, ειδικά με εικόνες υψηλής ανάλυσης. Ο αλγόριθμος PatchMatch είναι εξ ορισμού απαλλαγμένος από το fronto-parallel bias λόγω της χρήσης των local tangent planes και αποφεύγει τη δημιουργία DSI, δύο χαρακτηριστικά που τον καθιστούν ιδιαίτερα ανταγωνιστικό ως προς τις υπόλοιπες μεθόδους που αναφέρθηκαν παραπάνω.

Σχετικά με τη σημασιολογική κατάτμηση, αρχικά μελετήθηκε η σχετική βιβλιογραφία, δίνοντας ιδαίτερη βάση στους σύγχρονους αλγορίθμους βαθιάς μηχανικής μάθησης. Το θεωρητικό υπόβαθρο γύρω από τα νευρωνικά δίκτυα (neural networks) παρουσιάζεται συνοπτικά, δίνοντας έμφαση στις αρχιτεκτονικές που είναι σχεδιασμένες για σημασιολογική κατάτμηση σε εικόνες και στα διαθέσιμα benchmark σύνολα δεδομένων (datasets) για φωτογραμμετρικές εφαρμογές και εφαρμογές όρασης υπολογιστών. Κατόπιν, εντοπίζοντας την έλλειψη στη σύγχρονη βιβλιογραφία ενός benchmark υψηλής ανάλυσης για σημασιολογική κατάτμηση σε προσόψεις ιστορικών κτηρίων, προτείνεται ένα νέο benchmark dataset, το *3DOM Semantic Facade*, που περιέχει 227 εικόνες υψηλής ανάλυσης με αντίστοιχους αληθείς σημασιολογικά κατατετμημένους χάρτες (segmentation maps). Στο *3DOM Semantic Facade* διαχωρίζονται οι εξής κατηγορίες (classes): τοίχος, παράθυρο, πόρτα, ουρανός και εμπόδιο. Στη συνέχεια, περιγράφεται μια state-of-the-art διαδικασία βασισμένη στην αρχιτεκτονική U-Net για την εκπαίδευση ενός αλγορίθμου που μπορεί να προβλέπει την κατηγορία του κάθε pixel ακόμη και σε άγνωστα (unseen) δεδομένα. Η προτεινόμενη διαδικασία αξιολογείται ως προς τα αληθή δεδομένα και επιτυγχάνει υψηλά ποσοστά επιτυχίας (precision, recall, IoU score etc.). Παράλληλα, αναπτύσσεται και προτείνεται μια μέθοδος για την αξιοποίηση των κατατετμημένων εικόνων στην 3Δ ανακατασκευή (semantic photogrammetry). Πράγματι, χρησιμοποιώντας, ταυτόχρονα με τις πραγματικές εικόνες και τις αντίστοιχες κατατετμημένες, δίνεται η δυνατότητα να υπολογιστούν απευθείας κατατετμημένα πυκνά νέφη σημείων, δηλαδή νέφη τα οποία εμπεριέχουν, μαζί με τη γεωμετρία, και τη σημασιολογική πληροφορία για τον 3Δ χώρο. Αυτή η σημασιολογική πληροφορία μπορεί να χρησιμοποιηθεί και για την επιλεκτική ανακατασκευή μόνο των περιοχών ενδιαφέροντος, αποκλείοντας, λ.χ. από τη διαδικασία ανακατασκευής όλα τα pixel που ανήκουν στην κατηγορία «ουρανός». Η μέθοδος αυτή βασίστηκε στη βιβλιοθήκη ανοικτού κώδικα OpenMVS, αναπτύσσοντας και ενσωματώνοντας σε αυτή μία επιπλέον λειτουργία για το σκοπό αυτό.

Αναπόφευκτα, προκύπτει ο προβληματισμός γύρω από το εάν και πώς αυτή η σημασιολογική πληροφορία μπορεί να υποβοηθήσει τον καθαυτό υπολογισμό του βάθους, ιδιαίτερα στις προβληματικές περιοχές που αναφέρθηκαν παραπάνω. Πράγματι, τόσο κατά τη διάρκεια της βιβλιογραφικής έρευνας, αλλά και των πρώτων πειραματικών εφαρμογών, διαπιστώθηκε ότι ένα από τα πιο σημαντικά και άλυτα προβλήματα στον υπολογισμό του βάθους είναι η ύπαρξη ασαφειών στη συνταύτιση (matching ambiguities). Το πρόβλημα αυτό, παρουσιάζεται πολύ συχνά σε πρακτικές εφαρμογές, και η πλειοψηφία των state-of-the-art μεθόδων δεν καταφέρνει να υπολογίσει αξιόπιστες τιμές βάθους σε περιοχές χωρίς υφή ή ανακλαστικές επιφάνειες, περιπτώσεις που πολύ συχνά συναντώνται σε εφαρμογές επίγειων αποτυπώσεων, εσωτερικών και εξωτερικών χώρων. Ο αλγόριθμος PatchMatch, που υιοθετείται εδώ για την πολυεικονική ανακατασκευή, δε διαφέρει από τις υπόλοιπες state-of-the-art μεθόδους σε αυτό, καθώς ο υπολογισμός του κόστους στηρίζεται και εδώ σε συνήθη visual appearance metrics. Για το λόγο αυτό, στο πλαίσιο αυτής της διδακτορικής διατριβής, προτείνεται μια πρωτότυπη μέθοδος κατά την οποία, η σημασιολογική πληροφορία χρησιμοποιείται για τον υπολογισμό γεωμετρικών δεσμεύσεων, με σκοπό να αντιμετωπιστούν τα matching ambiguities και να υπολογιστούν πιο αξιόπιστες τιμές βάθους και προσανατολισμού σε αυτές τις προβληματικές περιοχές (Κεφάλαιο 5). Πιο συγκεκριμένα, επιλέγονται οι σημασιολογικές κατηγορίες που είναι πιο πιθανό να περιέχουν επιφάνειες που μπορούν να περιγραφούν με γεωμετρικά σχήματα στο χώρο, όπως για παράδειγμα η κατηγορία «τοίχος» είναι πολύ πιθανό να μπορεί να περιγραφεί με 3Δ επίπεδα. Λαμβάνοντας υπ'όψιν μόνο τις προβολές των pixel αυτών των κατηγοριών στον 3Δ χώρο, ανιχνεύονται επίπεδα με την μέθοδο Efficient RANSAC. Ύστερα, προτείνεται μια νέα, σύνθετη συνάρτηση κόστους, που λαμβάνει υπ'όψιν τόσο την υφιστάμενη πληροφορία για τα επίπεδα, όσο και την τοπική υφή, μαζί με το σύνηθες κόστος συνταύτισης (photometric cost). Οι πειραματικές εφαρμογές τόσο σε benchmark datasets (*ETH3D*) όσο και σε custom δεδομένα, αποδεικνύουν το ότι η προτεινόμενη μέθοδος (semantic PatchMatch) υπολογίζει πιο αξιόπιστες τιμές βάθους στις προβληματικές περιοχές, χωρίς να υπολείπεται ακριβείας στις περιοχές με πλούσια υφή. Η μέθοδος συγκρίνεται με την μέθοδο αναφοράς OpenMVS καθώς και με άλλους state-of-the-art αλγορίθμους και παρουσιάζει ανταγωνιστική απόδοση.

Παρόλο που η παραπάνω μέθοδος αποδείχθηκε αποτελεσματική, μπορεί να εφαρμοστεί μόνο σε περιπτώσεις που η σημασιολογική πληροφορία είτε είναι εκ των προτέρων διαθέσιμη είτε μπορεί να εξαχθεί εύκολα. Αυτό είναι αρκετά πιθανό για κάποιες συγκεκριμένες εφαρμογές, ιδαίτερα για εκείνες για τις οποίες υπάρχει πλήθος διαθέσιμων δεδομένων για εκπαίδευση. Με σκοπό τη γενίκευση της μεθόδου και σε άλλες εφαρμογές, προτείνεται μια δεύτερη, πρωτότυπη και καθολική μέθοδος που είναι ανεξάρτητη από τη σημασιολογική πληροφορία, αλλά στηρίζεται σε υποθέσεις για την τοπική δομή (local structure assumptions) (Κεφάλαιο 6). Πιο συγκεκριμένα, υπολογίζονται quadtree δομές πάνω στις εικόνες. Η εικόνα χωρίζεται, δηλαδή, σε υποσύνολα (block) με βάση την τοπική υφή. Τα block αυτά διαφέρουν στο μέγεθος, καθώς μια περιοχή χωρίς υφή περιγράφεται από ένα μεγάλο block ενώ περιοχές με πλούσια υφή θα είναι «κατακερματισμένες» σε πολλά μικρά quadtree block. Παρόμοια με πριν, ανιχνεύονται επίπεδα στον 3Δ χώρο με τη μέθοδο Efficient RANSAC και προτείνεται τα τοπικά επίπεδα να μεταδίδονται με βάση τα quadtree block (quadtree-guided plane propagation), κάνοντας την παραδοχή ότι γειτονικά block που έχουν παρόμοια μέση τιμή χρώματος είναι πιθανό να ανήκουν στο ίδιο τοπικό επίπεδο. Και εδώ, χρησιμοποιείται η σύνθετη συνάρτηση κόστους που προτάθηκε στην προηγούμενη μέθοδο με σκοπό να υποβοηθήσει την μετάδοση αξιόπιστων τιμών βάθους και προσανατολισμού κατά τη διάρκεια του PatchMatch. Η προτεινόμενη μέθοδος αξιολογείται σε όλες τις σκηνές (scenes) του benchmark dataset (*ETH3D*) (13 training και 12 testing) και σε custom δεδομένα, όπου παρουσιάζει σταθερή βελτίωση της μεθόδου αναφοράς, και ανταγωνιστική απόδοση σε σχέση με άλλες state-of-the-art μεθόδους, τόσο συμβατικές όσο και μηχανικής μάθησης.

Οι παραπάνω μέθοδοι προτείνουν καινοτόμες λύσεις και state-of-the-art στρατηγικές για την εκμετάλλευση της σημασιολογικής πληροφορίας στην 3Δ ανακατασκευή και επεκτείνουν το πεδίο εφαρμογών ακόμα και σε περιπτώσεις που αυτή η πληροφορία δεν είναι διαθέσιμη. Εφαρμόστηκαν τόσο σε benchmark όσο και σε custom δεδομένα πραγματικών φωτογραμμετρικών εφαρμογών με υποσχόμενα αποτελέσματα. Αποτελούν απλές, αλλά αποτελεσματικές συμβατικές λύσεις που μπορούν να εφαρμοστούν σε μεγάλο εύρος εφαρμογών. Η τελευταία γενικευμένη μέθοδος δεν εξαρτάται από το πεδίο εφαρμογής όπως συχνά συμβαίνει με τις μεθόδους μηχανικής μάθησης, είναι δηλαδή domain-independent. Οι ανωτέρω αλγόριθμοι αναπτύχθηκαν χρησιμοποιώντας λύσεις ανοιχτού κώδικα και ενσωματώθηκαν στην βιβλιοθήκη OpenMVS και είναι διαθέσιμες στην επιστημονική κοινότητα για περαιτέρω μελέτη και βελτιώσεις. Τα γενικά συμπεράσματα και οι μελλοντικές προεκτάσεις παρουσιάζονται στο Κεφάλαιο 7, μαζί με το γενικότερο πλαίσιο της έρευνας και τις σχετικές δημοσιεύσεις της συγγραφέως.

Contents

List of Figures

List of Tables

Acronyms

AD Absolute differences. 22, 28, 37

ANN Artificial Neural Network. 79–81

BF Brute Force. 29

BP Belief Propagation. 36, 64, 65

CNN Convolutional Neural Network. 7, 50–54, 74, 75, 80–82, 94

CRF Conditional Random Fields. 50–53, 56, 83, 84, 87, 113

CT Census Transform. 25

DLT Direct Linear Transform. 17

DOF Degrees of Freedom. 17, 19, 21

DoG Difference of Gaussians. 30

DSI Disparity Space Image. 26–28, 63

DSM Digital Surface Model. 86

EM Expectation Maximation. 65

FCN Fully Convolutional Network. 83

GAN Generative Adversarial Network. 84, 93

GCP Ground Control Points. 47, 104

GPU Graphics Processing Unit. 37, 63, 67, 95, 157

GRU Gated Recurrent Units. 53

HDR High Dynamic Range. 165

LSM Least Squares Matching. 32, 35

LSTM Long Short Term Memory. 53

MI Mutual Information. 26, 33, 40

MRF Markov Random Field. 29, 31, 33, 41, 44, 51, 52, 55, 65, 66, 137, 166

MVS Multiple View Stereo. 3, 5–12, 25, 26, 38–40, 42, 44–50, 52, 53, 55, 57, 59–61, 65–71, 75, 79, 86, 104, 108, 111–114, 123, 125, 126, 128, 130, 132, 133, 135, 137, 138, 144–146, 148, 151, 152, 155, 159–162, 166

NCC Normalized Cross Correlation. 24, 25, 28, 37, 40, 57, 70, 73, 137

NN Nearest Neighbor. 28, 62, 63

ORB Oriented FAST and Rotated BRIEF. 4, 26

PDE Partial Differential Equation. 29, 42

PnP Perspective-n-Point. 4

PSV Plane Sweep Volume. 52, 62

R-CNN Region-based Convolutional Neural Network. 83

RANSAC Random Sample Consensus. 4, 10, 19, 112, 117–119, 131, 132, 136, 139–141, 147, 155, 156, 160, 163, 164

RNN Recurrent Neural Network. 53, 84

RT Rank Transform. 24, 25

SAD Sum of Absolute Differences. 23, 25, 37, 40

SD Squared Differences. 23

SfM Structure from Motion. 3–6, 26, 39, 46–48, 59, 66, 72, 79, 104, 114, 115, 137, 138, 146, 159

SGM Semi Global Matching. 33–37, 45, 51, 75, 76, 166

SIFT Scale Invariant Feature Transform. 4, 26

SSD Sum of Squared Differences. 24

SURF Speeded Up Robust Features. 4, 26

TSDF Truncated Signed Distance Field. 113

UAV Unmanned Aerial Vehicle. 165

WTA Winner Takes All. 27, 28, 35, 44

ZNCC Zero-mean Normalized Cross Correlation. 24, 70, 73, 115, 121, 122, 137

Chapter 1

Introduction

Obtaining 3D representations of the physical world is a fundamental research topic with numerous application fields spanning from mapping, autonomous navigation, and localization to cultural heritage documentation and augmented or virtual reality. Indeed, various research fields such as photogrammetry, computer vision, robotics, and other engineering sectors aim to obtain realistic, precise, complete, and visually pleasing 3D representations of scenes. A plethora of specially designed sensors for this scope, both active and passive, have been developed in recent years, such as specially designed camera systems, laser and structured light scanners, which, mounted on diverse platforms, can be used to obtain data of various scales, ranging from close-range laboratory measurements to airborne acquisitions.

3D reconstruction of scenes using multiple, overlapping images, often called image-based modeling, is currently one of the most widely used and cost-effective techniques to obtain such 3D representations. It has been a well-studied problem and has seen tremendous evolution in recent years. Given the popularity of camera sensors and the recently developed user-friendly software implementations, image-based 3D reconstruction enables potentially everyone to generate realistic 3D digital replicas of scenes and objects.

Image-based 3D reconstruction aims to recover the structure of scenes given overlapping projections of the 3D space on 2D images as observations. It typically relies on correspondence search based on visual appearance between the images and subsequent depth estimation and reconstruction of the matches. In principle, it is inspired by the human vision system, observing the same scene from two different viewing points, with the camera being the hardware equivalent of the human eye. Humans are capable of perceiving effortlessly spatial information, depth, and semantic cues for every observed scene, even if seen for the first time, and subsequently employ knowledge interpretation. However, this straightforward cognitive process of the human brain cannot be easily abstracted in a way interpretable by computers.

Hence, extensive research has been conducted in the last decades for robust 3D reconstruction from RGB images, in both photogrammetry and computer vision, independently in the beginning, yet continuously converging recently. Researchers have been trying first to understand the underlying theoretical geometric principles in-depth and then implement robust algorithms to enable such a process. Undoubtedly, some of the greatest problems have been undertaken accordingly, and algorithms are now mature enough to provide reliable reconstructions even in extreme scenarios, e.g., by exploiting the vast amount of unorganized images available in the cloud and thus generating quality 3D reconstructions using crowd-sourced images. Nonetheless, image-based 3D reconstruction is an inherently hard and, by definition, ill-posed problem due to the information loss during the inverse mapping from 3D to 2D. Indeed, various 3D scenes may result in the same set of images; for instance, objects of arbitrary scales may be projected in the same way on the images. Prior knowledge about the scene and the image acquisition would be needed to fully pose the problem. Therefore, it can be said that despite the great evolution of the methods, so far, among the developed solutions, there is no general-purpose system, either based on conventional or learning-based methods, that can efficiently and simultaneously tackle all challenges in 3D reconstruction for real-world scenarios.

In fact, the problem of geometric 3D reconstruction is highly related to the broader field of 3D scene understanding, i.e., the analysis of the important features of the scene and its semantic reasoning using images or directly in the 3D space. Extensive research has been recently conducted in computer vision regarding scene understanding, mainly due to the advancements in artificial intelligence and particularly in semantic segmentation using deep learning. Semantic segmentation refers to assigning a semantically meaningful label (i.e., class) to each pixel or 3D point of the scene, and numerous robust algorithms have been recently introduced to the literature. Such high-level scene priors can potentially support image-based 3D reconstruction, especially in cases where plain geometric and visual appearance information is not enough. Moreover, semantically segmented 3D models can facilitate scene understanding; nonetheless, despite the recent advancements, semantic reasoning directly in the 3D space is still a non-trivial problem because of the computational complexity and the limited generalization ability of the algorithms.

In the context of this doctoral dissertation, multi-view 3D reconstruction is comprehensively discussed, acknowledging the open challenges and proposing efficient and easy-to-implement methodologies to address them based on the recent advances in the field. Advanced scene priors, semantic or structure-based, are exploited in this direction.

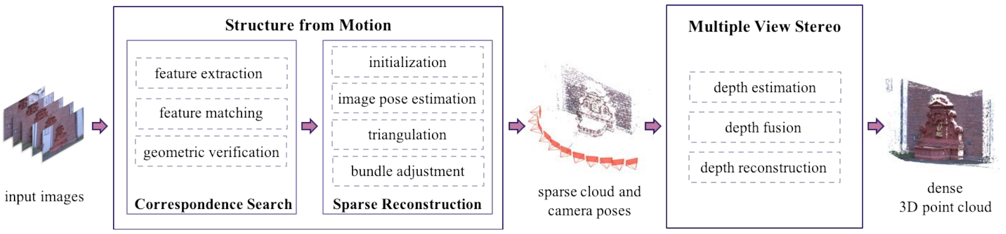


Figure 1.1: **Overview of the image-based 3D reconstruction pipeline.** Incremental SfM and MVS with depth fusion. Inspired by: [Schönberger and Frahm, 2016], data: Fountain-P11 dataset, [Strecha et al., 2008].

1.1 Overview of image-based 3D reconstruction

Correspondence search and depth estimation have been thoroughly studied for the stereo scenario (also called binocular or two-view), i.e., given only two images of the same scene [Hannah, 1974; Yang et al., 1993; Kanade and Okutomi, 1994]. Consequently, the earlier approaches for 3D reconstruction referred to the two-view scenario, e.g., [Longuet-Higgins, 1981]. Later on, the benefits of the redundancy in the multi-view scenario have also been exploited for efficient and robust 3D reconstruction [Szeliski and Kang, 1994; Beardsley et al., 1997], exploring also self-calibration [Fitzgibbon and Zisserman, 1998; Pollefeys, 1999]. This progress led to the development of algorithms able to process a massive amount of unstructured data harvested from the internet [Snavely et al., 2006; Frahm et al., 2010; Agarwal et al., 2011; Wu, 2013]. In the past decade, several non-commercial or open-source solutions have been released to the public such as Bundler [Snavely et al., 2006], VisualSfM [Wu et al., 2011; Wu, 2013], MVE [Fuhrmann et al., 2014], OpenMVG [Moulon et al., 2016], Colmap [Schönberger and Frahm, 2016] along with commercial software implementations.

Efficient dense 3D reconstruction of rigid scenes can be divided into two well-established workflows, namely Structure from Motion (SfM) and Multiple View Stereo (MVS) (Figure 1.1). SfM searches the best image pairs based on the network geometry and the scene structure and performs feature detection, description, and matching among the images. Using abundant features and epipolar geometry constraints, the relative camera poses, meaning the rotation and position in the 3D space, equivalent to the camera external and internal parameters, can be estimated along with the projection of these points in the 3D space and be jointly optimized using bundle adjustment. Given the camera poses and calibration, MVS techniques aim to reconstruct, if possible, pixel by pixel correspondences in the 3D space resulting in richer scene representations, i.e., dense point clouds or 3D surfaces/meshes. Each sub-step of this pipeline is undoubtedly a standalone research field, and scientists are working towards optimizing each of them to enhance the robustness of the results.

1.1.1 Structure from Motion

Equivalent to the traditional image orientation process for an image block in photogrammetry, SfM mainly aims to calculate the camera poses in the 3D space, while recovering also a sparse scene structure. It tries to solve the correspondence problem and identify invariant features across the potential overlapping images, resulting in a scene graph to express the relationship between images and scene points. Feature detection and description are commonly performed with algorithms such as SIFT [Lowe, 2004], SURF [Bay et al., 2006], ORB [Rublee et al., 2011], and AKAZE [Alcantarilla and Solutions, 2011], or more recently with learned descriptors [DeTone et al., 2018; Ono et al., 2018]. Once the features are extracted, feature matching can be performed naively by searching all potential correspondences exhaustively, following a standard brute-force approach or based on techniques such as kd-trees [Muja and Lowe, 2009] and cascade hashing [Cheng et al., 2014]. Good potential image pairs are considered the ones that have a sufficient amount of common features between the two images. For image matching, sophisticated solutions are based on vocabulary trees, e.g., [Nister and Stewenius, 2006] and global image descriptors to identify the most visually similar images on a global level and even enable efficient processing of large-scale datasets [Agarwal et al., 2011; Wu, 2013; Moulon et al., 2016]. Typically a geometric verification step is needed to evaluate the putative feature matches and filter the non-overlapping image pairs based on a geometric multi-view model, e.g., the homography or the fundamental matrix models [Hartley and Zisserman, 2003] and robust fitting methods such as RANSAC [Fischler and Bolles, 1981]. A valid transformation is sought to map a sufficient amount of features between the images.

Images passing the geometric verification step are inserted into the scene graph; given this graph, the reconstruction step can be performed in an incremental, i.e., initializing from a two-view [Moulon et al., 2012; Wu, 2013; Schönberger and Frahm, 2016] or global, i.e., as a joint optimization, fashion [Moulon et al., 2013; Sweeney et al., 2015b]. Incremental reconstruction tends to be more robust in practical applications [Stathopoulou et al., 2019] and is therefore often preferred over global optimization [Schönberger and Frahm, 2016]. In an incremental paradigm, starting from an initial pair, images are registered repeatedly to the scene solving the Perspective-n-Point (PnP) problem using 2D-3D correspondences of the previously registered images and calculating thus the camera poses, i.e., position in the 3D space and calibration, of the new images; minimal solvers [Lepetit et al., 2009] RANSAC-based approaches are typically used here for outlier removal. Once the camera poses are recovered, point triangulation [Hartley and Zisserman, 2003] takes place, reconstructing the points visible by each view. New images should have common points in the 3D space with the already existing ones. New points are triangulated from each newly registered image. Finally, bundle adjustment [Triggs et al., 1999], performing linear refinement and minimizing the reprojection error of the reconstructed 3D points on the images and is a core module of SfM pipelines for extra robustness, is commonly performed by the widely-used Ceres solver [Agarwal et al., 2012]. For an extensive review on the SfM algorithms, the reader is referred to [Schönberger and Frahm, 2016].

1.1.2 Multiple view stereo

SfM typically yields an abstract representation of the scene consisting of few, high-fidelity 3D points along with the camera poses. Multiple (or multi-) view stereo (MVS) algorithms aim to generate a rich, dense 3D model of the scene in the form of a dense point cloud or a triangulated mesh. A rough categorization of the existing methods would cluster them into two large groups; some works parameterize the problem in the image space, recovering the depth for every pixel, e.g., [Goesele et al., 2007; Gallup et al., 2007; Campbell et al., 2008]. Other works perform directly in the scene space, e.g., [Furukawa and Ponce, 2009; Häne et al., 2013; Ulusoy et al., 2015; Zach, 2008]. That being said, this section is introductory and does not aim to present a comprehensive taxonomy of the methods; the reader is referred to Chapter 2 (Section 2.5) of this dissertation for a detailed review of the algorithms and the respective challenges and limitations.

In a typical MVS pipeline, the robust estimations for the camera poses along with the sparse points obtained during the reconstruction step are used as input. During this process, the depth of, if possible, every pixel of the scene is to be calculated. In the two-view scenario, epipolar geometry constraints simplify the correspondence search by restricting the search space along one dimension. Several methods have been developed for solving this correspondence search problem, either local [Scharstein, 1994; Hosni et al., 2012; Bleyer et al., 2011], global [Faugeras and Keriven, 1998; Strecha et al., 2004] or hybrid semi-global [Hirschmuller, 2008] methods. Nonetheless, multi-view reconstruction is a more complicated problem than the classic two-view approach due to the ray redundancy resulting from the multiple observations and the strong occlusions. Many specially designed algorithms have been developed to efficiently solve multi-view reconstruction in recent years, achieving impressive results [Strecha et al., 2006; Galliani et al., 2015; Schönberger et al., 2016]. In the last decade, the PatchMatch algorithm [Bleyer et al., 2011] has been established as the standard MVS approach for its robustness, efficiency, and scalability, and the most widely-used implementations are based on it [Schönberger et al., 2016; Cernea, 2020; Xu and Tao, 2019]. The PatchMatch algorithm, being a core part of this thesis, is extensively discussed in Chapter 3.

1.2 Motivation and current challenges

Despite the widespread evolution of the algorithms, yielding complete, accurate, and aesthetically pleasing 3D representations of a scene remains an open issue in real-world and large-scale applications. Although SfM pipelines typically give robust solutions for pose estimation and sparse clouds, especially for proper image networks acquired for the scope of 3D reconstruction, dense scene reconstruction is still a great challenge. Indeed, finding dense pixel correspondences and robustly reconstructing the depth in the 3D space depends on a sequence of variables; thus, finding a solution that will simultaneously undertake all challenges is not trivial. Research has been active in the field of depth estimation in both stereo and multi-view scenarios in the last decades, with MVS being particularly interesting also for practical, real-world applications in photogrammetry and computer vision. Commonly, such image-based 3D reconstruction projects have high standards in completeness, accuracy, and overall visual representation of the results.

MVS methods certainly provide more robust depth estimations than the two-view scenario, given the redundant observations from the overlapping images, resulting in generally robust reconstructions. At the same time, occlusions are mostly handled properly based on several developed strategies. However, scene and image acquisition properties can be critical; image network geometry should sufficiently cover the scene of interest, minimizing the occlusions and optimizing the intersection angles and the perspective differences between the images. As a matter of fact, dense and properly acquired image networks tend to yield more robust 3D representations. Image acquisition conditions such as illumination changes can also negatively affect the quality of the results. Moreover, the nature of the scene itself also plays an important role in the quality of the reconstruction. MVS solutions work by definition with static scenes; thus, moving objects cannot be reconstructed properly and typically result in noisy point clouds. Traditional MVS methods [Strecha et al., 2006; Hirschmuller, 2008; Rothermel et al., 2012] rely on global or local smoothness assumptions to establish pixel correspondences and recover the scene depth. Hence, drastic depth discrepancies and surface discontinuities are challenging, making it harder to estimate reliable depth values around the crease edges. Such smoothness formulations often also imply fronto-parallel surfaces (i.e. surfaces parallel to the camera baseline) and fail to reconstruct slanted surfaces without additional cues (see Chapter 2 for more details). Moreover, they typically construct memory-consuming cost volumes since they evaluate the matching cost in every possible disparity and often require depth range priors. The high memory requirements cause a severe scalability problem since large-scale scenes are prohibiting in such scenarios. Plane-sweep [Gallup et al., 2007] was the first approach to efficiently tackle the fronto-parallel bias, yet global cost volumes were still computed. PatchMatch [Bleyer et al., 2011] is a robust alternative to these limitations of the standard approaches; using local planes, the depth can be recovered even for slanted scene surfaces. Reliable depth estimates are propagated to their neighboring pixels based on the natural spatial coherence of the images, skipping the computationally expensive global cost volumes and the requirement for a pre-defined scene depth range. Chapter 3 provides a comprehensive review of the principles of the PatchMatch algorithm for depth estimation and the immense progress that the field has seen. Nonetheless, the greatest limitations nowadays in MVS are the ones related to the nature of the surface; large textureless areas, reflective or transparent materials, also known as non-Lambertian surfaces[1], are challenging due to the high degree of visual similarity and, thus, inevitable matching ambiguities.

The advent of deep learning techniques has pushed forward the research into tackling these challenges and brought new perspectives to solving the problem by incorporating semantic cues into the depth estimation problem; however, MVS remains a deeply geometric task. Up to now, learning-based algorithms have not demonstrated the capability to process high-resolution images. Therefore, depth maps are typically calculated on lower resolution images and upsampled in a post-processing step, but this inevitably implies detail loss and a compromise in accuracy with respect to conventional, handcrafted methods. Such detail loss is prohibitive in real-world scenarios, especially in photogrammetric applications with high-quality requirements. Moreover, most CNN-based methods, e.g., [Yao et al., 2018; Xu and Tao, 2020c], are based on plane sweeps and construct global cost volumes, adding a computational burden, especially during regularization. A more detailed survey of the recent learning-based methods for depth estimation is provided in Chapter 2.

Acknowledging the aforementioned observations, this dissertation aims, on the one hand, to better understand and investigate the underlying concepts and challenges in image-based 3D reconstruction and, more particularly, the depth estimation in real-world multi-view scenarios. Instead of more traditional methods, a PatchMatch-based approach is followed as the most efficient and robust state-of-the-art solution. However, PatchMatch is, in practice, a local cost computation method relying only on standard similarity metrics that have proven to be sensitive in the presence of matching ambiguities. Focusing particularly on this limitation, this work explicitly addresses one of the major, real-world failure cases, the non-Lambertian surfaces. Such problematic regions commonly occur in man-made scenes with large textureless surfaces or highly reflective materials; they tend to be problematic in 3D reconstruction pipelines, as the ambiguity of matches does not facilitate depth estimation. This problem has recently been of high interest to the research community with conventional [Romanoni and Matteucci, 2019; Xu and Tao, 2020b] and learning-based methods [Wang et al., 2020b]. In earlier approaches [Furukawa et al., 2010; Shen, 2013; Schönberger et al., 2016], such surfaces were commonly reconstructed with few, sparse points that could be filtered out similarly to noise. In practical image-based 3D reconstruction applications with high-quality requirements, such problematic surfaces are often reconstructed using hardware assistance, i.e., laser scanners that may operate in a complementary manner. The observation that additional higher-level scene

[1] As "Lambertian" are defined surfaces that follow the Lambertian Law and exhibit, thus, Lambertian reflectance; in other words, perfectly "matte" surfaces with diffuse reflectance. On the contrary, non-Lambertian surfaces have a different appearance depending on the viewpoint.

understanding information (e.g., from object detection and segmentation) is needed to tackle this problem, as stated in Schöps et al. [2017], being investigated mostly in the volumetric reconstruction domain [Häne et al., 2016], led to the motivation to use advanced scene cues to guide the depth reconstruction process.

1.3 Overall goal, objectives and contributions

1.3.1 Overall goal

The overall goal of this dissertation, apart from comprehensively studying the nature of the depth estimation problem in both the stereo and the MVS case, is to identify the open challenges on a theoretical basis yet also in practical, real-world applications. The particular challenges should be critically investigated, and innovative methodologies should be proposed to address them in a constructive way. Although having a strong theoretical background, the novel methodologies should be functional and aim to be easily adopted and further developed by other researchers for similar applications. Accordingly, the proposed improvements should be integrated into an open-source and well-known MVS framework to enable reproducibility. Such a framework should rather be domain-independent, robust, and adequately scalable for arbitrary large, real-world datasets. Overall, it can be said that the proposed approaches intend to generate more reliable depth estimates, especially under challenging scenarios, and thus, more complete, accurate, and visually appealing 3D point clouds.

1.3.2 Objectives

Considering the above, the following research question summarizes the aim of the thesis adequately:

Research question: "Can advanced scene priors such as semantic cues be leveraged in the multi-view scenarios, enrich the delivered data and support the depth estimation and 3D reconstruction on particularly challenging areas where matching ambiguities occur? Can this approach generalize to a method independent from semantic reasoning and undertake the matching ambiguities in an unsupervised, non-data-driven way?"

To undertake the arisen issues of these questions, the following objectives have been defined:

Objective 1. Exploit the recent advances in deep learning for efficient semantic segmentation on images using data acquired with high-resolution cameras. Contribute with a new dataset, targeting the specific problem of facade segmentation and employ a straightforward pipeline producing satisfying results for real-world and high-resolution scenarios.

Objective 2. Propose a functional image-based 3D reconstruction pipeline with the additional module for generating semantically enriched point clouds using label transfer from pixels to 3D points. Within this module, class-specific reconstruction should be enabled in such a way to selectively reconstruct the depth for only the pixels that are semantically meaningful for each application.

Objective 3. Leverage the semantic cues deriving from 2D semantic masks to the MVS reconstruction to tackle matching ambiguities and improve depth estimation in challenging scene surfaces, and derive complete and visually appealing 3D dense clouds.

Objective 4. Develop an adaptable, generic, non-data-driven strategy to efficiently target the same problem of matching ambiguities in cases where no semantic information for the scene can be obtained. The proposed method has to fulfill certain scalability requirements, i.e., it should be able to process arbitrary large datasets efficiently.

Objective 5. Integrate the proposed methodologies of Objectives 2, 3, and 4 in an end-to-end framework in such a way to make it easily adopted by the research community and facilitate reproducibility. Opt for a non-learning-based pipeline to ease the requirement for the enormous amount of training data and be domain-independent and thus enable easy generalization. The functionalities should be employed in a robust, widely-used and open-source library.

1.3.3 Original contributions

The major contributions of this dissertation can be summarized:

- A comprehensive overview of the depth estimation concept under both stereo and multi-view scenarios. To this end, state-of-the-art algorithms are discussed in detail to better understand their principles and impact, along with their major open challenges. Practical experience in this field has been gained recently, among others, during the research works of [Stathopoulou et al., 2019] for the dense reconstruction and [Nocerino et al., 2020] for mesh reconstruction under the MVS scenario. Both aforementioned publications presented a throughout experimental evaluation of various state-of-the-art algorithms under diverse scenarios and provided the ground for the further developments discussed in this dissertation.

- The integration of two fundamental research pillars in photogrammetry and computer vision: image-based 3D reconstruction and semantic segmentation. Motivated by the recent advancements in deep learning, a robust pipeline is proposed for efficient semantic segmentation on images; such segmentation masks can be used to generate semantically enriched 3D point clouds and enable class-specific 3D reconstruction. Within this context:
  - a new benchmark for facade semantic segmentation on historic building images *3DOM Semantic Facade* has been introduced in [Stathopoulou and Remondino, 2019a]. *3DOM Semantic Facade* is used in an effective semantic segmentation pipeline based on deep learning to train a model for semantic segmentation of building facades that can be easily generalized on unseen data. Detailed experiments and evaluations are presented. This contribution is relevant to **Objective 1**.
  - after the generation of the segmentation masks, label transfer from 2D to 3D is proposed to enable the direct and efficient generation of enriched 3D outputs, rather than following a semantic segmentation strategy in the 3D space [Stathopoulou and Remondino, 2019a; Stathopoulou et al., 2021b]. Similarly, experiments on class-specific 3D reconstruction demonstrate the effectiveness of the proposed approach, as presented in [Stathopoulou and Remondino, 2019b]. Although the conceptualization of this idea initially referred to the building facade scenario, it is demonstrated that the method can be employed in a large variety of applications, either airborne or terrestrial. This contribution is relevant to **Objective 2**.

- The potential of semantic reasoning is exploited to improve the quality of the 3D reconstruction. Therefore, a method for leveraging a priory obtained semantic cues into the standard PatchMatch-based MVS is proposed in Stathopoulou et al. [2021b]. The novel pipeline specifically targets the matching ambiguities problem in the presence of textureless, reflective, and generally non-Lambertian surfaces in order to yield, with respect to the standard approach, more complete and accurate 3D point clouds. Within this context and relevant to **Objectives 3** and **5**:
  - the *ETH3D* MVS benchmark has been extended by semantic equivalents on three of its sets (*courtyard, terrace, pipes*), including outdoor and indoor scenarios;
  - a class-specific prior generation in the 3D space method is proposed. A RANSAC-based approach is followed and the semantic masks are used to guide the plane search and define dominant planar areas, commonly consisting of large textureless surfaces;
  - a novel adaptive cost function is introduced to seamlessly leverage the planar priors deriving from the semantic cues and according to the local textureness information;

  - evaluations of diverse scenarios on benchmark and custom datasets demonstrate the efficiency of the proposed method.

- The above method is generalized to undertake cases where semantic cues are not available or hard to obtain; a novel approach based solely on local structure and textureness information and quadtree structures guidance is proposed [Stathopoulou et al., 2022] to improve the completeness and the accuracy of the results. Within this context and relevant to **Objectives 4** and **5**:

  - local structure information as described by a quadtree-based image decomposition is used as guidance for the plane prior generation and the depth hypothesis in the 3D space. Large areas of uniform color are potentially more probable to be described by local planes. The adaptive cost function is also adopted here to seamlessly integrate the prior hypotheses and the photometric cost.
  - experiments have been made on the large-scale benchmark dataset *ETH3D* [Schöps et al., 2017] and custom datasets, demonstrating state-of-the-art performance in line with the high-performing state-of-the-art conventional and learned methods.

- the deployment of the aforementioned improvements in a well-established and broadly-used open-source library for image-based 3D reconstruction in an end-to-end fashion. The OpenMVS [Cernea, 2020] library was chosen as a representative example of a robust framework for image-based 3D reconstruction.

1.4 Thesis outline

This dissertation can be roughly divided in two parts; the first part provides an introduction to the topic, describes the theoretical background and discusses the state-of-the-art methods divided in two chapters for depth estimation and reconstruction (Chapter 2) and the PatchMatch algorithm in particular (Chapter 3). The second part presents the proposed methodologies and improvements, as well as the relative experiments and evaluations on semantic segmentation and 2D to 3D label transfer (Chapter 4), integration of semantic priors in MVS (Chapter 5) and the use of solely structure priors for improving MVS depth estimation and reconstruction (Chapter 6).

Chapter 2. Theoretical background and literature review on depth estimation and reconstruction in stereo and multi-view scenarios is discussed. Fundamental concepts are explained along with an in-depth categorization of the available methods, both conventional and learning-based. The main challenges and limitations are comprehensively discussed, along with improvement proposed in related work. Finally, the available benchmarks are introduced.

Chapter 3. The PatchMatch algorithm is comprehensively presented. Its applicability in depth estimation is outlined in stereo and multi-view cases. An exhaustive literature review on the state of the art methods is presented and related works are categorized. The advantages of using PatchMatch over traditional methods are highlighted, and the remaining open challenges are discussed.

Chapter 4. Semantic segmentation toward scene understanding is briefly discussed, and the basic principles of deep learning methods are outlined in a constructive fashion. A short literature review on facade segmentation methods is presented. The *3DOM Semantic Facade* benchmark is introduced along with a working pipeline for efficient generation of 2D semantic maps for historic building facades. Finally, the integrated 3D reconstruction pipeline is introduced with respective experimental results, yielding semantically augmented point clouds and enabling class-specific reconstruction.

Chapter 5. A novel MVS framework is proposed for integrating semantic reasoning in multi-view stereo reconstruction and achieving more complete point clouds in problematic regions. Semantically-guided plane hypotheses are generated to support the PatchMatch algorithm and propagate reliable depth estimates in textureless and reflective areas where typically matching ambiguities occur. Experimental results on benchmark (*ETH3D*) and custom datasets are presented and evaluated.

Chapter 6. In cases where no semantic information is available, the method proposed in Chapter 5 is generalized to rely only on local structure and textureness information. Quadtree image decomposition is used to guide the 3D plane hypotheses and support PatchMatch depth estimation in challenging scene areas. Experimental results on benchmark (*ETH3D*) and custom datasets are presented and evaluated.

Chapter 7. The final chapter summarizes the work presented in this dissertation. The main contributions are recapped, and future directions are given. Finally, the publications leading to this dissertation are listed along with other relevant publications by the author.

Chapter 2

Depth estimation and reconstruction

The complete image-based 3D reconstruction of a scene requires estimating the depth, i.e., the distance from the camera in the 3D space, for potentially every pixel. To this end, pixel correspondences are established between two or multiple images. Several computer vision and image processing problems are based on pixel correspondences such as stereo matching [Scharstein and Szeliski, 2002], optical flow [Bailer et al., 2015] and computational photography applications such as deblurring [Hacohen et al., 2013], and inpainting [Guillemot and Le Meur, 2013] to name but a few.

Depth estimation from RGB images is one of the fundamental problems in photogrammetry and computer vision; it has been an active research topic for decades with high-level applications in robotics [Samadi and Othman, 2013], autonomous driving [Geiger et al., 2012], medical imaging [Nam et al., 2012], augmented reality [Baričević et al., 2014], among others, under binocular, i.e., two-view, multi-view or even monocular scenarios. In the past decades, several algorithms and techniques have been developed for automated depth estimation and, consequently, 3D reconstruction from images. Various other active or passive methods for 3D scene recording also exist, such as structured light scanning, triangulation laser scanning, shape from silhouette, as well as shape from shading methods, or photometric stereo. However, image-based 3D reconstruction is a widely used technique as, apart from its robustness, it is also time and cost-effective.

Stereo matching has been one of the dominant methods for depth estimation due to its strong connection with the human vision system. It relies on matching corresponding pixels across two or multiple images, for which the relative geometry is known, also called the correspondence problem. Depth estimation from pixel correspondences is an inverse, hence ill-posed, problem given the large ambiguities introduced by potential occlusions and surface appearance variations across different views. The fundamental cue for stereo matching is the surface's visual appearance, represented by the pixel color as a function of object material, scene illumination, and the 3D geometry of the scene being captured. In other words, the rays observing the same scene point should convey similar photometric information or be "photometrically consistent". The photometric consistency, or photo-consistency measure, is used to quantify this similarity. Once valid pixel correspondences are found, the depth of every pixel can be calculated, making, thus, its reprojection into the 3D space possible. Despite the extensive relative research in the last decades, accurate depth estimation and complete reconstruction remain an open challenge due to de facto unsolved issues such as illumination changes, occlusions, textureless areas, and non-Lambertian surfaces. Eventually, the developed algorithms often fail to find reliable correspondences and correctly reconstruct the depth in the areas of low texture, as photo-consistency measures alone are not robust enough to deal with depth inconsistencies and, consequently, the matching ambiguities.

In this chapter, some basic principles and notation on depth estimation from RGB images will be given. First, the basics of camera geometry will be shortly discussed, followed by an extensive overview of the stereo and multi-view methods for depth estimation and scene reconstruction with a critical view of the open

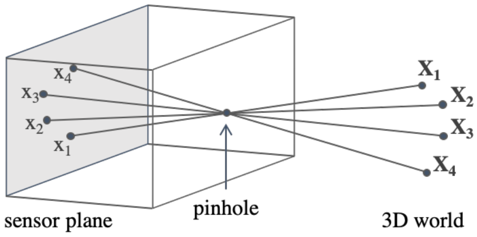


Figure 2.1: **Abstract representation of the pinhole camera model.** Points of the 3D world are mapped to the sensor plane via the pinhole.

challenges and limitations. Learning-based depth estimation approaches will also be reviewed, and finally, the most commonly used benchmarks in the field will be presented.

2.1 Camera geometry basics

A camera is essentially a many-to-one mapping between the real 3D world and a 2D image. Precise depth estimation from RGB images requires reliable information about the camera geometry, i.e., the shape of the bundle of rays connecting the object in the 3D space and its traces on the image. Accordingly, some basic background in camera geometry will be introduced to describe the fundamental concepts and notation relevant to this thesis; the reader is referred to [Hartley and Zisserman, 2003; Förstner and Wrobel, 2016; Szeliski, 2010] for a more comprehensive understanding of these geometric concepts.

2.1.1 Camera models

Camera models describe the association between observation rays and pixels in an abstract mathematical way. During the calibration process, the exact parameters of this model are defined, known as intrinsic parameters of the camera or parameters of the interior orientation [McGlone, 2004; Förstner and Wrobel, 2016]. Instead of modeling the rays for every pixel, camera models make certain assumptions to reduce the number of the parameters.

A widely used model is the pinhole camera; it is an ideal, abstract model assuming that all light rays from the scene forming the image on the camera sensor pass without deviation through the same 3D point, as shown in Figure 2.1. This point is commonly called optical center, camera (lens) center, perspective center, center of projection, or simply pinhole. Assuming a viewing image plane in front of the pinhole **C** (Figure 2.2), the projection of the pinhole, i.e., the perspective center on the image plane, is called the principal point $c(c_x, c_y)$. The distance from **C** to the image plane is called focal length f, or principal distance. In an *ideal pinhole*

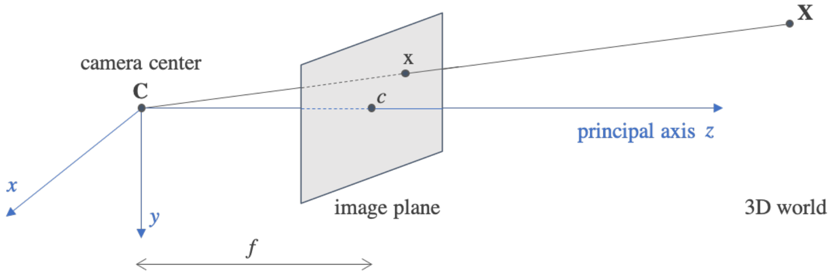


Figure 2.2: **Geometry of a camera.** The camera center **C** is the origin of a cartesian coordinates system. A 3D point scene **X**, its trace on the 2D image x and the camera center **C** are collinear.

camera with a planar sensor, the pinhole **C** is the origin of the coordinate system, and thus the camera is described only by the focal length f. If the principal point deviates from the origin of the coordinate system, the a camera model would be described by the three parameters c_x, c_y and f (*Euclidean camera*). A more generalized version of this model is the *perspective camera or camera with affine sensor*, including a skew value s for non-square pixels, and/or scale difference (aspect ratio) a. For a more detailed overview of the commonly used models, the interested reader is referred to [Hartley and Zisserman, 2003; Förstner and Wrobel, 2016].

Camera models are commonly described with the camera calibration matrix **K**, a 3 × 3 matrix that transforms the rays into homogeneous image coordinates; homogeneous representations are typically used in this geometric context as they are convenient for linear algebra calculations [Förstner and Wrobel, 2016]. Hence, **K** encodes the transformation from the normalized image coordinates, measured on an ideal plane, to image coordinates. The general form of this matrix refers to the perspective camera model; it expresses the affine transformation containing the focal length f expressed by two components f_x and f_y differing by an aspect ratio a, the principal point coordinates $c(c_x, c_y)$ and a skewness parameter (shear) s for the pixels:

$$\mathbf{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.1}$$

The above models assume a linear projection according to which straight lines in the 3D world are mapped as straight lines in the image (ideal perspective models) [Förstner and Wrobel, 2016]. In the real world, lenses are characterized by de facto imperfections, causing, among others, distortion effects like radial distortion. To compensate for systematic errors, the estimation of the distortion coefficients is required for the correct mapping between object points and pixels, and the camera model is generalized to the *perspective camera with nonlinear* *distortions*. Distortion compensation is commonly performed according to the model introduced by Brown [1965]. Typically, 1-3 radial distortion coefficients (k_1, k_2, k_3) are accounted for, along with two tangential coefficients p_1 and p_2.

Let the center of projection $\mathbf{C}$ be the origin of a cartesian coordinate system, with the z-axis being the axis perpendicular to the image plane (principal axis). A 3D point $\mathbf{X}(X, Y, Z)^T$, its trace on the 2D image x and the camera center $\mathbf{C}$ lie on the same line (Figure 2.2). The point $\mathbf{X}$ from the Euclidean space $\mathbb{R}^3$ is mapped on the Euclidean space $\mathbb{R}^2$ (image plane) by similar triangles as:

$$(X, Y, Z)^T \mapsto (f\frac{X}{Z}, f\frac{Y}{Z})^T. \tag{2.2}$$

Using homogeneous coordinates for x, the matrix describing this mapping is a 3×4 homogeneous matrix known as the projection matrix $\mathbf{P}$. It includes all information about the camera parameters, intrinsic and extrinsic, and is equivalent to the collinearity equations used in photogrammetry [Das, 1949]. It is given by:

$$\mathbf{P} = \mathbf{K}[\mathbf{I}|0], \tag{2.3}$$

with $\mathbf{K}$ being the camera calibration matrix and $\mathbf{I}$ the identity matrix. Considering general rotation $\mathbf{R}$ and translation $\mathbf{t}$:

$$\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]. \tag{2.4}$$

Therefore, a point on the image x is connected with a point $\mathbf{X}$ in 3D by:

$$\mathrm{x} \sim \mathbf{PX}, \tag{2.5}$$

where $\sim$ indicates equality up to scale; $\mathbf{P}$ is homogeneous as its scale can be arbitrarily chosen. It has 11 degrees of freedom DOF, 5 of the intrinsic and 6 of the extrinsic geometry of the bundle. The mapping between x and $\mathbf{X}$ can be solved via the direct linear transformation (DLT) algorithm [Abdel-Aziz and Karara, 1971; Hartley and Zisserman, 2003].

2.1.2 Two-view geometry

Similar to the human vision system, to obtain 3D measurements from 2D images, two (at least) overlapping views of the same scene are needed; such systems are known as two-view or binocular. The intrinsic projective geometry between two camera views that relates the cameras, the points in 3D, and the observations in 2D is described by the epipolar geometry [McGlone, 2004; Hartley and Zisserman, 2003; Förstner and Wrobel, 2016]. Let two images taken from camera centers $\mathbf{C}$ and $\mathbf{C}'$, related by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, and let $\mathbf{X}$ be a

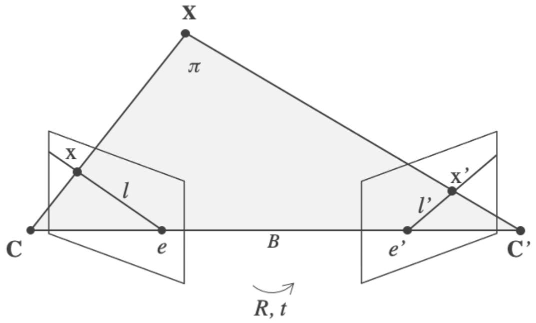


Figure 2.3: **Epipolar geometry.** For two images with centers $\mathbf{C}$ and $\mathbf{C}'$ and relative pose transformation defined by $\mathbf{R}, \mathbf{t}$. The epipolar plane π is defined by the 3D point $\mathbf{X}$, and the two image centers $\mathbf{C}$ and $\mathbf{C}'$.

point in the 3D object space (Figure 2.3). Then x and x' will be the projections of this point onto the left and right image planes, respectively. The plane π defined by $\mathbf{X}$ and the two camera centers $\mathbf{C}$ and $\mathbf{C}'$ is called the epipolar plane. The line that joins the two camera centers $\mathbf{C}$ and $\mathbf{C}'$ corresponds to the baseline B. The traces of π on the image planes are called epipolar lines l, and l', and the intersections of these lines with the baseline are the epipoles e and e'. All epipolar lines go through the camera's epipole. The coplanarity constraint implies that the viewing rays through corresponding points are coplanar.

Corresponding points x and x' of a stereo pair must lie on corresponding epipolar lines l and l'. Therefore, there is a mapping relationship between a point and a line, also known as epipolar constraint:

$$\mathrm{x} \mapsto l', \tag{2.6}$$

which represents a singular (and thus not proper) correlation. Indeed, this mapping is singular because, for every point of the first image, an epipolar line on the second one exists. There is no inverse mapping; that is, to every epipolar line of the second image corresponds a line on the first image. The mapping between each point and its corresponding epipolar line is described for the general uncalibrated camera case by the fundamental matrix $\mathbf{F}$:

$$l' = \mathbf{F}\mathrm{x}. \tag{2.7}$$

Similarly, $l = \mathbf{F}^T\mathrm{x}'$. In other words, $\mathbf{F}$ is the algebraic representation of the epipolar geometry [Hartley and Zisserman, 2003]. For any pair of correspondences x and x', $\mathbf{F}$ must satisfy the condition:

$$\mathrm{x}'^{\mathrm{T}}\mathbf{F}\mathrm{x} = 0. \tag{2.8}$$

It is a 3×3 matrix with zero determinant ($\det \mathbf{F} = 0$) and rank=2 such that:

$$\mathbf{F}e = \mathbf{F}^{\mathbf{T}}e' = 0. \tag{2.9}$$

F has 7 DOF and can be computed by a set of corresponding points with homogeneous coordinates. Commonly, the 8-point algorithm or the 7-point algorithm are used for its estimation, typically refined with RANSAC-based model fitting methods [Zhang, 1998; Faugeras and Luong, 2001; Hartley and Zisserman, 2003]. **F** refers to the general case where no information about the camera intrinsics is available, i.e., to uncalibrated cameras. Therefore, just a projectively distorted model of the scene can be computed, or, in other words, **F** contains a projective ambiguity. If the intrinsic parameters are known, the fundamental matrix **F** is equivalent to the essential matrix **E**. Introduced by Longuet-Higgins [1981], it is given by:

$$\mathbf{E} = [\mathbf{t}]_{\mathbf{x}}\mathbf{R}, \tag{2.10}$$

where **R** is the 3×3 rotation matrix and $[\mathbf{t}]_{\mathbf{x}}$ is the skew-symmetric (cross-product) matrix of the translation vector **t**:

$$[\mathbf{t}]_{\mathbf{x}} = \begin{bmatrix} 0 & -t_x & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}. \tag{2.11}$$

Equation 2.8 will be in case of the essential matrix:

$$\mathrm{x}'^{\mathrm{T}}\mathbf{E}\mathrm{x} = 0. \tag{2.12}$$

Alternatively, with the use of camera model **K**:

$$\mathbf{E} = \mathbf{K}'^{\mathbf{T}}\mathbf{F}\mathbf{K}, \tag{2.13}$$

where **K** is the camera matrix that includes the intrinsic parameters of the camera model. Distortion (radial and tangential) is assumed to be zero. Similar to the fundamental matrix, **E** has rank=2. However, the essential matrix has 5 DOF and can be computed by only 5 points [Nistér, 2004]. With given intrinsic parameters, a Euclidean 3D model of the scene can be estimated.

2.1.3 Stereo rectification

Once the relative poses of the two images are estimated, the epipolar constraint can be used to limit the search space for corresponding pixels in the other image. An efficient way can be to first rectify, i.e., warp, the input images so that

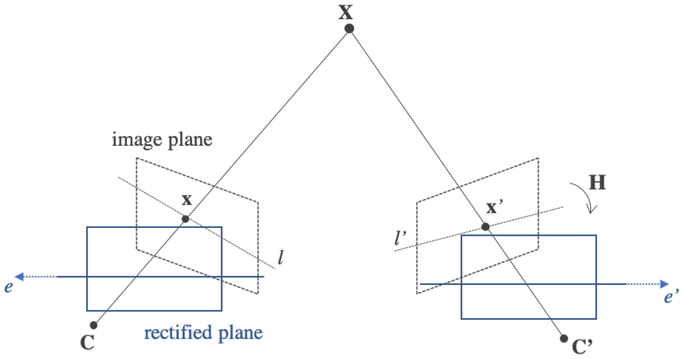


Figure 2.4: **Epipolar rectification.** Image planes become parallel and epipoles are located at infinity.

the horizontal scanlines are aligned and correspond to the epipolar lines while minimizing distortions, a process known as stereo or epipolar rectification [Loop and Zhang, 1999; Hartley and Zisserman, 2003]. Stereo rectification generates coplanar images and ensures that potential correspondences are located in the same row of pixels in the reference and source image. In other words, rectified images will be transformed such that $\mathbf{R} = \mathbf{R}'$, i.e., the baseline will be parallel to the x-axis. Therefore, a projective transformation $\mathbf{H}$ needs to be found such that the epipoles e and e' in the two images are mapped to the infinite point $[1, 0, 0]^T$. Evidently, such a configuration represents the "stereo normal case" in photogrammetry performed on calibrated cameras.

Various approaches for different camera configurations have been proposed for this scope, trying to minimize the image distortion [Pollefeys et al., 1999; Fusiello et al., 2000; Abraham and Förstner, 2005; Fusiello and Irsara, 2008]. In the widely used method of Fusiello et al. [2000] a linear way to rectify the two images of known poses is proposed, by first rotating both cameras so that they are perpendicular to the baseline and subsequent rotation of the two images to the coordinate system of their baseline B. A quasi-Euclidean rectification is implemented for general uncalibrated cases in [Fusiello and Irsara, 2008].

Stereo rectified images fulfill two conditions: (1) all epipolar lines should be parallel to the horizontal axis, and thus the baseline (2) corresponding points have identical vertical coordinates i, so the disparity, i.e., the coordinate difference, in y-direction will be equal to zero.

2.2 Stereo matching

Stereo matching, also known as binocular stereo, relies on the same principles as the human vision system; the two cameras, (equivalent to the eyes) observing the same scene space have the ability to perceive depth information. Nonetheless, while the human brain can effortlessly create such associations, designing an algorithm to simulate this task is not trivial.

The basic principle of stereo matching can be explained as follows: given two stereo-rectified images I and I' of known relative poses $(\mathbf{R}, \mathbf{t})$, with the I being the reference (left) image and I' the source (right) image, optimal pixel correspondences, i.e., matches, are to be established. Matching pixels represent the same point of the 3D scene projected onto both images. Thus, the depth d of, if possible, every pixel p of the reference image can be calculated. In other words, the aim is to find pairs of 2D pixels that correspond to the same 3D point $\mathbf{X}$. From these correspondences, the 3D coordinates can be estimated via triangulation.

Stereo methods typically refer to small baselines and use stereo rectified images, although solutions for not rectified images also exist. For establishing reliable pixel correspondences, the epipolar constraint is exploited; according to this constraint, corresponding pixels should lie on the same scan line, which is parallel to the baseline and the horizontal axis for stereo rectified images. Hence, the problem has only one degree of freedom (DOF), the offset of the x-axis, also known as x-parallax, and the search space is reduced to a 1D horizontal line compared to the standard 2D optical flow case. The apparent horizontal shift or displacement d in the x-direction between corresponding pixels p and p$'$ on the reference and source images is called disparity:

$$\mathrm{p}(x, y) \leftrightarrow \mathrm{p}'(x + d, y) \tag{2.14}$$

Consequently, the problem can be reformulated by finding the disparity $d = x' - x$ for each pixel correspondence. Inversely proportional to the disparity, the depth Z of each pixel is calculated by similar triangles (Figure 2.5) $Z = f\frac{B}{d}$, where f is the camera's focal length and B is the baseline, i.e., the distance between the two optical centers.

During stereo matching, corresponding disparity and depth maps are typically generated for each view, storing the disparity or depth values of each pixel for a stereo pair. Considering precisely known relative poses and camera calibration parameters, the depth reconstruction accuracy in the 3D space is directly related to the quality of the matches; consequently, depth refinement methods and sub-pixel interpolation crucial for an accurate 3D reconstruction. Many algorithms in this direction have been introduced in the latest decades. According to Scharstein and Szeliski [2002], a typical stereo matching pipeline can be described by (1) matching cost computation (2) cost aggregation (3) disparity estimation (4) disparity refinement.

This sequence of steps is the most prominent one, yet other combinations are also possible. Generally, stereo matching algorithms can be roughly divided into local and global methods. Local methods consider support windows around

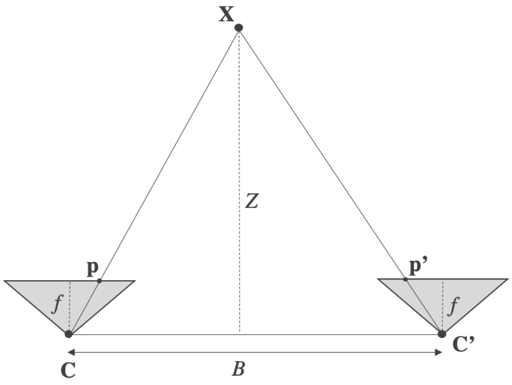


Figure 2.5: **Stereo correspondences.** Two cameras with centers $\mathbf{C}$, $\mathbf{C}'$ and baseline B. Matching pixels p and p$'$ correspond to the same 3D point $\mathbf{X}$.

each pixel and make implicit constant smoothness assumptions over this window during cost aggregation. On the other hand, global methods model explicit smoothness assumptions based on surface priors and try to solve an energy minimization problem, commonly skipping the cost aggregation step. In the following paragraphs, an overview of the stereo matching procedure is given, along with a survey of the related work in the field in the past decades and the current open challenges and limitations of the methods.

2.2.1 Matching cost computation

The human vision system has the ability to locate corresponding patterns between two images simply by studying their visual appearance, even if certain geometric or radiometric transformations are present. Moreover, humans can perceive contextual information of the overall scene, which tends to be helpful while making these associations. In the natural world, surfaces are piecewise smooth and continuous; therefore, the 2D projections of these regions on the images will inherit this property. Stereo matching algorithms aim to simulate this principle, using appearance relationships and natural coherency.

The matching cost is a measure to quantify dissimilarity or distance between general queries (in this case pixels or patches). Visual appearance is the most commonly used criterion for stereo matching, also known as photometric or photo consistency. Photometric consistency assumes that the corresponding pixels have the same visual appearance across different views. In the most simplified case, the matching cost for a certain disparity d can be assumed as the absolute difference (AD) [Kanade et al., 1995] of the gray values between of corresponding pixels p(x, y) and p$'(x + d, y)$ in the reference image I and the source image I' respectively:

$$AD(p) = |I(\mathrm{p}) - I'(\mathrm{p}')|. \tag{2.15}$$

A low matching cost value implies high visual similarity and vice versa. In stereo matching, the correspondence with the highest similarity and thus lowest matching cost is sought. Similarly, the squared differences (SD) can be assumed [Hannah, 1974] instead of simply the absolute difference.

In practice, the difference of the gray values alone is insufficient due to appearance differences between the images, even if they are sequential in time and space. Indeed, deciding if two pixels are corresponding, i.e., photometrically consistent, is a quite challenging task as many factors may affect the final appearance of an image, like viewpoint changes, variations in the overall illumination of a scene, occlusions, shading, reflections, noise, and differences in the camera settings. Accordingly, patches, i.e., small pixel windows $n \times n$, are used to consider the close neighborhood and enforce distinctiveness. Since there is no guarantee that the visual appearance of every single pixel is unique across the search space, larger patch size is more likely to lead to unique correspondences. A variety of similarity measures has been proposed in the literature, either parametric or non-parametric, including:

Sum of absolute differences (SAD). It is a simple measure based on the L1 norm of all pixels q in a local window $n \times n$ around a pixel $\mathrm{p}(x, y)$ [Kanade and Okutomi, 1994] between the reference image and the source image:

$$SAD(\mathrm{p}) = \sum_{\mathrm{q} \in [n \times n]} |I(\mathrm{q}) - I'(\mathrm{q}')|. \tag{2.16}$$

SAD is typically truncated to become more robust to outliers. Yet, it is still sensitive to gain and bias (equivalent to contrast and brightness, respectively) and, therefore, to illumination changes. However, SAD is a fast measure and can be used in real-time applications where capturing conditions are similar across the images.

Sum of squared differences (SSD). It computes the difference of intensities between two local windows q and q′ by first calculating the squared distance pixelwise and subsequently summing them (L2 norm) [Hannah, 1974; Matthies et al., 1989; Okutomi and Kanade, 1993]:

$$SSD(\mathrm{p}) = \sum_{\mathrm{q} \in [n \times n]} (I(\mathrm{q}) - I'(\mathrm{q}'))^2. \tag{2.17}$$

It is sensitive to outliers and gain and bias. By definition, it is not bounded yet is often truncated to exclude outliers or mapped through an exponential function, becoming bounded to $[0, 1]$. SSD is sensitive to radiometric changes, especially while unbounded, since the squared difference L2 norm may take particularly large values.

Normalized cross-correlation (NCC). The cross-correlation operation is similar to convolution, applying a filter (kernel) across the image; hence this metric is also known as "normalized sliding dot product" [Hannah, 1974]. NCC measures the differences in a normalized way to compensate for gain and bias, increasing, in this manner, the robustness in linear illumination changes that can be beneficial in stereo matching. Statistically, it is the optimal metric for cases where Gaussian noise is present [Hirschmuller and Scharstein, 2008]. It encourages the matching of underlying patterns rather than raw intensity values. Since it is normalized by the product of the standard deviation of the two patches, it calculates a correlation value bounded between $[-1, 1]$. An NCC cost equal to 1 means that the potential correspondences are identical, while -1 means that they are totally irrelevant. Although it is well-suited for perspective differences, it often fails on low-textured surface cases and repetitive patterns. To increase robustness, a zero-mean NCC (ZNCC) is commonly preferred since, due to the subtraction of the zero mean μ of the neighboring intensities, it can also handle affine intensity changes:

$$NCC(\mathrm{p}) = \frac{\sum\limits_{\mathrm{q}\in[n\times n]} [I(\mathrm{q}) - \mu_I] \cdot [I'(\mathrm{q}') - \mu_{I'}]}{\sqrt{\sum\limits_{\mathrm{q}\in[n\times n]} [I(\mathrm{q}) - \mu_I]^2} \cdot \sqrt{\sum\limits_{\mathrm{q}\in[n\times n]} [I'(\mathrm{q}') - \mu_{I'}]^2}}. \tag{2.18}$$

Rank transform (RT). Non-parametric measures apply some ordering transformation to the data and use this information instead of the original data values [Zabih and Woodfill, 1994]. Rank transform, introduced by Zabih and Woodfill [1994] is a non-parametric measure applied to both images before the matching cost computation to enhance the edges and reduce the noise. To do so, pixel intensities are sorted, i.e., ranked, within a local window, and their scalar rank substitutes the actual intensity. Accordingly, RT is defined as the number of pixels q whose intensities are smaller than the one of the current pixel p:

$$RT(\mathrm{p}) = \sum_{\mathrm{q}\in[n\times n]} T(\mathrm{p}, \mathrm{q}), \tag{2.19}$$

where the function T expresses the relationship between the pixel intensities p and q:

$$T(\mathrm{p},\mathrm{q}) = \begin{cases} 0, & \text{if } I(\mathrm{p}) \leq I(\mathrm{q}) \\ 1, & \text{if } I(\mathrm{p}) > I(\mathrm{q}). \end{cases} \tag{2.20}$$

For a window $n \times n$, the pixel intensities are replaced with values between $[0, (n \times n) - 1]$; thus, the chosen window size is fundamental. After this transformation, common metrics such as SAD can be used for cost computation. RT is invariant to changes in gain and bias and hence more robust to global radiometric changes since it depends only on the ordering of the methods and not their original intensities. However, similar to other visual similarity metrics, is still sensitive to local radiometric changes and thus to matching ambiguities commonly occurring in textureless areas. To increase robustness, Hirschmuller and Scharstein [2008] used the Soft Rank Transform by defining a linear, soft transition zone between 0 and 1 for values that are close together.

Census transform (CT). Another non-parametric image transformation is Census Transform (CT) [Zabih and Woodfill, 1994]. For CT, a binary descriptor vector (bitstring) is assigned to each pixel based on intensity differences of every pixel q in a patch neighborhood $n \times n$ around pixel $\mathrm{p}(x, y)$. Using the same definition of T as before, CT is finally expressed as:

$$CT(\mathrm{p}) = \otimes T(\mathrm{p},\mathrm{q}), \tag{2.21}$$

with $\otimes$ denoting concatenation.

Compared with RT, CT also encodes the spatial relation of the pixels apart from their intensities. CT is robust against changes in gain and bias, and thus global radiometric/illumination changes [Hirschmuller and Scharstein, 2008]. Moreover, since the actual pixel intensities affect only partially the binary descriptors, CT is more robust to outliers and can achieve better results on textureless areas than other metrics. The robustness increases by increasing the mask size, yet a larger mask implies higher computational time. Nevertheless, it is still sensitive to strong perspective changes, as it is based on fixed geometry of the compared pixels. CT is less discriminative than NCC, leading to more ambiguous matches, yet it handles better depth discontinuities in object boundaries. In stereo normal cases, CT has become popular since it often outperforms NCC, yet in MVS scenarios commonly strong perspective changes exist [Ruf et al., 2021] has limited performance. The final matching cost is calculated as the Hamming distance between the two binary vectors in the query.

Mutual information (MI). Often used in information theory, mutual information cost proposed by Viola and Wells III [1997] measures the amount of information contained in one pattern about the other as a joint probability. In other words, it measures how dependent two patterns are. Operating on full images and with a given initial disparity estimation, it considers the entropy of the probability distributions H_I and $H_{I'}$ in two images I and I' as well as the entropy of the joint probability distribution $H_{I,I'}$ of pixel-wise correspondences of both images, using a Parzen window method [Parzen, 1962].

$$MI = H_I + H_{I'} - H_{I,I'}. \tag{2.22}$$

The probability distributions are calculated from the histograms of the intensities of the two images. Typically, high-fidelity image poses imply low joint entropy $H_{I,I'}$, as one image can be predicted by the other. Lower joint entropy means higher MI. It favors finding the most complex regions (maximizing individual entropies) that explain each other as well as possible (minimizing the joint entropy). It typically operates better when a large, more representative support region is provided, yet such operations are time-expensive and often cause blurring effects. MI, typically insensitive to radiometric changes along with gain and bias [Kim et al., 2003], can be adopted in both global and local methods and has been used in the literature [Hirschmuller and Scharstein, 2008; Kim et al., 2003; Campbell et al., 2008], although it is not a common measure.

Matching, in the broader context, may refer to establishing correspondence between salient points, also known as feature matching, using descriptors such as SIFT [Lowe, 2004], SURF [Bay et al., 2006], ORB [Rublee et al., 2011] etc. Such methods generally demonstrate geometric and photometric resilience in finding good correspondences but typically lead to abstract, sparse representations of the scene. Therefore, they find applicability in sparse matching scenarios, e.g., in Structure from Motion (SfM) as briefly explained in Chapter 1. However, specialized descriptors have been used in the past for finding dense or semi-dense correspondences in the MVS scenario [Tola et al., 2009, 2012], but such methods are beyond the scope of this dissertation.

Local methods use support windows (patches) for cost computation and aggregation, while global methods calculate the cost pixelwise and do explicitly perform cost aggregation. The matching costs may refer to grayscale or color images or their respective gradients to increase robustness in radiometric changes [Scharstein, 1994], or a combination of both [Hosni et al., 2012]. Pre-processed versions of images have also been used in the literature [Di Stefano et al., 2004; Hirschmuller and Scharstein, 2008].

2.2.2 Cost aggregation

The disparity space image (DSI) [Yang et al., 1993; Scharstein and Szeliski, 2002], often also referred to as cost volume, is the representation of the per-pixel disparity cost over all possible disparities in $[d_{min}, d_{max}]$. It is a 3D array of size: $W \times H \times d_{max}$, where H is the image height, W the image width, and d_{max} the maximum value of the disparity range.

In local methods, the cost aggregation step connects the costs within a certain neighborhood in the DSI [Scharstein and Szeliski, 2002], i.e., a local window or patch for regularization and refinement purposes, and improves performance when matching ambiguities are present. For local methods, based on the assumption that neighboring pixels share the same disparity, matching costs are summed over a 2D support region, commonly a window $n \times n$, around each image pixel p [Mühlmann et al., 2002; Di Stefano et al., 2004]. So, the aggregated cost of a pixel $c(\mathrm{p})$ is the sum of the costs of all pixels q in its neighborhood:

$$c(\mathrm{p}) = \sum_{\mathrm{q} \in [n \times n]} c(\mathrm{q}). \tag{2.23}$$

A major challenge is the definition of the optimal window size; it should be large enough to consider distinguishable enough disparities but also not too large to avoid edge fattening and oversmoothing, and detail loss. Although rectangular windows are mostly used because of their efficiency and ease of implementation, aggregation schemes with adaptive support windows have also been introduced [Kanade and Okutomi, 1994; Fusiello et al., 1997; Yang et al., 2008] particularly to avoid edge fattening. In fact, a simple summation of the costs within a local window implies a fronto-parallel surface and may result in oversmoothing as edges are not taken into consideration; thus, bilateral filters [Tomasi and Manduchi, 1998; Richardt et al., 2010] or guided filters [He et al., 2010; Hosni et al., 2012] have also been employed for edge preservation. Other methods adopt weights attributed to window pixels based on color similarity or geometric proximity between reference and neighbor pixels [Yoon and Kweon, 2006; Bleyer et al., 2011; Hosni et al., 2013], or histogram-based aggregation schemes for weights [Min et al., 2011], yet these approaches are usually computationally slow. Shiftable windows [Kang et al., 2001] and linearly expanded cross-skeleton windows [Stentoumis et al., 2014] have also been suggested in the literature. Cost aggregation can generally be performed using 2D windows or 3D convolutions (box filters in the case of square windows).

2.2.3 Disparity estimation

Disparity estimation methods can be roughly divided into local and global methods, with the latter being referred also as disparity optimization algorithms. Early disparity estimation methods were local, based on the "winner takes all" (WTA) rule, where the correspondence with the minimum cost is selected among the matching candidates. On the other hand, global methods are optimization methods that use energy function minimization for all image pixels, commonly expressed with a data term and a smoothness term, representing the summation of pixel costs and the local smoothness support, respectively.

Local methods. Local methods are based on correlation and use support windows to estimate the best correspondences in a limited local neighborhood, sliding across the image or search area till the best score is achieved. A DSI or cost volume **S** is typically built to store the information of all possible disparity values. The disparity estimation itself is straightforward since they follow the WTA approach, i.e., for each pixel p the disparity $\hat{d}$ with the lowest matching cost is selected:

$$\hat{d} = \operatorname{argmin} \mathbf{S}(\mathrm{p}, d). \tag{2.24}$$

The neighborhood information is not considered until the cost aggregation step. However, they model an implicit smoothness assumption within the support window, which often leads to blurry object boundaries [Hirschmuller, 2006]. That being said, and if no support window is used, the problem becomes a pixelwise matching one [Birchfield and Tomasi, 1998] that consequently has to deal with stronger matching ambiguities. Local methods are generally fast but typically provide inferior quality results than the global methods. Yet, the quality can be significantly improved by using an elaborated cost aggregation step.

Local algorithms vary from simple correlation-based methods [Faugeras et al., 1993; Scharstein and Szeliski, 2002; Hu and Mordohai, 2012] to more sophisticated approaches aiming to confront limitations such as fronto-parallel bias. Such methods refer to using predefined planes like the plane-sweep method [Collins, 1996; Gallup et al., 2007] or using local adaptive planes for each window [Bleyer et al., 2011; Hosni et al., 2012]. In order to achieve sub-pixel accuracy, least-square methods have also been proposed already since the early years [Grün, 1985].

Global methods. Global methods typically skip the cost aggregation step and form, instead, an energy function for the whole image, enforcing spatial consistency between neighboring pixels. This global cost function is defined with a unary and a pairwise term, and the best disparity value that minimizes it is sought. Let d be a possible disparity solution:

$$E(d) = E_{Data}(d) + \lambda E_{Smooth}(d). \tag{2.25}$$

E_{Data} is the data or unary term expressed as a photo-consistency metric (e.g., AD, NCC, etc.) and practically indicates how well the solution agrees with both images. E_{Smooth} is the smoothness term, incorporating the smoothness assumptions of each algorithm. The unary term alone simulates nearest neighbor (NN) search like in local methods, yet is performed pixelwise. The first order smoothness term or pairwise (i.e., implying pixel interaction) term E_{Smooth} encourages similar disparity values between neighboring pixels, enforcing spatial consistency, and λ controls the influence of E_{Smooth} over the data term E_{Data}. This 2D optimization problem is considered NP-hard[1] for computing the exact minimum, and various global approximation algorithms have been proposed to minimize the energy function, typically based on a Markov Random Fields[2] (MRF) formulation. Early methods used simulated annealing [Barnard, 1989] or tried to solve the problem as independent scanline optimization [Birchfield and Tomasi, 1998], reducing the complexity to 1D or using dynamic programming and imposing a piecewise smoothness constraint along the scanline [Ohta and Kanade, 1985]. Yet, since spatial consistency in the 2D space is disregarded, such methods inevitably suffer from lack of coherence, the so-called "stair-case" effects.

The energy minimization problem was extended by adding 2D smoothness constraints, implying local smoothness and penalizing discrepancies [Scharstein and Szeliski, 1998]. Graph cut methods for stereo or multi-view [Boykov et al., 2001; Kolmogorov and Zabih, 2002; Wei and Quan, 2005] employ a min/max cut 2D optimization, with the nodes of the graph representing the pixels and the edges their connections with the neighbors. The multi-labeling problem, where each "label" is actually a disparity value, is solved iteratively by α-expansion. Yet, the solution is often limited to integer disparities, and the method is computationally costly. Other approaches were based on belief propagation (BF) [Pearl, 1988] to employ 2D optimization on a graph [Sun et al., 2003; Felzenszwalb and Huttenlocher, 2006; Klaus et al., 2006; Besse, 2013]. BF solves the labeling issue by iteratively passing messages between neighboring pixels to find the optimal surface. Sun et al. [2003] formulate the stereo matching problem using MRFs and estimate the optimal solution using BF. Acknowledging the inefficiency of the problem, some BF methods aim at improving the computational time [Yang et al., 2008; Felzenszwalb and Huttenlocher, 2006]. According to Tappen and Freeman [2003], both graph cut and BF demonstrate comparable performances while solving the same MRF. Different from discrete labels in probabilistic methods, variational approaches form the problem in the continuous space; for instance, partial differential equations (PDE) have been used in the past in the multi-view scenario [Faugeras and Keriven, 1998; Strecha et al., 2004]. Inspired by optical flow, Ranftl et al. [2012] adopted a variational inference model for energy minimization in stereo matching. Later dynamic programming methods try to solve the problem with both vertical and horizontal scanlines applied on a tree structure to reduce the stair-case artifacts (also referred to as streaking artifacts) [Veksler, 2005]. In general, global methods are more computationally costly, but the quality is typically superior to the local methods as they perform relatively better in textureless areas due to the incorporated smoothness prior to the smoothness term.

[1] States for non-deterministic polynomial time. NP-hard problems are at least as hard as any NP one.

[2] In stereo matching, each pixel is represented by a graph node. Neighboring pixels are expected to have similar disparities. This assumption is modeled using the conditional relationships between proximate nodes to enforce the smoothness. The problem is modeled as an indirect graph, and there is no guarantee for convergence.

2.2.4 Disparity refinement and filtering

Depth maps for each reference image are the final outcomes of the depth estimation process, assigning the chosen disparity to each pixel. In a typical stereo matching pipeline, disparity refinement is performed as a post-processing step to improve the quality of the estimated depth values. Disparity refinement generally may refer to sub-pixel refinement, outlier removal, occlusion handling, left-right consistency check, gap interpolation, or confidence check. Most stereo methods operate in discrete space, and correspondences are located at integer pixel locations. To obtain floating-point sub-pixel accuracy, quadratic functions, i.e., curves, are used for interpolation over the cost volumes [Di Stefano et al., 2004; Mizukami et al., 2012] or least-squares correlation [Grün, 1985] should be performed. Depth filtering, also known as speckle filtering, is also commonly applied during the refinement step to exclude outliers. Simple techniques such as median filter [Birchfield and Tomasi, 1999; Hosni et al., 2012], mean filter, Gauss filter, or Difference of Gaussians (DoG) filter are applied. A rather challenging issue refers to occlusion handling, i.e., detecting the areas that are not visible in both views, which tend to be assigned with erroneous, over-smoothed estimates [Egnal and Wildes, 2002; Sun et al., 2003; Yang et al., 2008], yet this problem is more extensively studied under the multi-view scenario (see Section 2.5.2). Another common refinement step is the left-right consistency check, which, as implied by the name, enforces constant disparity values between the depth maps for both the left and right image to filter out inconsistent depth values and limit the erroneous estimates. Gap filling is also commonly employed by interpolation [Hirschmuller, 2008]. Finally, confidence checks can improve the quality of the depth maps, filtering out the depth estimates with low confidence [Hu and Mordohai, 2012]. A representative approach to successfully incorporating the various refinement steps into the same pipeline can be found in Hosni et al. [2011].

2.2.5 Limitations and challenges in stereo matching

Stereo matching is a highly ill-posed problem and therefore relies on several constraints for its efficient implementation. First and foremost, the *epipolar constraint* is taken into consideration as it limits the search space for the correspondence problem. Consequently, errors in the camera geometry are inherited from the stereo correspondences; some works have tried to evaluate this influence [Scharstein et al., 2014], although, in most scenarios, the geometry is considered to be known with high fidelity. The *visual similarity constraint* is a basic underlying assumption of stereo matching, implying that two correspondences appear similarly across the two views. Surface properties and scene conditions may cause corresponding pixels to be dissimilar or, in contrast, matching ambiguities may occur. Indeed, in the ideal scenario, the *uniqueness constraint* requires that a point in one image would have at most one correspondence in the other one. The *continuity constraint* assumes that disparity values are piecewise smooth. Local methods model this constraint by using a constant disparity window while global methods incorporate it in an energy function; however, this does not model sufficiently the underlying geometry as it does not hold in the presence of slanted surfaces or in the case that the window falls within surface boundaries. The *visibility constraint* is used to avoid physically impossible matches due to occlusions; a pixel in the left image should be visible in both images if it has at least one match in the right image. The visibility constraint is more flexible than the uniqueness constraint as it allows many-to-one matching [Sun et al., 2005]. Finally, the *ordering constraint* assumes surface continuity and preserves the relative ordering of pixels along corresponding epipolar lines, particularly useful in dynamic programming approaches [Ohta and Kanade, 1985]. In an ideal stereo case, all the aforementioned constraints should be fulfilled. Nevertheless, in real-world scenarios, inevitably, some of them cannot be satisfied. In the following paragraphs, the most common issues and challenges are discussed.

Challenging surfaces. The fundamental underlying assumption in conventional stereo matching is that a point in the 3D space will visually appear similar in the two images (or more, in the multi-view case), i.e., they will be photometrically consistent. This hypothesis is satisfied for Lambertian or near-Lambetrian surfaces where, indeed, good results are typically achieved. Yet, visual similarity generally cannot be fulfilled in cases of specular, reflective, or transparent objects; the same holds for textureless surfaces and repetitive patterns, since multiple ambiguous local minima may occur, resulting in noisy estimates or information gaps. Local methods, in particular, rely on photometric consistency metrics alone and therefore tend to be uninformative and inadequate on such challenging surfaces where matching ambiguities occur. Indeed, photometric consistency is ambiguous in such surface scenarios since large depth variations may lead to small cost changes due to appearance similarity. The consequences of this assumption are more evident in the 3D reconstruction result rather than in the quality of the depth maps. Generally speaking, an optimal stereo matching algorithm would take into consideration not only the photometric consistency among the images but also the natural coherence, typically described by local surface smoothness. Designing such a robust metric that will encapsulate the above constraints is a non-trivial task, and thus, typically, regularization approaches or cost volume filtering [Hosni et al., 2011, 2012] are required thereafter to optimize the results. Global and semi-global methods, on the other hand, formulate the problem as an MRF optimization and adopt a smoothness term to enforce spatial consistency and reassure smoothness by penalizing disparity changes [Kolmogorov and Zabih, 2002; Sun et al., 2003; Hirschmuller, 2008]. Although achieving satisfying results, even these methods are still generally incompetent in textureless areas. In fact, despite being an active research field for decades, algorithms still struggle with challenging surfaces and generally yield inaccurate depth estimates and incomplete reconstructed point clouds in these areas. A research direction proposed adaptive support windows to cope with this problem in the stereo case [Kanade and Okutomi, 1994; Fusiello et al., 1997; Yang et al., 2008; Stentoumis et al., 2014].

Fronto-parallel bias. In conventional local matching approaches, the principal implicit assumption is made on constant disparity within the same window around a pixel. In other words, all pixels within a support window should have the same distance from the camera, i.e., they would lie on a fronto-parallel surface, that is, a surface parallel to the sensor plane and the epipolar lines. In practice, this assumption is unlikely to hold; the window may contain pixels that belong to a different surface than the center pixel in the case of physical surface edges. Besides, the window may lie on a slanted scene surface and, thus, not fronto-parallel. Hence, such solutions cannot perform efficiently in real-world scenarios as they are prone to errors. Most local methods try to tackle this issue during the cost aggregation [Zhang et al., 2008; Hosni et al., 2012]. In global and semi-global optimization techniques, on the other hand, the cost is measured on a pixel basis, yet the fronto-parallel assumption underlies the first-order smoothness term, introducing again fronto-parallel bias. Generally, slanted surfaces are not supported under these scenarios of stereo matching, as they are mapped and sampled differently across the views. Therefore, most methods based on smoothness assumptions and other fronto-parallel shape priors such as scanline optimization and semi-global methods [Hirschmuller, 2008] (see Section 2.3) also suffer from this problem. To compensate for this challenge, least-squares methods (LSM) based on affine transformations [Grün, 1985] have been proposed. Second-order smoothness terms have also been suggested in the literature [Ishikawa and Geiger, 2006; Woodford et al., 2009]; however, using such a complex pairwise model adds a significant computational expense. In the same line of thought, some works have also considered curved surfaces [Lin and Tomasi, 2003; Li and Zucker, 2008]. Segmentation-based approaches have also been implemented, approximating second-order priors with segmented image parts that are piecewise smooth, yet typically hard constraints are applied, leading to rough results [Hong and Chen, 2004]. Adaptive support weights can help to manage the different disparity values included in every window [Yoon and Kweon, 2006; Bleyer et al., 2011]. However, the most efficient methods for tackling the fronto-parallel bias are plane sweeping, based on direct wrapping of the image to one common plane [Collins, 1996; Gallup et al., 2007], discussed in Section 2.4, and local plane fitting based on the PatchMatch algorithm [Barnes et al., 2009], extensively explained in Chapter 3.

Occlusions. Occlusion detection and handling is yet another particularly challenging task in stereo correspondences, either in two-view or multi-view scenarios; it aims to avoid the establishment of false correspondences due to occlusions. Occluded pixels are the ones visible only in one image, whose depth inevitably cannot be reconstructed due to the violation of the visibility constraint. This can commonly occur around natural object edges, where background pixels are occluded by the foreground ones. Depending on the algorithm and the extent of the occluded area, artifacts, erroneous depth estimates, or information gaps may appear in such regions. Approaches have been proposed to confront this issue mainly globally but also locally [Sun et al., 2005; Hu and Mordohai, 2012; Hosni et al., 2012]. This topic has been thoroughly studied also under the multi-view scenario with the formulation of visibility models (see Section 2.5.2 for details).

Other challenges. Apart from the aforementioned challenges and limitations, other open issues in stereo matching are more relevant to the implementation of the algorithms and their performance. For instance, computational complexity is a problem encountered mostly in global methods since local methods are, by definition, cheaper. This is mainly due to the expensive global cost volumes and the need to search throughout the whole disparity range. Indeed, a pre-defined range $[d_{min}, d_{max}]$ should be rigorously selected and adaptively tuned for each application, e.g., as in [Wenzel, 2016; Rothermel et al., 2012]. To cope with the computational burden of global cost volumes, hierarchical approaches have been proposed in the literature applying Gaussian scale-space image pyramids [Quam, 1987]. Such methods work in multi-resolution schemes to guide the search from coarse to fine scales and consequently reduce the search space and thus the computational cost [Hirschmuller, 2006; Rothermel et al., 2012]. Dynamic cost structures have also been exploited [Rothermel et al., 2012]. Finally, most conventional stereo algorithms operate in the discrete disparity space and account for full integer (quantized) disparity values only; consequently, they do not achieve sub-pixel accuracy without further refinement and interpolation step [Yang et al., 2008].

2.3 Semi global matching

In the seminal work of Hirschmuller [2006, 2008], semi-global matching (SGM) was introduced as a hybrid solution between global and local methods. It approximates a global optimization inspired by scanline optimization (NP-hard 2D graph partitioning) [Ohta and Kanade, 1985] and dynamic programming [Veksler, 2005] at a lower computational cost. The pixelwise matching cost and the smoothness assumptions are formulated similarly to the global methods into one energy function $E(D)$ as a 2D MRF problem. This MRF is approximated; using up to 16 independent cardinal directions $r = \{(0,1), (0,-1), (1,0), \dots\}$ along 1D scanlines. Optimized by dynamic programming to minimize the energy function and enforce smoothness, it provides a trade-off between computational complexity and quality.

The original SGM proposes to use the MI metric [Viola and Wells III, 1997] for initial matching cost computation due to its robustness to gain and bias and radiometric changes. For disparity estimation, as most standard global minimization approaches, SGM defines the energy function for the disparity image D:

$$E(D) = \sum c(\mathrm{p}) + \sum_{\mathrm{p},\mathrm{q}} T(\mathrm{p}, \mathrm{q}). \tag{2.26}$$

The first term is the sum of the photometric consistency costs c at all pixel locations p for a disparity $d \in D = \{d_{min}, \dots, d_{max}\}$. Nonetheless, the second term in SGM represents the smoothness as a combination of two linear truncated penalty terms:

$$T(\mathrm{p}, \mathrm{q}) = \begin{cases} 0, & \text{if } |D_\mathrm{p} - D_\mathrm{q}| = 0 \\ P_1, & \text{if } |D_\mathrm{p} - D_\mathrm{q}| = 1 \\ P_2, & \text{if } |D_\mathrm{p} - D_\mathrm{q}| > 1. \end{cases} \tag{2.27}$$

This pairwise term aims to penalize disparity changes between neighboring pixels. P_1 is a constant penalty for locations q in the neighborhood of p if they have small disparity discontinuities $|D_\mathrm{p} - D_\mathrm{q}| = 1$. P_2 adds a larger constant penalty for all larger disparity changes $|D_\mathrm{p} - D_\mathrm{q}| > 1$.

The computation of the optimal disparities is done in two steps. A disparity space image is first constructed, which is actually a cost volume (cubic cost structure), with d being a discrete value in a constant range D defining the potential correspondences along the epipolar line. For k discrete levels of disparity, the matching cost forms a 3D disparity volume of size $W \times H \times D$. In a second step, all costs along several image paths, typically 8 or 16 (Figure 2.6), are accumulated and stored in another 3D volume of the same dimensions. This cost volume also contains noisy depth estimates as wrong hypotheses can potentially have better scores than the correct ones, especially in the presence of textureless surfaces. Therefore, additional prior information about the scene surface should be considered, like local smoothness in the same fashion as in the global methods. However, in SGM, the smoothness constraint is approximated by the recursive aggregation on multiple linear path directions (cost aggregation) that can be horizontal, vertical, and diagonal. It is enforced by adding penalties for high disparity differences (jumps) for each path and aggregating them recursively for each direction. As expected, the more the paths are, the better the cost function approximation, yet typically eight paths are chosen. The aggregated costs are summed for each pixel, resulting in an aggregated cost volume $\mathbf{S}_{agg}$. The per-pixel minimum cost in the aggregated cost volume is finally chosen as the winning disparity value $\hat{d}$ based on the Equation 2.24.

In Hirschmuller [2006, 2008], sub-pixel accuracy is achieved by cost approximation using a second-order function. A second-order function is fitted to the minimum

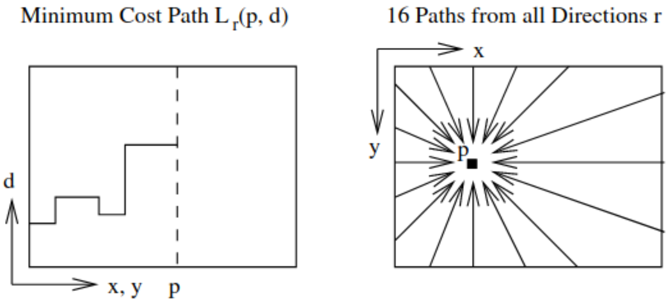


Figure 2.6: **Cost aggregation in SGM.** Left: minimum cost path, right: the 16 considered paths for cost aggregation. Source: [Hirschmuller, 2008].

cost and the costs of two neighboring disparities (limited approximation). Additional methods like LSM could be potentially applied for better results. A left-right consistency check is also commonly performed so that for the same pixel, the disparity difference between the left and right disparity images does not exceed a certain threshold. Disparity values that exceed this threshold are removed from the final depth map. Originally, SGM was developed to process image pairs; for multiple view scenarios, the algorithm can be extended by calculating matching costs across multiple images. To efficiently treat the occlusions issue, however, in multi-view SGM scenarios, it is preferred to fuse the pairwise disparity maps after a pairwise occlusion filtering. Further disparity filtering, e.g., using median filtered, is commonly performed.

2.3.1 Limitations of SGM and further improvements

Optimization via SGM is a robust solution for accurate depth estimates and has been widely used in applications like mapping [Hirschmuller, 2008; Rothermel et al., 2012], assisted driving [Gehrig et al., 2009] or robot navigation [Schmid et al., 2013] while being particularly famous within the photogrammetric community. However, as a global method, the implication via the first-order smoothness term of a fronto-parallel shape prior limits the applicability of the algorithm when strongly slanted surfaces are present. Although this is not the case for nadir airborne images, it is particularly true in close-range applications with wide baselines. Similarly, matching ambiguities due to multiple local cost minima may occur in large, low-textured areas, repetitive patterns or generally non-Lambertian surfaces. In the original SGM scenario, such ambiguities are partially handled by the smoothness term interpolating depth values by the neighboring pixels, yet, often, various inconsistencies occur. Moreover, cost aggregation from the different paths and the WTA-fashioned depth estimation may be problematic when costs along the various paths are inconsistent [Schönberger et al., 2018]. Summing the costs over the various paths was originally introduced to avoid the stair-case effect [Hirschmuller, 2006], but it has proven inefficient for handling matching ambiguities and slanted surfaces. Indeed, as explained in Schönberger et al. [2018], when the photometric cost is unreliable, the smoothness term tends to propagate equally likely depth estimates along the propagation direction. Especially in the case of slanted textureless areas, there is no clear outlier, and cost summation leads to a biased estimate and consequently to noisy depth maps. Also, during occlusion handling, the smoothness constraint is likely to propagate correct depth hypotheses only when the occluded region is fronto-parallel. Several methods have been proposed in the literature to efficiently tackle the stair-case effects, like using BP and truncated smoothness terms [Facciolo et al., 2015], individually weighted penalties for each path [Michael et al., 2013], or weights based on surface priors [Spangenberg et al., 2013]. More recently, random forests are used to predict per-pixel weights for each path [Poggi and Mattoccia, 2016] or for the efficient fusion of disparity proposals [Schönberger et al., 2018].

The original SGM implementation suggested using a constant P_2 penalty value or defining it as a function of the gradient of the image between the current and the previous pixel. Zhu et al. [2011], based on this formula, also introduced a weight to adjust the penalty value. Banz et al. [2012] further elaborated on this idea and introduced three more parameters for P_2. Other works extended the approach by incorporating second-order terms [Hermann et al., 2009; Ni et al., 2018]. Explicit priors have been used to confront matching ambiguities, either in a simple plane fitting fashion [Humenberger et al., 2010; Sinha et al., 2014] or by incorporating surface normals as soft constraints by penalizing the deviation of the calculated surface orientation from the assumed prior [Scharstein et al., 2017].

SGM requires a disparity range prior $[d_{max} - d_{min}]$ and is, consequently, computationally expensive when large-scale variations are present in the scene since every disparity in the cost volume needs to be evaluated for each pixel. That being said, the assumption of constant disparity range for the entire scene may not be an issue for airborne nadir images, yet it is not the case for close-range high-resolution images with large disparity ranges. Hirschmuller [2008] suggested processing the large images in tiles for a more efficient implementation in a coarse-to-fine scheme to initialize and update the matching cost. Later, Hirschmüller et al. [2012] suggested a quite simplified yet memory-efficient version of SGM, storing only the minimum costs for each pixel. Related literature has also proposed approaches to ease the memory requirements, like dividing the image into stripes [Humenberger et al., 2010; Spangenberg et al., 2014]. Calculating disparities in lower resolution and using them as a prior has also been popular for reducing the search space and easing the memory requirements [Gehrig et al., 2009; Hermann and Klette, 2012]. Similar to other methods, also in SGM hierarchical approaches have proven to be efficient for complexity reduction. Image pyramids are used in a coarse-to-fine scheme, and disparity ranges from lower resolution levels are passed to the higher ones to narrow down the search space. The works of Rothermel et al. [2012] and Wenzel et al. [2013] proposed an extension by a hierarchical strategy that allows for arbitrary depth variation and resolution in a more efficient way; the cost volume has a dynamic structure with an individual range for each pixel rather than fixed dimensions. Toward the same goal, Sinha et al. [2014] combined local plane sweeping (see Section 2.4) and SGM to confine the disparity ranges and reduce the computational complexity. Bethmann and Luhmann [2014] performed cost calculation and optimization in the object space using a voxel structure and facilitating the computational process in multi-view scenarios.

2.4 Plane sweeping

The plane sweeping algorithm was originally introduced by Collins [1996] as a method to sparsely match multiple images that are not required to be stereo rectified. For each depth, the source images are projected onto fronto-parallel planes upon the camera frustum of the reference image. Yang and Pollefeys [2003] further elaborated on this approach for dense depth estimation using the GPU, but the method gained particular popularity when Gallup et al. [2007] extended the approach to explicitly handle multiple slanted planes; they demonstrated that robustness is improved when the sweeping plane is aligned with the predominant directions of the scene in multi-view scenarios. Plane sweeping is, in principle, a local stereo method. Yet, in the approach of Gallup et al. [2007], the fronto-parallel bias, commonly present in stereo methods, is undertaken since, for each pixel, a different plane direction may be assigned. In more detail, first, the scene's principal plane orientations, i.e., the surface normals, are identified using the priorly calculated sparse points, then it estimates the depth by sweeping a family of planes along each normal direction generating multiple depth hypotheses for each pixel. The best depth and normal values combination is selected as the correct plane hypothesis. Multiple plane sweeps are performed in order to reconstruct planar surfaces having a particular normal. Sweeping through a series of disparity hypotheses corresponds to warping or projecting each input image onto the virtual planes. Photo-consistency values are evaluated for each plane. For cost computation, standard photo-consistency measures are applied like absolute differences (AD), sum of absolute differences (SAD), and normalized cross-correlation (NCC). Instead of building a cost volume by matching patches across epipolar lines like in other stereo methods, plane sweep-based methods build a cost volume for a sampled set of plane hypotheses of the 3D scene. However, plane hypothesis direction is important; its normal should match the actual surface direction to minimize distortions on the resulting warped image and support reliable and photo-consistent matches. Among the strengths of the algorithm is the real-time performance, yet inaccurate results may occur in general camera configurations. It works well in urban scene reconstructions, where dominant plane directions are present in the scene. However, plane-sweep is computationally consuming in multi-view and high-resolution scenarios and typically relies on GPU optimization for time-efficiency.

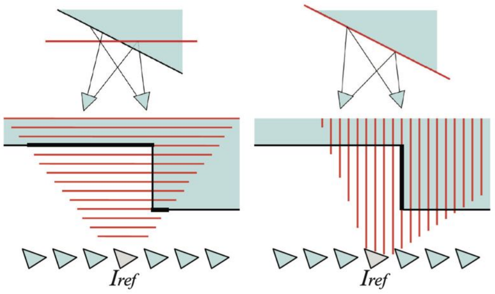


Figure 2.7: **Sweeping planes.** Fronto-parallel (left) and slanted sweeping planes (right) extracted from the sparse point cloud for the reference image I_{ref}, as introduced in [Gallup et al., 2007]. Source: [Furukawa and Hernández, 2015].

Plane sweeping is an efficient solution and sufficiently recovers the depth of rich textured areas; however, in problematic textureless areas, other engineered optimization solutions like graph cuts [Kolmogorov and Zabih, 2002], belief propagation [Sun et al., 2003] or cost filtering [Hosni et al., 2012] can obtain better results. Dominant plane orientation is used to assist stereo reconstruction in the literature, like [Pollefeys et al., 2008] or using the Manhattan World assumption, i.e., scenes that consist of piecewise planar surfaces with dominant directions [Furukawa et al., 2009] or focused on architectural scenes [Cornelis et al., 2008]. Sinha et al. [2009] combined plane candidates and 3D line segments and computed piecewise planar depth maps using energy minimization.

2.5 Multiple view stereo

Multiple (or multi-) view stereo (MVS) algorithms address the problem of generating a dense and complete 3D representation of the scene, using multiple ($i \geq 3$) overlapping images of known viewpoints simultaneously, generalizing the standard stereo case for more than two views. In the last decade, MVS methods have demonstrated a great potential to efficiently reconstruct complex and large scenes and hence have been a hot research topic in photogrammetry and computer vision.

MVS basically relies on the same principles as classic binocular stereo; however, the problem cannot be extended directly from the two-view to the multi-view scenario since, for the estimating the depth of one reference image, correspondences across multiple source images should be considered simultaneously, enforcing photo-consistency between them (multi-photo-consistency). This redundancy is an essential part of the several MVS methods, as it is utilized for more efficient cost aggregation, tackling the commonly occurring matching ambiguities of two-view scenarios and yielding more robust depth estimates. Trivial stereo matching for all possible pairs with a subsequent fusion of disparities or 3D points can be followed, yet the multi-view redundancy will not be fully exploited. Indeed, early approaches try to solve the problem as multiple pairwise matching in order to utilize the well-known epipolar constraints from the standard stereo case [Grün and Baltsavias, 1988; Okutomi and Kanade, 1993; Kanade et al., 1995] while similar constraints can be used based on the trifocal tensor for image triplets [Fitzgibbon and Zisserman, 1998; Hartley and Zisserman, 2003]. In recent years, enforcing the consistency across multiple images simultaneously is one of the open challenges in MVS research and is typically evaluated in sophisticated ways as a function of scene geometry, viewpoints, materials, and illumination.

Classic stereo approaches typically handle small baselines; on the contrary, one fundamental property of MVS algorithms is their ability to handle arbitrary varying viewpoints, i.e., drastic angle and scale changes that inevitably cause geometric and radiometric distortions and, thus, increase the difficulty to find corresponding patches. However, such strongly variant viewpoints also inherit occlusions that need, in turn, to be identified based on visibility reasoning, i.e., visibility models and suitable neighboring view selection for optimal appearance consistency, also considering the scene complexity. MVS methods typically can handle images that are not stereo rectified; in fact, in multi-view cases, searching for correspondences along aligned epipolar lines can be performed successfully only in some special camera configurations to avoid excessive deformations, e.g., aligned cameras centers [Nozick, 2011]. Actually, in multi-view camera networks, often the camera network geometry is such that the epipoles lie within or are particularly close to the image borders, where epipolar rectification typically fails [Häne et al., 2014].

In multi-view scenarios, for the image set $\mathcal{I}$ the camera poses $(\mathcal{R}, t)$ along with a sparse representation of the scene are obtained by priory performed Structure from Motion (SfM) pipelines [Snavely et al., 2006; Agarwal et al., 2011; Schönberger and Frahm, 2016]. Thus, the relative geometry for every camera in the 3D space is known $\mathbf{R}, \mathbf{t}$, while a different calibration matrix $\mathbf{K}$ for each camera may also be considered. Additional scene cues such as lighting [Langguth et al., 2016], color [Gallup et al., 2010], scene structure [Furukawa et al., 2009], and semantics [Häne et al., 2013; Blaha et al., 2017] have been incorporated to improve the quality of the multi-view 3D reconstruction; yet, the joint use of all of them is non-trivial and remains an open challenge.

2.5.1 Taxonomy in MVS

Seitz et al. [2006] proposed a taxonomy based on six criteria to categorize the MVS algorithms according to: scene representation, photo-consistency measure, visibility model, shape prior, reconstruction algorithm, and initialization requirements. Scene representation may be described by depth maps, level sets, triangulated meshes, or voxels. Depth maps are considered point-wise representations since depth reprojection leads to the 3D point cloud equivalent [Galliani et al., 2015].

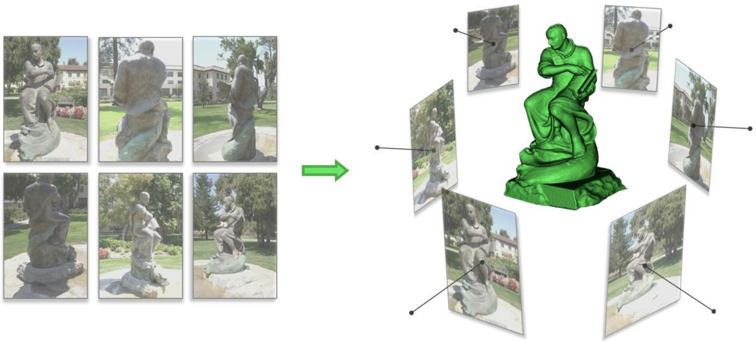


Figure 2.8: **An example MVS reconstruction pipeline**, where a 3D model of the scene is reconstructed using multiple overlapping images of known poses. Data: Ignatius, *Tanks and Temples* benchmark [Knapitsch et al., 2017].

Photo-consistency measures refer to the choice of the matching cost metric, e.g., NCC, SAD, MI (see also Section 2.2.1), and can be enforced in the image space or the object space. The visibility model connects each point with the cameras; it is the way the occlusions are handled and is used accordingly to select appropriate stereo pairs. Shape priors or assumptions about the surface geometry can be used in cases of limited redundancy and to support local planarity in case of ambiguities (e.g., due to the lack of texture). The reconstruction algorithm is the series of steps followed for depth estimation and geometry retrieval. Finally, initialization requirements represent the necessary initial information for the reconstruction, such as the bounding box, a pre-defined disparity range, or other scene priors.

Since this taxonomy was introduced, algorithms have significantly evolved, and the vast literature on MVS reconstruction has grown rapidly, and keeping clear boundaries between methods is tough. Other categorization schemes exist, such as the one proposed by Aanæs et al. [2016], who divided MVS algorithms into three main categories: point cloud-based, volume-based, and mesh-based methods. However, within this dissertation, the popular categorization based on their reconstruction algorithm is adopted, according to which MVS algorithms may refer to (1) voxel-based methods, (2) surface evolution-based methods, (3) feature point growing-based methods, and (4) depth map-based methods. Some methods, combining different steps and procedures, can potentially fit into more than one category.

Volumetric reconstruction. Volumetric mesh reconstruction in multi-view, introduced in the seminal work of Curless and Levoy [1996], was first applied in the computer graphics field; it may have as input 3D information deriving from different sources such as 3D volumes, depth maps, or point clouds. The general underlying idea is to leverage ray visibility information, connecting the 3D space with the camera poses, to understand which part of the scene is free-space and which part is matter, i.e., occupied. Among the first approaches to recover the object geometry from arbitrary cameras were methods assigning occupancy information to each voxel while considering photo-consistency like space carving [Kutulakos and Seitz, 2000] and voxel coloring [Seitz and Dyer, 1999].

Volumetric approaches first compute a bounding box, then divide the 3D space into a regular voxel grid and compute a cost function, typically based on photometric consistency metrics [Zach et al., 2007; Zach, 2008] or also exploiting silhouette constraints [Cremers and Kolev, 2010]. Image visibility information is kept and inherited, and costs represent the possibility of the voxel being part of the surface. A 3D volume is calculated from which the optimal surface will be extracted using, e.g., graph cuts [Hornung and Kobbelt, 2006; Vogiatzis et al., 2007; Hernández et al., 2007; Sinha et al., 2007] or the signed distance function will be calculated [Newcombe et al., 2011b; Werner et al., 2014] posing the surface reconstruction as zero iso-surface extraction problem [Curless and Levoy, 1996; Zach et al., 2007]. Some works extend this approach by modeling the visibility along the full viewing ray for each pixel, using an MRF formulation [Liu and Cooper, 2010; Savinov et al., 2016; Ulusoy et al., 2015, 2016]. Ray potentials efficiently model occlusions and enforce consistency; however, higher-order ray potentials are formed per pixel and increase the memory consumption. Further regularization in volumetric integration has also been investigated using the Total Variation norm (TV-L1) [Zach, 2008; Häne et al., 2012] to suppress noise and reject outliers. Multi-scale extensions have been proposed to handle varying scale depth maps using hierarchical signed distance fields [Fuhrmann and Goesele, 2011] or independent subsets [Kuhn et al., 2013]. For real-time applications, low-resolution RGB-D images of the Kinect sensor have been used [Newcombe et al., 2011b]. Nonetheless, volumetric methods have the drawback that space discretization per se is memory-consuming. Moreover, the reconstruction accuracy is restricted to the voxel size; hence scalability issues occur for large-scale scenes with high accuracy requirements. In order to skip the requirement for predefined scene extent, other approaches deviate from the regular partitioning of the volume using voxels; they first reconstruct a sparse point cloud that will be converted to an irregular mesh via Delaunay Tetrahedralization by solving a volumetric MRF [Labatut et al., 2007; Sugiura et al., 2013]. The tetrahedrons are then labeled empty or occupied, typically formulated as a graph-cut problem. The idea was extended by [Jancosek and Pajdla, 2011, 2014], starting from an input point cloud and exploiting visibility to also recover the so-called "weakly-supported" objects. These methods are inherently more scalable than the voxel-based ones since Delaunay triangulation adapts its density according to the point cloud.

Final surfaces are produced from volumetric representations using Poisson triangulation from oriented point clouds [Kazhdan et al., 2006; Kazhdan and Hoppe, 2013] or the marching-cubes algorithm [Lorensen and Cline, 1987]. However, such approaches usually generate a rough surface which can be further refined by using photo-consistency metrics on the images to recover scene details [Jancosek and

Pajdla, 2011]. Optimization in volumetric approaches works in object space and can be limited by the memory availability, also restricting its applicability in large-scale and high-resolution scenes. More recently, other methods have tried to incorporate semantic cues in volumetric reconstruction, treating it as a joint problem of reconstruction and semantic segmentation [Häne et al., 2013; Savinov et al., 2016; Blaha et al., 2017]. Recent deep-learning methods also adopt such scene representations [Paschalidou et al., 2018], as discussed in Section 2.6.

Surface evolution. Surface evolution-based methods require a good initial guess to initialize the surface evolution and iteratively improve it guided by multi-view photometric consistency in image space. The initial surface guess can be as rough as a triangulated version of a sparse point cloud [Hiep et al., 2009] or a visual hull [Furukawa and Ponce, 2006], but also a dense point cloud or a mesh generated with volumetric methods. Hence, surface evolution can potentially be added as a final step in any MVS pipeline to refine the mesh representation and recover details. The images are again involved in this step to guide the refinement based on photo-consistency, and the vertex positions are optimized to minimize the reprojection error.

Visual hulls, introduced by Laurentini [1994] were one of the first approaches to infer shapes from images, separating the background objects relying on silhouettes, i.e., contours. Level-set methods were also initially used in the early steps of these approaches for variational refinement [Faugeras and Keriven, 1998; Pons et al., 2007], employing partial differential equations (PDEs) but had a prohibitive computational and memory cost. Space carving can also be considered a surface evolution method since it progressively eliminates inconsistent voxels from an initially reconstructed volume [Kutulakos and Seitz, 2000]. Silhouettes defining a visual hull have been integrated with texture in a deformable model by Hernández and Schmitt [2004] and Furukawa and Ponce [2006] yet these methods have limited applicability. In other methods, $s-t$ cut global optimization is used for a visibility consistent initial volume, and details are recovered via variational mesh refinement [Labatut et al., 2007; Hiep et al., 2009]. Jancosek and Pajdla [2011] added a visual hull component to recover challenging surfaces, starting from a given point cloud. Delaunoy and Prados [2011] also considered visibility constraints combined with gradient flows, and Delaunoy and Pollefeys [2014] formulated the problem in a geometric bundle adjustment. [Li et al., 2016] divided the scene into significant and insignificant regions to recover details adaptively. More recently, Romanoni et al. [2017] optimized camera view selection to treat occlusions during model refinement. Variational multi-view mesh refinement formalizes the disagreement between the triangulated mesh and the image data in an energy function. This energy is typically minimized with gradient descent and produces highly detailed results [Vu et al., 2011; Li et al., 2016; Heise et al., 2015]. A drawback of surface evolution methods is the requirement of a reliable enough initial surface which is usually not feasible for real-world outdoor scenarios. In this regard, Cremers and

Kolev [2010] define a convex functional minimization problem that does not need initialization.

Feature point growing. Another research branch casts the refinement on patches, i.e., pairs of normal and depth, instead of surfaces via mesh evolution. Feature point growing methods, also known as patch-based algorithms, initially parse point clouds based on discriminative features on highly textured areas and subsequently expand them to reconstruct the whole scene instead of reconstructing each point independently. PMVS is a seminal patch-based approach introduced by [Furukawa and Ponce, 2009], achieving groundbreaking results in its time, further developing on the idea of Lhuillier and Quan [2005], who introduced a quasi-dense approach. Small rectangular patches (i.e., points with local region support) are used as local approximations of a tangent plane, generated by sparse feature correspondences, and are grown repeatedly in a region-growing fashion. Sparse features are matched, then expanded, creating a dense set of patches, and finally filtered iteratively while enforcing local photometric consistency across multiple views and global visibility constraints. Patches and correspondences are stored in a grid structure, consisting of grid cells, for every image. With this seed-and-expand scheme, the patches are densified until each grid cell is full. No explicit regularization is applied, yet patches are reconstructed sufficiently complete and smooth. Oriented point clouds are generated, and, optionally, a surface can be reconstructed with standard methods such as Poisson reconstruction and refined using energy minimization with geometric smoothness and photometric consistency terms. In PMVS, no surface initialization is required, and normals are also considered to avoid fronto-parallel bias. A drawback of such an approach is the scalability problems, as they tend to reconstruct the complete scene at once. In order to handle the high memory consumption of this method, CMVS was proposed [Furukawa et al., 2010] as a follow-up method using clusters. Other works have been proposed inspired by PMVS [Wu et al., 2010; Locher et al., 2016], also incorporating scene priors to overcome the textureless areas' problem [Furukawa et al., 2009]. However, among the limitations of these methods is the inefficient spread of points in textureless regions. It is to be noted that patch-based PMVS should not be confused with PatchMatch-based approaches, which are based on similar patches yet operate diversely since they generate depth maps; PatchMatch will be extensively discussed in Chapter 3, as it is a core part of this thesis.

Depth map fusion. Depth map fusion algorithms have been widely used in recent years under large-scale, high-resolution applications with high accuracy requirements due to their overall efficiency and scalability and are, thus, typically preferred over the other methods. Indeed, voxel-based methods are constrained by the predefined resolution and are thus not applicable in large-scale reconstructions, e.g., outdoor scenarios in photogrammetric applications; surface-evolution methods depend on a reliable surface initialization, and feature-based methods rely on seed points, limiting their completeness. On the contrary, depth map fusion methods decouple the complex MVS problem into per-view depth estimation tasks and subsequent fusion. They infer the depth map for each reference image considering several source images as input, simultaneously or individual pairs, enabling scalability while also yielding more robust depth estimates and 3D reconstructions in the form of point clouds [Merrell et al., 2007] or mesh representations [Curless and Levoy, 1996; Newcombe et al., 2011a; Vu et al., 2011; Heise et al., 2015]. However, depth map fusion has particular challenges, such as view selection, occlusion handling, object boundaries, and depth discontinuities.

Depth maps, containing the disparity values of the scene points and are typically derived by stereo matching methods described in Section 2.2 [Hirschmuller, 2008; Gallup et al., 2007; Bleyer et al., 2011]. For each generated depth map, given the camera poses, the 3D scene geometry can be generated by projecting the depth in the 3D space. For complex scenes, a proper depth sampling scheme needs to be adapted to maintain constant depth accuracy along with efficiency [Gallup et al., 2008] since accuracy is typically inversely proportional to the distance to the surface. Considering that for each view, one depth map is generated, and the views are overlapping, depth maps inherently also overlap. Redundancy is exploited to optimize the accuracy and completeness of the final fused point cloud. Undoubtedly, using more than two intersecting rays for each point will increase the quality of the reconstruction. Depth maps are fused together in order to derive the optimal surface representation while taking into consideration geometric or visibility criteria, e.g., intersection angles, image scale [Merrell et al., 2007; Hu and Mordohai, 2012] but also radiometric or image criteria, e.g., quality of the texture, image blur [Vu et al., 2011] to select the best viewpoints. Algorithms proposing WTA depth estimation [Hernández and Schmitt, 2004; Hu and Mordohai, 2012], truncated costs [Goesele et al., 2006], or using more robust photo-consistency metrics based on Parzen windows [Vogiatzis et al., 2007]. Enforcing consistency simultaneously across multiple views is actually a large optimization problem, while outliers and noise also need to be considered, along with smoothness. A standard approach would be to solve a linear system $\mathbf{AX} = 0$ to calculate the optimal point $\mathbf{X}$ in the 3D space while considering redundant depth estimates [Li et al., 2010]. Outliers are subsequently treated statistically; thresholds in reprojection errors are set to exclude erroneous measurements, improving accuracy. However, this solution is efficient only if the number of correct estimates is significantly smaller than the number of outliers. Global optimization methods use an MRF formulation and optimize via graph cuts [Kolmogorov and Zabih, 2002; Campbell et al., 2008] of belief propagation [Strecha et al., 2006], similarly to the ones discussed in Section 2.2. Such methods combine multiple hypotheses to improve the depth maps for multiple view stereo; first, depth labels are extracted and then assigned to the pixels with MRF optimization. Expectation maximization (EM) for joint depth and occlusion estimation has also been applied [Strecha et al., 2006; Tola et al., 2009], but such implementations generally suffer from high complexity and are accordingly limited to a small number of images. Tola et al. [2012] relied on dense DAISY features to generate pairwise depth maps in an efficient manner; redundant depth maps are merged by consistency checks among neighboring views, working on the full resolution of high-resolution images. Such dense descriptors were, at the time, a good solution to avoid matching windows and increase robustness to distortions. Depth maps can be finally combined in volumetric representations [Goesele et al., 2006; Zach et al., 2007; Fuhrmann and Goesele, 2011; Kuhn et al., 2013, 2017].

Other methods focused on processing extremely large, internet-scale collections, also exploiting GPU implementations that were revolutionary in their time [Goesele et al., 2007; Furukawa et al., 2010; Frahm et al., 2010]. Plane sweeping MVS algorithms [Gallup et al., 2007] are also depth map fusion methods (see also Section 2.4). Similarly, PatchMatch-based MVS approaches [Shen, 2013; Galliani et al., 2015; Schönberger et al., 2016; Xu and Tao, 2019], discussed in detail in Chapter 3, exploit multi-view redundancy for the computation of each depth map and perform subsequent fusion; they have been proven to work efficiently, especially when it comes to accurate depth estimation of slanted surfaces due to the usage of support windows to eliminate fronto-parallel bias.

In photogrammetric applications, traditional methods followed least-squares minimization approaches for multi-view scenarios [Grün, 1985; Helava, 1988; Grün and Baltsavias, 1988]. Recently, SGM-based approaches are typically performed for all overlapping image pairs separately, with subsequent consistency checks to eliminate outliers and fusion in the 3D space [Hirschmuller, 2008; Rothermel et al., 2012; Wenzel et al., 2013]; however such approaches, although having lower computational complexity, do not fully exploit multi-view redundancy as they enforce photo-consistency only within each pair. Another approach performing SGM directly in the object space using voxel grids has been proposed [Bethmann and Luhmann, 2014], relaxing the requirement for stereo rectified images. Potentially, the point clouds can subsequently converted to meshed surfaces with Poisson reconstruction [Kazhdan et al., 2006], e.g., in [Furukawa and Ponce, 2009] or Delaunay Tetrahedralization [Vu et al., 2011; Tola et al., 2012; Jancosek and Pajdla, 2014]. Delaunay Tetrahedralization is commonly preferred since it adapts to point density and is, thus, more scalable. Although their great applicability, depth map fusion methods commonly suffer from the inherited deficiencies of stereo matching and fail to recover information in weakly-supported regions of the scene.

2.5.2 Visibility models

Visibility models in MVS define physically impossible surface states, i.e., surfaces violating the visibility constraint, and are thus designed to identify and handle occlusions. As aforementioned, additional ray redundancy in multi-view methods is used to resolve, to some extent, the inherited occlusions of the two-view methods. However, a robust selection of the subset of best neighboring views $\mathcal{I}_{neigh}$ of the image set $\mathcal{I}$ for each reference image I_{ref} is a crucial step for occlusion handling and distortion minimization between corresponding image patches; at the same time, appearance similarity should be reassured, and suitable baselines should be chosen for accurate triangulation. Photo-consistency measures assume prior knowledge of the camera geometry but are agnostic to the captured 3D scene, creating a dependency loop (Figure 2.9); for this reason, some approaches make the initial assumption of no occlusions and re-estimate visibilities and depth iteratively [Kang et al., 2001].

In some methods, an initial surface reconstruction is used to estimate the visibility, as commonly done in surface evolution algorithms [Hiep et al., 2009; Faugeras and Keriven, 2002]. Space carving [Seitz and Dyer, 1999; Kutulakos and Seitz, 2000] also starts from an initial volume that is iteratively carved out by removing voxels that are not photometrically consistent, meaning that textureless surfaces will not be recovered. According to this model, constraints on the camera centers are imposed such that it is guaranteed that occluder voxels are visited before their potential occluded voxels, a method commonly named visibility reasoning. Other methods use simple heuristics to select the best views for every reference image, like minimum photometric costs [Kang et al., 2001; Gallup et al., 2008; Galliani et al., 2015] or viewing angles, scale differences, and baseline length criteria to restrict the number of "good" image pairs [Tola et al., 2012; Shen, 2013]. In some cases, SfM points are often used to support the visibility information [Goesele et al., 2007; Vu et al., 2011; Furukawa et al., 2010]. However, since matched images during SfM do not always imply also "good" image pairs for depth estimation, the visibility information needs to be further exploited here. Defining the best views can be formulated as a problem as such or jointly exclude the most improbable views with a clustering process to handle a vast number of images, e.g., in crowdsourced collections [Goesele et al., 2007; Furukawa et al., 2010]. Global approaches often cast the problem into an energy function considering jointly photometric and geometric consistency, and thus solving depth and visibility, for every pixel [Kolmogorov and Zabih, 2002; Strecha et al., 2006; Campbell et al., 2008; Tola et al., 2009; Savinov et al., 2016; Romanoni and Matteucci, 2021]. More sophisticated recent approaches reformulate the problem as a joint optimization for depth, normal and pixelwise view selection in a probabilistic framework based on PatchMatch MVS [Zheng et al., 2014; Schönberger et al., 2016]. More details on such approaches are given in Chapter 3.

2.5.3 Open-source implementations for MVS reconstruction

Among the pioneer works that revolutionized the field of image-based 3D reconstruction, Snavely et al. [2006] introduced Photo Tourism, later known as Bundler, an interface to sparsely reconstruct random community photos. VisualSfM was

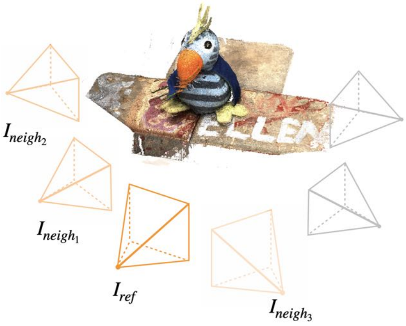


Figure 2.9: **Visibility problem.** Depth estimation algorithms assume known camera poses and visibility information, however, the scene surface is not known beforehand. Here, a specific 3D point in the object visible by the reference camera I_{ref} is seen by the $\mathcal{I}_{neigh}$ set of cameras (in orange), but is occluded for the gray ones. Data: *DTU Robotics* [Aanæs et al., 2016].

one of the first widely used GUI solutions [Wu et al., 2011; Wu, 2013] integrating the PMVS/CMVS algorithms for dense reconstruction [Furukawa and Ponce, 2009; Furukawa et al., 2010]. Another end-to-end solution offering a user interface for SfM, MVS and surface reconstruction is MVE [Fuhrmann et al., 2014].

More recently, Schönberger and Frahm [2016] and Schönberger et al. [2016] introduced two frameworks for SfM and MVS reconstructions respectively, integrated in the all-in-one solution COLMAP, which is still considered as the state-of-the-art for conventional methods, due to its high accuracy results. It encapsulates efficient implementations for global and incremental SfM reconstruction and PatchMatch stereo for MVS, while allowing for a high level of flexibility also for the intermediate tasks such as feature detection and description. However, the frameworks OpenMVG [Moulon et al., 2016] for SfM reconstruction and OpenMVS [Cernea, 2020] for MVS are also widely used and often combined together for 3D reconstruction applications [Stathopoulou et al., 2019] and have become the baselines for further research. Several other libraries offering robust SfM implementations exist, such as Theia [Sweeney et al., 2015a] or the OpenSfM library [Adorjan, 2016] and other for MVS, such as HPMVS [Locher et al., 2016], Gipuma [Galliani et al., 2016] and CMPMVS [Jancosek and Pajdla, 2011]. ACMP is a recently released open-source implementation with promising results for MVS reconstruction [Xu and Tao, 2019, 2020b].

More adjacent to the needs of the photogrammetric community, MicMac [Pierrot-Deseilligny and Paparoditis, 2006; Rupnik et al., 2017] is an open-source pipeline for SfM and MVS reconstruction enabling GCPs usage in the bundle adjustment and camera constraints. Robust commercial software solutions for photogrammetric applications also exist in the market, yet they typically rely on black-box solutions, and on that account, are not further discussed in this dissertation.

2.5.4 Limitation and challenges in MVS

Despite the overall success of the MVS algorithms, there are still some open challenges regarding the complete and accurate 3D reconstruction of a scene. The challenges of establishing dense correspondences using photometric consistency are mostly equivalent to those discussed in stereo matching (Section 2.2), as the limitations of the algorithms are naturally inherited from the two-view to the multi-view problem. However, the ray redundancy in multi-view cases generally enables a most efficient scene recovery. That being said, MVS faces other particular challenges regarding the efficient analysis of large-scale data, heterogeneous illumination conditions, varying viewpoints, etc. Moreover, potential image misalignments due to inaccurate pose estimation during SfM would inevitably deteriorate the results affecting consistency check and fusion. Apart from the scene nature and acquisition conditions, the particular limitations also depend on the adopted reconstruction algorithm and principally on how the correspondence problem is formulated. Some of the most broadly discussed limitations in the literature refer to:

Occlusions and slanted surfaces. The occlusions, as well as the information gaps typically encountered in the stereo cases, in MVS are mostly treated via the visibility models exploiting the multiple observations and achieving more complete scene reconstructions, as discussed in Section 2.5.2. The assumption of constant depth across the support region commonly decreases the efficiency of the algorithm in the presence of slanted surfaces. This limitation can be undertaken with the use of more advanced algorithms also considering the local normal information [Zabulis and Daniilidis, 2004; Bleyer et al., 2010; Gallup et al., 2007].

Challenging surfaces. Nonetheless, the major failure cases are due to the presence of challenging, non-Lambertian surfaces. Similar to stereo matching, photometric consistency measures are the basis for finding correspondences in the vast majority of MVS methods. Photometric consistency is based on the assumption of diffuse or Lambertian reflectance on the surfaces; as a matter of fact, in cases of near-Lambertian surfaces with rich textures, MVS algorithms have achieved impressive 3D reconstruction results [Furukawa et al., 2010; Tola et al., 2012; Fuhrmann et al., 2014; Schönberger et al., 2016]. However, in most real-world scenes, surfaces that variate from this standard are often encountered, such as large textured areas, specular, transparent, or reflective surface materials, and thin structures like wires. For instance, in most indoor scenes, there are several walls of uniform color or reflective metallic structures, thus “weakly-supported” [Jancosek and Pajdla, 2011]. These surfaces are still an unsolved problem in MVS reconstruction as the photometric consistency metric alone typically cannot provide reliable depth estimates due to matching ambiguities. Again, unlike the two-view case, the redundant hypotheses in MVS may be beneficial to partially recover depth estimates in these problematic areas by enforcing multi-view consistency, yet in practice, the presence of such areas remains an open challenge.

Scalability. MVS algorithms are often used to reconstruct large-scale datasets consisting of a vast amount of high-resolution images. This inevitably creates a scalability and runtime performance problem due to the huge cost volumes and the need to compare all possible disparities and the pre-defined search ranges that are often required. Most standard methods become impractical, and several engineering solutions like hierarchical approaches have been proposed in the literature [Rothermel et al., 2012; Wenzel et al., 2013].

Due to the aforementioned limitations, the reformulation of the MVS problem has become implicit; indeed, one of the most efficient state-of-the-art method to tackle this problem is the PatchMatch algorithm [Bleyer et al., 2011; Shen, 2013; Schönberger et al., 2016], discussed in detail in Chapter 3.

2.6 Learning-based methods

Deep neural networks have been recently used in several high-level visual recognition tasks such as image classification [Krizhevsky et al., 2012; He et al., 2016], object detection [Girshick et al., 2014; He et al., 2017], and semantic segmentation [Long et al., 2015; Chen et al., 2017; Badrinarayanan et al., 2017], as well as low-level visual tasks such as optical flow prediction [Dosovitskiy et al., 2015; Sun et al., 2018]. This success has fostered the exploitation of such architectures also in the field of depth estimation and 3D reconstruction. Under this concept, learning algorithms try to infer a depth map from the set of input images. Conventional methods for depth estimation, either binocular or multi-view, highly depend on the handcrafted features used in their cost functions, particularly the photometric consistency. Learning methods try to reformulate the problem while also leveraging semantic cues of the scene, closer to the human vision system. Indeed, they learn more complex feature representations, combining photoconsistency and context, that tend to be robust, as they are commonly able to incorporate global semantic context. Particularly in depth estimation, these cues could potentially be useful in case of sparse image signal, which tends to cause matching ambiguities, i.e., weakly supported textureless surfaces, or even facilitate occlusion reasoning [Dai et al., 2019; Khot et al., 2019]. In fact, correspondence matching and depth estimation may benefit from the global semantic context instead of relying solely on local visual appearance and geometry information.

Generally speaking, learning methods can be supervised, that is, providing ground truth data, and unsupervised, where no ground truth data is available. In supervised methods, the loss function in the training process tries to minimize the discrepancy between the ground truth and the estimated depth along with a regularization smoothness term while taking into consideration the supervision cues, i.e., the ground truth data. Particularly for depth inference problems, ground truth data are commonly captured with depth sensors. Unsupervised or self-supervised methods do not rely on ground truth data and use other cues for training. Additional cues for training, used both in supervised and unsupervised methods, may be smoothness constraints, left-right consistency, maximum depth, or scale-invariant gradient loss. Other works incorporate auxiliary semantic cues such as normal, segmentation labels, or edge maps [Eigen and Fergus, 2015]. Finally, depth map inference is typically achieved by direct regression of the depth values or by treating the problem as an inverse depth classification [Xu and Tao, 2020a; Peng et al., 2022]. Direct depth regression samples uniform depth hypotheses achieving sub-pixel estimation but lacks robustness. On the other hand, inverse depth classification is prone to stair-case noise, so further refinement in a post-processing step is commonly performed. Another commonly adopted categorization scheme, tailored for multi-step problems such as depth estimation, relies on whether the process is formed as an end-to-end or a non-end-to-end pipeline [Zhou et al., 2020].

In this section, the most prominent methods of the recent literature are discussed, either supervised and unsupervised depth estimation for stereo, MVS and monocular depth estimation; a brief overview on depth refinement and completion using neural networks is also given. For an exhaustive review on learning-based depth estimation, the interested reader is referred to the relative review articles [Zhou et al., 2020; Laga et al., 2020]. It is to be noted that, while this section covers the related work in the field, an introduction to the basic concepts and background of neural networks is given in Chapter 4.

2.6.1 Supervised stereo

The first learning approach for stereo treated the problem as a Conditional Random Field (CRF) training task in order to model the relationship between penalty terms and local gradients [Scharstein and Pal, 2007]. In the early applications of deep learning, CNNs were introduced to substitute one or more components of the legacy stereo pipeline in a non-end-to-end manner. The pioneers in the field used robust CNN features to calculate the matching cost between image patches instead of using loose handcrafted photoconsistency metrics, implementing a two-stream Siamese network [Zbontar et al., 2016; Han et al., 2015] or exploring the efficiency of varying architectures [Zagoruyko and Komodakis, 2015]. Unary features are extracted for left and right image patches and then concatenated and passed through fully connection layers to predict matching scores. In a similar way of thought, Chen et al. [2015] introduced multiscaled features for the matching cost, while Luo et al. [2016] used unary features to learn a probability distribution for faster cost computation, but instead of concatenation, they used an inner product layer for direct correlation via photometric similarity, accelerating computational efficiency. Depth inference is treated as multi-label classification. A further regularization step is applied in CNN-based depth estimation as in the conventional counterparts to deliver the final depth maps. MRF-based methods have been used, either handcrafted [Zbontar and LeCun, 2015; Chen et al., 2015; Luo et al., 2016] or learned [Seki and Pollefeys, 2017; Schönberger et al., 2018] to predict the penalties for cost regularization and disparity refinement in an SGM fashion. Knobelreiter et al. [2017] proposed a method to learn smoothness penalties by combining CNNs with CRF optimization in a hybrid representation to avoid post-processing. [Gidaris and Komodakis, 2017] confront stereo matching as a pixel labeling and refinement problem; instead of handcrafted disparity refinement functions, they use a three-stage network that detects, replaces, and refines erroneous label estimations embedded in the same architecture.

The aforementioned methods were robust enough with respect to the baseline conventional ones, yet they were limited by their computational deficiency due to the multiple forward passes for every disparity value. Parts of the pipeline were still hand-engineered functions and limitations similar to the conventional methods, such as the matching ambiguities problem, were, even so, present. For this reason, exploiting the incorporation of contextual information via end-to-end approaches was crucial. The end-to-end methods refer to procedures that seamlessly integrate all steps of the stereo pipeline; they actually divide the pipeline into sub-steps of differential blocks, allowing end-to-end training. A plethora of such algorithms with a more powerful representation ability has been developed and became popular after large benchmark datasets [Mayer et al., 2016]. The early works use the 2D encoder-decoder architecture to directly regress the disparity maps [Mayer et al., 2016; Yang et al., 2018] from cost volumes without requiring an explicit feature matching module. To improve the performance, Pang et al. [2017] suggested a cascade residual learning (CRL) framework with a stacked hourglass aggregation network. Nevertheless, the seminal work of Kendall et al. [2017] was the first to incorporate feature extraction, cost aggregation, and disparity estimation in an end-to-end deep architecture based on a plane sweep volume. Multi-scale 3D convolutions were used to regularize the cost volume, and the best disparity values were directly regressed out of a stereo pair using a soft argmin operation. Chang and Chen [2018] further improved the accuracy using spatial pyramid pooling to construct the cost volume in a multi-scale concept and employ 3D CNNs for regularization. To substitute the expensive 3D convolutions and increase the efficiency, Zhang et al. [2019a] proposed two new layers, a semi-global and a local, for guided cost aggregation. Khamis et al. [2018], aiming at sub-pixel precision, used a coarse-to-fine approach based on Siamese networks and a hierarchical refinement for edge preservation. Guo et al. [2019] construct the cost volume using group-wise correlation to better measure similarities across features while reducing memory consumption by modifying the refinement step.

Learning-based methods based on cost volumes are particularly expensive and, therefore, often applied to downsampled images to compensate for the cost with an extra interpolation module to recover the required resolution. Cascade formulations have been exploited to improve the efficiency [Gu et al., 2020]. However, end-to-end learning-based methods have outperformed the conventional ones in the *KITTY* benchmark [Menze and Geiger, 2015] already since the earlier works [Mayer et al., 2016; Kendall et al., 2017].

2.6.2 Supervised MVS

Learning-based stereo methods cannot be easily generalized to multi-view for basically (1) applying pair-wise rectification and fusion of the 3D result would not fully utilize the redundancy of MVS (2) arbitrary camera geometries add extra complexity for learning [Yao et al., 2018]. The pioneer learning methods in the MVS field, considering the aforementioned limitations were based on volumetric scene representations, i.e., learning voxel occupancy, typically either by surface fusion [Ji et al., 2017] or by fusing the feature grids [Kar et al., 2017]. Such architectures encode camera parameters implicitly by mapping the image appearance on the 3D voxels or unproject image features into 3D grids by perspective geometry. Paschalidou et al. [2018] combined a CNN that learns surface appearance variations with an MRF in order to consider also the physical properties and occlusion models, along with the perspective projection. However, as in their conventional voxel-based counterparts, the learning-based algorithms for volumetric reconstruction with regular grids are computationally expensive, limiting, thus, the applicability of such approaches with high-resolution, real-world images.

To address this issue and enable better scalability, plane-sweep volumes (PSV) were introduced to infer depth maps per view, using planes at different depth values. One of the first approaches, directly learned a multi-patch similarity metric using an average pooling layer in a Siamese network architecture to replace the traditional cost metric and then reconstructed the depth maps by a standard plane-sweep stereo [Hartmann et al., 2017]. Huang et al. [2018] introduced an approach using pre-computed plane-sweep volumes similar to Kendall et al. [2017] but generalized to an arbitrary number of views with an encoder-decoder architecture. The system pre-warps the images in the 3D space and treats the depth reconstruction as a multi-class classification problem using CRFs to refine the raw predictions. In the first end-to-end method for learning depth map inference in MVS, Yao et al. [2018] extracted deep features and built the 3D cost volumes upon the reference camera viewing frustum instead of the regular 3D Euclidean space as in the volumetric methods using the plane-sweep algorithm. The differentiable homography mapping operation, which implicitly encodes the camera geometry, is used to warp features from multiple views; the built cost volumes are further regularized with 3D CNNs to finally either regress the depth or use inverse multi-label classification for depth inference. MVSNet indeed became influential, and most subsequent end-to-end approaches inherited these steps.

To explicitly define the smoothness constraints, [Xue et al., 2019] incorporated CRFs in the pipeline. Im et al. [2019] construct the plane sweep volumes within the network and regress the depth maps in an end-to-end manner. Similar to stereo methods, cost volume regularization is an essential step before depth map inference, particularly useful for noisy data. Cost volumes are computationally costly and therefore often used in a fixed downsampled resolution with a subsequent upsampling or post-refinement module [Yao et al., 2018; Im et al., 2019; Chen et al., 2019; Luo et al., 2019] for the final output.

To overcome the high memory requirements of direct cost volume filtering, which is cubic to the image resolution due to the use of 3D CNNs, recurrent networks with gated recurrent units (GRUs) were adopted for the cost volume regularization [Yao et al., 2019]. In this way, instead of 3D convolutions, 2D cost maps are sequentially regularized along the depth direction, reducing the memory requirements to quadratic. To improve accuracy, a variational refinement is performed in a post-processing step. However, this method cannot efficiently leverage global contextual information and consequently has inferior performance than the standard MVSNet [Yao et al., 2018]. To efficiently capture long-range dependencies apart from the adjacent ones considered in GRU and LSTM methods, and thus global context information, Xu et al. [2021b] employs a non-local RNN for cost regularization. Yet, RNN approaches generally suffer from extensive computational times in exchange for memory performance.

Recent works proposed strategies to reduce the memory requirements and computational times, exploiting feature pyramids to extract multi-scale features and allow the applicability of learning methods in higher resolution datasets. Such coarse-to-fine approaches typically build the cost volume in the early layers of the networks using sparse sampling and combine them adaptively with the denser cost volumes of the finer resolution layers. Yang et al. [2020], introduced a cost volume pyramid to implement a coarse-to-fine approach for depth inference via thin cost volumes. Chen et al. [2019] suggested a coarse-to-fine method that starts from a rough depth map and iteratively corrects the 3D point cloud predicting the residual along visual rays using edge convolutions. In the same line of thought, Gu et al. [2020] and Cheng et al. [2020] proposed cascade cost volume representations for coarse-to-fine approaches. Uniform sampling of fronto-parallel planes is implemented to build the cost volume, followed by iterative depth map refinement, sharing insight with Chen et al. [2019]. Similarly, Yi et al. [2020], proposed a multi-scale method along with self-adaptive view aggregation to guide depth estimation. However, coarse-to-fine methods rely mainly on coarse-scale estimation, leading to detail information loss. Luo et al. [2019], based on MVSNet, proposed an end-to-end network for learning patch-wise matching confidence aggregation for MVS, using isotropic and anisotropic plane-sweep volumes in a hybrid 3D U-Net. In order to propose a more lightweight cost volume representation, Xu and Tao [2020a] used a group-wise correlation similarity measure in a similar fashion as in the stereo methods [Guo et al., 2019]. A cascade 3D U-Net was used for regularization achieving better scalability and depth inference is treated as an inverse depth regression task for robustness. After getting unary features for left and right image patches, these features are concatenated and passed through fully connected layers to predict matching scores. To adaptively consider pixelwise visibility and be more robust in strong viewpoint variations, Xu and Tao [2020c] proposed a network to learn the neighbors for each source image. Cost aggregation is performed using this information via 2D visibility maps and depth is inferred with 3D CNNs. Focusing on time efficiency, Yu and Gao [2020], in a lightweight architecture, learn sparse depth estimates using a sparse cost volume, which is further densified using joint bilateral upsampling propagation.

Transformer architectures have also been recently proposed for more efficient incorporation of the global context [Zhu et al., 2021; Ding et al., 2021]; similarly, attention-based mechanisms have been adopted to capture long-range dependencies [Zhang et al., 2021].

Nevertheless, learning methods for the multi-view scenario remain an open challenge in the cases of large-scale scenes with high-resolution images, as they have limited scalability due to the high computational costs. As a matter of fact, they have high memory requirements due to the use of 3D CNNs for cost volume representation and typically need further refinement and postprocessing. Therefore, such methods typically are evaluated over low or medium resolution datasets [Yang et al., 2021b; Yu et al., 2021; Sormann et al., 2020] and it was not till recently that they were tested in higher resolution [Xu and Tao, 2020c; Ma et al., 2021]. Moreover, learning for depth reconstruction has a large number of parameters and in order to generalize appropriately needs a lot of GT depth maps for training, which are tedious to obtain, therefore commonly provided by synthetic datasets, e.g., Yao et al. [2020]. However, methods trained only on synthetic data inevitably suffer from domain differences with real-world scenarios. For this reason, the combination of large synthetic and small real-world scenarios of the target domain for fine-tuning has also been investigated. These methods are trained on publicly available data, commonly synthetic or real-world scenarios with calibrated cameras, and have, thus, a poor generalization ability in complex, real-world scenes.

2.6.3 Unsupervised stereo and MVS

To surpass the requirement of GT depth maps for training, research has begun to exploit unsupervised and self-supervised methods. Originally used for depth estimation in optical flow problems [Jason et al., 2016], some works focus on monocular reconstruction [Garg et al., 2016; Zhou et al., 2017; Luo et al., 2018], however, the scale ambiguity problem introduces errors in the process. As in the supervised methods, introducing constraints such as left-right consistency has been proven to improve the results and be comparable to, or even outperform the supervised methods [Godard et al., 2017]. In binocular and MVS scenarios, the loss typically aims to minimize the photometric consistency error across the views in an unsupervised way while considering occlusions. Zhong et al. [2017] combined two GC-Nets for disparity estimation directly from stereo inputs, learning to minimize the warping error. Related literature proposed novel view synthesis where the input image and the predicted map are used to reconstruct another view [Flynn et al., 2016; Xie et al., 2016; Luo et al., 2018]. In the MVS case, Khot et al. [2019], relying only on the available images, dynamically aggregates informative clues from selected nearby views to train a photoconsistency loss. Dai et al. [2019] predict the depth maps in a symmetric way, enforcing cross-view consistency and filter the occluded regions. Huang et al. [2021] implement a multi-metric loss function also considering object features, i.e., keeping both photometric and geometric consistency for improvement. Since most of these training strategies heavily rely only on photometric consistency cues that commonly cause matching ambiguities, Xu et al. [2021a] used clustered maps to guide the semantic consistency and data augmentation, and Yang et al. [2021a] introduced a self-supervised network to infer depth maps as pseudo labels to overcome the matching ambiguities problem. However, these methods still underperform in the presence of large textureless areas.

Certain works focus rather on weak supervision, i.e., providing few supervision cues. These can be sparsely estimated depth maps, for instance, from conventional stereo matching techniques [Tonioni et al., 2017], sparse ground truth depth maps [Kuznietsov et al., 2017], priorly predicted depth values in an iterative fashion as self-guidance [Zhou et al., 2017], or using GT LiDaR data [Smolyanskiy et al., 2018]. Nevertheless, unsupervised methods, although not requiring ground-truth data for supervision, are generally memory-consuming and thus evaluated over low-resolution datasets.

2.6.4 Learning for monocular depth estimation

Monocular depth estimation methods aim to recover distances between scene objects and camera parameters from a single image; it is, by definition, an ill-posed problem since redundant 3D scenes can be projected to the same 2D image. Indeed, an efficient depth map recovering from a single image would require rich scene prior cues, commonly used in conventional methods. Early methods for monocular depth estimation relied on handcrafted features and used complementary cues in an MRF formulation to recover the depth since limited information about the scene geometry can be directly extracted from a single image [Saxena et al., 2008]. In the deep learning era, monocular depth prediction refers to the single image inference during test time and is typically formulated either as a regression or a classification problem. The seminal work of Eigen et al. [2014] proposed a scale-invariant loss function in a coarse-to-fine context using a VGG network [Simonyan and Zisserman, 2014]; as a follow-up, Eigen and Fergus [2015] also predicted surface normals and semantic maps in a similar framework. Since then, the problem has been studied in the literature as a supervised [Laina et al., 2016; Xu et al., 2018; Fu et al., 2018] or unsupervised problem [Garg et al., 2016; Godard et al., 2017; Tosi et al., 2019]. In supervised methods, GT is often obtained by sparse depth maps generated using LiDaR point clouds, since rich GT depth annotations are costly to obtain for every pixel. For unsupervised methods, on the other hand, binocular cues such as left-right consistency are used to circumvent the need for GT data. Loss functions are formed either based on pixel-wise photometric loss, either L1 or L2, [Garg et al., 2016], or by combining more sophisticated cues such as structural information [Godard et al., 2017; Watson et al., 2019]. Conditional Random Fields (CRFs) have also been used to exploit neighbor relations and include a more global context [Liu et al., 2015a]. Skip connections in a ResNet fashion are used to preserve the fine-grained features of the first layers [Laina et al., 2016]. Depending on the available training data, the scene depth can be estimated as ordinal, i.e., relative [Fu et al., 2018] or Euclidean [Eigen et al., 2014; Yin et al., 2019]. Local planar priors have also been incorporated as guidance [Lee et al., 2019].

Even though achieving excellent results in depth map prediction, the respective reconstructions in the 3D space suffer from significant distortions and the presence of artifacts. Only recently, few works have tried incorporating 3D awareness into the methods. Since most man-made scenes can be decomposed in planar structures, plane detection can be used as a prior for monocular depth estimation [Lee et al., 2019]. However, the 3D structure was not explicitly considered until recently; Yin et al. [2019] formulated a joint loss function using virtual normals to enforce high-order geometric consistency between surface patches in a large range. The work was further extended by considering affine-invariant depth [Yin et al., 2020] and adding an extra training module for scene 3D reconstruction [Yin et al., 2021]. These state-of-the-art methods, although promising, still suffer from generalization limitations in diverse scenarios. In a recent method, the potential of monocular depth estimation for 3D reconstruction showed limited generalization ability [Welponer et al., 2022].

2.6.5 Learning for refinement and completion

Removing unreliable depth estimates during filtering can lead to sparse depth maps and, therefore, incomplete 3D reconstructions. To improve the completeness, regularization approaches are commonly used on the cost volume in conventional [Kolmogorov and Zabih, 2002; Hirschmuller, 2008; Sun et al., 2003] or learning approaches [Huang et al., 2018; Yao et al., 2018; Guo et al., 2019]. However, such regularization techniques are memory-consuming for high-resolution datasets. Depth completion methods aim to improve the completeness of the depth maps while guided by RGB images. Thus, pixels with unknown depth need to be assigned a depth estimate using the neighboring estimates and the original color images. Straightforward interpolation or other handcrafted methods can be trivial in this case, as they may result in undesired artifacts, oversmoothing, and unreliable estimates in depth discontinuities. Image inpainting methods with deep learning have been investigated for color images [Pathak et al., 2016] however, such approaches are not suitable for depth maps as they lack robust features [Zhang and Funkhouser, 2018]. Hence, deep learning architectures specifically designed for this scope have been developed, starting from sparse depth estimations derived by depth sensors [Chen et al., 2018b; Ma and Karaman, 2018] or MVS [Liu et al., 2020a]. A more sophisticated approach to respect depth discontinuities in object boundaries was introduced by Imran et al. [2019] based on the so-called depth coefficients. Following a different approach, [Kuhn et al., 2019] proposed plane primitive fitting followed by filtering in a post-processing step on high-resolution datasets. However, as in all supervised learning tasks, the lack of complete GT depth maps for training such models remains an open challenge.

2.6.6 Confidence map prediction

Confidence or uncertainty estimation of the pixel correspondences is a crucial step of 3D reconstruction. In conventional methods, confidence values can be used to detect unreliable depth estimates and therefore remove potential noise from the depth maps; in fact, erroneous depth estimates are characterized by low confidence values [Hu and Mordohai, 2012]. In local methods, confidence represents the matching cost of each pixel, calculated by the chosen photo-consistency metric such as NCC. Similarly, in global methods, the confidence is derived from the globally optimized cost volume and takes into account global smoothness. [Hirschmuller, 2008] and [Hu and Mordohai, 2012] provide an analysis of the influence of diverse confidence metrics on the final cost.

Recent deep learning methods have incorporated confidence maps in their pipelines [Seki and Pollefeys, 2017; Gidaris and Komodakis, 2017; Jie et al., 2018]. Confidence prediction aims to calculate the probability that a depth estimate lies within a reasonable noise range. Methods calculate the confidence from left-right confidence checks [Seki and Pollefeys, 2017; Jie et al., 2018] or pose a regression problem to estimate confidence from raw disparity maps [Poggi and Mattoccia, 2016]. To consider a larger receptive field, Tosi et al. [2018] implemented both a local and a global network in ConfNet, and Kim et al. [2019] used a combined adaptive network with an attention module to include the information of the entire cost volume. Extending the applicability of confidence prediction to the MVS scenario, Kuhn et al. [2020] introduced DeepC-MVS for depth map filtering and refinement, avoiding global cost volumes.

2.6.7 Open issues in depth inference

Depth inference in monocular, two-view, and multi-view methods is typically solved as a regression or classification problem. Both problem formulations have recently achieved satisfying results, yet they still suffer from certain limitations; regression methods tend to overfit the training data due to the indirect learning cost volume, and classification methods, being discrete, cannot directly infer the exact depth values [Xu and Tao, 2020a; Peng et al., 2022].

Recent research in depth estimation has moved toward learning-based methods mainly due to their capability to consider global semantic context; trainable photometric costs and cost volume regularization is employed, typically outperforming the conventional methods in dense camera networks with small depth ranges [Lee et al., 2021], especially in confronting the matching ambiguities problem in challenging areas. However, deep learning methods generally have high memory requirements due to their high dimensional cost volumes. Consequently, they cannot directly handle high-resolution images and they are usually either evaluated on lower resolution scenarios like *DTU* [Aanæs et al., 2016] or *Tanks and Temples* datasets [Knapitsch et al., 2017] for instance [Yang et al., 2021b; Yu et al., 2021; Sormann et al., 2020] or use strongly downsampled versions of high-resolution datasets like *ETH3D* [Schöps et al., 2017] [Xu and Tao, 2020c; Wang et al., 2021], often accompanied with an additional spatial upsampling module [Wang et al., 2022] compromising detail recovery. In fact, the inefficient reconstruction of details and thin structures may limit the applicability of learning methods in real-world applications of high accuracy requirements. Similarly, scenes with large depth ranges are often prohibiting due to memory limitations. Moreover, learning methods are data-driven, i.e., they heavily depend on training data for supervision, thus limiting their applicability and generalization in real-world high-resolution scenarios. These ground truth data are commonly acquired using depth sensors, a fact that restricts the applicability of the methods to scenarios where depth sensors can easily collect reliable data, such as close-range indoor scenarios. Hence, domain adaptation is a non-trivial challenge. Ground truth depth maps can also be generated with standard photogrammetric workflows, inheriting, however, the limitations of the method while requiring considerable computational time for data preparation. To relax the requirement of ground truth (GT) training data, unsupervised and self-supervised methods have also been developed, employing self-supervision with left-right consistency [Godard et al., 2017] or novel view synthesis [Dai et al., 2019; Huang et al., 2021]. Still, memory consumption and high-resolution images are a real challenge for these methods. On the contrary, the second approach proposed in this dissertation relies on local textureness cues, and therefore it can easily generalize to indoor and outdoor scenarios and does not require training data.

2.7 Benchmarking in MVS

Toward the evaluation of the image-based 3D reconstruction algorithms on a common framework, several benchmark datasets have been released to the public in the last decades, gradually inspiring the research community to achieve efficient depth estimation and 3D reconstruction. Benchmarks may vary based on the purpose, the nature of input data, the available ground truth (GT) as well as the evaluation metrics used.

Middlebury sequences was a pioneer benchmark, released, and served to evaluate two-view stereo [Scharstein and Szeliski, 2002; Scharstein et al., 2014] and multi-view stereo algorithms [Seitz et al., 2006]. In a dedicated platform, they invited submissions of results from reconstruction algorithms which were publicly ranked against each other. Their early releases contain low resolution for today's standards scenes under controlled laboratory conditions along with GT depth maps. While the stereo dataset consists of real-world scenes, the optical flow dataset is a mixture of real-world scenes and rendered scenes. Later, a new, higher resolution version from Scharstein et al. [2014] was made available (6 MPixels).

EPFL dataset release followed [Strecha et al., 2008], containing few real-world outdoor scenes for MVS purposes. GT mesh models are available, deriving from laser scans. The sequences are characterized by small camera networks and simple camera configurations of mostly well-textured surfaces in medium resolution, hence not particularly challenging for today's algorithms. *EPFL* datasets are still widely used, however, the support is deprecated.

The *KITTI* dataset [Geiger et al., 2012; Menze and Geiger, 2015] is still a widely used multi-purpose benchmark for binocular stereo, optical flow, visual odometry, tracking and semantics. It contains stereo videos of road scenes from a mobile platform, i.e., a calibrated pair of cameras and a laser scanner mounted on a car. While the dataset contains real data, the application scenarios are limited to read-like scenes, and the acquisition method restricts the ground truth only to static parts. The ground truth data is sparse and up to a certain distance and height. In the most recent version, 3D models of cars were fitted to the point clouds to obtain denser ground truth.

DTU robotics dataset [Aanæs et al., 2016] is a laboratory-made MVS evaluation dataset of relatively low resolution images. Various image sequences are acquired from the same poses under different illumination conditions. GT data from a structured light scanner are available.

Tanks and Temples is a modern 3D reconstruction dataset providing a variety of training and testing sequences of indoor and outdoor scenes [Knapitsch et al., 2017]. The dataset aims to provide ground for evaluating both SfM and MVS algorithms. The acquired data are video sequences, but the extracted frames are also provided. Thus, viewpoint chances are small, demonstrating great overlap.

Table 2.1: Widely-used benchmarks for depth estimation and 3D reconstruction.

Dataset	Year	Purpose	Resolution	Scene Type	GT data
Middlebury	2001, 2003, 2005, 2016, 2014, 2021	stereo, optical flow	varying	laboratory	depth maps
Middlebury	2006	MVS	640 × 480	laboratory	3D mesh
EPFL	2008	MVS	3072 × 2048	outdoor	3D mesh
KITTY	2012, 2015	stereo, optical flow visual odometry	1240 × 376, 1242 × 375	outdoor	depth maps
KITTY	2015	MVS	1242 × 375	outdoor	depth maps
DTU	2014	MVS	1600 × 1200	laboratory	3D point cloud
Tanks and Temples	2017	MVS	1920 × 1080	outdoor, indoor	3D point cloud
ETH3D	2017	MVS	6048 × 4032	outdoor, indoor	3D point cloud

ETH3D [Schöps et al., 2017] is a widely used MVS reconstruction benchmark with high resolution scenes of real-world scenarios, indoor and outdoor. It is characterized by strong viewpoint variations and the presence of many challenging surfaces (non-Lambertian) and thin objects. A summary of the properties of the aforementioned benchmarks can be found in Table 2.1.

In recent years, a growing body of benchmarks providing data for deep learning purposes is also being released. For such purposes, also RGBD benchmarks [Silberman et al., 2012; Sturm et al., 2012] acquired by commodity sensors like Kinect have also been widely used, however, they commonly do not scale well on high-resolution real-world scenarios. However, in deep learning, using synthetic data for training is a common strategy. Rendered views of urban scenes from 3D CAD models or from video games [Huang et al., 2018] have been introduced, typically of low or medium resolution. A more recent large-scale synthetic dataset is *BlendedMVS* [Yao et al., 2020]. Considering real-world scenes and originally designed for single view estimation, yet also applied in the MVS case, *MegaDepth* [Li and Snavely, 2018] is a generic dataset containing scenes from highly varying scenarios.

Chapter 3

The PatchMatch algorithm

MVS algorithms aim to establish valid pixel correspondences across multiple, overlapping views and estimate thus the scene depth based on local or global assumptions. The underlying principle of depth estimation is the local smoothness, expressed either using a support window of constant disparity or based on pairwise smoothness terms while penalizing depth discrepancies. These smoothness formulations assume fronto-parallel surfaces and, hence, fail to robustly reconstruct the depth in areas where slanted surfaces are present (Chapter 2). Besides, these algorithms are typically computationally costly due to the resulting enormous 3D cost volumes, making them impractical in large-scale and high-resolution scenarios. At the same time, they mainly depend on depth range priors as they generate candidates with equal intervals in a specific, predefined depth range, impairing their scalability.

Notable alternatives to these approaches have been proposed in the literature, considering surface normals [Furukawa et al., 2010] or second-order smoothness terms [Woodford et al., 2009] to avoid the fronto-parallel bias. To reduce the search space and improve efficiency, coarse-to-fine schemes have also been applied [Rothermel et al., 2012; Wenzel et al., 2013]. The plane-sweep method [Gallup et al., 2007] was certainly a breakthrough; nevertheless, it was not until the proposal of the PatchMatch stereo algorithm [Bleyer et al., 2011] that these limitations were simultaneously and robustly circumvented. Indeed, PatchMatch tackles the problem of matching ambiguities in the presence of slanted surfaces by fitting a local support patch to every pixel. A patch π is essentially a tangent 3D plane, locally approximating the surface. Both pixel depth d and normal $\mathbf{n}$ of this plane patch are taken into consideration for the algorithm convergence. Given such a function definition, reconstructing a patch is simply achieved by maximizing the photo-consistency function with respect to those parameters. Global cost volumes, either classic, i.e., for each patch along epipolar lines, or plane-sweep-based (PSV), i.e., for each set of plane hypotheses, are computationally expensive. PatchMatch discards the idea of global cost volumes and reduces the computational and memory cost by propagating the depth hypotheses across the image; it provides an alternative to examining all possible disparities, pruning out the search space by exploiting the natural spatial coherence of the images. Pre-defined disparity ranges are also avoided since a stochastic search over the continuous depth space is adopted. Considering these advantages in robustness and performance, PatchMatch-based algorithms gradually have replaced the standard ones in the latest years in state-of-the-art implementations [Galliani et al., 2016; Schönberger et al., 2016; Cernea, 2020]. The proposed method in this dissertation also relies on PatchMatch and builds upon its functionalities; accordingly, in this chapter, the details of the original algorithm and its variants will be discussed.

3.1 The PatchMatch algorithm

The original PatchMatch [Barnes et al., 2009] algorithm was introduced to establish valid matches between patches as a randomized and iterative method. The term “matches” here is more general and not constrained to the correspondence search; originally, it referred to image editing purposes, including inpainting, image denoising, and object detection.

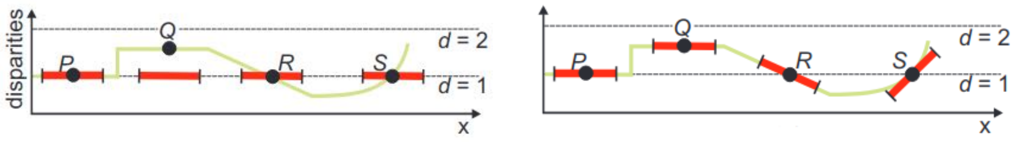
The valid matches were calculated by performing an efficient nearest neighbor (NN) search. Actually, the core idea was based on random initialization, performed once, and spatial propagation followed by random refinement, running iteratively until convergence. Although a random choice would not probably be a good guess, the intuition is that a large number of random initial assignments is likely to converge to at least one good match; this is particularly true in the case of high-resolution images, as there is a higher possibility that one reliable guess is made. That being said, if prior knowledge is available about the NN, it can be used to guide the initialization. Regarding spatial propagation, the underlying assumption is that due to the natural local consistency of the images, good matches can be propagated to the neighboring pixels, spreading best estimates across the whole image. In fact, the algorithm starts with an initial guess and propagates the best-scored values to neighboring pixels. Finally, in the random refinement step, randomly sampled values, both far different and closely similar to the current match, are tested against the current match to define an optimal solution and escape from local minima. Spatial propagation and random refinement are performed iteratively for a certain number of iterations or until a predefined criterion is met, e.g., the total error over the image. PatchMatch, although simple, has been proven to perform surprisingly well [Barnes et al., 2009]. Generally, with such a sequence of steps, PatchMatch is a high-performance algorithm, and, especially in GPU implementations, the performance is almost real-time.

More particularly, for patch correspondences, Barnes et al. [2009] define a nearest-neighbor (NN) field as such: let a function $f : I \mapsto \mathbb{R}^2$ of random "offsets" over the range of image I, for some distance function of two patches. For a patch center coordinate a in reference image I and its potential corresponding NN patch center b in source image I', $f(a) = b - a$; all possible values of f are the correspondence vectors, or "offsets", defining the NN field. The Generalized PatchMatch [Barnes et al., 2010] extended the approach across k nearest neighbors, instead of finding only the nearest one, using a heap data structure[1]. Varying scales and rotations are also considered and calculated on the fly, while arbitrary similarity measures can also be used.

3.1.1 Overview of PatchMatch for depth estimation

Tailoring the idea of PatchMatch in the context of stereo matching, the seminal work of Bleyer et al. [2011] used photometric consistency measures and slanted support windows, i.e., 3D oriented planes, instead of single disparity values assigned to every pixel p. In other words, surfaces are modeled with local planes, and the 1D search is replaced by a more complete geometric model; the nearest neighbor on the epipolar line according to a plane is calculated, avoiding global cost volumes. Thus, it allows for a quick solution without browsing through all possible solutions. As a matter of fact, using PatchMatch stereo has been proven more efficient than local methods or semi-global matching approaches since both require evaluating the full disparity space image (DSI), i.e., computing matching costs at each pixel for all disparities under consideration. PatchMatch stereo, on the other hand, avoids exploring the full disparity space by propagating good disparities from an initial set of guesses to neighbors, resulting in low-memory requirements. Consequently, scalability and runtime performance are both improved, especially with large sets of high-resolution images. Besides, sub-pixel depth accuracy is de facto obtained since the operations are made in the continuous space and do not use discrete values like in more traditional global and local methods (see also Chapter 2). The initial plane assignments, although random, quickly lead to convergence since, for high-resolution images, there are good chances that at least one random hypothesis will be close to the correct one. Even a single acceptably good guess is enough to spread the correct estimates to all pixels that belong to the same plane through propagation. In [Bleyer et al., 2011], propagation is performed iteratively, starting from the top-left to the bottom-right pixel for odd iterations and vice versa for even ones.

[1] A heap data structure in computer science is a complete binary tree structure. Each node must satisfy a heap property, either max-heap or min-heap, i.e., that the max (or min) always is always in the root node.



(a) fronto-parallel windows

(b) slanted support windows

Figure 3.1: **Fronto parallel and slanted windows.** (a) standard local methods assume fronto-parallel support windows at integer disparities (in red) while (b) the PatchMatch-based approach of Bleyer et al. [2011] introduced slanted 3D support windows in the continuous space. Source: [Bleyer et al., 2011].

One of the most substantial properties of the algorithm is its robustness in the presence of slanted surfaces, a feature that established PatchMatch as a state-of-the-art practice in recent years. As discussed in Chapter 2, local and global methods typically suffer from the fronto-parallel bias as they imply constant disparity in the pixel neighborhood; yet, this assumption does not hold in real-world scenarios due to the presence of discontinuities or inclined surfaces. In PatchMatch, this limitation was undertaken using oriented support windows; the object's surface is approximated locally with oriented patches, allowing for efficient depth estimation even in the presence of slanted surfaces (Figure 3.1). Plane refinement is performed by assigning random values iteratively and checking if these values are better estimates than the current one. Occlusion handling is performed via left-right consistency checks at a post-processing step, which is a common practice in local stereo methods. Although efficient against fronto-parallel bias, PatchMatch is essentially a local method, inheriting the limitations of such algorithms. In fact, energy minimization consists only of a unary term measuring the photometric consistency; hence, smoothness is not modeled explicitly. Indeed, in Bleyer et al. [2011], refinement is performed only for the pixels that did not pass the left/right consistency check in a post-processing step by plane extrapolation to consider also slanted surfaces, followed by filtering to reduce artifacts.

Due to its effectiveness in treating slanted surfaces and reducing the search space, PatchMatch has drawn the attention of the research community. Several improvements followed the original greedy[2] PatchMatch algorithm introducing regularization. To achieve sub-pixel accuracy for stereo depth estimation and optical flow, while explicitly introducing smoothness constraints, Besse et al. [2014] used belief propagation (BP) [Pearl, 1988; Yedidia et al., 2005] and PatchMatch in a joint fashion. BP is a message-passing algorithm used for graph optimization and, therefore, energy minimization; in the approach of Besse et al. [2014], the pairwise terms, in the form of penalties, encourage smoothness in the field of the 3D planes that the PatchMatch optimizes considering both the normal and the disparity value. As a deduction, PatchMatch is extended to a continuous

[2] A greedy algorithm is a simple, heuristic algorithm that estimates the best choice at each step, disregarding global information.

MRF inference formulation; in the absence of the pairwise term, the algorithm reduces to standard PatchMatch. Li et al. [2015] also modeled a continuous MRF, approximated with BP, and added a cost aggregation module. They proposed an accelerated scheme based on superpixel-level graphs to handle critical computational bottlenecks. Rather than using BP, Heise et al. [2015] integrated PatchMatch in an explicit variational smoothing formulation; a data term and a regularization term are combined to alleviate the problems occurring from the implicit smoothing model of the standard algorithm.

Eventually, PatchMatch quickly became standard practice for stereo matching and was consequently also adopted in the multi-view scenario. The pioneering works introduced such an extension by selecting the best neighboring views for every reference image based on geometric criteria [Bailer et al., 2012; Shen, 2013]. These best views are considered for depth map computation and filtering for outlier removal. Wei et al. [2014] extended the work of Bailer et al. [2012], employing cross-view filtering based on depth variance to enforce consistency across different views for outlier removal. Later works rather focused on more efficient view selection; Zheng et al. [2014] proposed an Expectation-Maximization (EM) probabilistic graphical model to solve the joint pixel-level view selection problem and simultaneously perform depth estimation, yet it suffered from fronto-parallel bias. In the seminal work of Schönberger et al. [2016], this approach was extended to exploit normal guidance in the photometric cost and work efficiently also in slanted surfaces. Indeed, additional geometric consistency constraints were imposed, also considering the normals for the matching cost guidance, resulting in high accuracy dense clouds in rich texture regions. Galliani et al. [2015], sharing insight with Shen [2013], formulated the PatchMatch in the scene space and adjusted the cost aggregation for the multi-view scenario. The propagation scheme was modified following a red-black checkerboard pattern to achieve computational efficiency. After these groundbreaking works, PatchMatch became the state-of-the-art method for MVS implementations. Toward efficient view selection, adaptive checkerboard sampling propagation and multi-hypothesis have also been investigated, along with multi-scale feature guidance to solve matching ambiguities [Xu and Tao, 2019]. Similarly, Wei et al. [2014] employed cross-view consistency in a hierarchical coarse-to-fine scheme, while the multi-scale approach of Xu et al. [2020] estimated the optimal scale for every pixel guided by the epipolar constraint. Multi-scale approaches can better alleviate the ambiguities since, in coarser scales, texture information is more discriminative, and reliable depth values are propagated in the finer scales. However, these approaches are limited to pre-defined scales and may lead to information loss, especially the fine details.

Romanoni and Matteucci [2019] assumed piecewise planarity on image superpixels [Van den Bergh et al., 2015] while Xu and Tao [2020b] added direct planar priors to assist PatchMatch with planar compatibility constraints for the matching cost. In a similar fashion, Kuhn et al. [2019] used superpixels to perform depth completion where textureless areas are treated with multi-scale geometric consistency guidance, yet as a post-processing step. Later, a trainable post-processing module for regularization based on confidence prediction was proposed by Kuhn et al. [2020]. Recently, plane hypothesis inference using MRF was also proposed as a post-processing step after initial depth estimation and filtering [Sun et al., 2021]. Except for standard stereo matching and MVS, PatchMatch was also efficiently exploited in optical flow tasks [Besse et al., 2014; Bao et al., 2014; Li et al., 2015; Hu et al., 2016].

Support plane parametrization. Generally, 3D support patches π are described by two components; the point $\mathbf{X}(X, Y, d)$ with d being the disparity value and the corresponding normal vector $\mathbf{n}$. Bleyer et al. [2011] define the support plane in image coordinates. The first MVS frameworks proposed strategies that use one depth and two spherical coordinates, i.e., angles, to model the 3D support plane in the object space [Shen, 2013; Bailer et al., 2012]. Later, a strategy proposed by Galliani et al. [2015] and also adopted by Schönberger et al. [2016] used a local 3D plane with normal $\mathbf{n}$ and depth d in the Euclidean space as the support domain for the corresponding pixels. They define a tangent plane for every scene point; such an explicit formulation of 3D planes represents plane-induced homographies [Hartley and Zisserman, 2003]. Accordingly, the epipolar rectification is skipped, as well as the tracing of epipolar lines during correspondence search, allowing for efficient cost aggregation in the MVS scenario. Zheng et al. [2014] used single-oriented planes, i.e., planes oriented in one direction instead of multiple oriented ones [Bailer et al., 2012] to reduce the search space; hence, they applied fronto-parallel homographies to map patches across images, a fact that inevitably leads to artifacts in slanted surfaces. Zhu et al. [2015] adopted a support plane described by one depth and two depth offsets.

Random initialization. Original PatchMatch stereo is based on random initialization, i.e., a random 3D plane normal $\mathbf{n}$ and a random depth value d are assigned to each pixel [Bleyer et al., 2011]. Still, these values should be selected carefully, especially while working in the object space [Galliani et al., 2015]. If sparse depth estimates and point normals are available, e.g., from the SfM sparse cloud, they are used for faster convergence.

3.2 PatchMatch MVS

The multi-view depth estimation requires robust handling of significant viewpoint variations concerning angles and baselines; thus, spatial propagation must be carefully designed. Generally, multi-view reconstruction with PatchMatch relies on efficient view selection and propagation scheme, depth computation and refinement, and fusion.

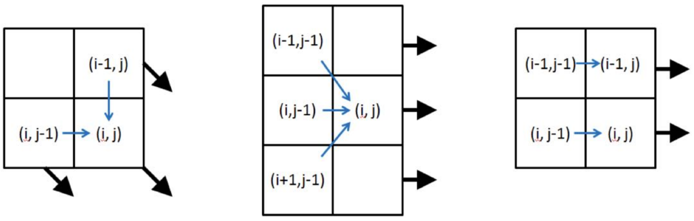


Figure 3.2: **Sequential propagation schemes in PatchMatch.** Sampling (blue arrows) and propagation direction (black arrows). Bleyer et al. [2011] (left), Bailer et al. [2012] (middle), Zheng et al. [2014] (right). Source: [Zheng et al., 2014].

3.2.1 Propagation schemes

The propagation scheme and the view selection are crucial for an accurate, complete, and time-efficient 3D reconstruction. Two main approaches exist for propagation schemes in PatchMatch MVS scenarios, sequential and diffuse-like.

Sequential propagation. In the standard implementation [Bleyer et al., 2011], depth propagation is performed sequentially, propagating information diagonally across the image (Figure 3.2, left). In particular, starting from the top left corner, odd-numbered iterations propagate diagonally good support plane estimates to the lower and right neighbors of the current pixel if the lower cost criterion is fulfilled. Once all pixels of the images are processed, even-numbered iterations do the same process with reverse iteration direction, i.e., starting from the bottom-right pixel of the image and working upwards. In such propagation schemes, every pixel depends on the previous one(s), and good estimates can propagate arbitrarily far in only one pass. Apart from the original approach, several works have adopted this practice [Heise et al., 2013; Shen, 2013; Wei et al., 2014]. Zheng et al. [2014] considered only one previous neighbor and applied upward/downward and leftward/rightward propagation (Figure 3.2, right). Typically, two or three iterations are enough for the algorithm to converge. Such schemes are more sensitive to textureless regions, as only pixels in the close neighborhood of the current one are considered for depth propagation. Parallelization attempts have also been performed, yet a sequential scheme cannot take full advantage of the GPU architecture, as it can be parallelized only at the row or column level. For instance, Bailer et al. [2012] implemented a sequential scheme that used only horizontal and vertical scan lines to partially exploit GPU usage; three previous neighbors are considered in downward/upward and leftward/rightward directions for odd and even iterations, respectively (Figure 3.2, middle).

Diffusion-like propagation. Galliani et al. [2015] proposed a new scheme to enable the full use of GPU computation. It simultaneously updates half of the pixels of the image with a checkerboard, red-black pattern. In more detail, image

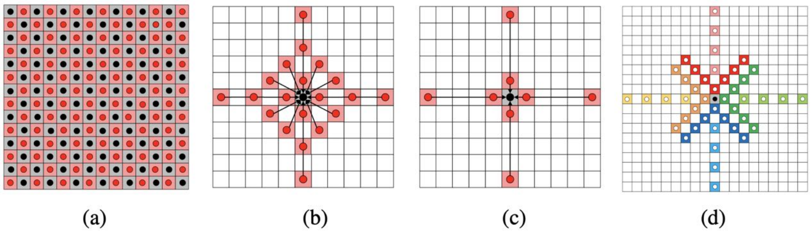


Figure 3.3: **Diffusion-like propagation schemes in PatchMatch.** (a) red-black pattern (b), (c) standard and fast checkerboard propagation scheme with the current pixel in black and the considered neighbors in red Galliani et al. [2015], asymmetric checkerboard propagation with the V-shaped neighbors in different colors [Xu et al., 2017]. Source: [Xu et al., 2017].

pixels are divided into a red and black grid (Figure 3.3a); in every iteration, the red or the black pixels are updated simultaneously based on the hypothesis of their black and red neighbors, respectively, achieving better efficiency in parallelization. In the original implementation, a total of 20-pixel neighbors are considered as candidates to update the current pixel, while a fast version using eight neighbors is also proposed (Figure 3.3b). This scheme is inspired by a common practice in message-passing algorithms [Felzenszwalb and Huttenlocher, 2006]. The used local neighborhood is broad enough, and consequently, information is diffused relatively far, achieving convergence within a few iterations; however, given that the propagation is regular and symmetric, reliable depth estimates will expand to a certain degree. As reported in Xu and Tao [2019], checkerboard propagation may lead to inferior results in challenging areas due to ineffective view selection, i.e., they do not consider that the good hypotheses should have priority in propagation and employ a simple heuristic scheme instead. This fact inspired further improvements by using asymmetric and adaptive checkerboard propagation to spread, in the continuous regions, the good hypothesis even beyond [Xu and Tao, 2019] based on the message-passing scheme of Sun et al. [2003]. In this adaptive scheme, neighboring pixels are grouped into close-region (V-shaped) pixels, representing the possibility for depth continuity, and distant-region pixels (linear), representing the possibility for depth variation (Figure 3.3d). Adopting such a propagation scheme has proven to be more efficient for large homogeneous regions and has inspired further developments; e.g., Zhou et al. [2021] substituted the V-shaped areas with long strips to better recover thin foreground objects.

3.2.2 View selection

In MVS, robust depth estimation for a given pixel in a reference image I_{ref} demands the selection of a subset of neighboring views $\mathcal{I}_{neigh} \subset \mathcal{I}$ that will contribute to the most effective cost calculation. A carefully designed view selection scheme is needed, especially for unordered images, as it is a way to enforce photometric consistency between neighboring views and handle occlusions. It is commonly treated with simple rules on visibility information, such as keeping the 50% best views [Kang et al., 2001] or excluding the most improbable neighboring views based on global criteria and photometric cost [Goesele et al., 2007; Furukawa et al., 2010]. If such selection generates huge image subsets, commonly, only the best views will be considered for a more robust solution. In the Patchmatch MVS scenario, views were originally pre-selected, relying on simple heuristics based on global viewing angles and baselines. For instance, Bailer et al. [2012] extended the heuristic view selection approach of Goesele et al. [2007], while Shen [2013] selected a fixed number of best neighboring views for each reference image by setting thresholds for the viewing angles and baselines; Galliani et al. [2015] followed a similar idea. More sophisticated approaches jointly implement pixel-wise view selection and depth estimation based on robust probabilistic graphical models; in other words, a subset of good neighboring images is selected for each pixel individually and not globally for every view [Zheng et al., 2014; Schönberger et al., 2016]. Good neighboring views are selected based on visual similarity metrics, i.e., the per-pixel photometric cost [Zheng et al., 2014]. However, considering only this metric inevitably favors short baselines and small viewing angle variations, which may be less informative. Thus, geometric priors and temporal smoothness have also been introduced to sample from diverse viewpoints and increase robustness [Schönberger et al., 2016].

It is worth noting that the propagation scheme, either sequential or diffuse-like, can be potentially combined with both pixel-wise probabilistic or simple global heuristic models for view selection. Some methods use straightforward sequential propagation schemes combined with a sophisticated probabilistic model for view selection [Zheng et al., 2014; Schönberger et al., 2016]. Others use simple heuristic schemes to pre-select global aggregation view subsets with minimal matching costs combined with parallelized diffuse-like propagation [Galliani et al., 2015]. An advanced recent method combined adaptive checkerboard patterns with a multi-hypothesis strategy for pixel-wise view selection to improve the results [Xu and Tao, 2019]. In this way, they avoid the bias due to different aggregation view subsets for different hypotheses; the best views are selected for each pixel via a voting decision scheme based on matching cost and confidence considering the pixel's neighborhood. The best estimate is the hypothesis with the minimum multi-view aggregated cost. Although this scheme has been proven robust enough, view selection may not be reliable for datasets with particularly strong viewpoint variations.

3.2.3 Cost computation

In most local methods, the cost for every pixel is typically calculated by a similarity measure based on photometric consistency accumulated across the support window. Generally, the squared differences or absolute differences of color values often were used in some real-time multi-view stereo [Hosni et al., 2011], while the normalized cross-correlation and census transform were adopted to consider the bias and gain changes across multiple images on a window basis [Shen, 2013; Zheng et al., 2014; Li et al., 2015; Zhu et al., 2015]. To achieve a reliable correspondence, aggregation of the image similarities was often necessary, i.e., applying a smooth filter over the image similarity space [Hosni et al., 2013]. The adaptive support-weight approaches improved the robustness of the similarity metric and achieved structure-preserving property [Yoon and Kweon, 2006; Hosni et al., 2011, 2013].

In the context of PatchMatch stereo, the original work of Bleyer et al. [2011] accumulates the cost across an adaptive weight window around each pixel [Yoon and Kweon, 2006]. The weights control the influence of the window pixels according to their proximity to the central one and overcome the edge-fattening problem. They represent the likelihood of the two pixels to lie on the same plane based on color similarity. In the cost function, absolute color differences and differences in magnitude are combined. Galliani et al. [2015] substituted the color values with intensity differences and used a sparse census transform [Zinner et al., 2008], evaluating every other row and column in the window for speed. However, the normalized cross-correlation (NCC) and its variants have been proven competent enough in PatchMatch scenarios for high-resolution images regarding performance and computational efficiency. The standard NCC has been successfully used in the MVS scenarios [Bailer et al., 2012; Zheng et al., 2014], while Shen [2013] employed the zero-mean version of it (ZNCC). Schönberger et al. [2016] and Xu and Tao [2019] applied a bilaterally weighted NCC to treat depth discontinuities efficiently.

Shen [2013] formulated the PatchMatch in the scene space in contrast to previous works that run in the disparity space, enabling efficient cost aggregation across different views without rectified images; the approach was further adopted also by Galliani et al. [2015], Schönberger et al. [2016] and subsequent works.

3.2.4 Refinement

Random refinement is necessary to help the result avoid local minima and, consequently, filter noisy depth and normal estimates. For each pair of hypotheses, three possibilities exist: both normal and depth are correct, one of them is correct, or neither of the two is correct [Schönberger et al., 2016; Xu and Tao, 2019]. An extra set of hypotheses is generated to be compared with the current estimate; thus, for every PatchMatch iteration, the latter is not only compared with the estimates of the pixel neighbors but with the extra hypotheses as well, completely random or perturbed by the current one. Finally, the hypothesis with the minimum cost across all comparisons is selected.

3.2.5 Depth fusion

Since a depth map is calculated for every reference view, a subsequent fusion of the overlapping projected depth estimates is required to generate a final, merged 3D point cloud. During depth fusion, outliers are eliminated, and the noise is reduced; points with an insufficient number of supporting views are excluded. Generally, an inlier value should be photometrically stable across multiple views; therefore, consistency checks are performed. Consistency checks often include the reprojection of a view's 3D points to the overlapping views and the subsequent check for constant disparity values; normal consistency is also taken into consideration [Shen, 2013; Galliani et al., 2015; Xu and Tao, 2019]. A recent method additionally performed cross-checking with neighboring pixels, assuming local coherency [Xu et al., 2020]. Finally, average depth values of the remaining inlier points across the consistent views are calculated to avoid artifacts and reduce noise. Typically, depth fusion is, indeed, an engineered part of the pipeline, as many handcrafted heuristics, e.g., thresholds, are applied for outlier removal and noise reduction. That being said, the methods generally aim to generate the best possible depth maps to enable straightforward fusion [Galliani et al., 2015]. Schönberger et al. [2016] formulated depth fusion as a graph-based problem, considering both depth and normal maps recursively; the reprojection error, indicating geometric consistency, as well as the photometric consistency, are considered.

3.3 A PatchMatch MVS algorithm explained

In this section, the PatchMatch MVS method proposed by Shen [2013] is discussed in detail to introduce basic notation and context since the proposed methods in this dissertation build upon these principles. This algorithm is based on the standard PatchMatch approach for stereo, yet it is particularly tailored to the multi-view cost aggregation problem, defining the support planes in the 3D space. Given a set of $i = \{1, 2, \ldots, n\}$ overlapping images $\mathcal{I}$ of known camera poses, the pipeline can be summarized in the following steps:

Propagation and view selection. Depth propagation follows a simple sequential scheme similar to the one of Bleyer et al. [2011]. Candidate neighboring (source) views for every reference image are chosen based on global visibility criteria. For the sake of robustness, a good potential pair should fulfill the dual criterion of similar viewing direction and adequate baseline length. As a matter of fact, views with long baselines typically suffer from insufficient overlap as well as considerable perspective and radiometric differences; on the other hand, views with a short baseline may have a similar appearance yet involve unfavorable intersection angles. The best angles between the principal viewing

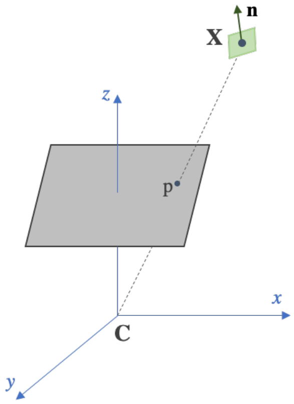


Figure 3.4: **Slanted 3D tangent planes.** To each pixel p a 3D plane is assigned, described by its center coordinates **X** in the camera reference system and its normal vector **n**. Adapted from [Shen, 2013].

directions of reference and neighboring views are selected using the visibility of the already available sparse 3D points, commonly calculated during the SfM step. An acceptable such angle θ is between 5° and 60°. For the images that meet this requirement, the median distance b between neighboring optical centers is computed, and acceptable distances are considered the ones whose $b < 2\bar{b}$ or $b > 0.05\bar{b}$. The final set of pairs is sorted in ascending order, and the best k neighboring images are considered.

Depth map computation. For every image I_i of the input set $\mathcal{I}$ with camera parameters $\mathbf{K}_i, \mathbf{R}_i, \mathbf{C}_i$, a rough depth map is approximated by interpolating the 3D sparse point cloud resulting from SfM. To each pixel p with homogeneous coordinates $\mathrm{p} = [x, y, 1]^T$ slanted support plane π with normal $\mathbf{n}$ and center coordinates $\mathbf{X}$ are randomly assigned, aiming to find the best support plane that corresponds to the minimal aggregated photometric cost (Figure 3.4). The 3D world point $\mathbf{X}$ lies on the viewing ray of p. Given the camera calibration matrix $\mathbf{K}_i$, for any randomly selected depth value d in the range $[d_{min}, d_{max}]$, the 3D coordinates of $\mathbf{X}$ are computed in the camera coordinate system:

$$\mathbf{X} = d\mathbf{K_i^{-1}}\mathrm{p}, \tag{3.1}$$

and a random plane normal $\mathbf{n}$ is assigned to it. According to the basic principle of PatchMatch, this random initialization is likely to have at least one good hypothesis for each depth value. In the case of high-resolution images, this is even more robust since every scene plane contains more pixels and thus more guesses.

Since the homography mapping between the images is already known from the pose estimation, potential pixel correspondences are established for all image pairs.

The aggregated matching cost is calculated using NCC, particularly a weighted zero-mean version of it (ZNCC) as given in Equation 2.18, which integrates the subtraction of the local mean μ to the NCC and tends thus to be more robust to illumination changes and depth discontinuities. This measure is considered reliable enough, especially for high-resolution images, and in this way, more complex aggregation costs are avoided for time efficiency.

Accordingly, every pixel p is associated with a rough 3D plane that is to be further refined during the PatchMatch iterations. Two procedures are performed during each PatchMatch iteration on each image pixel, namely spatial propagation and refinement. Spatial propagation is based on the idea that the neighboring pixels $\mathrm{p_N}$ are likely to belong to the same plane with p and have a similar depth value. Therefore, during the iterations, the assigned planes between neighboring pixels are compared to ensure depth smoothness among them and propagate correct estimates; the $(\hat{d}, \hat{\vec{n}})$ combination with the highest photometric score (meaning the lowest cost in the cost function) is kept and propagated, as it is considered to be a better estimate. To further refine these values, random assignment is performed, i.e., several randomly assigned planes are iteratively compared with the current estimate to potentially reduce the matching cost. In such a way, the search range is progressively reduced, and pixels with high aggregated matching costs are removed.

Depth map filtering. During the filtering step, consistency between neighboring views is enforced for every depth map to refine the depth values and remove potential outliers. To this end, for each pixel p, a point $\mathbf{X}$ is reconstructed in 3D using the assigned depth value $\hat{d}$, the camera intrinsic parameters $\mathbf{K}_i$, the rotation matrix $\mathbf{R}_i$, and the projection center $\mathbf{C}_i$:

$$\mathbf{X} = \hat{d}\mathbf{R}_i^T\mathbf{K}_i^{-1}\mathrm{p} + \mathbf{C}_i. \tag{3.2}$$

Subsequently, $\mathbf{X}$ is back-projected to all neighboring k views and is considered as a valid estimate only if its depth d is consistent across the views, that is, only if the depth difference between d and d_k is small enough. In [Shen, 2013], a minimum number of consistent views is set to $m = 2$ and the threshold for closeness is defined as:

$$\frac{|\hat{d} - d_k|}{d_k} < \tau, \tag{3.3}$$

where τ is a constant. Otherwise, if $\mathbf{X}$ does not fulfill this consistency criterion, it gets discarded. This refinement process handles occlusions and significantly reduces the errors in the final, filtered depth map for each view.

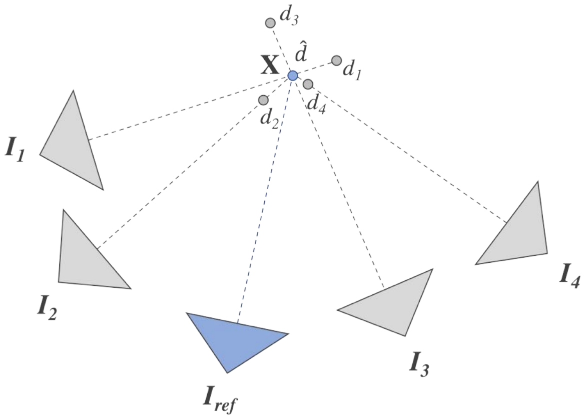


Figure 3.5: **Neighboring depth map test for depth map merging,** as proposed by [Shen, 2013]. Depth redundancy is treated by merging points that are close enough or removing occluded points. Adapted from [Shen, 2013].

Depth map merging. The various depth maps referring to overlapping parts of the scene are fused together to remove redundant depth values for every 3D point $\mathbf{X}$ back-projected using in Equation 3.2. $\mathbf{X}$ is then reprojected to all $\mathcal{I}_{neigh}$ neighboring views. Redundant points are merged together or are eliminated accordingly, based on the neighboring depth map test. Let a reference camera I_{ref} with center $\mathbf{C}_i$ have $m = 4$ neighboring views as shown in Figure 3.5 and $\hat{d}$ be the depth with respect to the reference camera. Values d_m are computed for $\mathbf{X}$ based on the depth maps of the neighboring views I_m. The depth valued d_4 from view I_4 is close enough to $\hat{d}$, so these points are considered identical and merged. On the contrary, if $\hat{d} < d_m$, as for instance for d_3 and d_1, the point $\mathbf{X}$ is considered occluded on the respective neighboring views and gets discarded from their depth maps. Lastly, all depth maps are projected to the 3D space resulting into a single, fused dense cloud.

3.4 Learning-based PatchMatch

Learning-based methods for depth estimation and reconstruction have been gradually introduced in the literature and achieved promising results, often outperforming standard handcrafted approaches by employing trainable photometric costs and cost-volume regularization (Chapter 2). Most methods typically rely on frontal plane sweeps [Huang et al., 2018; Yao et al., 2018; Guo et al., 2019; Luo et al., 2019; Xu and Tao, 2020a]; however, the inevitable cost volume regularization using 3D CNNs adds a computational burden, limiting the applicability of the methods in low-resolution applications and small depth ranges. To overcome this scalability barrier, some of these methods downsample the input and compute the cost volume in low-resolution [Yao et al., 2018; Xu and Tao, 2020a], often using an additional spatial upsampling module later [Wang et al., 2022]. Other works employed recurrent 2D cost map regularization [Yao et al., 2019], attention architectures [Luo et al., 2020] for refinement and regularization or coarse-to-fine schemes for efficiency [Yu and Gao, 2020; Cheng et al., 2020; Gu et al., 2020]. More recent solutions aim to exploit the PatchMatch algorithm, in stereo [Duggal et al., 2019] or multi-view scenarios [Wang et al., 2021; Lee et al., 2021].

Nonetheless, it is to be noted that the iterative and sampling part of PatchMatch is non-differentiable and thus not trivial to be incorporated into an end-to-end pipeline. In the first such approach for stereo matching, Duggal et al. [2019] reduce the search space by using differentiable PatchMatch to obtain a lightweight cost volume that is further refined by a 3D CNN. In multi-view, some PatchMatch-based methods use extra training modules for confidence prediction alone [Kuhn et al., 2020] or combined with mesh guidance [Wang et al., 2020b]. The first end-to-end cascade formulation of PatchMatch in the MVS scenario was PatchMatchNet, minimizing the sum of per-iteration losses [Wang et al., 2021]; no cost volume regularization is applied. However, this approach is still not competitive in high-resolution datasets. Recently, Lee et al. [2021] used a reinforcement learning technique for the PatchMatch algorithm, predicting depths, normals, and visibility information while applying pixel-wise regularization. Yet, the implementation is limited to downsampled images compared to the respective handcrafted methods, demonstrating reduced performance. As a matter of fact, accuracy and fine detail recovery remain a challenge in such methods. Finally, similar to other learning depth estimation approaches, the ones relying on PatchMatch are heavily data-driven and suffer from generalization and domain adaptation, apart from the computational complexity problems.

3.5 Discussion

PatchMatch is a robust algorithm for efficient processing of high-resolution images in large-scale MVS scenarios; indeed, many of the state-of-the-art frameworks rely on PatchMatch principles and provide efficient, open-source solutions in image-based 3D reconstruction. Accordingly, it has gradually replaced other depth estimation methods in research scenarios in the past decade. Despite the great success of PatchMatch-based MVS reconstructions, conventional photogrammetric methods with high precision requirements often prefer to follow semi-global (SGM) approaches [Hirschmuller, 2008]. This choice is debatable, yet, our intuition is that PatchMatch has great potential for efficient depth estimation.

In terms of memory and time efficiency, PatchMatch is beneficial for high-resolution datasets and well-suited for memory-constrained environments, as its runtime complexity increases linearly with the image resolution ($W \times H$) and is independent of the disparity range. On the other hand, semi-global approaches have a polynomial complexity ($\mathcal{O}(W \times H \times D)$) considering the additional disparity dimension D in the cost volume calculation. That being said, SGM does not rely on pixel-wise propagation in an iterative fashion, a fact that compensates for the computational cost up to some extent. Regarding the reconstruction quality, the unbiased oriented tangent planes in PatchMatch typically result in higher quality and more complete reconstructions in case of slanted surfaces, as they disregard the fronto-parallel assumption [Bleyer et al., 2011; Galliani et al., 2015; Schönberger, 2018]. Thus, slanted scene areas are efficiently reconstructed, while traditional local and SGM methods still struggle with this challenge. On the other hand, PatchMatch lacks an explicit smoothness term and imposes implicit constraints based on plane propagation.

The major open challenge for PatchMatch remains the inefficiency of the similarity measures to deal with matching ambiguities, a common problem in all depth estimation approaches based on visual appearance similarity, either two-view or multi-view. In large textureless areas and non-Lambertian surfaces, photometric consistency measures alone often struggle to recover reliable depth estimates. Nevertheless, due to its robust optimization scheme, PatchMatch performs better in such areas than most global algorithms, as the latter can easily get trapped into local minima. Unfortunately, reliable depth estimates cannot be efficiently calculated in particularly low discriminative areas, even with such a robust propagation. Many efforts have been made in this direction, either by using more efficient propagation and sampling patterns [Xu and Tao, 2019] or coarse-to-fine approaches [Wei et al., 2014; Xu and Tao, 2019; Xu et al., 2020]. However, it is not guaranteed that matching ambiguities will be efficiently tackled even with these improvements. Therefore, current research focuses on incorporating higher-level semantic or structure priors for the scene in conventional or learning approaches. Incorporating such information through hand-crafted cost functions is non-trivial, yet, in this dissertation, two frameworks are introduced in this direction (Chapters 5 and 6).

Chapter 4

Semantic segmentation in 3D reconstruction

Semantic segmentation is a key topic in computer vision and a fundamental component towards visual scene understanding. The term scene understanding refers to the broad research field that attempts to perceive and analyze a 3D scene on different levels of abstraction, directly in 3D space or using images. Often performed in a dynamic way, the scene is analyzed with respect to its geometric structure, functional and spatial relationships between objects as well as their semantic reasoning. By definition, it is a non-trivial problem, and it may combine image-based 3D reconstruction with classification, semantic segmentation, and object detection. Along these lines, scene understanding refers not only to the detection of visual and geometric features of a specific scene but also to their enhancement with information about the physical world in a humanly meaningful way; in other words, scene understanding aims to resemble, up to some extent, the human perception system.

Different from *image classification*, where typically a dominant object present in the scene is recognized and the respective single label is assigned to the entire image (e.g., “cat”, “dog”, “building”), *semantic segmentation* is defined as the fine-grained process of assigning a semantically meaningful label to every single pixel of the scene; also known as pixel-wise classification. That is, each pixel may have a different label, categorizing it in a separate semantic class. In this way, semantic segmentation can separate, group, highlight, and extract clusters of pixels with similar attributes across an image. It has a vast field of applications spreading from medical imaging to autonomous navigation, city mapping, and localization. Early methods included simple techniques relying on low-level vision cues, such as thresholding [Otsu, 1979]; region-growing [Nock and Nielsen, 2004], and clustering [Dhanachandra et al., 2015] methods have also been used, while more sophisticated approaches formed the problem as a Markov Random Field minimization [Plath et al., 2009]. Handcrafted features and flat classifiers have also been widely used in the past for semantic segmentation [Shotton et al., 2009, 2008; Fulkerson et al., 2009]. However, in recent years, the rise of deep learning has revolutionized the field; robust algorithms have been generated that can efficiently segment images [Long et al., 2015] as well as point clouds [Qi et al., 2017]. Closely related to semantic segmentation are the research fields of *object detection*, where the goal is to identify each scene object and typically localize it with a bounding box [Girshick et al., 2014; Ren et al., 2015], and *instance segmentation*, where all the entities of the semantic class are identified with a separate mask [Dai et al., 2016; He et al., 2017]. More recently, *panoptic segmentation* has also been introduced, combining both semantic and instance segmentation [Kirillov et al., 2019].

Higher-level semantic cues are considered essential for various computer vision applications, where plain geometric and visual appearance information is not enough. In the concept of 3D reconstruction from images, semantically segmented 3D models would enable better scene understanding. Indeed, a semantic 3D scene reconstruction can allow for further analysis and re-utilization, for instance, class-specific operations or semantic completion of parts of the scene that are not captured. Yet, semantic segmentation directly in the 3D space typically requires an independent post-processing module, which is rather computationally costly as it typically involves complex mathematical operations.

In this chapter, based on previous work [Stathopoulou and Remondino, 2019a,b], an integrated image-based 3D reconstruction pipeline is proposed, exploiting semantic information to generate semantically enriched point clouds and thus, facilitate scene understanding. Towards this end, rather than working in the 3D space, semantic segmentation is performed on the image level; during 3D reconstruction, each input image is accompanied by its semantic equivalent. Image semantic segmentation can be achieved using standard supervised learning techniques, yet the large amount of training data needed is considered a burden. In that respect, a high-resolution benchmark dataset with rich annotations is developed and introduced in this thesis, particularly aided to historic building facade segmentation, to enable efficient segmentation in similar scenarios. The 3D reconstruction is performed with a combination of open-source SfM and MVS practices that have been proven to achieve high-quality results [Stathopoulou et al., 2019]. In the integrated pipeline, for every image, the semantic information of each pixel is carried along with the whole reconstruction procedure. At the time of reconstruction, based on the label of each pixel, rules can be applied, enabling selective reconstruction of any semantic class by demand. In this way, the resulting 3D models are free of the classes that typically add noise and are uninformative, such as the sky. Meanwhile, class-specific reconstruction of particular areas of interest is enabled, e.g., only the building openings. Finally, the label of each pixel can be projected on the reconstructed 3D points, resulting in semantically enhanced point clouds. Although using facade segmentation as a proof of concept, the proposed semantic 3D reconstruction approach can be generalized in diverse scenarios, spanning from building modeling to urban mapping, and hence adequate for terrestrial or airborne applications.

4.1 Semantic segmentation on images

The implementation proposed in this chapter uses pixel-wise semantic labels for each input image of the 3D reconstruction. To generate these semantic equivalents, a deep learning segmentation pipeline on images of building facades is proposed. Therefore, in the next paragraphs, the basic functionalities of such learning architectures in general as well as specifically for semantic segmentation purposes are briefly described. For a more comprehensive study on deep learning in computer vision, the reader is referred to a plethora of high-valued textbooks and publications in the field [Goodfellow et al., 2014; Zeiler and Fergus, 2014].

4.1.1 Neural networks

Deep learning methods based on artificial neural networks (ANNs) have gained popularity in recent years due to the increase in computer power together with the expanded availability of training sets and have, consequently, been applied in different fields of data science. Particularly while undertaking visual recognition problems such as image classification and segmentation, as well as object detection and localization, the use of ANNs has become common practice and generally outperforms conventional methods or even the standard supervised machine learning ones. An ANN is a group of interconnected neurons, also called nodes, that have been observed to simulate, in a loose analogy, the neurons in the human brain; in fact, they aim to mimic the way humans perceive signals. Learning methods have been developed based on such stacked, densely connected groups of nodes, also known as architectures. A minimal architecture would include an

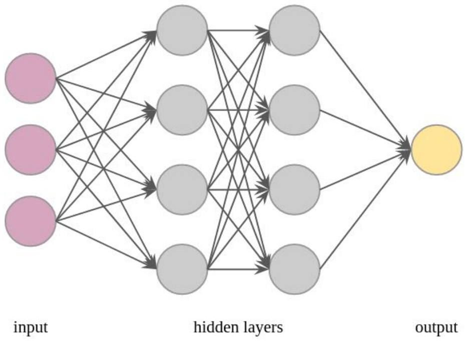


Figure 4.1: **An example neural network** with two hidden layers.

input layer X, an output layer, and some (at least one) hidden layers in between (Figure 4.1). Hidden layers represent mathematical operations z based on weights W and a bias b and have an activation function $a(z)$; these intermediate layers are at the same time input and output layers. Input layers are the input features, and the final output layers provide the prediction result. Activation functions in practice decide if a node will be activated or not, based on thresholds. The primary network form, having a single hidden layer, is conceptualized in the oldest neural network, the perceptron model [Rosenblatt, 1958]. Networks with few hidden layers are called shallow networks, in contrast to the deep ones that include a large number of hidden layers and are thus used for solving more complex problems. Neural networks are commonly used in supervised scenarios and are, hence, data-driven; that is, the algorithm observes the training data in order to learn. A forward propagation step and a backward propagation step are performed during the training procedure, interdependent on each other. Forward propagation is the sequence of mathematical operations from the input to the output and storage of intermediate variables; backward propagation refers to the calculation of the gradients of the variables, traversing the network in reverse order, based on gradient descent. ANNs have been proven robust, even with noisy training samples, and have thus been widely applied for prediction and analyses in various data science fields, spanning from speech recognition to computer vision. The timeline of some of the architectures mostly used in computer vision, along with their basic ideas, is briefly outlined in the next paragraphs.

Convolutional neural networks (CNNs) are a specific class of ANNs, based on convolutional operations with different filters, i.e., kernels, of weights between the network layers. They typically contain several convolutional blocks and non-linear activation function layers, alternating with pooling layers to reduce the size of the representations. After a sequence of convolutional and pooling layers, at the end of the architecture typically some dense layers follow. During these steps, the multidimensional layers are flattened into a single vector; dense layers are fully connected, i.e., not convolutional layers, and the output is derived by a final classifier.

CNN architectures, mainly due to parameter sharing and sparse connections, have by definition fewer parameters than plain, fully connected ANNs, providing a computational advantage and are hence commonly preferred for efficiency. They especially target image problems and have enjoyed great success as they tend to outperform other hand-crafted methods in efficiency and accuracy [Zeiler and Fergus, 2014; LeCun et al., 2015]. The idea of CNNs was originally introduced in the scientific community in the 1980s [Fukushima and Miyake, 1982], and was partially explored in, e.g., document recognition in the LeNet-5 architecture [LeCun et al., 1998], using few convolutional layers combined with average pooling layers, and a small total number of parameters (60K). However, CNNs did not gain popularity till the technological advances and the use of high-performance systems allowed the, for that time, very deep architecture AlexNet with around 60M parameters [Krizhevsky et al., 2012]. AlexNet was the first architecture to achieve considerable accuracy, over 80%, and win the famous ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) challenge, a pioneer effort to collect an extensive image dataset harvested from the web and generate respective annotations to enable training for image classification purposes [Deng et al., 2009; Russakovsky et al., 2015]. Compared to LeNet-5 [LeCun et al., 1998], AlexNet is a relatively complex architecture containing more layers of subsequent convolutions combined with max-pooling layers. A more simplified yet effective, deep architecture is VGG [Simonyan and Zisserman, 2014], with over 130M parameters and fixed kernel sizes; multiple variants of the VGG network exist, e.g., containing 16 or 19 layers. However, very deep networks containing many layers are typically difficult to train due to vanishing gradients caused by the small derivative values during back-propagation [He et al., 2016]. To address this problem, ResNet was introduced [He et al., 2016], where residual blocks, or "shortcut connections", allow for training very deep networks (over 100 layers) without losing performance. In a more complicated, deeper architecture, the GoogleNet/Inception family networks are based on the concatenation of many operations, i.e., convolutions with various filter sizes, max pooling, etc., in the same block, the "inception module"; the high computational cost of such an increased network width is compensated using one-by-one convolutions and reducing, in this way, the number of parameters [Szegedy et al., 2015, 2016, 2017]. In a different line of thought, DenseNet [Huang et al., 2017] connects each layer with all the others exploiting feature re-use throughout the network resulting in a condensed model with a relatively low number of parameters and is easy to train. More recently, the demand for reducing the computational cost led to the design of lightweight networks such as MobileNet [Howard et al., 2017] that appoint depth-wise separable convolutions to reduce the number of parameters. To meet the need to deploy standard architectures in low compute environments, EfficientNet allows for uniform scaling of network dimensions based on the available resources [Tan and Le, 2019].

CNN architectures have evolved rapidly across the years and have employed different techniques to optimize their performance, including (1) testing different activation functions, i.e., tanh, sigmoid, or ReLU, to solve the vanishing gradients problem (2) experimenting with the number of stacked layers, (3) applying diverse pooling operations, i.e., average or max pooling, (4) using appropriate classifiers, i.e., softmax for multi-categorical tasks or sigmoid for binary classification. During training, state-of-the-art methods also experiment with various regularization techniques, especially in very deep architectures, to prevent overfitting. Such approaches include L2-regularization of the parameters or the incorporation of dropout layers [Hinton et al., 2012] to set to zero the activation functions of some randomly chosen hidden units in each training step, based on a probability. With such a regularization, the total number of parameters decreases, simulating a smaller network. Data augmentation is a rather engineering solution that is also commonly applied to increase the training set size. Batch normalization has also been introduced [Ioffe and Szegedy, 2015] to normalize the values of the hidden units and speed up, thus, the learning procedure while also bringing a slight normalization effect. Independent of the complexity of the architecture, training per se remains relatively simple, as it mostly relies on stochastic gradient descent as an optimization method with a chosen learning rate. Adaptive optimization algorithms have been introduced to adapt the learning rate and achieve faster convergence, like RMSProp [Tieleman and Hinton, 2012] and Adam optimizer [Kingma and Ba, 2014].

4.1.2 Semantic segmentation using CNNs

Directly applying such deep CNN architectures designed for image classification to semantic segmentation would yield coarse results due to the repeatedly reduced feature map resolution through the recurrent convolutional and pooling operations. Yet, semantic segmentation is a particular task that needs to provide pixel-wise prediction labels as an output; these segmentation maps must be of the same size as the input image, while arbitrary-sized images are often used. One of the challenges of semantic segmentation is the combination of global and local context; global information is used to decide the label ("what") while local information helps define the location ("where"). CNNs, by definition, have a limited receptive field; that is, layer nodes are only locally connected through the weights, inevitably missing the global context.

Early works in the field typically used handcrafted features and flat classifiers such as Random Forest [Shotton et al., 2008], Boosting [Shotton et al., 2009; Ladický et al., 2010] or Support Vector Machines [Fulkerson et al., 2009] to predict class probabilities, often combined with Conditional Random Fields for smoother results [Ladický et al., 2010], yet the performance of such methods was constrained by the limited representation ability of the features.

In the deep learning era, semantic segmentation was typically addressed as a joint task with object detection and its bounding box proposals [Girshick et al.,

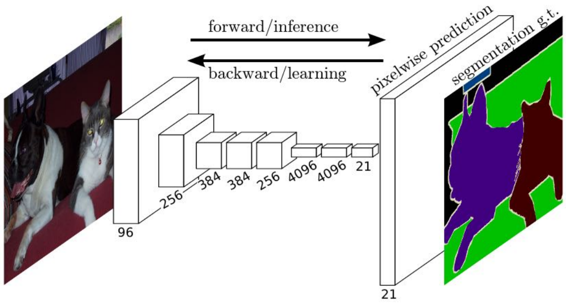


Figure 4.2: **The original FCN network design.** Only convolutional layers are used. Source: [Long et al., 2015].

2014; Gupta et al., 2014] or superpixel segmentation [Mostajabi et al., 2015] followed by region based-classification (R-CNN). Eigen et al. [2014] discarded the image segmentation step and implemented a coarse-to-fine approach using two different networks, transforming the fully connected layers with convolutional ones. The revolutionary work of Long et al. [2015] proposed the use of a Fully Convolutional Network (FCN) architecture particularly designed for semantic segmentation, starting from image classification and exploiting transfer learning. Such architectures contain only convolutional layers and can manage arbitrary-sized images and directly output segmentation maps in an end-to-end manner. Hence, there are no fully connected (dense) layers but are rather replaced by 1×1 (depth-wise) convolutions. FCNs have a downsampling part (encoder), an upsampling part (decoder) of typically one layer, and lateral (skip) connections (Figure 4.2). A standard part similar to VGG [Simonyan and Zisserman, 2014] can be adopted without its last fully connected layers; since such an architecture drastically reduces the feature map size, an upsampling part must be added to map the downscaled feature representations to pixel-wise predictions in the original image resolution. Upsampling is typically performed using deconvolution, i.e., an inverse convolutional operation, to upscale the spatial resolution. However, since the last downscaled layers have limited spatial resolution, details will inevitably be lost during the upscaling; to address this issue, skip connections are introduced. Skip connections concatenate encoder feature maps with their corresponding decoder ones, allowing the combination of high-level information from the deep layers and the appearance information of the shallow layers. Several works based on symmetric encoder-decoder architectures have been proposed in the literature to increase performance of FCNs; U-Net [Ronneberger et al., 2015] uses skip connections with symmetric and direct feature map concatenation, while SegNet [Badrinarayanan et al., 2017] proposes the re-use of pooling indices instead.

Other methods focus on incorporating global context information by integrating Conditional Random Fields (CRFs) into FCN architectures [Zheng et al., 2015], using global average pooling vector across the image [Liu et al., 2015b]. Spatial pyramid pooling structure [Lin et al., 2017a] for multi-scale analysis has also been introduced; PSPNet [Zhao et al., 2017] uses a pyramid pooling module with four pyramid levels to exploit multi-scale feature information and, thus, global context. In the same line of thought, DeepLab [Chen et al., 2017] proposed the “atrous convolution”, i.e., convolution with upsampled features, to extract features in various scales while also incorporating CRFs for more accurate boundaries between the classes.

Semantic segmentation can also be solved by using Generative Adversarial Networks (GANs) [Goodfellow et al., 2014], e.g. [Luc et al., 2016; Isola et al., 2017; Hung et al., 2018] or by adopting attention mechanisms [Chen et al., 2016; Li et al., 2018; Fu et al., 2019]. Recurrent Networks (RNNs) [Byeon et al., 2015; Liang et al., 2016] have also been explored, although they are typically more appropriate for sequence models and applied, therefore, in problems such as natural language understanding. Weakly supervised approaches have also been studied using bounding box annotations or image-level annotations [Papandreou et al., 2015; Dai et al., 2015], and a few fully unsupervised methods also exist applied in small-scale datasets [Ji et al., 2019]. Since an extensive review of all the proposed architectures in the literature is out of scope for this thesis, for a more detailed overview, the interested reader is referred to Minaee et al. [2020].

4.1.3 The U-Net architecture

Initially proposed for biomedical imaging applications, the U-Net [Ronneberger et al., 2015] architecture has become a common practice for semantic segmentation in various domains due to its efficiency. U-Net improves upon the fully convolutional architecture and contains two symmetric encoding-decoding parts (Figure 4.3), and is commonly referred as “hourglass architecture”; the first part captures low-level, i.e., detailed, contextual information but has low spatial resolution due to the repetitive downsampling of the images during the convolutional operations. The second part uses inverse convolutions for upsampling the feature maps reaching high resolutions, up to the original image size, and thus precise localization while encoding high-level, i.e., rough, contextual information. Each layer contains a block of consecutive convolutional operations and activation functions with some max-pooling layers, along with depthwise convolutions. One drawback of the standard encoder-decoder architecture is the loss of fine-grained information about the image through the encoding process. To overcome this limitation, a fundamental part of U-Net architecture is the usage of skip connections which connect the early layers with the latest ones and concatenate these specific sets of activations. By doing this, they leverage the fine-grained spatial information of the low-level features from the encoder layers with the high-resolution and high-level contextual information of the decoder ones. Finally, a one-by-one convolution layer outputs a segmentation map with the original input image width and height and channels equal to the number of classes. Given its good performance in many applications [Zhou et al., 2018; Zhang et al., 2018], even with relatively few

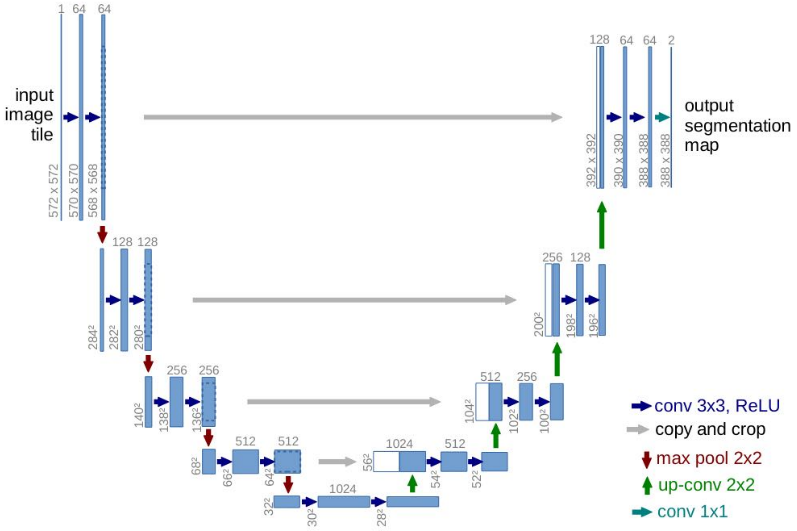


Figure 4.3: **The original U-Net network design.** U-Net is a symmetric-shape architecture with skip connections. Source: [Ronneberger et al., 2015].

samples, in the experiments within this research work, the U-Net architecture is employed for model training.

4.1.4 Semantic segmentation benchmarks

Semantic segmentation has been investigated as a standalone research task or as an auxiliary module for other pipelines such as navigation and obstacle avoidance. Commonly, pre-trained models on generic image classification or semantic segmentation tasks are used to enable easier convergence and are especially benefiting when limited training data is available. Such models are trained on large-scale image recognition milestone datasets such as the *ImageNet* [Deng et al., 2009; Russakovsky et al., 2015], the *PASCAL Visual Object Classes (VOC)* dataset [Everingham et al., 2010], the *MS COCO* dataset [Lin et al., 2014] and so on.

In recent years, many semantic segmentation algorithms have been developed toward scene understanding for autonomous driving. Accordingly, several benchmark datasets with ground-truth 2D data for semantics of urban street-level scene analysis have been introduced, often accompanied by relevant depth maps or 3D information. Among the most prominent efforts are *CamVid* [Brostow et al., 2009], *CityScapes* [Cordts et al., 2016] datasets offering a variety of street scenes and semantic classes, while the *KITTI* road estimation [Fritsch et al., 2013] has also been used for road segmentation. The *Mapillary Vistas* [Neuhold et al., 2017] dataset is a larger scale dataset of this category in terms of the number of images and respective classes. Such datasets are specific to urban scene understanding and have, thus, limited scene variation while certain classes are considered. Other than street scenes, considerable work has also been done on semantic RGB-D datasets from indoor scenarios, which are more complex and include a larger spectrum of semantic classes. RGB-D scenes refer to video sequences acquired using commercial sensors like the pioneer *NYU Depth v2* [Silberman et al., 2012] on depth estimation and 2D semantics that bootstrapped progress on the topic. However, the continuously growing demand for more training data with their respective ground truth annotations led to the release of larger-scale datasets. For instance, the *SUN RGB-D* [Song et al., 2015] or the more recent *Stanford 2D-3D semantics* dataset [Armeni et al., 2017], the *ScanNet* dataset [Dai et al., 2017] and the synthetic *SceneNet RGB-D* [McCormac et al., 2017], are more complete initiatives towards this end, since they provide ground truth 2D, 2.5D, and 3D data. However, these datasets are limited to indoor scenarios restricted by the capabilities of RGB-D sensors, have mostly low-resolution data, and are rather acquired with SLAM-like trajectories, hence typically cannot be used for MVS purposes.

The photogrammetric community provides semantic airborne image datasets for geospatial scenarios; the *UDD5* [Chen et al., 2018a] and the *UAVid* [Lyu et al., 2020b] series focus on drone image sets, while the *ISPRS 2D semantic labeling* dataset provides very high-resolution orthophotos with semantic classes for remote sensing applications [Rottensteiner et al., 2012; Niemeyer et al., 2014] and has been widely used by the community. More recently, benchmark datasets shifted the attention towards 3D point cloud segmentation deriving from image sets [Hu et al., 2021] or acquired by LiDaR sensors in street-level [Hackel et al., 2017] or airborne [Zolanvari et al., 2019; Kölle et al., 2021].

Undoubtedly, benchmarking plays a fundamental role in computer vision tasks as they offer a challenge that drives the research towards novel directions and establishes a common baseline on which new algorithms are evaluated. Moreover, model training relies on massive amounts of annotated ground-truth data; using sparse or mislabeled data samples typically leads to domain-dependent models that may not generalize well to diverse scenarios.

4.2 The 3DOM Semantic Facade benchmark

In photogrammetry and remote sensing applications, research has focused primarily on the semantic segmentation of airborne and satellite images, often accompanied by Digital Surface Models (DSM) or airborne laser scanning data for urban-level mapping and land cover purposes. Earlier methods used handcrafted features, simple classifiers [Volpi et al., 2013], or shallow networks [Bischof et al., 1992] for semantic segmentation of images. In more recent years, challenged by the introduced benchmarks, the photogrammetric community proposed methods tailored for object detection and building extraction [Marmanis et al., 2016; Cheng et al., 2016]. Segmentation of 3D point clouds derived by photogrammetry or acquired by laser scanning has also been recently investigated [Özdemir et al., 2021; Can et al., 2021]. Such large-scale applications are challenging, mainly because of the need to manage a massive amount of 2D or 3D data and potentially extreme class imbalance. This fact is particularly true when dealing with satellite images where related research has focused mainly on building extraction and change detection [Vakalopoulou et al., 2015; Huang et al., 2016; Zhang et al., 2019b] or land cover applications [Helber et al., 2019]. Within the context of this dissertation, a new benchmark for facade segmentation is proposed. In this chapter, rather than using airborne data for city-scale mapping, terrestrial images are used for semantic segmentation and building-scale 3D reconstruction. Indeed, even with oblique images, the reconstruction of the facade surfaces is challenging in airborne applications, hence street-level acquisitions are often used for complete reconstructions of urban scenarios. To this end, *3DOM Semantic Facade*, a novel facade segmentation benchmark is introduced.

4.2.1 Motivation and overview

Automatic facade segmentation and parsing is an important research topic in scene understanding with applications in street scene reconstruction, urban-scale modeling, and building analysis. In such a process, each pixel is assigned to a semantically meaningful class for the specific application. For instance, structural components, such as windows and doors, commonly should be identified to enable further analysis. For historic building facades, in particular, semantic segmentation is also important for identifying specific characteristics of architectural styles. In general, building facades provide a rich test bed for evaluating semantic segmentation methods as they can be highly variant in architectural styles and feature different characteristics. That being said, and especially due to this variability, facade segmentation is considered a non-trivial task.

Building facades, like the vast majority of man-made architectures, are highly structured scenes. Hence, prior knowledge cues about facade appearance and layout can be used for semantic segmentation. Facades are mostly regular and symmetric and follow a "Manhattan world" layout assumption; that is, they are characterized by structural regularities and dominant directions. Moreover, in such scenes humans easily perceive other semantic cues; for instance, the sky is expected to be in the upper part of the image, the street level is in the lower part, and windows are typically placed in repetitive patterns in grid-like arrangements, etc. Earlier works typically used such a priori knowledge using shape grammar and symmetry for facade parsing and predicting the structure in unseen data with procedural modeling [Müller et al., 2007; Zhang et al., 2013]. Random forests and CRF classifiers have also been implemented with or without grammars [Teboul et al., 2010; Riemenschneider et al., 2012; Martinovic et al., 2015; Mathias et al., 2016; Rahmani et al., 2017]. However, the design of hand-crafted features and shape grammar rules limit the applicability of such traditional approaches in

Table 4.1: Available benchmarks for semantic segmentation on building facades.

dataset	year	#images	image resolution	#classes
eTrims	2009	60	512 × 768	4, 8
LabelMe	2010	945	varying, max dim: 703	8
ECP	2010	104	varying, max dim: 646	8
Graz50	2012	50	2590 × 1715	4
CMP	2013	400	varying, max dim: 1024	12
RueMonge	2014	428	800 × 1067	8
3DOM Semantic Facade	2019	428	varying, max dim: 6048	5

more complex scenarios and architectural styles. Recent works employ facade segmentation with deep learning features, either utilizing the structure knowledge rules for regularization [Liu et al., 2017] or exploiting full automation for easier generalization [Jampani et al., 2015; Schmitz and Mayer, 2016; Liu et al., 2020b; Ma et al., 2020], achieving improved results. Alternatively, other works, use object detection approaches to identify facade openings [Hensel et al., 2019].

Given the interest of the community, several facade segmentation benchmarks have been proposed in the literature in the past decades. One of the pioneer efforts, the *eTRIMS* dataset [Korc and Förstner, 2009], consists of 60 images depicting facades of simple architectural styles and considers eight semantic classes along with their instances. Adopting a similar nomenclature, Fröhlich et al. [2010] proposed the *LabelMe* dataset that contains a significantly larger amount of facade images. The *Ecole Centrale Paris (ECP)* dataset [Teboul et al., 2010] provides rectified images of facades in Paris with eight defined classes. Riemenschneider et al. [2012] introduced the *Graz50* dataset consisting of 50 facade images of various architectures from buildings in Graz, Austria, while four semantic classes are defined. The *CMP facade* dataset [Tyleček and Šára, 2013] contains around 400 rectified facades from different cities featuring diverse styles and 12 semantic classes. Towards combined 2D and 3D segmentation, the *RueMonge2014* dataset [Riemenschneider et al., 2014] provides semantic labels on 428 images, along with the reconstructed 3D point cloud and mesh. The details of these datasets are summarized in Table 4.1.

During the research work leading to this thesis, the *3DOM Semantic Facade*[1] dataset was introduced in Stathopoulou and Remondino [2019a] as a new real-world, densely-annotated, semantic segmentation benchmark for building facades. The dataset has been publicly released and updated since then. It contains a selection of 227 previously collected high-resolution images across various cities in Italy. They depict historic building facades from the street level. The buildings, although of similar height, feature a significant diversity in architectural styles

[1] https://github.com/3DOM-FBK/3DOM-Semantic-Facade

Table 4.2: Acquisition details of the 3DOM Semantic Facade benchmark.

scene	#images	resolution	camera model
Palazzo Albergati	21	4608 × 3072	Nikon D3100
Bologna Portici	55	4608 × 3072	Nikon D3100
Piazza del Campidoglio	52	4416 × 3312	Canon PS G10
Palazzo Chigi	20	4416 × 3312	Canon PS G10
Lecce Teatini	12	6000 × 4000	Nikon D5300
Lecce Duomo	9	6000 × 4000	Nikon D5300
Piazza Navona	15	4000 × 3000	Samsung ST45
Piazza Duomo-Trento	43	6048 × 4032	Nikon D3X

and structural characteristics spreading from traditional historic center buildings to cathedrals, which was one of the prime motivations for creating such a dataset. The *3DOM Semantic Facade* dataset is the first public benchmark for such a purpose using high-resolution images and pixel-level annotations.

The images were carefully selected based on a predefined rationale, fulfilling some basic criteria. First, they should be of high resolution to be suitable for 3D reconstruction purposes of high requirements. To this end, they also should demonstrate enough overlap for dense reconstruction purposes. Moreover, they should be diverse enough to cover a wide variety of building structures in typical historic city centers of Italy and similar style cities across Europe as well as acquisition conditions, such as lighting, weather conditions, distance to the object, camera sensor, etc. Each image needs to contain objects belonging to different semantic classes for balance. The details of the acquired data are presented in Table 4.2.

4.2.2 Classes and nomenclature

The dataset aims at a generic facade segmentation and identification of the basic components; each class should have a clear and unambiguous semantic meaning, while all classes should be unique with respect to their geometric characteristics and visual appearance. Therefore, five semantic classes were defined, namely "wall", "sky", "obstacle", "window", and "door"; the class "obstacle" includes all parts of the scene that are typically unwanted in photogrammetric scenarios, e.g., moving objects such as cars, bikes, and pedestrians, but also trees, vegetation, traffic signs, and the street itself. Such a nomenclature facilitates the isolation of objects considered noise, e.g., the sky and obstacles, while enabling the selective reconstruction of specific semantic classes of interest such as the facade walls or the openings "window" and "door". In the *3DOM Semantic Facade* benchmark, the sky is considered a separate class and the background is not classified. Images were selected in such a way to include all semantic classes if possible; that being said, classes are inevitably imbalanced since the number of pixels for every class varies a lot.

4.2.3 Ground truth annotation

Rich data annotation is essential for any supervised learning problem, as the training algorithms heavily rely on accurate ground truth labels for training. In semantic segmentation, data annotation refers to the manual assignment of a semantic label to each pixel, whereas in object detection, typically, each object is annotated by a box. It is an indispensable, although laborious and time-consuming stage that cannot be fully automated yet. Labeling should be performed with caution to avoid gross errors that would affect the training quality. Common image processing software can be used for manual annotation; however, specially designed interfaces also exist, e.g., Labelbox[2], LabelMe[3]. Recently, the high demand for data labeling led to the launching of specialized services that perform this task by demand, or even dedicated crowdsourcing platforms.

For 3DOM Semantic Facade dataset, image annotation was performed manually in-house using image processing software[4]. A custom coloring policy is employed, as shown in Figure 4.4. Ground truth (GT) data were manually annotated in a fine-grained way by selecting and grouping together pixels that are visually identified to share the same properties and, thus, belong to the same semantic class. All labeled images have been manually cross-checked, guaranteeing consistency and high quality of the annotations. Pixels along class borders, commonly assigned with ambiguous labels, were further refined and finally assigned to either a class.

4.3 Network architecture and results

Deep learning methods have demonstrated great success and remarkable performance in semantic segmentation. However, in practice, there are some limitations due to the excessive computational time, the high requirement for memory resources as well as the demand for a large amount of training data in order to generalize well in diverse scenarios. In this section, a deep learning pipeline for an efficient generation of segmentation maps for building facades is presented, using the proposed benchmark. Due to the architectural complexity of the facades, symmetry rules are not considered in the scenario presented in this thesis; instead, following a deep learning approach, a generic semantic segmentation method is applied without any prior cue. In this way, the potential of the proposed

[2] https://labelbox.com/

[3] https://github.com/wkentaro/labelme

[4] https://www.gimp.org/, https://www.adobe.com/products/photoshop.html

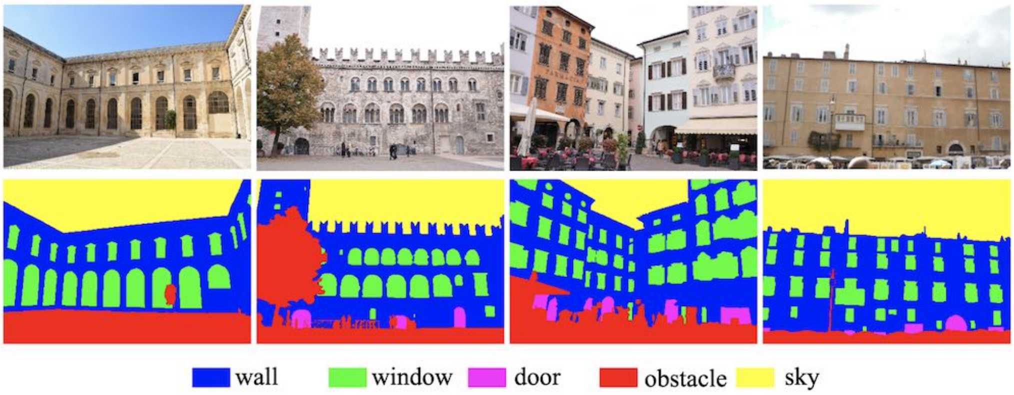


Figure 4.4: **The *3DOM Semantic Facade* benchmark**. RGB images and respective ground truth labels. From letf to right: Lecce Teatini, Piazza Duomo - Trento (Duomo), Piazza Duomo (other buildings), Piazza Navona.

benchmark in the facade segmentation of heritage building facades is explored. In particular, a standard U-Net network combining an encoder and a decoder part is exploited. The U-Net architecture was chosen as it has been proven to perform efficiently for similar semantic segmentation tasks; indeed, one of its most important properties is that output images can have the same resolution as the input images since the deconvolution operations restore the output feature maps to the original input resolution, resulting in a class label corresponding to each pixel. The network is trained as a multi-categorical classification problem in a supervised way using the ground truth manually annotated labels. It is to be noted that multi-categorical classification stands for the procedure when each pixel can belong to only one of the considered classes, while binary classification refers to a problem with only two classes, commonly one class of interest and one background class.

Some images, depicting close-up views of the facades and thus containing mostly one dominated class, were excluded from the set. The dataset considered finally contains 211 images, from which 14 images are kept for the test set and the remaining images are split into training (90%) and validation (10%); these operations are performed with a random shuffling approach to keep the data distribution relatively homogeneous. Given the memory constraints, the original high-resolution images must either be downsampled or divided into tiles. Severe downsampling would inevitably cause detail loss; on the contrary, tiling preserves the fine details but can be ambiguous due to limited global context. A combination of the two strategies could potentially provide a compromise between both downsides. Following this strategy, in this experiment the images were first downsampled by a factor of four to each dimension. Subsequently, crops of 256×256 pixels were extracted with horizontal and vertical stride of 128, creating a total of 9120 samples. A pseudo-class "background" is also used, containing possible unassigned

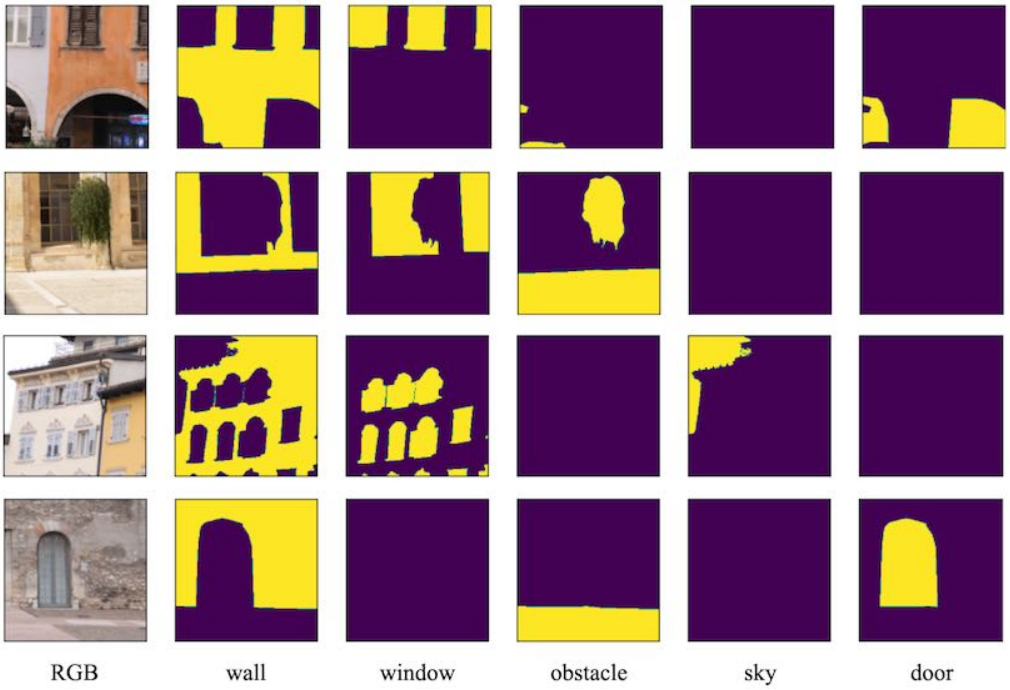


Figure 4.5: **Input RGB image crops and the respective classes.** Each class is shown as binary segmentation maps, where yellow indicates the respective label and purple indicates everything else.

pixels, yet in our dataset, this case was rare (less than 1% of the total number of pixels). The training procedure is based on Tensorflow (version 1.15.0); the used models are as implemented in the open-source library *segmentation_models*[5]. The particular library is chosen as a standard, broadly-used and open-source library for semantic segmentation on Tensorflow, but other libraries can be used as well.

4.3.1 Data augmentation

Data augmentation, already used in early neural network applications [LeCun et al., 1998] can be considered as a regularization method such as L2 regularization or dropout. It is applied to increase the number of training samples, performing a set of simple transformations to the already existing ones. Standard data augmentation practices include geometric transformations such as translation, rotation, flipping, cropping, scaling, warping, or even affine and perspective transformations. Radiometric manipulation refers to brightness and contrast manipulations, color space shifting, noise application, image blurring, and sharpening, among others. Data augmentation has proven to improve algorithm performance and reduce overfitting [Perez and Wang, 2017; Shorten and Khoshgoftaar, 2019], especially

[5] https://github.com/qubvel/segmentation_models

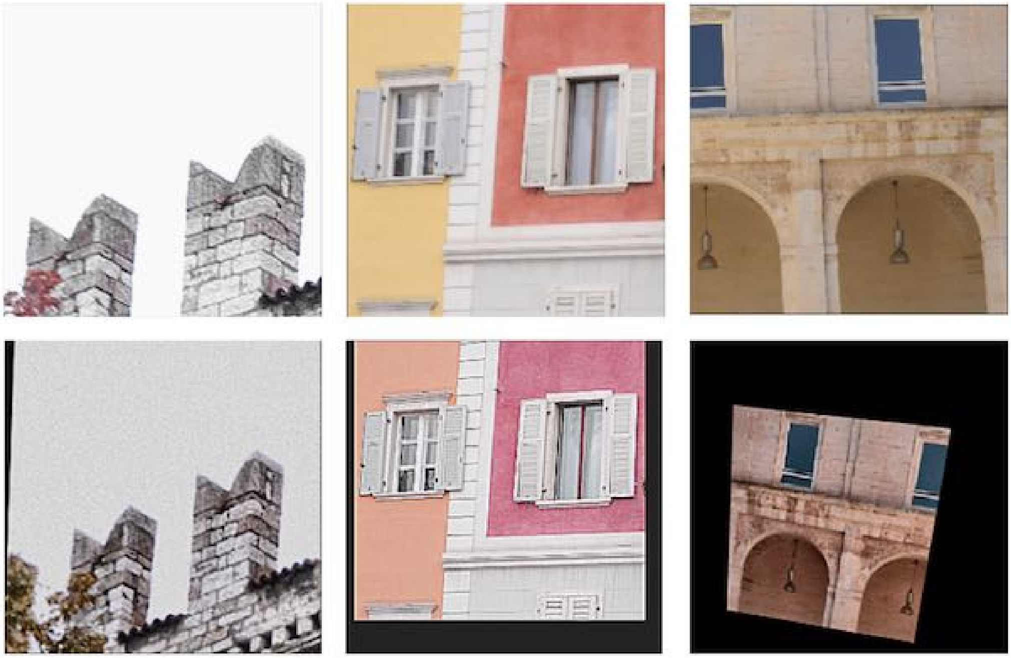


Figure 4.6: **Data augmentation examples.** Original RGB image samples (upper) undertaking random combinations of data augmentation techniques (lower).

for applications in domains where large datasets are inaccessible or unexisting, for instance in the medical imaging field. In fact, this effectiveness has also launched research towards learning data augmentation, e.g. by augmenting the feature space [DeVries and Taylor, 2017] or GAN-based augmentation [Frid-Adar et al., 2018], particularly used for unbalanced data. For a comprehensive taxonomy of current data augmentation techniques, the interested reader is referred to the detailed review article of Shorten and Khoshgoftaar [2019].

In the experiment presented in this chapter, data augmentation was applied to decrease overfitting and achieve faster training convergence and better generalization in unseen datasets (Figure 4.6). The applied data augmentation strategy includes randomly applied horizontal flips, affine transformation, perspective transformation, random crops, brightness, gamma, contrast, hue manipulations, image sharpening and blurring, and, Gaussian noise. To be noted that geometric transformations are applied to both RGB images and their respective masks, whereas radiometric manipulation is performed only on the RGB image.

4.3.2 Training

The proposed dataset containing RGB images and their corresponding ground truth segmentation maps are used to train the network based on stochastic

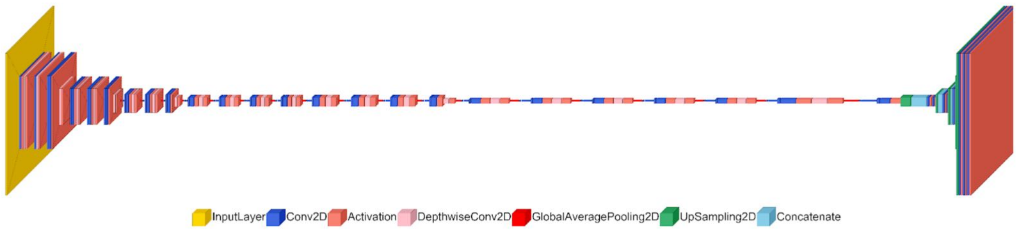


Figure 4.7: **The employed architecture EfficientNet-B2 for the proposed pipeline.** Dimensions are scaled to 512×512 input size for better visualization.

gradient descent optimization in Tensorflow (version 1.15.0).

Architecture. Training a deep convolutional neural network (CNN) from scratch is challenging, since extremely large datasets should be used for efficient learning and generalization. An alternative to full training is transfer learning, commonly also named domain adaptation [Weiss et al., 2016; Zhuang et al., 2020]; in transfer learning, a network that has been trained on large datasets such as *ImageNet* [Deng et al., 2009; Russakovsky et al., 2015] or *Microsoft COCO* [Lin et al., 2014] is fine-tuned for another application or another domain. This technique is effective and standard practice in visual recognition tasks since they mostly share the low-level image features that are better learned with large datasets; the knowledge is then transferred from one task, e.g., image classification to a similar one, e.g., semantic segmentation. There are various techniques for transfer learning, such as feature representation transfer, fine-tuning, and pre-training. In this experiment, a pre-trained EfficientNet-B2 [Tan and Le, 2019] was used as a backbone (Figure 4.7). This model was chosen as a compromise between the available resources and performance for the given crop size. In fact, EfficientNet is a family of models B0-B7; the basic model B0 was introduced as a clean network architecture with a compound scalable strategy of all three dimensions of the network based on the available resources achieving computational efficiency. Models B1-B7 are increasingly scaled up from the base model B0 using different compound coefficients. The top layers are retrained instead of being frozen, thus weights are updated. The total number of parameters for EfficientNet-B2 is 14K. As commonly performed in semantic segmentation tasks, U-Net architecture [Ronneberger et al., 2015] is followed, enabling prediction map generation in the same resolution as the input images. A softmax classifier is chosen, appropriate for such multi-categorical tasks. Since softmax is between $[0, 1]$, all probabilities add up to 1.

Loss function. In learning tasks, the choice of an appropriate loss/objective function is of utmost importance for efficient training convergence. Various domain-specific loss functions have been proposed in the literature. In semantic segmentation, the loss function should measure the difference between the predicted label $\hat{y}$ and the ground truth label y_{GT} during training. Popular strategies include using distribution-based losses examining the similarity between two distributions, like the binary cross-entropy loss [Yi-de et al., 2004], or its weighted equivalent [Pihur et al., 2007] or the focal loss [Lin et al., 2017b]. Region-based losses are also very common in segmentation tasks measuring the overlap between the training samples, as for instance the dice loss [Sudre et al., 2017] and its generalized weighted version Tversky loss Salehi et al. [2017] as well as the focal Tversky loss [Abraham and Khan, 2019]. Boundary-based losses like the Haudorff distance loss have also been used recently [Karimi and Salcudean, 2019]. An extensive survey on the state-of-the-art loss functions in semantic segmentation can be found in [Jadon, 2020]. In the proposed approach of this dissertation, the specific semantic segmentation task is considered a multi-categorical classification problem; that is, each pixel can belong to only one out of many possible classes. A custom joint loss function $\mathcal{L}$ is used, combining categorical cross-entropy loss $\mathcal{L}_{CE}$ and the Jaccard index $\mathcal{L}_{JAC}$:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathrm{w}\mathcal{L}_{JAC}, \tag{4.1}$$

where w is a weight factor, here set to 0.1. The categorical cross-entropy loss is defined as:

$$\mathcal{L}_{CE} = -\sum (y_{GT} \log(\hat{y})). \tag{4.2}$$

And the $\mathcal{L}_{JAC}$ loss based on the Jaccard index, is similar to dice loss, is calculated as the ratio between the overlap of the positive instances between two sets:

$$\mathcal{L}_{JAC} = 1 - \frac{A \cup B}{A \cap B}. \tag{4.3}$$

Cross entropy is a pixel-wise loss function. The Jaccard loss is a global function that provides better perceptual quality. Dice loss is particularly useful for segmentation problems where there is class imbalance. Softmax is used as an activation function, appropriate for multi-categorical classification problems.

Hyperparameter settings. The adaptive moment estimator (Adam) optimizer [Kingma and Ba, 2014] with a learning rate α reduction strategy was used in this experiment. Initial learning rate was set to $\alpha = 0.0001$ with a decay factor of 0.2. A mini-batch size of 8 was followed, as a compromise between GPU resources and fast convergence. Training is performed for 25 epochs.

4.3.3 Results and evaluation

Evaluation Metrics. Apart from the visual quality of the results, based on human perception, quantitative metrics are also used for evaluating a trained model. The predicted labels are divided into true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Common evaluation metrics are the overall accuracy, the precision, recall, and the F_1 score. Precision actually gives the percentage of correct predictions, while recall depicts the percentage of the correctly predicted positives. The overall accuracy is the ratio of the correct predictions to the total predicted labels, both correct and incorrect.

$$Overall\_Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.4}$$

Precision actually gives the percentage of correct predictions

$$Precision = \frac{TP}{TP + FN}, \tag{4.5}$$

while recall depicts the percentage of the correctly predicted positives:

$$Recall = \frac{TP}{TP + FP}. \tag{4.6}$$

And their harmonic mean:

$$F_1 = \frac{2(Precision \times Recall)}{Precision + recall}. \tag{4.7}$$

Intersection over Union (IoU), also called Jaccard Index, is one of the most commonly used metrics in visual recognition tasks. In semantic segmentation, it is defined as the area of intersection between the predicted segmentation map and the ground truth label, divided by the area of their union. The confusion matrix, showing a combination of predicted and actual values is an effective way to visualize the prediction performance.

Prediction on the test set. The model is evaluated on the test image tiles; this images come from the same distribution as the training and the validation sets, however, the exact images have been excluded from the training procedure and are, thus, unseen. The achieved mean IoU and F_1 scores are satisfying high, 90.3% and 92.1% respectively 4.3. In Figure 4.8 shows some indicative results on the test tiles. It can be observed that the model fits on the most dominant classes (e.g., wall, window, sky) but fails to learn efficiently the features of minor categories that are not well represented in the data (i.e., obstacle, door).

Table 4.3: Prediction metrics on the image tiles.

loss	0.069
mean IoU (%)	90.30
mean F_1 (%)	92.01

Prediction on the test set - full sized images. Full sized images are the ones from which the test set tiles have been generated, e.g., the downsampled by four original sized images; a tiling procedure is followed with certain overlaps for optimal results. Predictions are performed for 512 × 512 crops and different strides. The results are shown in Figure 4.8 and the respective metrics for each semantic class are summarized in Table 4.4, while the normalized confusion matrix in Figure 4.9 summarizes in a graphical way the recall for each semantic class. It can be observed that the highest scores are achieved for the major classes wall, window, sky, while the underrepresented classes door and obstacle show lower performance. The good performance on the facade class is expected since it well represented in any building-related semantic segmentation.

Table 4.4: Per-class metrics (%) for the full images.

	wall	window	obstacle	sky	door
pixel accuracy	96.89	98.35	98.65	99.87	99.02
precision	96.91	94.49	96.72	99.43	73.23
recall	97.55	92.78	95.06	99.52	78.44
F_1 score	97.23	93.63	95.88	99.47	75.75
IoU	94.61	88.02	92.09	98.95	60.96
overall acc.					96.39

Prediction on data from different distributions. In order to test the generalization ability of the proposed method and the effectiveness of the *3DOM Semantic Facade* benchmark, the trained model is used for prediction on unseen images from different distributions without any additional fine tuning. In this experiment, test images of historic building facades from the city of Limassol, Cyprus were used[6]. The original images have slightly different resolutions (6720 × 4480 and 6016 × 4016 pixels). They are downsampled by four as the training and test images for a fair comparison. Although the buildings feature varying architectural styles that are highly variant from the ones used for the model training, the predicted masks identify most classes of the scene overall in a satisfactory level. For these images, no ground truth labels are available, allowing

[6] Data acquired in the acquisition campaign of the Periscope project in June 2021, Limassol, Cyprus.

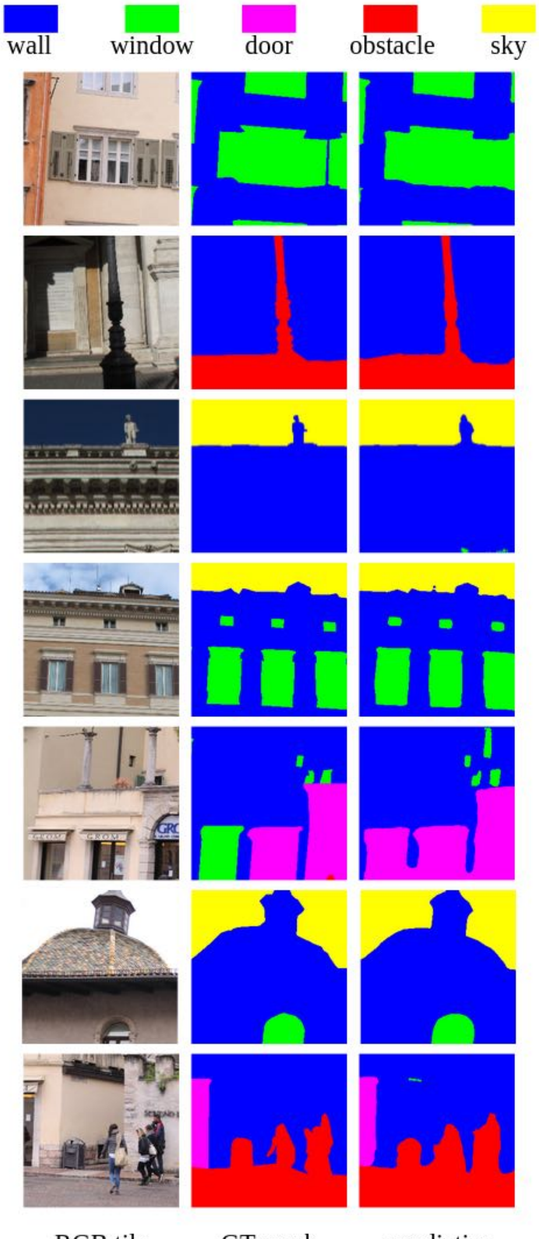
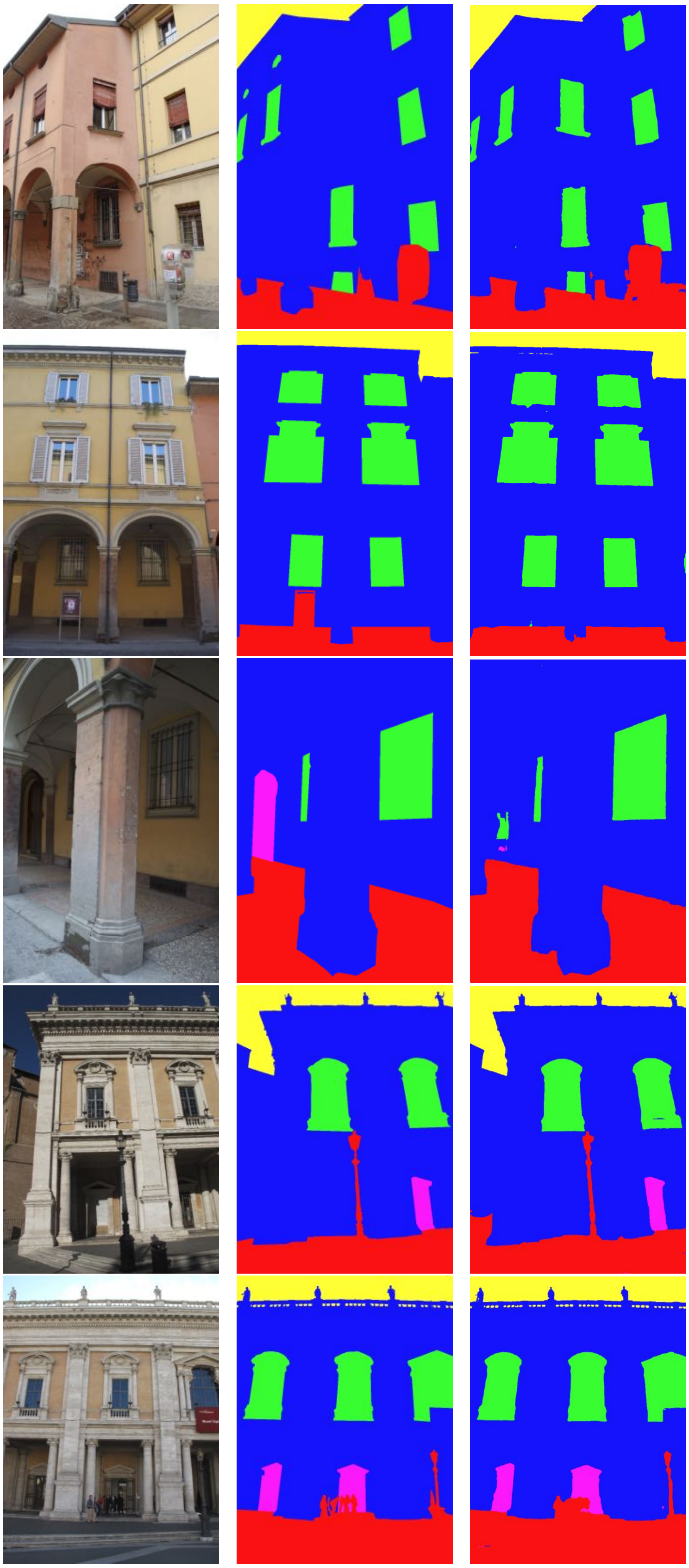
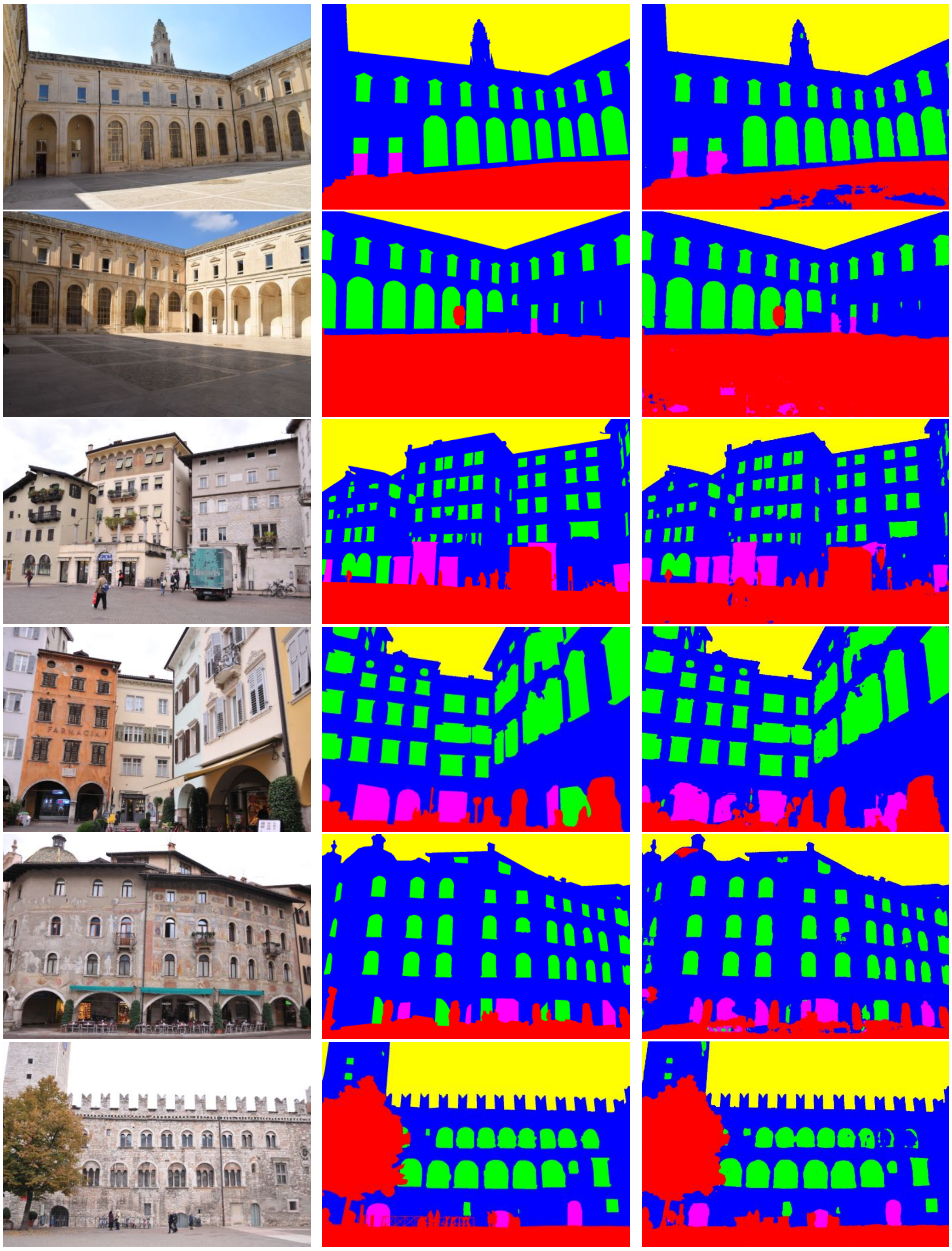
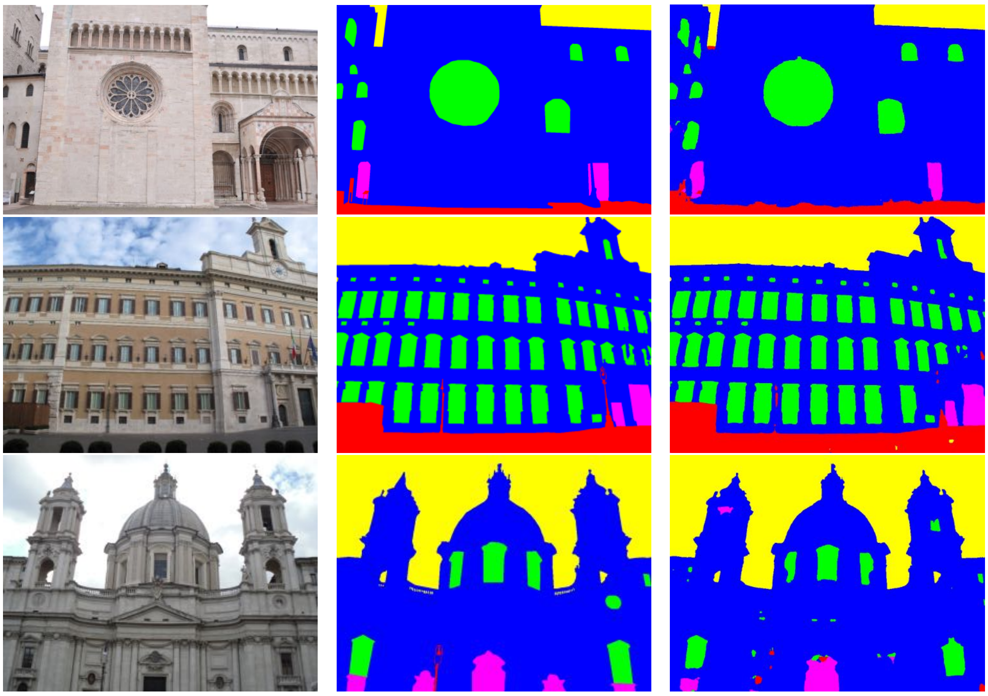


Figure 4.8: **Label prediction results on test set samples.** RGB images, ground truth masks, and prediction masks.

only for qualitative evaluation of the results. The best results are achieved for the sky and wall classes, whereas windows and doors are often interchangeable. This can be explained by the highly different architectural components of the openings.







(a) RGB image (b) GT mask (c) predicted mask

Figure 4.8: Predicted labels on the images of the test set.

4.4 Semantic photogrammetry

Motivated by the increased demand for semantically enriched 3D data in various application domains, within the context of this dissertation, the standard image-based 3D reconstruction pipeline is enhanced with semantic information. The proposed approach aims to generate semantically enriched point clouds while enabling, at the same time, the selective reconstruction of semantically meaningful classes for the various applications. Apart from the methodologies discussed in this Chapter, the derived semantic information can also be integrated in the depth estimation algorithm to improve the completeness and overall quality of the derived 3D point cloud, a concept explained in detail in Chapter 5.

4.4.1 Rationale on 2D segmentation

Semantic segmentation algorithms on 2D images has been proven robust enough in recent years, with high-performance scores using machine and deep learning techniques [Plath et al., 2009; Long et al., 2015; Chen et al., 2015]; however, the respective algorithms for 3D data are still an open challenge. As a matter of fact, semantic segmentation of point clouds requires complex mathematical operations

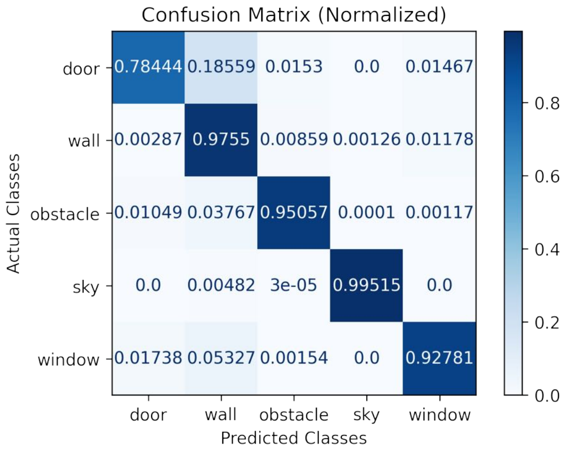
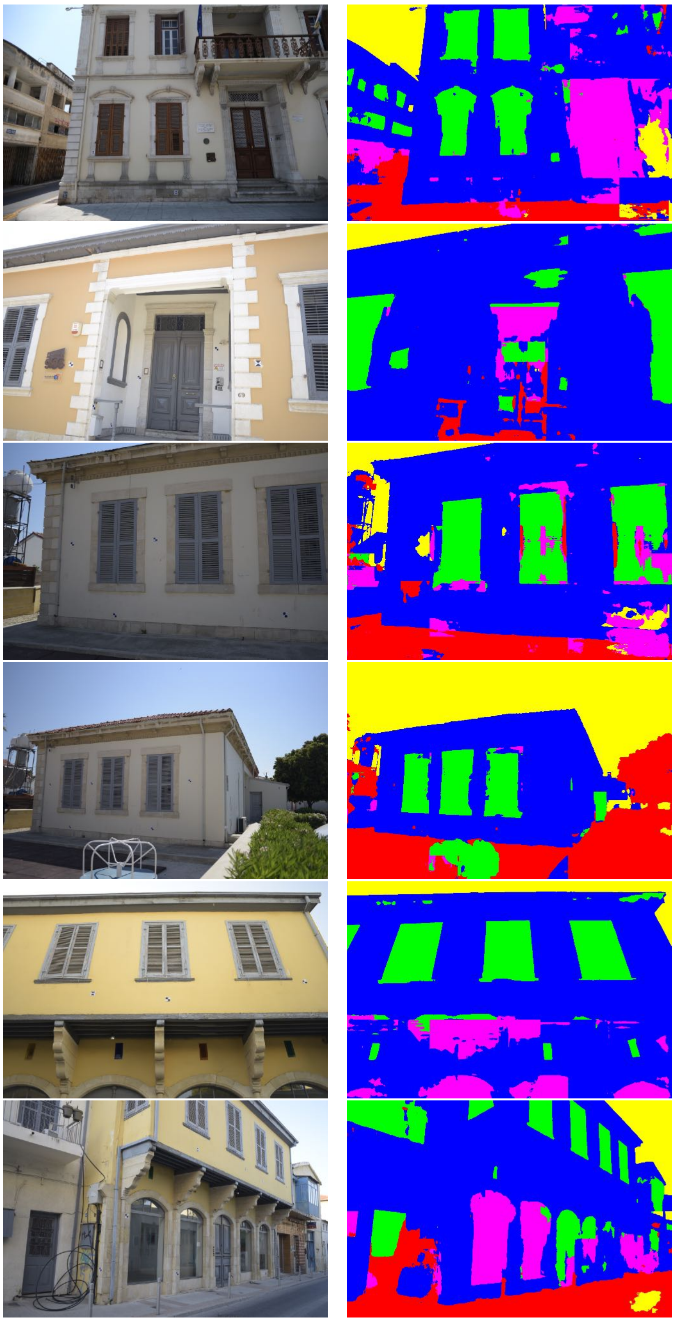


Figure 4.9: **The calculated confusion matrix** for the all classes on the test images.

and huge computational power due to the 3D nature of the data. Moreover, the available GT labels needed for supervision are expensive to obtain and have consequently limited availability with respect to their 2D equivalents. State-of-the art 3D semantic segmentation methods can be roughly categorized into 2D projection-based [Kalogerakis et al., 2017; Wang et al., 2020a; Lyu et al., 2020a], voxel-based [Tchapmi et al., 2017; Liu et al., 2019], and point-based methods [Qi et al., 2017; Thomas et al., 2019; Guo et al., 2021]. Point-based methods tend to be more computationally efficient and deliver accurate point-wise labels, yet they typically suffer from domain dependence and cannot, thus, generalize well in unseen scenarios [Hu et al., 2021]. Moreover, due to the unavoidable noise in the 3D space, deciding the correct label for each pixel is in general non-trivial. Recently, weakly supervised methods have also been proposed to relax the requirement for dense point annotations [Wei et al., 2020].

To circumvent these problems, researchers have been exploiting the massive amount of 2D labels for label propagation in 3D already some years now [Wang et al., 2013]. Particularly for facade segmentation, most works perform segmentation in the image domain, however, few approaches exploit also the 3D domain towards facade modeling [Martinovic et al., 2015]. In photogrammetric applications, semantic segmentation in the 3D space is commonly applied to urban scenarios and the proposed methods are typically tested on the few existing benchmarks, hence having limited generalization ability [Kölle et al., 2021].

In the following paragraphs, an alternative strategy is proposed for label propagation from 2D to 3D and enhancing, thus, 3D reconstruction with semantic labels; semantically segmented images obtained by procedures similar to the one described in Section 4.3 are used.



(a) RGB image

(b) predicted mask

Figure 4.10: **Inference results on unseen images of different distributions.** The RGB images (left) and the predicted semantic maps (right). Images of cases study buildings in Limassol, Cyprus, acquired during the Periscope project campaign (July 2021).

4.4.2 Image-based 3D reconstruction pipeline

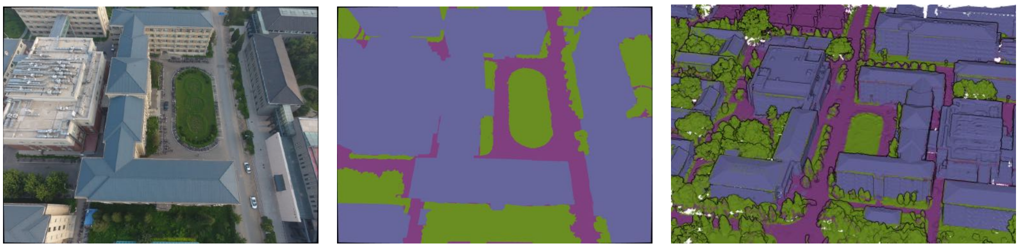
Multi-view 3D reconstruction has achieved impressive results in recent years and several methods have been proposed in the literature as described in Chapters 2 and 3. In an application level, a dense 3D reconstruction of a scene requires a set of images $\mathcal{I}$ as an input along with the respective metadata, whereas GCP points may be also used for scaling and georeferencing the model. The image network should be properly acquired, with sufficient overlap to guarantee redundancy in pixel correspondences and proper intersection angles and enable, thus, a robust reconstruction in the 3D space. For details and guidelines on proper image acquisition, the reader is referred to relative literature [Wenzel et al., 2013; Furukawa and Hernández, 2015]. A typical dense 3D reconstruction pipeline consists of two distinct, yet highly linked, workflows, Structure from Motion (SfM) and Multiple View Stereo (MVS), as outlined in Chapter 1.

In this work, a 3D reconstruction pipeline combining an SfM and an MVS module based on open-source libraries [Moulon et al., 2016; Schönberger and Frahm, 2016; Cernea, 2020] is implemented, inspired by previous work [Stathopoulou et al., 2019]. Sparse reconstruction and camera pose estimation is calculated using the incremental method of [Moulon et al., 2012] followed by bundle adjustment using the Ceres solver [Agarwal et al., 2012]. Given these, a PatchMatch-based approach is adopted for multi-view stereo reconstruction as implemented in the OpenMVS framework [Cernea, 2020].

4.4.3 Semantically enriched point clouds

Efficient segmentation with direct methods in the 3D space cannot be easily applied due to computational complexity and the lack of large GT training 3D datasets. Given these limitations, a straightforward solution would be to exploit the robust 2D segmentation results and transfer the labels on the 3D point cloud in the context of multi-view reconstruction. Thus, the final point clouds, apart from their real RGB color, are enhanced with semantic attributes, enabling 3D semantic segmentation. The proposed method takes as input only the images for 3D reconstruction and their corresponding segmentation masks. The aim is to obtain high-quality dense reconstructions of the scene enhanced by the semantic information.

To achieve this goal, a system has been developed, using the open-source library OpenMVS [Cernea, 2020] as the baseline method. Standard PatchMatch-based MVS reconstruction is performed, but an extra module is added to load and keep in memory the corresponding semantic segmentation map for every input image. Once the final 3D point cloud is fused, the labels are retrieved and every 3D point is assigned to a semantic class. In an MVS scenario, there are redundant ray intersections and hence redundant label information available for each pixel. Therefore, the "best" image across overlapping views needs to be selected to inherit



(a) RGB image (b) GT labels (c) semantically enriched point cloud

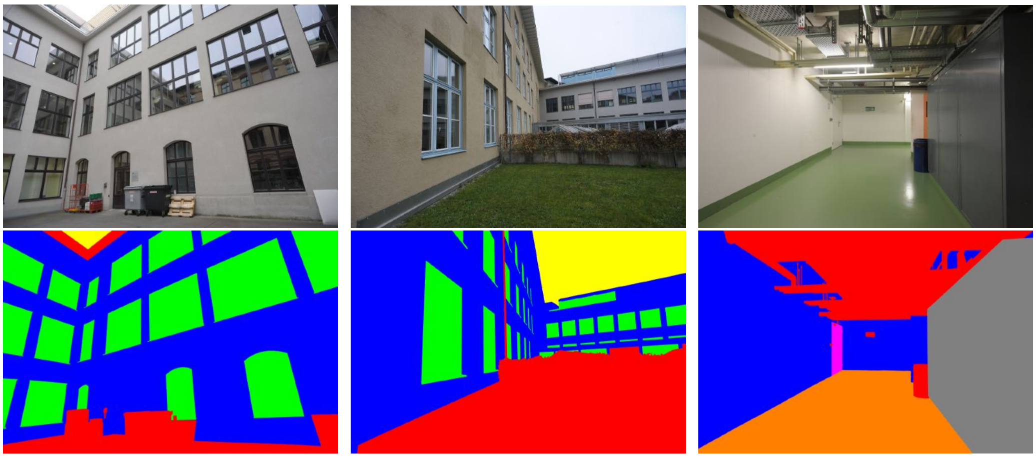
Figure 4.11: **Semantically enriched point cloud of the *UDD5* benchmark** using GT 2D labels.

the appropriate label to each pixel. In this study an approach inspired by the color information assignment in standard texturing procedures is followed [Cernea, 2020].

Recently, a method motivated by the present work proposed a similar semantic photogrammetry approach using orthophotos [Murtiyoso et al., 2021] or the acquired images [Murtiyoso et al., 2022] for class-specific reconstruction and automatic masking purposes, yet their implementation differs from ours.

The experiments discussed in this section refer mainly to semantic segmentation applications for facade segmentation and street-level mapping. In the case of facade segmentation, such semantically enriched results enable the direct identification and segmentation of areas of interest, e.g., building openings, and, at the same time the identification of undesired areas such as the sky. For urban mapping scenarios, a variety of applications exist, such such building or vegetation extraction. However, the proposed method, being easily scalable since the labels do not add a computational burden in the 3D reconstruction process, can be extended also in larger scale scenarios such as urban mapping applications. For example, in Figure 4.11 results on the *UDD5* dataset are also presented, using the GT labels of [Chen et al., 2018a] that include the semantic classes "vegetation" (green), "road" (magenta), "building" (purple), "vehicle" (blue) and "other" (black).

Meanwhile, three selected sequences of the *ETH3D* benchmark were manually annotated, namely *courtyard*, *terrace*, and *pipes* to proof the applicability of the proposed semantic photogrammetry method in indoor and outdoor scenarios (Figure 4.12). The annotated images were further used for the method proposed in Chapter 5 and are to be publicly release to enable further research on the topic. A similar nomenclature as the one decided for *3DOM Semantic Facade* was followed for the sequences *courtyard* and *terrace*. For the indoor sequence *pipes*, ad-hoc classes were defined: "wall" (blue), "floor" (orange), "door" (pink), "wardrobe" (grey) and "other" (red). Figures 4.13 and 4.14 show the output semantically segmented point clouds for *3DOM Semantic Facade* and *ETH3D* benchmarks respectively.



(a) courtyard (b) terrace (c) pipes

Figure 4.12: **GT annotations for *ETH3D* sequences.** Outdoor and indoor scenarios are included. The outdoor scenarios follow the same nomenclature as the *3DOM Semantic Facade* benchmark; the for indoor sequence *pipes*, ad-hoc classes were defined: “wall” (blue), “floor” (orange), “door” (pink), “wardrobe” (grey) and “other” (red).

4.4.4 Class-specific reconstruction

Class-specific reconstruction implies the selective reconstruction of particular semantic classes that are of interest for each application scenario while excluding (i.e., filtering) undesired or poorly defined and fuzzy parts of the scene, e.g., the sky, obstacles and trees, improving, in this way, the overall quality of the reconstruction. The proposed strategy takes as input the corresponding segmentation maps and their link to the original images with a direct pixel-to-pixel mapping. During depth fusion, the semantic criterion is taken into account, selectively generating point clouds based on their semantic label (Figure 4.15). In the proposed approach, the available semantic information is used directly, different from explicitly using masks, an approach commonly followed in the literature [Murtiyoso et al., 2022].

4.5 Discussion

In this chapter, two fundamental research tasks of photogrammetry and computer vision are studied and interlinked; semantic segmentation and image-based 3D reconstruction. First, a new benchmark, *3DOM Semantic Facade* is introduced for facade segmentation on historic buildings. The generalization ability of this

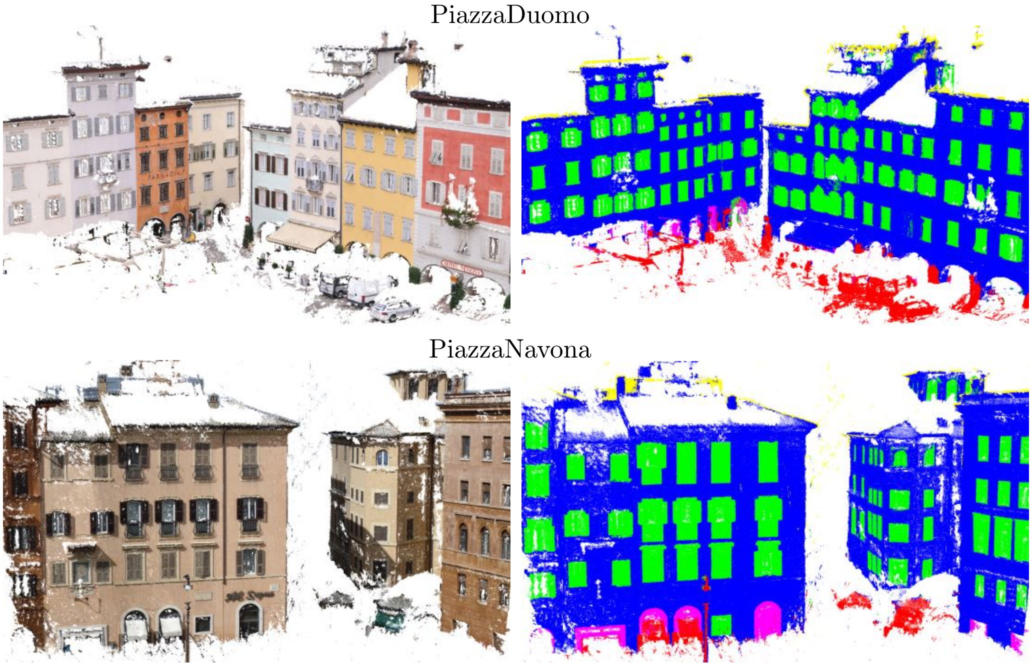


Figure 4.13: **Semantically enriched point clouds for facade segmentation applications (*3DOM Semantic Facade*)**. Left: standard dense cloud right: dense cloud with assigned labels to each pixel.

dataset is proven by implementing a standard deep learning pipeline for model training, achieving high performance scores in label inference. Moreover, a novel functionality is proposed for label transfer from 2D to 3D, generating semantically enhanced point clouds and enabling class-specific reconstruction, built on the open-source library OpenMVS [Cernea, 2020].

Deep learning algorithms for semantic segmentation on 2D images are considered mature enough and have achieved impressive results in recent years. However, although the great development, as most deep learning methods, they still are highly dependent on the training data and have, thus, limitations on domain generalization. Indeed, most approaches achieve high performance on public benchmarks, yet precise inference in diverse, real-world scenes is still an open challenge. Toward the expansion of the available data and the ease of generalization in real-world photogrammetric applications, in the context of this work a new benchmark dataset for facade segmentation, the *3DOM Semantic Facade*, has been presented. The benchmark includes 227 high-resolution images of historic building facades of diverse architectural styles. To prove the usability of the introduced benchmark, a straightforward and time- and memory-efficient training procedure is followed, based on EfficientNet [Tan and Le, 2019] and a U-Net architecture [Ronneberger et al., 2015]. The trained model is used to infer semantic labels on data from the same distribution (test set) as well as on completely unseen images from different distributions. The proposed method, as shown in Table 4.4, achieves high scores in terms of pixel accuracy, recall, completeness, F_1 score and IoU, and can be followed as an example procedure for similar problems also in other domains.

The obtained semantic masks are then used for label transfer in 3D yielding semantically segmented point clouds. A novel, ready-to-use MVS pipeline is introduced based on the open-source library OpenMVS [Cernea, 2020]. The developed system takes as input the RGB images along with their respective semantic maps and results dense 3D point clouds enhanced with semantic attributes. The proposed framework is scalable and domain independent; providing a priory calculated semantic labels for any input set of images $\mathcal{I}$, the labels can be transfer from the 2D space to 3D and enabling semantic segmentation in the 3D space even for large scenes, with minimal additional computational cost. This method can be particularly beneficial for cases where semantically enhanced 3D point clouds are needed, but direct segmentation in the 3D space is prohibiting due to lack of computational resources or the inability to gather a big amount of 3D GT data for training.

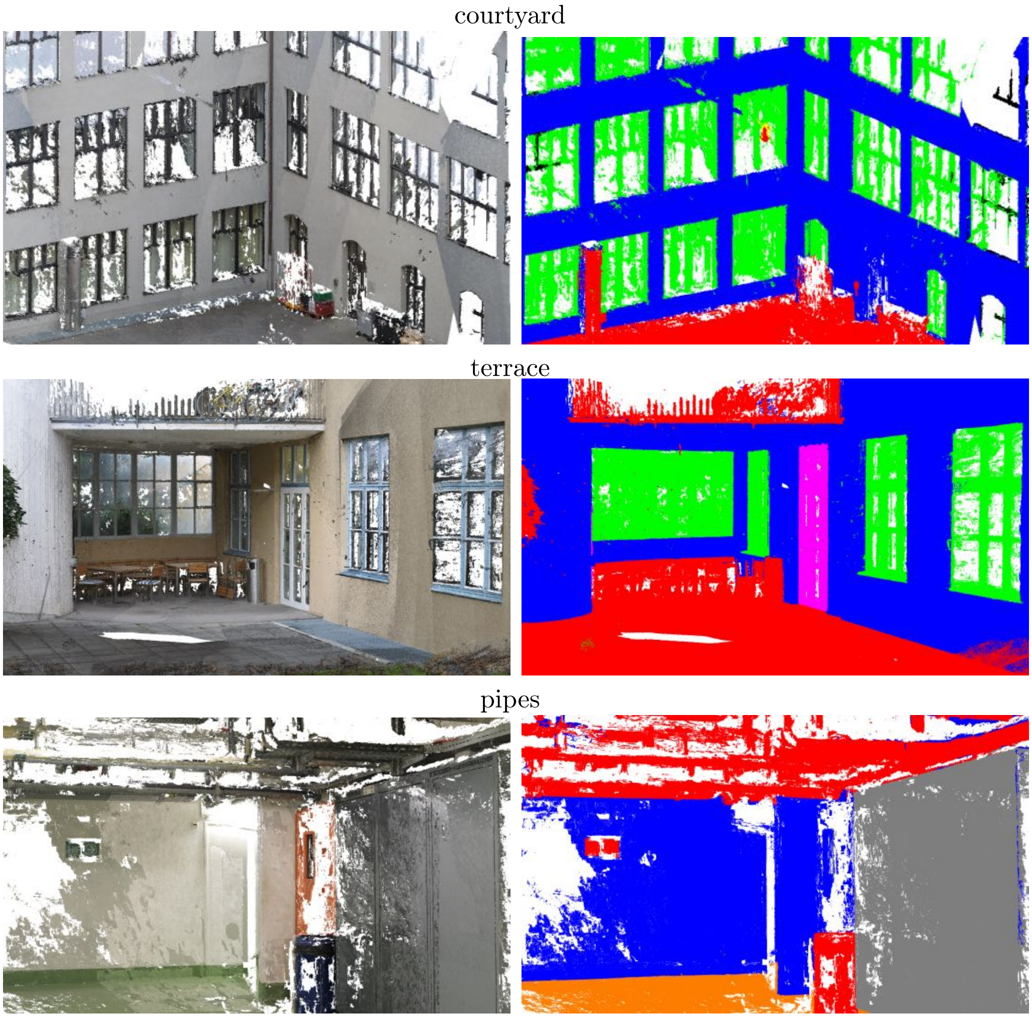


Figure 4.14: **Semantically enriched point clouds for the *ETH3D* benchmark sequences (indoor and outdoor)**. Left: standard dense cloud. Right: dense cloud with assigned semantic labels to each pixel.

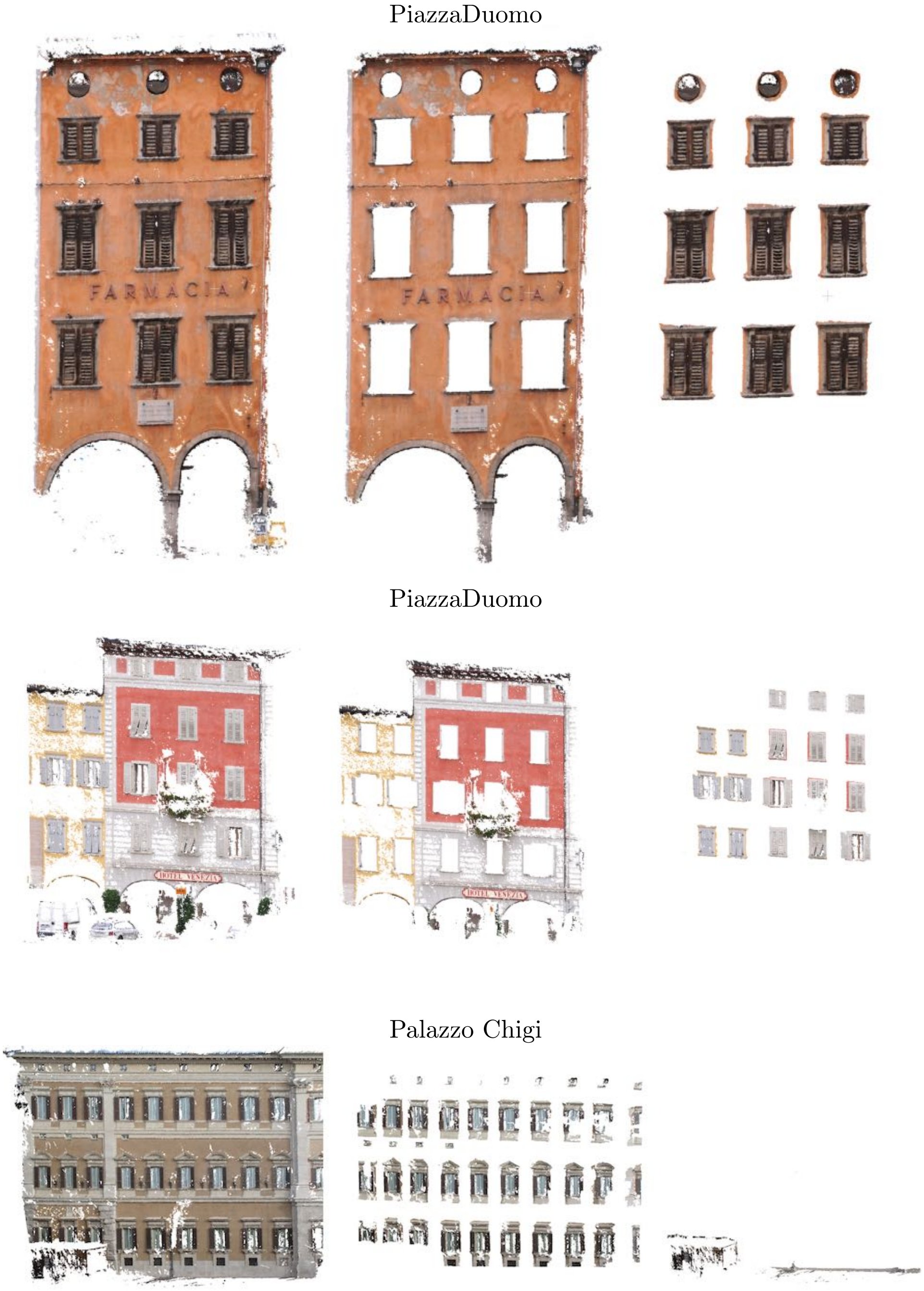


Figure 4.15: **Selective 3D reconstruction based on the semantic label on the 2D images.** Upper and middle, from right to left: all classes, wall, windows. Lower, from right to left: all classes, windows, obstacles.

Chapter 5

Semantic cues in depth estimation

This chapter focuses on the MVS reconstruction part and proposes a solution to confront the often occurring matching ambiguities problem in man-made indoor and outdoor scenarios. The main motivation is the potential of the advanced scene priors, and in particular semantic reasoning, in supporting the depth estimation process when pure geometric and radiometric information is not enough. Based on this insight, a new methodology is introduced for leveraging explicit semantic cues into MVS under a PatchMatch scenario, as originally introduced in [Stathopoulou et al., 2021b].

Semantic segmentation has become increasingly popular in recent years, and algorithms have demonstrated great potential in generating semantic maps on images, especially in well-studied contexts such as airborne mapping, street scenes, and indoor spaces for navigation and mapping. Hence, semantic masks can be generated relatively easily for similar real-world tasks, particularly when few and well-represented classes are defined. In the presented approach, a priori generated semantic segmentation masks, obtained with methods similar to those presented in Chapter 4, are used to support the depth estimation process and improve, thus, 3D point cloud completeness and overall quality, particularly in challenging areas where commonly matching ambiguities exist. This is achieved by exploiting constraints derived from the semantically cues to imply additional class-specific shape priors during matching cost computation in PatchMatch MVS (Figure 5.1). The idea is based on the fact that semantics can often successfully indicate textureless and other non-Lambertian areas derived by the class label of the scene area (e.g., “wall”); in such regions, frequently, depth miscalculations occur due to matching ambiguities. To confront this limitation, geometric constraints can be

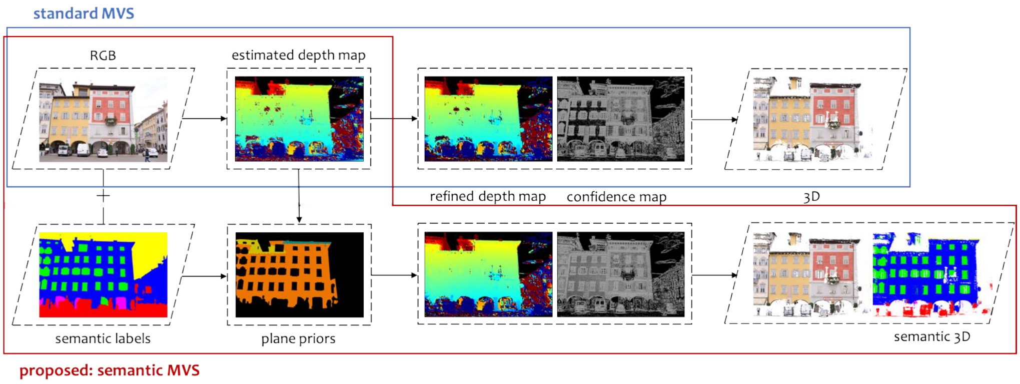


Figure 5.1: **Overview of the proposed semantic PatchMatch MVS pipeline.** A priory obtained semantic cues for the scene are integrated into the MVS reconstruction to promote reliable depth estimates in challenging areas and generate more complete point clouds with respect to the standard method. A semantically enhanced point cloud can be generated as described in Chapter 4.

implied to yield more reliable depth values and, therefore, more complete in the final 3D point cloud while object boundaries and depth details are preserved.

Standard PatchMatch approaches, as most local methods, make, however, a priori regularization assumptions to ensure smoothness within the local window (see also Chapters 2 and 3), yet the matching ambiguities in textureless and other non-Lambertian surfaces severely decrease the quality of the reconstructed point cloud, resulting in outliers and information gaps. Hence, additional geometric constraints have been implied directly in the recent literature to support depth estimation, e.g., local surface planarity [Romanoni and Matteucci, 2019; Xu and Tao, 2019, 2020b]. In the same line of thought, the method proposed in this chapter formulates geometric constraints based on semantic priors and benefits from the class-specific geometric properties. RANSAC 3D planes are detected for all dominant surfaces presented in the scene, e.g., under “wall”, “floor”, etc. labels which are assumed to be planar. Then a new, adaptive cost function is introduced to integrate depth prior hypotheses and texture information of pixel neighborhood to the standard photometric cost. Regarding its implementation, the developed algorithm builds upon the open-source MVS library OpenMVS [Cernea, 2020] and extends its functionality by adding the *semantic PatchMatch* module. Finally, the effectiveness of the framework is evaluated over selected scenes of the *ETH3D* benchmark dataset [Schöps et al., 2017] and other custom sequences.

5.1 Semantic reasoning and 3D reconstruction

Algorithms for semantic segmentation on images have become increasingly popular in recent years, achieving robust results in various fields of applications as discussed in Chapter 4. Meanwhile, it has been observed that advanced scene priors can potentially help to overcome certain deficiencies in image-based 3D reconstruction. As a matter of fact, in recent years, several works couple 3D reconstruction and semantics; they either refer to joint segmentation and reconstruction optimization for multi-view [Ladickỳ et al., 2012; Schneider et al., 2016] and monocular setups using conditional random fields (CRFs) [Kundu et al., 2014] or to the use of depth maps to support 2D segmentation [Zhang et al., 2010]. In the volumetric representation domain, Häne et al. [2013, 2016] proposed a rigorous solution to jointly confront volumetric 3D with semantics in multi-view scenarios with variational optimization, while Savinov et al. [2016] applied a ray potential computation method in a semantic context. Blaha et al. [2016, 2017], inspired by [Häne et al., 2013], enabled semantic segmentation and volumetric reconstruction jointly for surface refinement of large-scale scenes, updating shapes and labels simultaneously. Similarly, Romanoni et al. [2017] implemented joint optimization of mesh refinement and semantic segmentation, also combining the photometric consistency. Cherabier et al. [2018] learned semantic priors for TSDF volumetric reconstruction and joint optimization, while Yingze Bao et al. [2013] and Ulusoy et al. [2017] used learned data-driven geometric shape priors for volumetric reconstruction without aiming for a semantically enhanced output.

Closer to the present work, regarding the optimization of the depth estimates, research has been shifted towards introducing priors in MVS. Assumptions may vary among the studies, yet a great part of them implicitly impose geometric constraints along with semantics. Man-made objects usually conform to clearly defined geometric shapes and belong to certain semantic classes. Introduced as "object knowledge information constraints", common semantic labels indicate the sharing of geometric properties along with local smoothness and can therefore facilitate 3D reconstruction. Indeed, some studies adopt the hypothesis that scene objects are piecewise planar [Furukawa et al., 2009; Gallup et al., 2010] or that all pixels belonging to the same semantic label must necessarily share also the same disparity value to guide depth computation for challenging, poorly textured surfaces [Chen et al., 2014]. Similarly, other works use a group representation of pixels with common properties, the so-called semantic stixels [Schneider et al., 2016] or 2.5D shape samples known as displets [Guney and Geiger, 2015] to boost efficiency in depth calculation.

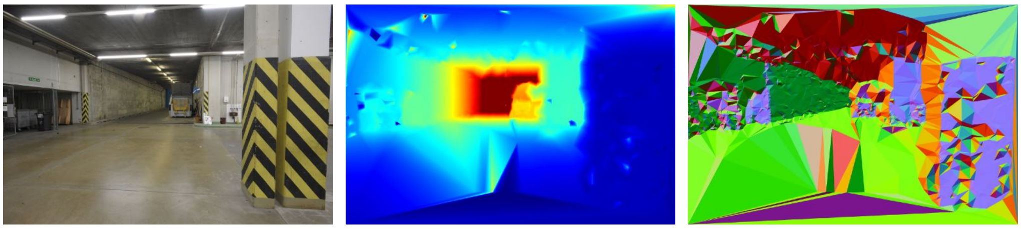
The problem of weakly-supported, textureless areas under PatchMatch MVS scenarios has been recently undertaken in the literature towards large-scale applications with a high overlapping percentage. TAPA-MVS [Romanoni and Matteucci, 2019] assumed piecewise planarity on image superpixels for joint PatchMatch and view selection. Kuhn et al. [2019] extended this framework and achieved depth completion as a post-processing step using hierarchical superpixel clustering. On the contrary, the method introduced in this chapter considers depth estimation optimization as an integrated problem, while plane hypotheses are detected in the 3D space. Other recent works handle textureless areas with multi-scale geometric consistency guidance [Xu and Tao, 2019] or consider direct planar priors based on the sparse reconstruction [Xu and Tao, 2020b]. Recent deep learning methods tackle this problem with coarse-to-fine schemes, often requiring additional detail restorer modules [Wang et al., 2020b]. A more detailed overview of these methods particularly designed to solve the matching ambiguities problem based solely on structure cues is presented in Chapter 6.

5.2 Proposed methodology: semantic PatchMatch

The proposed approach for semantically-guided PatchMatch MVS links the input images with their semantic mask equivalent using a direct pixel-to-pixel mapping (Figure 5.1). Following a PatchMatch MVS approach based on [Shen, 2013], it extends the initial idea presented in Section 3.3 by imposing class-specific geometric constraints during the depth map computation step. These geometric constraints are used in the matching cost computation, supporting the propagation of reliable depth estimates in textureless areas while preserving the fine details on the resulting point cloud. As a matter of fact, semantic information can generally imply geometric constraints, and pixels belonging to the same class often have common geometric properties. Other recent works assume local planarity in the form of triangles [Xu and Tao, 2020b] or superpixels [Romanoni and Matteucci, 2019], yet in this chapter, semantic info is explicitly used to derive geometric constraints by assuming planarity for larger, dominant planar areas of the scene. For instance, semantically segmented images in urban scene scenarios can provide structure hypotheses for building facades. Planar walls are assumed to be more likely textureless areas, commonly made of flat surfaces of the homogeneous color. However, the method is potentially extendable to other shape priors as well, such as cylinders, spheres etc. A priori labels are generated as described in Chapter 4 and in [Stathopoulou and Remondino, 2019a,b].

Typically, the input of each MVS process is a sequence $\mathcal{I} = \{I_0, \ldots, I_m\}$ of RGB images of known camera poses $(\mathcal{R}, t)$ and the sparse point cloud, previously calculated with standard SfM techniques. In the extended *semantic PatchMatch* approach, priorly calculated semantic maps for each input image $\mathcal{L} = \{L_0, \ldots, L_m\}$ are also required to leverage the semantic cues into the MVS depth calculation (Figure 5.1). All images share the same intrinsics given by the calibration matrix $\mathbf{K}$.

View selection. As explained in Section 3.3, the first step of PatchMatch MVS is stereo pair selection; in the proposed method, a heuristic approach is followed



(a) RGB image (b) depth map (c) normal map

Figure 5.2: **Depth and normal map initialization example.** RGB image, initial depth and normal maps generated by interpolating the sparse SfM points.

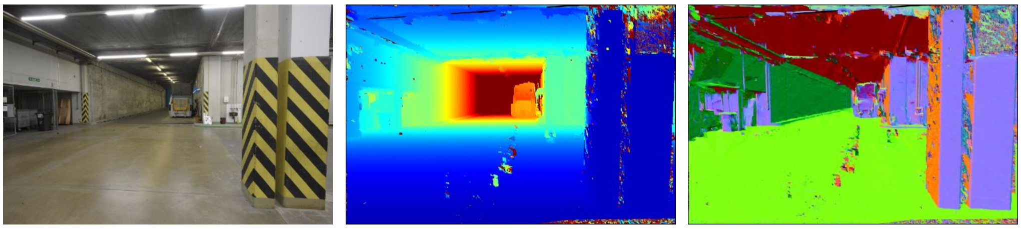
as implemented in the OpenMVS library [Cernea, 2020]. The best neighboring views are selected and sorted for each reference image based on visibility criteria, i.e., baseline, intersection angle, scale, similarly to [Goesele et al., 2007]. Then, the scene graph is generated; the views are vertices connected with their neighbors with edges, and the best pairs are selected globally.

Depth map initialization. After the selection of the best neighboring views for each reference image, depth map initialization follows. Here, instead of completely random values, initial depth and normal estimates are assigned to each pixel using the sparse cloud derived in the SfM process and interpolating the depth for all pixels. The resulting depth and normal maps are rough but provide a generally good initial estimate for the PatchMatch iterations (Figure 5.2).

Depth map estimation. The coarse values from the initialization step are then refined using PatchMatch spatial propagation iteratively. First, the matching cost for each pixel is calculated based solely on the photo-consistency metric, in this case, zero-mean cross-correlation (ZNCC). The window size is of utmost importance in such metrics; typically, a window radius of 5 or 7 pixels is experimentally proven to be efficient for high-resolution images. Subsequently, spatial propagation and refinement follow.

Spatial Propagation. The current estimates for depth and normal are compared to those of the neighboring pixels and, if the latter have lower photometric cost, are considered more reliable and replace the current estimates; otherwise, the current estimate is kept as such. Since a sequential propagation scheme is followed, neighbors are considered the adjacent pixels top, bottom, right, and left pixels and the propagation direction is top left to bottom right for the odd iterations and bottom right to top left for the even ones.

Random refinement. At the end of every iteration, the estimates are compared with random values in a pre-defined range (as calculated in the initialization step), and the ones with the best scores are kept to further refine the results and



(a) RGB image (b) depth map (c) normal map

Figure 5.3: **Depth and normal map computation example.** RGB image, depth and normal maps after the first PatchMatch iteration.

eliminate potential outliers. Figure 5.3 shows the respective depth and normal maps for an image after the first PatchMatch iteration.

This process will most probably converge after 2-3 iterations to reliable depth estimates in near-Lambertian, rich-texture areas, especially for high-resolution images [Shen, 2013]. However, in weakly-supported, textureless regions, the photometric cost alone is sensitive to local minima, resulting in matching ambiguities and, consequently, in wrongly reconstructed 3D points. For instance, in Figure 5.3 it can be observed that even for only one PatchMatch iteration, rich texture areas generate mostly reliable depth and normal estimates. Still, areas with particularly weak texture (white part of the column on the right) or reflective regions (floor) are dominated by noise.

5.2.1 Semantically-guided 3D plane prior hypotheses

Standard PatchMatch iterations, as explained above, estimate the depth values d and assign normal vectors $\mathbf{n}$ to each pixel p, generating a depth $\mathcal{D}_{tmp}$ and a normal map $\mathcal{N}_{tmp}$ for every view. These maps contain some outliers and noise that can be partially refined in the following steps or filtering and fusion. In the proposed *semantic PatchMatch* method, these depth estimates, although relatively noisy, are used to generate depth d_{prior} and normal $\mathbf{n}_{prior}$ hypotheses in the 3D space for a pixel that belongs in a dominant plane of the scene and the respective $\mathcal{D}_{prior}$ and N_{prior} maps for each view. Planar surfaces are adopted here, yet the approach is extendable also to other primitives; planes are commonly encountered in the majority of man-made scenes, indoor or outdoor. For instance, in indoor scenarios, commonly, planar walls, floors or ceilings are present. Outdoor scenarios, either close-range or airborne, also typically contain several man-made structures, i.e., buildings with dominant planar features such as facades and roofs.

Point cloud filtering. After the computation of the depth maps, depth estimates are calculated in the 3D space using the camera projection matrix $\mathbf{P}$ (Equation 2.4) and generating intermediate point clouds for each view. In the proposed approach, instead of projecting all scene pixels, a "semantic label check" is performed at the beginning of the prior hypotheses generation algorithm. That is, the semantic segmentation maps are used to constrain the area and project in 3D only the subset of pixels under specific labels that are more likely to include planes (i.e., "wall", "floor"), adjusted to the needs of each application (Figure 5.4b,c). As a first rough filtering step, for the semantic labels of interest, only depth values of high confidence are projected in the 3D space. Confidence, in this context, directly corresponds to the photometric cost value; low photometric cost implies high confidence. Naturally, the most reliable points are the ones around the crease edges. More details on confidence measures for stereo matching can be found in [Hu and Mordohai, 2012].

Although the 3D points of the most reliable depth estimates are reconstructed, the estimated point clouds are still evidently noisy. To confront this problem, a two-step filtering strategy is applied before plane detection to eliminate outliers and keep the more reliable points that can potentially belong to planar surfaces.

First, the covariance features derived from the covariance matrix of the 3D point coordinates in a given local neighborhood are exploited, as they are proven robust to directly classify points with certain geometric characteristics. For instance, linear parts of the scene will generally have high linearity values, and planar regions will include points with high planarity values. These covariance features are expressed as combinations of the eigenvalues and eigenvectors of the covariance matrix for each point and commonly include *planarity, surface variation, sphericity, omnivariance, anisotropy,* and *linearity* as formulated in [Hackel et al., 2016]. Using the eigenvalues $\lambda_1, \lambda_2, \lambda_3$, in a local neighborhood of $k = 10$ points, point planarity ρ is calculated as:

$$\rho = \frac{\lambda_2 - \lambda_3}{\lambda_1}. \tag{5.1}$$

The points with low planarity values ($\rho < \mathit{threshold}$) are eliminated since they most probably do not belong to representative dominant planes of the scene and are considered outliers. Planarity filtering has proven particularly robust and has increased the robustness of the shape extraction method in the experiments.

Finally, an additional filtering method is applied based on the average point spacing $\bar{s}$ in a close 3D neighborhood (e.g., $k = 10$ or 24 points). In this way, remaining sparse or isolated points are removed from the final point cloud. An overview of the prior generation algorithm is presented in Algorithm 1.

RANSAC plane detection. Subsequently, 3D planes in every view are detected using a RANSAC-based method. RANSAC [Fischler and Bolles, 1981] is a popular, model-fitting method that starts from a random minimal paradigm set, i.e., the smallest set of observations required for a solution. Being robust,

Algorithm 1 Point cloud filtering and RANSAC shape detection

Input: depth and normal map $\mathcal{D}_{tmp}$,$\mathcal{N}_{tmp}$, semantic map L
Output: set of m RANSAC shapes $\mathcal{S} = \{S_1, \ldots, S_m\}$

```
for all pixel p in $\mathcal{D}_{tmp}$ do
    if semantic label check OK AND conf $\geq$ threshold then
        project point in 3D
    else
        skip pixel
    end if
end for
for all 3D points do
    calculate planarity $\rho$ for $k$ neighbors
    if planarity $\rho \geq$ threshold AND point average spacing check OK then
        keep point
    else if  then
        discard point
    end if
end for
find best-fitting RANSAC shapes $\mathcal{S} = \{S_1, \ldots, S_m\}$ for the filtered point set
following [Schnabel et al., 2007] given $\epsilon$, $c_\epsilon$, $\mathbf{n}_\epsilon$, $M_{min}$
return  $\mathcal{S} = \{S_1, \ldots, S_m\}$
```

it is applied in a wide range of applications and is commonly used as a shape extraction method for primitives, e.g., planes, cylinders, spheres, cones, and tori. Hence, starting from a randomly selected point set, it tests the remaining points against the model to determine whether the model represents efficiently the set of points. Being an iterative method, it converges to the model that approximates the best the set of points and continues with the rest of the data.

In the proposed approach, the remaining 3D points after the filtering procedures described above are used as input to the RANSAC algorithm. In particular, the Efficient RANSAC solution [Schnabel et al., 2007] as enfolded in CGAL library [The CGAL Project, 2021] is used for the experiments. RANSAC parameters are adjusted accordingly based on the average spacing of every point cloud so that only significantly large planes are considered valid and avoid, thus, over-segmentation (Figure 5.4d). In more detail, the maximum tolerance for Euclidean distance between a point and a shape ϵ and the connectivity measure c_ϵ are calculated as a function of the average spacing $\bar{s}$ for scalability, while the normal deviation $\mathbf{n}_\epsilon$ is set to a constant value. The minimum number of points M_{min} needed to form a shape is defined based on the total number of 3D points for that view. In such manner, the most dominant planes of the scene are detected.

Prior depth and normal map generation. For every detected plane hypothesis the weighted 3D centroid $C_{R_{plane}}$ and the normal $\mathbf{n}_{R_{plane}}$ is calculated. The boundaries of each plane in 3D are defined using the minimum bounding rectangle

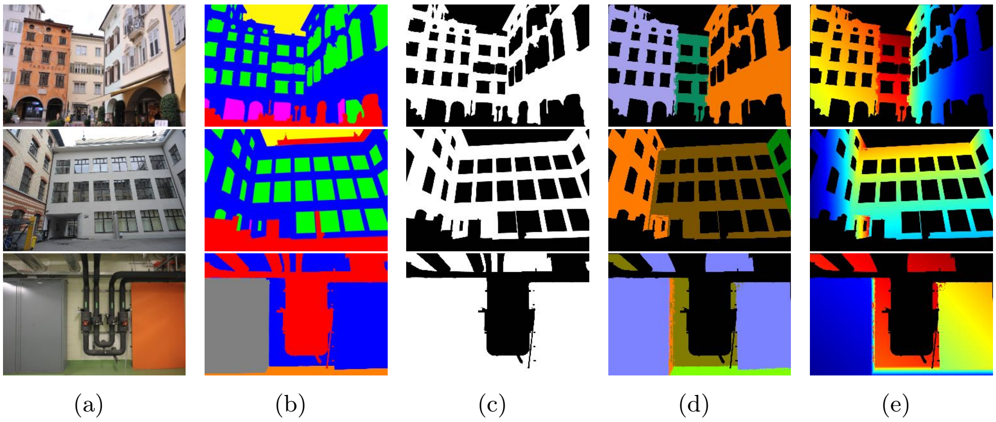


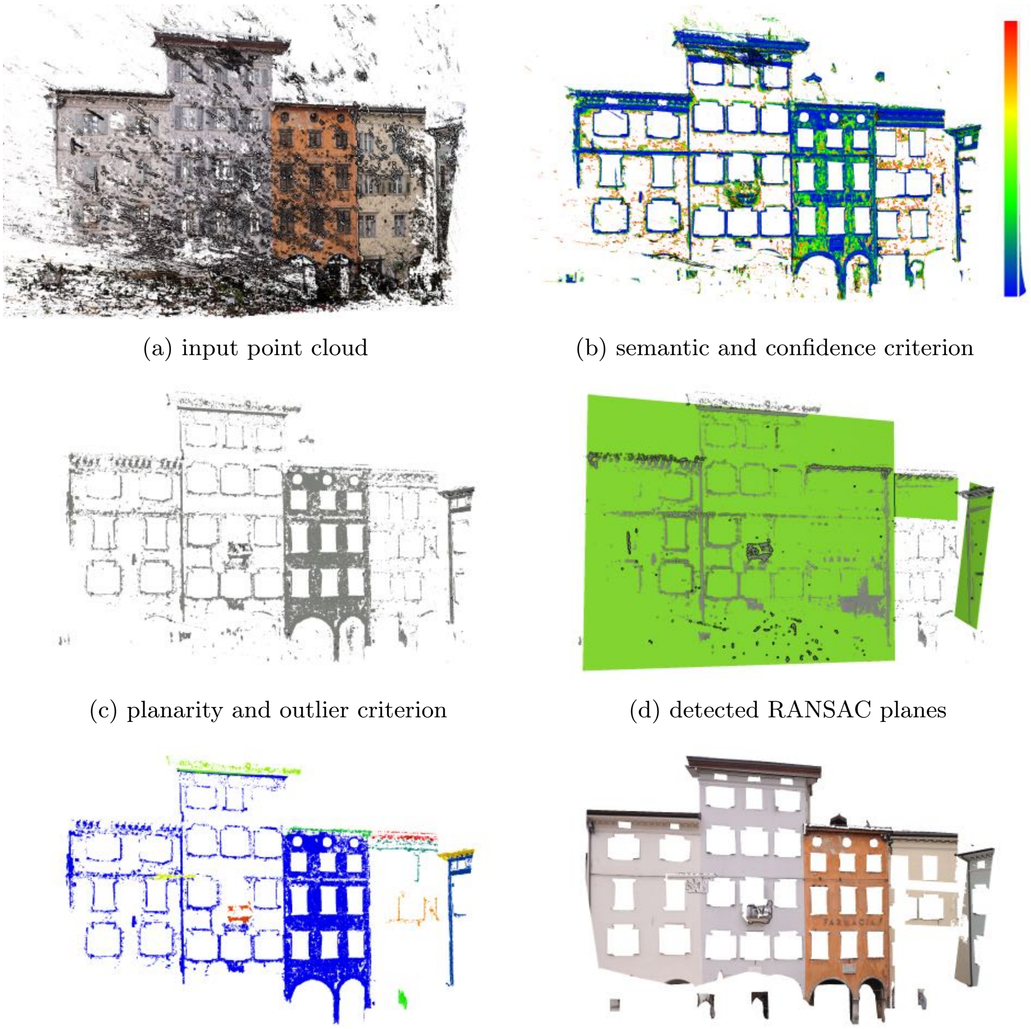
Figure 5.4: **Plane prior estimation with semantic guidance.** Input image (5.4a); respective labels (5.4b) for semantic classes: sky (yellow), wall (blue), window (green), door (purple), other (red); binary mask for planar classes (5.4c); the estimated normal prior map for the RANSAC planes detected in 3D (5.4d) and their respective depth priors in color scale (5.4e) with blue being the closest and red the farthest.

(extent) of each point set belonging to the same RANSAC plane based on the max and min coordinates of the set. Each image pixel that passes the semantic label check (i.e., belongs to a class with a potentially planar shape) is projected in 3D and assigned to the closest 3D plane id, e.g. the plane with the minimum absolute distance. Pixels that belong to non-planar classes are not considered. Then, for each eligible pixel, its depth prior d_{prior} is calculated using ray-plane intersection; starting from the camera origin, a ray is casted through every pixel of the view to the 3D space, until it intersects with the assigned 3D plane prior. The $\mathbf{n}_{prior}$ of each pixel is the normal of its assigned 3D plane hypothesis $\mathbf{n}_{R_{plane}}$.

Eventually, prior hypotheses are generated only for the semantic classes that are considered locally planar, e.g., facade walls. Both d_{prior} and $\mathbf{n}_{prior}$ of each planar region pixel are stored (Figure 5.4d, 5.4e) in corresponding maps $\mathcal{D}_{prior}$ and $\mathcal{N}_{prior}$; these hypotheses are further used to adaptively guide the cost computation in the next step. All the aforementioned steps for point cloud filtering and planar prior generation are visualised in Figure 5.5.

5.2.2 Adaptive cost calculation

PatchMatch highly relies on the photometric consistency measure to correctly select the value from the random estimates that represents the best hypothesis. The proposed method starts from the mostly good depth and normal estimates (d_{tmp}, $\mathbf{n}_{tmp}$ calculated by the first iterations of the standard PatchMatch (commonly set to $2-4$ iterations) and refines the results using the plane prior hypotheses during spatial propagation while doing a couple of additional PatchMatch iterations. The



(a) input point cloud

(b) semantic and confidence criterion

(c) planarity and outlier criterion

(d) detected RANSAC planes

(e) assigned points to planes

(f) plane priors

Figure 5.5: **The proposed semantically-guided prior hypotheses generation strategy.** For each view, the intermediate point clouds from PatchMatch iterations are used (5.5a). Only the points that belong to potentially planar classes (here class "wall") are considered. Points are first filtered based on their confidence value; points of high confidence (low photometric cost) are considered as more reliable points (5.5b). Points are color-coded, with blue being the points of low, i.e. robust photometric cost (high confidence) and red being the less reliable points. Planarity and outlier filtering based on average spacing follow (5.5c). Following Efficient RANSAC, the most dominant scene planes are detected (5.5d); here, only three planes are shown for better visualization. The corresponding points of each plane are color-coded based on the assigned plane id (5.5e). Finally, the plane priors are generated with ray-plane intersection and guided by the semantic label (5.5f).

matching cost of the PatchMatch estimates of the baseline approach relies simply on the photometric consistency measure, in this case, the zero-mean normalized cross-correlation ZNCC as defined in Equation 2.18. ZNCC is used as a common robust metric that performs well in cases of Gaussian noise, perspective and intensity changes. In particular, the photometric cost c_{ph} for a pixel p is defined:

$$c_{ph} = 1 - ZNCC. \tag{5.2}$$

For high $ZNCC$ values, the cost will tend to zero since $-1 \leq ZNCC \leq 1$. On the contrary, the higher the cost, the lower the photometric consistency, resulting in erroneous and ambiguous depth estimates.

Even though the simple photometric cost produces generally accurate results in near-Lambertian surfaces, in textureless and highly reflective areas, it often causes matching ambiguities and promotes, thus, the propagation of wrong depth estimates. One possible solution for this would be to directly substitute the estimated depth and normal values $[d_{tmp}, \mathbf{n}_{tmp}]$ in problematic areas with the depth and normal values of the plane hypotheses $[d_{prior}, \mathbf{n}_{prior}]$ calculated in the previous step. However, this would again result in unreliable outcomes, as it would completely disregard the photometric matching cost in planar areas, forcing planarity and smoothing out fine details. Instead, it is proposed to introduce a novel, adaptive cost function that leverages the plane prior hypotheses, if previously generated for this area, with the standard photometric cost. It introduces two additional coefficients and integrates them in the cost function as additional terms in a combined formulation. The first one is a pixel-wise coefficient v^2 based on the local textureness, defined as the variance of the intensity values $I(\mathrm{q})$ in a $n \times n$ pixel neighborhood:

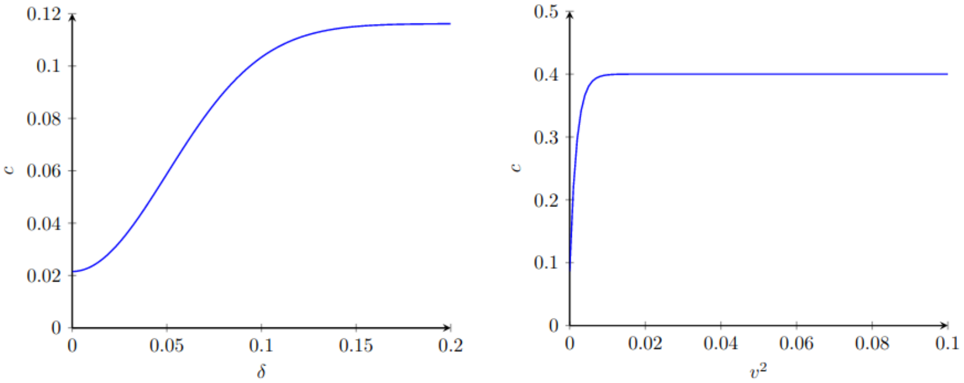
$$v^2 = \frac{\sum\limits_{\mathrm{q} \in [n \times n]} (I(\mathrm{q}) - \mu)^2}{n}, \tag{5.3}$$

where μ the mean local intensity. In the presence of evenly colored surfaces, the local variance will have very small values. The textureness coefficient will have the following formulation based on a Gaussian kernel as in Stathopoulou et al. [2021b]:

$$t = e^{\frac{-v^2}{2\sigma_t^2}}, \tag{5.4}$$

where σ_t a constant. Likewise, for a smooth integration of the depth hypotheses derived by the semantically-guided plane priors a second coefficient is defined:

$$s = e^{\frac{-\delta^2}{2\sigma_s^2}}, \tag{5.5}$$



(a) (b)

Figure 5.6: **Cost function behavior.** The combined cost c given a standard photometric cost $c_{ph} = 0.4$ (relatively low confidence) with respect to: (a) the depth difference δ for $v^2 = 0.001$ (textureless area) and $\{\sigma_t = 0.03, \sigma_s = 0.05, \gamma = 0.1\}$; (b) the variance v^2 for $\delta = 0.01$ and $\{\sigma_t = 0.03, \sigma_s = 0.05, \gamma = 0.1\}$.

where σ_s is a constant fixed experimentally. δ is given as the percentage of difference between the depth prior hypothesis d_{prior} and the current PatchMatch estimate d_{tmp}:

$$\delta = \frac{|d_{prior} - d_{tmp}|}{d_{prior}}. \tag{5.6}$$

The cost function should seamlessly leverage the terms, i.e., integrate the prior depth hypotheses with the ZNCC metric and result in more reliable depth estimates that will subsequently be spread across the image and generate more complete 3D point clouds in cases where matching ambiguities occur. Thus, s (Equation 5.5) and the textureness coefficient t (Equation 5.4) with the original photometric matching cost c_{ph} (Equation 5.2) are combined to give the total cost c:

$$c = c_{ph}(1 - t) + \gamma(1 - s)t, \tag{5.7}$$

where γ is a weight factor.

Large, textureless regions are very likely to have been assigned with high photometric cost values during the standard PatchMatch iterations due to the presence of matching ambiguities. In such regions, the variance v^2 of the intensity of the pixel neighborhood will probably be close to zero since color similarity is maximized. In the proposed approach, textureless regions will be most probably assigned with plane prior hypotheses. For these planar regions with high scores, the new cost function will prioritize the prior hypotheses $[d_{prior}, \mathbf{n}_{prior}]$. On the contrary, for the regions where the original photometric cost is calculated reliably enough (i.e.

low photometric cost), the standard PatchMatch depth and normal estimates $[d_{tmp}, \mathbf{n}_{tmp}]$ will be trusted more. In such a way, plane priors are alleviated with the photometric consistency, and erroneous estimates tend to vanish, resulting in more reliable depth maps. In other words, when the surface deviates from the plane but has a significant texture variance, the photometric cost is trusted more. Possible outliers will be filtered out from PatchMatch because of no coherence with the neighborhood, and in the worst-case scenario, it will degenerate to the standard case. Example behavior of the cost function with respect to δ and v^2 variations are shown in Figure 5.6.

Recent methods have proposed analogous priors in varying formulations; for instance, Romanoni and Matteucci [2019] proposed another textureness metric as a weight for the hypotheses during the cost calculation under a probabilistic framework. Xu and Tao [2020b] used planar prior constraints by triangulating sparse points without explicitly considering the local textureness. The proposed approach leverages planar priors, textureness information and photometric consistency in a simple, yet efficient way.

The cost function affects directly not only the depth maps but also the normal and confidence ones as shown also in Figures 5.1, 5.3. Noisy regions of the normal maps are also smoothed, and information gaps are filled in since estimated normals are leveraged with the normal prior information. The same holds for confidence maps that reflect the depth estimate reliability of every pixel (i.e., the cost). In the performed experiments, it was demonstrated that only two additional PatchMatch iterations with the proposed adaptive cost function were enough to significantly improve the depth and normal map quality, as well as the confidence of every pixel and, finally, the generated dense 3D point cloud. Indeed, the proposed solution converges relatively fast.

5.3 Experiments and results

5.3.1 Datasets

Benchmark datasets of high-resolution images for accurate 3D reconstruction using MVS techniques providing GT pixel-level semantic masks are not publicly available. Hence, the proposed method cannot be directly evaluated with respect to common benchmarks such as *ETH3D* [Schöps et al., 2017] or *Tanks and Temples* [Knapitsch et al., 2017] as other state-of-the-art MVS algorithms typically do [Xu and Tao, 2019; Romanoni and Matteucci, 2019; Kuhn et al., 2019], due to the fact that these datasets lack accompanied labeled data. However, to be in line and comparable with the other state-of-the-art techniques, three representative *ETH3D* datasets are used for which the GT labels were manually annotated (Chapter 4). Along with the *ETH3D* sequences, the proposed algorithm is evaluated on two custom datasets obtained from the *3DOM Semantic Facade* dataset Stathopoulou and Remondino [2019a,b] and the airborne benchmark dataset *UDD5* [Chen et al., 2018a].

ETH3D. Three sequences from the high-resolution (6048 × 4032) datasets for which GT 3D data is available were chosen for these experiments: *ETH3D-courtyard* (38 images) and *ETH3D-terrace* (23 images) as typical outdoor scenarios and *ETH3D-pipes* (13 images) for the indoor one. Manual labeling is performed for the building facades in order to extract planar regions, (Figure 5.4). Class nomenclature is the same with the one followed in [Stathopoulou and Remondino, 2019a,b] for the *ETH3D-courtyard* and *ETH3D-terrace* datasets while for the interior scenario *ETH3D-pipes* the semantic labels “wall”, “floor”, “door”, “closet” and “other” are introduced. In this specific dataset, plane estimation is performed within the classes “wall”, “floor” and “closet”, whereas for *ETH3D-courtyard* and *ETH3D-terrace* only “wall” is considered as a class with potentially planar instances. To be comparable with the publicly available results of the other state of the art methods tested on the benchmark, images are resampled to 3200 pixels as in Xu and Tao [2019]. This is a common practice in order to reduce the computational cost and handle large-scale datasets, and although dense cloud density is, as expected, partially affected it is considered to be enough for such datasets. For these datasets qualitative comparisons for depth maps (Figure 5.7) and confidence maps (Figure 5.8) are performed. The resulting 3D dense clouds are evaluated qualitatively (Figure 5.9) and quantitatively (Table 5.1). Results derived with the proposed method are compared against the baseline OpenMVS [Cernea, 2020], as well as COLMAP [Schönberger et al., 2016] and four recent methods that use geometric prior-assisted PatchMatch: TAPA-MVS [Romanoni et al., 2017], ACMM [Xu and Tao, 2019], ACMP [Xu and Tao, 2020b] and PCF-MVS [Kuhn et al., 2019].

Custom datasets. Two more scenarios are used for evaluation, namely *Piazza Duomo* (12 high resolution images, 6048 × 4032 px) and *Piazza Navona* (5 high resolution images, 4000 × 3000 px). Again, images are resampled to 3200 pixels. Ground truth semantic labels are available from the previous work presented in Chapter 4 and [Stathopoulou and Remondino, 2019a] and “wall” is considered a class with potential planar areas. For the dataset *Piazza Duomo*, a ground truth 3D point cloud from terrestrial laser scanning is also available. In this scenario, results are compared with COLMAP [Schönberger et al., 2016], TAPA-MVS [Romanoni et al., 2017], ACMM [Xu and Tao, 2019] and ACMP [Xu and Tao, 2020b] (Table 5.2, Figure 5.10)[1]. Qualitative comparison for the dense and confidence maps is also presented (Figure 5.11).

[1] Process run by Andrea Romanoni (TAPA-MVS) and Qingshan Xu (ACMM/ACMP) in November 2021, when the respective implementations were not open-source.

UDD5. *UrbanDrone Dataset (UDD5)* [Chen et al., 2018a] is a large scale benchmark dataset for segmentation of airborne urban scenarios. The training data for which the images labels are given as ground truth are used. *UDD5* labels are defined as “vegetation”, “building”, “vehicle”, “road” and “other”. Plane priors are estimated for the class “building” (which includes roofs and facades). Since 3D ground truth data are not available for this dataset, it is used only for qualitative evaluation (Figure 5.11). For computational efficiency, depth maps are generated in 1/4 of the original resolution, i.e., 2000 × 1500 pixels.

5.3.2 Implementation details

The evaluation tests are performed on an AMD Ryzen 2950X CPU running on 3.5GHz. For a fair comparison with the baseline PatchMatch MVS approach as implemented in OpenMVS [Cernea, 2020] following [Shen, 2013], the same parameter configuration is kept. The combined total cost is computed using a pixel window size $N = 7$ and $\{\gamma = 0.1$, $\sigma_s = 0.05$ and $\sigma_t = 0.03\}$ that experimentally were proven to be the best trade-off values across the datasets. Following the baseline implementation of OpenMVS [Cernea, 2020], $N_{source} = 8$ source images for each reference image are used for the *ETH3D* dataset. The confidence measure as implemented in the baseline method [Cernea, 2020] is actually an inverse metric that directly corresponds to the photometric cost and is defined in the interval $[0, 2]$ with 0 being the lower cost (totally correlated patterns). Hence, low values on this measure actually correspond to robust photometric costs (high confidence); threshold is set to $conf > 0.18$. During the planarity filtering, points with $\rho \leq 0.3$ (loose threshold) are excluded. RANSAC parameters $\{\epsilon, c_\epsilon\}$ are, as aforementioned, calculated adaptively based on the average spacing $\bar{s}$, in this case set to $\{\epsilon = 5\bar{s}, c_\epsilon = 8\bar{s}\}$, while the normal deviation is set to $\mathbf{n}_\epsilon = 0.92$. The minimum number of points needed to form a shape is a function of the total number of points for each view $M_{min} = \frac{M_{total}}{200}$.

The depth map filtering step is skipped for the reconstruction of this experimental setup, and it is substituted with point cloud filtering, following the OpenMVS parameter settings for the published results available in the *ETH3D* website for a fair comparison with the baseline method. Depth map fusion is then used as enfolded in OpenMVS library [Cernea, 2020].

5.3.3 Evaluation metrics

According to [Schöps et al., 2017; Knapitsch et al., 2017], completeness, or recall, is calculated as the amount of GT points for which the distance to the MVS reconstructed points are below a certain threshold τ. On the contrary, accuracy, or precision, refers to the ratio of reconstructed points which are within the threshold distance τ from the ground truth points without taking into consideration the

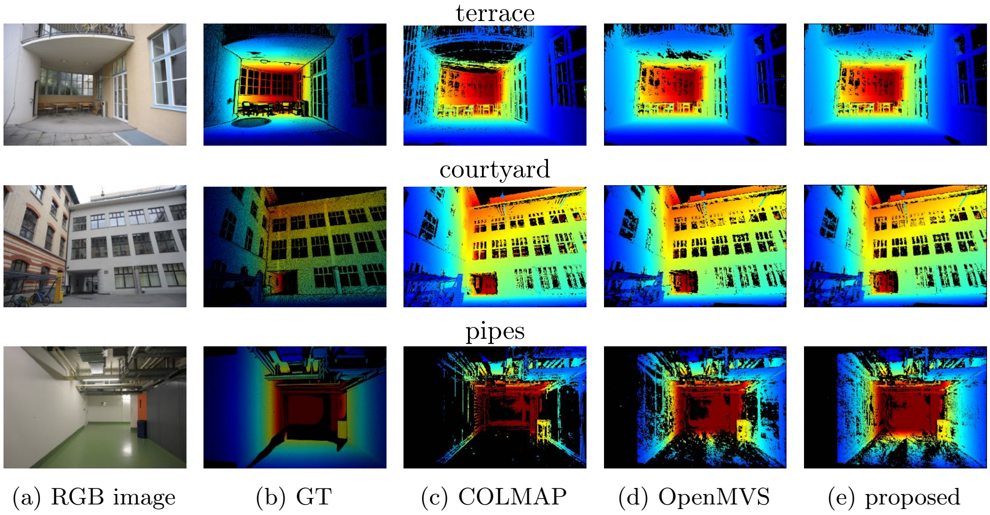
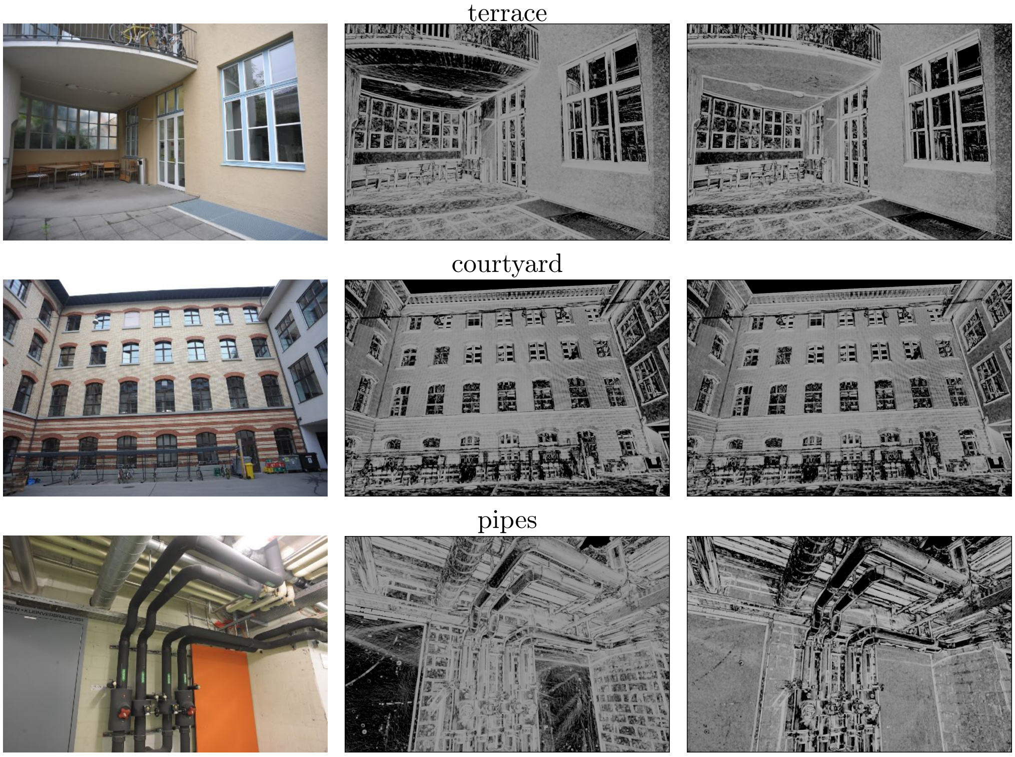


Figure 5.7: **Qualitative depth map comparison of the proposed method with other state-of-the-art algorithms for the *ETH3D* benchmark sequences**. GT depth maps look sparse as they contain empty pixels [Schöps et al., 2017]. All other depth maps are scaled to the GT color scale.

GT information gaps. Both accuracy and completeness are considered important for the effectiveness of the methods, while F_1 score their harmonic mean, defined as $F_1 = 2(precision * recall)/(precision + recall)$.

5.3.4 Evaluation on the ETH3D benchmark

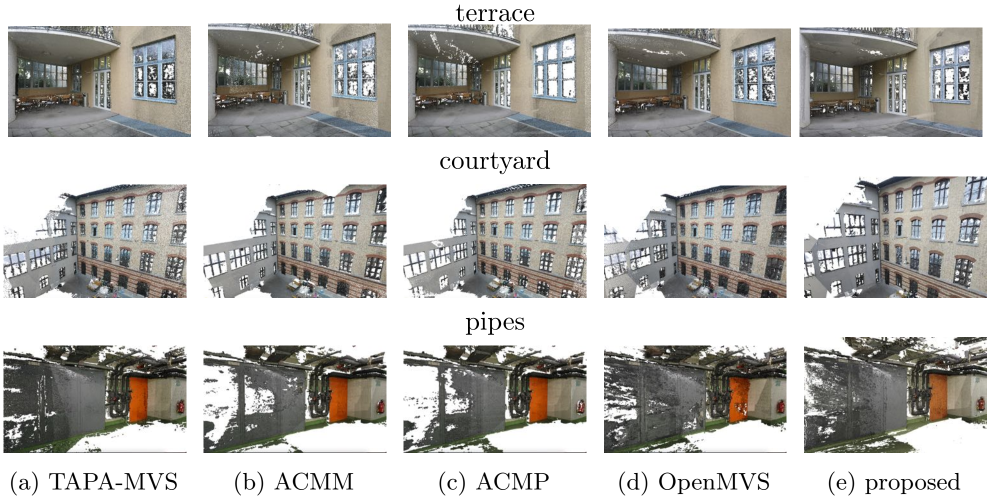
Experimental results on both *ETH3D* sequences show the potential of the proposed approach in handling efficiently textureless areas and generating more complete point clouds. It is to be noted that *ETH3D-courtyard* and *ETH3D-terrace* are generally complete sequences acquired with dense image networks of high overlap where no particularly problematic textureless areas are present. Indeed, most state-of-the-art algorithms perform well on them, as shown in Table 5.1. Even in this case, the proposed approach performs in a competitive way. On the other hand, *ETH3D-pipes* is one challenging sequence, featuring lower overlap, strong viewpoint changes, and large textureless areas or reflective surfaces. The presented method outperforms the other algorithms in completeness and F_1 score in this particular sequence. Overall, the proposed method generates more complete depth maps (Figure 5.7) and higher confidence values (Figure 5.8) for the *ETH3D* datasets with respect to other MVS methods and the respective point clouds contain less gaps (Figure 5.9). As shown in Table 5.1, better completeness results are achieved with respect to all other methods in all three *ETH3D* datasets for $\tau = 2$ cm and for $\tau = 10$ cm except for *ETH3D-courtyard* in $\tau = 10$ cm



(a) RGB image (b) OpenMVS (c) proposed

Figure 5.8: **Qualitative confidence map comparison of the proposed method with respect to the baseline OpenMVS on the *ETH3D* datasets.** Scale black to white, where black means lower confidence (high photometric cost). It is evident that the plane priors increase the confidence, and this is particularly important where textureless areas are present in *ETH3D-terrace* (ceiling) and *ETH3D-pipes* (orange panel, closet). For *ETH3D-courtyard* where the not evident textureless areas exist, the confidence maps remain mostly the same.

where *semantic PatchMatch* is ranked second. Accuracy and F_1 score values are significantly higher than the baseline OpenMVS for *ETH3D-pipes*, marginally better for *ETH3D-terrace*, and slightly lower for *ETH3D-courtyard* for both τ = 2 cm and τ = 10 m. However, in terms of F_1 score, the proposed method is always among the best ones: other methods that perform well in accuracy (such as COLMAP [Schönberger et al., 2016]) suffer in completeness since they generate significantly sparser point clouds. ACMM [Xu and Tao, 2019] and ACMP [Xu and Tao, 2020b], following the COLMAP framework, typically have similar high accuracy values. The qualitative comparisons of the depth and confidence maps, where available, show that the proposed *semantic PatchMatch* method delivers more complete depth maps and higher confidence values even in textureless areas where other algorithms fail (Figures 5.7 and 5.8).



(a) TAPA-MVS (b) ACMM (c) ACMP (d) OpenMVS (e) proposed

Figure 5.9: **Qualitative point cloud comparison of the proposed method with other state-of-the-art algorithms** for the *ETH3D* benchmark sequences *terrace*, *courtyard* and *pipes*. Dense reconstructions for the state of the art methods are as in ETH3D evaluation site.

Ablation study. As an ablation study, the input labels are removed to evaluate the performance of the method in the absence of semantic guidance; hence, the search for valid dominant planes is extended across all image regions (proposed-w/o labels). The results show similar performance with the *semantic PatchMatch* method for the *ETH3D-courtyard* and *ETH3D-terrace* sequences with typically marginally lower accuracy, completeness and F_1 score values (Table 5.1). For the *ETH3D-pipes* sequence, more evident improvement is proven while using the labels (proposed), especially in completeness ($2-4\%$) and F_1 score ($1-2\%$) with respect to the variant without the labels (proposed-w/o labels) as well as compared with the baseline OpenMVS.

5.3.5 Evaluation on custom datasets and the UDD5 dataset

For *Piazza Duomo* and *Piazza Navona* datasets, the proposed approach generates more complete point clouds with respect to the baseline and other MVS methods. Especially in the low textured regions, satisfying results in gap-filling in the depth maps and higher confidence values are achieved (Figure 5.11, first two rows), while the 3D point clouds lack less information in textureless areas (Figure 5.10). This is also proven by the completeness score, which outperforms all other methods; the proposed method has the second-best accuracy after COLMAP (Table 5.2) that, however, produces very sparse results (Figure 5.10a). The *Piazza Duomo* and *Piazza Navona* datasets have been proven to be challenging, as they were acquired with sparse image networks and feature, thus, relatively small overlap with respect

Table 5.1: Accuracy, completeness and F_1 score (%) comparisons for $\tau = 2cm$ and $\tau = 10cm$ for the *ETH3D* benchmark datasets (the higher the better). Values for the other methods are taken from the *ETH3D* evaluation site. Best values in bold. Second-best values are underlined.

		$\tau = 2cm$			$\tau = 10cm$		
	Method	Acc. ↑	Compl. ↑	F_1 ↑	Acc. ↑	Compl. ↑	F_1 ↑
ETH3D - terrace	COLMAP	**96.79**	75.67	84.94	**99.29**	93.83	96.48
	TAPA-MVS	94.00	82.37	87.80	98.45	98.15	98.30
	ACMM	96.19	84.13	89.76	99.13	96.16	97.62
	ACMP	96.14	84.45	**89.92**	99.14	96.42	97.76
	PCF-MVS	92.72	84.75	88.56	98.09	97.46	97.78
	OpenMVS	88.72	87.52	88.12	98.00	98.53	98.27
	proposed	89.81	**88.83**	89.32	98.28	**98.98**	**98.63**
	proposed-w/o labels	89.77	88.65	89.21	98.26	98.94	98.60
ETH3D - courtyard	COLMAP	88.98	73.47	80.49	99.14	92.20	95.54
	TAPA-MVS	84.69	77.04	80.68	97.64	96.14	96.89
	ACMM	**91.35**	82.85	**86.89**	**99.51**	91.90	95.56
	ACMP	90.83	80.96	85.61	99.43	90.80	94.92
	PCF-MVS	86.12	83.67	84.88	98.43	94.44	96.39
	OpenMVS	80.46	90.10	85.01	97.85	**97.63**	**97.74**
	proposed	79.66	**90.58**	84.77	97.61	97.22	97.41
	proposed-w/o labels	79.69	90.43	84.72	97.60	97.04	97.32
ETH3D - pipes	COLMAP	**97.77**	34.24	50.72	99.18	62.75	76.86
	TAPA-MVS	93.71	63.80	75.91	97.90	86.70	91.96
	ACMM	96.63	53.97	69.26	98.89	66.25	79.34
	ACMP	97.65	53.54	69.16	**99.20**	65.80	79.12
	PCF-MVS	90.40	69.18	78.38	98.48	88.47	93.21
	OpenMVS	82.33	64.55	72.36	95.95	85.42	90.38
	proposed	85.33	**73.50**	**78.97**	96.89	**93.63**	**95.23**
	proposed-w/o labels	84.19	69.88	76.37	97.32	91.08	94.10

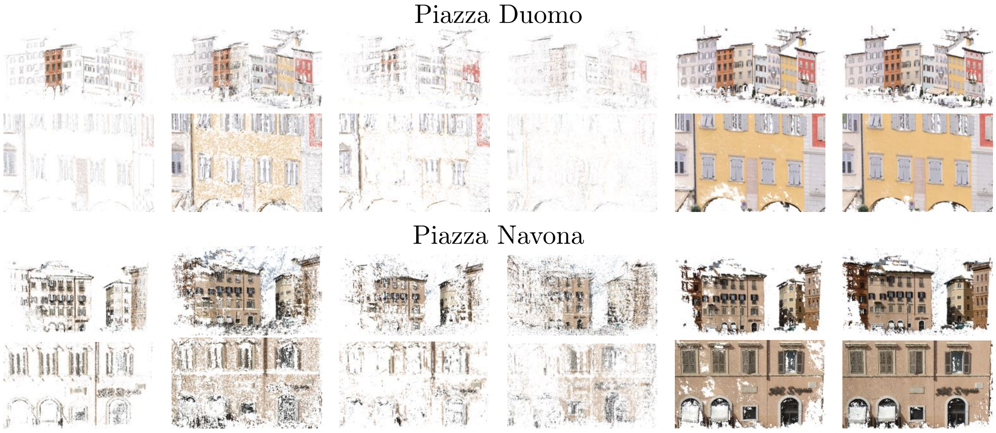
to *ETH3D* datasets. Moreover, the scenes include large textureless regions, as they depict building facades with walls of mostly homogeneous color (Figure 5.10). *UDD5* dataset [Chen et al., 2018a], on the other hand, is a generally dense sequence of 200 images of high overlap. The standard OpenMVS reconstruction in this case was not particularly problematic since the scene was generally well-textured. Small gaps in depth maps still exist, though they are mainly caused by occlusions. In such cases, the proposed algorithm performs equally well as the standard approach (Figure 5.11, lower row).

5.4 Discussion

Coupling semantic reasoning and image-based 3D reconstruction has caught the attention of the research community in recent years since it has been observed that higher-level scene semantics can potentially help to overcome problems

Table 5.2: Accuracy, completeness and F_1 score (%) comparisons of the Piazza Duomo dataset for $\tau = 10cm$ (the higher the better). Best values in bold. Second-best values are underlined.

	Method	Acc. ↑	Compl. ↑	F_1 ↑
Piazza Duomo	COLMAP	**88.89**	38.00	52.24
	TAPA-MVS	25.56	23.74	24.62
	ACMM	50.87	50.51	50.69
	ACMP	40.92	25.93	31.75
	OpenMVS	70.53	<u>68.55</u>	<u>69.52</u>
	proposed	<u>71.08</u>	**69.38**	**70.22**

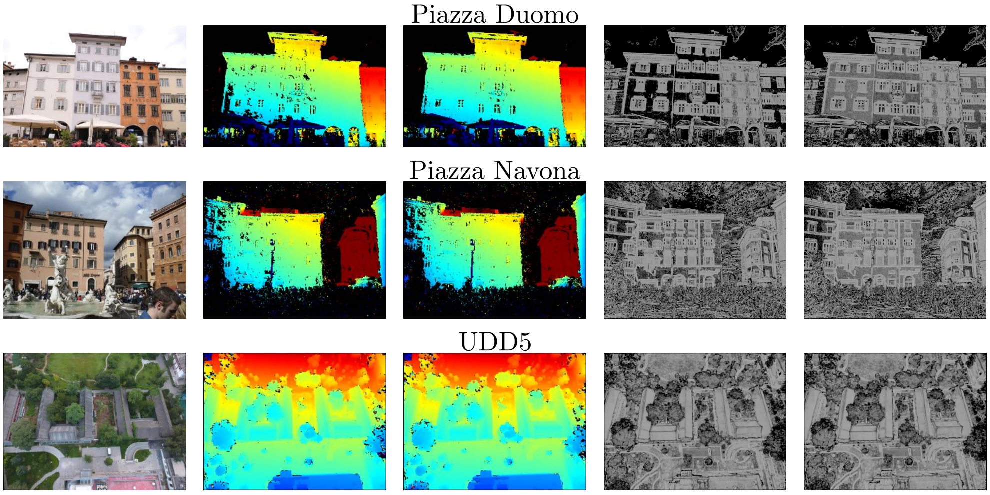


(a) COLMAP (b) ACMM (c) ACMP (d) TAPAMVS (e) OpenMVS (f) proposed

Figure 5.10: **Qualitative point cloud comparison for the custom datasets.** Overview and detailed views of *Piazza Duomo* (first two rows) and *Piazza Navona* (last two rows): state-of-the-art results versus the proposed method.

that plain image information, either visual appearance or geometric, cannot solve. Indeed, conventional MVS approaches based solely on photo-consistency measures are generally robust yet often fail in calculating valid depth pixel estimates due to matching ambiguities in non-Lambertian parts of the scene. Real-world applications often face this problem, as in several man-made scenarios containing building facades, indoor scenes, or airborne (mostly oblique) datasets, such challenging areas are present.

In this chapter, a novel approach is proposed to specifically undertake this challenge by leveraging semantic priors into a PatchMatch-based MVS, targeting high-resolution and real-world photogrammetric applications, *semantic PatchMatch*. Semantic reasoning is used in the form of a priory-generated semantic labels for each pixel of the scene. Such cues are used to impose class-specific geometric



(a) RGB image (b) depth (standard OpenMVS) (c) depth (proposed) (d) confidence (OpenMVS) (e) confidence (proposed)

Figure 5.11: **Qualitative depth and confidence map comparison of the proposed method with respect to the baseline OpenMVS on custom datasets and the *UDD5* dataset.** The proposed method improves depth estimations and achieve higher confidence scores in problematic planar areas for *PiazzaDuomo* and *PiazzaNavona*. In *UDD5*, where no evident textureless areas exist, it performs like standard OpenMVS.

constraints during multi-view stereo, optimizing the depth estimation on weakly supported, textureless areas. Guided by the segmentation masks, dominant shapes, e.g., planes, are detected directly in 3D with RANSAC-based techniques. A new, adapted cost function is introduced that combines and weights both photometric cost and plane prior hypotheses for each pixel, propagating, thus, more accurate depth estimates across the image. Being adaptive, it fills in apparent information gaps and smooths local roughness in problematic regions while at the same time preserving important geometric details. Experiments on the *ETH3D* benchmark and custom datasets demonstrate the effectiveness of the presented approach. The experiments were designed as such to include real-world and high-resolution scenarios, typically the case for various photogrammetric applications. Although the good performance scores and the potential of the proposed method, there some limitations; in the following paragraphs, such challenges are critically discussed.

Semantic guidance. Semantic reasoning is used in the presented approach to guide the prior generation. Indeed, plane fitting is guided by the semantic labels, making class-specific assumptions and restricting the search to the regions where it is more probably to find dominant planar structures. However, the detected planes in the 3D space have no clearly-defined boundaries and the semantic guidance may not be always effective. That is, plane boundaries are limited by the pixel label, yet, multiple planes within the potentially planar classes may intersect with each other in the 3D space. This problem is solved in the proposed approach by using the closest plane in 3D, but this criterion may not always assign the correct plane to the point and is considered a limitation of the method. The semantic masks, even roughly estimated, are considered to be obtained a priory and be given as input to the integrated MVS pipeline. Nowadays, semantically segmented data become increasingly more available, and the generation of such masks is feasible for a variety of real-world scenarios; however, the method is highly dependent on this prerequisite, and its applicability may be restricted in cases where no such cues cannot be obtained. Nonetheless, the ablation study showed promising results for the robustness of the method even in absence of semantic guidance. Toward the generalization of the presented approach also in case where semantic segmentation masks are not available or cannot be easily generated, a more powerful approach is presented in Chapter 6, based only in local textureness and structure priors.

RANSAC performance. The proposed approach extracts RANSAC shapes to support depth estimation in problematic areas. RANSAC is typically sensitive to its parameters and often requires fine-tuning based on the application, yet in this approach, an adaptive tuning of the parameters based on the average point spacing $\bar{s}$ is proposed to enhance robustness and generalization. Given appropriate parameter tuning, RANSAC is generally a robust model-fitting algorithm and can yield satisfying results in the presence of some outliers. The proposed method estimates planar priors in the 3D space on the intermediate point clouds generated by the roughly estimated depth maps after a few standard PatchMatch iterations. However, regions with matching ambiguities result in outliers in 3D, and, consequently, the plane fitting procedure is less robust. Indeed, even with the semantic guidance and the proposed point filtering strategy, in particularly noisy point clouds, RANSAC is not able to cope with these errors. Hence, plane prior hypotheses cannot be accurately generated, and wrong depth estimations will still exist in a similar way as the standard PatchMatch approach of OpenMVS.

Cost function. The proposed combined cost function has an adaptive formulation and gives priority to the standard photometric cost in rich-texture areas, while trusting more the prior hypotheses in the textureless one. It has proven efficient enough under different scenarios in the experiments and yields promising results. However, highly reflective areas, such as some surfaces in the *ETH3D-pipes* dataset, seem to be very difficult to treat since the matching ambiguities there cause erroneous depth and normal estimates over large areas and even such an adaptive formulation struggles to propagate reliable values to the neighboring pixels. An intuition about this problem could be that the normal inconsistencies are not tackled efficiently, or that in the implementation, RANSAC fails to assign always correct normals to its plane hypotheses.

Runtime performance. The proposed method behaves similar to standard OpenMVS, as the semantic map loading and the two additional PatchMatch iterations add little extra computational cost to the entire MVS procedure.

Chapter 6

Quadtree-guided priors in depth estimation

This chapter introduces a novel, generic and robust method to support depth estimation under a PatchMatch-based MVS scenario and generate more complete dense representations using high-resolution images in real-world scenarios. The framework is similar to the semantic PatchMatch method presented in Chapter 5, yet it is extended by introducing texture-guided structure/shape priors. The new approach goes beyond semantic reasoning, acknowledging that in several real-world applications, such cues are not easy to obtain, although the vast training data availability and the robustness of the state-of-the-art algorithms, as discussed in Chapter 4.

The basic insight for this method is the fact that textureless areas commonly belong to local planar structures, which is particularly true in the case of man-made environments; indeed, similar priors have been used in the past for depth reconstruction in stereo or MVS scenarios [Furukawa et al., 2009; Romanoni and Matteucci, 2019]. Unlike previous works, the proposed method adopts a quadtree structure to organize the input image according to the local texture so that pixels with similar intensities are grouped together under the same quadtree block. The hypothesis is made that pixels within a block belong to the same local plane rather than making the strong assumption that they share a constant depth value. A clear advantage of this method is that the block size is adaptive to local texture; thus, textureless areas are grouped in large blocks, while rich textured areas are represented by smaller block sizes down to the size of one pixel. Moreover, a quadtree structure allows exploiting neighbors in an efficient way differently than

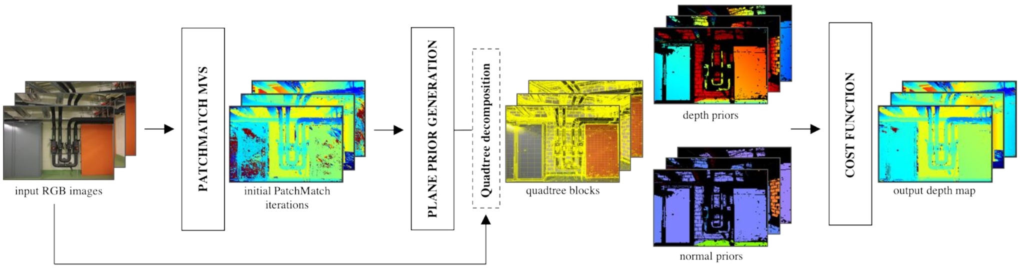


Figure 6.1: **The proposed pipeline based on quadtree-guided priors.** First, some initial standard PatchMatch iterations are performed to generate rough depth estimates. Plane prior generation follows using quadtree block guidance. Plane and normal priors are then taken into consideration in a combined cost function to generate more complete and accurate final depth maps.

standard image segmentation techniques such as superpixels [Van den Bergh et al., 2015]. Dominant plane priors are detected in 3D using a RANSAC-based approach, and for each pixel of the block, a plane prior hypothesis is made. The proposed approach does not rely on training data and is, thus, domain-independent, i.e., it can generalize in diverse scenarios. In contrast to recent learning approaches that try to solve this issue [Yang et al., 2021b; Wang et al., 2021], this method can directly handle high-resolution images and is computationally efficient.

6.1 Prior-assisted PatchMatch

Despite the recent advances in conventional and learning-based algorithms, most methods lack completeness in textureless areas due to matching ambiguities and imply, thus, the need for advanced scene cues along with the standard photometric consistency measures. Piecewise planar constraints have been used in traditional stereo scenarios in the past [Gallup et al., 2010; Furukawa et al., 2009]. Recently, Romanoni and Matteucci [2019], acknowledging the challenge of textureless regions, especially under large-scale applications, assumed piecewise planarity on image superpixels [Van den Bergh et al., 2015] for joint PatchMatch and view selection. Superpixels were derived in multi-scale resolution for preserving depth details, and a textureness term was added to the cost function. Similarly, Kuhn et al. [2019] extended this framework, achieving depth completion where textureless areas are treated with multi-scale geometric consistency guidance, yet as a post-processing step. Close to these works, rather than superpixels, in the presented method, quadtree-based image decomposition is used. The basic intuition for this choice is that quadtree blocks are robust in aggregating pixels of similar intensities, and, most importantly, their size is adaptive to the local texture. Moreover, quadtree structures allow for the exploitation of neighboring relations among blocks. Xu and Tao [2020b] added direct planar priors to assist PatchMatch with planar compatibility constraints for the matching cost. Their approach is adjacent to the idea presented in this chapter, yet the priors are triangular primitives derived by sparse correspondences without explicit textureness constraints. Wang et al. [2020b] used a pyramid architecture for coarse-to-fine MVS with mesh guidance and a confidence prediction network for depth refinement as an extra module. However, although efficient for large textureless areas, such multi-scale schemes are limited by the predefined scales and often fail to preserve fine details. Recently, plane hypothesis inference using Markov Random Fields (MRFs) is proposed in [Sun et al., 2021] as a post-processing step after initial depth estimation and filtering. The work presented in Chapter 5 exploits semantic reasoning to detect dominant plane priors in the object space and improve depth reconstruction in textureless areas as in [Stathopoulou et al., 2021b]. This formulation has promising results but heavily relies on semantic priors that are not always available or easy to obtain, especially in real-world application public datasets designed for evaluating MVS algorithms. Building upon this work, this chapter proposes a more generic, robust scheme that can be implemented independently from semantic label guidance, as it relies only on local structure information and can be therefore generalized under diverse scenarios.

6.2 Proposed methodology: quadtree-guided PatchMatch

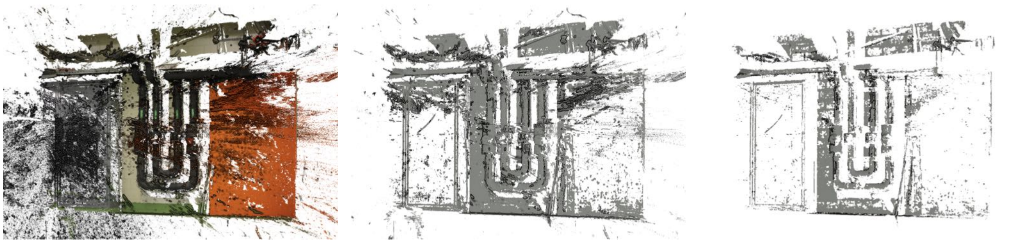
Given a set of input images $\mathcal{I} = \{I_0, \dots, I_m\}$ with known camera poses along with a sparse point cloud, typically derived via standard SfM methods, MVS algorithms aim to estimate a reliable depth value for almost every pixel and generate, thus, a complete dense 3D reconstruction of the scene. A standard procedure for depth estimation using the PatchMatch algorithm under a multi-view scenario, as described in detail in Sections 3.3 and 5.2 can be outlined in four steps: (1) view selection and initialization for selecting the best-overlapping pairs and initialize the depth and normal estimations for each pixel (2) depth estimation based on photometric consistency, depth propagation, and random refinement, (3) depth map filtering and (4) fusion of the individuals 3D point clouds into one. Each step is crucial for the quality of the 3D reconstruction, and multiple works have been proposed to improve them. Yet, depth estimation itself is commonly based on simple photometric consistency measures to calculate the matching costs between two corresponding pixels. Such a common measure is the normalized cross-correlation (NCC) and, in particular, one of its variants, the zero-mean NCC (ZNCC) as defined in Equation 2.18, which is typically robust to illumination changes, as the mean intensity of the neighborhood is subtracted. As explained in the previous chapter, in highly textured areas, pixel estimates are propagated from the neighboring pixels to the current one if their matching cost is lower than the current one, and in this way, more reliable depth estimates are spread using spatial propagation. Nonetheless, in textureless areas, multiple local minima may exist; in such challenging regions, standard photometric consistency measures fail to discriminate patches since, for a certain patch, an ambiguous number of good candidate matches may exist, resulting in noisy depth estimations or information gaps.

The proposed approach specifically tackles the matching ambiguities challenge in depth estimation using depth hypotheses derived by structure/shape priors (Figure 6.1). Different from standard PatchMatch approaches that make a priori planar assumptions to ensure local smoothness, in this method, plane priors are detected in the object space for dominant planes of the scene. To assign a plane prior to each pixel, a quadtree scheme is adopted to organize the image pixels in groups, based on the intuition that similar intensity pixels are likely to belong to the same plane. Thus, in the case of highly textured areas, local planes are degenerated to the standard support planes of PatchMatch, preserving the curvature and fine details. Each quadtree block is assigned to its corresponding 3D plane, i.e., all pixels among a quadtree block are set to a common normal value $\mathbf{n}_{quad}$. By leveraging the depth prior hypotheses in the standard matching cost function in a similar fashion as in Chapter 5, more reliable depth estimates are propagated in textureless regions from the rich textured ones, which commonly occur near the natural crease edges. The proposed method significantly increases the completeness of the depth maps, achieving similar results with depth completion techniques [Kuhn et al., 2019] which, however, act as depth map refinement step in post-processing.

As in Chapter 5, also this pipeline builds upon the standard PatchMatch MVS approach following [Shen, 2013] as implemented in OpenMVS library [Cernea, 2020], which is considered the baseline method. View selection is performed using visibility criteria, i.e., baseline and angles between images, and depth maps are initialized by triangulating the sparse SfM point cloud. A simple sequential PatchMatch propagation scheme is followed as in [Cernea, 2020]. However, the proposed approach can be easily integrated into other algorithms employing different view selection and propagation schemes, e.g., checkerboard or red-black propagation [Xu and Tao, 2020b].

6.2.1 Quadtree-guided 3D plane prior hypotheses

A set of input images $\mathcal{I} = \{I_0, \ldots, I_m\}$ is provided, with known camera poses $(\mathcal{R}, t)$, sharing the same calibration matrix $\mathbf{K}$ along with the sparse point cloud of the scene. In a similar line of thought as in Chapter 5, it is assumed that at least a couple of PatchMatch iterations have been executed in such a way that a depth d_{tmp} and normal $\mathbf{n}_{tmp}$ estimation are available for (if possible) all scene pixels. Among these depth estimations, generally, some correspond to high-fidelity matches and are already reliable, at least for the highly textured areas and the areas near the crease edges. Plane prior hypotheses are generated in the object



(a) input point cloud (b) after confidence criterion (c) after planarity criterion and outlier removal

Figure 6.2: **The proposed point cloud filtering strategy.** The intermediate point clouds for each view are filtered to generate more robust plane priors. First, points of high confidence, e.g., low photometric cost, are considered; then, the points are further filtered based on their planarity value, to distinguish the points that are more likely to belong to a planar neighborhood.

space, i.e., in the individual 3D point clouds for each view; these clouds are the 3D equivalents of the intermediate depth maps $\mathcal{D}_{tmp}$ computed by the first few PatchMatch iterations. Also here, a RANSAC-based technique is used as an effective method to detect dominant plane structures in the scene. Inevitably, the point clouds also include a vast number of outliers and erroneously estimated points, especially in the textureless areas that inevitably affect the robustness of the algorithm. A filtering strategy is followed, in the same line of thought with the semantic PatchMatch method, yet applied to all pixels of each view since no semantic guidance is considered in this method. Then, plane hypotheses are detected, guided by the quadtree blocks on the images.

Point cloud filtering. An analogous strategy to the one presented in Chapter 5 is followed. However, in the absence of semantic guidance, all image pixels are initially considered and projected in the 3D space (Figure 6.2a). To select only the most reliable depth estimates out of all scene pixels, a confidence criterion is set; the term confidence here is equivalent to the matching cost. A low matching cost implies highly correlated pixels and thus reliable correspondences (i.e., of high confidence). Hence, based on this criterion, only estimates of low photometric cost qualify for being inliers. In the experiments we use the matching cost definition of the baseline method OpenMVS [Cernea, 2020] defined between $[0, 2]$ where 2 indicates uncorrelated patches and 0 indicates highly correlated ones. To exclude unreliable correspondences, we consider only 3D points that correspond to matches with $cost < 0.20$. Given that the most reliable points are the ones around the crease edges, these points are more likely to be kept after the confidence criterion (Figure 6.2b). Then, similar to the semantically-guided prior generation (Chapter 5), the points with low planarity value as defined in Equation 5.1 are excluded; finally, gross outlier removal is performed based on the average point spacing within a $k = 20$ point neighborhood (Figure 6.2c).

Quadtree decomposition. A quadtree data structure is generated to guide the plane hypothesis; the input images are decomposed based on the standard deviation of the pixel intensities, where each "parent" has four "children". Quadtrees are selected as an efficient, regular data structure to group the image pixels into groups of similar intensities while exploiting the neighboring relations rigorously. Region splitting stops when the desired minimum standard deviation (here is used $std = 1$) or the minimum block size (here, the minimum dimension corresponds to 3 pixels) is reached, and this process is repeated recursively until all blocks meet one of the two criteria. Generally, the deeper the quadtrees are, the more details are kept in the final structure. For each block, its adjacent neighbors can be identified and stored in memory. The generated quadtree blocks are used as guidance regions of interest (ROIs) during the prior generation, making the assumption that pixels within the same block approximately belong to the same 3D plane. Experimentally, this method was found to be superior to using image clustering methods such as, e.g., superpixels [Van den Bergh et al., 2015] used in the literature for similar cases [Romanoni and Matteucci, 2019; Kuhn et al., 2019]. The intuition for this is that originally, superpixel areas are mostly of similar size, whereas quadtree block size is highly irregular and adaptive based on the textureness of the pixel neighborhood (Figure 6.3). In this way, highly textured areas with harsh intensity changes are represented by dense blocks of smaller size, while extended low-textured regions are grouped under the larger, sparser blocks (Figure 6.3). Adaptive superpixel segmentation has recently been introduced in the literature [Uziel et al., 2019], yet tree structures have the strong advantage of efficiently storing the information of neighbor relations, thus enabling the robust propagation in the block neighborhood.

RANSAC plane detection. RANSAC plane detection is performed subsequently, using the Efficient RANSAC optimization algorithm [Schnabel et al., 2007] as implemented in the CGAL library [The CGAL Project, 2021]. In a similar fashion as in the semantic PatchMatch approach (Chapter 5), adaptive parameter setting is used based on average spacing $\bar{s}$ for the Euclidean distance between a point and a shape ϵ and the connectivity measure c_ϵ, while the minimum number of points M_{min} needed to form a valid shape is also adaptively set based on the total number of 3D points for each view.

Depth and normal prior generation. Prior generation exploits quadtree guidance to robustly propagate the plane prior hypothesis to the textureless areas of the scene. To do so, the quadtree blocks that include inlier RANSAC points are assigned to their corresponding 3D plane, and a depth hypothesis is assigned to each block center via ray-plane intersection as a prior. The normal prior hypothesis $\mathbf{n}_{prior}$ is given by the assigned normal of the 3D plane $\mathbf{n}_{R_{plane}}$, under the constraint that the direction should point towards the camera to avoid impossible normal directions (i.e., flipped normals). Each assigned block Q is

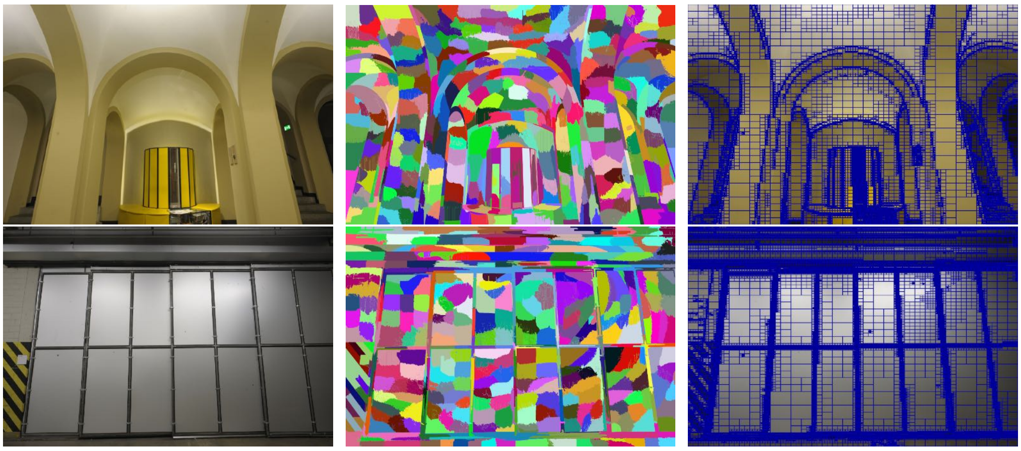


Figure 6.3: **Superpixel segmentation and quadtree decomposition.** Example images from *ETH3D* datasets (left) with the calculated superpixels following Van den Bergh et al. [2015] and our proposed quadtree decomposition (right). Using quadtrees, block size is adaptive based on the textureness of the area while neighbor relations are kept and inherited, supporting the depth propagation in neighboring blocks with high color similarity.

compared to its adjacent neighboring blocks in 4 directions (top, down, left, and right), and, if they have a high color similarity s_{color}, the 3D plane hypothesis is propagated to the neighbors (Figure 6.4). The similarity metric s_{color} is the absolute distance of the a and b channels of the mean pixel intensities of each block in the CIE-Lab color space. CIE-Lab is preferred over RGB as it is tuned to human perception and is more robust in illumination changes and shadow effects [Wang et al., 1981; Tomasi and Manduchi, 1998]. In this way, blocks with no inlier RANSAC plane points will be assigned the plane hypothesis of their adjacent blocks if they have similar color appearance. Accordingly, the 3D planes will be expanded using 2D guidance based on the local textureness (Figures 6.5, 6.6). However, handling the hypotheses propagation is a crucial step since blocks with similar appearance may belong to different planes. Since this commonly happens around crease edges, all neighboring blocks except those of maximum level, i.e., the small blocks generated in highly textured areas or near the edges, are considered eligible for propagation. The propagation should start from the smallest blocks towards the bigger ones since the bigger blocks, inevitably corresponding to large, textureless areas, are more likely to contain noisier points and be assigned with less reliable plane hypotheses. If plane propagation is performed efficiently, all image pixels that belong to quadtree blocks with similar appearance will be assigned to the respective depth prior d_{prior} via ray-plane intersection, while the pixel normal prior $\mathbf{n}_{prior}$ is the normal of the plane $\mathbf{n}_{R_{plane}}$ (Figure 6.6b). For unassigned pixels, no prior hypothesis will be generated. An overview of this quadtree-guided propagation is shown in Algorithm 2. Both d_{prior} and $\mathbf{n}_{prior}$ for

Algorithm 2 Quadtree-guided RANSAC plane propagation

Input: RANSAC 3D shapes $\mathcal{S} = \{S_1, \ldots, S_m\}$ for each view $I \in \mathcal{I}$
Output: $\mathcal{D}_{prior}$, $\mathcal{N}_{prior}$ for each view

```
for all RANSAC shapes $\mathcal{S} = \{S_1, \ldots, S_m\}$ do
  get assigned set of points $\mathcal{S}$
  assert plane normal direction $\mathbf{n}_{R_{plane}}$
  if consistent with camera viewing direction then
    keep normal
  else
    invert normal
  end if
  for all assigned points to planes do
    get the block in which they belong
    if color similarity is high AND block size is big then
      propagate to neighboring block
    else if  then
      stop propagation
    end if
  end for
  for all assigned blocks to planes do
    calculate $d_{prior}$ with ray-plane intersection
    assign plane normal $\mathbf{n}_{R_{plane}}$ as $\mathbf{n}_{prior}$
  end for
end for
return  $\mathcal{D}_{prior}, \mathcal{N}_{prior}$
```

each pixel are stored in corresponding maps $\mathcal{D}_{prior}$ and $\mathcal{N}_{prior}$.

It is to be noted that the propagation scheme is defined in such a way that is unlikely to infinitely expand the plane boundaries. Each quadtree block is directly assigned to the corresponding 3D plane, in such a way that no closest plane criterion, as implemented in Chapter 5, is needed (Figure 6.6).

6.2.2 Adaptive cost calculation

In the same line of thought as in the semantic PatchMatch method described in Chapter 5, rather than implying hard constraints and forcing planarity by directly assigning the plane hypothesis $[d_{prior}, \mathbf{n}_{prior}]$ to every pixel p of a particularly high cost, the plane prior is leveraged into the standard photometric matching cost given in Equation 5.2 in an adaptive fashion also considering the local textureness described by the coefficient given in Equation 5.4. Similar to the definition of the semantic coefficient s in Equation 5.5, a quadtree-guided plane prior coefficient g is defined as:

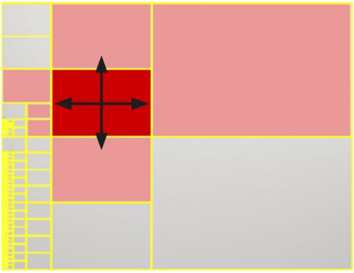
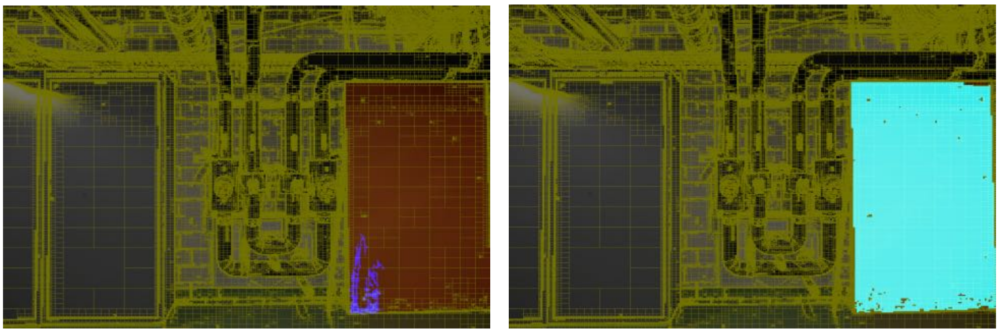


Figure 6.4: **Quadtree-guided plane propagation scheme.** A quadtree block Q with inlier RANSAC points (in dark red) is assigned to a plane prior; this hypothesis is propagated to its adjacent neighbors across 4 directions (in light red), if the latter have similar color appearance with Q and no inlier RANSAC points. If a block is assigned to a plane, prior hypotheses are generated for all pixels belonging to the block.



(a) seed points

(b) plane hypothesis propagation

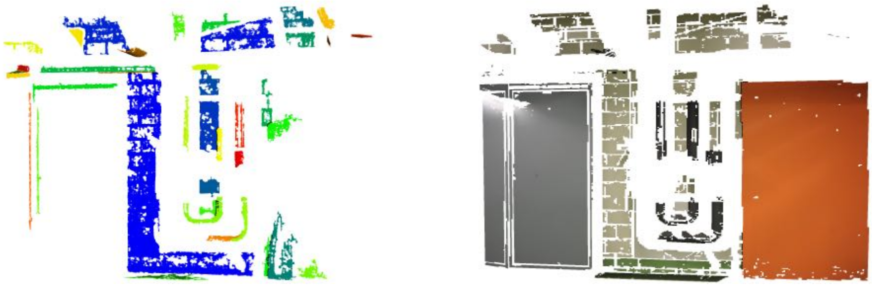
Figure 6.5: **Example of plane propagation in a textureless area.** Inlier RANSAC points (in blue) belonging to the same RANSAC plane (6.5a), and plane hypothesis propagation to neighboring blocks (in cyan) of similar appearance that have no inlier points (6.5b).

$$g = e^{\frac{-\delta^2}{2\sigma_g^2}}, \tag{6.1}$$

where σ_g is a constant fixed experimentally. Also here, δ is given as the percentage of difference between the quadtree-guided depth prior hypothesis d_{prior} and the current PatchMatch estimate d_{tmp}:

$$\delta = \frac{|d_{prior} - d_{tmp}|}{d_{prior}}. \tag{6.2}$$

Finally, the cost function is defined in an adaptive formulation similar to Equation 5.7:



(a) 3D point sets assigned to RANSAC planes

(b) plane priors in 3D

Figure 6.6: **Example of plane prior generation in 3D.** Points assigned to the detected 3D planes with RANSAC; point sets are color-coded where each color corresponds to a plane id (6.6a). Final plane priors in 3D after the quadtree guidance (6.6b).

$$c = c_{ph}(1 - t) + \gamma(1 - g)t, \tag{6.3}$$

where γ being a weight factor.

As discussed in Section 5.2.2, this cost function formulation encourages the propagated depth estimation to be close to the plane prior. The goal is to favor the hypotheses deriving from the plane prior in the textureless areas. However, it has to be underlined that the photometric cost remains the main term of the cost function in such a way that photometric consistency will be trusted, and $[d_{tmp}, \mathbf{n}_{tmp}]$ will be prioritized under highly textured regions. On the contrary, the variance would tend to zero for low textured areas with high cost, and the prior depth hypothesis $[d_{prior}, \mathbf{n}_{prior}]$ will be favored.

Depth propagation is performed from top-left to bottom-right and vice versa in a sequential propagation scheme [Shen, 2013; Cernea, 2020]. The PatchMatch proceeds with random refinement circles, assigning random depth values and comparing them to the current estimate to avoid the convergence to local minima. In the proposed method, this set of hypotheses is extended by also including $[d_{prior}, \vec{n}_{prior}]$ values in the sample. Such a combination of random and prior values increases the possibility of sampling a correct estimate and reduces the outliers.

Increasing the completeness of the depth maps is a crucial step in an MVS scenario as, if a pixel is visible in enough views, there are more chances the point will fulfill the consistency checks and thus be correctly reconstructed in 3D during depth map fusion. Depth map filtering within a post-processing step is commonly applied in MVS methods to remove inconsistent depth values between neighboring views, typically based on the re-projection error and the photometric consistency. While this increases the accuracy of the final point cloud, it can significantly reduce

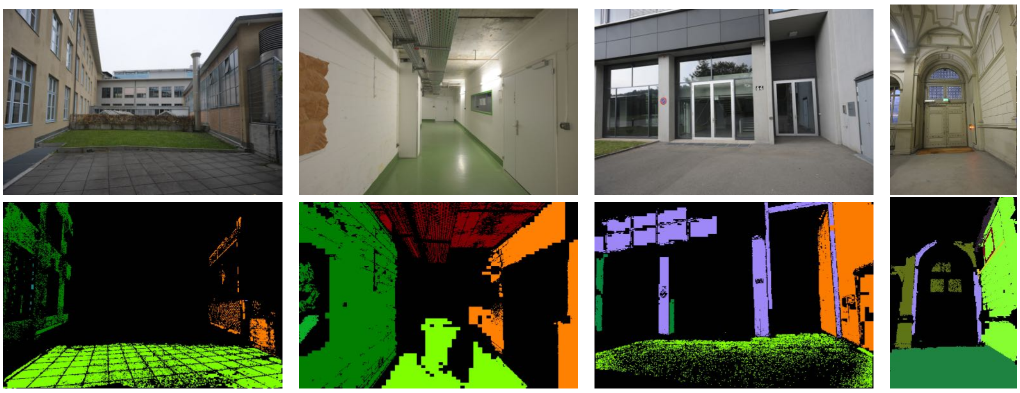


Figure 6.7: **Generated plane prior hypotheses.** Example images from *ETH3D* dataset (up) and the corresponding 3D plane hypothesis detected using the proposed RANSAC-based approach (bottom). Prior planes are color-coded by normal vector. Black regions refer to pixels with no assigned prior hypothesis.

its completeness. In the implementation of the proposed method, since most of the inconsistent depth estimates have already been refined by the integration of the plane priors, the depth map filtering step is skipped; point cloud filtering is rather performed based on visibility criteria on the fused 3D point cloud to remove the outliers directly in the 3D space in a post-processing step. Regarding depth map fusion, the scheme of [Shen, 2013] is adopted, where overlapping depth maps are merged together, comparing the depth values across the neighboring views using back projection. For both depth map fusion and point cloud filtering, the baseline implementation of [Cernea, 2020] is followed for a fair comparison between the methods. It is to be noted that the proposed approach leverages the prior information in the cost computation and hence significantly deviates from depth completion and refinement techniques that assign the new depth hypothesis directly at the pixel [Kuhn et al., 2019].

6.3 Experiments and results

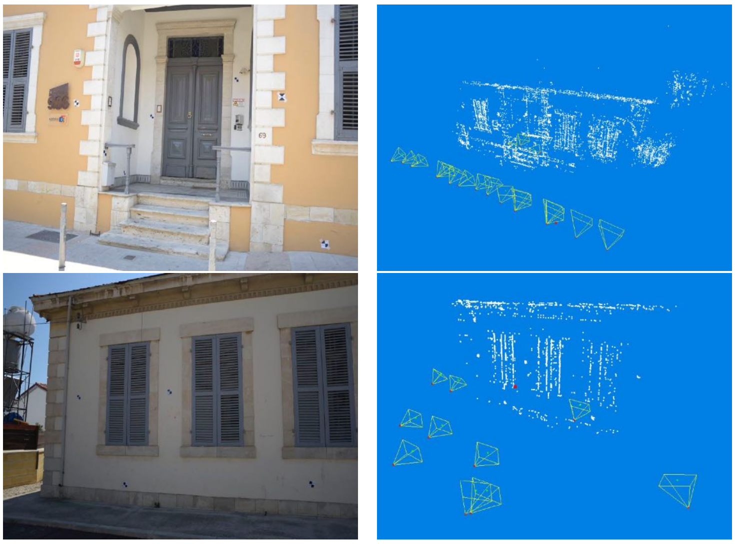
6.3.1 Datasets

The proposed method of this chapter is evaluated on the complete *ETH3D* high-resolution (6048 × 4032 pixels) MVS benchmark dataset and on two additional custom datasets of real-world photogrammetric scenarios.

ETH3D. It contains 13 and 12 sequences in the training and test sets respectively of real-world indoor and outdoor scenarios. Most *ETH3D* sequences are particularly challenging due to the image network geometry, as strong viewpoint variations exist, and the presence of large textureless areas and reflective surfaces. Typically, state-of-the-art MVS methods struggle to generate complete and accurate point clouds in such challenging scenarios, making it thus a suitable dataset to evaluate the performance of our method. For this reason, in this study *ETH3D* scenarios are considered the most challenging ones and suitable to evaluate such methods. Other datasets exist, such as the *DTU* robotic dataset [Aanæs et al., 2016] or the *Tanks and Temples* dataset [Knapitsch et al., 2017], but they are of lower resolution (1920 × 1080 pixels for *Tanks and Temples* and 1600 × 1200 for *DTU*), with less challenging image network while the majority of their scenes do not contain evidently textureless regions. *ETH3D*, on the contrary, consisting of real-world and high-resolution scenes, is closer to common photogrammetric 3D reconstruction scenarios. For *ETH3D*, camera extrinsic and intrinsic parameters, as well as the sparse point clouds derived from standard SfM [Schönberger and Frahm, 2016] are provided as GT in a scaled reference system. For the training datasets, GT 3D data acquired by laser scanning are publicly available for evaluation, while the GT data for the test set are not publicly available to prevent overfitting. For depth estimation, images are resampled to 3200 pixels keeping the aspect ratio in accordance with previous works [Xu and Tao, 2019; Schönberger et al., 2016] for a fair score comparison. The obtained 3D point clouds with the presented method are compared against the baseline method OpenMVS [Cernea, 2020], COLMAP [Schönberger et al., 2016] and other state-of-the-art algorithms that use prior-assisted PatchMatch, namely TAPA-MVS [Romanoni and Matteucci, 2019], ACMP [Xu and Tao, 2020b] and MAR-MVS [Xu et al., 2020] as well as the depth completion method of PCF-MVS [Kuhn et al., 2019]. Both conventional approaches and learning-based methods are considered, although state-of-the-art learning methods are typically applied in resized versions (lower than the 3200 used here and in the other conventional algorithms) of the *ETH3D* high-resolution images [Xu and Tao, 2020c; Wang et al., 2021].

Custom Datasets. To prove the generalization ability of the presented quadtree-guided method in other real-world photogrammetric applications, two additional custom scenes of building facades are considered, featuring evident large, textureless areas, *House 1* and *House 2*. The two sequences contain 22 and 10 high-resolution (6016 × 4016 pixels) images respectively (Figure 6.8). The camera extrinsic and intrinsic parameters are calculated using the SfM approach of Moulon et al. [2016]. The GT 3D point cloud was acquired using terrestrial laser scanning[1]. For consistency with the *ETH3D* experiments, during depth estimation, images are resampled to 3200 pixels keeping the aspect ratio.

[1] Data acquired in the acquisition campaign of the Periscope project in June 2021, Limassol, Cyprus.



(a) RGB images

(b) image network

Figure 6.8: **The custom datasets used in the experiments.** Image acquisition network with the sparse point cloud for the two custom datasets, *House 1* (top) and *House 2* (bottom).

6.3.2 Implementation details

Experiments are executed on an AMD Ryzen 2950X CPU running on 3.5GHz and 48GB of RAM. The proposed method is implemented in C++ and executed in 32 parallel threads on the CPU. The parameters used for the cost function are $\{\gamma, \sigma_t, \sigma_p, s_{color}\} = \{0.1, 0.08, 0.15, 1.5\}$ that experimentally were proven to be the best trade-off values. The combined total cost is computed using a pixel window size $N = 7$. Following the baseline implementation of OpenMVS [Cernea, 2020], $N_{source} = 8$ source images for each reference image are used. The confidence threshold value is set to threshold is set to $conf < 0.20$. During the planarity filtering, points with $\rho \leq 0.75$ were excluded (strict threshold). Quadtree propagation starts from the smallest quadtree blocks towards the largest, as they typically correspond to highly textured areas and crease edges. In the presented experiments, the 4 deeper quadtree levels that include points assigned to a RANSAC plane are considered and propagated toward their neighbors of equal or larger size. RANSAC parameters in this case are set to $\{\epsilon = 1.5\bar{s}, c_\epsilon = \bar{s}\}$, while the normal deviation is set to $\mathbf{n}_\epsilon = 0.92$. The minimum number of points needed to form a shape is a function of the total number of points for each view $M_{min} = \frac{M_{total}}{200}$.

Table 6.1: Completeness scores for tolerance $\tau = 2cm$ and $\tau = 10cm$ for the benchmark high-resolution training set. Values from website. Higher scores are better. Best values in bold, second-best values are underlined. The proposed approach outperforms the other methods in several datasets.

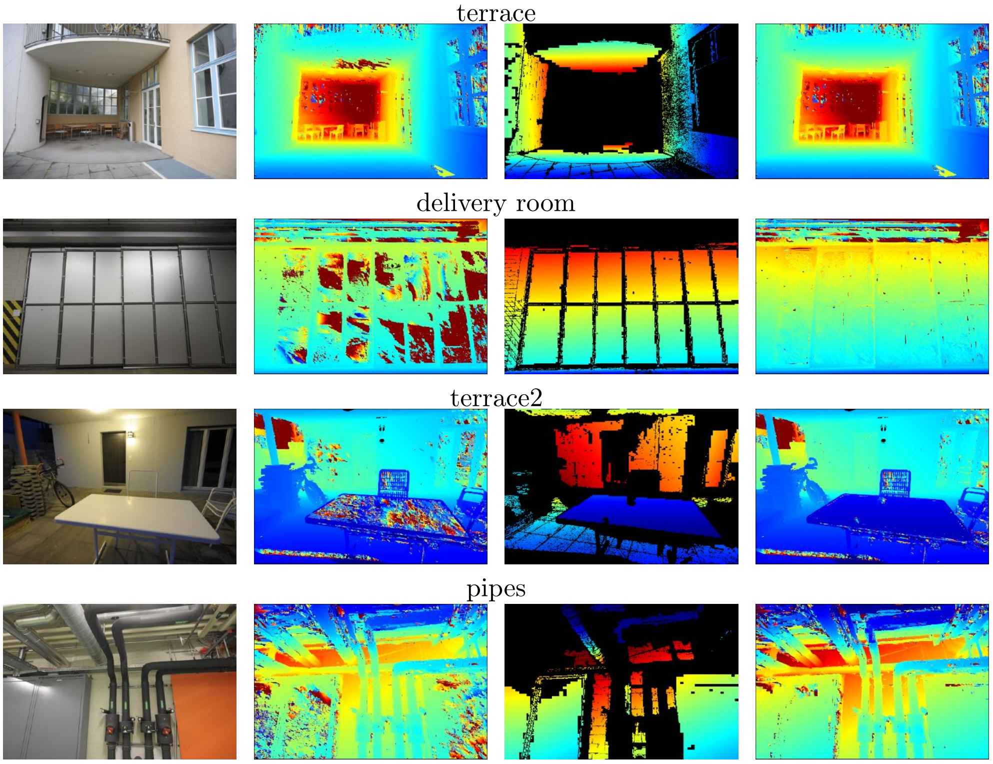
			indoor							outdoor					
	method	avg.	deliv.	kick.	off.	pipes	rel.	rel.2	terr.	court.	elec.	fac.	mead.	playgr.	terr.
2cm	COLMAP	55.13	67.21	47.83	31.6	34.24	63.88	62.52	63.02	73.47	62.13	55.82	34.39	44.9	75.67
	TAPA-MVS	71.45	80.78	**80.8**	60.91	63.8	74.59	72.42	80.97	77.04	73.81	62.1	51.28	67.98	82.37
	ACMP	72.15	80.87	69.29	**64.04**	53.54	73.74	75.27	88.98	80.96	77.93	64.13	60.79	63.91	84.45
	PCF-MVS	75.73	88.04	69.24	60.59	69.18	77.11	77.65	**94.01**	83.67	**81.47**	70.62	57.52	70.65	84.75
	MAR-MVS	77.19	90.58	77.62	59.24	67.98	82.37	**83.12**	87.67	89.87	79.33	65.56	58.33	**74.83**	86.95
	OpenMVS	74.92	90.38	66.09	46.33	64.55	79.92	80.88	89.2	90.1	78.01	66.35	**62.8**	71.77	87.52
	proposed	**77.51**	**92.00**	66.22	54.7	**71.79**	**82.46**	82.04	92.79	**91.52**	81.41	**70.63**	58.05	74.49	**89.46**
10cm	COLMAP	79.47	90.53	78.27	58.26	62.75	82.41	82.13	83.98	92.2	85.3	83.69	61.47	78.28	93.83
	TAPA-MVS	90.98	97.66	94.74	80.22	86.7	91.65	90.17	93.83	96.14	94.8	89.00	77.42	92.27	98.15
	ACMP	87.15	94.45	89.15	**86.35**	65.8	88.45	87.52	97.56	90.8	89.41	83.09	76.98	86.94	96.42
	PCF-MVS	90.42	98.18	84.55	75.84	88.47	90.37	90.66	98.88	94.44	94.35	**94.33**	76.61	91.29	97.46
	MAR-MVS	90.44	97.92	**95.07**	80.83	79.1	93.16	94.46	97.63	97.48	90.96	79.8	77.03	**94.68**	97.59
	OpenMVS	89.84	98.51	88.14	65.06	85.42	93.06	93.7	98.37	97.63	94.64	85.47	**78.39**	91.04	98.53
	proposed	**91.68**	**98.72**	90.86	76.07	**90.35**	**94.10**	**94.6**	**99.09**	**97.86**	**95.60**	85.72	77.14	92.69	**99.04**

6.3.3 Evaluation metrics

Following the established protocol by Schöps et al. [2017], accuracy, completeness, and F_1 score are used to evaluate the obtained 3D reconstruction results. The description of these metrics can be found in Section 5.3.3.

6.3.4 Evaluation on the ETH3D benchmark

The proposed method aims at improving the completeness and overall 3D reconstruction quality of challenging scenes with evident textureless surfaces. Thus, it is compared against state-of-the-art algorithms with similar principles to enable a direct and fair comparison. Our method uses relatively low computational cost using only CPU implementation, achieving competitive results. Ideally, an efficient MVS algorithm would achieve a high score in both accuracy and completeness, so F_1 score is a good approximation of the efficiency of the algorithm. The improvement in the depth estimates is shown in Figure 6.9; the depth maps deriving from the standard PatchMatch are compared with the resulting ones of our method after the integration of the priors in the cost calculation. It is evident that for the low textured areas such as the walls or the metallic/reflective surfaces like the table in the *terrace2* dataset, the closet in *pipes*, etc., the standard PatchMatch cannot propagate reliable estimates (Figure 6.9b). However, after the integration of the generated prior hypothesis in the cost calculation (Figure 6.9c), the final depth maps have been significantly improved in these problematic regions (Figure 6.9d). For the regions where no priors were generated, typically non-planar surfaces, the proposed approach behaves like the standard PatchMatch

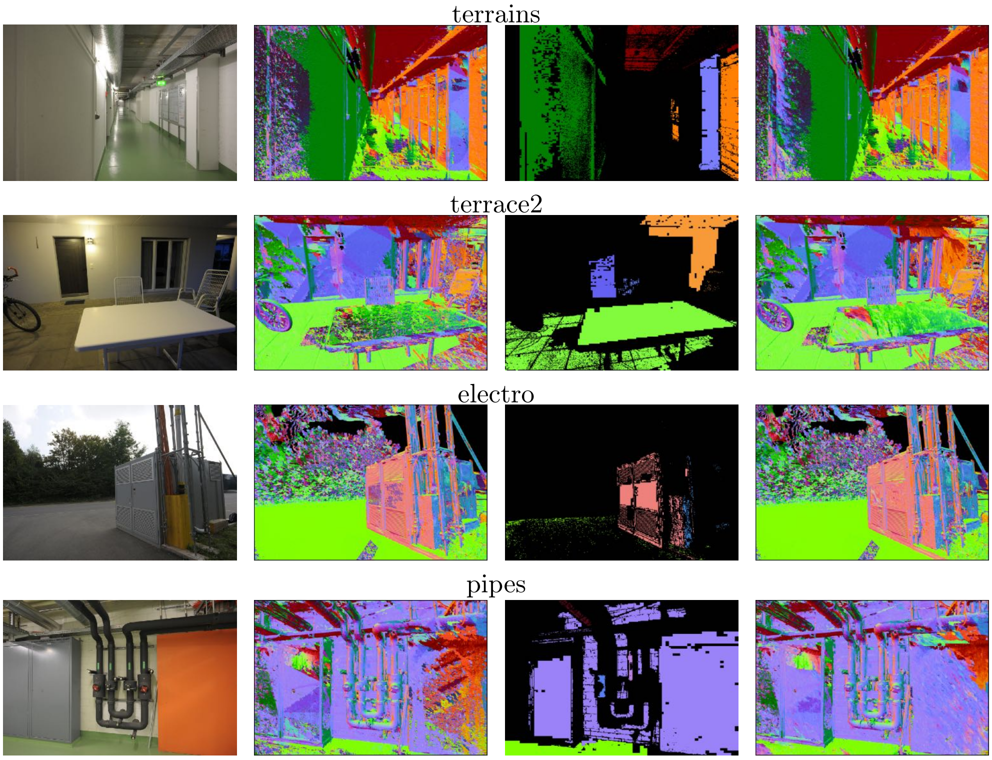


(a) RGB image (b) depth map (standard PatchMatch iterations) (c) depth priors (d) depth map (proposed approach)

Figure 6.9: **Qualitative depth map comparison of the baseline and the proposed method.** RGB images (6.9a), the intermediate depth maps after a couple of PatchMatch iterations (6.9b), the prior depth hypotheses (6.9c) and the improved depth maps using the proposed method (6.9d). Sequences from both training and test *ETH3D* sets.

method, relying only on the photometric cost.

To quantify the efficiency of the proposed method, the achieved completeness scores are compared to the other state-of-the-art methods, as reported in Table 6.1. It is observed that for $\tau = 2cm$ in many scenarios, the proposed method either outperforms other algorithms or comes second for a small margin (*delivery room, pipes, relief, courtyard, electro, facade, playground, terrace2*). The average completeness score considering both indoor and outdoor sequences ranks first, outperforming recent algorithms such as ACMP, MAR-MVS, or PCF-MVS. It is to be noted that PCF-MVS is a depth completion method, performing depth refinement as a post-processing step; on the contrary, the quadtree-guided method integrates the priors during depth estimation in an end-to-end way. It is worth underlying that for the majority of the sequences, the proposed approach significantly improves (average improvement of about 2.5%) the completeness scores of the baseline method OpenMVS, which were anyway among the highest. For $\tau = 10cm$,



(a) RGB image

(b) depth map (standard PatchMatch iterations)

(c) depth priors

(d) depth map (proposed approach)

Figure 6.10: **Qualitative normal map comparison of the baseline and the proposed method.** RGB images (6.10a), the intermediate normal maps after a couple of PatchMatch iterations (6.10b), the prior normal hypotheses (6.10c) and the improved normal maps using the proposed method (6.10d). Sequences from both training and test *ETH3D* sets.

the proposed method again achieves the best average score across all training sequences. Similarly, the F_1 score, reflecting both accuracy and completeness, is constantly superior to the baseline while achieving competitive results with respect to other methods across most datasets of the training set (Figure 6.11). This explicitly reflects the fact that the proposed method improves completeness without compromising accuracy and hence detail loss. The most challenging sequences for the quadtree-guided method are *kicker, office*, and *meadow*; the lower performance scores in these particular scenes seem to be in line with the general tendency of most state-of-the-art algorithms and the baseline method.

To further evaluate the robustness of the proposed approach, experiments are performed on the test set for which GT data are not publicly available; therefore, it does not allow for parameter tuning and thus prevents overfitting. Table 6.2 demonstrates that the quadtree-guided PatchMatch approach achieves the

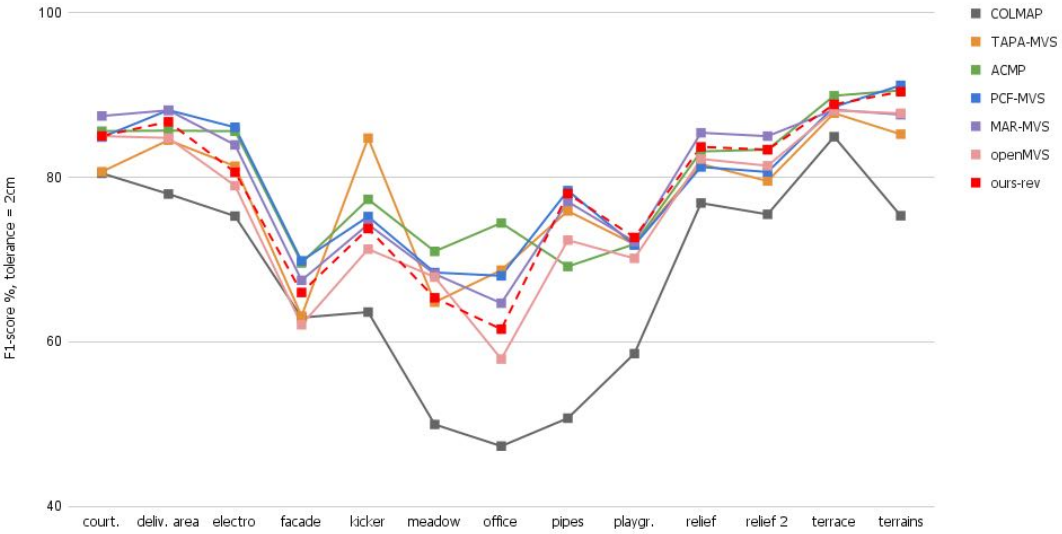


Figure 6.11: F_1 **scores for tolerance** $\tau = 2cm$ **for the high-resolution training datasets of the *ETH3D* benchmark.** The scores of the proposed approach is shown with a dash line.

second-best F_1 score for both indoor and outdoor scenarios. Also, in the test set, the completeness scores rank among the highest in the majority of the scenarios (Figure 6.12) while typically outperforming the baseline method. The presented approach is a non-data-driven, end-to-end pipeline; current data-driven methods for depth estimation in the MVS scenario perform plane-sweeps and construct global cost volumes [Yang et al., 2021b; Xu and Tao, 2020c] and thus struggle to handle high-resolution images and operate on lower resolutions. The proposed method, instead, can efficiently handle high-resolution images as it follows the PatchMatch algorithm and avoids the explicit construction of cost volumes. For a more complete evaluation, the results of the proposed method in comparison with recent deep learning approaches based on the PatchMatch algorithm [Lee et al., 2021; Wang et al., 2021] are briefly reported. In Table 6.3 the performance in accuracy, completeness, and F_1 scores are presented for both training and testing *ETH3D* datasets; the proposed method, working on higher resolution images, outperforms both learning-based approaches.

Figure 6.13 shows the point clouds of some sample sequences from the training and test sets generated by the quadtree-guided method and other state-of-the-art approaches. By visually comparing the results, it can be observed that the proposed method achieves to generate depth estimates on challenging areas of the scene while keeping the noise level low and yields visually pleasing results, even in the case that its scores are relatively lower than the other methods, i.e., for the *office* dataset. A good balance between noise level and completeness is particularly

Table 6.2: Accuracy, completeness and F_1 scores (%) for tolerance $\tau = 2cm$ for the *ETH3D* test benchmark. Values from *ETH3D* website. Best values in bold, second-best values are underlined.

	Method	Accuracy ↑	Completeness ↑	F_1 ↑
indoor	COLMAP	**91.95**	59.65	70.41
	TAPA-MVS	84.83	73.53	77.94
	ACMP	90.60	74.23	80.57
	PCF-MVS	80.92	77.63	78.84
	MAR-MVS	78.74	**83.43**	**80.70**
	OpenMVS	82.00	75.92	78.33
	proposed	82.59	77.39	79.50
outdoor	COLMAP	**92.04**	72.98	80.81
	TAPA-MVS	88.37	79.17	82.79
	ACMP	90.35	79.62	84.36
	PCF-MVS	85.84	84.29	85.01
	MAR-MVS	84.73	86.44	**85.27**
	OpenMVS	81.93	86.41	84.09
	proposed	82.58	**87.64**	85.03

important since most methods struggle in this regard. Indeed, an exceptionally performing MVS algorithm would achieve a good trade-off between completeness and accuracy, thus high F_1 scores. The proposed method yields more complete point clouds than the baseline OpenMVS, for example, in the whiteboard region of the *office* scene, the reflective closet in the *pipes* dataset, or the columns of the *old computer* room. At the same time, it typically generates fewer noisy points than MAR-MVS for most datasets. Compared to TAPA-MVS and ACMP, the visual results are similar in various scenes, for instance, *terrace2* and *lecture room*. In the presence of large, metallic surfaces, e.g., in *delivery room* and *pipes*, the quadtree-guided method yields more complete and visually appealing results.

6.3.5 Evaluation on custom datasets

Additionally, two scenes from real-world applications consisting of high-resolution images with sufficient overlap are considered. The performance of the quadtree-guided PatchMatch algorithm is compared with the baseline method [Cernea, 2020]. As shown in Table 6.4 for both scenes, the proposed approach outperforms the baseline method in F_1-score. The clear improvements of the proposed method in the problematic, textureless areas can be further studied in Figure 6.14; indeed, information gaps are filled in with reliably reconstructed points, resulting in point clouds with improved accuracy and completeness.

Table 6.3: Accuracy, completeness and F_1 score (%) comparison of the proposed approach with the recent PatchMatch-based deep learning methods PatchMatchNet [Wang et al., 2021] and PatchMatch-RL [Lee et al., 2021] for the *ETH3D* training and test sequences for $\tau = 2cm$. Note that the deep learning methods work on lower resolution images. Values from *ETH3D* website. Best values in bold.

			Indoor	Outdoor
	Method	Resolution	Acc./ Compl./F_1 ↑	Acc./ Compl./F_1 ↑
training	PatchMatchNet	2688×1792	63.74/67.71/64.65	66.06/62.78/63.69
	PatchMatch-RL	1920×1280	76.64/60.69/66.65	75.36/64.01/69.10
	proposed	3200×2132	**84.34/77.43/80.22**	74.87/**77.59/76.43**
testing	PatchMatchNet	2688×1792	68.83/74.63/71.33	72.33/85.96/78.49
	PatchMatch-RL	1920×1280	73.22/69.98/70.90	78.27/78.28/76.80
	proposed	3200×2132	**82.59/77.39/79.50**	**82.58/87.64/85.03**

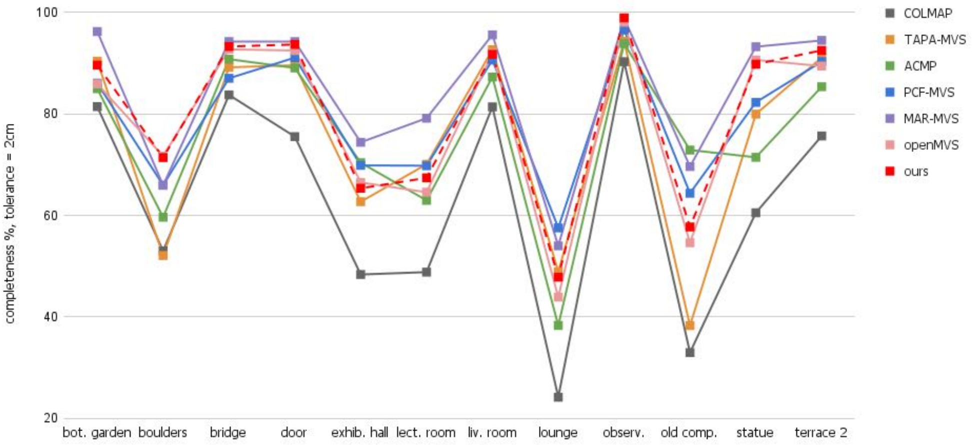
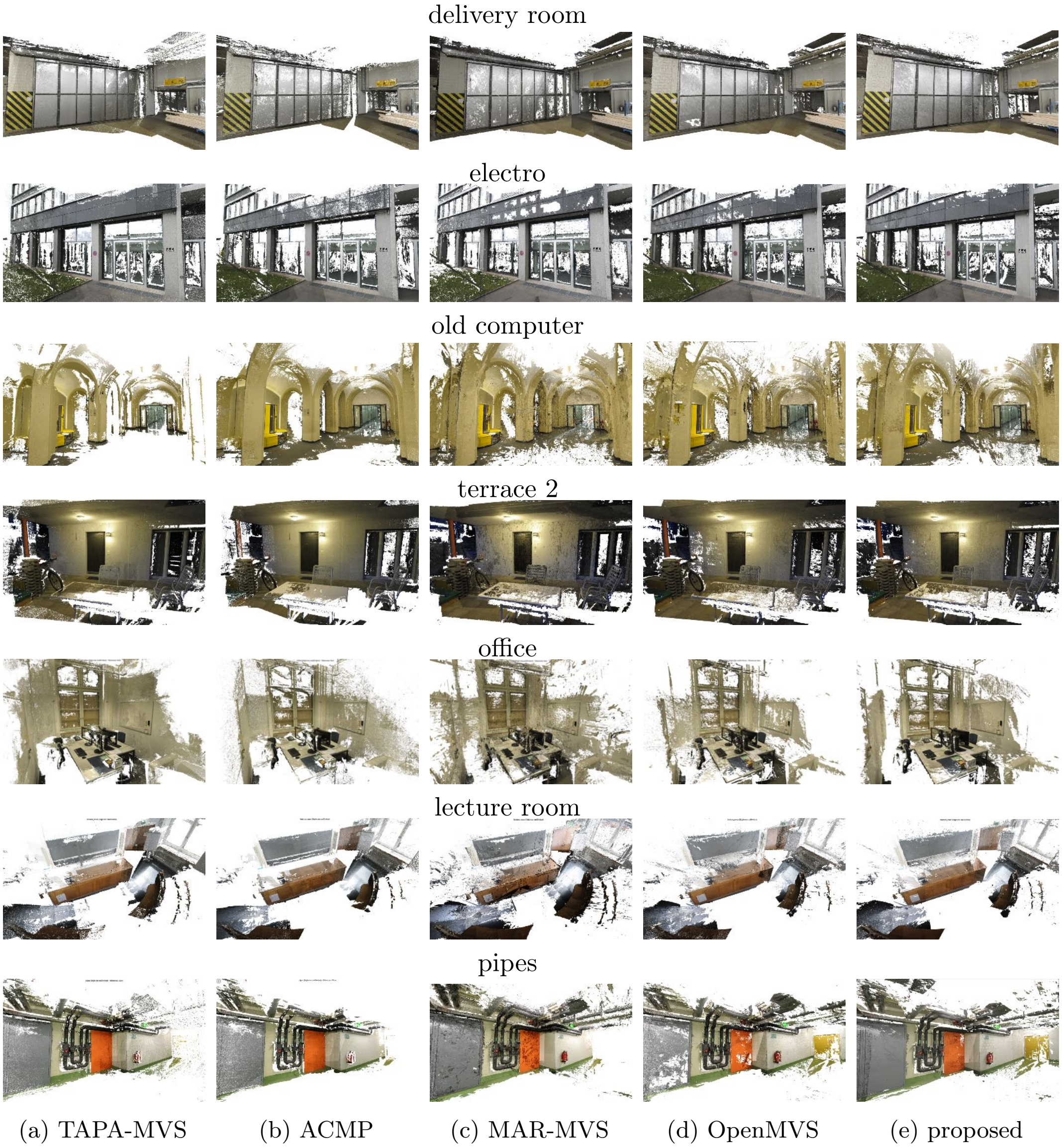


Figure 6.12: **Completeness scores (%) for tolerance $\tau = 2cm$ for the high-resolution test datasets of the *ETH3D* benchmark.** The proposed method is shown in dash red line.

Table 6.4: Performance scores (%) of the proposed approach with respect to the baseline method [Cernea, 2020] for the two custom scenes. Best values in bold.

		$\tau = 5cm$	$\tau = 10cm$
		F_1 ↑	F_1 ↑
House 1	OpenMVS	70.67	74.11
	proposed	**71.60**	**75.83**
House 2	OpenMVS	87.02	86.09
	proposed	**87.61**	**87.30**



(a) TAPA-MVS (b) ACMP (c) MAR-MVS (d) OpenMVS (e) proposed

Figure 6.13: **Qualitative point cloud comparison of our method with the state of the art baselines for the *ETH3D* training and test sets.** Dense models for the state of the art methods are as in *ETH3D* evaluation site.

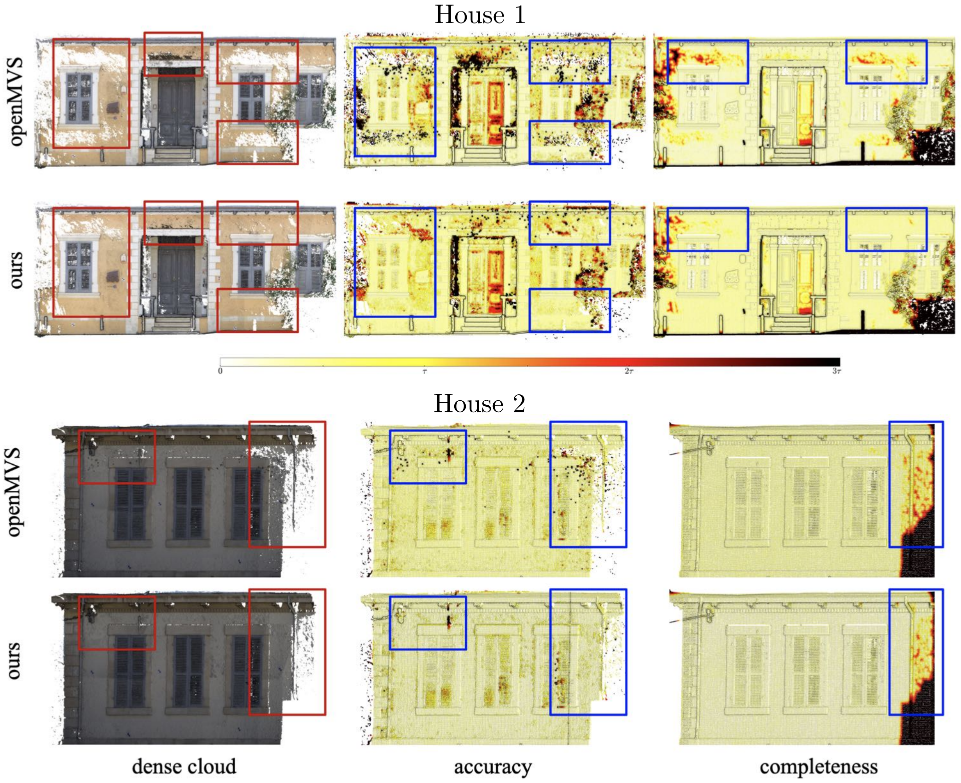


Figure 6.14: **Evaluation of the proposed approach with respect to the baseline method on custom datasets *House 1* and *House 2*.** From left to right: dense point clouds, accuracy and completeness for $\tau = 5cm$. The most evidently improved areas are highlighted in red and blue boxes.

6.4 Discussion

In this chapter, an extended PatchMatch-based MVS approach is presented to generate more complete 3D representations of indoor and outdoor scenarios using high-resolution images. Conventional MVS reconstruction algorithms rely solely on photometric consistency measures and thus fail to generate correct depth estimates in the presence of matching ambiguities, commonly occurring under non-Lambertian surfaces like textureless or highly reflective regions. Given that in man-made scenarios, textureless surfaces generally belong to planar structures, plane priors are generated to support depth reconstruction. Using adaptive quadtree blocks based on local texture, pixels with similar intensities are grouped together and considered to be part of the same local planar surface. The dominant planes of the scene are detected with a RANSAC-based sampling method on the initial rough 3D point clouds of each view. The final matching cost is calculated within an adaptive formulation in a way that the depth prior hypothesis is leveraged with the initial photometric cost, spreading more reliable depth estimates in the problematic areas. Experiments were performed on the high-resolution *ETH3D* benchmark dataset, a particularly challenging dataset featuring large textureless areas in indoor and outdoor scenes. The results prove the efficiency of the proposed method, outperforming in completeness other methods in various scenarios and achieving competitive reconstruction results both qualitatively and quantitatively. Additional experiments on custom scenes demonstrate the generalization ability of our method in real-world photogrammetric applications featuring such problematic areas. The approach is easily employable on standard machines, even for high-resolution images.

The proposed strategy is built upon the widely used open-source library OpenMVS [Cernea, 2020]; yet the presented ideas are easily transferable in other frameworks, extending thus their functionality and enabling further research. Experiments on high-resolution benchmark images demonstrate that the proposed method tackles the problem of incomplete reconstructions, achieving competitive completeness scores. Especially when large, textureless areas or reflective surfaces are present in the scene, it efficiently alleviates matching ambiguities, achieving high completeness scores in the final 3D point clouds, competitive with the state-of-the-art. In this section, the challenges and limitations of the presented method are discussed, and potential future work is outlined.

RANSAC performance. The quality of the plane hypotheses and thus the depth priors highly depend on the performance of the RANSAC algorithm. The random initialization of the algorithm may affect the generated hypotheses, although generally, it tends to converge similarly. RANSAC can be sensitive to the parameter setting, often implying fine-tuning for each case study for optimal results. Hence, in the experiments an adaptive parameter setting is proposed, relative to the average spacing s and total number of 3D points for each view, to compensate for the individual properties across the various sequences ($\epsilon = 1.5\bar{s}, c_\epsilon = \bar{s}, \mathbf{n}_\epsilon = 0.92, M = \frac{M_{total}}{250}$). For a fair comparison, the same parameters are kept for the complete *ETH3D* dataset (indoor and outdoor), as well as for the custom scenes, and were proven to work efficiently across different scenarios.

Quadtree guidance performance. Plane priors are propagated across the quadtree blocks based on their color similarity. The absolute distance in CIE-Lab color space and, in particular, the one between the values of the channels a and b have been chosen as a robust metric to propagate the planar prior hypothesis. Indeed, it was observed that the plane priors were propagated efficiently; yet, similar to the semantic PatchMatch method (Chapter 5), intense reflections would not be overcome, causing errors in prior plane propagation and ambiguous normal directions, especially under particularly large areas. During preliminary experiments, the Bhattacharyya distance of the histograms was also used as a similarity metric for plane propagation across the quadtree blocks, but its performance was found to be similar to the absolute distance of the values a and b while the time efficientcy decreased significantly (ca 5 times slower). Nonetheless, other perception-related similarity metrics can also be considered in the future for more robust plane propagation and illumination invariance.

Compared to the semantic PatchMatch, in the quadtree-based method each quadtree block is assigned directly to its corresponding 3D plane and no closest plane criterion needs to be used. That being said, the semantic guidance excludes in an early step the non-planar regions and potentially achieves more reliable RANSAC hypothesis, while also it efficiently restricts the plane expansion. On the other hand, the quadtree plane expansion relies on the block size to avoid infinite plane expansion and erroneous assignment of blocks with similar appearance that belong to different planes, a case that is often encountered over crease edges where inevitably small size blocks will be generated. As a matter of fact, the independence of the a-priori known semantic labels for the scene lead to a more generic method that can be applied across various domains.

Furthermore, the plane hypotheses generation is performed per single view; a future direction could be to work toward multi-view plane detection to enforce hypothesis consistency across the views and potentially generate more reliable planes.

Runtime performance. For a reference image of 3200 × 2132 pixels, the processing time for depth map generation is approximately 450 seconds in a single CPU thread running on 3.5GHz. Given that no GPU optimization is used, this performance is considered satisfying and feasible to be also implemented on low computational power devices, although further improvements are to be investigated.

Chapter 7

Conclusion and outlook

This final chapter concludes the work presented in this dissertation by first presenting a summary of the material discussed previously, followed by a discussion on the conclusions of the thesis; remarks and future outlook are also briefly presented. Finally, the overall research framework is introduced along with the most relevant publications of the author on the topics related to this dissertation.

7.1 Summary

This dissertation focuses on the image-based 3D reconstruction process and particularly on the multiple view stereo (MVS) part, during which dense 3D representations of the world are reconstructed. It proposes innovative methods to integrate prior scene cues in the reconstruction pipeline and confront the challenge of matching ambiguities that commonly occur in large non-Lambertian surfaces. The main goal and objectives of the thesis, as defined in Chapter 1, refer to the development of novel practical approaches toward this end. The proposed methodologies are integrated into a well-established, open-source framework to promote usability and reproducibility.

The 3D reconstruction pipeline can be roughly divided into two main parts as briefly explained in Chapter 1; Structure from Motion (SfM) where the camera poses and a sparse scene representation are obtained, and Multiple View Stereo (MVS), which aims to generate dense 3D representations of the scene. Therefore, this thesis focuses on depth estimation under the MVS scenario. Comprehensive theoretical background and relative literature review on depth estimation for both stereo and multi-view scenarios are provided (Chapter 2), whereas an in-depth survey of the PatchMatch-based methods for depth estimation is also presented (Chapter 3).

Nonetheless, geometric 3D reconstruction is closely related to scene understanding, another hot topic in the computer vision research field. Indeed, advanced scene prior cues can potentially support the efficiency of image-based 3D reconstruction and vice versa. Semantic reasoning directly in the 3D space is non-trivial mainly due to the limited availability of training data and the computational complexity;

on the contrary, algorithms for 2D semantic segmentation are mature enough to obtain robust results, and the existence of large-scale datasets facilitates the generalization of the trained models. However, there are few available large-scale and high-resolution benchmarks. This work introduces a new benchmark for semantic segmentation for historic building facades, *3DOM Semantic Facade*, acknowledging the lack of existing, high-resolution benchmarks for similar purposes. A straightforward pipeline for training is proposed, and the inference results are evaluated. Moreover, a new functionality is built upon the open-source MVS pipeline OpenMVS [Cernea, 2020] to enable label transfer from 2D to 3D, yielding semantically enhanced point clouds (Chapter 4).

Toward confronting the inevitable matching ambiguities in large, non-Lambertian surfaces, the obtained semantic maps in the 2D space can be leveraged into the MVS reconstruction as guidance for more reliable depth estimation. In particular, such advanced scene priors can indicate important cues for the 3D scene structure. For instance, parts of the scene with the label “wall” commonly imply textureless, planar structures. PatchMatch-based algorithms for depth estimation, although efficient and robust in the presence of slanted surfaces, highly depend on visual similarity measures that typically fail to reliably recover the depth due to matching ambiguities. A novel strategy is proposed to guide the depth propagation in such challenging surfaces; after a few standard PatchMatch iterations, RANSAC-based plane hypotheses in the 3D space are calculated from the rough scene reconstruction. Then, a novel, combined cost function is introduced to adaptively promote more reliable depth estimates across the image in the subsequent PatchMatch iterations (Chapter 5).

Semantic segmentation on the 2D images has recently been very popular, and a large variety of benchmark datasets exist for various applications, either indoor, outdoor, or airborne. However, in real-world applications it is non-trivial to obtain such semantic cues for every scene; a large amount of additional GT data may be required, and model training or fine-tuning is often a laborious task. Thus, an alternative, generic and domain-independent solution is also proposed, guided only by local textureness cues. Based on quadtree structures, groups of pixels with similar color attributes are grouped together. Similar to the previous method, planar hypotheses are extracted and guided by the quadtree blocks, assuming that such blocks, grouping together large areas of pixels of the same color, belong roughly to the same plane. The adaptive cost function is also used here to support PatchMatch propagation (Chapter 6).

7.2 Contributions and concluding remarks

Motivated by the current open challenges in the MVS reconstruction field, the main goal of this thesis was to answer the research question formed in the introduction (Chapter 1) regarding integrating advanced scene priors in the MVS process toward more complete and accurate 3D reconstruction. To this end, first the necessary theoretical background was comprehensively studied and the relative research was exhaustively surveyed. Subsequently, the open challenges were identified and new practical strategies were proposed. In this direction, several objectives were defined and successfully undertaken by the respective contributions as outlined in Section 1.3. The most important contributions along with concluding remarks on the results are outlined below:

Introduction of a novel benchmark for semantic segmentation of historic building facades and proposal of an efficient strategy for network training. The *3DOM semantic Facade*, a new high-resolution benchmark for facade segmentation for historic buildings, was proposed and released to the community; the benchmark consists of 227 images with respective GT segmentation masks and defining 5 classes: wall, window, door, sky and obstacle. It is hoped that this benchmark will further push the limits of research in semantic segmentation based on deep-learning techniques. To prove the effectiveness and generalization capability of the benchmark, a straightforward learning process was engineered based on a U-Net architecture [Ronneberger et al., 2015]. Practical insights on data preparation and hyperparameter tuning were given, which can also be useful in solving similar semantic segmentation problems. During inference, the segmented maps achieved high-score performance metrics as shown in Table 4.4 on the test set and show satisfying generalization on data of different distributions as shown in Figure 4.10.

Development a new module for robust label transfer and selective image-based 3D reconstruction integrated in a well-established and open-source MVS framework. Building upon the OpenMVS library [Cernea, 2020], a new module was developed that integrates the semantic information into the standard PatchMatch-based 3D reconstruction pipeline. The new functionality takes as input the corresponding semantic masks for every input image; the semantic information in inherited in the generated 3D point cloud via label transfer, directly yielding semantically enhanced 3D point clouds (Figure 4.13,4.14). At the same time, selective (class-specific) reconstruction is made possible based on the semantic label of each scene pixel; in this way, the user can reconstruct only the areas of interest of the scene according to the needs of each application. This functionality can be generalized in every MVS scenario for which semantic masks are available.

Proposal of an innovative method to exploit semantic scene cues to confront the matching ambiguities and improve the overall quality of the dense point cloud in challenging areas. Employ and integrate the proposed algorithm in a well-established and open-source MVS

framework. A novel approach was developed that leverages semantic priors into the MVS depth estimation; semantically derived planar priors are estimated and taken into consideration during matching cost calculation in the PatchMatch iterations. In this way, more reliable depth estimates are spread across the areas where the matching ambiguities occur, and the final dense point clouds achieve higher completeness. The new algorithm is integrated in the open-source library OpenMVS [Cernea, 2020] as an additional functionality. During the experiments on the *ETH3D* benchmark as well as on custom scenes, the proposed algorithm achieved constantly better results than the baseline method in completeness as shown in Table 5.1 and 5.2. Given the growing availability on semantically segmented data, this approach can be implemented in a variety of scenarios, indoor and outdoor.

Implementation of a non-data-driven, generalized approach based solely on local structure and texture information to undertake matching ambiguities and improve the overall quality of the dense point cloud in challenging areas. Employ and integrate the proposed algorithm in a well-established and open-source MVS framework. The semantic PatchMatch approach was further extended to provide robust solutions also in the absence of explicit semantic cues; instead, guidance based on local textureness information is proposed. Image pixels are organized into groups based on neighborhood and color similarity using quadtree structures. Quadtree blocks of similar color are assumed to lie approximately on the same plane. Planar priors in the 3D space are calculated and leveraged into the score function to promote more reliable depth estimates. Results on the entire training and test set of the *ETH3D* dataset demonstrate the effectiveness of the proposed approach show a clear improvement in completeness scores with respect to the baseline method as shown in Tables 6.1 and 6.2 and visually demonstrated in Figure 6.14. To further prove the applicability of the new method to different scenarios, two additional custom datasets were considered, on which, similar improvements were achieved (Table 6.4 and Figure 6.14).

7.3 Outlook

The work of this thesis considers the challenges in multiple view stereo reconstruction and proposes some novel methodologies for integrating advanced scene priors in the procedure. First, the generation of semantic scene priors in the 2D space is investigated using state-of-the-art deep learning approaches (Chapter 4) and proposing a novel benchmark dataset for facade segmentation of historic buildings. Such priors can be used to generate semantically enriched dense 3D point clouds (Chapter 4) or to undertake the matching ambiguities problem during depth estimation (Chapter 5) and obtain accurate, complete, and visually pleasing results. A second, more generalized approach was also proposed toward the same goal, based solely on local structure and texture information and being thus easily applicable also in case where no semantic masks are available (Chapter 6). During the latest few years of research for this thesis, several innovative ideas have been defined, employed and investigated, yet the strategies presented in this dissertation are the ones that have proven more effective and practical.

***3DOM Semantic Facade* benchmark.** Model training using the proposed benchmark has demonstrated robustness and generalization ability. However, a possible extension of the benchmark by adding more GT data and possibly expand the semantic classes to consider also, e.g., architectural details, pedestrians, cars, etc. would enable the generalization of the benchmark in a broader field of application scenarios.

Semantic segmentation strategy. A state-of-the-art network is used for semantic segmentation, obtaining satisfying results. Recently developed and upcoming architectures, could, however, achieve higher performance scores and should thus be investigated for further research.

Label transfer and selective reconstruction. An efficient, straightforward approach is presented built upon the well-established, open-source library OpenMVS Cernea [2020]. Following a similar strategy, other state-of-the-art frameworks may be considered for further research.

Matching ambiguities.

Plane hypotheses in the 3D space. In both proposed methods for treating matching ambiguities, plane hypotheses are estimated in the 3D space in a RANSAC-based fashion. First, the rough dense point clouds generated by the initial PatchMatch iterations are filtered to exclude outliers. Then, an Efficient RANSAC approach is followed [Schnabel et al., 2007], with adaptive parameters based on the local point density. The proposed adaptive strategy appears to work efficiently across diverse datasets and extracted planes are generally robust; however, noisy points may affect the quality of the resulting prior hypotheses. Moreover, plane detection in a multi-view formulation instead of independently for each view could potentially improve the quality of the resulting hypotheses.

PatchMatch propagation scheme. In the proposed approaches for PatchMatch depth estimation, a sequential propagation scheme is adopted following the baseline implementation of OpenMVS [Cernea, 2020]. Sequential PatchMatch schemes are simple to implement and have proven efficient enough for most applications. However, adaptive checkerboard patterns similar to those proposed in [Galliani et al., 2015; Xu and Tao, 2019; Zhou et al., 2021] can also be employed to achieve more robust propagation across larger regions.

Cost function. The proposed compound cost function consists of the basic photometric consistency term, a textureness coefficient, and the geometric prior coefficient; all these terms are combined adaptively to promote reliable depth and normal estimates and undertake the matching ambiguities in the regions where the photometric cost alone is not enough. This formulation has proven robust enough during the experiments, although particularly large and reflective surfaces cannot always be undertaken efficiently. The integration of additional terms enforcing geometric consistency can also be beneficial, yielding more reliable depth estimates and thus less noisy point clouds with higher performance in accuracy.

Depth filtering and completion. The presented strategies directly integrate scene priors in the PatchMatch depth estimation. After the first couple of PatchMatch iterations, plane priors are generated in the 3D space, and the depth and normal hypotheses for these planes are leveraged into a compound cost function in an adaptive way to confront the matching ambiguities problem in large textureless or reflective surfaces. Additional post-processing modules aiming for depth completion [Kuhn et al., 2019] or extra speckle filtering [Romanoni and Matteucci, 2019] could also be applied, potentially increasing the performance of 3D reconstruction in both accuracy and completeness.

Time efficiency.

Semantic segmentation. The proposed strategy for semantic segmentation on images is time efficient and scalable to the available computational resources, given that the backbone is EfficientNet [Tan and Le, 2019], a model specifically designed for efficiency scalability.

2D to 3D label transfer. The integrated module for label transfer from 2D to 3D and the generation of semantically enhanced point clouds adds practically no additional computational cost to the standard OpenMVS [Cernea, 2020] pipeline.

Leveraging scene priors in PatchMatch depth estimation. For the first presented approach relying on the semantic cues, given that the segmentation masks are calculated a priory, the additional computational cost refers mainly to the 3D plane hypotheses estimation. RANSAC is a quite efficient solution, yet the filtering in the 3D space inevitably adds some extra cost. Similarly, for the quadtree-assisted plane generation, the calculation of the quadtree structures is straightforward and efficient; the necessary 3D point filtering, though, is more time-consuming. For both approaches, a couple of extra PatchMatch iterations are implemented with the new cost function; yet, PatchMatch converges generally fast adding no severe cost in the procedure.

Deep learning methods. In recent years, the tremendous development of learning-based methods has brought new perspectives to the field of depth estimation and reconstruction. The interest of the research community has been shifted toward such methods; however, for real-world applications, especially in the photogrammetry domain where typically high accuracy requirements are set, learning-based methods have limited applicability mainly due to the lack of a large amount of GT data for training, their difficulty in generalization and the unbearable computational cost. As a matter of fact, current state-of-the-art learning methods typically employ severe downsampling and have limited performance in comparison with conventional ones (see also Table 6.3). In the future, considering the continuous increase of the computational power and the data availability, it will be more feasible to exploit the applicability of such learning-based methods in real-world scenarios.

7.4 Research overview

This doctoral dissertation is the result of years of efforts in research on the broader topic of 3D reconstruction in the fields of photogrammetry and computer vision. Various methods and algorithms for efficient image-based 3D reconstruction have been investigated, focusing on the main challenges and potential improvements, taking advantage of the recent advances in computer vision and machine learning. An in-depth study of the individual steps of the 3D reconstruction pipeline has been performed in the past years resulting in several pertinent publications. In more detail, during the early steps of this thesis, a survey on the available 3D recording techniques has been made [Georgopoulos and Stathopoulou, 2017]. Various capturing techniques combing different sensors and platforms have been investigated, i.e., using High Dynamic Range Images (HDR) over standard ones [Kontogianni et al., 2015; Suma et al., 2016] or experimenting with depth sensors on UAV platforms [Deris et al., 2017]. The efficiency of commercial software implementations [Stathopoulou et al., 2015; Georgopoulos et al., 2016] and open-source frameworks [Stathopoulou et al., 2019] in real-world 3D reconstruction scenarios has been exploited and experimentally evaluated. Moreover, exhaustive research has been conducted on the state-of-the-art feature detectors and descriptors in the concept of sparse image matching, experimenting with their applicability and efficiency over various datasets, e.g., terrestrial or fused airborne and terrestrial data [Stathopoulou et al., 2019; González-Aguilera et al., 2020] using open-source implementations. That being said, such open-source methods have been of particular interest for this research work, and the interchangeability between their file formats and standards has been investigated, aiming for a straightforward execution of experiments in a combined pipeline on a large scale [Stathopoulou et al., 2019]. At the same time, an extensive investigation of state-of-the-art surface reconstruction methods under photogrammetric scenarios has been done, considering the meshing procedure as a joint part of the image-based 3D reconstruction pipeline imposing visibility constraints or as an independent, subsequent step [Nocerino et al., 2020]. More recently, 3D edge extraction methods have been explored for enhancing edge details in mesh representations derived from MVS methods [Stathopoulou et al., 2021a]. Currently, deep learning algorithms for depth estimation and reconstruction are being investigated, considering monocular systems [Welponer et al., 2022]. These activities are considered relevant to the broader topic of 3D reconstruction and have contributed to narrowing down the problem and set the main objectives of the work, yet are not included directly in this dissertation.

Eventually, the author's research has mainly focused on depth estimation in the MVS scenario. Depth estimation and subsequent reconstruction, both in stereo and MVS, is a fundamental problem in computer vision and counts many decades of research since it is a core component of numerous applications. In the framework of this work, an exhaustive investigation of the existing methods has been made, and comparative experiments have been performed with well-known open-source 3D reconstruction frameworks and libraries, mainly the ones using PatchMatch-based approaches for depth estimation. PatchMatch has been proven to be more efficient in MVS reconstruction for practical applications than traditional Markov Random Field (MRF) optimization problems such as semi-global matching (SGM) and has, therefore, been preferred as a baseline for this study. Toward optimizing the depth estimation step and the 3D final output, the leverage of a priori known semantic information into the image-based 3D reconstruction pipeline has been investigated. In its initial steps, this research has elaborated on the generation of the a priori data (i.e., the semantic maps) using learning-based methods for semantic segmentation to improve the feature matching results and the generation of semantically enriched point clouds [Stathopoulou and Remondino, 2019a]. Later, the semantic priors have been used to enable the selective reconstruction of the desired regions and the exclusion of the unwanted ones [Stathopoulou and Remondino, 2019b]. Finally, an innovative framework has been introduced to support depth estimation under challenging scenarios, particularly large textureless and non-Lambertian surfaces, where commonly matching ambiguities occur. Using high-level scene semantics on the 2D images, geometric priors are generated to support depth estimation and reconstruction and promote correct depth hypothesis in such problematic regions. At the same time, semantic label transfer from 2D to 3D has also been enabled in the same integrated framework to generate semantically augmented 3D outputs Stathopoulou et al. [2021b]. Generalizing this idea to generate geometric prior information independently from the semantic information for the scene, a novel approach has been proposed based solely on the local textureness. Priors in the 3D space are guided by adaptive quadtree structures on the image to support depth estimation and reconstruction on large non-Lambertian areas of the scene Stathopoulou et al. [2022].

The main contributions and the content of this dissertation are primarily based on the material published and the expertise gained in the following peer-reviewed journal articles and conference proceedings:

Confronting matching ambiguities

Journal articles:

- Stathopoulou, E. K., Battisti, R., Cernea, D., Remondino, F., and Georgopoulos, A. Multi view stereo with quadtree-guided priors. *ISPRS Journal of Photogrammetry and Remote Sensing* (under review), 2022.
- Stathopoulou, E. K., Battisti, R., Cernea, D., Remondino, F., and Georgopoulos, A. Semantically derived geometric constraints for MVS reconstruction of textureless areas. *Remote Sensing*, 13(6):1053, 2021.

3D reconstruction with open-source frameworks

Journal articles:

- Nocerino, E., Stathopoulou, E. K., Rigon, S., and Remondino, F. Surface reconstruction assessment in photogrammetric applications. *Sensors*, 20(20):5863, 2020.

Conference proceedings:

- Stathopoulou, E. K., Rigon, S., Battisti, R., and Remondino, F. Enhancing geometric edge details in MVS reconstruction. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2021:391–398, 2021.
- Stathopoulou, E. K., Welponer, M., and Remondino, F. Open-source image-based 3D reconstruction pipelines: review, comparison and evaluation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W17:331–338, 2019.

Semantic segmentation, label transfer and selective reconstruction

Conference proceedings:

- Stathopoulou, E. K. and Remondino, F. Multi-view stereo with semantic priors. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W15:1135–1140, 2019.

- Stathopoulou, E. K. and Remondino, F. Semantic photogrammetry – boosting image-based 3D reconstruction with semantic labelling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9:685–690, 2019.

Apart from the aforementioned publications closely related to the topic of this dissertation, the author has a long publication record in the broader fields of photogrammetry and computer vision. The following list includes selected publications mostly relative to data acquisition, sensor fusion, 3D reconstruction and semantic segmentation:

Journal articles:

- Wang Y., James S., Stathopoulou E. K., Beltrán-González C., Konishi Y., del Bue A. Autonomous 3-D reconstruction, mapping, and exploration of indoor environments with a robotic arm. *IEEE Robotics and Automation Letters*, 4(4):3340-7, 2019.
- Suma R., Stavropoulou G., Stathopoulou E. K., Van Gool L., Georgopoulos A., Chalmers A. Evaluation of the effectiveness of HDR tone-mapping operators for photogrammetric applications. *Virtual Archaeology Review*, 7(15):54-66, 2016.

Conference proceedings

- Welponer, M., Stathopoulou, E. K., and Remondino, F. Monocular depth prediction in photogrammetric applications. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022:469–476, 2022.
- Remondino, F., Morelli, L., Stathopoulou, E. K., Elhashash, M., and Qin, R. Aerial triangulation with learning-based tie points, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022:77–84, 2022 (best poster paper award).
- Kyriakaki-Grammatikaki, S., Stathopoulou, E. K., Grilli, E., Remondino, F., and Georgopoulos, A. Geometric primitive extraction from semantically enriched point clouds, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVI-2/W1-2022:291–298, 2022.
- González-Aguilera, D., Ruiz de Oña, E., López-Fernandez, L., Farella, E. M., Stathopoulou, E. K., Toschi, I., Remondino, F., Rodríguez-Gonzálvez, P., Hernández-López, D., Fusiello, A., and Nex, F. PhotoMatch: An open-source multi-view and multi-modal feature matching tool for photogrammetric applications, *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B5-2020:213–219, 2020.

- Deris, A., Trigonis, I., Aravanis, A., and Stathopoulou, E. K. Depth cameras on UAVs: a first approach, *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W3:231–236, 2017.

- Stathopoulou, E. K., Georgopoulos, A., Panagiotopoulos, G., and Kaliampakos, D. Crowdsourcing Lost Cultural Heritage, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5/W3:295–300, 2015.

- Kontogianni, G., Stathopoulou, E. K., Georgopoulos, A., and Doulamis, A. HDR imaging for feature detection on detailed architectural scenes, *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5/W4:325–330, 2015.

Book chapters

- Georgopoulos, A. and Stathopoulou, E. K. Data acquisition for 3D geometric recording: state of the art and recent innovations. *Heritage and archaeology in the digital age*, pages 1–26, 2017.

Bibliography

Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., and Dahl, A. B. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.

Abdel-Aziz, Y. and Karara, H. Direct linear transformation into object space coordinates in close-range photogrammetry. In *Proc. Symp. Close-Range Photogrammetry*, pages 1–18, 1971.

Abraham, N. and Khan, N. M. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging*, pages 683–687. IEEE, 2019.

Abraham, S. and Förstner, W. Fish-eye-stereo calibration and epipolar rectification. *ISPRS Journal of photogrammetry and remote sensing*, 59(5):278–288, 2005.

Adorjan, M. *OpenSfM: A Collaborative Structure-From-Motion System.* PhD thesis, Wien, 2016.

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., and Szeliski, R. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

Agarwal, S., Mierle, K., et al. Ceres solver. 2012.

Alcantarilla, P. F. and Solutions, T. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7):1281–1298, 2011.

Armeni, I., Sax, S., Zamir, A. R., and Savarese, S. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Bailer, C., Finckh, M., and Lensch, H. Scale robust multi view stereo. In *European Conference on Computer Vision*, pages 398–411. Springer, 2012.

Bailer, C., Taetz, B., and Stricker, D. Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 4015–4023, 2015.

Banz, C., Pirsch, P., and Blume, H. Evaluation of penalty functions for semi-global matching cost aggregation. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 39, pages 1–6, 2012.

Bao, L., Yang, Q., and Jin, H. Fast edge-preserving patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3534–3541, 2014.

Baričević, D., Höllerer, T., Sen, P., and Turk, M. User-perspective augmented reality magic lens from gradients. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, pages 87–96, 2014.

Barnard, S. T. Stochastic stereo matching over scale. *International Journal of Computer Vision*, 3(1):17–32, 1989.

Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

Barnes, C., Shechtman, E., Goldman, D. B., and Finkelstein, A. The generalized patchmatch correspondence algorithm. In *European Conference on Computer Vision*, pages 29–43. Springer, 2010.

Bay, H., Tuytelaars, T., and Van Gool, L. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

Beardsley, P. A., Zisserman, A., and Murray, D. W. Sequential updating of projective and affine structure from motion. *International journal of computer vision*, 23(3): 235–259, 1997.

Besse, F., Rother, C., Fitzgibbon, A., and Kautz, J. Pmbp: Patchmatch belief propagation for correspondence field estimation. *International Journal of Computer Vision*, 110(1): 2–13, 2014.

Besse, F. O. *PatchMatch Belief Propagation for Correspondence Field Estimation and its Applications*. PhD thesis, University College London, 2013.

Bethmann, F. and Luhmann, T. Object-based multi-image semi-global matching-concept and first results. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(5):93, 2014.

Birchfield, S. and Tomasi, C. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4): 401–406, 1998.

Birchfield, S. and Tomasi, C. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, 1999.

Bischof, H., Schneider, W., and Pinz, A. J. Multispectral classification of landsat-images using neural networks. *IEEE transactions on Geoscience and Remote Sensing*, 30(3): 482–490, 1992.

Blaha, M., Vogel, C., Richard, A., Wegner, J. D., Pock, T., and Schindler, K. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3176–3184, 2016.

Blaha, M., Rothermel, M., Oswald, M. R., Sattler, T., Richard, A., Wegner, J. D., Pollefeys, M., and Schindler, K. Semantically informed multiview surface refinement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3819–3827, 2017.

Bleyer, M., Rother, C., and Kohli, P. Surface stereo with soft segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1570–1577. IEEE, 2010.

Bleyer, M., Rhemann, C., and Rother, C. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11): 1222–1239, 2001.

Brostow, G. J., Fauqueur, J., and Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

Brown, D. C. Decentering distortion and the definitive calibration of metric cameras. In *Annual Convention of the American Society of Photogrammetry, Washington DC*, volume 29, 1965.

Byeon, W., Breuel, T. M., Raue, F., and Liwicki, M. Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3547–3555, 2015.

Campbell, N. D., Vogiatzis, G., Hernández, C., and Cipolla, R. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008.

Can, G., Mantegazza, D., Abbate, G., Chappuis, S., and Giusti, A. Semantic segmentation on swiss3dcities: A benchmark study on aerial photogrammetric 3d pointcloud dataset. *Pattern Recognition Letters*, 150:108–114, 2021.

Cernea, D. OpenMVS: Multi-view stereo reconstruction library. 2020. URL https://cdcseacave.github.io/openMVS.

Chang, J.-R. and Chen, Y.-S. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

Chen, L.-C., Yang, Y., Wang, J., Xu, W., and Yuille, A. L. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Chen, R., Han, S., Xu, J., and Su, H. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019.

Chen, W., Hou, J., Zhang, M., Xiong, Z., and Gao, H. Semantic stereo: integrating piecewise planar stereo with segmentation and classification. In *2014 4th IEEE International Conference on Information Science and Technology*, pages 200–204. IEEE, 2014.

Chen, Y., Wang, Y., Lu, P., Chen, Y., and Wang, G. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 347–359. Springer, 2018a.

Chen, Z., Sun, X., Wang, L., Yu, Y., and Huang, C. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 972–980, 2015.

Chen, Z., Badrinarayanan, V., Drozdov, G., and Rabinovich, A. Estimating depth from rgb and sparse sensing. In *Proceedings of the European Conference on Computer Vision*, pages 167–182, 2018b.

Cheng, G., Zhou, P., and Han, J. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016.

Cheng, J., Leng, C., Wu, J., Cui, H., and Lu, H. Fast and accurate image matching with cascade hashing for 3d reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–8, 2014.

Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L. E., Ramamoorthi, R., and Su, H. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

Cherabier, I., Schonberger, J. L., Oswald, M. R., Pollefeys, M., and Geiger, A. Learning priors for semantic 3d reconstruction. In *Proceedings of the European conference on computer vision (ECCV)*, pages 314–330, 2018.

Collins, R. T. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363. IEEE, 1996.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

Cornelis, N., Leibe, B., Cornelis, K., and Van Gool, L. 3d urban scene modeling integrating recognition and reconstruction. *International Journal of Computer Vision*, 78(2):121–141, 2008.

Cremers, D. and Kolev, K. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1161–1174, 2010.

Curless, B. and Levoy, M. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

Dai, J., He, K., and Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

Dai, J., He, K., and Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3150–3158, 2016.

Dai, Y., Zhu, Z., Rao, Z., and Li, B. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In *2019 International Conference on 3D Vision (3DV)*, pages 1–8. IEEE, 2019.

Das, G. A mathematical approach to problems in photogrammetry. *Empire Survey Review*, 10(73):131–137, 1949.

Delaunoy, A. and Pollefeys, M. Photometric bundle adjustment for dense multi-view 3d modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1493, 2014.

Delaunoy, A. and Prados, E. Gradient flows for optimizing triangular mesh-based surfaces: Applications to 3d reconstruction problems dealing with visibility. *International journal of computer vision*, 95(2):100–123, 2011.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Deris, A., Trigonis, I., Aravanis, A., and Stathopoulou, E. Depth cameras on uavs: A first approach. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:231, 2017.

DeTone, D., Malisiewicz, T., and Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

DeVries, T. and Taylor, G. W. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.

Dhanachandra, N., Manglem, K., and Chanu, Y. J. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.

Di Stefano, L., Marchionni, M., and Mattoccia, S. A fast area-based stereo matching algorithm. *Image and vision computing*, 22(12):983–1005, 2004.

Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., and Liu, X. Transmvsnet: Global context-aware multi-view stereo network with transformers. *arXiv preprint arXiv:2111.14600*, 2021.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

Duggal, S., Wang, S., Ma, W.-C., Hu, R., and Urtasun, R. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4384–4393, 2019.

Egnal, G. and Wildes, R. P. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE Transactions on pattern analysis and machine intelligence*, 24(8):1127–1133, 2002.

Eigen, D. and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Facciolo, G., De Franchis, C., and Meinhardt, E. Mgm: A significantly more global matching for stereovision. In *BMVC 2015*, 2015.

Faugeras, O. and Keriven, R. Complete dense stereovision using level set methods. In *European conference on computer vision*, pages 379–393. Springer, 1998.

Faugeras, O. and Keriven, R. *Variational principles, surface evolution, PDE's, level set methods and the stereo problem.* IEEE, 2002.

Faugeras, O. and Luong, Q.-T. *The geometry of multiple images: the laws that govern the formation of multiple images of a scene and some of their applications.* MIT press, 2001.

Faugeras, O., Hotz, B., Mathieu, H., Viéville, T., Zhang, Z., Fua, P., Théron, E., Moll, L., Berry, G., Vuillemin, J., et al. Real time correlation-based stereo: algorithm, implementations and applications. Technical report, Inria, 1993.

Felzenszwalb, P. F. and Huttenlocher, D. P. Efficient belief propagation for early vision. *International journal of computer vision*, 70(1):41–54, 2006.

Fischler, M. A. and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Fitzgibbon, A. W. and Zisserman, A. Automatic camera recovery for closed or open image sequences. In *European conference on computer vision*, pages 311–326. Springer, 1998.

Flynn, J., Neulander, I., Philbin, J., and Snavely, N. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016.

Förstner, W. and Wrobel, B. P. *Photogrammetric computer vision.* Springer, 2016.

Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., et al. Building rome on a cloudless day. In *European conference on computer vision*, pages 368–381. Springer, 2010.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Fritsch, J., Kuehnl, T., and Geiger, A. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems*, pages 1693–1700. IEEE, 2013.

Fröhlich, B., Rodner, E., and Denzler, J. A fast approach for pixelwise labeling of facade images. In *2010 20th International Conference on Pattern Recognition*, pages 3029–3032. IEEE, 2010.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

Fuhrmann, S. and Goesele, M. Fusion of depth maps with multiple scales. *ACM Transactions on Graphics (TOG)*, 30(6):1–8, 2011.

Fuhrmann, S., Langguth, F., and Goesele, M. Mve-a multi-view reconstruction environment. In *GCH*, pages 11–18. Citeseer, 2014.

Fukushima, K. and Miyake, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

Fulkerson, B., Vedaldi, A., and Soatto, S. Class segmentation and object localization with superpixel neighborhoods. In *2009 IEEE 12th international conference on computer vision*, pages 670–677. IEEE, 2009.

Furukawa, Y. and Hernández, C. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

Furukawa, Y. and Ponce, J. Carved visual hulls for image-based modeling. In *European Conference on Computer Vision*, pages 564–577. Springer, 2006.

Furukawa, Y. and Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. Manhattan-world stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429. IEEE, 2009.

Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. Towards internet-scale multi-view stereo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1434–1441. IEEE, 2010.

Fusiello, A. and Irsara, L. Quasi-euclidean uncalibrated epipolar rectification. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

Fusiello, A., Roberto, V., and Trucco, E. Efficient stereo with multiple windowing. In *Proceedings of IEEE Computer Society conference on computer vision and pattern recognition*, pages 858–863. IEEE, 1997.

Fusiello, A., Trucco, E., and Verri, A. A compact algorithm for rectification of stereo pairs. *Machine vision and applications*, 12(1):16–22, 2000.

Galliani, S., Lasinger, K., and Schindler, K. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015.

Galliani, S., Lasinger, K., and Schindler, K. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25(361-369):2, 2016.

Gallup, D., Frahm, J.-M., Mordohai, P., Yang, Q., and Pollefeys, M. Real-time plane-sweeping stereo with multiple sweeping directions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

Gallup, D., Frahm, J.-M., Mordohai, P., and Pollefeys, M. Variable baseline/resolution stereo. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

Gallup, D., Frahm, J.-M., and Pollefeys, M. Piecewise planar and non-planar stereo for urban scene reconstruction. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1418–1425. IEEE, 2010.

Garg, R., Bg, V. K., Carneiro, G., and Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016.

Gehrig, S. K., Eberli, F., and Meyer, T. A real-time low-power stereo vision engine using semi-global matching. In *International Conference on Computer Vision Systems*, pages 134–143. Springer, 2009.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

Georgopoulos, A. and Stathopoulou, E. K. Data acquisition for 3d geometric recording: state of the art and recent innovations. *Heritage and archaeology in the digital age*, pages 1–26, 2017.

Georgopoulos, A., Oikonomou, C., Adamopoulos, E., and Stathopoulou, E. Evaluating unmanned aerial platforms for cultural heritage large scale mapping. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41, 2016.

Gidaris, S. and Komodakis, N. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5248–5257, 2017.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Godard, C., Mac Aodha, O., and Brostow, G. J. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

Goesele, M., Curless, B., and Seitz, S. M. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2402–2409. IEEE, 2006.

Goesele, M., Snavely, N., Curless, B., Hoppe, H., and Seitz, S. M. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

González-Aguilera, D., Ruiz de Oña, E., López-Fernandez, L., Farella, E., Stathopoulou, E. K., Toschi, I., Remondino, F., Rodríguez-Gonzálvez, P., Hernández-López, D., Fusiello, A., et al. Photomatch: An open-source multi-view and multi-modal feature matching tool for photogrammetric applications. In *XXIV ISPRS Congress (2020 edition)*, volume 43, pages 213–219, 2020.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Grün, A. Adaptive least squares correlation: a powerful image matching technique. *South African Journal of Photogrammetry, Remote Sensing and Cartography*, 14(3):175–187, 1985.

Grün, A. and Baltsavias, E. P. Geometrically constrained multiphoto matching. *Photogrammetric engineering and remote sensing*, 54(5):633–641, 1988.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., and Tan, P. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

Guillemot, C. and Le Meur, O. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144, 2013.

Guney, F. and Geiger, A. Displets: Resolving stereo ambiguities using object knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4165–4175, 2015.

Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R. R., and Hu, S.-M. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

Guo, X., Yang, K., Yang, W., Wang, X., and Li, H. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019.

Gupta, S., Girshick, R., Arbeláez, P., and Malik, J. Learning rich features from rgb-d images for object detection and segmentation. In *European conference on computer vision*, pages 345–360. Springer, 2014.

Hackel, T., Wegner, J. D., and Schindler, K. Contour detection in unstructured 3d point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1610–1618, 2016.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K., and Pollefeys, M. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847*, 2017.

Hacohen, Y., Shechtman, E., and Lischinski, D. Deblurring by example using dense correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2384–2391, 2013.

Han, X., Leung, T., Jia, Y., Sukthankar, R., and Berg, A. C. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3279–3286, 2015.

Häne, C., Zach, C., Zeisl, B., and Pollefeys, M. A patch prior for dense 3d reconstruction in man-made environments. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 563–570. IEEE, 2012.

Häne, C., Zach, C., Cohen, A., Angst, R., and Pollefeys, M. Joint 3d scene reconstruction and class segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 97–104, 2013.

Häne, C., Heng, L., Lee, G. H., Sizov, A., and Pollefeys, M. Real-time direct dense matching on fisheye images using plane-sweeping stereo. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 57–64. IEEE, 2014.

Häne, C., Zach, C., Cohen, A., and Pollefeys, M. Dense semantic 3d reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1730–1743, 2016.

Hannah, M. J. *Computer matching of areas in stereo images.* Stanford University, 1974.

Hartley, R. and Zisserman, A. *Multiple view geometry in computer vision.* Cambridge university press, 2003.

Hartmann, W., Galliani, S., Havlena, M., Van Gool, L., and Schindler, K. Learned multi-patch similarity. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1586–1594, 2017.

He, K., Sun, J., and Tang, X. Guided image filtering. In *European conference on computer vision*, pages 1–14. Springer, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Heise, P., Klose, S., Jensen, B., and Knoll, A. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2360–2367, 2013.

Heise, P., Jensen, B., Klose, S., and Knoll, A. Variational patchmatch multiview reconstruction and refinement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–890, 2015.

Helava, U. Object-space least-squares correlation. In *(ACSM and American Society for Photogrammety and Remote Sensing, Annual Convention, Saint Louis, MO, Mar. 14-18, 1988) Photogrammetric Engineering and Remote Sensing,*, volume 54, pages 711–714, 1988.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Hensel, S., Goebbels, S., and Kada, M. Facade reconstruction for textured lod2 citygml models based on deep learning and mixed integer linear programming. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2019.

Hermann, S. and Klette, R. Iterative semi-global matching for robust driver assistance systems. In *Asian Conference on Computer Vision*, pages 465–478. Springer, 2012.

Hermann, S., Klette, R., and Destefanis, E. Inclusion of a second-order prior into semi-global matching. In *Pacific-Rim Symposium on Image and Video Technology*, pages 633–644. Springer, 2009.

Hernández, C., Vogiatzis, G., and Cipolla, R. Probabilistic visibility for multi-view stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

Hernández, C. E. and Schmitt, F. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

Hiep, V. H., Keriven, R., Labatut, P., and Pons, J.-P. Towards high-resolution large-scale multi-view stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1430–1437. IEEE, 2009.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Hirschmuller, H. Stereo vision in structured environments by consistent semi-global matching. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2386–2393. IEEE, 2006.

Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008.

Hirschmuller, H. and Scharstein, D. Evaluation of stereo matching costs on images with radiometric differences. *IEEE transactions on pattern analysis and machine intelligence*, 31(9):1582–1599, 2008.

Hirschmüller, H., Buder, M., and Ernst, I. Memory efficient semi-global matching. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3: 371–376, 2012.

Hong, L. and Chen, G. Segment-based stereo matching using graph cuts. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004.

Hornung, A. and Kobbelt, L. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 503–510. IEEE, 2006.

Hosni, A., Bleyer, M., Rhemann, C., Gelautz, M., and Rother, C. Real-time local stereo matching using guided image filtering. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011.

Hosni, A., Rhemann, C., Bleyer, M., Rother, C., and Gelautz, M. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2012.

Hosni, A., Bleyer, M., and Gelautz, M. Secrets of adaptive support weight techniques for local stereo matching. *Computer Vision and Image Understanding*, 117(6):620–632, 2013.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., and Markham, A. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4977–4987, 2021.

Hu, X. and Mordohai, P. A quantitative evaluation of confidence measures for stereo vision. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2121–2133, 2012.

Hu, Y., Song, R., and Li, Y. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5712, 2016.

Huang, B., Yi, H., Huang, C., He, Y., Liu, J., and Liu, X. m3vsnet: unsupervised multi-metric multi-view stereo network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3163–3167. IEEE, 2021.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Huang, P.-H., Matzen, K., Kopf, J., Ahuja, N., and Huang, J.-B. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., and Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In *2016 IEEE International Geoscience and Remote Sensing Symposium*, pages 1835–1838. IEEE, 2016.

Humenberger, M., Engelke, T., and Kubinger, W. A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 77–84. IEEE, 2010.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., and Yang, M.-H. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

Im, S., Jeon, H.-G., Lin, S., and Kweon, I.-S. Dpsnet: End-to-end deep plane sweep stereo. In *7th International Conference on Learning Representations*, 2019.

Imran, S., Long, Y., Liu, X., and Morris, D. Depth coefficients for depth completion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12438–12447. IEEE, 2019.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Ishikawa, H. and Geiger, D. Rethinking the prior model for stereo. In *European Conference on Computer Vision*, pages 526–537. Springer, 2006.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Jadon, S. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2020.

Jampani, V., Gadde, R., and Gehler, P. V. Efficient facade segmentation using auto-context. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 1038–1045. IEEE, 2015.

Jancosek, M. and Pajdla, T. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR 2011*, pages 3121–3128. IEEE, 2011.

Jancosek, M. and Pajdla, T. Exploiting visibility information in surface reconstruction to preserve weakly supported surfaces. *International scholarly research notices*, 2014, 2014.

Jason, J. Y., Harley, A. W., and Derpanis, K. G. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016.

Ji, M., Gall, J., Zheng, H., Liu, Y., and Fang, L. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017.

Ji, X., Henriques, J. F., and Vedaldi, A. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

Jie, Z., Wang, P., Ling, Y., Zhao, B., Wei, Y., Feng, J., and Liu, W. Left-right comparative recurrent model for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3838–3846, 2018.

Kalogerakis, E., Averkiou, M., Maji, S., and Chaudhuri, S. 3d shape segmentation with projective convolutional networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3779–3788, 2017.

Kanade, T. and Okutomi, M. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE transactions on pattern analysis and machine intelligence*, 16(9):920–932, 1994.

Kanade, T., Kano, H., Kimura, S., Yoshida, A., and Oda, K. Development of a video-rate stereo machine. In *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, volume 3, pages 95–100. IEEE, 1995.

Kang, S. B., Szeliski, R., and Chai, J. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2001.

Kar, A., Häne, C., and Malik, J. Learning a multi-view stereo machine. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 364–375, 2017.

Karimi, D. and Salcudean, S. E. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019.

Kazhdan, M. and Hoppe, H. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013.

Kazhdan, M., Bolitho, M., and Hoppe, H. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., and Bry, A. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., and Izadi, S. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–590, 2018.

Khot, T., Agrawal, S., Tulsiani, S., Mertz, C., Lucey, S., and Hebert, M. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019.

Kim, J., Kolmogorov, V., and Zabih, R. Visual correspondence using energy minimization and mutual information. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1033–1040. IEEE, 2003.

Kim, S., Kim, S., Min, D., and Sohn, K. Laf-net: Locally adaptive fusion networks for stereo confidence estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 205–214, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

Klaus, A., Sormann, M., and Karner, K. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *18th International Conference on Pattern Recognition*, volume 3, pages 15–18. IEEE, 2006.

Knapitsch, A., Park, J., Zhou, Q.-Y., and Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

Knobelreiter, P., Reinbacher, C., Shekhovtsov, A., and Pock, T. End-to-end training of hybrid cnn-crf models for stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2339–2348, 2017.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J. D., and Ledoux, H. The hessigheim 3d (h3d) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from uav lidar and multi-view-stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 1:100001, 2021.

Kolmogorov, V. and Zabih, R. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002.

Kontogianni, G., Stathopoulou, E., Georgopoulos, A., and Doulamis, A. Hdr imaging for feature detection on detailed architectural scenes. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2015.

Korc, F. and Förstner, W. etrims image database for interpreting images of man-made scenes. *Dept. of Photogrammetry, University of Bonn, Tech. Rep. TR-IGG-P-2009-01*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Kuhn, A., Hirschmüller, H., and Mayer, H. Multi-resolution range data fusion for multi-view stereo reconstruction. In *German Conference on Pattern Recognition*, pages 41–50. Springer, 2013.

Kuhn, A., Hirschmüller, H., Scharstein, D., and Mayer, H. A tv prior for high-quality scalable multi-view stereo reconstruction. *International Journal of Computer Vision*, 124(1):2–17, 2017.

Kuhn, A., Lin, S., and Erdler, O. Plane completion and filtering for multi-view stereo reconstruction. In *German Conference on Pattern Recognition*, pages 18–32. Springer, 2019.

Kuhn, A., Sormann, C., Rossi, M., Erdler, O., and Fraundorfer, F. Deepc-mvs: Deep confidence prediction for multi-view stereo reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 404–413. IEEE, 2020.

Kundu, A., Li, Y., Dellaert, F., Li, F., and Rehg, J. M. Joint semantic segmentation and 3d reconstruction from monocular video. In *European Conference on Computer Vision*, pages 703–718. Springer, 2014.

Kutulakos, K. N. and Seitz, S. M. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000.

Kuznietsov, Y., Stuckler, J., and Leibe, B. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017.

Labatut, P., Pons, J.-P., and Keriven, R. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

Ladickỳ, L., Sturgess, P., Alahari, K., Russell, C., and Torr, P. H. What, where and how many? combining object detectors and crfs. In *European conference on computer vision*, pages 424–437. Springer, 2010.

Ladickỳ, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., and Torr, P. H. Joint optimization for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision*, 100(2):122–133, 2012.

Laga, H., Jospin, L. V., Boussaid, F., and Bennamoun, M. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

Langguth, F., Sunkavalli, K., Hadap, S., and Goesele, M. Shading-aware multi-view stereo. In *European Conference on Computer Vision*, pages 469–485. Springer, 2016.

Laurentini, A. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, J. H., Han, M.-K., Ko, D. W., and Suh, I. H. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

Lee, J. Y., DeGol, J., Zou, C., and Hoiem, D. Patchmatch-rl: Deep mvs with pixelwise depth, normal, and visibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6158–6167, 2021.

Lepetit, V., Moreno-Noguer, F., and Fua, P. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166, 2009.

Lhuillier, M. and Quan, L. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):418–433, 2005.

Li, G. and Zucker, S. W. Differential geometric inference in surface stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):72–86, 2008.

Li, H., Xiong, P., An, J., and Wang, L. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

Li, J., Li, E., Chen, Y., Xu, L., and Zhang, Y. Bundled depth-map merging for multi-view stereo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2769–2776. IEEE, 2010.

Li, Y., Min, D., Brown, M. S., Do, M. N., and Lu, J. Spm-bp: Sped-up patchmatch belief propagation for continuous mrfs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4006–4014, 2015.

Li, Z. and Snavely, N. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

Li, Z., Wang, K., Zuo, W., Meng, D., and Zhang, L. Detail-preserving and content-aware variational multi-view stereo reconstruction. *IEEE Transactions on Image Processing*, 25(2):864–877, 2016.

Liang, X., Shen, X., Feng, J., Lin, L., and Yan, S. Semantic object parsing with graph lstm. In *European Conference on Computer Vision*, pages 125–143. Springer, 2016.

Lin, M. H. and Tomasi, C. Surfaces with occlusions from layered stereo. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017a.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017b.

Liu, F., Shen, C., Lin, G., and Reid, I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015a.

Liu, H., Zhang, J., Zhu, J., and Hoi, S. C. Deepfacade: A deep learning approach to facade parsing. IJCAI, 2017.

Liu, H., Tang, X., and Shen, S. Depth-map completion for large indoor scene reconstruction. *Pattern Recognition*, 99:107112, 2020a.

Liu, H., Xu, Y., Zhang, J., Zhu, J., Li, Y., and Hoi, S. C. Deepfacade: A deep learning approach to facade parsing with symmetric loss. *IEEE Transactions on Multimedia*, 22(12):3153–3165, 2020b.

Liu, S. and Cooper, D. B. Ray markov random fields for image-based 3d modeling: Model and efficient inference. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1530–1537. IEEE, 2010.

Liu, W., Rabinovich, A., and Berg, A. C. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015b.

Liu, Z., Tang, H., Lin, Y., and Han, S. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Locher, A., Perdoch, M., and Van Gool, L. Progressive prioritized multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3244–3252, 2016.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Longuet-Higgins, H. C. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981.

Loop, C. and Zhang, Z. Computing rectifying homographies for stereo vision. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 125–131. IEEE, 1999.

Lorensen, W. E. and Cline, H. E. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Luc, P., Couprie, C., Chintala, S., and Verbeek, J. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

Luo, K., Guan, T., Ju, L., Huang, H., and Luo, Y. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019.

Luo, K., Guan, T., Ju, L., Wang, Y., Chen, Z., and Luo, Y. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1590–1599, 2020.

Luo, W., Schwing, A. G., and Urtasun, R. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016.

Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., and Lin, L. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018.

Lyu, Y., Huang, X., and Zhang, Z. Learning to segment 3d point clouds in 2d image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12255–12264, 2020a.

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., and Yang, M. Y. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119, 2020b.

Ma, F. and Karaman, S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE international conference on robotics and automation*, pages 4796–4803. IEEE, 2018.

Ma, W., Ma, W., Xu, S., and Zha, H. Pyramid alknet for semantic parsing of building facade image. *IEEE Geoscience and Remote Sensing Letters*, 18(6):1009–1013, 2020.

Ma, X., Gong, Y., Wang, Q., Huang, J., Chen, L., and Yu, F. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5732–5740, 2021.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, 3:473–480, 2016.

Martinovic, A., Knopp, J., Riemenschneider, H., and Van Gool, L. 3d all the way: Semantic segmentation of urban scenes from start to end in 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2015.

Mathias, M., Martinović, A., and Van Gool, L. Atlas: A three-layered approach to facade parsing. *International Journal of Computer Vision*, 118(1):22–48, 2016.

Matthies, L., Kanade, T., and Szeliski, R. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–238, 1989.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

McCormac, J., Handa, A., Leutenegger, S., and Davison, A. J. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687, 2017.

McGlone, J. *Manual of Photogrammetry Fifth Edition, the American Society for Photogrammetry and Remote Sensing.* 2004.

Menze, M. and Geiger, A. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015.

Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nistér, D., and Pollefeys, M. Real-time visibility-based fusion of depth maps. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Michael, M., Salmen, J., Stallkamp, J., and Schlipsing, M. Real-time stereo vision: Optimizing semi-global matching. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 1197–1202. IEEE, 2013.

Min, D., Lu, J., and Do, M. N. A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy? In *2011 International Conference on Computer Vision*, pages 1567–1574. IEEE, 2011.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image segmentation using deep learning: A survey. *arXiv preprint arXiv:2001.05566*, 2020.

Mizukami, Y., Okada, K., Nomura, A., Nakanishi, S., and Tadamura, K. Sub-pixel disparity search for binocular stereo vision. In *Proceedings of the 21st International Conference on Pattern Recognition*, pages 364–367. IEEE, 2012.

Mostajabi, M., Yadollahpour, P., and Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3376–3385, 2015.

Moulon, P., Monasse, P., and Marlet, R. Adaptive structure from motion with a contrario model estimation. In *Asian Conference on Computer Vision*, pages 257–270. Springer, 2012.

Moulon, P., Monasse, P., and Marlet, R. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3248–3255, 2013.

Moulon, P., Monasse, P., Perrot, R., and Marlet, R. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016.

Mühlmann, K., Maier, D., Hesser, J., and Männer, R. Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision*, 47(1):79–88, 2002.

Muja, M. and Lowe, D. G. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009.

Müller, P., Zeng, G., Wonka, P., and Van Gool, L. Image-based procedural modeling of facades. *ACM Trans. Graph.*, 26(3):85, 2007.

Murtiyoso, A., Lhenry, C., Landes, T., Grussenmeyer, P., and Alby, E. Semantic segmentation for building façade 3d point cloud from 2d orthophoto images using transfer learning. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:201–206, 2021.

Murtiyoso, A., Pellis, E., Grussenmeyer, P., Landes, T., and Masiero, A. Towards semantic photogrammetry: Generating semantically rich point clouds from architectural close-range photogrammetry. *Sensors*, 22(3):966, 2022.

Nam, K. W., Park, J., Kim, I. Y., and Kim, K. G. Application of stereo-imaging technology to medical field. *Healthcare informatics research*, 18(3):158–163, 2012.

Neuhold, G., Ollmann, T., Rota Bulo, S., and Kontschieder, P. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011a.

Newcombe, R. A., Lovegrove, S. J., and Davison, A. J. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011b.

Ni, J., Li, Q., Liu, Y., and Zhou, Y. Second-order semi-global stereo matching algorithm based on slanted plane iterative optimization. *IEEE Access*, 6:61735–61747, 2018.

Niemeyer, J., Rottensteiner, F., and Soergel, U. Contextual classification of lidar data and building object detection in urban areas. *ISPRS journal of photogrammetry and remote sensing*, 87:152–165, 2014.

Nistér, D. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004.

Nister, D. and Stewenius, H. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2161–2168. IEEE, 2006.

Nocerino, E., Stathopoulou, E. K., Rigon, S., and Remondino, F. Surface reconstruction assessment in photogrammetric applications. *Sensors*, 20(20):5863, 2020.

Nock, R. and Nielsen, F. Statistical region merging. *IEEE Transactions on pattern analysis and machine intelligence*, 26(11):1452–1458, 2004.

Nozick, V. Multiple view image rectification. In *2011 1st International Symposium on Access Spaces (ISAS)*, pages 277–282. IEEE, 2011.

Ohta, Y. and Kanade, T. Stereo by intra-and inter-scanline search using dynamic programming. *IEEE Transactions on pattern analysis and machine intelligence*, (2): 139–154, 1985.

Okutomi, M. and Kanade, T. A multiple-baseline stereo. *IEEE Transactions on pattern analysis and machine intelligence*, 15(4):353–363, 1993.

Ono, Y., Trulls, E., Fua, P., and Yi, K. M. Lf-net: Learning local features from images. *Advances in neural information processing systems*, 31, 2018.

Otsu, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Özdemir, E., Remondino, F., and Golkar, A. An efficient and general framework for aerial point cloud classification in urban scenarios. *Remote Sensing*, 13(10):1985, 2021.

Pang, J., Sun, W., Ren, J. S., Yang, C., and Yan, Q. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895, 2017.

Papandreou, G., Chen, L.-C., Murphy, K. P., and Yuille, A. L. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

Parzen, E. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Paschalidou, D., Ulusoy, O., Schmitt, C., Van Gool, L., and Geiger, A. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2018.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

Pearl, J. Belief updating by network propagation. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference; Morgan Kaufmann Publisher: San Francisco, CA, USA*, pages 143–190, 1988.

Peng, R., Wang, R., Wang, Z., Lai, Y., and Wang, R. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, June 2022.

Perez, L. and Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Pierrot-Deseilligny, M. and Paparoditis, N. A multiresolution and optimization-based image matching approach: An application to surface reconstruction from spot5-hrs stereo imagery. *Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36(1/W41):1–5, 2006.

Pihur, V., Datta, S., and Datta, S. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13):1607–1615, 2007.

Plath, N., Toussaint, M., and Nakajima, S. Multi-class image segmentation using conditional random fields and global classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 817–824, 2009.

Poggi, M. and Mattoccia, S. Learning a general-purpose confidence measure based on o (1) features and a smarter aggregation strategy for semi global matching. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 509–518. IEEE, 2016.

Pollefeys, M. *Self-calibration and metric 3D reconstruction from uncalibrated image sequences*. PhD thesis, PhD thesis, ESAT-PSI, KU Leuven, 1999.

Pollefeys, M., Koch, R., and Van Gool, L. A simple and efficient rectification method for general motion. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 496–501. IEEE, 1999.

Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., et al. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2):143–167, 2008.

Pons, J.-P., Keriven, R., and Faugeras, O. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Quam, L. H. Hierarchical warp stereo. In *Readings in computer vision*, pages 80–86. Elsevier, 1987.

Rahmani, K., Huang, H., and Mayer, H. Facade segmentation with a structured random forest. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2017.

Ranftl, R., Gehrig, S., Pock, T., and Bischof, H. Pushing the limits of stereo using variational stereo estimation. In *2012 IEEE Intelligent Vehicles Symposium*, pages 401–407. IEEE, 2012.

Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Richardt, C., Orr, D., Davies, I., Criminisi, A., and Dodgson, N. A. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *European conference on Computer vision*, pages 510–523. Springer, 2010.

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D., and Bischof, H. Irregular lattices for complex shape grammar facade parsing. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1640–1647. IEEE, 2012.

Riemenschneider, H., Bódis-Szomorú, A., Weissenberg, J., and Van Gool, L. Learning where to classify in multi-view semantic segmentation. In *European Conference on Computer Vision*, pages 516–532. Springer, 2014.

Romanoni, A. and Matteucci, M. Tapa-mvs: Textureless-aware patchmatch multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10413–10422, 2019.

Romanoni, A. and Matteucci, M. Facetwise mesh refinement for multi-view stereo. In *2020 25th International Conference on Pattern Recognition*, pages 6794–6801. IEEE, 2021.

Romanoni, A., Ciccone, M., Visin, F., and Matteucci, M. Multi-view stereo with single-view semantic mesh refinement. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 706–715, 2017.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Rothermel, M., Wenzel, K., Fritsch, D., and Haala, N. Sure: Photogrammetric surface reconstruction from imagery. In *Proceedings LC3D Workshop, Berlin*, volume 8, 2012.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., and Breitkopf, U. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

Ruf, B., Weinmann, M., and Hinz, S. Fass-mvs–fast multi-view stereo with surface-aware semi-global matching from uav-borne monocular imagery. *arXiv preprint arXiv:2112.00821*, 2021.

Rupnik, E., Daakir, M., and Deseilligny, M. P. Micmac–a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards*, 2(1):1–9, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Salehi, S. S. M., Erdogmus, D., and Gholipour, A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

Samadi, M. and Othman, M. F. A new fast and robust stereo matching algorithm for robotic systems. In *The 9th International Conference on Computing and Information-Technology (IC2IT2013)*, pages 281–290. Springer, 2013.

Savinov, N., Häne, C., Ladicky, L., and Pollefeys, M. Semantic 3d reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5460–5469, 2016.

Saxena, A., Sun, M., and Ng, A. Y. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5): 824–840, 2008.

Scharstein, D. Matching images by comparing their gradient fields. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 572–575. IEEE, 1994.

Scharstein, D. and Pal, C. Learning conditional random fields for stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

Scharstein, D. and Szeliski, R. Stereo matching with nonlinear diffusion. *International journal of computer vision*, 28(2):155–174, 1998.

Scharstein, D. and Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., and Westling, P. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014.

Scharstein, D., Taniai, T., and Sinha, S. N. Semi-global stereo matching with surface orientation priors. In *2017 International Conference on 3D Vision (3DV)*, pages 215–224. IEEE, 2017.

Schmid, K., Tomic, T., Ruess, F., Hirschmüller, H., and Suppa, M. Stereo vision based indoor/outdoor navigation for flying robots. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3955–3962. IEEE, 2013.

Schmitz, M. and Mayer, H. A convolutional network for semantic facade segmentation and interpretation. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:709, 2016.

Schnabel, R., Wahl, R., and Klein, R. Efficient ransac for point-cloud shape detection. In *Computer graphics forum*, volume 26, pages 214–226. Wiley Online Library, 2007.

Schneider, L., Cordts, M., Rehfeld, T., Pfeiffer, D., Enzweiler, M., Franke, U., Pollefeys, M., and Roth, S. Semantic stixels: Depth is not enough. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 110–117. IEEE, 2016.

Schönberger, J. L. *Robust methods for accurate and efficient 3D modeling from unstructured imagery.* PhD thesis, ETH Zurich, 2018.

Schönberger, J. L. and Frahm, J.-M. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

Schönberger, J. L., Zheng, E., Frahm, J.-M., and Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

Schönberger, J. L., Sinha, S. N., and Pollefeys, M. Learning to fuse proposals from multiple scanline optimizations in semi-global matching. In *Proceedings of the European Conference on Computer Vision*, pages 739–755, 2018.

Schöps, T., Schonberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., and Geiger, A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017.

Seitz, S. M. and Dyer, C. R. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE computer society conference on computer vision and pattern recognition*, volume 1, pages 519–528. IEEE, 2006.

Seki, A. and Pollefeys, M. Sgm-nets: Semi-global matching with neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 231–240, 2017.

Shen, S. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE transactions on image processing*, 22(5):1901–1914, 2013.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Shotton, J., Johnson, M., and Cipolla, R. Semantic texton forests for image categorization and segmentation. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

Shotton, J., Winn, J., Rother, C., and Criminisi, A. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International journal of computer vision*, 81(1):2–23, 2009.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sinha, S., Steedly, D., and Szeliski, R. Piecewise planar stereo for image-based rendering. *ICCV*, 2009. 1881-1888.

Sinha, S. N., Mordohai, P., and Pollefeys, M. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Sinha, S. N., Scharstein, D., and Szeliski, R. Efficient high-resolution stereo matching using local plane sweeps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1589, 2014.

Smolyanskiy, N., Kamenev, A., and Birchfield, S. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1007–1015, 2018.

Snavely, N., Seitz, S. M., and Szeliski, R. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006.

Song, S., Lichtenberg, S. P., and Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

Sormann, C., Knöbelreiter, P., Kuhn, A., Rossi, M., Pock, T., and Fraundorfer, F. Bp-mvsnet: Belief-propagation-layers for multi-view-stereo. In *2020 International Conference on 3D Vision (3DV)*, pages 394–403. IEEE, 2020.

Spangenberg, R., Langner, T., and Rojas, R. Weighted semi-global matching and center-symmetric census transform for robust driver assistance. In *International Conference on Computer Analysis of Images and Patterns*, pages 34–41. Springer, 2013.

Spangenberg, R., Langner, T., Adfeldt, S., and Rojas, R. Large scale semi-global matching on the cpu. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 195–201. IEEE, 2014.

Stathopoulou, E. and Remondino, F. Semantic photogrammetry – boosting image-based 3d reconstruction with semantic labelling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9:685–690, 2019a.

Stathopoulou, E., Georgopoulos, A., Panagiotopoulos, G., and Kaliampakos, D. Crowdsourcing lost cultural heritage. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2, 2015.

Stathopoulou, E., Welponer, M., and Remondino, F. Open-source image-based 3d reconstruction pipelines: review, comparison and evaluation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W17:331–338, 2019.

Stathopoulou, E., Rigon, S., Battisti, R., and Remondino, F. Enhancing geometric edge details in mvs reconstruction. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2021:391–398, 2021a.

Stathopoulou, E. K. and Remondino, F. Multi-view stereo with semantic priors. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W15:1135–1140, 2019b.

Stathopoulou, E. K., Battisti, R., Cernea, D., Remondino, F., and Georgopoulos, A. Semantically derived geometric constraints for mvs reconstruction of textureless areas. *Remote Sensing*, 13(6), 2021b.

Stathopoulou, E. K., Battisti, R., Cernea, D., Remondino, F., and Georgopoulos, A. Multi view stereo with texture-guided priors. *ISPRS Journal of Photogrammetry and Remote Sensing (under review)*, 2022.

Stentoumis, C., Grammatikopoulos, L., Kalisperakis, I., and Karras, G. On accurate dense stereo-matching using a local adaptive multi-cost approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 91:29–49, 2014.

Strecha, C., Fransens, R., and Van Gool, L. Wide-baseline stereo from multiple views: a probabilistic account. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004.

Strecha, C., Fransens, R., and Van Gool, L. Combined depth and outlier estimation in multi-view stereo. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2394–2401. IEEE, 2006.

Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., and Thoennessen, U. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

Sugiura, T., Torii, A., and Okutomi, M. 3d surface extraction using incremental tetrahedra carving. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 692–699, 2013.

Suma, R., Stavropoulou, G., Stathopoulou, E. K., van Gool, L., Georgopoulos, A., and Chalmers, A. Evaluation of the effectiveness of hdr tone-mapping operators for photogrammetric applications. *Virtual Archaeology Review*, 7(15):54–66, 2016.

Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

Sun, J., Zheng, N.-N., and Shum, H.-Y. Stereo matching using belief propagation. *IEEE Transactions on pattern analysis and machine intelligence*, 25(7):787–800, 2003.

Sun, J., Li, Y., Kang, S. B., and Shum, H.-Y. Symmetric stereo matching for occlusion handling. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 399–406. IEEE, 2005.

Sun, S., Zheng, Y., Shi, X., Xu, Z., and Liu, Y. Phi-mvs: Plane hypothesis inference multi-view stereo for large-scale scene reconstruction. *arXiv preprint arXiv:2104.06165*, 2021.

Sweeney, C., Hollerer, T., and Turk, M. Theia: A fast and scalable structure-from-motion library. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 693–696, 2015a.

Sweeney, C., Sattler, T., Hollerer, T., Turk, M., and Pollefeys, M. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the IEEE international conference on computer vision*, pages 801–809, 2015b.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Szeliski, R. *Computer vision: algorithms and applications.* Springer Science & Business Media, 2010.

Szeliski, R. and Kang, S. B. Recovering 3d shape and motion from image streams using nonlinear least squares. *Journal of Visual Communication and Image Representation*, 5(1):10–28, 1994.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Tappen, M. F. and Freeman, W. T. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Computer Vision, IEEE International Conference on*, volume 3, pages 900–900. IEEE Computer Society, 2003.

Tchapmi, L., Choy, C., Armeni, I., Gwak, J., and Savarese, S. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017.

Teboul, O., Simon, L., Koutsourakis, P., and Paragios, N. Segmentation of building facades using procedural shape priors. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3105–3112. IEEE, 2010.

The CGAL Project. *CGAL User and Reference Manual.* CGAL Editorial Board, 5.3 edition, 2021. URL `https://doc.cgal.org/5.3/Manual/packages.html`.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., and Guibas, L. J. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.

Tieleman, T. and Hinton, G. Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning. Technical report, 2012.

Tola, E., Lepetit, V., and Fua, P. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32 (5):815–830, 2009.

Tola, E., Strecha, C., and Fua, P. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012.

Tomasi, C. and Manduchi, R. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998.

Tonioni, A., Poggi, M., Mattoccia, S., and Di Stefano, L. Unsupervised adaptation for deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1605–1613, 2017.

Tosi, F., Poggi, M., Benincasa, A., and Mattoccia, S. Beyond local reasoning for stereo confidence estimation with deep learning. In *Proceedings of the European Conference on Computer Vision*, pages 319–334, 2018.

Tosi, F., Aleotti, F., Poggi, M., and Mattoccia, S. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.

Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

Tyleček, R. and Šára, R. Spatial pattern templates for recognition of objects with regular structure. In *German conference on pattern recognition*, pages 364–374. Springer, 2013.

Ulusoy, A. O., Geiger, A., and Black, M. J. Towards probabilistic volumetric reconstruction using ray potentials. In *2015 International Conference on 3D Vision*, pages 10–18. IEEE, 2015.

Ulusoy, A. O., Black, M. J., and Geiger, A. Patches, planes and probabilities: A non-local prior for volumetric 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3280–3289, 2016.

Ulusoy, A. O., Black, M. J., and Geiger, A. Semantic multi-view stereo: Jointly estimating objects and voxels. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4531–4540. IEEE, 2017.

Uziel, R., Ronen, M., and Freifeld, O. Bayesian adaptive superpixel segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8470–8479, 2019.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., and Paragios, N. Building detection in very high resolution multispectral data with deep learning features. In *2015 IEEE international geoscience and remote sensing symposium*, pages 1873–1876. IEEE, 2015.

Van den Bergh, M., Boix, X., Roig, G., and Van Gool, L. Seeds: Superpixels extracted via energy-driven sampling. *International Journal of Computer Vision*, 111(3):298–314, 2015.

Veksler, O. Stereo correspondence by dynamic programming on a tree. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 384–390. IEEE, 2005.

Viola, P. and Wells III, W. M. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

Vogiatzis, G., Esteban, C. H., Torr, P. H., and Cipolla, R. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2241–2246, 2007.

Volpi, M., Tuia, D., Bovolo, F., Kanevski, M., and Bruzzone, L. Supervised change detection in vhr images using contextual information and support vector machines. *International Journal of Applied Earth Observation and Geoinformation*, 20:77–85, 2013.

Vu, H.-H., Labatut, P., Pons, J.-P., and Keriven, R. High accuracy and visibility-consistent dense multiview stereo. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):889–901, 2011.

Wang, D. C., Vagnucci, A. H., and Li, C. Gradient inverse weighted smoothing scheme and the evaluation of its performance. *Computer Graphics and Image Processing*, 15 (2):167–181, 1981.

Wang, F., Galliani, S., Vogel, C., Speciale, P., and Pollefeys, M. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021.

Wang, F., Galliani, S., Vogel, C., and Pollefeys, M. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8606–8615, 2022.

Wang, H., Rong, X., Yang, L., Feng, J., Xiao, J., and Tian, Y. Weakly supervised semantic segmentation in 3d graph-structured point clouds of wild scenes. *arXiv preprint arXiv:2004.12498*, 2020a.

Wang, Y., Ji, R., and Chang, S.-F. Label propagation from imagenet to 3d point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3135–3142, 2013.

Wang, Y., Guan, T., Chen, Z., Luo, Y., Luo, K., and Ju, L. Mesh-guided multi-view stereo with pyramid architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2039–2048, 2020b.

Watson, J., Firman, M., Brostow, G. J., and Turmukhambetov, D. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171, 2019.

Wei, J., Resch, B., and Lensch, H. P. Multi-view depth map estimation with cross-view consistency. In *BMVC*, 2014.

Wei, J., Lin, G., Yap, K.-H., Hung, T.-Y., and Xie, L. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4384–4393, 2020.

Wei, Y. and Quan, L. Asymmetrical occlusion handling using graph cut for multi-view stereo. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 902–909. IEEE, 2005.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Welponer, M., Stathopoulou, E.-K., and Remondino, F. Monocular depth prediction in photogrammetric applications. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:469–476, 2022.

Wenzel, K. *Dense image matching for close range photogrammetry.* PhD thesis, ifp, 2016.

Wenzel, K., Rothermel, M., Haala, N., and Fritsch, D. Sure–the ifp software for dense image matching. In *Photogrammetric week*, volume 13, pages 59–70, 2013.

Werner, D., Al-Hamadi, A., and Werner, P. Truncated signed distance function: experiments on voxel size. In *International Conference Image Analysis and Recognition*, pages 357–364. Springer, 2014.

Woodford, O., Torr, P., Reid, I., and Fitzgibbon, A. Global stereo reconstruction under second-order smoothness priors. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2115–2128, 2009.

Wu, C. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013.

Wu, C., Agarwal, S., Curless, B., and Seitz, S. M. Multicore bundle adjustment. In *CVPR 2011*, pages 3057–3064. IEEE, 2011.

Wu, T.-P., Yeung, S.-K., Jia, J., and Tang, C.-K. Quasi-dense 3d reconstruction using tensor-based multiview stereo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1482–1489. IEEE, 2010.

Xie, J., Girshick, R., and Farhadi, A. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European conference on computer vision*, pages 842–857. Springer, 2016.

Xu, D., Ricci, E., Ouyang, W., Wang, X., and Sebe, N. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5354–5362, 2017.

Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., and Ricci, E. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3917–3925, 2018.

Xu, H., Zhou, Z., Qiao, Y., Kang, W., and Wu, Q. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2, page 6, 2021a.

Xu, Q. and Tao, W. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019.

Xu, Q. and Tao, W. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020a.

Xu, Q. and Tao, W. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12516–12523, 2020b.

Xu, Q. and Tao, W. Pvsnet: Pixelwise visibility-aware multi-view stereo network. *arXiv preprint arXiv:2007.07714*, 2020c.

Xu, Q., Oswald, M. R., Tao, W., Pollefeys, M., and Cui, Z. Non-local recurrent regularization networks for multi-view stereo. *arXiv preprint arXiv:2110.06436*, 2021b.

Xu, Z., Liu, Y., Shi, X., Wang, Y., and Zheng, Y. Marmvs: Matching ambiguity reduced multiple view stereo for efficient large scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5981–5990, 2020.

Xue, Y., Chen, J., Wan, W., Huang, Y., Yu, C., Li, T., and Bao, J. Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4312–4321, 2019.

Yang, G., Zhao, H., Shi, J., Deng, Z., and Jia, J. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision*, pages 636–651, 2018.

Yang, J., Mao, W., Alvarez, J. M., and Liu, M. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

Yang, J., Alvarez, J. M., and Liu, M. Self-supervised learning of depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7526–7534, 2021a.

Yang, Q., Wang, L., Yang, R., Stewénius, H., and Nistér, D. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):492–504, 2008.

Yang, R. and Pollefeys, M. Multi-resolution real-time stereo on commodity graphics hardware. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

Yang, Y., Yuille, A., and Lu, J. Local, global, and multilevel stereo matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 274–279. IEEE, 1993.

Yang, Z., Ren, Z., Shan, Q., and Huang, Q. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. *arXiv preprint arXiv:2104.13325*, 2021b.

Yao, Y., Luo, Z., Li, S., Fang, T., and Quan, L. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., and Quan, L. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., and Quan, L. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on information theory*, 51(7):2282–2312, 2005.

Yi, H., Wei, Z., Ding, M., Zhang, R., Chen, Y., Wang, G., and Tai, Y.-W. Pyramid multi-view stereo net with self-adaptive view aggregation. In *European Conference on Computer Vision*, pages 766–782. Springer, 2020.

Yi-de, M., Qing, L., and Zhi-Bai, Q. Automated image segmentation using improved pcnn model based on cross-entropy. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, pages 743–746. IEEE, 2004.

Yin, W., Liu, Y., Shen, C., and Yan, Y. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019.

Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C., and Renyin, D. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., and Shen, C. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.

Yingze Bao, S., Chandraker, M., Lin, Y., and Savarese, S. Dense object reconstruction with semantic priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1264–1271, 2013.

Yoon, K.-J. and Kweon, I. S. Adaptive support-weight approach for correspondence search. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):650–656, 2006.

Yu, A., Guo, W., Liu, B., Chen, X., Wang, X., Cao, X., and Jiang, B. Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:448–460, 2021.

Yu, Z. and Gao, S. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1949–1958, 2020.

Zabih, R. and Woodfill, J. Non-parametric local transforms for computing visual correspondence. In *European conference on computer vision*, pages 151–158. Springer, 1994.

Zabulis, X. and Daniilidis, K. Multi-camera reconstruction based on surface normal estimation and best viewpoint selection. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 733–740. IEEE, 2004.

Zach, C. Fast and high quality fusion of depth maps. In *Proceedings of the international symposium on 3D data processing, visualization and transmission*, volume 1. Citeseer, 2008.

Zach, C., Pock, T., and Bischof, H. A globally optimal algorithm for robust tv-l 1 range image integration. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Zagoruyko, S. and Komodakis, N. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015.

Zbontar, J. and LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015.

Zbontar, J., LeCun, Y., et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Zhang, C., Wang, L., and Yang, R. Semantic segmentation of urban scenes using dense depth maps. In *European Conference on Computer Vision*, pages 708–721. Springer, 2010.

Zhang, F., Prisacariu, V., Yang, R., and Torr, P. H. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019a.

Zhang, H., Xu, K., Jiang, W., Lin, J., Cohen-Or, D., and Chen, B. Layered analysis of irregular facades via symmetry maximization. *ACM Trans. Graph.*, 32(4):121–1, 2013.

Zhang, X., Hu, Y., Wang, H., Cao, X., and Zhang, B. Long-range attention network for multi-view stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3782–3791, 2021.

Zhang, Y. and Funkhouser, T. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018.

Zhang, Y., Gong, M., and Yang, Y.-H. Local stereo matching with 3d adaptive cost aggregation for slanted surface modeling and sub-pixel accuracy. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

Zhang, Z. Determining the epipolar geometry and its uncertainty: A review. *International journal of computer vision*, 27(2):161–195, 1998.

Zhang, Z., Liu, Q., and Wang, Y. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

Zhang, Z., Vosselman, G., Gerke, M., Persello, C., Tuia, D., and Yang, M. Y. Detecting building changes between airborne laser scanning and photogrammetric data. *Remote sensing*, 11(20):2417, 2019b.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

Zheng, E., Dunn, E., Jojic, V., and Frahm, J.-M. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

Zhong, Y., Dai, Y., and Li, H. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.

Zhou, C., Zhang, H., Shen, X., and Jia, J. Unsupervised learning of stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1567–1575, 2017.

Zhou, K., Meng, X., and Cheng, B. Review of stereo matching algorithms based on deep learning. *Computational intelligence and neuroscience*, 2020, 2020.

Zhou, L., Zhang, Z., Jiang, H., Sun, H., Bao, H., and Zhang, G. Dp-mvs: Detail preserving multi-view surface reconstruction of large-scale scenes. *Remote Sensing*, 13 (22):4569, 2021.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

Zhu, J., Peng, B., Li, W., Shen, H., Zhang, Z., and Lei, J. Multi-view stereo with transformer. *arXiv preprint arXiv:2112.00336*, 2021.

Zhu, K., d'Angelo, P., and Butenuth, M. A performance study on different stereo matching costs using airborne image sequences and satellite images. In *ISPRS Conference on Photogrammetric Image Analysis*, pages 159–170. Springer, 2011.

Zhu, Z., Stamatopoulos, C., and Fraser, C. S. Accurate and occlusion-robust multi-view stereo. *ISPRS Journal of Photogrammetry and Remote Sensing*, 109:47–61, 2015.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

Zinner, C., Humenberger, M., Ambrosch, K., and Kubinger, W. An optimized software-based implementation of a census-based stereo matching algorithm. In *International Symposium on Visual Computing*, pages 216–227. Springer, 2008.

Zolanvari, S., Ruano, S., Rana, A., Cummins, A., da Silva, R. E., Rahbar, M., and Smolic, A. Dublincity: Annotated lidar point cloud and its applications. *arXiv preprint arXiv:1909.03613*, 2019.