National Technical University of Athens
School of Rural, Surveying & Geoinformatics Engineering
Remote Sensing Laboratory

Unlocking the Potential of Artificial Intelligence for Satellite Image Processing

Doctoral Dissertation

Viktoria Kristollari
Rural, Surveying & Geoinformatics Engineer N.T.U.A.

Supervisor:
Dr. Vassilia Karathanassi
Prof. NTUA

Athens, July 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων & Τοπογράφων Μηχανικών – Μηχανικών Γεωπληροφορικής
Εργαστήριο Τηλεπισκόπησης

Αξιοποίηση των Δυνατοτήτων της Τεχνητής Νοημοσύνης στην Επεξεργασία Δορυφορικών Εικόνων

Διδακτορική Διατριβή

Βικτωρία Κριστολλάρη
Αγρονόμος & Τοπογράφος Μηχανικός – Μηχανικός Γεωπληροφορικής ΕΜΠ

Επιβλέπουσα:
Δρ. Βασιλεία Καραθανάση
Καθηγήτρια ΕΜΠ

Αθήνα, Ιούλιος 2024

Examination Committee:

Dr. Vassilia Karathanassi Prof. NTUA	Dr. Dimitris Argialas Emeritus Prof. NTUA	Dr. Nikolaos Doulamis Prof. NTUA
Dr. Konstantinos Karantzalos Prof. NTUA, SG Ministry of Digital Governance	Dr. Maria Vakalopoulou Assist. Prof. CentraleSupélec	Dr. Ioannis Papoutsis Assist. Prof. NTUA
Dr. Athanasios Voulodimos Assist. Prof. NTUA		

Abstract

Artificial intelligence (AI) encompasses the execution of tasks typically associated with intelligent entities, carried out by machines. Advances in neuroscience since the late 19th century inspired the creation of the "perceptron" in 1958, which is a mathematical model of a biological neuron. Since then, artificial neural networks (ANNs), an AI method that is inspired by the human brain, have shown great progress in various tasks. The increased computational power provided in the last decade was among the main triggers of the field.

ANNs exhibit their potential mainly in "Big Data" tasks where they outperform other methods. Thus, research attention has been attracted to satellite Earth observation (EO) where large data volumes are frequently collected. The main positive points are independence from feature engineering, high flexibility, and spatial perception in image processing, while negative points are the time-consuming creation of annotations and the low interpretability. In this doctoral dissertation, the capabilities of ANNs were investigated in four EO applications: cloud masking in Sentinel-2 (S2) data, VHR change detection (CD), marine plastic litter detection through image fusion, and RGB-to-NIR image-to-image translation (ITIT).

Cloud masking is a crucial pre-processing step in EO data analysis because it excludes clouds from optical imagery. Threshold-based methods, which are still the golden rule in this task, exhibit difficulties in challenging cases which include the presence of thin clouds (omission error) and bright non-cloud objects (commission error). To mitigate the above-mentioned challenges, three studies were performed in this thesis on S2 data. In the first study, a multi-layer perceptron (MLP) architecture was implemented that yielded superior results compared to state-of-the-art rule-based and muti-temporal methods in the separation of clouds from deep water spectra with noise and sunglint. Directional reflectance effects were also considered. For the purpose of the study, a relevant manual dataset was created and publicly released since equivalent datasets do not exist in the literature. Interesting findings were the possibility of producing a positive effect when applying feature scaling by using the parameters of the test set instead of the training set, and the definition of the important bands in mitigating the spectra with noise and sunglint by employing the network weights. In the second study, a novel fine-tuning methodology for self-organizing maps (SOMs) was developed that successfully rectified the misclassified predictions of bright non-cloud spectra in land areas. SOMs are ANNs that carry topological properties. The proposed approach, applied to the output of the non-fine-tuned network, is task-independent and requires only small amounts of data. In addition, it eliminates the necessity for additional training. It is performed by pinpointing the neurons that correspond to the incorrectly predicted bright non-cloud objects and altering their labels. In the third study, a patch-to-pixel convolutional neural network (CNN) was created that effectively identified semi-transparent clouds and separated bright clouds from bright non-cloud objects. CNNs are ANNs that are inspired by the human vision. The model underwent evaluation on the first publicly available annotated cloud masking image dataset, which allowed for a robust and objective evaluation. The study further reinforced the value of CNNs in applications where spatial context is crucial and demonstrated that lightweight architectures can be successful in cloud masking.

CD in the context of EO is the task of monitoring land cover transitions through time. When performed in VHR data, it is possible to detect changes in smaller objects such as buildings. However, the complexity of the task increases because of heightened within-class variance and geometric registration errors. Traditional pixel and object-based methods are more successful in high/medium resolution data where residual misregistration is less important. Recently, convolutional deep learning (DL) has contributed to the VHR CD since CNNs inherently possess spatial context perception. However, the research has predominantly concentrated on images with minor co-registration errors and the evaluation is typically not conducted on highly dissimilar test datasets. In an attempt to reduce this research gap, in this thesis, a comparative study was conducted where several state-of-the-art DL CD methods and automatic co-registration methods were assessed on VHR images with severe co-registration errors. The images were collected from European areas with versatile urban patterns and the challenges included geometric distortions and radiometric differences, as well as seasonal and vehicle-related changes. The diversity between the training sets and the study data also posed a challenge for the supervised methods. The evaluation was reinforced by a novel proposed score that provides a better understanding of the magnitude of the commission error. It was shown that an FFT-based method that uses phase correlation produced the most satisfactory co-

registration results and a network called STANet outperformed the other methods in building-related changes. The STANet performance can be credited to the synergy between the spatial attention mechanism and a substantial annotated dataset.

Marine litter exerts a wide spectrum of both adverse environmental and socio-economic effects. Plastic, which is its dominant component, constitutes the most significant hazard. Recent research employing satellite sensors has shown promising results in the detection of large-sized marine debris, however, the field is still in its infancy. High spatial and spectral resolutions are two critical factors in enhancing the detection and discrimination ability. However, in the present satellite sensors there are critical trade-offs in this regard. In this thesis, under the assumption that this issue could be alleviated by image fusion, two studies were performed that focused on the increase of spatial resolution in either the PRISMA or the S2 satellites. In the first study, the potential of HS satellite imagery in marine plastic litter detection was investigated for the first time through PRISMA data. The research centered on identifying small-sized targets ($\leq$ 5 m) specifically designed for the experiment, adding an extra layer of complexity. Several literature conventional and state-of-the-art DL pansharpening approaches were evaluated. A PCA-based substitution method showed the best results as it efficiently separated plastic from water spectra without producing distortions on the output images. In the DL methods (originally introduced in the literature for VHR data), spatial distortions were encountered due to the large difference between the spatial resolutions of the PAN and the HS bands and the lack of ground-truth data. However, the importance of histogram clipping as a pre-processing step was established since random water spectra were effectively separated from the target spectra. In the absence of SWIR features, spectral VNIR characteristics were exploited and an intersection of the outputs of three novel marine plastic indexes was applied on the PCA image that detected plastics with size equal to 8% HS pixel coverage. In the second study, S2 and WV-3 images were fused since the SWIR information available in S2 and absent in WV-3 is valuable for the detection and identification of plastics. Several conventional and DL image fusion approaches were evaluated in terms of spatial and spectral accuracy on artificial plastic targets. CNMF showed the best performance overall and a GAN- and a ResNet-based model (created for the purpose of the study) outperformed all methods in terms of spectral similarity. Important findings were: a) the adequacy of VNIR WV-3 information in generating the most effective output, enhancing the chances of achieving temporally close acquisitions, b) the reinforcement of the significance of the SWIR information in detecting plastic, and c) the observation of dissimilarities in the spectral regions of S2 bands between the signatures of the various plastic materials. It is noted that the conventional image fusion methods in both studies were carried out by M. Kremezi.

Image-to-image translation (ITIT) refers to an image processing technique that aims to learn the mapping functions between an input and an output image and can be either performed in a paired (co-registered input and output) or an unpaired setting. Lately, the EO community has exhibited a heightened interest in DL paired ITIT by typically employing conditional GANs (cGANs) to synthesize missing information in several applications. RGB-to-NIR ITIT, where this thesis focused on, has been either addressed indirectly in the context of creating HS outputs (spectral super-resolution) or has been exclusively directed towards vegetation applications. Regarding unpaired ITIT, it has predominantly been utilized in VHR data as an intermediary step to improve the quality of cross-domain semantic segmentation (unsupervised domain adaptation). In this thesis, attempting to contribute to the RGB-to-NIR ITIT literature, a thorough study was performed that focused on three main land cover categories (impervious, vegetation, ground) on heterogeneous bi-temporal VHR images. Through a three-stage GAN framework, both paired and unpaired data were exploited and several network configurations were explored in order to predict satisfactory NIR outputs in in- and out-domain data (do not belong to the domain of the training set). The paired data experiments, which were run in an in- and out-domain setting, showed that cGANs produced adequate NIR predictions even in out-domain cases when the domain gap was not significantly high and that instance normalization performed better than batch normalization, especially on data with low representation on the training set. The unpaired data experiments managed to enhance the NIR prediction in the vegetation category when high dissimilarities existed in the respective RGB domains.

Περίληψη (Abstract in Greek)

Η τεχνητή νοημοσύνη αφορά την εκτέλεση εργασιών που τυπικά συνδέονται με έξυπνες οντότητες, από μηχανές. Οι εξελίξεις στη νευροεπιστήμη από τον 19ο αιώνα και μετά ενέπνευσαν τη δημιουργία του "perceptron" το 1958, το οποίο αποτελεί ένα μαθηματικό μοντέλο ενός βιολογικού νευρώνα. Από τότε, τα τεχνητά νευρωνικά δίκτυα (ΤΝΔ), μία μέθοδος τεχνητής νοημοσύνης που εμπνέεται από τον ανθρώπινο εγκέφαλο, έχουν επιδείξει μεγάλη πρόοδο σε διάφορες εργασίες. Η αυξημένη υπολογιστική ισχύς που παρέχεται την τελευταία δεκαετία αποτέλεσε έναν από τους κύριους παράγοντες ώθησης του τομέα.

Τα ΤΝΔ επιδεικνύουν τις δυνατότητές τους κυρίως σε εργασίες "Μεγάλων Δεδομένων", όπου υπερτερούν σε σχέση με άλλες μεθόδους. Επομένως, υψηλό είναι το ερευνητικό ενδιαφέρον στην επιστήμη της Δορυφορικής Τηλεπισκόπησης καθώς συχνά συλλέγονται μεγάλοι όγκοι δεδομένων. Τα κύρια πλεονεκτήματα είναι η ανεξαρτησία από τη χειρωνακτική εξαγωγή χαρακτηριστικών, η υψηλή ευελιξία και η χωρική αντίληψη στο πεδίο της επεξεργασίας εικόνων, ενώ μειονεκτήματα συνιστούν η χρονοβόρα δημιουργία επισημασμένων δεδομένων και η χαμηλή ερμηνευσιμότητα. Στην παρούσα διδακτορική διατριβή, διερευνήθηκαν οι δυνατότητες των ΤΝΔ σε τέσσερις εφαρμογές της Τηλεπισκόπησης: αφαίρεση νεφών σε Sentinel-2 (S2) δεδομένα, ανίχνευση μεταβολών σε εικόνες πολύ υψηλής ανάλυσης, ανίχνευση θαλάσσιων πλαστικών απορριμμάτων μέσω συγχώνευσης εικόνων και τέλος μετάφραση φυσικών έγχρωμων εικόνων (ΦΕΕ) σε εικόνες εγγύς υπέρυθρου (ΕΥ).

Η ανίχνευση νεφών (cloud masking) είναι ένα κρίσιμο βήμα προεπεξεργασίας στην ανάλυση τηλεπισκοπικών δεδομένων καθώς αποκλείει τα νέφη από τις οπτικές εικόνες. Οι μέθοδοι κατωφλίωσης, οι οποίες εξακολουθούν να είναι ο χρυσός κανόνας στην επίλυση του συγκεκριμένου προβλήματος, παρουσιάζουν μειωμένη απόδοση σε περιπτώσεις αυξημένης δυσκολίας όπως είναι η παρουσία ημιδιαφανών νεφών (σφάλμα παράλειψης) και φωτεινών μη-νεφωδών αντικειμένων (σφάλμα συμπερίληψης). Για να αντιμετωπιστούν οι παραπάνω προκλήσεις πραγματοποιήθηκαν τρεις μελέτες σε αυτήν τη διατριβή σε S2 εικόνες. Στην πρώτη μελέτη εφαρμόστηκε μία αρχιτεκτονική πολυεπίπεδων perceptron η οποία παρήγαγε καλύτερα αποτελέσματα σε σύγκριση με κοινά αποδεκτές μεθόδους κατωφλίωσης και πολυχρονικές μεθόδους όσον αφορά τη διάκριση νεφών από φασματικές υπογραφές (ΦΥ) βαθιάς θάλασσας με επιδράσεις θορύβου και ανάκλασης. Για τις ανάγκες της μελέτης δημιουργήθηκε ένα σετ δεδομένων το οποίο δημοσιεύθηκε προς ελεύθερη χρήση καθώς αντίστοιχα σετ δεδομένων δεν διατίθενται στη βιβλιογραφία. Ενδιαφέροντα ευρήματα αποτέλεσαν η πιθανή θετική επίδραση της εφαρμογής κανονικοποίησης στο σετ δοκιμής αντί του σετ εκπαίδευσης και ο καθορισμός των σημαντικών καναλιών στην αντιμετώπιση ΦΥ με επιδράσεις θορύβου και ανάκλασης μέσω της χρήσης βαρών του δικτύου. Στη δεύτερη μελέτη αναπτύχθηκε μία νέα μεθοδολογία ρύθμισης (fine-tuning) αυτό-οργανωμένων χαρτών (ΑΟΧ), η οποία διόρθωσε επιτυχώς τις εσφαλμένες ταξινομήσεις φωτεινών μη-νεφώδων ΦΥ στην ξηρά. Οι ΑΟΧ είναι ΤΝΔ τα οποία διαθέτουν τοπολογικές ιδιότητες. Η προτεινόμενη προσέγγιση, η οποία εφαρμόζεται στο αποτέλεσμα του μη ρυθμισμένου δικτύου, είναι ανεξάρτητη από την εκάστοτε εφαρμογή και απαιτεί μόνο μικρή ποσότητα δεδομένων. Επιπλέον, εξαλείφει την ανάγκη για περαιτέρω εκπαίδευση του δικτύου. Πραγματοποιείται με τον εντοπισμό των νευρώνων που αντιστοιχούν στα εσφαλμένα ταξινομημένα φωτεινά μη νεφώδη αντικείμενα και την τροποποίηση των επισημάνσεων τους. Στη τρίτη μελέτη δημιουργήθηκε ένα συνελικτικό νευρωνικό δίκτυο (ΣΝΔ) (patch-to-pixel) το οποίο εντόπισε τα ημιδιαφανή νέφη και διέκρινε τα φωτεινά νέφη από τα φωτεινά μη-νεφώδη αντικείμενα. Τα ΣΝΔ είναι ΤΝΔ τα οποία εμπνέονται από την ανθρώπινη όραση. Το μοντέλο υποβλήθηκε σε αξιολόγηση στο πρώτο δημόσια διαθέσιμο επισημασμένο σετ αφαίρεσης νεφών, επιτρέποντας την αξιόπιστη και αντικειμενική αξιολόγηση. Η μελέτη ενίσχυσε περαιτέρω την αξία των ΣΝΔ σε εφαρμογές όπου η χωρική πληροφορία είναι κρίσιμη και έδειξε ότι λιγότερο σύνθετα δίκτυα μπορούν να έχουν ικανοποιητική απόδοση στο πεδίο της αφαίρεσης νεφών.

Η ανίχνευση μεταβολών (ΑΜ) στην επιστήμη της Τηλεπισκόπησης αφορά την παρακολούθηση των καλύψεων γης μέσα στο χρόνο. Όταν εκτελείται σε δεδομένα πολύ υψηλής ανάλυσης (ΠΥΑ), είναι δυνατή η ΑΜ σε μικρότερα αντικείμενα, όπως κτίρια. Ωστόσο, η πολυπλοκότητα του προβλήματος αυξάνεται λόγω της έντονης διασποράς εντός κλάσεων και γεωμετρικών σφαλμάτων συνταύτισης. Οι παραδοσιακές μέθοδοι σύγκρισης εικονοστοιχείων ή κατάτμησης είναι πιο αποτελεσματικές σε δεδομένα υψηλής/μεσαίας ανάλυσης όπου τα

υπολειπόμενα σφάλματα συνταύτισης είναι λιγότερο σημαντικά. Πρόσφατα η συνελικτική βαθιά μηχανική μάθηση (ΒΜΜ) συνέβαλε στην ΑΜ σε εικόνες ΠΥΑ καθώς τα ΣΝΔ διαθέτουν έμφυτη αντίληψη της χωρικής πληροφορίας. Ωστόσο, η έρευνα έχει επικεντρωθεί κυρίως σε εικόνες με μικρά σφάλματα συνταύτισης και η αξιολόγηση συνήθως δεν διεξάγεται σε πολύ ανόμοια σετ δοκιμών. Σε απόπειρα μείωσης του συγκεκριμένου κενού στην έρευνα, στην παρούσα διατριβή διεξήχθη μία συγκριτική μελέτη, όπου διάφορες προηγμένες μέθοδοι ΒΜΜ που ανιχνεύουν μεταβολές και αυτόματες μέθοδοι συνταύτισης αξιολογήθηκαν σε εικόνες ΠΥΑ με σοβαρά σφάλματα συνταύτισης. Οι εικόνες συλλέχθηκαν από ευρωπαϊκές περιοχές με ποικίλα αστικά μοτίβα και οι προκλήσεις περιλάμβαναν γεωμετρικές παραμορφώσεις και ραδιομετρικές διαφορές, καθώς και μεταβολές συσχετισμένες με εποχές και κίνηση οχημάτων. Η διαφοροποίηση των σετ εκπαίδευσης από τα σετ μελέτης αποτέλεσε επίσης πρόκληση για τις επιβλεπόμενες μεθόδους. Η αξιολόγηση ενισχύθηκε από ένα νέο προτεινόμενο ποσοτικό δείκτη που βελτιώνει την αντίληψη του μεγέθους του σφάλματος συμπερίληψης. Η μελέτη έδειξε ότι μία μέθοδος που χρησιμοποιεί συσχέτιση φάσης παρήγαγε τα πιο ικανοποιητικά αποτελέσματα συνταύτισης και το δίκτυο STANet παρουσίασε την καλύτερη απόδοση όσον αφορά μεταβολές που σχετίζονται με κτίρια. Η απόδοση του δικτύου πιθανώς οφείλεται στη συνέργεια μεταξύ του μηχανισμού χωρικής προσοχής και του συνοδευτικού επισημασμένου σετ δεδομένων.

Τα θαλάσσια απορρίμματα προκαλούν ένα ευρύ φάσμα ανεπιθύμητων περιβαλλοντικών και κοινωνικοοικονομικών επιπτώσεων. Το πλαστικό, το οποίο αποτελεί το κυρίαρχο συστατικό, συνιστά το σημαντικότερο κίνδυνο. Πρόσφατες έρευνες οι οποίες χρησιμοποίησαν δορυφορικά δεδομένα έδειξαν ελπιδοφόρα αποτελέσματα στην ανίχνευση θαλάσσιων απορριμμάτων μεγάλου μεγέθους αλλά το συγκεκριμένο πεδίο είναι ακόμη στα πρώτα του βήματα. Η υψηλή χωρική και φασματική ανάλυση είναι δύο κρίσιμοι παράγοντες για τη βελτίωση της ικανότητας εντοπισμού και ταυτοποίησης των πλαστικών. Ωστόσο, στους παρόντες δορυφορικούς αισθητήρες υπάρχουν κρίσιμες συμβιβαστικές λύσεις όσον αφορά το συγκεκριμένο θέμα. Στην παρούσα διατριβή, υπό την υπόθεση ότι το προαναφερθέν πρόβλημα θα μπορούσε να αντιμετωπιστεί μέσω συγχώνευσης εικόνων, πραγματοποιήθηκαν δύο μελέτες οι οποίες επικεντρώθηκαν στην αύξηση της χωρικής ανάλυσης των δορυφόρων PRISMA και S2. Στην πρώτη μελέτη διερευνήθηκαν για πρώτη φορά οι δυνατότητες της υπερφασματικής (ΥΦ) δορυφορικής Τηλεπισκόπησης μέσω PRISMA δεδομένων στον εντοπισμό θαλάσσιων πλαστικών απορριμμάτων. Η έρευνα επικεντρώθηκε στον εντοπισμό στόχων μικρού μεγέθους ($\leq$ 5 m) που σχεδιάστηκαν αποκλειστικά για το πείραμα, αυξάνοντας τη δυσκολία του προβλήματος. Αξιολογήθηκαν διάφορες συμβατικές μέθοδοι καθώς και προηγμένα δίκτυα ΒΜΜ της βιβλιογραφίας με στόχο τη συγχώνευση του παγχρωματικού καναλιού με τα υπερφασματικά. Τα καλύτερα αποτελέσματα παρήχθησαν από μία συμβατική μέθοδο αντικατάστασης κύριων συνιστωσών, όπου διαχωρίστηκαν αποτελεσματικά οι ΦΥ του πλαστικού από του νερού χωρίς να προκαλούνται παραμορφώσεις στη συγχωνευμένη εικόνα. Στις μεθόδους ΒΜΜ (σημειώνεται ότι στην προγενέστερη βιβλιογραφία είχαν εφαρμοστεί σε εικόνες ΠΥΑ), χωρικές παραμορφώσεις εντοπίστηκαν στις συγχωνευμένες εικόνες λόγω της μεγάλης διαφοράς στις χωρικές αναλύσεις μεταξύ του παγχρωματικού και των ΥΦ καναλιών και της έλλειψης αληθών δεδομένων (ground-truth). Ωστόσο, η σημασία της αποκοπής του ιστογράμματος καθιερώθηκε, καθώς τυχαίες ΦΥ νερού διαχωρίστηκαν αποτελεσματικά από τις αντίστοιχες των στόχων πλαστικού. Λόγω απουσίας διακριτών χαρακτηριστικών στο μέσο υπέρυθρο (short-wave infrared (SWIR)), αξιοποιήθηκαν χαρακτηριστικά στο ορατό και εγγύς υπέρυθρο τμήμα του φάσματος και εφαρμόστηκε η τομή των αποτελεσμάτων τριών νέων δεικτών θαλάσσιων πλαστικών στη συγχωνευμένη εικόνα της μεθόδου των κυρίων συνιστωσών. Η ελάχιστη διάσταση ανιχνεύσιμου πλαστικού ήταν 8% του εικονοστοιχείου της ΥΦ εικόνας. Στη δεύτερη μελέτη πραγματοποιήθηκε συγχώνευση S2 και WV-3 εικόνων καθώς η πληροφορία του μέσου υπέρυθρου (διαθέσιμη στον S2 και απούσα στον WV-3) είναι πολύτιμη στον εντοπισμό και στην ταυτοποίηση των πλαστικών. Αξιολογήθηκαν διάφορες συμβατικές μέθοδοι συγχώνευσης καθώς και προηγμένα δίκτυα ΒΜΜ ως προς την ακρίβεια της απεικόνισης της χωρικής και φασματικής πληροφορίας τεχνητών στόχων πλαστικού. Η μέθοδος CNMF επέδειξε την καλύτερη συνολικά απόδοση, ενώ δύο μοντέλα βασισμένα σε ανταγωνιστική μάθηση (GANs) και υπολειπόμενες συνδέσεις αντίστοιχα (δημιουργήθηκαν για την πραγματοποίηση της μελέτης), ξεπέρασαν σε επιδόσεις όλες τις μεθόδους ως προς τη φασματική ομοιότητα. Σημαντικά ευρήματα ήταν: α) η επάρκεια της πληροφορίας στο εγγύς υπέρυθρο του WV-3 για την παραγωγή του καλύτερου συγχωνευμένου αποτελέσματος, βελτιώνοντας τις πιθανότητες επίτευξης χρονικά κοντινών λήψεων, β) η ενίσχυση της σημασίας της πληροφορίας

του μέσου υπέρυθρου στον εντοπισμό πλαστικών και γ) η παρατήρηση ανομοιοτήτων στις συγχωνευμένες ΦΥ των διάφορων πλαστικών υλικών. Σημειώνεται ότι οι συμβατικές μέθοδοι συγχώνευσης εκτελέστηκαν από τη Μ. Κρεμεζή.

Ο αγγλικός όρος image-to-image translation (ITIT) αναφέρεται σε μία τεχνική επεξεργασίας εικόνας που στοχεύει στην εκμάθηση των συναρτήσεων αντιστοίχισης μεταξύ μίας εικόνας εισόδου και μίας εικόνας εξόδου. Μπορεί να εκτελεστεί είτε σε paired δεδομένα (συνταύτιση εικόνας εισόδου και εξόδου) είτε σε unpaired. Το τελευταίο διάστημα, η κοινότητα της Τηλεπισκόπησης έχει εκδηλώσει αυξημένο ενδιαφέρον για το ITIT με paired δεδομένα, χρησιμοποιώντας συνήθως δίκτυα ανταγωνιστικής μάθησης (ΔΑΜ) υπό συνθήκη (conditional GANs) για να συνθέσουν την πληροφορία που λείπει σε διάφορες εφαρμογές. Η πρόβλεψη εικόνων ΕΥ από ΦΕΕ, στην οποία επικεντρώθηκε η παρούσα διατριβή, είτε έχει προσεγγιστεί έμμεσα στο πλαίσιο της δημιουργίας ΥΦ προϊόντων (spectral super-resolution), είτε έχει κατευθυνθεί αποκλειστικά σε εφαρμογές βλάστησης. Όσον αφορά το ITIT σε unpaired δεδομένα, έχει χρησιμοποιηθεί ως επί το πλείστον ως ενδιάμεσο βήμα για τη βελτίωση των αποτελεσμάτων της σημασιολογικής κατάτμησης μεταξύ διαφορετικών πεδίων (μη επιβλεπόμενη προσαρμογή πεδίου). Η παρούσα διατριβή επιχείρησε να συμβάλει στην έρευνα πρόβλεψης ΕΥ από ΦΕΕ, πραγματοποιώντας μία εμπεριστατωμένη μελέτη που επικεντρώθηκε σε τρεις κύριες κατηγορίες κάλυψης γης (μη διαπερατό, βλάστηση, έδαφος) σε ετερογενείς δίχρονες ΠΥΑ εικόνες. Μέσω μίας μεθοδολογίας τριών βημάτων με χρήση ΔΑΜ αξιοποιήθηκαν αντιστοιχιζόμενα και μη αντιστοιχιζόμενα δεδομένα, ενώ εξετάστηκαν και διαφορετικές παραλλαγές δικτύου, με στόχο την ικανοποιητική ΕΥ πρόβλεψη σε δεδομένα εντός και εκτός πεδίου (δεν ανήκουν στο πεδίο του σετ εκπαίδευσης). Τα πειράματα των paired δεδομένων, τα οποία εκτελέστηκαν σε δεδομένα εντός και εκτός πεδίου, έδειξαν ότι τα ΔΑΜ υπό συνθήκη παρήγαγαν επαρκείς προβλέψεις ΕΥ ακόμα και στις περιπτώσεις εκτός πεδίου, όταν οι ανομοιότητες των πεδίων (domain gap) δεν ήταν πολύ υψηλές. Επιπλέον προέκυψε ότι η κανονικοποίηση ανά περίπτωση (instance normalization) απόδωσε καλύτερα από την κανονικοποίηση ανά σύνολο (batch normalization), ιδιαίτερα σε δεδομένα με χαμηλή εκπροσώπηση στο σετ εκπαίδευσης. Στα πειράματα των unpaired δεδομένων κατέστη δυνατή η βελτίωση της πρόβλεψης του ΕΥ στην κατηγορία της βλάστησης σε περιπτώσεις υψηλών ανομοιοτήτων στα αντίστοιχα φυσικά έγχρωμα πεδία.

Acknowledgments

The research of this doctoral dissertation was conducted in the NTUA Remote Sensing Lab. I feel obliged to express my appreciation to all those that contributed in some unique way in its realization.

First and foremost, I would like to express my sincere gratitude to the supervisor of my doctoral dissertation, Prof. Vassilia Karathanassi for her constant interest and insightful feedback on my research, and for the exciting opportunities she provided in terms of international collaborations that significantly broadened my horizons.

I would also like to thank the two advisors of my doctoral dissertation, Emeritus Prof. Dimitris Argialas and Prof. Nikolaos Doulamis for their encouraging comments and useful advice in the reviewing stage, and the rest of the members of the examination committee (Prof. Konstantinos Karantzalos, Assist. Prof. Maria Vakalopoulou, Assist. Prof. Ioannis Papoutsis, Assist. Prof. Athanasios Voulodimos) for the constructive discussion conducted during the examination process and their positive feedback.

In continuation, I would like to acknowledge the NTUA Special Account for Research Funds for providing a scholarship to financially support my research, and it is also important to mention that this work has been supported by research projects of the NTUA Remote Sensing Lab that received funding from European Union's Horizon 2020 program (SEO-DWARF: no. 691071, HYPERION: no. 821054) and the European Space Agency (REACT: no. 4000131235/20/NL/GLC, 4000131040/20/NL/GLC).

My special thanks are extended to my colleague and dear friend Maria Kremezi for our excellent cooperation in the plastic litter detection application and for her support in general as a valuable friend. I am also grateful to other members of the NTUA Remote Sensing Lab (Dr. Kleanthis Karamvasis, Dr. Pol Kolokoussis, Maria Adepli, Milly Vassiliou, Dr. Vassilis Andronis) for the pleasant interaction and the willingness to assist.

Above all, I would like to thank my parents, Zorika and Foti, for imprinting in me their admiration for knowledge and the virtues of hard work, patience, and perseverance. I am also grateful to them for encouraging the pursuit of my goals even if that imposes geographical separation.

Last but not least, there are not enough words to express the positive impact that the presence of Antonios Liamis has exerted in every aspect of my life. His noble attitude and charismatic glow provided unwavering motivation, optimism and serenity throughout this journey. I feel incredibly fortunate to have crossed paths with him.

Dedicated to Antonios

Table of Contents

List of Abbreviations

Abbreviation	Definition	Abbreviation	Definition
ABI	Advanced Baseline Imager	MI	Meteorological Imager
ACCA	Automatic Cloud Cover Assessment	MLP	Multi-layer Perceptron
Adam	Adaptive Moment Estimation	MODIS	Moderate Resolution Imaging Spectroradiometer
AHI	Advanced Himawari Imager	MRA	Multiresolution Analysis
AI	Artificial Intelligence	MS	Multispectral
ALI	Advanced Land Imager	MSG	Meteosat Second Generation
ANN	Artificial Neural Network	MSE	Mean Squared Error
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer	MSI	Multispectral Instrument
AVHRR	Advanced Very-High-Resolution Radiometer	MTF	Modulation Transfer Function
BMU	Best Matching Unit	NDSI	Normalized Difference Snow Index
BN	Batch Normalization	NDVI	Normalized Difference Vegetation Index
BOW	Bag-of-Words	OBCD	Object-Based Change Detection
BRDF	Bidirectional Reflectance Distribution Function	PAN	Panchromatic
CAE	Convolutional Autoencoder	PET	Polyethylene Terephthalate
CC	Correlation Coefficient	PCA	Principal Component Analysis
CCA	Canonical Correlation Analysis	PP	Polypropylene
CVA	Change Vector Analysis	PS	Polystyrene
CD	Change Detection	PVC	Polyvinyl chloride
CDSS	Cross-domain Semantic Segmentation	RCAN	Residual Channel Attention Network
CNMF	Coupled Nonnegative Matrix Factorization	RMSE	Root Mean Squared Error
CNN	Convolutional Neural Network	ReLU	Rectified Linear Unit
COMS	Communication, Ocean and Meteorological Satellite	RBF	Radial Basis Function
CS	Component Substitution	SAD	Spectral Angle Distance
DL	Deep Learning	SAR	Synthetic Aperture Radar
DTM	Digital Terrain Model	SFIM	Smoothing Filter-based Intensity Modulation
DSM	Digital Surface Model	SISR	Single Image Super-resolution
EO	Earth Observation	SNR	Signal-to-Noise Ratio
EO-1	Earth Observing One	SeaWifs	Sea-Viewing Wide Field-of-View Sensor
EPS	Expanded Polystyrene	S2	Sentinel-2
EM	Expectation Maximization	SEVIRI	Spinning Enhanced Visible and Infrared Imager
FAI	Floating Algae Index	SIFT	Scale-Invariant Feature Transform
FFT	Fast Fourier Transformation	SNR	Signal-to-Noise Ratio
FCN	Fully Convolutional Network	SOM	Self-Organized Map
FDI	Floating Debris Index	SRF	Spectral Response Function
Fmask	Function of mask	SRGAN	Super-resolution GAN
GAN	Generative Adversarial Network	SSR	Spectral Super-resolution
GDD	Guided Deep Decoder	SVM	Support Vector Machine
GEMS	Geostationary Environment Monitoring Spectrometer	SVR	Support Vector Regression
GLP	Generalized Laplacian Pyramid	SWIR	Shortwave Infrared
GPU	Graphical Processing Unit	TIR	Thermal Infrared
GCOM-C	Global Change Observation Mission	TVR	Total Variation Regularization
GOES	Geostationary Operational Environmental Satellite	VIS	Visible
HDPE	High-density polyethylene	UAV	Unmanned Aerial Vehicle
HI	Hydrocarbon Index	UDA	Unsupervised Domain Adaptation
HPM	High Pass Modulation	VHR	Very High-Resolution
HS	Hyperspectral	VNIR	Visible-Near Infrared
IN	Instance Normalization	WV	WorldView
InSAR	Interferometric SAR	XAI	Explainable Artificial Intelligence
ITIT	Image-to-image Translation		
LDPE	Low-density Polyethylene		
Lidar	Light detection and ranging		
LMM	Local Mean Matching		
LMVM	Local Mean and Variance Matching		
MAD	Multivariate Alteration Detection		
MERIS	Medium Resolution Imaging Spectrometer		

CHAPTER 1

Introduction

The current PhD thesis investigates the capabilities of artificial neural networks (ANNs) in four Remote Sensing applications: cloud masking in Sentinel-2 (S2) data, VHR change detection (CD), marine plastic litter detection through image fusion and RGB-to-NIR image-to-image translation (ITIT). In section 1.1 a brief historical background of ANNs is outlined. In section 1.2 the current scientific challenges and motivations are described. In section 1.3 the main and specific objectives for each application are stated. Finally, in sections 1.4 and 1.5 the contributions and the PhD thesis outline are presented.

1.1 A brief history of ANNs

The term "Artificial Intelligence (AI)" originated in 1956 in a summer workshop in Dartmouth [1]. In the proposal, the idea that "every feature of intelligence can, in principle, be so precisely defined that a machine can simulate it" was introduced.

ANNs are an AI method which is inspired by the human brain. Thus, it naturally follows that the progress in the ANN field is interconnected with the progress in the field of neuroscience. Santiago Ramón y Cajal first formulated the theory of the individuality of the nerve cell by observations of brain tissue through a microscope in 1891 [2]. Till then, it was believed that the nervous system is a network of continuous fibers. He also deduced that signals enter the neuron through dendrites and exit through the axon, and that information transmission is performed through separation gaps called "synapses" (Figure 1.1). The invention of electroencephalography (EEG), an electrical activity recording method of the brain, by Hans Berger in 1924 [3], further advanced the neuroscience field. Significant was also the development of the Hodgkin–Huxley model in 1952 [4], a mathematical approximation of the electrical engineering properties of excitable cells such as neurons.

The first mathematical model of a biological neuron was the McCulloch- Pitts neuron proposed in 1943 [5], a precursor of the "perceptron" proposed by Frank Rosenblatt in 1958 [6], which is a more generalized computational model (Figure 2.5, section 2.2.2.2). However, criticism by the scientific community [7], mainly caused by the single layer property, and technological limitations led to very slow advances in the ANN field until 2012 when Alexnet [8], a network with 60 million weights (connections/synapses), won the ImageNet 2012 challenge by a large margin (9.8%) [9]. It is noted that the human brain has ~ 10^{14} synapses [10].

The main factors that cumulatively led to Alexnet's success over the years were: a) the stochastic gradient descent (SGD), first applied by Frank Rosenblatt [6], b) the introduction of convolutional neural networks (CNNs) in 1980 by Kunihiko Fukushima [11], c) the invention of backpropagation in 1986 [12], d) the release of CUDA

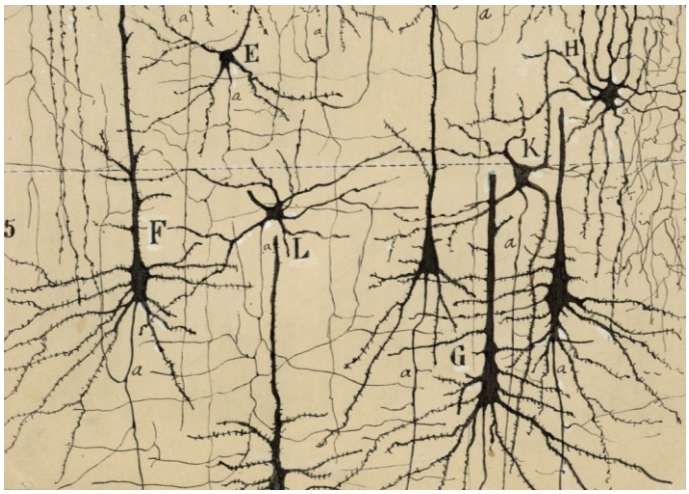


Figure 1.1. A drawing by Cajal depicting the nervous system. Cajal Legacy. Instituto Cajal (CSIC). Madrid (Spain)

[13] by Nvidia in 2006, an API that enables software access to highly-parallel GPU processing, and e) the creation of the ImageNet dataset [14] (12 million images) through a crowd-sourcing platform in 2009. CNNs are ANNs that are inspired by the human vision and several past studies have shown similarities in the hierarchical analysis of visual information [15].

1.2 Current scientific challenges and motivations

Remote Sensing is the acquisition of the physical properties of objects using radiation information. It is typically performed by satellite, airborne, or aerial sensors. Satellite Earth observation is characterized by the frequent collection of imagery by various satellite/sensor specifications, resulting in large volumes of complex, heterogeneous, and multi-modal data. Since Remote Sensing, "Big Data" [16], and image understanding are interconnected, ANNs with "deep" architecture (multiple layers) and mainly in convolutional form, have attracted the research attention. Among the positive points of the deep learning (DL) approaches are the independence from cumbersome feature engineering, the low sensitivity in noisy training data [17], and the flexibility in processing heterogenous information. However, the selection of the network hyperparameters and the training stage are time-consuming. In addition, DL methods perform better in the presence of large annotated training sets, which are difficult to create. A final negative point is the difficulty in interpreting the behavior of the networks. In the following sections, the particular challenges for each Remote Sensing application that was in the scope of this PhD thesis are discussed.

1.2.1 Cloud masking in Sentinel-2 (S2) data

Cloud masking is the process of excluding clouds from optical Remote Sensing imagery. It is an important pre-processing step required in every land and ocean study. Higher reflectance and lower brightness temperature of clouds compared to land cover are the common assumptions in current cloud detection rule-based methods where several thresholds (static or dynamic) are applied [18][19]. Threshold-based cloud detection is usually platform-specific and strongly linked to the geographical area and date of data collection [20][21]. The most well-known method in this category is Fmask [18][22] originally designed for Landsat data but it is also extensively applied in S2. Multi-temporal methods have also been proposed based on the idea that abrupt changes in image time series are mainly caused by the presence of clouds [23][24].

More recently, the cloud masking problem has been addressed by conventional machine learning [25][26] as well as ANN techniques [27][28][29]. The creation of large annotated datasets to boost the DL performance is still an ongoing process. The publicly available annotated datasets for the S2 satellite are shown in Table 1.1. Relevant studies have also been performed with self-organizing maps (SOMs) due to their high interpretative properties [30][31].

The main challenges in cloud masking are thin cloud omission and bright non-cloud object commission. Sunglint, high noise levels (random or periodic), and snow constitute bright non-cloud objects. Sunglint occurs mainly on the seawater and occurs when sunlight is reflected directly into the optical sensor [32][33]. Thermal information (not present in the S2 MS satellite) or the cirrus band are typically employed to detect high-level thin clouds [18][22], a process less demanding than low-level thin clouds [34][35][36].

To mitigate noise, researchers typically use spatial information and post-processing by taking advantage of textural properties and morphological operators [37][38][39]. Concerning sunglint, spectral, spatial, and geometric information is employed [40][41][42]. In addition, thermal bands (whenever available) are used because the cloud

Table 1.1. S2 Cloud masking publicly available datasets

Dataset	Year of Release	Size	Labels
Hollstein et al. [43]	2016	9 million	Polygon
Baetens et al. [44]	2018	38	Full-scene
S2 cloud mask catalogue [45]	2020	513	Full-scene
WHUS2-CD+ [46]	2021	36	Full-scene
CloudSEN12 [47]	2022	49,400	Partial-scene
S2 cloud cover segmentation dataset [48]	2022	22,728	Partial-scene

height can be estimated [49]. Although the current research has shown promising results, improvement is still required [35][50][51].

1.2.2 VHR change detection (CD)

Through the CD Earth observation task, land cover transitions through time can be monitored. In VHR data, changes in smaller objects (e.g. buildings) can be displayed. However, when the land cover is observed in VHR, its high complexity emerges. The main challenges are posed by the increased within-class variance and the geometric registration errors [52][53]. The high within-class variance is generated not only by the object properties but also by the variable lighting conditions. Severe co-registration errors are mainly caused by the oblique sensor viewing geometry in multi-modal CD where the data collection is performed by heterogeneous sensors [54][55].

The CD task was initially approached by pixel-based techniques (e.g. CVA [56], PCA [57]) and subsequently by object-based CD (OBCD) methods [58] where spatial information is exploited. Even though OBCD is less sensitive to co-registration errors, the segmentation accuracy plays an important role in the success of the CD task. The above-mentioned methods were mainly applied in high/medium resolution images. Pixel-based CD in medium resolution is more robust to residual misregistration and researchers have typically confronted the issue by employing local information and the polar domain [59][60].

Recently, the increased access to high processing power systems has allowed convolutional DL, which has an innate spatial context perception, to advance the CD field. Both unsupervised and supervised approaches have been performed. The unsupervised CD is generally based on the comparison of feature maps produced by the bitemporal images [61][62][63]. Concerning the supervised CD, which usually produces more accurate results, its progress has been motivated by the increased availability of annotated CD datasets (Table 1.2).

Since CNNs capture spatial information, they perform better than pixel-base methods in the misregistration problem. In the latest research, the spatial context perception is further enhanced by the adoption of spatial attention mechanisms that capture long-range spatial dependencies [64][65]. Although the current scientific research concerning DL CD with co-registration noise has shown promising results, it has mostly focused on images with minor co-registration errors. In addition, the evaluation of the networks is usually not performed on very dissimilar test datasets compared to the training sets.

1.2.3 Marine plastic litter detection through image fusion

Marine litter is composed of manufactured solid materials that are discarded in the marine and coastal environment [66] and it mainly originates from land-based activities (~80%) [67]. Marine litter has a broad range of negative environmental and socio-economic impacts [68][69]. Its dominant component is plastic, a durable material typically used for storage purposes because it does not react with the content. When plastic is discarded in the marine environment, it can pose a threat to the marine wildlife (e.g. entanglement) and subsequently to the human

Table 1.2. VHR CD publicly available datasets

Dataset	Year of Release	Number of pairs	Spatial resolution (m)	Spectral resolution (channels)	Types of changes	Time periods	Size
SZTAKI AirChange Benchmark set [70]	2008	13	1.5	3	Binary (multi-type)	3	952×640
Lebedev et al. [71]	2018	11	0.3-1	3	Binary (multi-type)	2	4725×2700 1900×1000
HRSCD [72]	2019	291	0.5	3	Multiclass	2	10,000×10,000
WHU building data set [73]	2019	1	0.3	3	Binary (buildings)	2	20.5 km^2
LEVIR-CD [65]	2020	637	0.5	3	Binary (buildings)	2	1024×1024
S2Looking [74]	2021	5000	0.5-0.8	3	Binary (buildings)	2	1024×1024
QFabric [75]	2021	2520	0.45	4	Multiclass	5	8192×8192
SECOND [76]	2022	4662	various sensors	3	Multiclass	2	512×512

health, as it degrades into microplastic and can access the food chain [68][69].

Marine plastic litter has become a global environmental concern in recent years causing an urgent demand for tools and techniques that enable waste detection and monitoring. Satellite Remote Sensing provides global and continuous temporal coverage [77][78] and has shown promising results in the detection of large-sized marine debris in initial experiments. However, challenges emerge regarding atmospheric and sea-surface effects, spectral/ spatial and temporal resolutions, and availability of ground-truth data [79].

Plastic litter detection studies conducted in commercial VHR data have shown distinctive SWIR absorption, NIR peak reflectance, variability in the visible spectrum, and lower reflectance of wet plastic compared to dry [80][81]. In addition, recent experiments with artificial floating targets demonstrated that 10×10 m^2 plastic targets are distinguishable in S2 data from the water due to their higher reflectance when the pixel coverage is at least 25% [82][83]. The authors asserted that the identification of the plastic types and shapes requires multi- to hyper-spectral imaging. In another S2 study, the Floating Debris Index (FDI) has been proposed with successful results in the detection of macroplastics [84]. Lately, an important contribution in this field is the creation of a large dataset based on S2 data which contains verified plastic debris events in several geographical regions globally [85]. This dataset has paved the way for the application of DL detection approaches.

Research has highlighted that high spatial and spectral resolutions are the two critical factors in enhancing the detection ability of the specific plastic spectral features and the potential discrimination of different types. However, presently, owing to the constraints imposed by satellite sensors in terms of technology and physical capabilities, there are critical trade-offs between the spectral and spatial resolution of satellite imagery. This fact puts focus on the exploitation of image fusion approaches to optimize current observing systems' potential to detect and identify plastic marine litter.

In the image fusion task, an image of higher spatial and lower spectral resolution is fused with an image of higher spectral and lower spatial resolution. The main challenge is the production of accurate results in the spectral range where no overlap exists [86]. In addition, in DL approaches, a significant factor is the ratio of the spatial resolutions between the two different types of data. It is noted that DL approaches have been mainly applied in the pansharpening of VHR MS data [87] and in the spatial super-resolution of high/medium resolution MS images [88][89].

1.2.4 RGB-to-NIR image-to-image translation (ITIT)

ITIT refers to an image processing technique that aims to learn the mapping functions between an input and an output image [90]. ITIT can be either performed in a paired (co-registered input and output) or an unpaired setting. Recently, increased interest has been shown in the Remote Sensing community for DL paired ITIT approaches by typically employing conditional GANs (cGANs) to generate missing information.

SAR-to-optical ITIT (and vice versa) dominates the Remote Sensing research in this field due to the independence of SAR data from atmospheric conditions. Due to higher accessibility, most of the studies process high-resolution information (>=5 m) [91][92] instead of VHR [90][93]. Other studies have focused on translating images in the visible spectrum [94][95] as well as TIR-to-visible [96] in geostationary meteorological satellites. In addition, research has been conducted for VHR visible-to-map IT [97][98] which is advantageous for timely updates, and for optical-to-elevation ITIT [99] which serves as a cost-effective alternative to methods that rely on Lidar, InSAR, or stereo pairs.

Concerning the generation of NIR information from RGB data, according to the present literature, it has been either approached indirectly in the framework of spectral super-resolution (RGB-to-HS (SSR)) [100][101] or the interest has exclusively been focused on vegetation applications. In the SSR studies, it has been shown that DL predicted HS outputs are less noisy than the ground-truth [102], CNNs outperform conventional regression methods [103] and cGANs are more robust than non-adversarial networks [100]. It is noted that the researchers have evaluated their paired methodologies in the output as a whole without isolating particular bands (e.g. NIR).

The RGB-to-NIR paired DL IT in vegetation applications has been motivated by the valuable information it provides in determining vegetation parameters [104][105] and has proved its usefulness in implementing cost-effective precision agriculture through low-cost RGB cameras on board lightweight UAVs [106] and in forest

monitoring [107]. It has also been demonstrated that RGB-to-NIR paired DL ITIT in S2 data is unaffected by corrupted pixels.

Regarding unpaired ITIT, it has primarily found application in the field of VHR Remote Sensing for unsupervised domain adaptation (UDA) where it serves as an intermediary step to improve the quality of cross-domain semantic segmentation (CDSS) outputs. The typical scenario involves available annotations in the source domain but not in the target domain. The domain shifts are mainly generated by the variability in lighting conditions and viewing angles and the rich structure diversity.

The RGB-to-NIR DL literature has shown promising results so far in vegetation applications but studies of the RGB-to-NIR ITIT performance in more categories of the complex urban environment (e.g. impervious materials) would be beneficial. In the vegetation applications, it has been highlighted that less satisfactory evaluation scores are produced in out-domain experiments even after fine-tuning [107][108]. The term "out-domain" refers to data that do not belong to the domain of the training set (different satellite/collection date/region). UDA techniques can produce data radiometrically closer to the training set and could thus be explored as a mitigation measure for this challenge. It should be noted though, that unlike the CDSS task, in order to predict a reliable NIR when applying UDA, the source and target data should be collected from the same geographic zone (higher chance to encounter similar spectral information in urban structures/materials/vegetation species) and in the same month to avoid seasonal changes.

1.3 Objectives

The general objective of this PhD thesis is the investigation of the capabilities of different types of ANNs in four Remote Sensing applications: Cloud masking in S2 data, VHR CD, marine plastic litter detection through image fusion, and RGB-to-NIR ITIT.

1.3.1 Cloud masking in S2 data

The general objective of this application is the mitigation of the cloud masking challenges (section 1.2.1) by employing light, time-efficient ANN architectures. The main specific objectives are:

- The investigation of the potential of MLPs to separate pixels of clouds from non-cloud deep water pixels with noise, sunglint, and directional reflectance effects (caused by the broad range of viewing geometries) in S2 data.
- The study of the effect of feature scaling on the MLP predictions.
- The development of a novel fine-tuning methodology for SOMs to mitigate the effect of bright non-cloud objects caused by sunglint in land areas.
- The thorough examination of the capabilities of patch-to-pixel CNNs to tackle all challenging cases in cloud masking including snow and thin clouds.
- The exploitation of the first publicly available annotated datasets [43][44].
- The observation of the network parameters.
- The comparison with state-of-the-art cloud masking rule-based [18][22] [109] and muti-temporal [110] methods.

1.3.2 VHR CD

The general objective of this application is the assessment of several state-of-the-art DL CD methods on VHR images with severe co-registration noise. The main specific objectives are:

- The evaluation of automatic co-registration methods on VHR images with severe co-registration noise.
- The study of multi-modal data collected on urban areas of heterogenous morphology.
- The exploration of each main category of unsupervised and supervised DL CD methods.

1.3.3 Marine plastic litter detection through image fusion

The general objective of this application is the increase of spatial resolution in either the PRISMA or the S2 satellites through image fusion to facilitate the detection and monitoring of marine plastic litter. The main specific objectives are:

- The implementation of the first study that employs satellite HS data to detect plastic litter.

▪ The evaluation of several state-of-the-art DL pansharpening networks in HS PRISMA data and the exploration of the possibility to detect small-sized marine plastic targets (≤5 m).
▪ The development of novel plastic litter indexes in pansharpened PRISMA data.
▪ The creation of DL networks for the fusion of different orbiting MS sensors (case study: S2 + WV-3).
▪ The adaptation of literature pansharpening/single image super-resolution (SISR) networks to the fusion problem.
▪ The comparison of the outputs of the DL pansharpening/fusion networks with the outputs of popular conventional pansharpening/fusion techniques (implemented by M. Kremezi).

1.3.4 RGB-to-NIR ITIT

The general objective of this application is the generation of the NIR band in in- and out-domain VHR RGB imagery by exploiting paired and unpaired GANs. The main specific objectives are:
▪ The investigation of the performance of different NIR prediction models (paired cGANs) in an in- and out-domain setting with heterogeneous bi-temporal data.
▪ The exploration of the potential of unsupervised domain adaptation (UDA/unpaired GANs) in improving the output of the NIR prediction models in out-domain data.
▪ The creation of a three-stage GAN framework in a paired and unpaired setting to generate NIR images.
▪ The investigation of the main thematic land cover categories.

1.4 Contributions

The work performed in the framework of this PhD thesis included novel experiments, methodologies, and conclusions that were published in order to assist the scientific progress. Besides the dissemination through scientific publications, relevant code and a dataset have been publicly released.

1.4.1 Technical contributions

1.4.1.1 Cloud masking in S2 data

▪ The creation of a dataset of 2,133,324 Sentinel-2 deep water spectra with noise/sunglint which was publicly released. The spectra were extracted by visual observation through polygons. The dataset can be accessed at https://doi.org/10.6084/m9.figshare.8075396.v1 /CC0 license
▪ The implementation of an MLP architecture that outperformed state-of-the-art rule-based and muti-temporal methods in the separation of clouds from deep water spectra with noise, sunglint, and directional reflectance effects (caused by the broad range of viewing geometries).
▪ The creation of a measure that indicates the bands that mitigate the influence of deep water spectra with noise/sunglint, based on the weights of the first MLP hidden layer.
▪ The investigation of the effect on the MLP output of applying feature scaling by using the parameters of the test set instead of the training set.
▪ The development of a novel fine-tuning methodology for SOMs which is task-independent and requires small amounts of data. The method mitigated the effect of bright non-cloud objects caused by sunglint in land areas.
▪ The implementation of a patch-to-pixel CNN that outperformed state-of-the-art rule-based methods in all challenging cases.

Code
- https://github.com/vkristoll/cloud-masking-ANNs /GPL-3.0 license
- https://github.com/vkristoll/cloud-masking-SOMs /GPL-3.0 license
- https://github.com/vkristoll/cloud-masking-CNNs /GPL-3.0 license

1.4.1.2 VHR CD

▪ The assessment of five state-of-the-art DL CD methods on VHR images with severe co-registration errors where the importance of spatial attention mechanisms and dataset similarity was reinforced.
▪ The evaluation of four co-registration methods on VHR images with severe co-registration errors where the superiority of an FFT-based method that uses phase correlation was shown.
▪ The creation of a novel score that provides a better understanding of the magnitude of the commission error.

Code
- https://github.com/vkristoll/change-detection-autoencoder /GPL-3.0 license

1.4.1.3 Marine plastic litter detection through image fusion

▪ The implementation of the first study that employs satellite HS data to detect plastic litter.
▪ The evaluation of three state-of-the-art pansharpening DL networks (CNNs) in HS PRISMA data for their potential in discriminating small-sized marine plastic targets (≤5 m) from water.
▪ The establishment of the importance of histogram clipping as a pre-processing step in DL methods.
▪ The development of novel plastic litter indexes in pansharpened PRISMA data by use of the VNIR spectrum.
▪ The creation of three lightweight CNNs for the fusion of S2 and WV-3 data and their evaluation in terms of spatial and spectral distortions.

Code
- https://github.com/vkristoll/Pansharpening-PRISMA-CNNs /GPL-3.0 license
- https://github.com/vkristoll/Fusion-Sentinel2-Worldview /GPL-3.0 license

1.4.1.4 RGB-to-NIR ITIT

▪ The demonstration of the ability of paired cGANs to produce adequate NIR predictions when the domain gap is not significantly high.
▪ The introduction of UDA in improving the output of the NIR prediction models in out-domain data, where promising results were shown for the high vegetation category..
▪ The creation of a three-stage GAN framework to generate NIR images where both paired and unpaired data were exploited.

1.4.2 Scientific publications

Journals

[J1] **Kristollari, V.** and Karathanassi, V., 2024. Exploiting paired and unpaired generative adversarial networks for NIR band generation in VHR RGB satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (under review) (IF: 5.5)*

[J2] Kremezi, M., **Kristollari, V.,** Karathanassi, V., Topouzelis, K., Kolokoussis, P., Taggio, N., Aiello, A., Ceriola, G., Barbone, E. and Corradi, P., 2022. Increasing the Sentinel-2 potential for marine plastic litter monitoring through image fusion techniques. *Elsevier Marine pollution bulletin*, *182*, p.113974. *(IF: 5.8)* doi:10.1016/j.marpolbul.2022.113974

[J3] Taggio, N., Aiello, A., Ceriola, G., Kremezi, M., **Kristollari, V.**, Kolokoussis, P., Karathanassi, V. and Barbone, E., 2022. A combination of machine learning algorithms for marine plastic litter detection exploiting hyperspectral PRISMA data. *MDPI Remote Sensing*, *14*(15), *p.3606. (IF: 5.0*) doi:10.3390/rs14153606

[J4] **Kristollari, V.** and Karathanassi, V., 2022. Change detection in VHR imagery with severe co-registration errors using deep learning: A comparative study. *IEEE Access*, *10*, *pp.33723-33741. (IF: 3.9)* doi:10.1109/ACCESS.2022.3161978

[J5] Kremezi, M., **Kristollari, V.**, Karathanassi, V., Topouzelis, K., Kolokoussis, P., Taggio, N., Aiello, A., Ceriola, G., Barbone, E. and Corradi, P., 2021. Pansharpening PRISMA data for marine plastic litter detection using plastic indexes. *IEEE Access*, *9*, *pp.61955-61971. (IF: 3.9)* doi: 10.1109/ACCESS.2021.3073903

[J6] **Kristollari, V.** and Karathanassi, V., 2020. Fine-tuning Self-Organizing Maps for Sentinel-2 imagery: Separating clouds from bright surfaces. *MDPI Remote Sensing*, *12(12), p.1923. (IF: 5.0)* doi:10.3390/rs12121923

[J7] **Kristollari, V.** and Karathanassi, V., 2020. Artificial neural networks for cloud masking of Sentinel-2 ocean images with noise and sunglint. *Taylor & Francis International Journal of Remote Sensing*, *41(11), pp.4102-4135. (IF: 3.4)* doi:10.1080/01431161.2020.1714776

Conferences

[C1] Karathanassi, V., Karamvasis, K., **Kristollari, V.,** Kolokoussis, P., Skamantzari, M., Georgopoulos, A., 2024, April. Remote sensing techniques for monitoring cultural heritage sites, *In EGU General Assembly 2024, Vienna, Austria*, doi:10.5194/egusphere-egu24-10181 *(abstract)*

[C2] Aiello, A., Barbone, E., Ceriola, G., Karathanassi, V., Kolokoussis, P., Kremezi, M., **Kristollari, V.,** Taggio, N., 2023, October. Unlocking the potential of Spectral Signature Unmixing and Machine Learning for detecting plastic marine litter: Insights from the REACT Project. *In ESA Remote Sensing of Marine Litter Workshop 2023, Netherlands (abstract + oral presentation)*

[C3] Kremezi, M., **Kristollari, V.,** Karathanassi, V., Kolokoussis P., 2022, September. Enhancing PRISMA and Sentinel 2 capabilities for marine plastic litter detection using Image Fusion techniques, Spectral Signature Unmixing and Spectral Indexes. *In 41st EARSeL Symposium, Cyprus. (abstract + poster + oral presentation) (peer-reviewed)*

[C4] **Kristollari, V.** and Karathanassi, V., 2020, August. Convolutional neural networks for detecting challenging cases in cloud masking using Sentinel-2 imagery. In *SPIE Eighth international conference on remote sensing and geoinformation of the environment (RSCy2020) (Vol. 11524, pp. 188-201) (peer-reviewed)* doi:10.1117/12.2571111

Author contributions
[J1][J6][J7][C4]: **Kristollari V.**: Conceptualization, methodology, data curation, code, writing – original draft/review and editing, Karathanassi V.: Conceptualization, writing – review and editing.
[J4]: **Kristollari V.**: Conceptualization, methodology, code, writing – original draft/review and editing, Karathanassi V.: Conceptualization, data curation, writing – review and editing.
[J2]: **Kristollari V.** Conceptualization, methodology, data curation, code, writing – original draft/review and editing (implementation of the DL-related work and creation of the novel pansharpening indexes), Kremezi M.: Conceptualization, methodology, data curation, code, writing – original draft/review and editing (implementation of the non-DL-related work), Karathanassi V.: Conceptualization, methodology, writing – review and editing, Topouzelis K.: Conceptualization, data curation, writing – review and editing, Kolokoussis P., Taggio N., Barbone E.: Conceptualization, writing – review and editing, Aiello A., Ceriola G.: Conceptualization, data curation, writing – review and editing, Corradi P.: writing – review and editing.
[J5]: **Kristollari V.** Conceptualization, methodology, code, writing – original draft/review and editing

(implementation of the DL-related work and creation of the novel pansharpening indexes), Kremezi M.: Conceptualization, methodology, code, writing – original draft/review and editing (implementation of the non-DL-related work), Karathanassi V.: Conceptualization, methodology, writing – original draft/review and editing, Topouzelis K.: Conceptualization, data curation, writing – review and editing, Kolokoussis P., Taggio N., Barbone E.: Conceptualization, writing – review and editing, Aiello A., Ceriola G.: Conceptualization, data curation, writing – review and editing, Corradi P.: writing – review and editing.

[C1]: **Kristollari V.,** Karamvasis, K, Kolokoussis, P., Skamantzari, M.: Conceptualization, methodology, code, writing – review and editing, Karathanassi, V., Conceptualization, writing – original draft/review and editing, Georgopoulos, A.: Conceptualization, writing –review and editing.

[C3]: **Kristollari V.**: Conceptualization, methodology, data curation, code, writing – review and editing (implementation of the DL-related work and creation of the novel pansharpening indexes), Kremezi M.: Conceptualization, methodology, data curation, code, writing – original draft/review and editing (implementation of the non-DL-related work), Kolokoussis P.: Conceptualization, writing – review and editing, Karathanassi V.: Conceptualization, methodology, writing – review and editing.

[J3]: **Kristollari V.**, Kremezi M., Kolokoussis P., Karathanassi V.: Data curation, writing – review and editing, Taggio N.: Conceptualization, methodology, code, data curation, writing – original draft/review and editing, Aiello A.: Conceptualization, writing – original draft/review and editing, Ceriola G.: Methodology, writing – review and editing.

[C2]: **Kristollari V.**: Conceptualization, data curation, methodology (implementation of the DL-related work and creation of the novel pansharpening indexes), code, writing – review and editing, Aiello A.: Conceptualization, data curation, writing – original draft/review and editing, Barbone E.: Conceptualization, writing – review and editing, Ceriola G.: Conceptualization, data curation, methodology, writing – review and editing, Karathanassi V.: Conceptualization, methodology, writing – review and editing, Kolokoussis P.: Conceptualization, data curation, writing – review and editing, Taggio N., Kremezi M.: Conceptualization, data curation, methodology, code, writing –review and editing.

1.5 Thesis Outline

The thesis is composed of six parts. In the first part (Chapter 1) at first a brief historical background of ANNs is outlined. Then, the motivations, objectives and contributions for each ANN application that was studied in the thesis are stated. In the second part (Chapter 2) three ANN approaches that were proposed in the framework of mitigating challenges in Sentinel-2 (S2) cloud masking are presented. In the third part (Chapter 3) the focus is put on evaluating state-of-the-art DL CD methods on VHR images with severe co-registration noise. In the fourth part (Chapter 4) two studies that were implemented in this PhD thesis concerning the increase of spatial resolution in either the PRISMA or S2 satellites through image fusion are described. In the fifth part (Chapter 5) an analysis is performed of a proposed methodology to generate the NIR band in VHR RGB imagery by exploiting paired and unpaired GANs. Finally, in the sixth part (Chapter 6) concluding comments and suggestions for future work are provided.

CHAPTER 2

Cloud masking in Sentinel-2 (S2) images

In Chapter 2 three approaches that were proposed in this PhD thesis in the framework of cloud masking in Sentinel-2 (S2) images are presented. The first puts focus on ocean data which suffer from noise and sunglint and makes use of MLPs. The second fine-tunes SOMs to mitigate the effect of bright non-cloud objects caused by sunglint in land areas. Finally, the third takes advantage of CNNs and attempts to tackle all challenging cases in cloud masking including snow and thin clouds. In section 2.1 a thorough literature review is provided regarding commonly utilized spectral ranges and the latest cloud masking methods proposed in the literature. The challenges of cloud masking are also presented in this section. In sections 2.2, 2.3, and 2.4 details about the three methodologies are respectively presented.

2.1 Related work

In optical satellite images, the presence of clouds is a crucial obstacle in land and ocean studies performed by image analysis tasks. Thus, the exclusion of clouds from the data is an important step that needs to be implemented prior to atmospheric correction. Two common assumptions that are employed in various cloud detection algorithms are that clouds are characterized by higher reflectance and lower brightness temperature than other types of surfaces [18][22][111][112]. Based on the aforesaid assumptions, most of the current cloud detection methods extract the clouds from the imagery through ruled-based classification which applies a set of thresholds (both static and dynamic) of reflectance and brightness temperature [19][113][114]. Threshold-based cloud detection is usually platform-specific and strongly linked to the geographical area and date of data collection [20][21]. The most well-known threshold methods are ACCA [115] and Fmask [18][22] which have been designed for Landsat imagery [116]. A threshold-based method is also used for the development of the S2 cloud masks provided by the level 2A product [109]. Multi-temporal methods have also been applied extensively by researchers and are based on the idea that abrupt changes in image time series are mainly caused by the presence of clouds since other types of surfaces follow smooth variations [23][24][117][118][119][120]. A well-known multi-temporal cloud masking algorithm is MAJA [110] designed for S2 images.

2.1.1 Main wavelengths used in cloud masking

Several VNIR and SWIR wavelengths have been selected by researchers for cloud masking applications since a variety of VNIR and SWIR bands carry useful information. Channel 2 (0.725–1.10 μm) of AVHRR is considered to provide high contrast between clouds and water [121]. The reflectance ratio of R0.87/R0.66 μm is used in MODIS data along with the 0.936-μm band for low cloud detection, while the SWIR band at 1,380 nm is used for the detection of high clouds (cirrus) [122]. The visible threshold test of the MSG/SEVIRI cloud mask is applied on the 0.8-μm band over the seas and on the 0.6-μm band over coasts [123]. Reflectance in the blue was used in the multi-temporal research of [23] in Formosat-2 and Landsat 5, 7 images, and in Proba-V cloud detection [124]. In Landsat 8, bands 3 (0.525–0.600 μm) and 4 (0.630–0.680 μm) were selected in [117] for the distinction between cloud and non-cloud. Finally, in [125] a SWIR threshold at 1,240 nm, 1,640 nm, and 2,130 nm was proposed for cloud masking in turbid waters instead of the 865/869 nm used in MODIS and SeaWifs respectively, which is considered more suitable for open oceans.

2.1.2 Machine learning approaches for cloud masking

Conventional machine learning as well as deep learning techniques have also been introduced for cloud masking of imagery collected by a wide variety of platforms and have indicated successful results. In [25] an SVM-RBF classification model was trained on fused multiple features of cloud and non-cloud regions of GaoFen-1 and

GaoFen-2 images. In [126] a linear kernel-based SVM was trained on images acquired from the commercial MS satellites: Geoeye, Ikonos, and WV-2. In [26] a BOW model was employed to construct compact features from dense local SIFT features extracted from RapidEye and Landsat imagery. In [127] cloud and cloud shadow were determined by training a total of 15 configurations of MLPs and the inclusion of spatial information was explored through the tassel-cap transformation in Landsat 7 scenes. In [128] the MSG/SEVIRI imager was used to detect cirrus clouds by utilizing a set of four MLPs trained on thermal observations and auxiliary data. In [129] the most significant band ratios and MLPs were combined to differentiate clouds from a background in Landsat ETM+ and MSG/SEVIRI data. In [130] deep extreme learning machines were used to detect cloud cover fraction and to distinguish thick from a thin cloud in HJ-1A/B satellite images.

In [37] a CNN architecture was compared with five MLPs applied to different spectral and spatial features extracted from Spot 6 images. In [27] pixel-level decision tree classifiers were trained on the database proposed in [43] and the labeled results were fed to a deconvolutional network by the use of the Alexnet-FCN model in S2 images. In addition, a patch-to-pixel CNN was combined with random forest in [131]. In [132] patch-to-pixel and patch-to-patch CNN architectures were studied for cloud masking of Proba-V MS images. In [133] multiscale convolutional features were integrated into a network based on FCN and Segnet which was trained on Gaofen-1 images. In [134] a CNN with two branches was designed and trained on Quickbird RGB patches of different sizes to distinguish thick from thin clouds. In [34] an ensemble method combining a lightweight U-Net with wavelet image compression was proposed for on-board cloud detection in small satellites. Finally, other encoder-decoder segmentation approaches have been implemented in Landsat 7,8 [49][135][136], S2 [28], and ZY-3 [137]. CNN approaches have been further proposed for the adaptation between different satellite platforms by Segal et al. [29] for WV-2 and S2 and by Mateo et al. [138] for Landsat-8 and Proba-V.

Besides the above, relevant studies have been performed with SOMs for Landsat 7 and MODIS [30][139] [140][141], taking advantage of their faster training/fine-tuning time and interpretative behavior. SOMs are also included in the operational cloud masking products of S2 [109] and Proba-V [142] satellites.

2.1.3 Approaches to mitigate challenging issues in cloud masking

In general, cloud masking methods usually suffer from thin cloud omission and bright non-cloud object commission. Sunglint, high noise levels, and snow constitute bright non-cloud objects. Sunglint is a transient anomaly that occurs when sunlight is reflected directly into the down-looking optical sensor [32][33]. It is influenced by the position of the sun, the viewing angle of the optical sensor, the water refractive index, the cloud cover, the wind direction, and the speed [143][144][145]. Sunglint occurs mainly on the seawater surface but can be also observed on buildings, desert, and coastal sand.

High noise levels in satellite images can appear as random (‘salt and pepper’) or periodic (vertical or oblique stripes) and can be optically recognized without difficulty. Directional reflectance effects caused by the configuration of the 12 detectors of the MSI of the S2 imaging mission [146] may also be considered as periodic noise with oblique wide stripes, whereas the S2 cirrus band (1.374 μm and relatively low SNR (50)) additionally presents periodic noise with linear stripes.

2.1.3.1 Thin cloud omission

For the detection of high-level thin clouds, the main approach is the use of thermal bands or the use of the cirrus band (1,374 nm) whenever brightness temperature is unavailable [18][22]. Low-level thin clouds are even harder to detect since their spectral signature is highly similar to the underlying surface. Some indicative studies that report the difficulty in correctly classifying this cloud category were conducted in [24][35] where multitemporal methods for Landsat were proposed, in [19][50] where threshold-based methods for Landsat and MODIS respectively were implemented, and in [36] where several conventional machine learning methods for Proba-V were tested.

Lately, CNNs have proven promising for the detection of thin clouds. In [136] an adaptation of Segnet was proposed and produced better results compared to CFmask for Landsat images, while in [34] higher accuracy was derived compared to Adaboost and Random Forest by applying a method based on UNET. Successful results were

also shown for Quickbird imagery in [147] where an encoder-decoder architecture was used and in [134] [148] where CNN patch-to-pixel architectures were used.

2.1.3.2 Bright non-cloud object commission

Concerning high noise levels, researchers usually use spatial information and post-processing methods. In [149] textural properties were used since they tend to be less sensitive to detector noise to train probability NNs and SOMs on GOES-8 images for cloud classification. On the same basis, in [37] textural features were also examined to train CNNs on Spot 6 images, and in [139] the authors experimented with using a spatial variation index to train SOMs on Landsat ETM+ images. In [127] spatial noise was removed with TVR before training MLPs and the masks were post-processed by applying the median filter. The median filter as a post-processing step was also applied in [38] where the proposed method was examined on MS and HS sensors and was based on the use of spectral indices, while opening and closing operators were applied in [39] on the output of their morphological method.

Concerning sunglint, researchers use spectral and spatial information, as well as geometric. Texture operators are often used in rule-based approaches as a pre-processing step, while morphology and geometry operators as a post-processing step. Moreover, when thermal bands are available, their use is supposed to improve non-cloud bright object commission error, since they lead to the estimation of cloud height [22][49][150][151]. However, such kinds of bands are unavailable in S2.

In [152] an algorithm was developed to discriminate sunglint from clouds based on its red characteristics by use of (469, 555, 1,240) nm MODIS bands. In [153] sunglint-affected measurements collected by unmanned/automated platforms were masked by setting thresholds in the 700–950 nm range on the premise that open seawater is assumed to absorb all light in the NIR. Based on the high variability of clouds, in [154] (Polder-2 instrument) and in [155] MODIS using a spatial variability threshold of reflectance at NIR was proposed, while in [40] MODIS better results were produced by examining the spatial variability at SWIR. In [156] the authors attempted to mediate the sunglint effect by the use of image enhancement techniques on AVHRR images. In [157] a combination of physical, statistical, and temporal approaches was used on SEVIRI images and managed not to overestimate cloudy pixels due to the sunglint. In [41] the authors trained SVMs on MODIS images and attempted to treat sunglint areas by use of the reflectance ratio of R0.905/R0.935 μm and a feature that combined R(0.87 μm), solar angle, and the satellite angle. Finally, in [158] MLPs were trained on textural features and gradient-filtered radiances on images collected by an airborne spectrographic imager. The authors observed that sunglint areas can be twice as bright as clouds of low brightness and used a single absorption-free wavelength (753 nm). They also decided not to include the sun and the viewing geometry as input parameters to avoid an incorrect correlation between the Sun zenith angle and cloudiness.

It should be noted that several cloud mask products of satellites with low spatial resolution define sunglint-affected areas geometrically. In more detail, the algorithm for the MODIS cloud masks defined the potential geometric sunglint region as being within 36 degrees of the specular direction and modified spectral tests on these areas [111]. In addition, the cloud masking algorithm used on the GCOM-C satellite identified sunglint areas as those whose cone angle between the solar incident and the satellite direction over the water surface is lower than 35 degrees [42]. Finally, the cloud mask product of the Himawari-8 satellite defined sunglint to be present in areas where the sun zenith angle is lower than 75 degrees [159]. They also took wind predictions into account.

Although the current research has shown promising results, improvement is still required as shown by several studies implemented in Landsat [24][35][50], Gaofen-1 [160], Proba-V [124], and MODIS [119] satellites that reported misclassification of bright built-up areas, soils, water bodies (ocean, lake) and snow (NDSI index is commonly calculated [22][18]). A methodology designed for S2 in [161] where a cloud displacement index was used based on the parallax effects of three highly correlated near-infrared (NIR) bands, has shown the most promising results until now. In addition, convolutional patch-to-pixel and encoder-decoder segmentation architectures have produced in general more successful and more effortless results for the separation of clouds from bright surfaces due to their inherent ability to perceive spatial information. Such a conclusion was reached in studies conducted in WV-2 [29], S2 [29], Landsat [34], and Gaofen-1 [133] where bright non-cloud object misclassification was not observed. In [29] CNN multi-modal patch-to-pixel method for WV-2 and S2 imagery

was proposed and misclassifications of wave-breaks did not occur. Incorrect bright object classification was also not observed in [34] and [133] where encoder-decoder architectures for Landsat and Gaofen-1 imagery were respectively used. As for the snow category, even convolutional deep learning approaches present difficulties in its separation from clouds [51][133][136][162].

2.2 Detecting clouds in Sentinel-2 (S2) ocean images with noise and sunglint through MLPs[1]

In section 2.2, the first cloud masking approach is presented. It employs MLPs on S2 data to separate clouds from ocean spectra with noise and sunglint. In section 2.2.1 the motivations and objectives of the study are provided. In section 2.2.2 the datasets employed in the study are described along with the theoretical and experimental framework. In section 2.2.3 the experimental results and their evaluation process are presented. Finally, in section 2.2.4 the conclusions are summarized and discussed, highlighting the main contributions and future work.

2.2.1 Introduction

The main advantage of deep learning techniques in comparison with conventional machine learning is their independence from the need for the extraction of human-engineered features which is a lengthy process. Another advantage concerns the fact that neural networks are reported as being less sensitive to noise in the training set [17][163]. This property is important since inaccuracy exists in the manual labeling of the ground-truth data [18] [22][23][24][25][26][27][116][120] [121][124][127][134]. However, when the complexity of the architecture of the network is high, combined with the laborious process of the selection of the optimal hyperparameters, deep learning techniques can also prove to be time-consuming. MLPs are characterized by simpler architectures but have proven to be a fast and very efficient method in a wide variety of applications. Thus, this study focuses on the use of MLPs for separating cloudy areas from deep water areas in S2 images with high noise levels, directional reflectance effects, and sunglint, a task which is still a challenge. The study makes use only of spectral information and proposes a simple and time-efficient method which produces satisfactory results. For its purpose: a) MLPs with different configurations are trained on two different databases: the public dataset produced manually in [43] (Hollstein et al. (2016)) and a dataset based on the images used in this study, which is also publicly provided, b) the possibility of improving results by making predictions using the feature scaling parameters of the test set instead of those of the training set is investigated in cases where the test set cannot be adequately represented by the training set and c) a measure that characterizes the importance of the bands according to the weights produced by the MLPs is defined and examined. The results are compared with cloud masks produced by three state-of-the-art algorithms: Fmask, MAJA, and Sen2Cor [109].

2.2.2 Materials and methods

2.2.2.1 Data description

Three datasets were used for analysis in this study. The first dataset consists of spectra extracted from the database created in [43], the second dataset contains 79 S2 satellite images analyzed for the purpose of this study and the third dataset contains spectra extracted from the second dataset. In this paper, the first dataset is named “Hollstein dataset”, the second dataset is named “S2 image dataset” and the third dataset is named “S2 spectra dataset”. The Hollstein dataset and the S2 spectra dataset were used in the training and evaluation process, while the S2 image dataset was used in the visual inspection process. The datasets are described in detail below.

A. Hollstein dataset

The Hollstein dataset is a manually created database with reflectance spectra collected around the globe from S2

[1] **Kristollari, V.** and Karathanassi, V., 2020. Artificial neural networks for cloud masking of Sentinel-2 ocean images with noise and sunglint. Taylor & Francis International Journal of Remote Sensing, 41(11), pp.4102-4135. doi:10.1080/01431161.2020.1714776

Figure 2.1. Location of scenes of the Hollstein et al. (2016) database

Table 2.1. Spectra comprising the Hollstein dataset

Class	Coverage	Number of spectra
Cloud	Opaque cloud	1,500,202
Cirrus	Cirrus and vapor trails	1,205,979
Water	Lakes, rivers, seas	1,435,003
Total		4,141,184

level 1C satellite images. To our knowledge, this was the only publicly available database of manually selected spectra from S2 images at the time that this study was conducted (2019) and contains the classes: “clear”, “cloud”, “shadow”, “snow”, “cirrus”, and “water”. The spectra were selected by use of spectral tools which included false-color composites, image enhancements, and graphical visualization of spectra. It is also clarified that when this study was conducted, publicly available annotated cloud masks for S2 images did not exist. The location of the scenes where the spectra were collected is depicted in Figure 2.1. The data were collected in 2016 and 2017 with 20-m spatial resolution. The selected spectra are 5,647,725 and 3,152,273 respectively and the database is stored in two separate .hd5 files. For this study, three classes were extracted from this database: ‘cloud’ (spectra from opaque clouds), ‘cirrus’ (spectra from cirrus and vapor trails), and ‘water’ (spectra from lakes, rivers, and seas). The number of spectra for each class is presented in Table 2.1. It is noted that the ‘cloud’ and ‘cirrus’ classes were joined in one class (cloud) in the experiments where this dataset was used since separating opaque and cirrus clouds was out of the scope of this study.

B. S2 image dataset

The second dataset used in this study contains 79 Sentinel-2A/2B level 1C images. These images refer to two tiles of the same orbit collected by the S2 MS Instrument in 2016 (four images), 2017 (40 images), and 2018 (35 images). The viewing geometries of the S2 detectors in these tiles range from 1° to 11° in zenith and from 21° to 316° in azimuth. The dates of collection covered all seasons of the year: 28 winter images (December, January, February), 24 spring images (March, April, May), 11 summer images (June, July, August) and 16 fall images (September, October, November). The collection time varied between 10:30 and 10:35 a.m. UTC. Depicting several noise levels and a wide variety of the percentages of cloud cover were the important factors during the selection of the dates. The noise analyzed in this study refers to the random and periodic noise (mainly caused by directional reflectance effects) of the S2 images. An example of the periodic noise caused by the detectors can be seen in Figure 2.9(a) and an example of the periodic noise of the cirrus band can be seen in Figure 2.9(b). A crucial factor in the selection of the study area was the availability of MAJA masks. These masks are highly accepted by the Remote-Sensing community and thus were considered significant for the evaluation process. It was decided to use the already available masks because running the binary code provided by the creators of the method requires high computational power. Figure 2.2 depicts with red color the scenes with available MAJA masks. In this Figure,

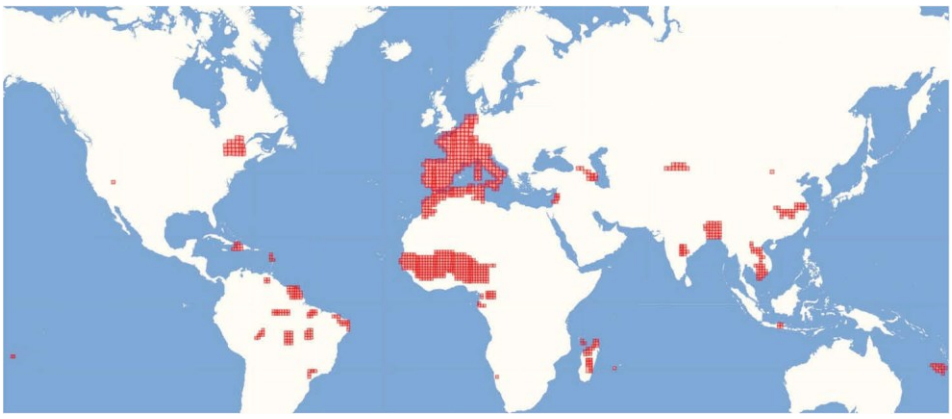


Figure 2.2. Scenes with available MAJA masks (red color)

Table 2.2. Wavelengths of the three spatial resolutions of the S2 instruments

		S2A	S2B
Spatial resolution (m)	Band number	Central wavelength (nm)	Central wavelength (nm)
10	2	496.6	492.1
	3	560	559
	4	664.5	665
	8	835.1	833
20	5	703.9	703.8
	6	740.2	739.1
	7	782.5	779.7
	8A	864.8	864
	11	1613.7	1610.4
	12	2202.4	2185.7
60	1	443.9	442.3
	9	945	943.2
	10	1373.5	1376.9

it can be observed that concerning ocean applications, these masks are at present scarce. Depicting a high percentage of water was also considered during the selection of the tiles of the study area.

S2 images contain 13 bands, three with 60-m spatial resolution, four with 10-m spatial resolution, and six with 20-m spatial resolution. The wavelengths of the three spatial resolutions of the S2 instruments are shown in Table 2.2. Before analysis, these images were processed. The bands with spatial resolution of 10 and 20 m were resampled to 60 m and then the images were cropped in order to remove the land and depict optically homogenous sea regions. The x-size (columns) of the cropped images was 1,830 pixels and the y-size (rows) was 1,130 pixels. Figure 2.3 depicts the study area, the location of the S2 tiles (white polygons (1,2)), and the cropped tiles (red polygons (3,4)).

C. S2 spectra dataset

This dataset includes spectra manually and randomly extracted from images of the S2 image dataset. In more detail, it includes:

a) Reflectance water spectra which were manually extracted by visual observation from 30 of the images of the S2 image dataset. These 30 images consisted of 8 winter images (December, January, February), 9 spring images (March, April, May), 5 summer images (June, July, August) and 4 fall images (September, October, November). These spectra were extracted from water areas with high noise levels and sunglint. Figure 2.4 depicts some example scenes from which spectra were obtained through regions of interest (ROIS). The spectra with high noise

Figure 2.3. The study area, the S2 tiles (white polygons (1, 2)), and the cropped tiles (red polygons (3, 4))

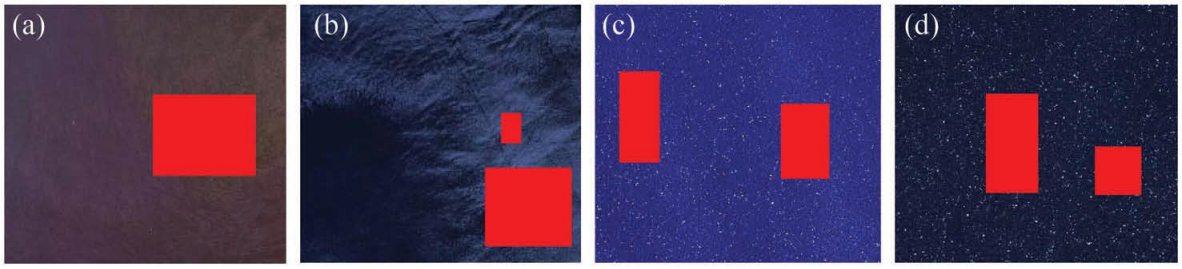


Figure 2.4. S2 scenes with sunglint (a, b) and noise (c, d)

levels were extracted from regions where the noise was visually recognized without difficulty, i.e. without the application of enhancement techniques (e.g. histogram stretching). The spectra with sunglint presence were discriminated from optically thin clouds by use of the cirrus band (1.374 μm) which is less affected by sunglint. The geometric pattern of sunglint was also taken into account. Public access is provided to the database created by the manually extracted water spectra. Spectra of cloud and water without visually obvious presence of noise and sunglint were not manually extracted. To our opinion, these spectra would not be characterized by lower omission and commission errors than those produced by the third experiment (sections 2.2.2.3.B.1 and 2.2.3.1.C), due to the fact that commission and omission errors usually occur in areas where the observer cannot with certainty label the correct class of a pixel, because of high visual similarity (e.g. very thin clouds). In addition, in such a scenario, the observer would choose 'easier' cases in order to increase the confidence of labeling which would probably lead to a less effective training set.

b) Cloud spectra and water spectra which were randomly extracted from 34 images of the S2 image set (different from the 30 images mentioned above). These 34 images consisted of 15 winter images, 10 spring images, 7 summer images, and 6 fall images and the water areas were characterized by low noise and no sunglint presence. These spectra were selected from the cloud masks which were successfully derived from the implementation of the third experiment. The number of manually and randomly extracted spectra is presented in Table 2.3. From each of the 34 images, 60,000 spectra were obtained for cloud and water, respectively, which accounts for 6% of each image spectra (120,000/(1,830 × 1,130)). This percentage of labeled areas corresponds to 4,080,000 spectra (120,000 × 34).

Table 2.3. Spectra comprising the S2 spectra dataset

Class	Number of spectra
Manually extracted water	2,133,324
Randomly extracted cloud	2,040,000
Randomly extracted water	2,040,000

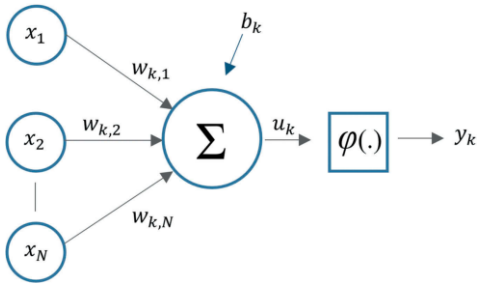


Figure 2.5. Model of a perceptron

2.2.2.2 Theoretical background

A. Multilayer perceptron neural network

MLPs consist of a number of neurons that exchange information in a similar manner as biological nerve cells transmit information via synapses in the human brain. An artificial neuron or perceptron [6] forms the basis for designing ANNs. A model of a perceptron is shown in Figure 2.5.

A neuron k can be described by the following pair of equations (Equations (2.1) (2.2)):

$$u_k = \sum_{i=1}^{N} w_{k,i} x_i \quad (2.1) \qquad y_k = \varphi(u_k + b_k) \quad (2.2)$$

where $x_1, \ldots, x_N$ are the input signals, $w_{k,1}, \ldots, w_{k,N}$ are the synaptic weights of neuron k, b_k is the bias, $\varphi(.)$ is the activation function, and y_k is the output signal of the neuron. The input signals in this study refer to the training spectra extracted from the Hollstein dataset and the S2 spectra dataset.

The MLP architecture consists of three units: input layer, output layer, and several hidden layers. The number of the nodes of the input layer is determined by its input parameters and the number of the nodes of the output layer is determined by its desired output. Neurons in successive layers are connected by weights which represent the importance of the connections in the network. The MLP model is a feed-forward ANN classifier. Each neuron receives inputs from the neurons in the previous layer and through a non-linear activation function converts them to input for the neurons in the next layer. MLPs utilize backpropagation for training the network. During the backward pass, the network's actual output is compared with the target output through an objective function (cost function (C)) (e.g. Equation (2.3)) that needs to be minimized.

$$C = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{i,j}\, w_j)^2 \quad (2.3)$$

where $y_0, \ldots, y_N$ are the true output values, $x_{0,0}, \ldots, x_{N,M}$ are the values of the neurons in the previous layer, $w_0, \ldots, w_M$ are the weights connecting the output layer with the previous layer, N is the sample size and M is the number of connections. The output values in this study refer to the class of the spectra. Output values over 0.5 were classified to the cloud class while output values below 0.5 were classified to the water class.

The error estimates are computed for the output units and the weights that connect the output units with the previous hidden layer are adjusted to reduce these errors. The error adjustment is propagated to the connections of the units in the hidden layers and the connections originating from the input units. The backpropagation process [12] is typically implemented by the stochastic gradient descent method which produces the updated weights for the learning rate:α by calculating the partial derivatives of the cost function with respect to each weight (Equation (2.4)).

$$w_j \leftarrow w_j - a\frac{\partial C}{\partial w_j} \qquad j \in [0,M] \qquad (2.4)$$

where w_j, C, M are defined in Equation 3 and $\frac{\partial C}{\partial w_0},\ldots,\frac{\partial C}{\partial w_M}$ are partial derivatives.

In this study, Adam [164] was used for the implementation of the backpropagation process which is an optimization algorithm of the stochastic gradient descent.

B. Adaptive moment estimation

Adam is an optimization algorithm of the stochastic gradient descent method for the calculation of the weights during the back propagation process. The method stores an exponentially decaying average of past gradients $\boldsymbol{m}_t$ (Equation (2.5)) and past squared gradients $\boldsymbol{v}_t$ (Equation (2.6)). The gradients $\boldsymbol{g}_t$ denote the vector of partial derivatives of the objective function (cost function) at timestep t. $\boldsymbol{m}_t$ and $\boldsymbol{v}_t$ are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients respectively. The zero bias of $\boldsymbol{m}_t$ and $\boldsymbol{v}_t$ is counteracted by computing bias-corrected first and second moment estimates ($\widehat{\boldsymbol{m}}_t$, $\widehat{\boldsymbol{v}}_t$) (Equations (2.7), (2.8)). These are used to update the parameters (weights ($\boldsymbol{\theta}_t$)) (Equation (2.9)).

$$\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \qquad (2.5) \qquad \boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t{}^2 \qquad (2.6)$$

$$\widehat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-\beta_1{}^t} \qquad (2.7) \qquad \widehat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1-\beta_2{}^t} \qquad (2.8) \qquad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{\sqrt{\widehat{\boldsymbol{v}}_t}+\epsilon}\widehat{\boldsymbol{m}}_t \qquad (2.9)$$

where β_1 and β_2 are exponential decay rates for the moment estimates and η is the learning rate.

C. Feature scaling

Feature scaling is a typical step of data pre-processing which is applied to independent variables or features in order to create a particular range of values. The implementation of this process impedes the dominance of the results by features of high magnitude and accelerates calculations. One of the methods widely used for feature scaling is standardization (or Z-score normalization) which is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with μ=0 (mean value) and σ=1 (standard deviation). This process was applied in this study for rescaling the features (spectra values) of the training and test sets. The rescaled values of the features (z) were calculated by Equation (2.10).

$$z = \frac{x-\mu}{\sigma} \qquad (2.10)$$

where x are the initial values of the features.

2.2.2.3 Method description

A. Description of the MLPs

In this study, a total of four MLP configurations were trained on the Hollstein dataset and the S2 spectra dataset, i.e. 8 trainings were implemented in total. The four configurations were differentiated by the use of different algorithms that prevent overfitting.

The architecture of the MLPs consisted of one input layer, two hidden layers, and one output layer. The input layer contained 13 neurons (the total number of S2 bands), each of the two hidden layers contained 20 neurons and the output layer contained one neuron since the classification is binary (cloud/water). It was decided to use the spectral information from all the bands of S2 images since the literature exploits the VIS, NIR, and SWIR bands. In addition, the analysis of the importance of the different wavelengths for the MLP was also a purpose of this study. The architecture and the number of neurons in the hidden layers were selected based on preliminary experiments conducted on the Hollstein dataset. The ReLU [165] (Equation (2.11)), was utilized as an activation function in the two hidden layers. Its main advantages are computational simplicity, its

linear behavior, and its sparse representation capability since it can output true zero value. The Sigmoid function [166] was used as an activation function in the output layer (Equation (2.12)). The graphs of the ReLU and the Sigmoid function are presented in Figure 2.6. It should be stated that this figure follows the nomenclature of the S2 products, i.e. the last band corresponds to the number '12' (Table 2.2).

$$\varphi(x) = \max(0, x) \quad (2.11) \qquad \varphi(x) = \frac{1}{1 + e^{-x}} \quad (2.12)$$

where $x \equiv u_k$ as described in Equation (2.1).

Adam optimization (Equation (2.9)) was selected for the back propagation process with the default values of the Keras library [167] (η=0.001, β_1=0.9, β_2=0.999, ϵ=10^{-8}). In the first configuration, the MLP was trained without applying any algorithm that prevents overfitting. In the second configuration, the dropout method [168] was applied, which ignores neurons at random during the training phase. This method was applied with a 0.3 value in both hidden layers, i.e. 30% of the neurons are ignored in each hidden layer. In the third and fourth configurations, the L1 (Equation (2.13)) and L2 regularizations (Equation (2.14)) [169] which add a regularization term in the cost function (Equation (2.3)), were respectively implemented in both hidden layers.

$$C = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{i,j}\, w_j)^2 + \lambda \sum_{j=0}^{M} |w_j| \quad (2.13) \qquad C = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{i,j}\, w_j)^2 + \lambda \sum_{j=0}^{M} {w_j}^2 \quad (2.14)$$

For the λ parameter, the value 0.001 was selected for the L1 regularization and the value 0.005 for the L2 regularization.

In all configurations, the MLPs were trained for 100 epochs, with batch size 1024. The weights from all the epochs were stored and the weights that produced the higher accuracy in the training set were used for predictions on the test sets. Training was implemented on the GPU: Nvidia GeForce GTX 960M and it lasted approximately 20 min for each of the 8 trainings. Figure 2.6 presents the proposed methodology. The ANNs were trained by using the Keras library and the Tensorflow [170] backend and were implemented in Python code. Tensorflow is an open-source software library for numerical computation developed by Google researchers. It uses a flexible data flow architecture that is suitable for parallel processing applications (e.g. neural networks). Keras is an open-source neural-network library written in Python and capable of running on top of Tensorflow. Creating neural-network models on Keras is simpler since emphasis was put on achieving user-friendliness.

B. Training the ANNs

1. Training on the Hollstein dataset

The Hollstein dataset was used in the training of all four configurations of the ANNs. The training set consisted of spectra extracted from the 'water' class, the 'cloud' class, and the 'cirrus' class. The number of labeled spectra for each class and their percentage which was calculated by use of the total number of spectra for each class of the Hollstein dataset (Table 2.1), is presented in Table 2.4. The purpose of the choice of the number of training spectra for each class was the exploitation of a large number of the available labeled spectra, by simultaneously preserving a balance between the size of the classes (Ncloud + Ncirrus ≈ Nwater). Retaining an adequate number of spectra for the test set (different from those of the training set) was also important (≥20%). As already mentioned the 'cloud' and 'cirrus' classes were joined in one class during the training. Three different experiments were implemented on the same training set (Table 2.4) which was rescaled using the average and standard deviation of the training set. The purpose of these experiments was to analyze the possibility of improving results by making predictions using the feature scaling parameters of the test set instead of those of the training set which is the usual practice. The motivation for this investigation was to maximize the exploitation of the Hollstein dataset since it contains a large number of publicly available spectra. In the first and second experiments, the spectra values of the test set were rescaled using the average value and standard deviation of the training set, while in the third experiment using the respective values of the test set. In addition, the experiments were differentiated by the test set used for the predictions. In the first experiment, the test set included spectra from the Hollstein dataset, in

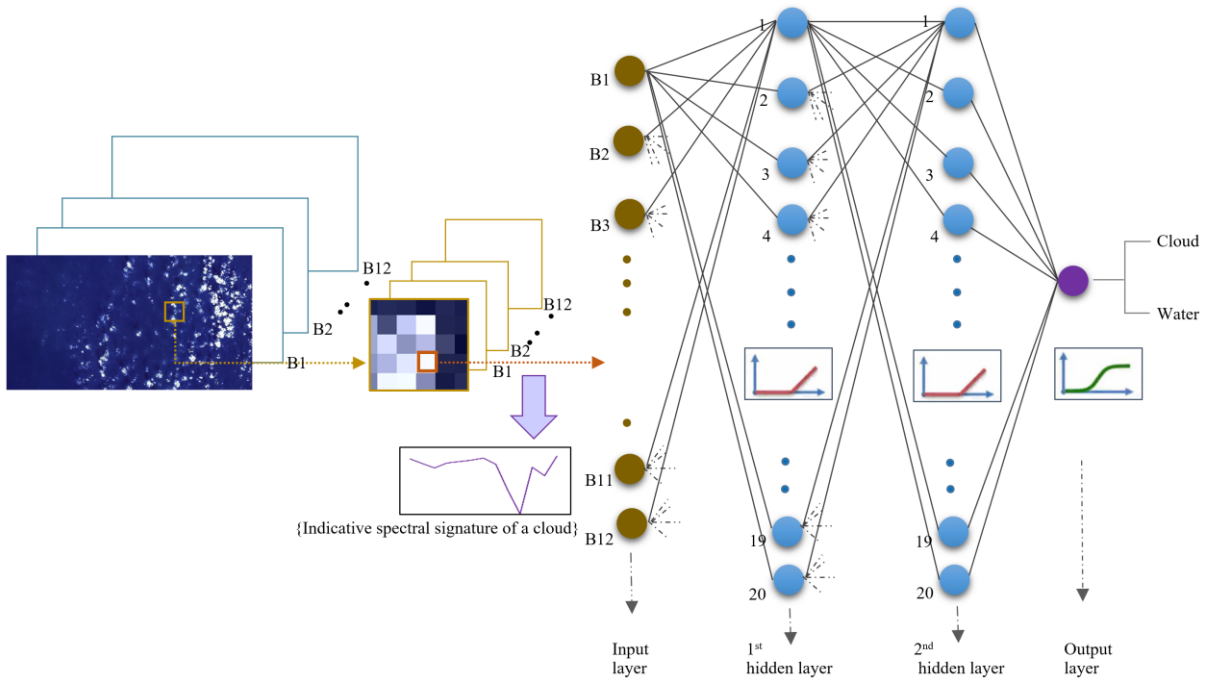


Figure 2.6. The proposed methodology

Table 2.4. Spectra comprising the Hollstein training set

Class	*A:*Spectra of Hollstein training set	*B:*Spectra of Hollstein dataset	*A* as a proportion of total (%)
Water	1,000,000	1,435,003	67%
Cloud	500,000	1,500,202	33%
Cirrus	500,000	1,205,979	41%

Table 2.5. Summary of MLP experiments (training on Hollstein dataset).

Experiment	Training set	Test set	Training set feature scaling	Test set feature scaling
1st	Hollstein training set	Hollstein test set	$z = \frac{x - \mu_{\text{training}}}{\sigma_{\text{training}}}$	$z = \frac{x - \mu_{\text{training}}}{\sigma_{\text{training}}}$
2nd	Hollstein training set	S2 spectra test set / S2 image dataset	$z = \frac{x - \mu_{\text{training}}}{\sigma_{\text{training}}}$	$z = \frac{x - \mu_{\text{training}}}{\sigma_{\text{training}}}$
3rd	Hollstein training set	S2 image dataset	$z = \frac{x - \mu_{\text{training}}}{\sigma_{\text{training}}}$	$z = \frac{x - \mu_{\text{image}}}{\sigma_{\text{image}}}$

the second experiment it included spectra from the S2 spectra dataset and the S2 image dataset, and in the third experiment it included the S2 image dataset. The experiments are described in detail below and are summarized in Table 2.5. It is noted that 'z' in Table 2.5 symbolizes the input of the MLP. It is also noted that the term 'Predictions' in the titles of the subsections below refers to the testing process of the MLPs after they are trained. During this stage spectra not included in the training process are given as an input to the MLP and the output is evaluated.

First experiment: Predictions on the Hollstein dataset by using for the test set the feature scaling parameters of the training set

In the first experiment, the test set included spectra from the 'cloud' class, the 'cirrus' class, and the 'water' class, which were extracted from the Hollstein dataset and were different from the training set. The number of spectra for each class and their percentage calculated by use of the total number of spectra for each class used for the experiment is presented in Table 2.6. The spectra values of the test set were rescaled by applying the average value and standard deviation of the training set. For this experiment, the results were evaluated for all four configurations by evaluation metrics.

Second experiment: Predictions on the S2 spectra dataset and the S2 image dataset by using for the test set the feature scaling parameters of the training set

In the second experiment, the test set included spectra from the S2 spectra dataset which as already mentioned (section 2.2.2.1.C) included cloud and water spectra randomly extracted by the successfully derived masks of the third experiment and manually extracted water signatures. The number of manually and randomly extracted spectra and their percentage calculated by use of the total number of spectra for each class used for the experiment is presented in Table 2.7. In the same table, the number of spectra for the training set explained in section 2.2.2.3.B.2 is also presented for easier understanding. For the test set of the randomly extracted spectra, it was decided to use the total number of unused remaining spectra of the S2 spectra dataset after subtracting the spectra of the training set. The spectra values of the test set were rescaled by applying the average value and standard deviation of the training set. For this experiment, the results were evaluated for all four configurations by evaluation metrics. In addition, the ANN trained with the first configuration was used to predict the class (cloud/water) of the reflectance signatures for the 79 images of the S2 image dataset. The cloud masks produced by these predictions were evaluated by visual observation.

Third experiment: Predictions on the S2 image dataset by using for the test set the feature scaling parameters of the test set

In the third experiment, the test set consists the 79 images of the S2 image dataset. In this experiment, instead of rescaling the spectra values of the test set by applying the average value and standard deviation of the training set, the predictions on the test set were carried out by rescaling the values with the average value and the standard deviation of the images. In more detail, when executing the predictions on the S2 image dataset, the (1,830 × 1,130) signatures of each image were rescaled according to the average value and standard deviation of this image, i.e. spectra of different images were differently rescaled. The cloud masks produced by these predictions were evaluated by visual observation. It is noted that these cloud masks were produced by the MLP trained with the first configuration. It should be also clarified that a test set consisting only of individual spectra (e.g. the test dataset used in the second experiment) cannot be used in this experiment since the key concept is feature scaling with the average value and standard deviation of the total number of spectra comprising a realistic cloud/water image.

Table 2.6. Spectra comprising the Hollstein test set.

Class	*A:*Spectra of Hollstein test set	*B:*Spectra of Hollstein training set	*A+B*	*A* as a proportion of total (%)
Water	300,000	1,000,000	1,300,000	23%
Cloud	150,000	500,000	650,000	23%
Cirrus	150,000	500,000	650,000	23%

Table 2.7. Spectra comprising the S2 spectra test set

Class	*A*:S2 spectra test set	*B*:S2 spectra training set	*A+B*	*A* as a proportion of total (%)
Manually extracted water	300,000	500,000	800,000	38%
Randomly extracted cloud	1,040,000	1,000,000	2,040,000	51%
Randomly extracted water	1,040,000	500,000	1,540,000	68%

2. Training on the S2 spectra dataset

Besides the Hollstein dataset, spectra from the S2 spectra dataset were also used in the training of all four configurations of the MLPs. This experiment is summarized in Table 2.8. The training set consisted of randomly extracted cloud signatures, randomly extracted water signatures, and manually extracted water spectra. The number of manually and randomly extracted spectra and their percentage calculated by use of the total number of spectra for each class of the S2 spectra dataset is presented in Table 2.9. It was decided that the number of training spectra for each class should be similar to the size of the Hollstein training set since it managed to produce satisfactory results in the first experiment (section 2.2.3.1.A). Moreover, it was considered appropriate to use an equal size of manually extracted water (high noise levels and sunglint) and randomly extracted water (low noise levels/no sunglint presence). From the remaining unused signatures of the S2 spectra dataset, the S2 spectra test set mentioned in the second experiment was created. The spectra values of the test set were rescaled by applying the average value and standard deviation of the training set. The results were evaluated for all four configurations by evaluation metrics. In addition, the MLP trained with the first configuration was used to predict the class of the reflectance signatures for the 79 images of the S2 image dataset. The cloud masks produced by these predictions were evaluated by visual and quantitative comparison with the results produced by the algorithms of Fmask, MAJA, and Sen2Cor.

2.2.3 Results

2.2.3.1 Results produced by training on the Hollstein dataset

A. Predictions on the Hollstein dataset by using on the test set the feature scaling parameters of the training set

Accuracy (Equation (2.15)), recall (producer's accuracy) (Equation (2.16)), precision (user's accuracy) (Equation (2.17)) and True Statistic Skill (TSS) (Equation (2.18)) were calculated for the Hollstein training and test set. Recall corresponds to omission error (100%-omission error) while precision corresponds to commission error (100%-commission error). TSS was chosen instead of Cohen's kappa (the most popular measure for the evaluation of presence-absence predictions), since besides taking random agreement into account, it is also independent of prevalence [171]. It is calculated by the use of sensitivity (recall) and specificity (True Negative Rate) (Equation (2.18)) and measures interrater reliability (agreement of prediction model with ground-truth). Table 2.10 presents the results of the predictions on the training set, while Table 2.11 presents the results of the predictions on the test set.

Table 2.8. Summary of MLP experiment (training on S2 spectra dataset)

Training set	Test set	Training set feature scaling	Test set feature scaling
S2 spectra training set	S2 spectra test set S2 image dataset	$z = \frac{x - \mu_{\text{training}}}{\sigma_{\text{training}}}$	$z = \frac{x - \mu_{\text{training}}}{\sigma_{\text{training}}}$

Table 2.9. Spectra comprising the S2 spectra training set

Class	*A:*S2 spectra training set	*B:*S2 spectra dataset	*A* as a proportion of total (%)
Manually extracted water	500,000	2,133,324	23%
Randomly extracted cloud	1,000,000	2,040,000	49%
Randomly extracted water	500,000	2,040,000	25%

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (2.15) \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2.16) \qquad \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2.17)$$

$$\text{TSS} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{(\text{TP} + \text{FN}) \times (\text{FP} + \text{TN})} = \text{sensitivity} + \text{specificity} - 1 \quad (2.18) \qquad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (2.19)$$

where TP: true positives, TN: true negatives, FP: false positives, and FN: false negatives.

It was observed that the evaluation metrics were very high for all four configurations, both for the training set and the test set.

B. Predictions on the S2 spectra dataset and on the S2 image dataset by using on the test set the feature scaling parameters of the training set

Accuracy, recall, precision, and TSS were calculated for the S2 spectra test set for all four configurations. From the evaluation metrics that are presented in Table 2.12 it was observed that a high number of water spectra was falsely classified as cloud (FN). The MLP trained with the first configuration was also used to predict the class of the reflectance signatures for the 79 images of the S2 image dataset. The cloud masks produced by these predictions were visually evaluated and in the majority of the images a large commission error was observed as expected by the evaluation metrics of Table 2.12 (Figure 2.7(e,f)). In more detail, the values of precision show a minimum commission error of 24%, which corresponds to the number of water pixels being incorrectly classified as cloud pixels. The values of recall show that the omission error is low, i.e. almost all the cloud pixels were correctly classified. Finally, the low TSS values confirm further the low reliability of the model.

These results led to the conclusion that the S2 image dataset cannot be adequately represented by the Hollstein dataset. As it can be noticed in Figure 2.1, the majority of the spectra have been collected from inland and coastal areas, while spectra from deep water areas are scarce. As a result, it could be naturally concluded that water spectra with high noise levels and sunglint are scarce in the Hollstein dataset as well.

C. Predictions on the S2 image dataset by using for the test set the feature scaling parameters of the test set

The predictions on the S2 image set of the MLP trained with the first configuration were evaluated by visual observation and it was observed that for 34 images, the produced cloud masks were satisfactory (Figure 2.7(c,d)), while the cloud masks produced on the rest 45 images were characterized by very high commission error (Figure 2.8(c,d)). The successful results for the 34 images led to the conclusion that the feature scaling process applied to these images (use of the parameters of the images instead of those of the training set) created spectra with statistical

Table 2.10. Evaluation metrics of the predictions on the Hollstein training set

Configuration	TP	FP	FN	TN	Accuracy	Precision	Recall	TSS
1st	999,856	82	144	999,918	0.9999	0.9999	0.9999	0.9998
2nd	997,556	365	2,444	999,635	0.9986	0.9996	0.9976	0.9972
3rd	998,968	44	1,032	999,956	0.9995	1.0000	0.9990	0.9989
4th	997,973	473	2,027	999,527	0.9988	0.9995	0.9980	0.9975

Table 2.11. Evaluation metrics of the predictions on the Hollstein test set

Configuration	TP	FP	FN	TN	Accuracy	Precision	Recall	TSS
1st	299,963	19	37	299,981	0.9999	0.9999	0.9999	0.9998
2nd	299,319	112	681	299,888	0.9987	0.9996	0.9977	0.9974
3rd	299,737	14	263	299,986	0.9995	1.0000	0.9991	0.9991
4th	299,427	148	573	299,852	0.9988	0.9995	0.9981	0.9976

Table 2.12. Evaluation metrics of the predictions on the S2 spectra test set

Configuration	TP	FP	FN	TN	Accuracy	Precision	Recall	TSS
1st	1,009,418	340,241	30,582	999,759	0.8442	0.7479	0.9706	0.7167
2nd	1,000,333	295,228	39,667	1,044,772	0.8593	0.7721	0.9619	0.7415
3rd	1,003,491	320,519	36,509	1,019,481	0.8500	0.7579	0.9649	0.7257
4th	1,003,853	320,751	36,147	1,019,249	0.8500	0.7579	0.9652	0.7259

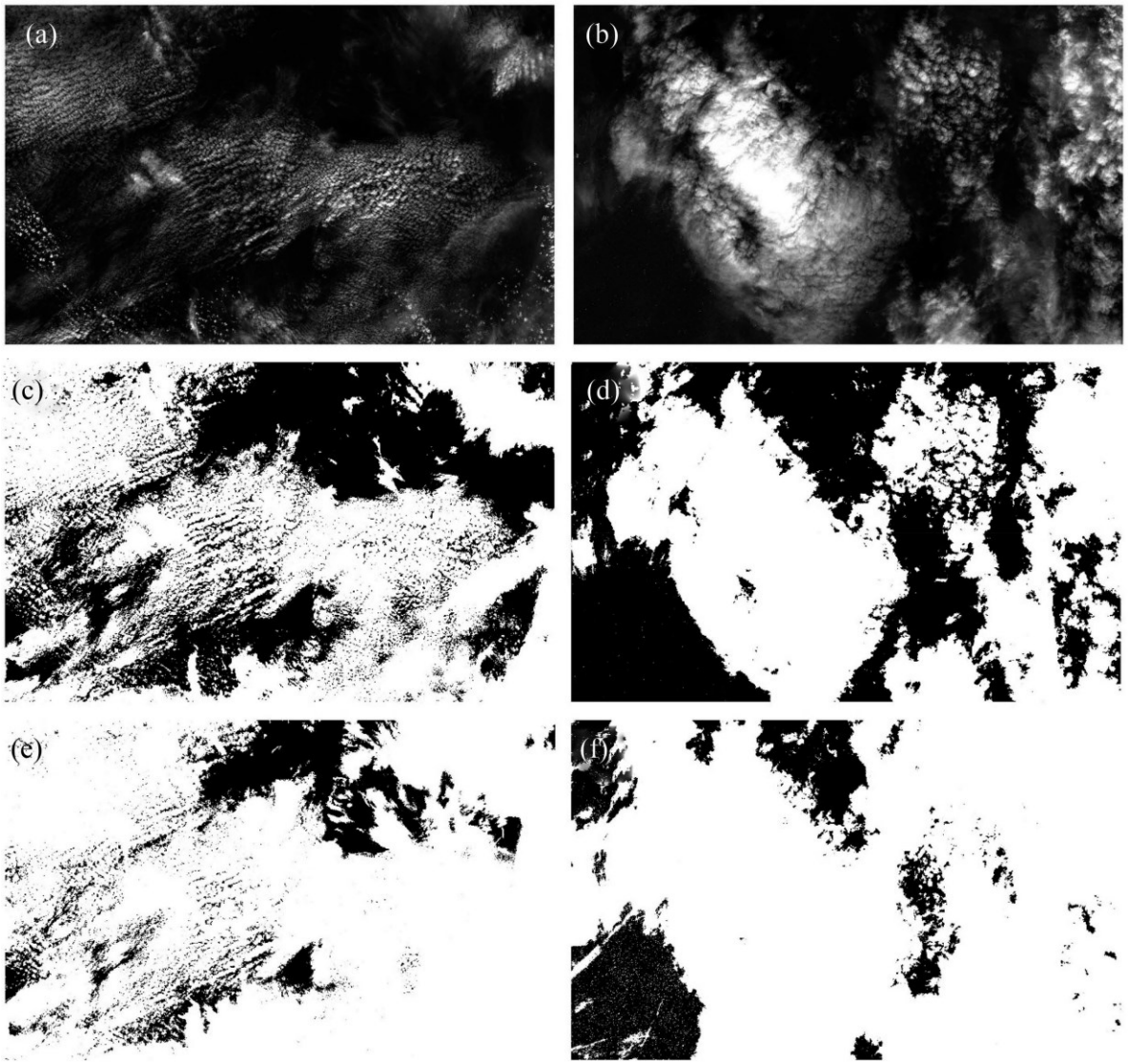


Figure 2.7. (a,b): 4-3-2 (RGB) natural color composite, (c,d): cloud mask produced by using on the test set the feature scaling parameters of the test set, (e,f): cloud mask produced by using on the test set the feature scaling parameters of the training set. The size of all figures is 109.8 × 67.8 km^2

parameters similar to those of the Hollstein dataset. As mentioned in section 2.2.2.1.C, spectra from these 34 cloud masks were randomly extracted and formed part of the S2 spectra dataset. Furthermore, it was observed that the majority of the 45 images (42/45) had high levels of oblique periodic noise in band 10 (cirrus band/1.374 μm) (Figure 2.9(c,d)) in contrast with the majority of the 34 images (31/34) (Figure 2.10(c,d)) which either depicted very low levels of oblique periodic noise or none. The magnitude of band 10 (after the implementation of FFT) [172] is presented in Figure 2.9(e,f) and Figure 2.10(e,f). It was also observed that the vast majority of the noisy images had much lower average reflectance values in band 10 (Figure 2.11). It is noted that the 30 images from

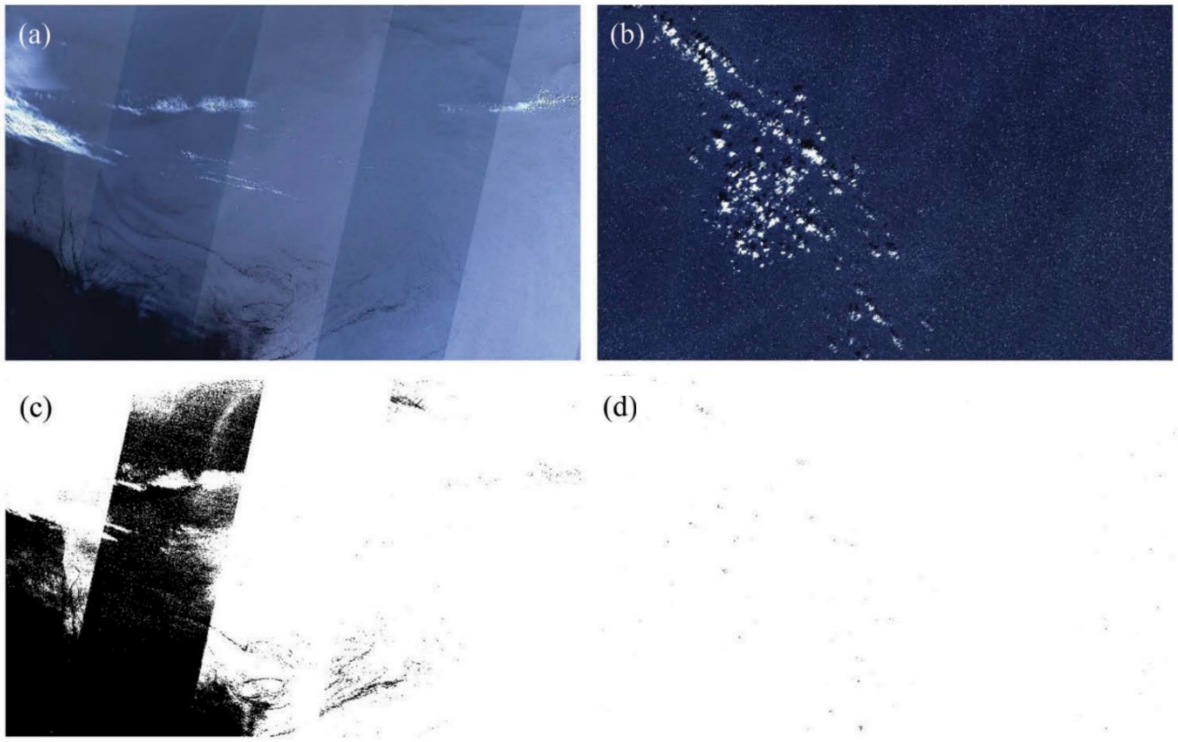


Figure 2.8. (a,b): 4-3-2 (RGB) natural color composite, (c,d): cloud mask produced by using on the test set the feature scaling parameters of the test set. The size of all the figures is 109.8 × 67.8 km^2

which the manually extracted spectra of the S2 dataset were extracted formed part of the 45 images mentioned above.

D. Observation of the weights of the first hidden layer

The weights of the first hidden layer for the four configurations were observed for the MLPs trained on the Hollstein dataset since they represent the importance of the bands for the MLP. Table 2.13 is created by calculating the importance of the bands which was defined as the sum of the absolute values of the 20 weights (equal to the number of neurons) corresponding to each of the 13 bands (Equation (2.20)). This table shows for each configuration in descending order the importance of the bands.

$$\mathrm{Im}_j = \sum_{i=1}^{20} |w_{i,j}| \qquad j \in [1,13] \qquad (2.20)$$

where $w_{1,1},\ldots,w_{20,13}$ are the weights of the first hidden layer.

It was observed that band 11 (1.6 μm) which is primarily used for cloud separation in turbid waters was given high weights in all configurations. As far as the rest of the bands are concerned, the ranking of importance greatly varied as described in section 2.1.1, since for the detection of clouds, a variety of VNIR and SWIR bands has proven to be useful.

2.2.3.2 Results produced by training on the S2 spectra dataset

A. Predictions on the S2 spectra dataset and on the S2 image dataset by using on the test set the feature scaling parameters of the training set

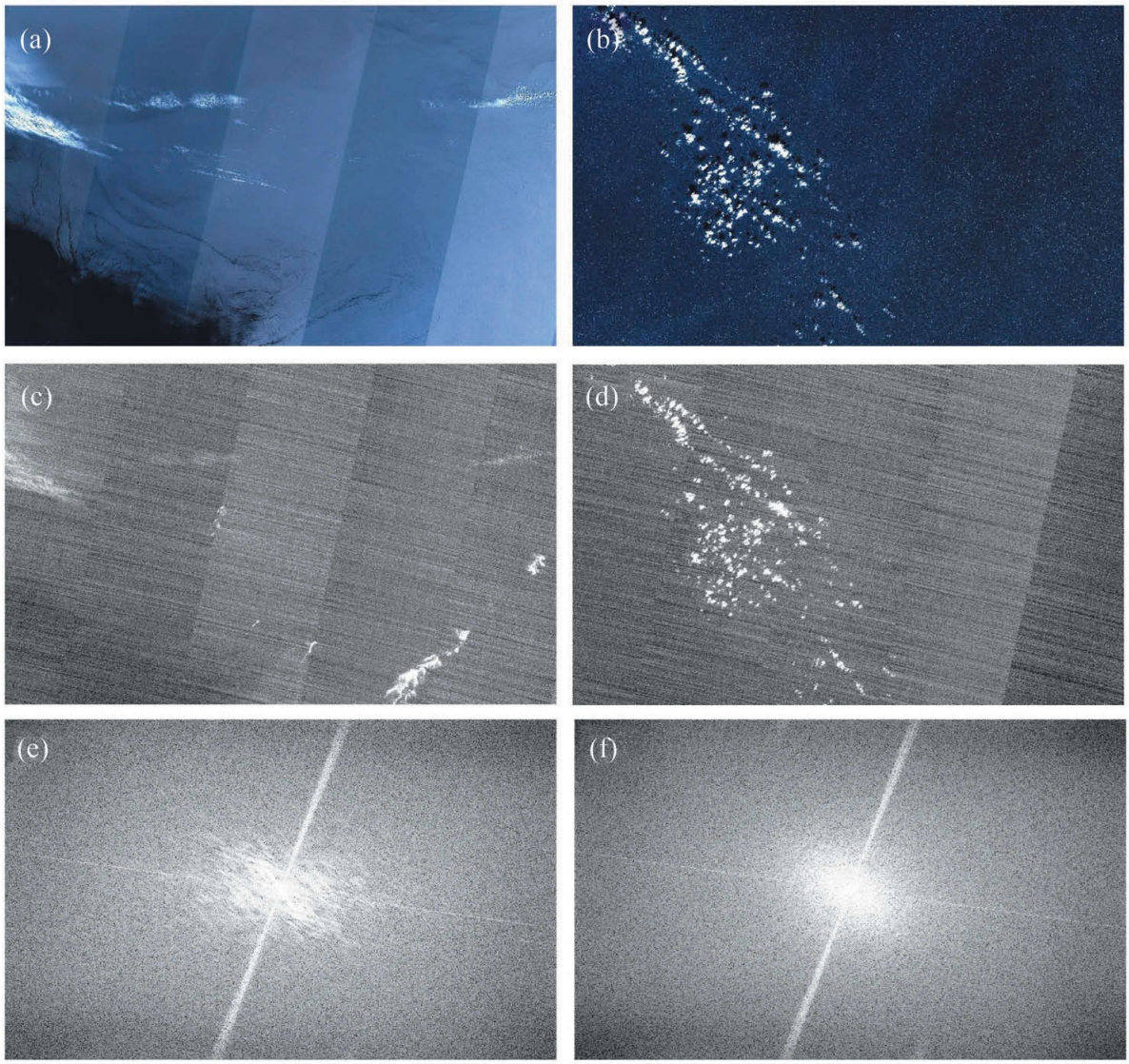


Figure 2.9. (a,b): 4-3-2 (RGB) natural color composite, (c,d): cirrus band (1.374 μm), (e,f): magnitude of cirrus band. The size of all figures is 109.8 × 67.8 km^2

1. Quantitative evaluation on the S2 spectra dataset – Comparison among MLP configurations

This section presents the results of the predictions of the MLPs trained on the S2 spectra dataset. Evaluation metrics were calculated for the S2 spectra training set and the S2 spectra test set. Table 2.14 presents the results of the predictions on the training set, while Table 2.15 presents the results of the predictions on the test set. It was observed that the evaluation metrics were high in all four configurations, both for the training set and the test set. Moreover, the MLP of the first configuration demonstrated the maximum values of accuracy, recall, precision, and TSS. The accuracy of this MLP on the test set was 92% and the TSS value was 0.86 which shows high model reliability. In addition, the recall value (~96%) shows that around 4% of cloud spectra were incorrectly classified to the water class (omission error), while the precision value (~88%) shows that 12% of water spectra were incorrectly classified to the cloud class (commission error).

2. Qualitative and quantitative comparison of the 1st configuration MLP with state-of-the-art methods

The MLP of the first configuration was also used to predict the class of the reflectance signatures for the 79 images of the S2 image dataset. The results (Figure 2.12, Figure 2.13, Figure 2.14) were at first evaluated by visual observation and were compared with the results produced by the algorithms of Fmask, MAJA, and Sen2Cor. In

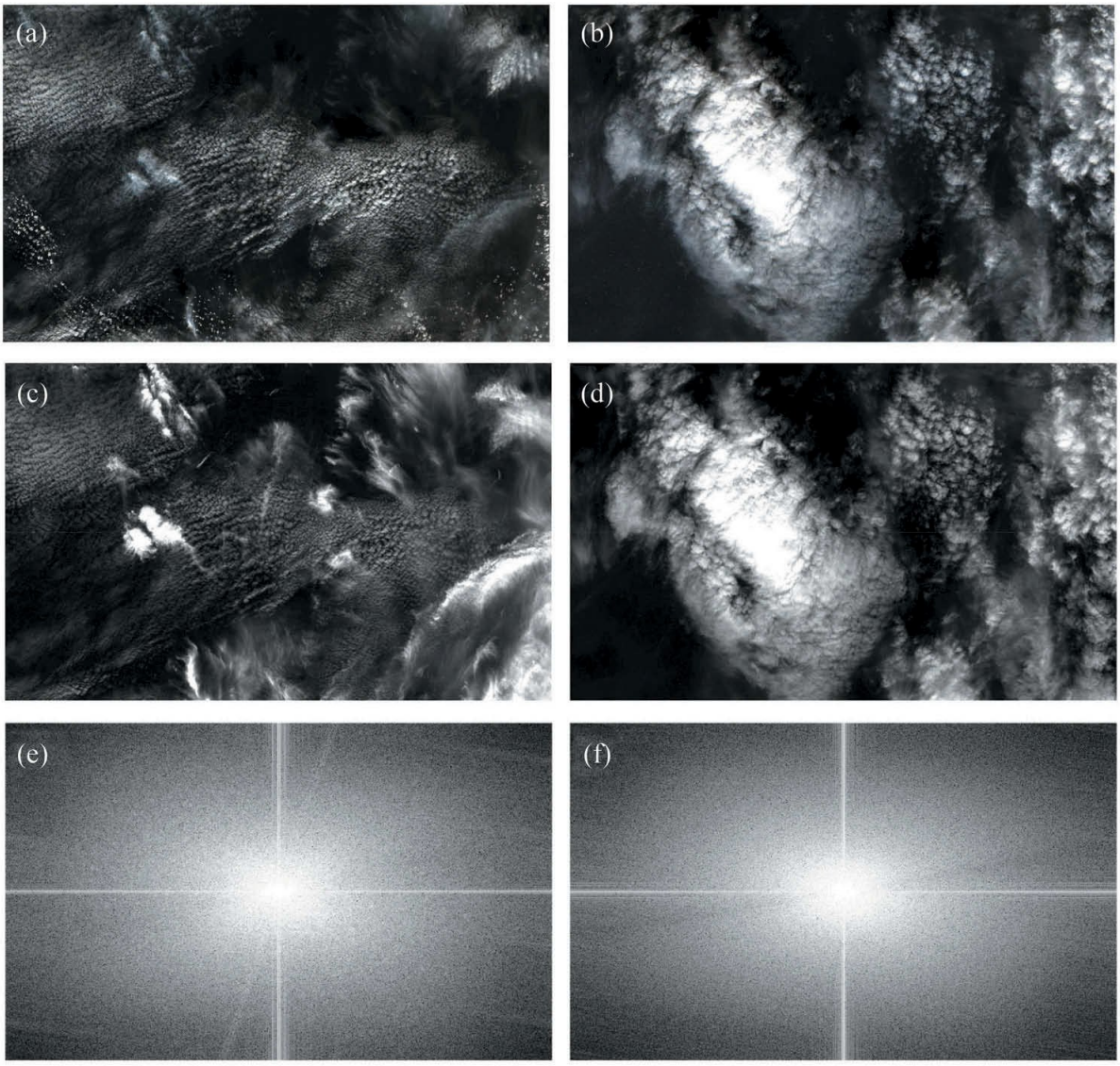


Figure 2.10. (a,b): 4-3-2 (RGB) natural color composite, (c,d): cirrus band (1.374 μm), (e,f): magnitude of cirrus band. The size of all figures is 109.8 × 67.8 km^2

addition, quantitative evaluation was applied by the calculation of evaluation metrics which included a) accuracy, recall, precision and TSS scores for the total number of the spectra of the S2 spectra dataset (labeled pixels) for the MLP and the three state-of-the-art algorithms and b) the phi coefficient (measures the degree of association between two binary variables [173] (Equation (2.21)) between the masks produced by the MLP and the respective masks produced by the above-mentioned algorithms. It is noted that Figure 2.12 presents images with low noise levels and no sunglint presence while Figure 2.13 and Figure 2.14 present more difficult cases (high noise levels and sunglint presence).

$$\Phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \quad (2.21)$$

where *A, D* are the diagonal values of a 2×2 contingency table and *B, C* are the non-diagonal values.

Qualitative comparison of the 1st configuration MLP with state-of-the-art methods

Concerning the visual evaluation, the results of the MLP in all of the 79 S2 images were considered to be very favorable compared to the above algorithms. MLP results were acceptable in all cases and unaffected by water areas with high noise levels and sunglint. In addition, the MLP proved to be robust since the results were homogenous and none of the 79 cases presented outlier classification output, i.e. classifying areas with opaque clouds as water (often observed in Sen2Cor masks) or classifying whole or large part of strips as cloud. It should be noted though that a small omission error was usually observed. The masks produced by the Sen2Cor algorithm

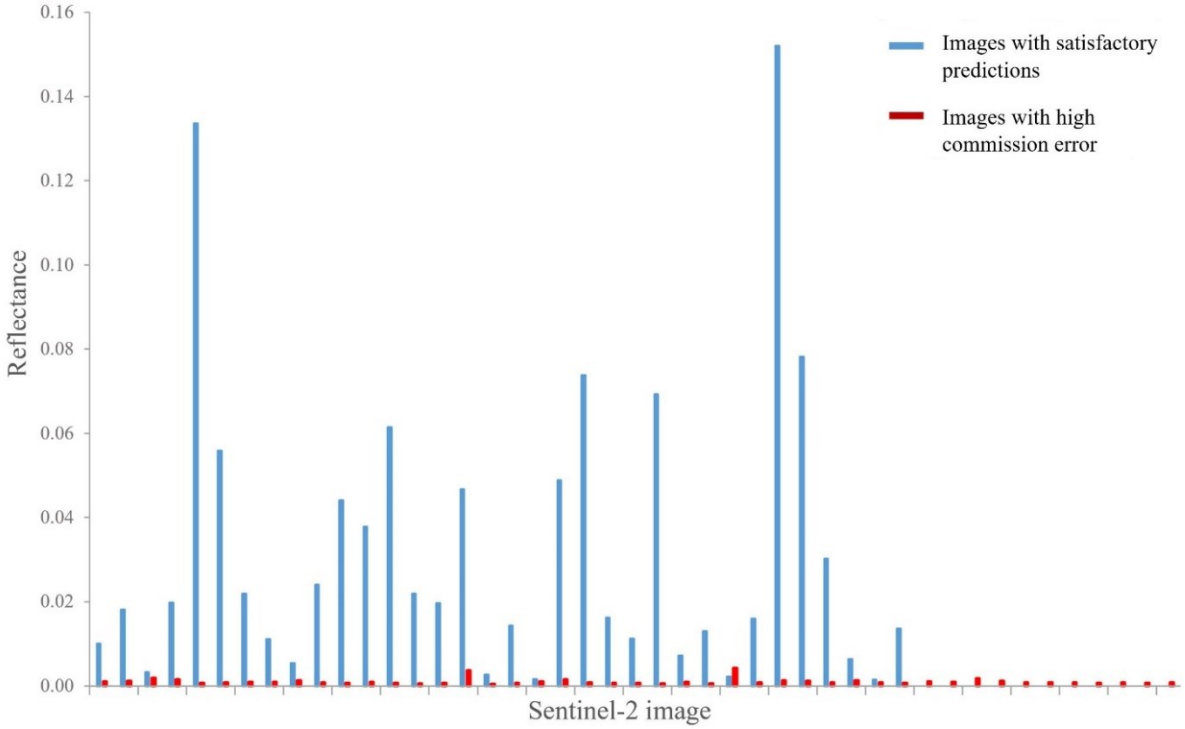


Figure 2.11. Average reflectance values of band 10 (1.374 μm) for the 79 S2 images. (blue): 34 images with satisfactory predictions. (red): 45 images with high commission error

Table 2.13. Importance of the S2 bands for the ANNs trained on the Hollstein dataset

1st configuration		2nd configuration		3rd configuration		4th configuration	
Bands	Importance	Bands	Importance	Bands	Importance	Bands	Importance
11	25.4059	11	18.7082	10	0.8123	11	0.4035
2	14.7908	2	8.8081	7	0.2332	12	0.1133
8A	13.1052	1	7.4302	11	0.2242	8	0.0949
1	8.4542	12	7.2015	8	0.1152	2	0.0867
8	7.4449	4	6.3049	1	0.0613	10	0.0784
10	7.0951	10	5.7476	6	0.0078	7	0.0691
3	6.9086	8	5.5324	5	0.0034	9	0.0606
12	5.08	8A	4.8357	2	0.0031	4	0.0595
9	4.7732	5	4.7502	9	0.003	6	0.0325
5	4.6562	3	4.7488	3	0.0028	5	0.0243
6	4.5948	7	4.3152	4	0.0027	1	0.0147
7	4.3817	9	4.1179	12	0.0023	8A	0.0095
4	4.3519	6	3.926	8A	0	3	0.0094

Table 2.14. Evaluation metrics of the predictions on the S2 spectra training set

Configuration	TP	FP	FN	TN	Accuracy	Precision	Recall	TSS
1st	960,126	66,741	39,874	933,259	0.9467	0.9350	0.9601	0.8934
2nd	950,654	73,443	49,346	926,557	0.9386	0.9283	0.9507	0.8772
3rd	937,267	68,792	62,733	931,208	0.9342	0.9316	0.9373	0.8685
4th	934,203	84,061	65,797	915,939	0.9251	0.9174	0.9342	0.8501

Table 2.15. Evaluation metrics of the predictions on the S2 spectra test set.

Configuration	TP	FP	FN	TN	Accuracy	Precision	Recall	TSS
1st	998,557	139,311	41,443	1,200,689	0.9241	0.8776	0.9602	0.8562
2nd	988,743	153,633	51,257	1,186,367	0.9139	0.8655	0.9507	0.8361
3rd	974,455	144,078	65,545	1,195,922	0.9119	0.8712	0.9370	0.8295
4th	971,340	175,500	68,660	1,164,500	0.8974	0.8470	0.9340	0.8030

demonstrated in general the least satisfactory results since they presented an overall omission error which in several cases was high. Fmask cloud masks showed better results than those of Sen2Cor but were characterized by commission error which in a few cases in water areas with a high presence of sunglint was high (Figure 2.13(e)). MAJA masks presented in our opinion better results than Sen2Cor and Fmask, although it should be stated that a small commission error was usually observed.

The images depicted in Figure 2.12, Figure 2.13, and Figure 2.14 present the results for the different types of cases of the study. For the case depicted in Figure 2.12(a), the MLP mask represented sufficiently the cloud presence of the image, the Fmask mask demonstrated high commission error, the MAJA mask showed high similarity with the MLP mask and Sen2Cor presented very high omission error. For the case depicted in Figure 2.12(b), the MLP mask showed the most acceptable, the MAJA mask presented a commission error which was higher for the Fmask mask and Sen2Cor presented a very high omission error. Regarding the cases with high noise levels and sunglint, for the case of Figure 2.13(a), the MLP mask was overall satisfactory since it was unaffected by sunglint and the oblique periodic noise but it omitted a few thin clouds, the Fmask mask incorrectly classified a large percentage of sunglint areas as clouds, the MAJA mask was similar with the MLP mask but omitted a higher cloud percentage and the Sen2Cor mask presented high omission error. For the case of Figure 2.13(b), the MLP mask was overall acceptable since it was unaffected by the random noise, but slightly underestimated the cloud presence, the Fmask mask presented a commission error which was higher for the MAJA mask and Sen2Cor presented an omission error. Finally, for the case of Figure 2.14, the MLP and MAJA masks were unaffected by sunglint, while Fmask, incorrectly classified a large sunglint area as cloud. Sen2Cor also misclassified a few sunglint pixels to the cloud category. Besides the above observations, the MLP masks seem to be the ones that better represent the natural shape of the clouds, since MAJA masks present the appearance of globs, while Sen2Cor masks show linear structure.

Quantitative comparison of the 1st configuration MLP with state-of-the-art methods

For the quantitative evaluation, evaluation metrics were at first calculated for the total number of the spectra of the S2 spectra dataset (Table 2.16). It was observed that the MLP showed the highest accuracy/TSS scores (~94%/0.89) followed by MAJA (~88%/0.80). The respective scores for Fmask (~82%/0.72) and Sen2Cor (~82%/0.57) were lower with Sen2Cor showing the minimum TSS value. Concerning recall values, except for Sen2Cor which showed high omission error (32%), the other algorithms produced low omission errors (~3%). Finally, regarding precision values, the MLP showed the highest value (~88%) corresponding to the lowest commission error of 12%. Sen2Cor and MAJA produced similar values (~75%/74%) while Fmask showed the lowest score (65%). The phi coefficient was also calculated. Values over 0.4 are considered to show a strong positive correlation while values over 0.7 show a very strong positive correlation.

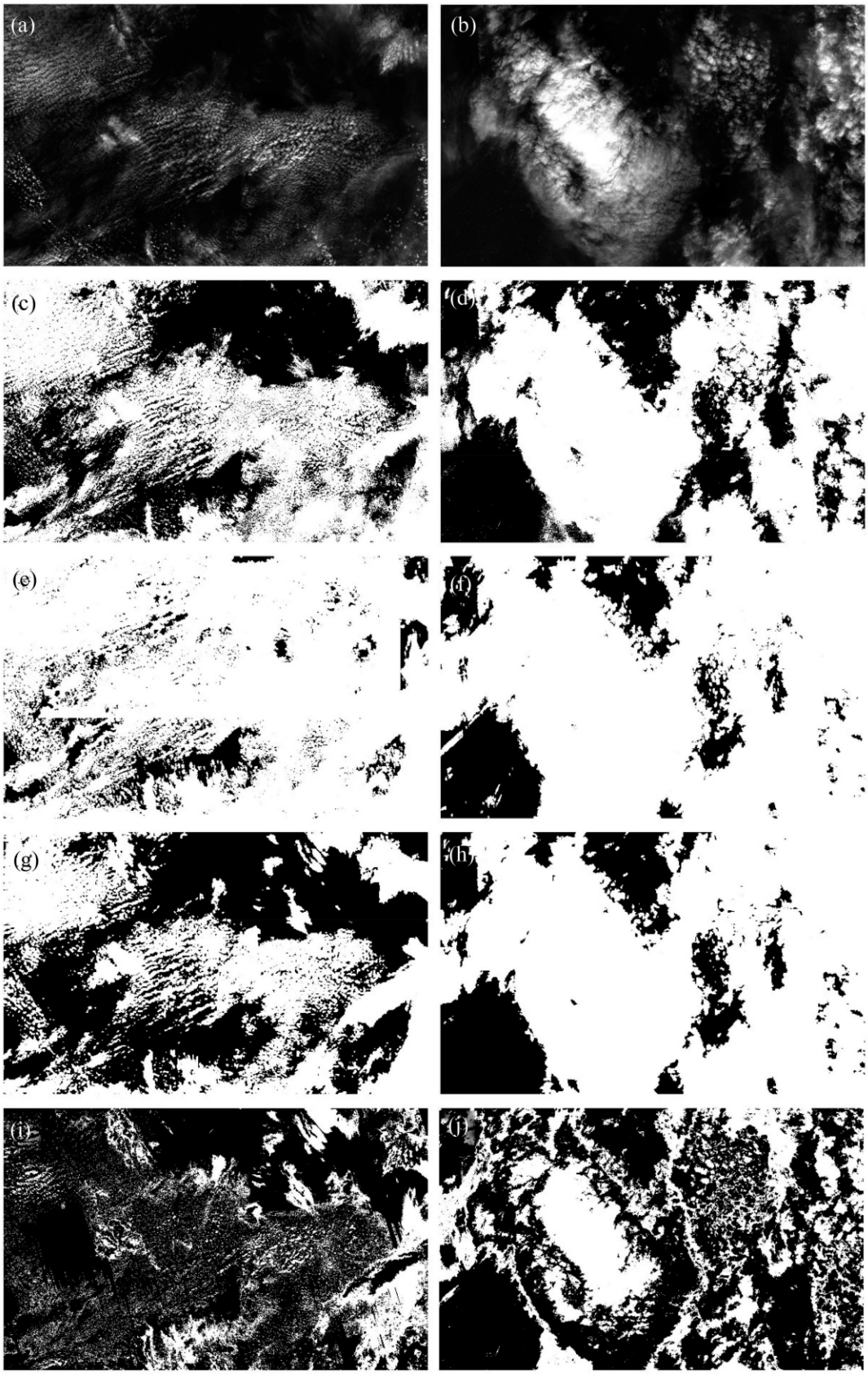


Figure 2.12. a,b): 4-3-2 (RGB) natural color composite, (c,d): MLP cloud mask, (e,f): Fmask cloud mask, (g,h): MAJA cloud mask, (i,j): Sen2Cor cloud mask. The size of all figures is 109.8 × 67.8 km^2

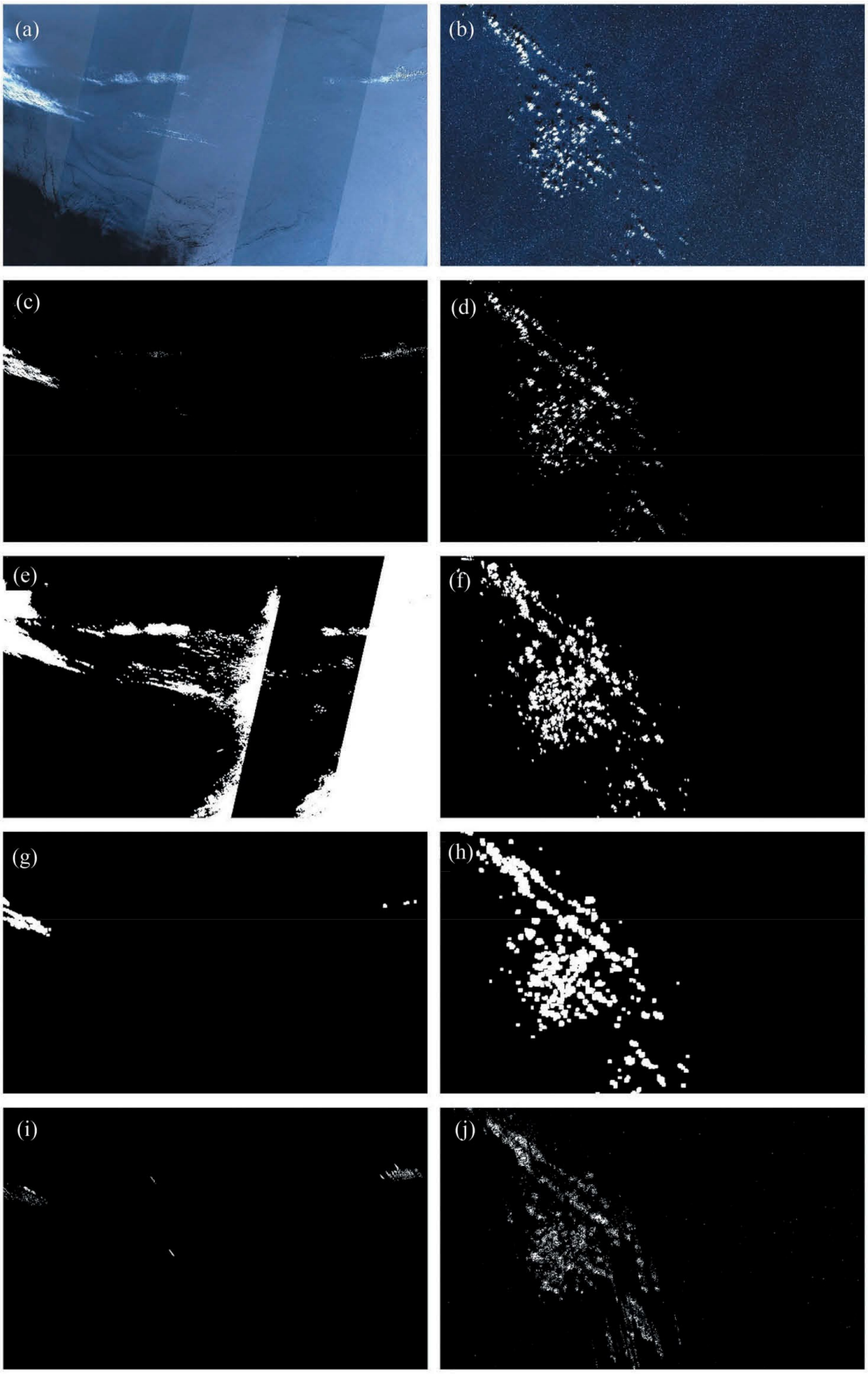


Figure 2.13. (a,b): 4-3-2 (RGB) natural color composite, (c,d): MLP cloud mask, (e,f): Fmask cloud mask, (g,h): MAJA cloud mask, (i,j): Sen2Cor cloud mask. The size of all figures is 109.8 × 67.8 km^2

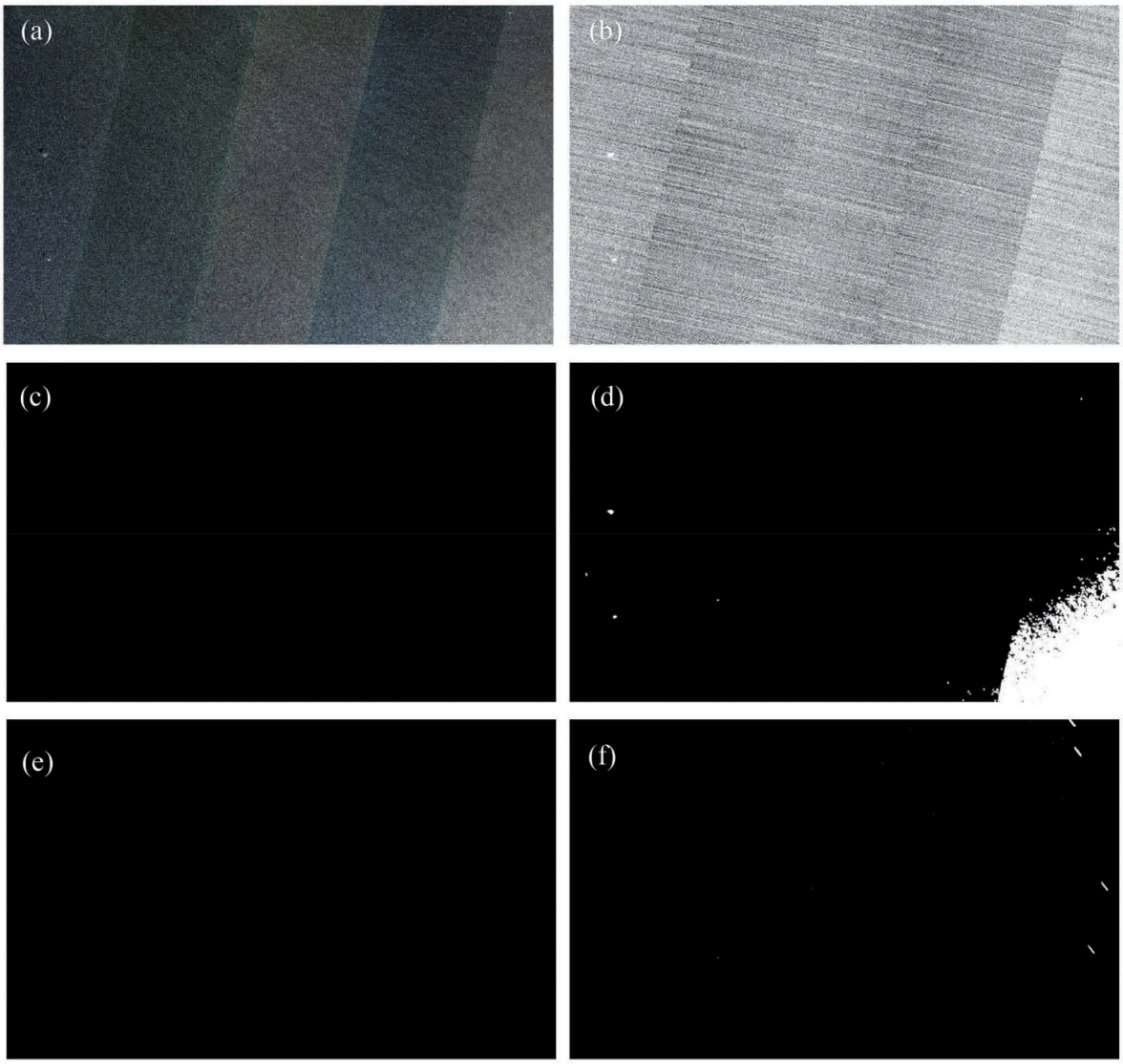


Figure 2.14 (a): 4-3-2 (RGB) natural color composite, (b): cirrus band (1.374 μm), (c): MLP cloud mask, (d): Fmask cloud mask, (e): MAJA cloud mask, (f): Sen2Cor cloud mask. The size of all the figures is 109.8 × 67.8 km^2

Table 2.16. Accuracy, precision, recall, and TSS scores for the S2 spectra dataset (Comparison of algorithms)

Method	TP	FP	FN	TN	Accuracy	Precision	Recall	TSS
MLP (1st configuration)	1,958,683	273,747	81,317	3,899,577	0.9429	0.8774	0.9601	0.8945
Fmask	1,989,338	1,085,049	50,662	3,088,275	0.8172	0.6471	0.9752	0.7152
MAJA	1,973,030	703,422	66,970	3,469,902	0.8760	0.7372	0.9672	0.7986
Sen2Cor	1,383,951	456,874	656,049	3,716,450	0.8209	0.7518	0.6784	0.5689

Figure 2.15 shows the phi coefficient values for the cloud masks of the S2 image dataset. It was observed that the mean value of the phi coefficient between the MLP masks and Fmask masks was 0.58 with a standard deviation of 0.19. Regarding the comparison with the MAJA masks, the mean value of the phi coefficient was 0.65 with a standard deviation of 0.19. Finally, the mean value of the phi coefficient between the MLP masks and Sen2Cor masks was 0.44 with a standard deviation of 0.27. Thus, MAJA masks are more positively correlated with the MLP masks followed by Fmask and Sen2Cor. Table 2.17 shows the phi coefficient values between the

MLP masks and the masks of the other algorithms for the masks depicted in Figure 2.12, Figure 2.13, and Figure 2.14.

It was observed that the MLP mask of Figure 2.12(a) showed a strong positive correlation with the MAJA mask, a lower positive correlation with the Fmask mask, and a negative correlation with the Sen2Cor mask. In addition, the MLP mask of Figure 2.12(b) showed a very strong positive correlation with the MAJA mask, a strong correlation with the Fmask mask, and a negative correlation with the Sen2Cor mask. Concerning the images with high noise levels and sunglint, the MLP mask of Figure 2.13(a) showed a strong positive correlation with the MAJA mask and a very low positive correlation with the Fmask and Sen2Cor masks, while the MLP mask of Figure 2.13(b) showed strong positive correlation for Fmask and MAJA masks and lower for the Sen2Cor mask. In addition, the MLP mask of Figure 2.14(a) showed a strong positive correlation with the MAJA mask followed by the Sen2Cor mask and no correlation with the Fmask mask.

From the above, it is concluded that the quantitative evaluation is in accordance with the visual evaluation. In more detail, the MLP masks present the highest accuracy/TSS scores and are more correlated with the MAJA masks which present the second best highest accuracy. In addition, the least satisfactory results are presented by Sen2Cor masks and Fmask presents the highest commission error. It should be also noted that an additional advantage of the MLP is that it is more time-efficient than Fmask, MAJA, and Sen2Cor since the mask can be created in seconds (inference time), while the other algorithms need at least 10 min (STEP 2016a) [175][176].

B. Observation of the weights of the first hidden layer

The weights of the first hidden layer for the four configurations were also observed for the MLPs trained on the S2 spectra dataset since as already mentioned they represent the importance of the bands for the MLP. Table 2.18

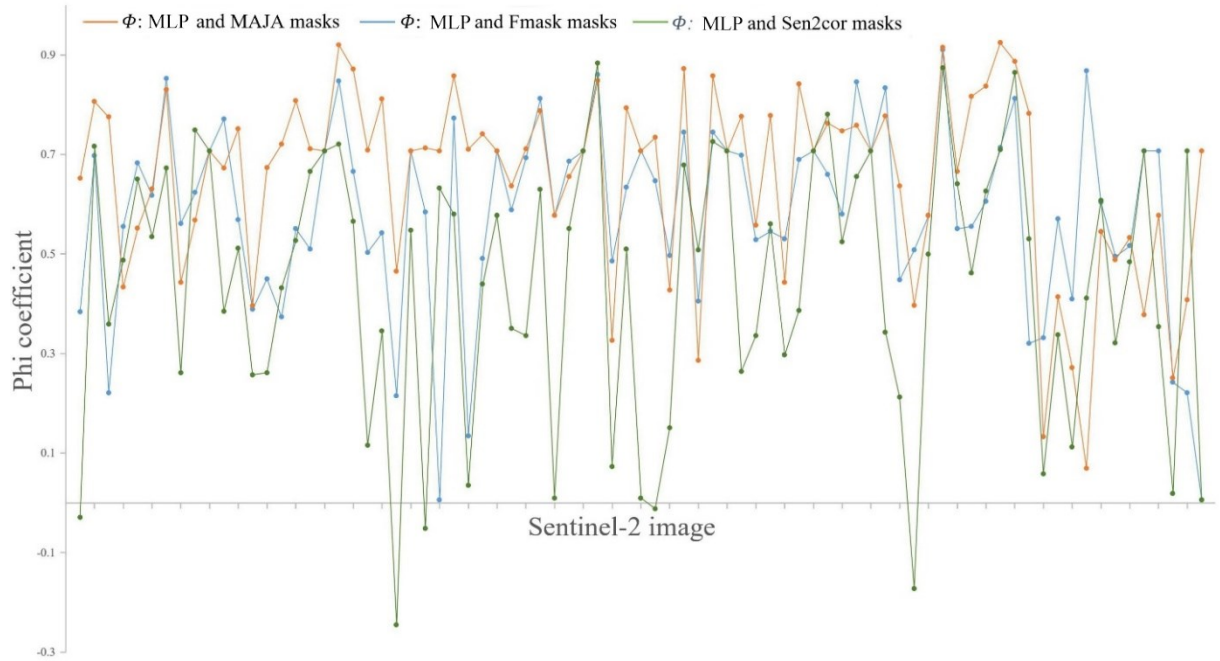


Figure 2.15. Phi coefficient for the cloud masks of the S2 image dataset. Brown: phi coefficient between ANN and MAJA masks, Blue: phi coefficient between ANN and Fmask masks, Green: phi coefficient between ANN and Sen2cor masks

Table 2.17. The phi coefficient values for the masks of the images depicted in Figures 2.12, 2.13, 2.14

Cloud mask	Phi Fmask	Phi MAJA	Phi Sen2Cor
Figure 2.12(*a*)	+0.38	+0.65	-0.03
Figure 2.12(*b*)	+0.65	+0.73	-0.01
Figure 2.13(*a*)	+0.13	+0.71	+0.03
Figure 2.13(*b*)	+0.53	+0.44	+0.29
Figure 2.14(*a*)	+0.00	+0.71	+0.63

is created in the same way as Table 2.13 and likewise shows for each configuration in descending order the importance of the bands (Equation (2.20)). It was observed that during the training with this dataset, band 11 (1.614 μm) is given high weights in all configurations, a behavior which is similar to the one previously observed in the training on the Hollstein dataset. However, it was also observed that the cirrus band (1.374 μm) acquired a high ranking, a behavior that can be explained by the fact that this band is less affected by sunglint since it corresponds to a strong absorption band of water vapor. High clouds have a high chance of being visible in this band in contrast to low clouds because incident and reflected light are highly absorbed. Bands 1 (coastal aerosol band/444 nm) and 9 (water vapor absorption band/945 nm) which are typically used for atmospheric correction purposes acquired high ranking as well. Compared to the ranking of the importance of the bands on the training on the Hollstein dataset, these changes directly indicate the bands that counteract the influence of the presence of spectra with high noise levels and sunglint in the training set.

2.2.4 Conclusions and discussion

In this study, MLPs were trained with four different configurations on a dataset extracted by the Hollstein et al. (2016) [43] database and on a dataset that was created by the extraction of spectra from 79 S2 level 1C images that were used for this study. The second dataset adequately represented different types of water, namely, it included water with high noise levels and sunglint. The four configurations were differentiated by the use of different algorithms that prevent overfitting but it was observed that these algorithms slightly affected the performance of the MLPs. Since the configuration of the MSI S2 imaging mission leads to a broad range of viewing geometries, the developed MLP was tested for directional reflectance effects only by the use of spectral information.

The MLPs trained on the Hollstein dataset were used in three experiments which were differentiated by the test set used for the predictions and by the feature scaling parameters of the test set. When the test set consisted of spectra from the Hollstein dataset, the evaluation metrics were over 0.99 in all configurations. However, the cloud masks of the S2 images produced by using the MLP of the first configuration presented high commission errors when applying to the spectra of the images the feature scaling parameters of the training set. This behavior leads to the conclusion that the dataset of the S2 images used in this study cannot be adequately represented by the Hollstein dataset, since spectra from deep water areas or spectra with high noise levels and sunglint are scarce. Interesting results were produced when applying on the spectra of the images the feature scaling parameters

Table 2.18. Importance of the S2 bands for the MLPs trained on the S2 spectra dataset

1st configuration		2nd configuration		3rd configuration		4th configuration	
Bands	Importance	Bands	Importance	Bands	Importance	Bands	Importance
1	14.5757	1	11.0918	10	1.4623	10	1.0388
9	11.3767	10	10.9864	1	1.2076	11	0.7415
11	10.2176	9	10.6795	11	0.5400	1	0.7076
10	8.9378	11	10.1314	12	0.4906	9	0.5820
12	8.7829	3	7.5562	9	0.1506	8A	0.4642
3	6.9809	12	6.7347	8	0.0944	12	0.4026
8A	6.9176	8	6.1217	3	0.0889	8	0.3834
5	5.4747	8A	5.3496	8A	0.0817	7	0.345
8	5.4138	4	4.1833	4	0.0747	2	0.3381
7	4.6918	5	4.1448	7	0.0084	3	0.3071
2	4.5922	2	4.0972	2	0.0077	6	0.3050
6	3.9374	7	3.4356	5	0.0074	5	0.3047
4	3.7180	6	1.7959	6	0.0066	4	0.2559

that corresponded to the images instead of the training set spectra, where acceptable results were produced for 34 images, which showed that using the feature scaling parameters of the test set could be a factor that could alter the predictions of an MLP towards a positive direction.

The overall accuracy of the MLP of the first configuration trained on the spectra dataset extracted from S2 images with several levels of noise and sunglint was over 0.92 on the test set. In addition, the TSS score was over 0.89. The predictions of this MLP were evaluated by visual observation and compared with the results produced by three state-of -the-art algorithms: Fmask, MAJA, and Sen2Cor. Its predictions were considered to be very favorable compared to the above-mentioned state-of-the-art algorithms. It produced robust results since none of the 79 cases presented outlier classification output and it proved to be unaffected by water areas with high noise levels and sunglint. In addition, its masks better represented the natural shape of the clouds. Although a small omission error was overall observed, the results were acceptable in all cases. The quantitative evaluation which was in accordance with the visual, showed that the MLP produced the highest accuracy/TSS scores and presented the strongest correlation with the MAJA masks followed by Fmask and Sen2Cor. As a general conclusion, this MLP showed the best performance not only concerning the quality of the masks but also the time efficiency.

The weights of the first hidden layer for the four configurations were observed since they represent the importance of the bands for the MLP and a simple importance measure was defined. It was observed that band 11 (1.6 μm) was given high weights in all configurations. In addition, when observing the weights of the MLPs trained on the dataset with high noise levels and sunglint extracted from the S2 images, it was demonstrated that the cirrus band which is less affected by sunglint and two bands typically used for atmospheric correction (444 and 945 nm) are the ones responsible for the mitigation of high noise levels and sunglint in the training set. Regarding the rest of the bands, the ranking of importance greatly varied, a behavior which can be explained by the fact that a variety of VNIR and SWIR bands carry useful information for cloud detection applications.

There is no perfect method to mask all clouds and retain water pixels, but this study proved that MLPs are a simple, fast, and effective cloud masking algorithm that can avoid the influence of deep-water areas with high noise levels and sunglint. The developed MLP successfully detects clouds on S2 images which present serious directional reflectance effects. It also showed that the database created by Hollstein et al. (2016) needs to be expanded with more spectra from deep water areas since successful MLP results are closely connected with training on a dataset that adequately represents the wide variability of cloud and water spectra. Finally, it was shown that there are cases where making predictions using the feature scaling parameters of the test set instead of those of the training set can improve the MLP results when the test set is not adequately represented by the training set. The possibility of generalizing this finding in other applications could be further investigated in future work.

The main contributions of this study are: a) the production of a manual dataset of water spectra with noise and sunglint which was made publicly available (2,133,324 spectra), b) the implementation of an MLP architecture that outperformed the state-of-the-art algorithms (Fmask (threshold-based), MAJA (temporal), Sen2Cor (threshold-based + SOMs)), c) the creation of a measure that indicates the bands that mitigate the influence of deep water spectra with noise/sunglint based on the magnitude of the weights of the first hidden layer, and d) the investigation of applying feature scaling on the test set.

2.3 Fine-Tuning SOMs for Sentinel-2 (S2) Imagery: Separating Clouds from Bright Surfaces[2]

In section 2.3, the second cloud masking approach is presented. SOMs are fine-tuned in S2 data to mitigate the effect of bright non-cloud objects caused by sunglint in land areas. In section 2.3.1 at first an overview of SOMs and their advantages is provided along with a definition of fine-tuning. Then, a general workflow of the proposed methodology is described. In this section, the motivations and objectives of the study can be derived. In section 2.3.2 a description of the datasets and the SOM theory is provided together with a detailed presentation of the

[2] **Kristollari, V.** and Karathanassi, V., 2020. Fine-tuning Self-Organizing Maps for Sentinel-2 imagery: Separating clouds from bright surfaces. MDPI Remote Sensing, 12(12), p.1923. doi:10.3390/rs12121923

methodology. In section 2.3.3 a thorough evaluation of the trained SOM and the resulting cloud masks before and after fine-tuning is presented. In section 2.3.4 the main points that distinguish the proposed method from common ANNs, threshold-based methods, and k-means are discussed. Finally, in section 2.3.5 the conclusions and the contributions are summarized. Future work is also stated.

2.3.1 Introduction

SOMs [177][178] are a type of competitive ANN that projects data of high dimensionality to a space of low dimensionality by simultaneously preserving topology relations. SOMs are related to vector quantization, with the difference of conservation of topologic information that makes them suitable for the organization and visualization of complex datasets. Their concept is based on the associative neural properties of the brain where neurons operate in a localized manner [179]. Contrary to other types of ANNs, SOMs do not perform error-correction learning but interpret the input information by the location of the response in the low-dimensional space without taking into account its magnitude. Even though SOMs are an unsupervised learning method, the produced clusters can be labeled given that available ground-truth data exists and consequently the clusters can be converted to classes. Majority voting is the common approach to define the labels of the classes, represented by the neurons of the produced map [177][180][181]. SOMs are weakly represented in current machine learning cloud masking research even though recent studies [30][139][140][141] report successful results with the additional advantage of faster training/fine-tuning time and more interpretative behavior (preservation of topologic relations) compared to other types of ANNs. This fact led to their inclusion in the creation of the operational cloud masking products of S2 [109] and Proba-V [142] satellites.

The term "fine-tuning" for other types of neural networks (e.g., CNNs) refers to the use of pre-trained neural networks for different applications [133] [137][138] than those that they were originally trained for. During fine-tuning, the pre-trained weights are used as initial weights and the network is further trained on the new dataset. As for SOMs, in the cases where fine-tuning is performed, the weights of the map neurons are updated through further training by taking into account the correctness or incorrectness of the prediction [177][181]. Fine-tuning is supposed to highly increase classification accuracy in SOMs [177].

This study evaluates a SOM for cloud masking S2 images and proposes a fine-tuning methodology based on the output of the non-fine-tuned network. The fine-tuning process does not require further training to correct the misclassified predictions of bright non-cloud spectra. It is important to note that the fine-tuning method follows a general procedure, thus its applicability is broad and not confined only to the field of cloud-masking. The study takes direct advantage of the similarities of the SOM to a brain map. In more detail, it is based on the fact that a detailed topographical map of the cerebral cortex of the brain can be deduced by various functional or behavioral impairments, or through stimulation of a particular site which leads to the disruption of a cognitive ability [177]. A SOM is trained on a spectral database created by Hollstein et al. 2016 [43] (the largest publicly available for S2 cloud applications at the time the study was conducted (2020)) and is tested on a truly independent (non-overlapping) database of S2 cloud masks created by Baetens et al. [44] in 2019 which is also publicly available. The trained SOM neurons are labeled through majority voting by use of the ground-truth labels provided in the training database. Finding the neuron with the minimum Euclidean distance from each pixel of the images of the test database leads to the production of the predicted cloud masks. In the next step, after observation of the cloud masks produced by the non-fine-tuned SOM, the fine-tuning process is applied. During fine-tuning, the SOM neurons that correspond to the bright misclassified non-cloud areas are detected by feeding the corresponding incorrectly classified spectral signatures into the network and consequently identifying the stimulated neurons. Then, the incorrect labels of the respective neurons are directly altered without applying further training. The network is evaluated both qualitatively and quantitatively with the interpretation of its behavior through multiple visualization techniques being a main part of the evaluation. The cloud masks are not only compared with ground-truth data but also with results produced by two state-of-the-art algorithms: Sen2Cor and Fmask. It is noted that in the context of this study, the term "bright non-cloud areas" refers to built-up areas, soils (e.g., desert), and coastal surfaces. It is also mentioned that the fine-tuning methodology proposed in this study was also applied in experiments that specifically targeted incorrectly classified snow pixels. Yet, the results were considered

unsatisfactory since the correct classification of the snow pixels led to a large omission error of clouds. Those experiments are not presented in the study.

2.3.2 Materials and methods

2.3.2.1 Data Description

A. Training Set

The training set consisted of 8,799,998 S2 reflectance spectra in total which form the database created by Hollstein et al. 2016. Information on this database has been provided in section 2.2.2.1.A. Likewise, information on the available spectral bands of S2 has been provided in section 2.2.2.1.B. Figure 2.16 depicts their location with black circles. Their processing level is 1C which denotes that they are not atmospherically corrected, thus they are suitable for the application of cloud masking methods. It contains six classes ("opaque cloud", "cirrus", "snow", "shadow", "water", "clear (land)"). Table 2.19 presents the number of spectra for each class.

B. Test Set

The test set consisted of 34 S2 level 1C images. Their corresponding cloud masks were provided by the database created by Baetens et al. [44]. This database was the only publicly available source of S2 ground-truth cloud masks at the time the study was conducted (2020). The creation of the masks was based on the application of Random Forest and their accuracy is reported to be 98%. The images cover different areas around the world with various land cover and cloud properties: three images were collected in North America, four in South America, nine in Africa, and 18 in Europe. The collection dates cover all seasons of the year: seven images were collected in winter (December, January, February), eight in spring (March, April, May), 11 in summer (June, July, August), and eight in fall (September, October, November) between seven a.m. and six p.m. UTC. Before feeding the spectra of the

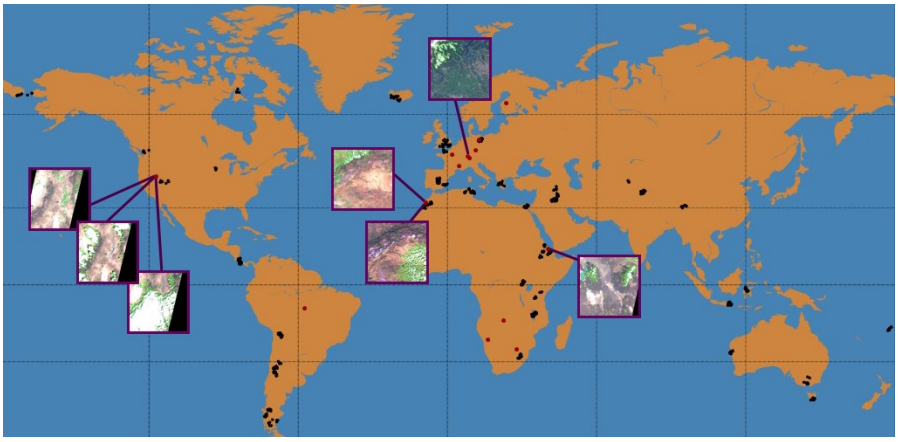


Figure 2.16. Location of the images of the training set (black circles) and the test set (red circles). The thumbnails depict cases with bright non-cloud objects

Table 2.19. Spectra comprising the training set

Class	Coverage	Number of Spectra
opaque cloud	opaque clouds	1,500,202
cirrus	cirrus and vapor trails	1,205,979
snow	snow and ice	1,271,143
shadow	shadows from clouds, cirrus, mountains, buildings, etc.	1,113,066
water	lakes, rivers, sea	1,435,003
land	remaining: crops, mountains, urban, etc.	2,274,605

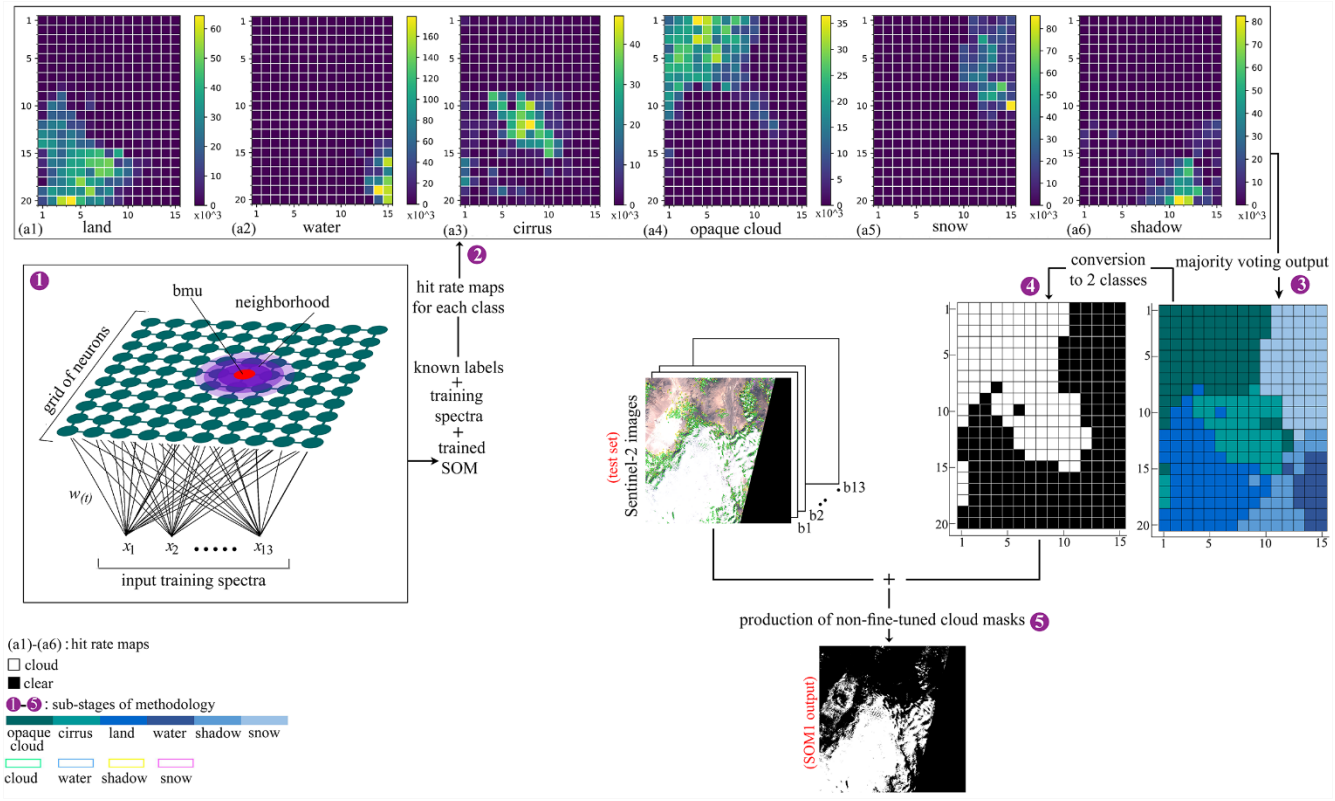


Figure 2.17. Sub-stages of the methodology of the study: self-organizing map (SOM) training (Stage 1) and production of non-fine-tuned cloud masks (Stage 2: sub-stages 2–5)

images into the SOM, the bands with spatial resolutions of 10 and 20 m were resampled to 60 m since cloud masking applications do not require higher spatial resolution. It is noted that the effect of the spatial resolution difference between the training set (20 m) and the test set (60 m) is expected to be insignificant. In addition, the lower resolution (60 m) significantly improves the inference time of the SOM network because the final size of the images is 1830 × 1830 pixels instead of 5490 × 5490. Figure 2.16 depicts the location of the images with red circles. It is mentioned that each image covers an area of 109.8 × 109.8 km^2. It is also worth noting that the test set is truly independent of the training set because the spectra do not overlap.

2.3.2.2 Theoretical Background

The SOM was introduced by Kohonen [177][178]. It is a shallow ANN architecture that consists of an input layer and an output layer depicted as a 2-dimensional (2D) grid. The output layer is fully connected to the input layer and is made up of a set of neurons (nodes) that represent feature vectors. The neurons are interconnected through receptive fields called neighborhoods and the coefficients of the vectors represent the weights of the network. The size of the vectors of the output layer is the same as the size of the vectors of the input layer. Before starting the training process, random initialization of the feature vectors of the neurons is commonly performed. During training, the network reads a random datapoint and the distance (e.g., Euclidean) of the input with all feature vectors (neurons) is computed. The neuron that presents the minimum distance is the winning neuron which is called the Best Matching Unit (BMU). The neurons of the neighborhood are also activated and the distance of their feature vectors from the input datapoint is reduced. The above process is repeated until a pre-defined stopping criterion (e.g., number of iterations) is met. In each iteration, both the learning rate and the neighborhood are decreased to ensure convergence. A graphical representation of the SOM training process is illustrated in the lower left part of Figure 2.17 (stage 1 of methodology).

2.3.2.3 Proposed Methodology

The proposed method consists of three main stages. During the first stage, the SOM is trained on the spectra of the training set. During the second stage, the non-fine-tuned cloud masks of the test set are created. This process involves the calculation of the hit rate maps for each class of the training set and the labeling of the SOM neurons through majority voting. The trained SOM before applying the fine-tuning process will be called "SOM1". Finally, during the third and final stage, the fine-tuning process is applied. It includes the manual sampling of incorrectly classified bright non-cloud pixels by SOM1 and the correction of the labels of their corresponding neurons. By use of the corrected SOM, the temporary fine-tuned cloud masks are produced. Applying a median and a dilation filter leads to the creation of the final fine-tuned cloud masks. We will refer to the final fine-tuned cloud masks as "SOM2" output. The analysis of the three main stages is written in sections 2.3.2.2.A, 2.3.2.2.B, and 2.3.2.2.C. In addition, the sub-stages of the methodology are depicted in Figure 2.17 and Figure 2.18 and are also listed below. The second stage includes the sub-stages represented by the numbers 2–5 of the list and the third stage includes the sub-stages represented by the numbers 6–9.

1. Training of the SOM with spectra randomly selected from the training set.
2. Production of hit rate maps for each class by feeding the complete training set with known spectra labels into the trained SOM and detecting the BMUs.
3. Labeling of the SOM neurons through majority voting.
4. Conversion of the six classes to two, which define the cloud and non-cloud classes.
5. Production of non-fine-tuned cloud masks (SOM1 output) for the entire test set.
6. Observation of the temporary non-fine-tuned cloud masks and manual sampling of incorrectly classified bright non-cloud pixels.
7. Detection of the neurons that correspond to the bright non-cloud pixels and correction of their labels.
8. Production of the temporary fine-tuned cloud masks.
9. Application of a median and a dilation filter and production of the final cloud masks (SOM2 output).

A. Training Process

This section presents the training process of the SOM which represents the first stage of the methodology (Figure 2.17). The study implemented a SOM according to the Python code for unsupervised learning created by Riese et al. (2019) [31]. Before starting the training of the network, a feature scaling process was applied on every wavelength of the S2 signatures of the training set (min-max normalization) (Equation (2.22)) in order to impede variables of higher magnitude to prevail over variables of lower magnitude.

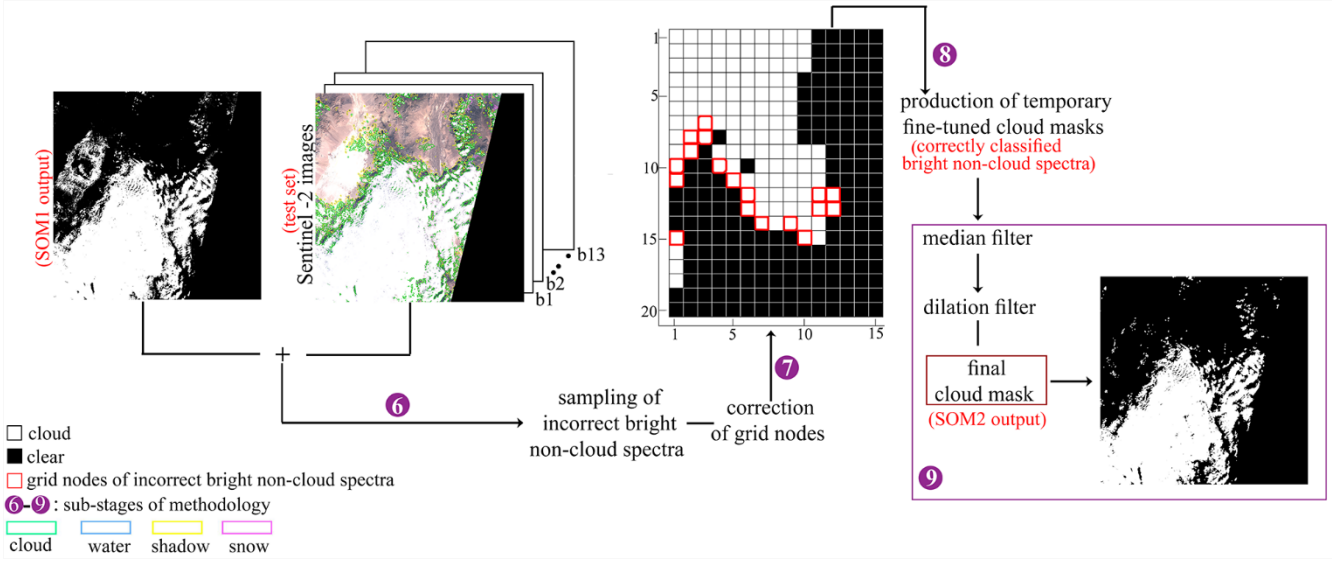


Figure 2.18. Sub-stages of the methodology of the study: Fine-tuning process (Stage 3: sub-stages 6–9)

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} + x_{\text{min}}} \quad (2.22)$$

where x: the value of a signature in a given band, x_{min}: minimum value of all signatures of the dataset in a given band and x_{max}: maximum value of all signatures of the dataset in a given band.

The training was performed by using all categories of the training set (six) and all S2 bands (13). The size of the 2D rectangular grid was selected to be $20_{rows} \times 15_{columns}$ nodes. There is no rule about the selection of the grid size, but it is noted that it is the SOM parameter that mostly affects the processing time. The values of the feature vectors of the neurons were initialized by taking random samples from a uniform distribution in the interval [0, 1). The distances between the input datapoints and the feature vectors of the neurons were calculated according to the Euclidean distance (Equation (2.23)) and the learning rate according to the equation proposed by Barreto and Araujo [182] (Equation (2.24)). The start value of the learning rate was set to 0.5 and the end value to 0.05.

$$d(\boldsymbol{x}, \boldsymbol{w}) = \sqrt{\sum_{j=1}^{n} (x_j - w_j)^2} \quad (2.23)$$

where x: the input datapoint, w: the SOM node, and n: the dimension of the vectors x, w

$$a(t) = a_0 \cdot \left(\frac{a_{\text{end}}}{a_0}\right)^{\frac{t}{t_{\text{max}}}} \quad (2.24)$$

where a_0: the start value of the learning rate, a_{end}: the end value of the learning rate, t: the number of the current iteration, and t_{max}: the number of maximum iterations.

As for the number of iterations the "rule of thumb" proposed by Kohonen [177] stating that "the number of maximum iterations should be at least 500 times the number of network units" was followed ($\geq (15 \times 20 \times 500)$, i.e., $\geq$ 150,000) and 1,000,000 iterations were performed. Decreasing learning rates are often implemented in ANNs to increase convergence and prevent oscillations [31].

The neighborhood radius was calculated according to the equation proposed by Matsusita et al. [183] (Equation (2.25)). The start value of the neighborhood function (radius) was chosen to be $\max(n_{rows}/2\ ,\ n_{columns}/2\ )$ as usually suggested.

$$\sigma(t) = \sigma_0 \cdot \left(1 - \frac{t}{t_{\text{max}}}\right) \quad (2.25)$$

where σ_0: start value of the neighborhood radius (t, t_{max}: as defined in Equation (2.24)).

The neighborhood distance weight which is dependent on the neighborhood radius and the Euclidean distance between the BMU and every other node on the SOM grid was calculated by Equation (2.26), proposed also by the researchers mentioned above. This equation is called "Pseudo-Gaussian".

$$h_{c,i}(t) = \exp\left(-\frac{d(c,i)^2}{2\sigma(t)^2}\right) \quad (2.26)$$

where $d(c, i)$: distance between the BMU c and node i on the SOM grid ($\sigma(t)$: as defined in Equation (2.25))

Algorithm 1: SOM training process

Input: training set
Input: start value of the learning rate a_0
Input: end value of the learning rate a_{end}
Input: start value of the neighborhood function σ_0
Input: number of maximum iterations t_{max}
Output: trained SOM
1: Generate random weights $\boldsymbol{w}_i(t)$
2: Set the number of the current iteration equal to 1 ($t = 1$)
3: **while** $t < t_{max}$ **do**
4: Get random input datapoint $\boldsymbol{x}(t)$
5: Find BMU $c(\boldsymbol{x})$
6: Calculate learning rate $a(t)$
7: Calculate neighborhood function $\sigma(t)$
8: Calculate neighborhood distance weights $h_{c,i}(t)$
9: Modify weights $\boldsymbol{w}_i(t + 1)$
10: $t \leftarrow t + 1$
11: **end while**

Finally, the weights of the SOM after each iteration were updated according to Equation (2.27).

$$w_i(t+1) = \boldsymbol{w}_i(t) + a(t) \cdot h_{c,i}(t) \cdot (\boldsymbol{x}(t) - \boldsymbol{w}_i(t)) \quad (2.27)$$

where $w_i(t)$: vector of the weights of node i at iteration t and $x(t)$: datapoint at iteration t ($a(t)$: as defined in Equation (2.24), $h_{c,i}(t)$: as defined in Equation (2.26)).

The steps of the training process are depicted in Algorithm 1. After defining the parameters of the network, the training was performed on a CPU (i7-8th generation, 3.7 GHz). It was a rapid process that lasted approximately two minutes.

B. Production of Non-Fine-Tuned Cloud Masks

This section presents the second stage of the methodology (Figure 2.18) which leads to the production of the non-fine-tuned cloud masks. The non-fine-tuned cloud masks of the test set were produced after the training process was completed. Labeling of the SOM neurons is the condition that needs to be fulfilled before the creation of the masks. The labeling was accomplished through majority voting which was applied on the hit rate maps of each class. The hit rate map (Figure 2.17 (a1-a6)) is a visualization technique that denotes the number of times a neuron was detected as a BMU. For their computation, every single signature of the training set (~nine million signatures) was fed into the network and the BMU was detected. As already explained in section 2.3.2.2, the BMU that corresponds to a spectral signature is the node that presents the minimum distance. Then, the "hits" were computed for each of the six classes. The creation of the hit rate maps for each class was possible because the labels of the training data were known. The computation of the BMUs for the training data lasted approximately 18 min. After the calculation of the hits, the majority voting was applied where each neuron was assigned the label of the class that corresponded to its maximum hits, e.g., in case the neurons were classified as opaque cloud, the following condition was true (Equation (2.28)):

$$N_{\text{opaque cloud}} > N_{\text{cirrus}} \,\&\, N_{\text{opaque cloud}} > N_{\text{land}} \,\&\, N_{\text{opaque cloud}} > N_{\text{water}} \,\&\, N_{\text{opaque cloud}} > N_{\text{shadow}} \,\&\, N_{\text{opaque cloud}} > N_{snow} \quad (2.28)$$

where N: Number of times a neuron was detected as a BMU for a class.

As a final step, the classes: opaque cloud and cirrus were joined to a class that will be called “cloud” and the classes: land, water, shadow, and snow were joined to a class that will be called “non-cloud”, to produce the final labeled SOM which contains two classes. The cloud masks of the test set were created by locating the BMU that corresponded to each signature (presented the minimum Euclidean distance) and retrieving the respective label. The process of locating the BMUs for all 1830 × 1830 = 3,348,900 pixels for each S2 image (inference time) lasted ~six min.

C. Production of Fine-Tuned Cloud Masks

This section presents the third and final stage of the methodology (Figure 2.18) which leads to the production of the fine-tuned cloud masks. The fine-tuning process aimed at correcting the incorrectly predicted labels of bright non-cloud objects by the trained SOM. The fine-tuning process was applied after the non-fine-tuned cloud masks of the entire test set had been created and observed. For its implementation, after the observation of the non-fine-tuned cloud masks, a sample of misclassified bright pixels (305,228 pixels in total) was selected from four images and the corresponding BMUs were detected. It is noted that the images that presented a high number of misclassified bright non-cloud spectra, were seven in number i.e., ~20% (7/34) of the global test set, which is a high percentage. The output of this process was the detection of the BMUs that represent the bright non-cloud object signatures of the sampled pixels. These BMUs were 18 in number, thus 6% of the total SOM neurons (18/15×20). After the detection of the location of these 18 neurons, their labels were altered from cloud to non-cloud and the temporary fine-tuned cloud masks were produced. It is noted that only the BMUs that corresponded to a number of hits larger than ~5% of the maximum number of hits for each image, were taken into account. This threshold was derived through a trial and error process which was based on evaluation of the temporary fine-tuned cloud masks. Since these 18 neurons did not exclusively represent the bright non-cloud objects but also a few cloud pixels, a median and a dilation filter of size 3 × 3 were implemented on the temporary fine-tuned cloud masks in order to compensate both for remaining omission and commission errors. Figure 2.18 depicts the locations of the altered BMUs. As expected they are located in the borders of the classes. In addition, Figure 2.19 depicts the BMUs and the number of hits for the sampled non-cloud spectra regarding two of the four images that were used in the sampling process. It can be observed that the number of the activated neurons (BMUs) differs, a fact that can be explained by the different spectral variability of the classes. This Figure also illustrates the corrected nodes in red rectangles. The thumbnails of the images where misclassified bright non-cloud pixels occurred are depicted in Figure 2.16.

2.3.3 Results

The non-fine-tuned network was at first evaluated (a) by employing several visualization techniques and (b) by calculating the confusion matrix on the training set. Then, an evaluation of the fine-tuned (SOM1) and the non-fine-tuned (SOM2) versions was performed on the cloud masks of the test set.

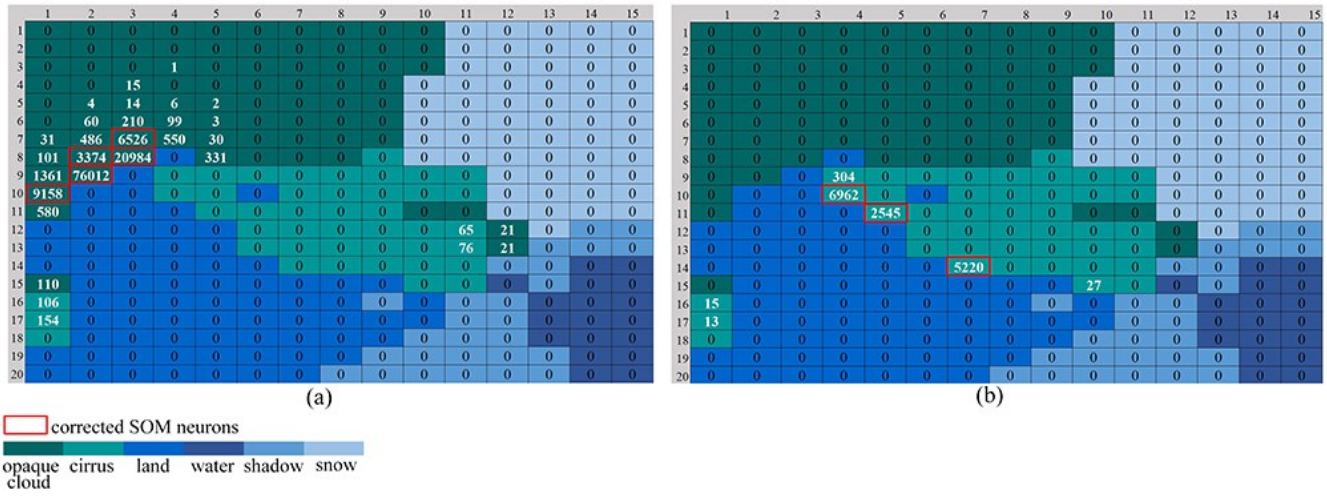


Figure 2.19. Number of hits for the sampled non-cloud spectra. (a) Image with a high number of activated neurons, (b) image with a low number of activated neurons

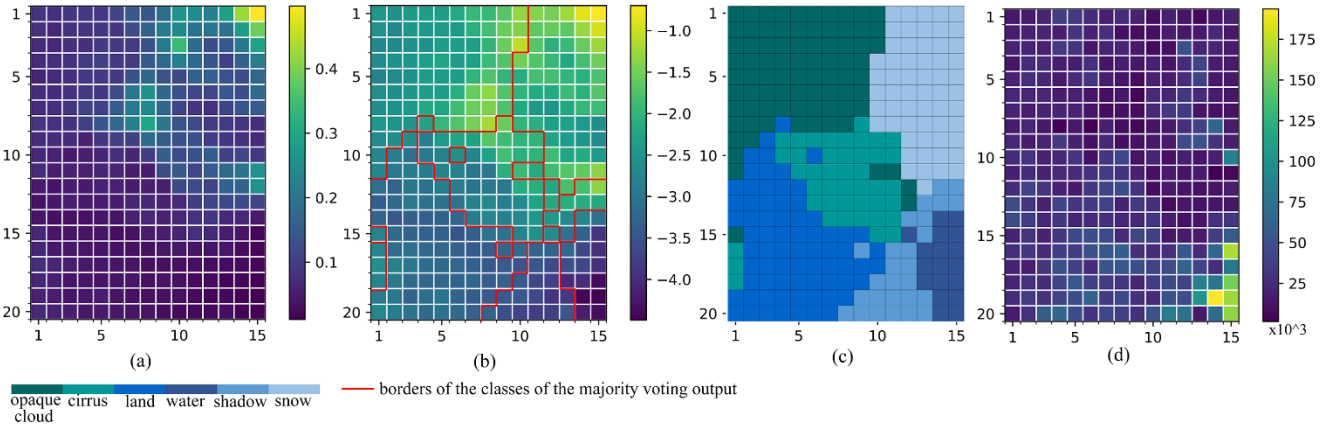


Figure 2.20. (a) U-matrix, (b) U-matrix with logarithmic scale, (c) majority voting output (d) hit rate map

2.3.3.1 Visualization Techniques

The trained SOM (SOM1) was evaluated through several visualization techniques which include: (a) the U-matrix [184], (b) the hit rate map, (c) the component planes, and (d) 2D scatterplots. These techniques are useful for the interpretation of the behavior of the network.

A. U-Matrix

The U-matrix, introduced by Ultsch (1990) [184], is a very widespread method for visualizing the SOM clusters. It is obtained by calculating the distance (Euclidean in this case) between the neurons that are neighbors. Small distances denoted by low U-matrix values are interpreted as similar data while high values occur in clusters of higher variability or borders of clusters. The U-matrix produced by the SOM implemented in this study is depicted in Figure 2.20(a). Figure 2.20(b) presents the U-matrix with a logarithmic scale with the borders of the classes of the majority voting output overlayed for easier visual interpretation. By observing simultaneously the U-matrix and the clusters of the majority voting output (Figure 2.20(c)), it can be concluded that by visual observation the borders between the clusters could be deduced. The water class appears to have the smallest variability.

B. Hit Rate Map

The hit rate map (Figure 2.20(d)) is a way to assess the success of the training process which is considered to be satisfactory when the majority of the cells of the hit rate map depict similar values. Such a scenario indicates that the SOM neurons were uniformly activated. In the case of this study, the neurons seem to have been uniformly activated (as BMUs) by the training data with the exception of the neurons that correspond to the water class (lower right corner). These neurons present a higher hit rate because as already observed by the U-matrix, the water spectra show lower variability than the rest of the classes, thus they can be represented by a lower number of neurons.

C. Component Planes

The component planes depict the coefficients of the feature vectors of the SOM neurons. Each coefficient stands for the spectral value of an S2 band, thus their number is equal to the number of S2 bands (13). For the purpose of the study, the component planes (Figure 2.21) were visualized and observed in synergy with the U-matrix and the majority voting output (Figure 2.20(c)). In Figure 2.21 they are grouped according to visual similarity. Thus, only one component plane is shown for bands 2–3, 4–8A, and 11–12, respectively because based on visual observation they presented similar spectral behavior. Since the component planes are essentially a quantized depiction of the training data with meaningful spatial relations, they are a useful and convenient way to extract the spectral properties of the training data. From their observation it can be seen that the bands that correspond to the blue (B1 (444 nm), B2 (497 nm)) and green (B3 (560 nm)) part of the spectrum, present higher spectral values in the classes:

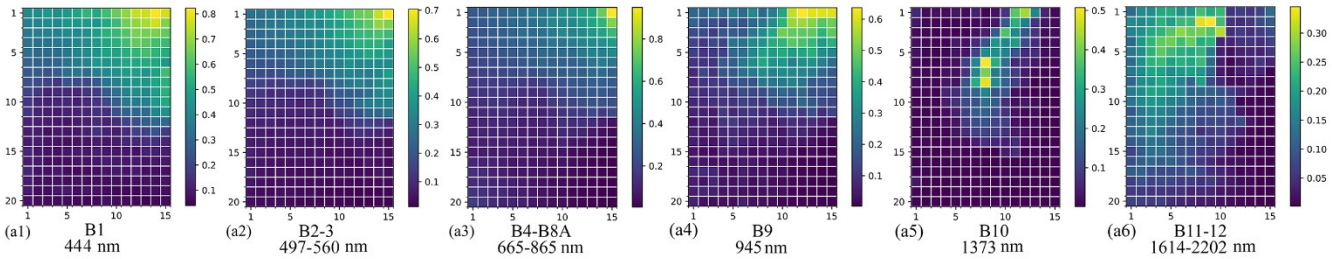


Figure 2.21. Component planes

snow, opaque cloud, and cirrus. The bands that correspond to the red (B4 (665 nm)–B7 (783 nm)) and NIR (B8 (835 nm), B8A (865 nm)) part of the spectrum, present higher spectral values in a small number of neurons representing the snow class. Similar behavior appears in the water absorption band (B9 (945 nm)), with the difference that high values are distributed to more neurons. Concerning the SWIR bands, the cirrus band (B10 (1374 nm)) where incident and reflected light are highly absorbed, presents very low values in most neurons. Slightly higher values are presented in many of the neurons that correspond to the cirrus class. In addition, an area that forms the border between the snow and the opaque cloud class presents the highest values, probably because these pixels appear in high altitudes. As for bands 11 and 12 (1614, 2202 nm) the neurons that correspond to the snow class present low values, and that is the reason why thresholds in these bands are commonly performed for the separation of snow from clouds.

D. Scatterplots

2D scatterplots are commonly visualized as a straightforward means to evaluate the distribution of the feature vectors of the SOM neurons among the training data. The better the shape of the envelope formed by the neurons simulates the shape of the training data, the greater the accuracy of the SOM training process. For the purpose of the study the 2D scatterplots between several S2 bands were observed both for the six classes of the training data in total and for each class separately. Figure 2.22 depicts the 2D scatterplots for the six classes between bands 3(560 nm)–8(835 nm) and 3–11(1614 nm). Figure 2.23 depicts respective 2D scatterplots for each class separately. As can be seen in Figure 2.22(a) the spectral values of the SOM nodes for bands 3 and 8 simulate very well the distribution of the training data. As for bands 3 and 11 (Figure 2.22(b)), the distribution of the nodes is less dispersed.

Similar conclusions are derived by observing Figure 2.23 which provides a clearer image for the interpretation of the distribution of the SOM nodes. As shown in Figure 2.23(a1), it is clear that the opaque cloud class is well represented as far as the spectral properties of the green (B3) and the NIR (B8) bands are concerned.

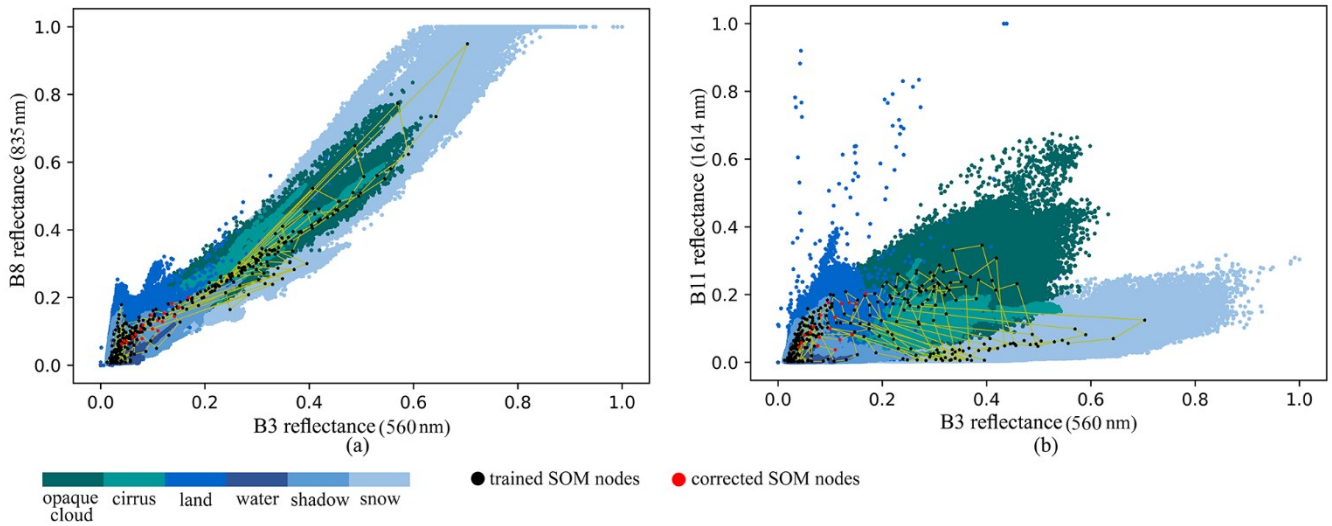


Figure 2.22 Scatterplots of the six classes of the training set

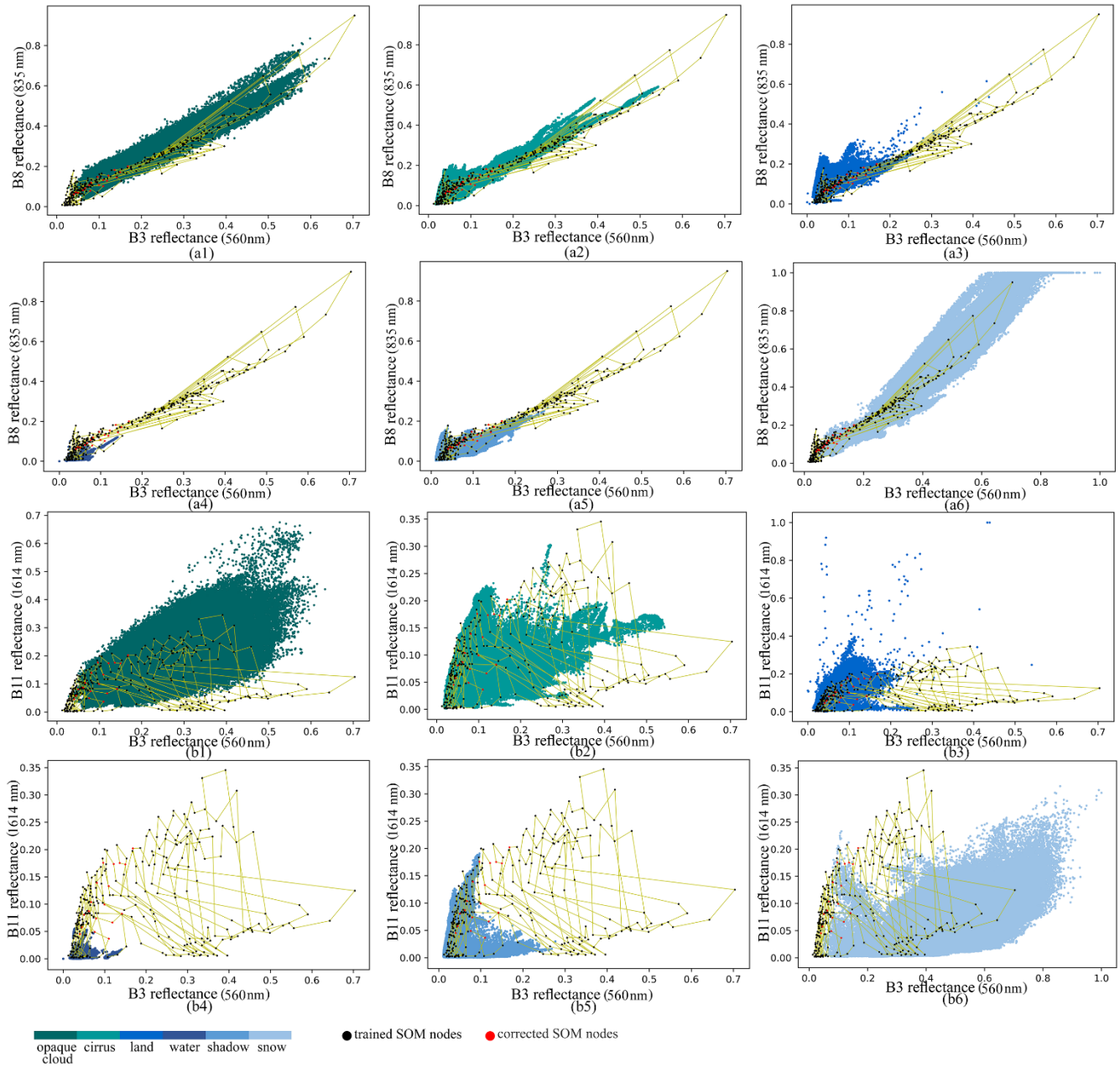


Figure 2.23. Separate scatterplots of the six classes of the training set

Table 2.20 Confusion matrix of trained SOM.

	Other	Cloud	Snow	Producer's Accuracy
Other	4,742,750	74,068	5856	0.983
Cloud	222,145	2,463,208	20,828	0.910
Snow	1048	8847	1,261,248	0.992
User's accuracy	0.955	0.967	0.979	

However, in Figure 2.23(b1), there is a spectral area at the top right that is not fully covered. Similar behavior is also observed for the snow class (Figure 2.23(a6,b6)). Regarding the land class (Figure 2.23(a3,b3)), even though the SOM neurons are distributed among the majority of the training data, their dispersion does not reach a portion on the top left. The cirrus (Figure 2.23(a2,b2)), water (Figure 2.23(a4,b4)) and shadow

Table 2.21. Evaluation metrics of S2 cloud masks (entire test set).

Method	Accuracy	Recall	Precision	Fscore
Sen2Cor	0.920	0.928	0.969	0.943
Fmask	0.922	0.917	0.984	0.945
SOM1	0.928	0.919	0.988	0.949
SOM2	0.928	0.919	0.986	0.949

$$\text{Fscore} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (2.29)$$

(Figure 2.23(a5,b5)) classes appear to be very well delineated by the scattering of the SOM neurons. It is noted that resembling behavior was shown for the other visible, NIR and SWIR S2 bands.

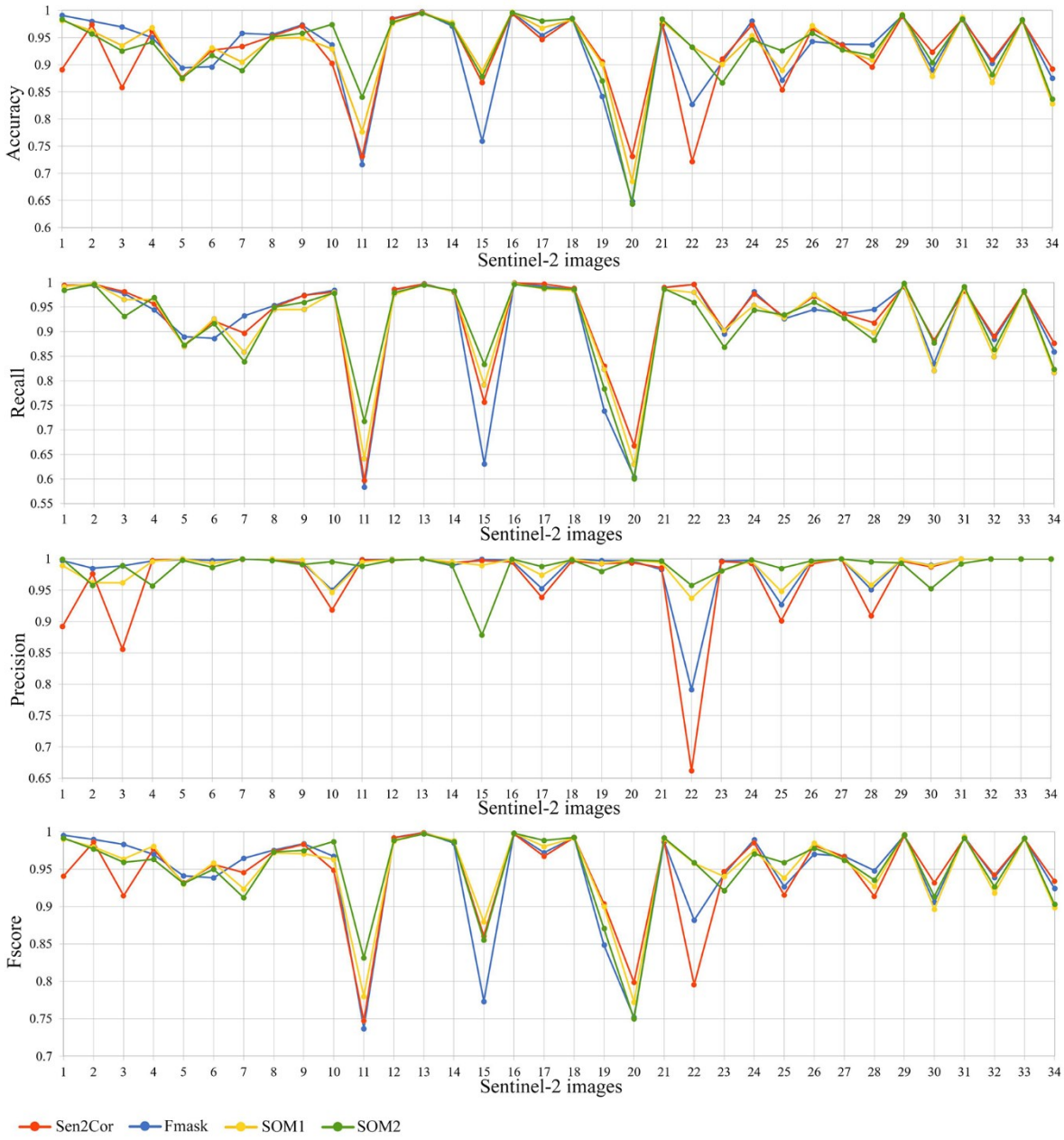


Figure 2.24. Evaluation metrics of the entire test set

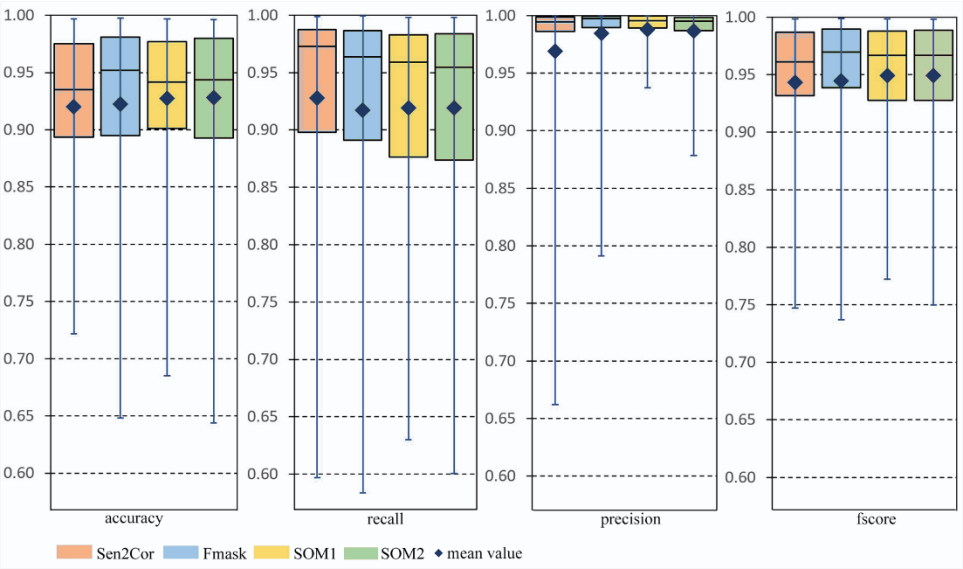


Figure 2.25. Box plots of the evaluation metrics of the entire test set

2.3.3.2 Training Set

The confusion matrix was created (Table 2.20) in order to evaluate the performance of the trained network (SOM1) on the training spectra. It was created by feeding into the non-fine-tuned trained network all the training spectral signatures and predicting their class. The equations for accuracy (Equation (2.15)), recall (producer's accuracy) (Equation (2.16)), and precision (user's accuracy) (Equation (2.17)) are mentioned in section 2.2.3.1.A.

The confusion matrix presented an overall accuracy of ~96%. By observing the values of the table it can be observed that the snow class presents low omission and commission errors (<1%, ~2%). As for the "other" class that includes the classes land, shadow, and water, the commission error is ~4% and the omission error is ~2%. Higher omission error is presented for the cloud class (~9%), while the commission error is close to the formerly mentioned classes.

2.3.3.2 Cloud Masks of the Entire Test Set

Several evaluation metrics were calculated for the quantitative evaluation of the cloud masks produced by the trained SOM before and after applying the fine-tuning process. In more detail, accuracy, recall, precision, and Fscore (Equation (2.29)) which combines recall and precision metrics were computed.

These metrics were also computed for two state-of-the-art algorithms: Sen2Cor and Fmask. Their average values are presented in Table 2.21. By observing the table values it can be deduced that the two SOM versions as well as Sen2Cor and Fmask perform similarly with differences often less than 1%. The average values of the evaluation metrics are: accuracy: ~93%, recall: ~92%, precision: ~98%, and Fscore: ~95%.

The similar behavior is also shown in the plots of Figure 2.24 where the evaluation metrics for each of the images of the test set are presented, as well as in the box plots depicted in Figure 2.25. A box plot is a diagram that illustrates the variance of the data. It consists of two boxes. The lower side of the lower box corresponds to the first quartile and the upper side to the second quartile. The vertical lines crossing the boxes denote the distance of the maximum or minimum value in comparison to the second quartile. The box plots of this study indicate a slightly greater variance of recall values for SOM1 and SOM2 with lower values of the first quartile. In addition, these algorithms present slightly higher precision values and a smaller distance of the minimum value from the second quartile. It is noted that Sen2Cor shows the highest average recall values (lowest omission error) but also the lowest average precision values (highest commission error).

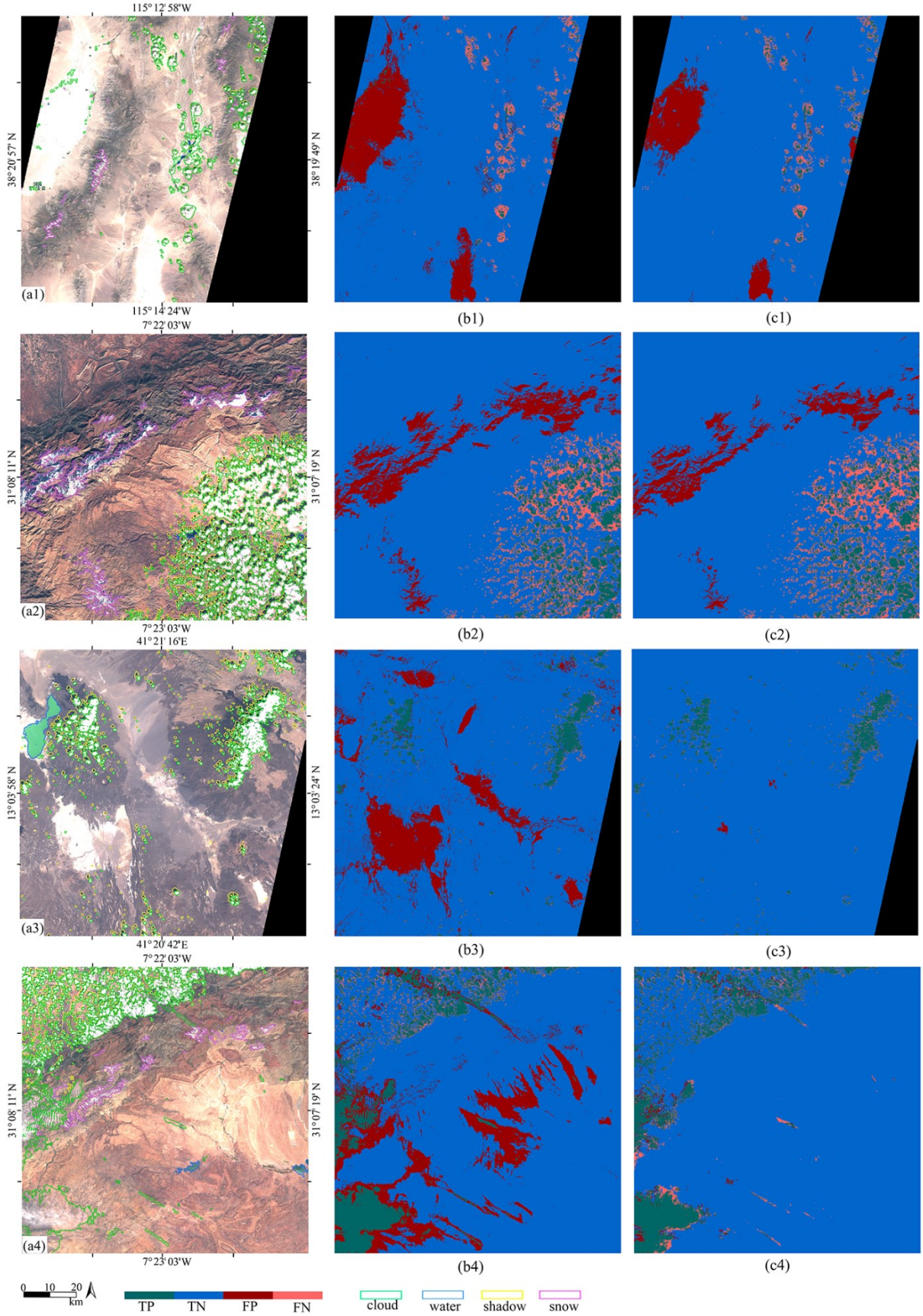


Figure 2.26. Cloud masks of S2 images with bright non-cloud objects. (a1–a4): RGB composites with delineation of categories, (b1–b4): Sen2Cor cloud masks, (c1–c4): Fmask cloud masks

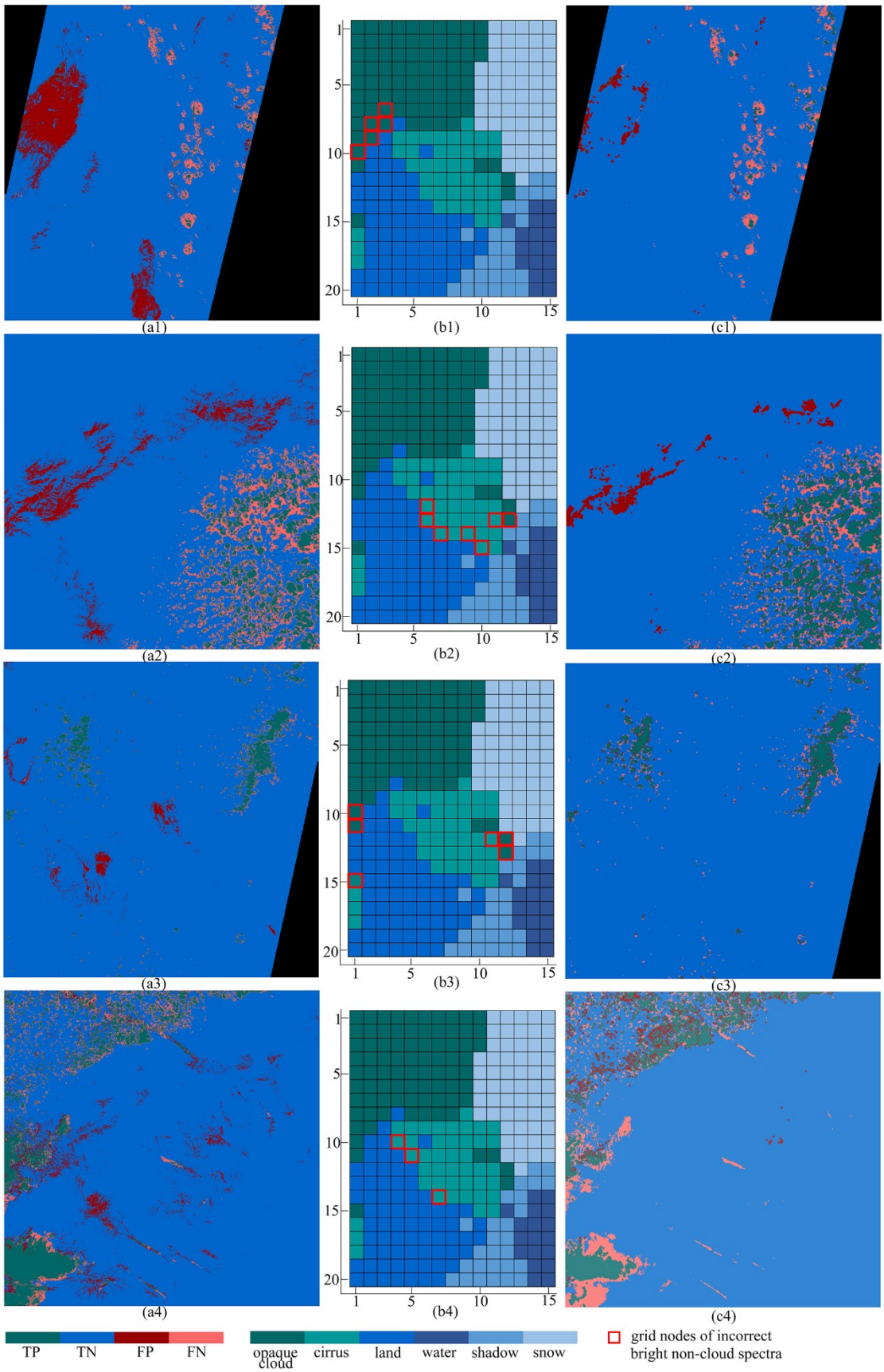


Figure 2.27. (a1–a4): SOM1 cloud masks, (b1–b4): neurons with altered labels, (c1–c4): SOM2 cloud masks

Table 2.22. Evaluation metrics of images with bright non-cloud objects (accuracy, recall)

	Accuracy				Recall			
	Sen2Cor	Fmask	SOM1	SOM2	Sen2Cor	Fmask	SOM1	SOM2
Figure 2.26(a1)	0.903	0.937	0.929	0.974	0.98	0.984	0.980	0.979
Figure 2.26(a2)	0.854	0.871	0.890	0.926	0.93	0.926	0.928	0.934
Figure 2.26(a3)	0.891	0.991	0.982	0.984	0.994	0.993	0.992	0.984
Figure 2.26(a4)	0.858	0.970	0.935	0.926	0.982	0.978	0.965	0.931
Figure 2.28(a1)	0.947	0.954	0.967	0.981	0.997	0.992	0.987	0.989
Figure 2.28(a2)	0.896	0.937	0.909	0.917	0.918	0.945	0.898	0.882
Figure 2.28(a3)	0.976	0.972	0.977	0.975	0.981	0.981	0.981	0.983
mean	0.903	0.947	0.941	0.954	0.969	0.971	0.961	0.954

Table 2.23 Evaluation metrics of images with bright non-cloud objects (precision, Fscore)

	Precision				Fscore			
	Sen2Cor	Fmask	SOM1	SOM2	Sen2Cor	Fmask	SOM1	SOM2
Figure 2.26(a1)	0.919	0.950	0.947	0.995	0.949	0.967	0.963	0.987
Figure 2.26(a2)	0.901	0.927	0.948	0.985	0.916	0.927	0.938	0.959
Figure 2.26(a3)	0.892	0.997	0.989	0.999	0.941	0.995	0.990	0.992
Figure 2.26(a4)	0.856	0.989	0.962	0.990	0.915	0.983	0.963	0.959
Figure 2.28(a1)	0.939	0.953	0.974	0.988	0.967	0.972	0.98	0.988
Figure 2.28(a2)	0.909	0.951	0.958	0.995	0.914	0.948	0.927	0.935
Figure 2.28(a3)	0.994	0.989	0.995	0.991	0.987	0.985	0.988	0.987
mean	0.918	0.965	0.968	0.992	0.941	0.968	0.964	0.972

2.3.3.3 Cloud Masks of Fine-Tuned Cases with Bright Non-Cloud Objects

A. Images Used in the Fine-Tuning Process

Figure 2.26 presents the cloud masks produced by Sen2Cor and Fmask for the four images with bright non-cloud objects which were used for the selection of the sample of incorrectly classified bright non-cloud pixels during the fine-tuning process (section 2.3.2.2.C). Respective results for SOM1 and SOM2 are presented in Figure 2.27. These figures show the RGB natural composite where the ground-truth categories were delineated by Baetens et al. (2019) [44], as well as correctly predicted pixels (for cloud (TP) and clear (TN) categories) along with omission (FN) and commission error (FP). The latter figure also shows the neurons that were altered in terms of their label during the fine-tuning process. These neurons as already mentioned corresponded to the signatures of the bright non-cloud objects that were incorrectly classified by SOM1. Based on the four images depicted in Figure 2.26(a1-a4), the labels of 18 neurons in total were altered from cloud to non-cloud. The evaluation metrics for these images are presented in Table 2.22 and Table 2.23.

For Figure 2.26(a1), the cloud masks produced by Sen2Cor and Fmask incorrectly classified two large bright areas of land (commission error: ~8% and ~5%) with Sen2Cor performing worse. The cloud mask produced by SOM1 was similar to the Fmask output but it can also be seen that two small snow areas were incorrectly detected. The cloud mask produced by SOM2 performed significantly better by correctly classifying the majority of the bright land pixels (commission error: <1%). Likewise, for Figure 2.26(a2), the cloud masks produced by Sen2Cor and Fmask misclassified a number of bright land and snow pixels (commission error: ~10% and ~7%). SOM1 showed a lower commission error (~5%) and SOM2 performed better than the previous methods (commission error: <2%). Regarding the SOM2 result, it should be observed that the misclassification of snow pixels is slightly lower than SOM1. This is due to the fact that a small percentage of snow pixels was selected along with the surrounding soil bright pixels during the fine-tuning process. As already clarified in the Introduction, experimental

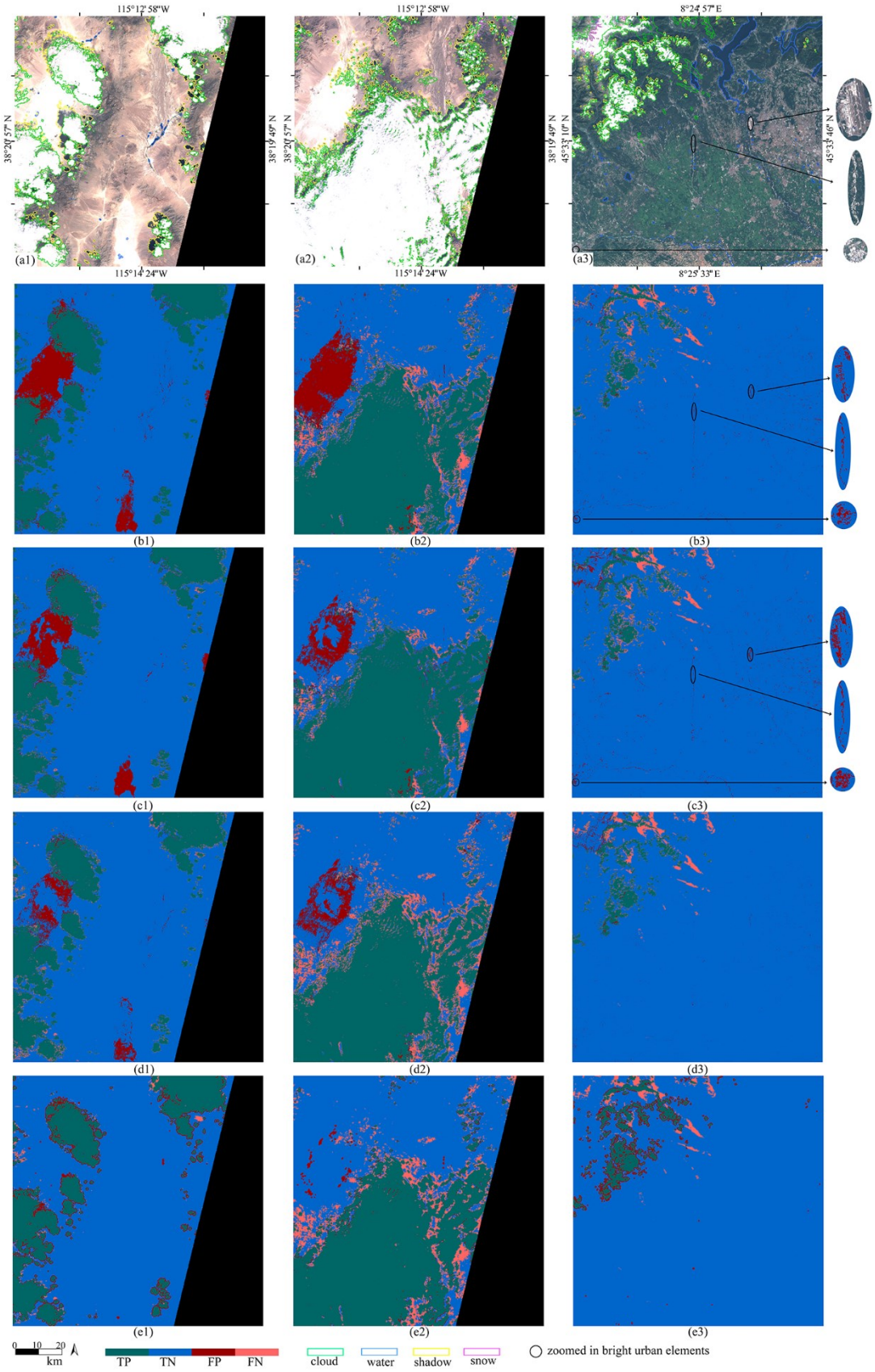


Figure 2.28. Cloud masks of S2 images with bright non-cloud objects. (a1–a3): RGB composites with delineation of categories, (b1–b3): Sen2Cor cloud masks, (c1–c3): Fmask cloud masks, (d1–d3): SOM1 cloud masks, (e1–e3): SOM2 cloud masks

attempts were also made for the alteration of the labels of more neurons that corresponded specifically to snow pixels. However, the results showed that a large omission error for the cloud class occurred and thus the results were considered unsatisfactory. Taking into account the fact that the training set produced <1% omission error for the snow class, it is safe to assume that the test set includes various snow spectra (e.g., wet, dry) of different thickness that do not appear in the training set.

For Figure 2.26(a3), Sen2Cor showed a high commission error (~11%) by incorrectly classifying several bright non-cloud areas. However, Fmask presented satisfactory results as it only misclassified two small bright land areas. Slightly lower performance was depicted by SOM1 which failed to correctly detect a few more bright non-cloud pixels. The best cloud mask was produced by SOM2 which managed to successfully predict the class of the bright non-cloud surfaces. Finally, for Figure 2.26(a4), Fmask and SOM2 presented the most successful results, followed by SOM1 (commission error:~4%) and Sen2Cor which produced the worst cloud mask (commission error (~14%)).

It is noted that as expected by the SOM theory, the neurons that were altered were located at the borders of the cloud class (opaque cloud + cirrus) with either the land or the snow class.

B. Images Not Used in the Fine-Tuning Process

Figure 2.28 shows the cloud masks produced by the four methods for the three of the seven images with bright non-cloud objects which were not used for the selection of the sample of incorrectly classified bright non-cloud pixels during the fine-tuning process. It was observed that the results are similar to those of the images presented in section 2.3.3.4.A.

For Figure 2.28(a1), Sen2Cor presents the least satisfactory results by incorrectly classifying two bright land areas (commission error: ~6%). Fmask performs slightly better (commission error: ~5%) and SOM1 appears to be more successful since it misclassifies a lower percentage of pixels (commission error: ~3%). The SOM2 cloud mask shows the most satisfactory results (commission error: ~1%). Likewise, for Figure 2.28(a2), SOM2 illustrates the best performance (commission error: <1%) followed by SOM1 (commission error: ~4%), Fmask (commission error: ~5%) and Sen2Cor (commission error: ~9%). Finally, for Figure 2.28(a3), Sen2Cor and Fmask fail to correctly detect the bright urban elements like buildings and streets as shown in the zoomed in areas delineated by black circles for easier perception. SOM1 and SOM2 cloud masks do not present such an issue.

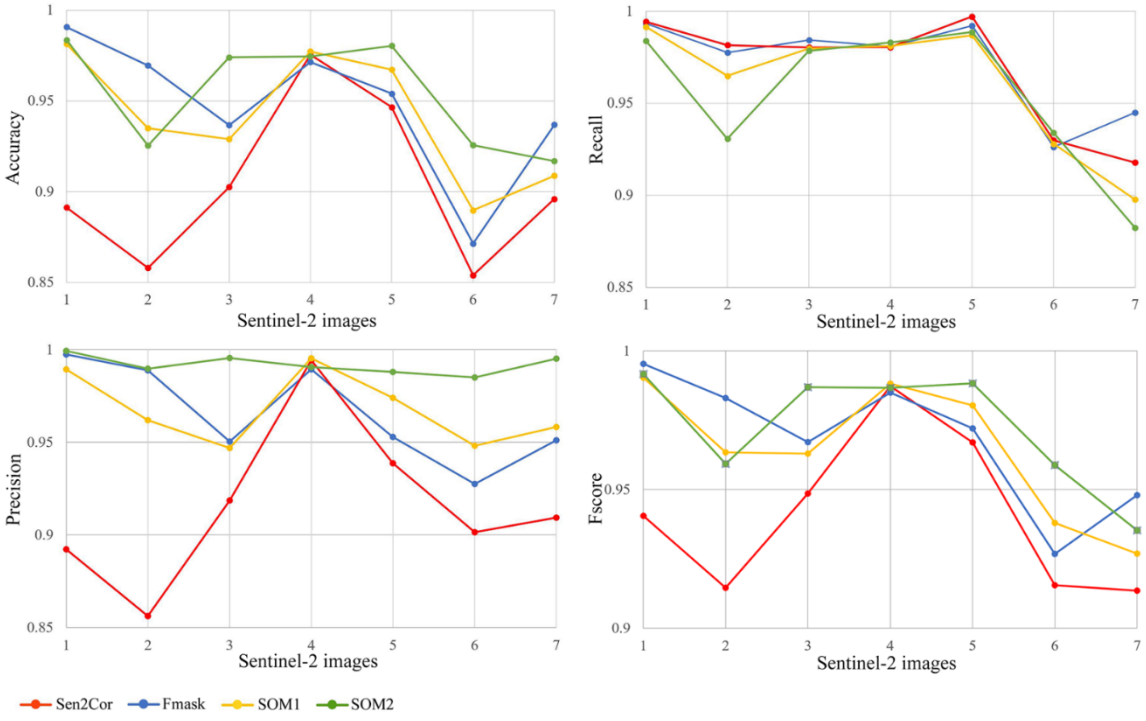


Figure 2.29. Evaluation metrics of the S2 images with bright non-cloud objects

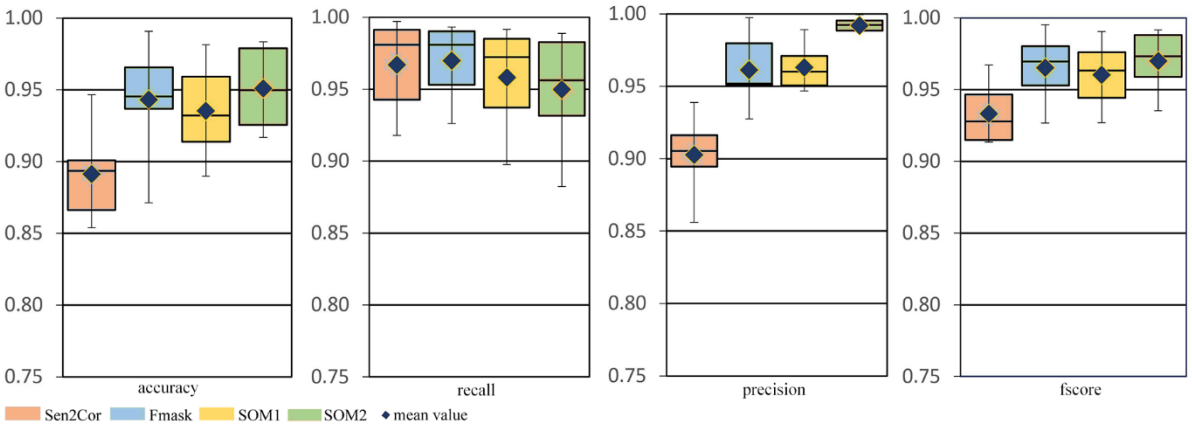


Figure 2.30. Box plots of the S2 images with bright non-cloud objects

Plots created by the values of the four evaluation metrics which are shown in Figure 2.29 are useful for the evaluation. By observing the line formed by the precision values, it is easily deduced that SOM2 produced by fine-tuning the SOM1 network, outperformed the other algorithms. This conclusion is also reached by observing the box plots of Figure 2.30. It is also noted that the SOM2 results produce only a slight decrease in recall values compared to SOM1.

Concerning the comparison of the time needed to produce a cloud mask for an S2 image, as already mentioned in section 2.3.2.2.B the SOM proposed in this study needs ~six minutes, while the versions of Fmask and Sen2cor (year:2020) which are much faster than the previous ones, when run from command line need ~four and ~two min, respectively. Nevertheless, even though Fmask and Sen2Cor run faster than the proposed SOM, they fail to distinguish the bright non-cloud objects of the test set, thus their time-efficiency becomes irrelevant in this case.

2.3.4 Discussion

This section highlights the main points that distinguish the proposed method from: (a) other types of ANNs (MLPs, CNNs), (b) the most commonly applied state-of-the-art algorithms (Sen2Cor, Fmask), and (c) the most widely used unsupervised classification method (k-means). The discussion also includes comments on the potential risk of the proposed fine-tuning approach.

To begin with, time-efficiency is one of the main benefits of SOMs compared to other types of ANNs such as MLPs and CNNs which are widely and most frequently applied in current research and industrial applications. MLPs and CNNs require multiple hours for training while SOMs usually need only a few minutes. As a matter of fact, the SOM proposed in this study was trained in two minutes and required 18 min to acquire its labeling in order to produce the cloud masks. Concerning fine-tuning for MLPs and CNNs, the main difference between the fine-tuning stage and the initial training stage is that during fine-tuning a smaller training set and fewer hidden layers are used, but the process is still slower compared to the proposed fine-tuning method (SOM2). In addition, the proposed fine-tuning method required only a small labeled dataset in order to detect the BMUs of the bright non-cloud spectra and then alter their labels.

Another advantage of SOMs is that their behavior is much more interpretative compared to MLPs/CNNs where the performance is mainly evaluated through the accuracy of the predictions on the test set. In contrast, in SOMs, the network can also be evaluated by useful visualization techniques that analyze the similarity/dissimilarity of the neighboring nodes, the uniformity of the activation of the neurons, the fast extraction of spectral properties from quantized data, and the distribution of the SOM nodes among the data.

As far as comparison with Fmask and Sen2Cor is concerned, the proposed fine-tuning method outperformed them by far in the separation of bright non-cloud objects from clouds. However, the newest versions of Fmask and Sen2cor produce an S2 cloud mask faster i.e., ~four minutes are required for Fmask and ~two minutes for Sen2Cor against ~six minutes for the proposed SOM.

Regarding the similarities of SOMs with the k-means, it can be indeed stated that the two methods are very similar with the main difference being that in SOMs the centers (neurons) interact with each other and create

neighborhoods that carry topologic information. The centers in k-means do not interact with each other. In practice, concerning our study, the main difference between the two methods lies in the fine-tuning process which would not be practically possible to be applied in the k-means algorithm. The reason is that it would require the number of clusters to be the same as the number of the SOM nodes, and thus the training process would be very slow. The common practice of fine-tuning the k-means algorithm is to run the method with an increasing number of classes until satisfying results are produced. This approach is still time-consuming and is even more cumbersome when the training set is different from the test set. In general, training a k-means is much slower than training a SOM, because in k-means every time the centers are updated, the distance between the new centers and all the data points needs to be calculated. In contrast, for the SOM training, data points are fed into the network one by one and every time a data point is fed into the network only the distance of this data point with the nodes of the SOM grid needs to be calculated.

A final point to be discussed is the potential risk of "altering the incorrect labeling" in our proposed fine-tuning approach. The effect of altering the incorrect labeling in the case study analyzed in our paper is that the altered nodes (as explained in section 2.3.2.2.C) are not only the BMUs of the bright non-cloud spectra but also of a few cloud pixels. In practice, that means that when we produce a cloud mask by using the fine-tuned SOM version (SOM2) it is probable that there are a few cloud pixels in the image that correspond to the altered neurons, and thus they will be misclassified to the non-cloud class. In this paper, we alleviated this issue by running a median and a dilation filter in order to retrieve the cloud pixels that SOM2 had misclassified. Thus, we have proven, that concerning the 34 images of the test set, the effect of altering the incorrect labeling can be overcome. Since these images were captured around the globe in different seasons and times of the day and represent a large variety of land cover, we believe that SOM2 would have a similar performance in images that were not included in our test dataset. Our opinion is reinforced by the fact that the proposed fine-tuned method alters the neurons of the borders of the opaque cloud and cirrus class with the land class and not the labels of the neurons that are situated in the center of the classes in the SOM grid.

2.3.5 Conclusions

This study evaluated a SOM for cloud masking S2 images and proposed a fine-tuning methodology based on the output of the non-fine-tuned network. The fine-tuning process managed to correct the misclassified predictions of bright non-cloud spectra without applying further training. The proposed fine-tuning method is the most important contribution of the study since it follows a general procedure, thus its applicability is broad and not confined only in the field of cloud-masking. It was performed by directly locating the neurons that correspond to the incorrectly predicted bright non-cloud objects and altering their labels. This process was chosen over the common practice of further training the network by feeding the labeled data (supervised training) since it was considered faster, simpler, and more efficient. Further training would probably also require more data than those available. A median and a dilation filter were performed as the final step of fine-tuning to compensate both for remaining omission and commission errors caused by the fact that the altered neurons represented also a percentage of cloud pixels.

The SOM was trained on approximately nine million spectral signatures extracted from the largest publicly available database (at the time the study was conducted (year:2020)) of S2 signatures for cloud masking applications. After the completion of the training, the non-fine-tuned network was at first evaluated (a) by employing several visualization techniques that illustrated essential spectral properties of the nodes based on topologic information and led to the interpretation of its behavior and (b) by calculating the confusion matrix on the training set where it produced an overall accuracy ~96%. Then, evaluation of the non-fine-tuned (SOM1) and the non-fine-tuned (SOM2) versions was performed on a truly independent test set of 34 S2 ground-truth cloud masks provided by the only publicly available source. By evaluating this entire test set through several evaluation metrics and plots and comparing them with two state-of-the-art algorithms (Sen2Cor, Fmask), it was deduced that both the two SOM versions and the two state-of-the-art algorithms produced similar results (accuracy: ~93%, recall: ~92%, precision: ~98% and Fscore: ~95%). However, when performing the quantitative and qualitative evaluation process for the cases with bright non-cloud objects, it was shown that the fine-tuned version performed more successfully with an average commission error of less than 1%. The respective values for the SOM before fine-tuning were ~3%, for Fmask ~4%, and for Sen2Cor ~8%. The fine-tuning method proposed in this study was also applied in experiments that specifically targeted incorrectly classified snow pixels. However, the results were considered unsatisfactory because a large omission error of clouds was produced. Thus, those experiments were

not presented in this study.

As a general conclusion, the study showed that the proposed method for fine-tuning SOMs is very effective for separating bright non-cloud objects from clouds, while the commonly used state-of-the-art algorithms failed in this task. In addition, the method is simple in its implementation and time-efficient, since it only involves the detection of the BMUs of interest and requires very few data points as input. Thus, in future work, the potential of the method in different scenarios could be investigated, especially for big data analysis where the processing time is crucial. Testing the method in datasets with greater availability of ground-truth data and comparison with supervised SOM approaches could also be considered in the future.

2.4 CNNs for detecting challenging cases in cloud masking using Sentinel-2 (S2) imagery[3]

In section 2.4, the third cloud masking approach is presented. It employs CNNs on S2 data and attempts to tackle all challenging cases in cloud masking including snow and thin clouds. In section 2.4.1 the motivations and the objectives of the study are provided. In section 2.4.2 a data and methodology description is provided. In section 2.4.3 the results and their evaluation are discussed. Finally, in section 2.4.4 the conclusions and contributions are summarized and future work is described.

2.4.1 Introduction

Convolutional deep learning approaches generally perform better than other approaches in the detection of challenging cases in cloud masking applications. It should be though highlighted that a crucial factor for achieving satisfactory performance is the high accuracy of the ground-truth cloud masks. The main technique for creating such masks is visual observation which is time-consuming. The recent public availability of the database provided by Baetens et al. (2019) [44] (first of its kind) motivated this study because it provided the opportunity to perform robust evaluation for S2 cloud masking methods. This dataset contains cloud-masks with 98% accuracy.

This study proposes a patch-to-pixel CNN architecture for mitigating thin cloud omission and bright non-cloud object commission which pose the main issues in cloud masking applications. For the purpose of the study, different hyperparameters are examined and the feature maps are observed. The results are compared qualitatively and quantitatively with ground-truth cloud masks and the outputs produced by state-of-the-art algorithms.

2.4.2 Proposed method

2.4.2.1 Data description

The study was performed by using in total 37 not atmospherically corrected images collected by the S2 satellite. Their corresponding cloud masks were provided by the database created by Baetens et al. [44]. Further information on the creation of this dataset has been provided in section 2.3.2.1.B. This database was the only publicly available source of S2 ground-truth cloud masks at the time the study was conducted (2020). The images were collected in: Europe (19), North America (three), South America (four), Africa (10), and Australia (one). The dates of the collection cover all seasons of the year: eight winter images (December, January, February), eight spring images (March, April, May), 12 summer images (June, July, August), and nine fall images (September, October, November). The collection time varies between seven a.m. and six p.m. UTC.

Information on the available spectral bands of S2 has been provided in section 2.2.2.1.B. Before analysis, these images were processed. The bands with spatial resolution of 10 and 20 m were resampled to 60 m, with x-size (columns):1,830 pixels and y-size (rows):1,830 pixels. Then, zero padding (size=eight pixels) was added around the images so that the size of the cloud masks produced by the CNN is the same as the S2 images, since the input patch x-size and y-size was 16×16. The training set consisted of 16 images and the test set of 21. A good representation of land cover and cloud variability was the main factor that was taken into account when selecting the images of the training set. Focus was also put on the inclusion of adequate samples of thin clouds and bright

[3] **Kristollari, V.** and Karathanassi, V., 2020, August. Convolutional neural networks for detecting challenging cases in cloud masking using Sentinel-2 imagery. In SPIE Eighth international conference on remote sensing and geoinformation of the environment (RSCy2020) (Vol. 11524, pp. 188-201) (peer-reviewed) doi:10.1117/12.2571111

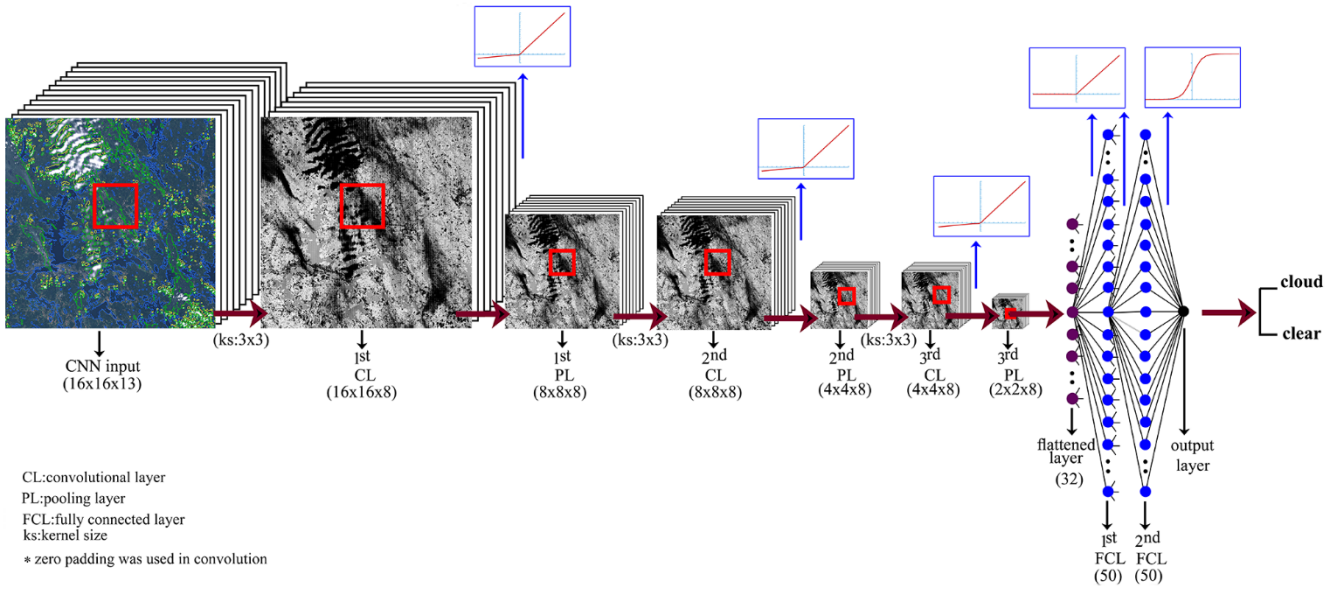


Figure 2.31. Architecture of the first version of the proposed CNN (Use of Leaky ReLU)

non-cloud objects.

2.4.2.2 CNN Architecture

The patch-to-pixel CNN architecture proposed in this study is composed of three convolutional layers and three pooling layers. Each convolutional layer is followed by a pooling layer that retains the maximum value of a window with size 2×2. The patch input size of the CNN is 16×16 and the output predicts the central pixel of the patch. A kernel of size 3×3 is applied for all three convolutional layers. It was decided to use zero padding before applying convolution, thus the size of the output of the operation is the same as the size of the input. The convolution operation is depicted in Equation (2.30). The CNN architecture is followed by a flattening layer, two fully connected layers each of which is composed of 50 neurons and an output layer. This architecture was investigated by implementing three different versions. In the first version (Figure 2.31) the Leaky ReLU activation function was applied after all three convolutional layers. The difference between this function and ReLU (Equation (2.11)) is the use of a small slope for negative values instead of zero. In the second version the ReLU activation function was applied and in the third BN [185] which normalizes input layers was combined with Leaky ReLU (BN was applied before Leaky ReLU). For all three versions, the ReLU function was used in the two fully connected layers and the Sigmoid function (Equation (2.12)) in the output layer. In addition, the dropout method with a 0.3 value was applied in the two fully connected layers.

$$G[i,j] = h * F = \sum_{u=-k}^{k} \sum_{v=-k}^{k} h[u,v]F[i-u,j-v] \quad (2.30)$$

where h is the image, F is the filter. u, v are row and column coordinates of the image and i, j are row and column coordinates of the filter.

2.4.2.3 Training and Inference

The training was performed on an NVIDIA Graphic Processing Unit (GPU) (NVIDIA 1070 Ti) using the Keras library [167] with Tensorflow [170] as the backend. Information on Keras and Tensorflow has been provided in section 2.2.2.3.A. Each of the three CNN models was trained for 30 epochs with 10,000 train steps. Training time was similar for the model that used Leaky ReLU and the model that used ReLU and it lasted approximately nine hours (the training time for the model that used ReLU was slightly faster). The training time for the model that used BN was 12 hours. Inference time was approximately two minutes for all models. A generator function was designed for the training with the purpose of feeding the CNN with batches of training data. Every time the generator was called, it selected randomly one of the 16 images of the training set and then it selected all the pixels

Table 2.24. Average values of accuracy and loss (30 epochs) for the training and test sets

CNN model	Accuracy		Loss	
	Training set	Test set	Training set	Test set
Leaky ReLU	0.9551	0.9612	0.1166	0.1442
ReLU	0.9539	0.9577	0.1192	0.0944
BN + Leaky ReLU	0.9614	0.8961	0.0993	0.3279

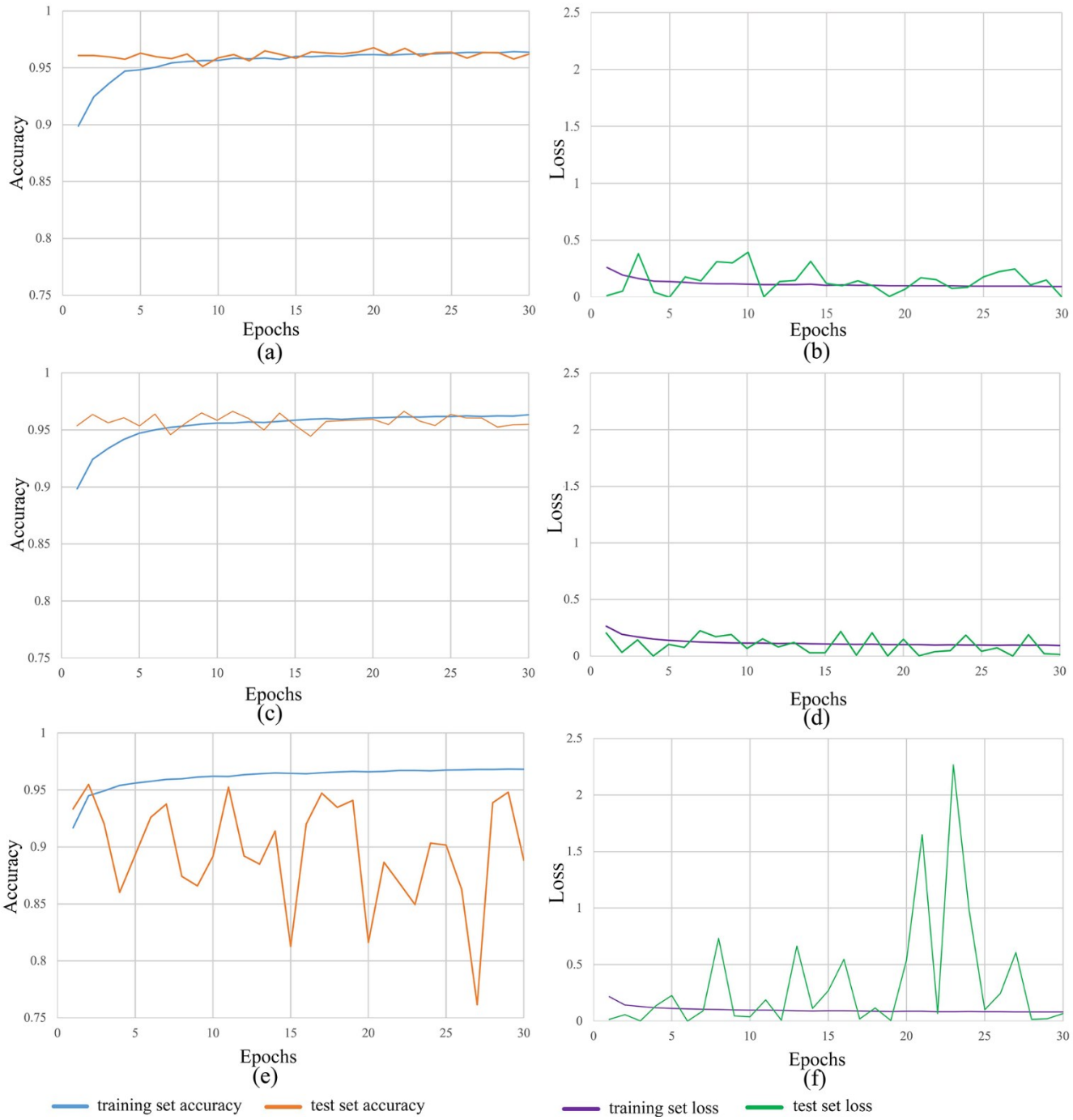


Figure 2.32. (a,b):Accuracy/Loss of model trained with Leaky ReLU, (c,d):Accuracy/Loss of model trained with ReLU, (e,f):Accuracy/Loss of model trained with BN and Leaky ReLU

of a random line as central pixels of a patch of size (16x16x13) where 13 is the number of the S2 bands. A similar generator was designed for the 21 images of the test set in order to compute accuracy and loss values for every epoch. During training, the weights were updated by applying Adam [164] with Equation (2.31) as the loss function. Information on Adam has been provided in section 2.2.2.2.B.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i log\big(p(y_i)\big) + (1-y_i) log(1-p(y_i)) \quad (2.31)$$

where y is the label and $p(y)$ is the probability of the central pixel of the input patch being classified as cloud.

2.4.3 Results

2.4.3.1 Training

The values of accuracy and loss function for the training and the test sets during 30 epochs are shown in Figure 2.32 for the three CNN models. In addition, the average accuracy and loss values are shown in Table 2.24. It can be observed that the CNN models trained by use of the Leaky ReLU and ReLU activation functions demonstrated high accuracy (~96%) and low loss values (<0.14) for both the training and the test. In contrast, the model that combined BN with Leaky ReLU performed by far less favorably since it showed high instability. In more detail, as can be seen in Figure 2.32, accuracy and loss values of the training set showed large differences since the former ranged between ~ 0.75 and ~ 0.95 and the latter between ~ 0 and ~ 2.5. Also, the lower performance of this model can be seen by the large difference in the average values of accuracy and loss function of the training set compared to the test set (~ 90%, ~ 96% and ~ 0.33, ~ 0.12). By observing the plots of Figure 2.32, it was decided to produce cloud masks only by use of the Leaky ReLU model of the last epoch since the accuracy of the test set for this epoch was slightly better than ReLU (Table 2.25).

2.4.3.2. S2 Cloud Masks

Evaluation metrics were computed for the cloud masks produced by the model trained with the Leaky ReLU and the respective cloud masks produced by Sen2Cor and Fmask. The metrics were calculated by considering as ground-truth masks those of the dataset produced by Baetens et al. (2019) [44]. The average values are presented in Table 2.26 and the values for each of the 37 images (16 training images, 21 test images) are presented in Figure 2.33. The metrics that were computed were accuracy (Equation (2.15)), recall (Equation (2.16)), precision (Equation (2.17)) and Fscore (Equation (2.29)).

Table 2.25. Last epoch values of accuracy and loss for the training and test sets

CNN model	Accuracy		Loss	
	Training set	Test set	Training set	Test set
Leaky ReLU	0.9638	0.9621	0.0939	0.0032
ReLU	0.9633	0.9549	0.0941	0.0148
BN + Leaky ReLU	0.9682	0.8884	0.0820	0.0667

Table 2.26. Evaluation metrics of S2 cloud masks

Method	Accuracy	Recall	Precision	Fscore
Sen2Cor	0.9170	0.9215	0.9713	0.9412
Fmask	0.9193	0.9115	0.9856	0.9424
CNN (training set)	0.9742	0.9762	0.9847	0.9804
CNN (test set)	0.9751	0.9800	0.9815	0.9805

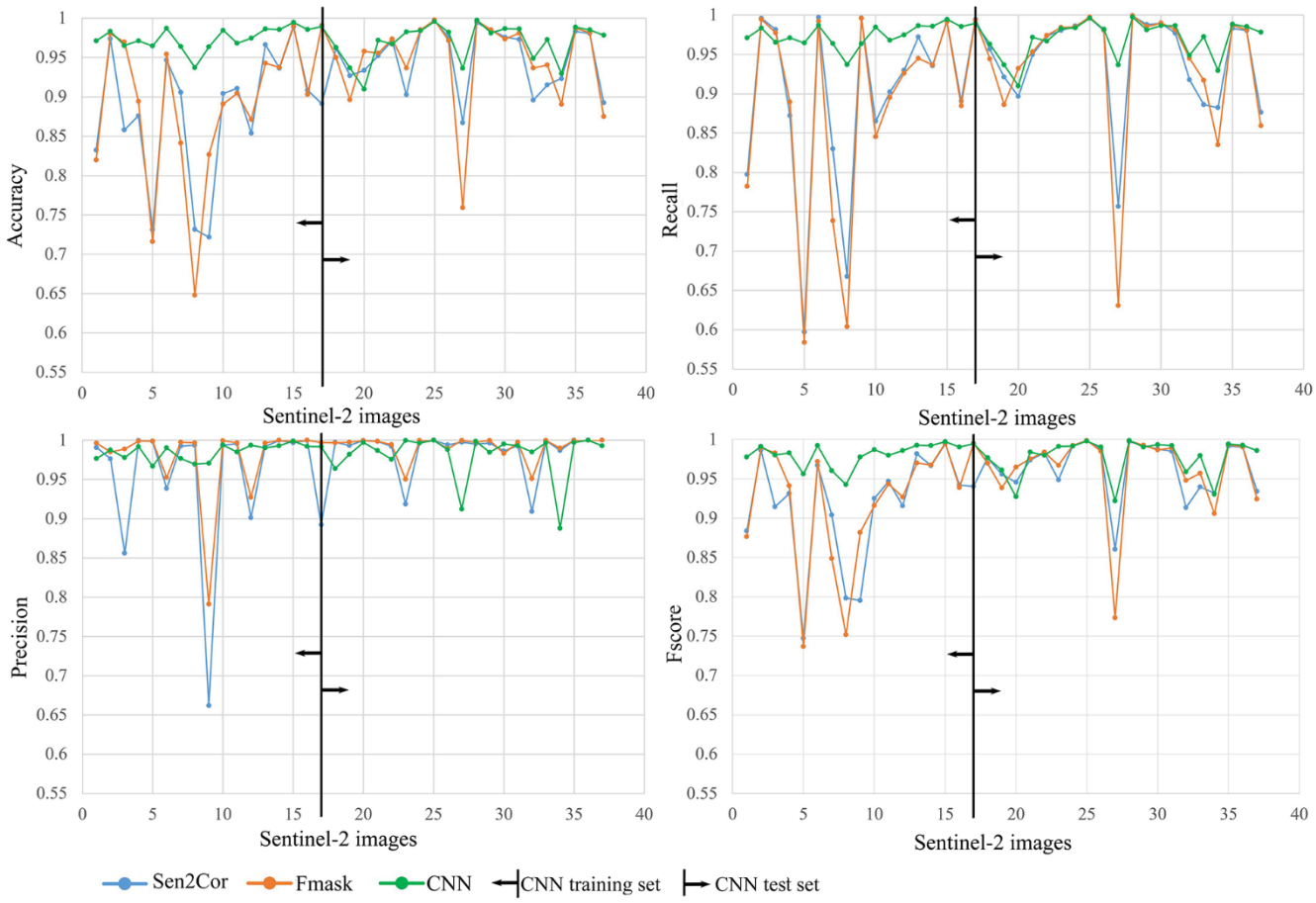


Figure 2.33. Evaluation metrics of the S2 images

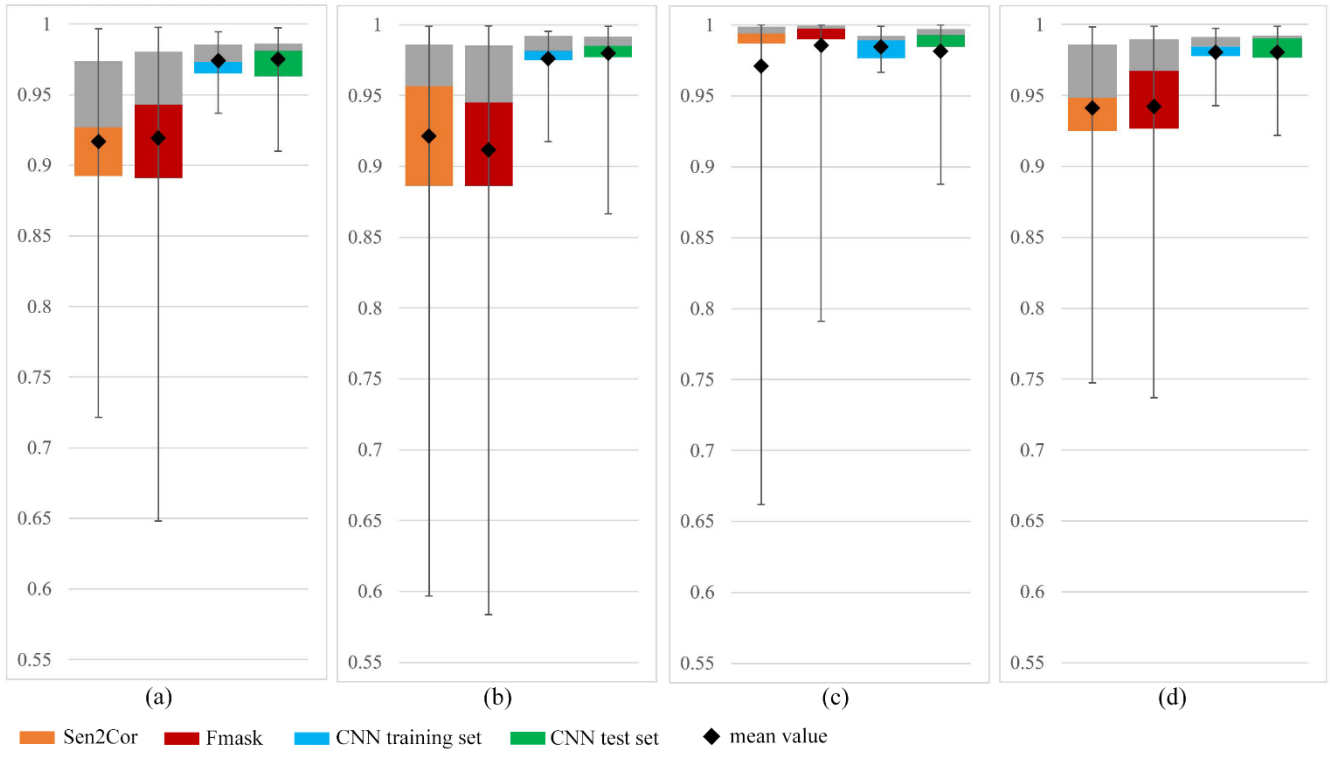


Figure 2.34. Box plots of the evaluation metrics of the S2 images. (a):Accuracy, (b):Recall, (c):Precision, (d):Fscore

For the CNN model, the evaluation metrics were calculated separately for the training and test sets. It was observed that the CNN showed exceptional performance both in the training set and the test set with all evaluation metrics having values ~98%. Concerning the state-of-the-art algorithms, the accuracy and recall values of Sen2Cor and Fmask were ~ 92%, the precision values were ~ 98% and the Fscore values were ~ 94%. Thus, these two state-of-the-art algorithms performed similarly and by far less favorably than the CNN model. The same conclusion can be reached by observing Figure 2.33 and the box plots of Figure 2.34. Information on box plots has been provided in section 2.3.3.2. In the box plots of this study it can be seen that for the CNN the values of all evaluation metrics are much closer to the mean value compared to Sen2Cor and Fmask.

2.4.3.3 Challenging cases

Figure 2.35 and Figure 2.36 present the cloud masks produced for indicative challenging cases by Sen2Cor, Fmask, and the CNN for the training set and the test set respectively. These figures show the RGB natural composite with delineation of the ground-truth categories by Baetens et al. (2019) [44] as well as correctly predicted pixels (for categories of cloud (TP) and clear (TN)) along with omission (FN) and commission error (FP). The evaluation metrics for these particular cases are stated in Table 2.27 and Table 2.28 for the training set and the test set respectively. Concerning the challenging cases of the training set, Figure 2.35(a1) and Figure 2.35(a2) depict cases with optically thin clouds where a high percentage is characterized by very high transparency. It is obvious that the omission error of the CNN is very small for Figure 2.35(a1)(<3%) in contrast to Sen2Cor (~13%) and Fmask (~11%). Similarly, the omission error for the CNN cloud mask of Figure 2.35(a2) is much smaller (~ 8%) than the respective cloud masks of Sen2Cor (~33%) and Fmask (~40%). Figure 2.35(a3) and Figure 2.35(a4) depict cases of non-cloud bright objects. From the produced cloud masks it can be observed that the snow area of Figure 2.35(a3) is correctly classified by the CNN while the other two methods incorrectly detect this snow area as cloud. It can also be seen that the CNN shows a much smaller omission error than the other algorithms. As for Figure 2.35(a4), it can be observed that the bright non-cloud area is successfully classified by the CNN, while Sen2Cor and Fmask fail to correctly categorize it.

Similar conclusions can be reached for the challenging cases of the test set Figure 2.36. Figure 2.36(a1) presents a case of semi-transparent clouds where the CNN cloud mask shows very low omission error (~2%) in contrast to Sen2Cor and Fmask which produce larger omission errors (~12%, ~14%). Figure 2.36(a2) presents a region with snow mountainous areas and an extended bright urban area. For this image, Sen2Cor and Fmask produce cloud masks that incorrectly classify a large part of the snow area as cloud and also incorrectly classify some bright urban elements (shown in a zoom-out circle). The respective CNN cloud mask performs more successfully both in the snow and in the urban area. In Figure 2.36(a3) it can be observed that the CNN can detect more cloud areas that have similar spectral signatures with the background in contrast to the other two methods. Finally, regarding the bright non-cloud objects of Figure 2.36(a4), it can be seen that the CNN and Fmask perform similarly while Sen2Cor produces a high commission error.

2.4.3.4 Feature Maps

Besides training the CNN, this study made an initial effort to investigate the feature maps produced by the convolutional layers, since it would be useful to extract kernels that could be used for the production of features for cloud masking. These kernels could potentially form a database that would enhance the performance of feature-based cloud masking methods. Figure 2.37 depicts an indicative example of an image of the training set which represents a very difficult case for cloud masking since it contains clouds of very high transparency. From visual observation, it can be assumed that the feature map depicted in Figure 2.37(a6) manages to detect more successfully this type of cloud. Figure 2.37 also shows the 13 kernels that were used in the convolution operation that produced the above-mentioned feature map. As already stated, a database composed of kernels of such kind could give the opportunity to easily recreate the feature maps without the need to have any prior information about the CNN. The images that these kernels would be applied should of course depict similar spectral range and potentially similar land cover to increase effectiveness.

2.4.4 Conclusions

This study proposed a CNN model that successfully detected semi-transparent clouds and separated bright clouds

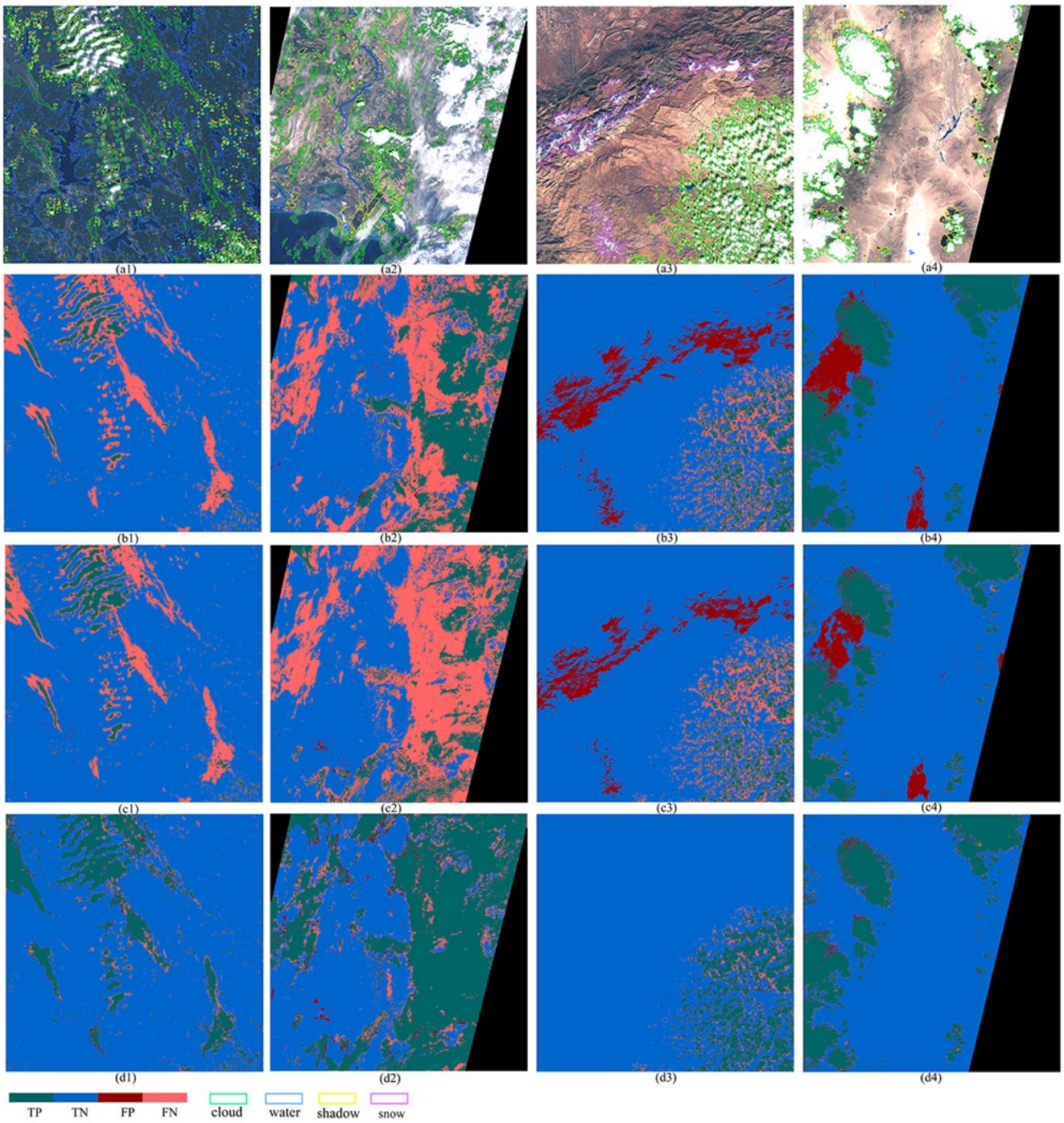


Figure 2.35. Cloud masks of the challenging cases of the training set. (a1-a4): RGB composite with delineation of categories, (b1-b4): Sen2Cor cloud masks, (c1-c4): Fmask cloud masks, (d1-d4): CNN cloud masks

Table 2.27. Evaluation metrics of the challenging cases of the training set

Fig.	Accuracy			Recall			Precision			Fscore		
	S2cor	Fmask	CNN	S2Cor	Fmask	CNN	S2Cor	Fmask	CNN	S2Cor	Fmask	CNN
a1	0.8760	0.8943	0.9711	0.8720	0.8894	0.9745	0.9998	0.9990	0.9917	0.9315	0.9410	0.9830
a2	0.7317	0.6482	0.9370	0.6675	0.6038	0.9173	0.9938	0.9966	0.9697	0.7986	0.7520	0.9428
a3	0.8539	0.8714	0.9747	0.9300	0.9263	0.9779	0.9014	0.9274	0.9936	0.9155	0.9268	0.9857
a4	0.9466	0.9541	0.9870	0.9971	0.9921	0.9939	0.9387	0.9527	0.9905	0.9670	0.9720	0.9922

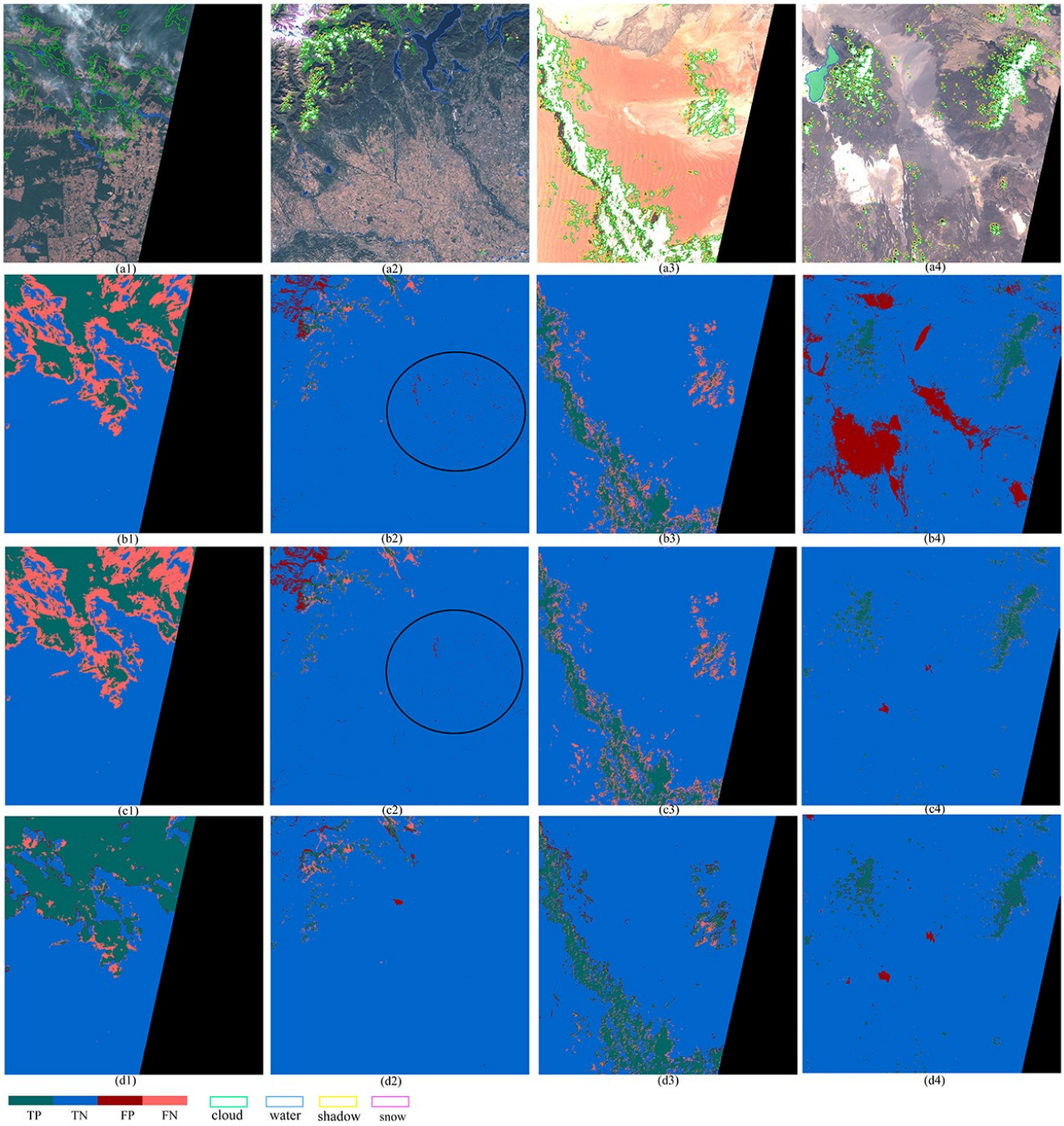


Figure 2.36. Cloud masks of the challenging cases of the test set.(a1-a4): RGB composite with delineation of categories, (b1-b4): Sen2Cor cloud masks, (c1-c4): Fmask cloud masks, (d1-d4):CNN cloud masks

Table 2.28. Evaluation metrics of the challenging cases of the test set

Fig.	Accuracy			Recall			Precision			Fscore		
	S2Cor	Fmask	CNN	S2Cor	Fmask	CNN	S2Cor	Fmask	CNN	S2Cor	Fmask	CNN
a1	0.8926	0.8752	0.9782	0.8764	0.8593	0.9788	1	1	0.9929	0.9341	0.9243	0.9858
a2	0.9759	0.9736	0.9866	0.9894	0.9898	0.9912	0.9860	0.9831	0.9951	0.9877	0.9865	0.9932
a3	0.9524	0.9555	0.9719	0.9499	0.9533	0.9813	0.9988	0.9986	0.9869	0.9737	0.9754	0.9841
a4	0.8914	0.9909	0.9893	0.9944	0.9933	0.9971	0.8922	0.9973	0.9918	0.9405	0.9953	0.9944

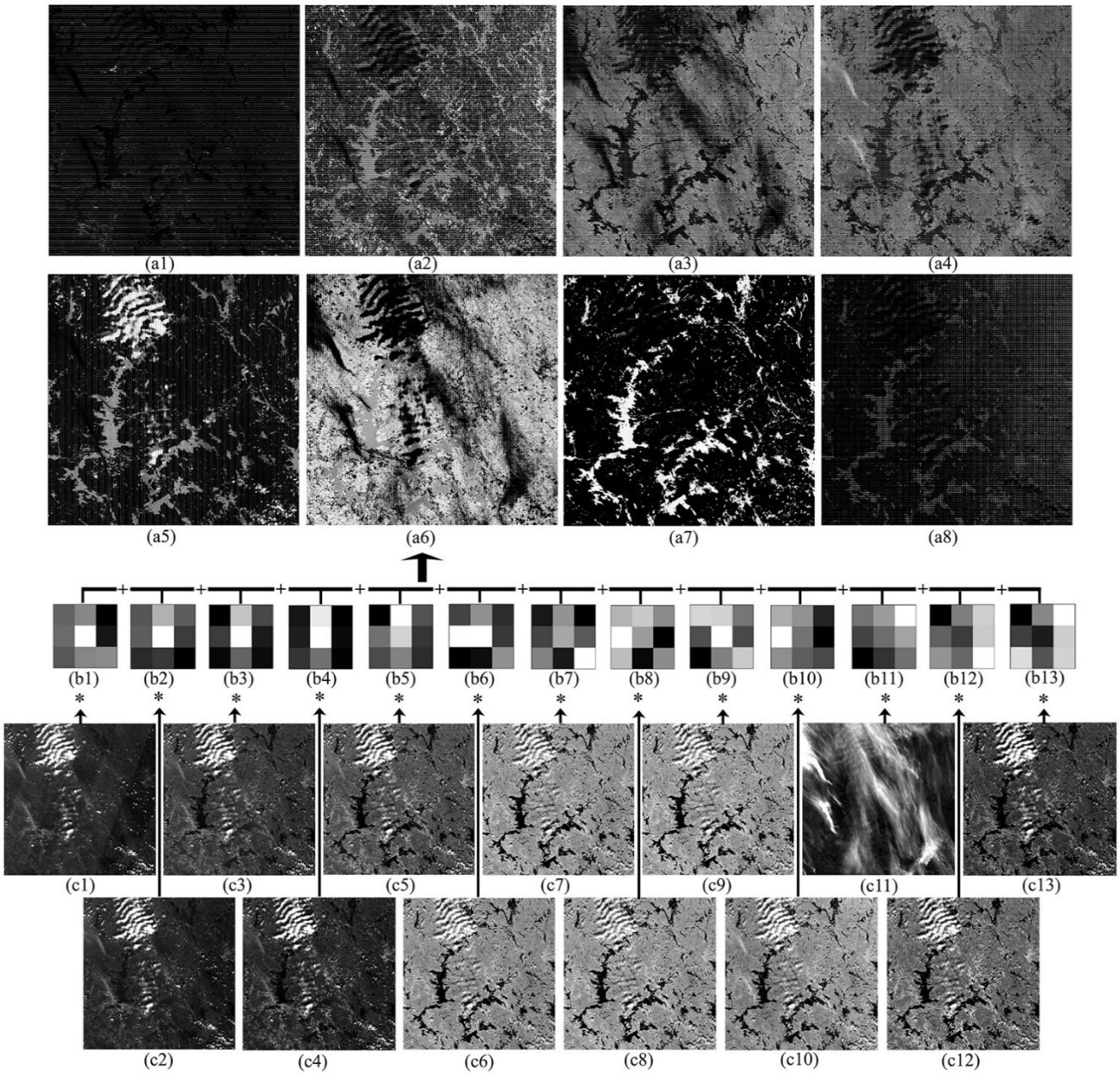


Figure 2.37. (a1-a8): Feature maps of the first convolutional layer for an indicative example, (b1-b13): Kernels used in the convolution operation that produced a6, (c1-c13): S2 bands

from bright non-cloud objects. The proposed method was applied on the first publicly available dataset of S2 ground-truth cloud masks which provides the opportunity for a robust and objective evaluation. Different versions of the proposed CNN architecture were investigated with the version using the Leaky ReLU activation function showing slightly higher accuracy in the test set than the version that used ReLU. The version that used BN produced the least accurate and more unstable results. The Leaky ReLU version was evaluated in the training and test sets quantitatively by calculating four evaluation metrics and qualitatively by visually observing the produced cloud masks. Comparison with cloud masks produced by Sen2Cor and Fmask was performed for both evaluations.

It was shown that the CNN produced exceptional results (~98%) both in the training and the test set compared to the state-of-the-art threshold-based methods which performed by far less favorably. In more detail, the CNN managed to detect even clouds of very high transparency and successfully separated clouds from snow as well as bright urban and desert areas. Thus, the study further reinforces the value of CNNs in applications where spatial context is very important and shows that an architecture that makes use of a smaller number of layers and

feature maps compared to recent deep learning literature, consequently being simpler and more time-efficient, can produce very satisfactory results in cloud masking.

Besides observing the produced cloud masks, an initial effort was performed to observe the feature maps produced by the convolutional layers aiming to extract the weights of the kernels. In our opinion, a database formed by such kernels would be very useful since it can easily provide crucial features that could be input to several algorithms outside of the context of neural networks. The creation of such a database could be part of future work.

CHAPTER 3

Change detection in VHR imagery with severe co-registration errors[4] [5]

In Chapter 3 a study that was performed in this PhD thesis in the framework of change detection in VHR satellite images is presented. The study evaluates several state-of-the-art deep learning (DL) change detection (CD) methods on VHR images with severe co-registration noise. In section 3.1 background information is provided on the change detection methods proposed in the literature, available annotated datasets, and mitigation techniques for co-registration errors. In section 3.2 the motivations and objectives are described. In section 3.3 a brief theoretical background of the evaluated co-registration and DL CD methods is stated. In section 3.4 the procured images and the study areas are presented. Finally, in section 3.5 the results are discussed and in section 3.6 the conclusions and contributions are summarized and future work is suggested.

3.1 Related work

CD is an important Earth observation task that aims at monitoring land cover transitions through time for a given area. In the recent past, attention has been drawn towards VHR images because smaller objects (e.g. buildings) can be displayed in detail. However, moving to VHR increases significantly the complexity of the problem, since these images present increased within-class variance and geometric registration errors [52][53][186][187]. The successful completion of the task becomes even more challenging when the data are collected from different sensors since their heterogeneity is heightened [54][55][188].

3.1.1 Conventional CD methods

Among the well-known traditional pixel-based CD methods are algebra methods such as CVA [56][189][190] and transformation methods such as PCA [191] and MAD [192]. CVA computes the spectral difference and provides change intensity and direction [193]. PCA implies the assumption of a linear relation between no-change pixels belonging to the two acquisitions [57] and selects a part of the principal components for the CD [194]. MAD, based on CCA [195], also exploits unchanged pixels and aims at identifying changes from the canonical difference of multivariate images [57]. Since exploring spatial information is significant, object-based CD (OBCD) methods were developed, where the basic unit consists of pixels with similar spectral signatures [58][196]. Although OBCD is less sensitive to co-registration noise [197], the performance of these methods highly depends on the accuracy of the segmentation process which generally alters the geometry of the objects [58].

3.1.2 Mitigation techniques for co-registration errors in conventional CD methods

Several techniques were used in pixel-based CD to improve robustness to residual misregistration. The majority of the techniques were applied to medium-resolution images (spatial resolution: ~30 m). In [59], the authors proposed image smoothing via an average or median filter and alternatively adaptive grey-scale mapping which calculates total excess and deficit with respect to the image mean in a pixel window. In [198], an approach was proposed which utilizes bands where the investigated changes are not detectable, since co-registration noise is generally visible in all spectral bands. In [199], residual misregistration was detected by introducing a modeling approach that makes use of spatial brightness gradients, assuming that misregistration effects are locally uniform.

---

[4] **Kristollari, V.** and Karathanassi, V., 2022. Change detection in VHR imagery with severe co-registration errors using deep learning: A comparative study. IEEE Access, 10, pp.33723-33741. doi:10.1109/ACCESS.2022.3161978

[5] Karathanassi, V., Karamvasis, K., **Kristollari, V.**, Kolokoussis, P., Skamantzari, M., Georgopoulos, A., 2024, April. Remote sensing techniques for monitoring cultural heritage sites, In EGU General Assembly 2024, Vienna, Austria, doi:10.5194/egusphere-egu24-10181 (abstract)

In [56], a method was presented that detects co-registration noise by representing spectral change vectors in the polar domain and exploiting the direction distribution information. The same approach was followed in [190] for VHR images with the difference that the pixels in the adjacent neighborhood were also considered. In [60], the symmetric local co-registration adjustment (SLCRA) scheme was developed for HR imagery (~5 m). The method chooses corresponding pixels by calculating the minimum dissimilarity in a window. Finally, in [57], the same approach was followed to reduce minor misregistration errors in statistically similar entities.

3.1.3 Unsupervised DL CD methods

Recently, convolutional DL CD methods have drawn very high attention because of their innate ability to detect spatial context from raw data and their flexibility in the combined processing of different types of information. Another reason is the technological progress that has increased access to higher processing power systems. Hence, both unsupervised and supervised approaches have been proposed. Unsupervised methods are generally based on the comparison of feature maps produced by the bitemporal images. In [61], CNN feature maps of the pre-trained CaffeNet on Imagenet [14] were concatenated and the change map was computed using pixel-wise Euclidean distance. The same authors in a different study [200] compared features extracted from different zooming levels of the pre-trained VGG-16 [201] on the same dataset to produce the final change map. As a pre-processing step, they applied PCA and segmented the three higher uncorrelated channels into superpixels. A similar approach was followed in [62] with the difference that the pre-trained VGG-16 deep change features were refined by a variance ranking-based method to retain only the relevant features. In [202], low-rank-based saliency computation and deep feature representation were combined. VGG-16 was fine-tuned on the AID dataset [203] and after extracting multilevel CNN features from superpixels, saliency maps that indicate pixel change probabilities were generated. In [63], the authors proposed the creation of a difference image of the feature maps produced by U-Net [204] pre-trained for semantic segmentation on the Vaihingen dataset [205]. By using networks pre-trained on the same dataset, transfer learning on U-Net was applied in [206] and an unsupervised context-sensitive deep CVA framework was proposed in [207]. Automatically selected features were combined into hypervectors that were compared pixel-wise to obtain deep change vectors for multiclass CD based on the direction of change. Finally, in [208], an unsupervised deep Siamese kernel PCA convolutional mapping network for binary and multiclass CD was designed. The multiclass CD was accomplished by a 2-D polar mapping.

3.1.4 DL CD methods for limited ground-truth data

Other studies have focused on approaches that avoid the costly annotation of samples. In [209], a Siamese version of VGG-16 pre-trained on AID was extended by adding a deep feature difference CNN and then transfer learning was applied by training on a small sample of VHR images with annotated changes. The final change map was created by a threshold. In [193], the authors applied an automatic pre-detection method of the training data and proposed a deep Siamese convolutional multiple-layers recurrent NN (RNN), which can be used both for homogeneous and heterogeneous images. Finally, in [210], pre-disaster OpenStreetMap building data were used to automatically generate training samples for a modified version of U-Net, where residual connections were added.

3.1.5 Available CD annotated datasets and mitigation techniques for co-registration errors in supervised DL CD methods

The increase in the availability of annotated CD datasets has greatly accelerated the research of supervised methods, which usually produce more accurate results. The SZTAKI AirChange Benchmark Set (1,5 m/px) [70] was the first VHR CD dataset that was made publicly available. This dataset has been used in many studies. In [211], the authors used it to train a Siamese CNN by the weighted contrastive loss. The changes in the image pair were detected by the distance of the feature vectors and the final output was produced by a threshold and a k-NN approach. In [212], three fully CNNs were trained on the SZTAKI dataset and instead of concatenating both connections of the encoding streams of the Siamese versions, the absolute value of their difference was concatenated. The same dataset was used in [213] to train the DeepLabV2 [214] network by an improved triplet loss function. The network was pre-trained on the PASCAL VOC 2012 dataset [215]. In addition, the SZTAKI and a building CD dataset were used in [216] to train a deep NN architecture based on the combination of an

attention mechanism with information transmission by the use of bidirectional LSTMs. Finally, in [217], a modified version of U-Net was trained on the SZTAKI dataset by using a depth-wise separable convolution making the network lighter and more efficient.

Lately, more datasets have been created to promote research in the field. In [72], the first large-scale VHR semantic CD dataset was presented and several fully CNNs for semantic CD were proposed. In [218], another dataset was used which is composed of multisource VHR images with annotated multitype changes [71]. In this study, a multiscale convolution module was incorporated into an FCN. The authors also proposed a combination of the weighted binary cross-entropy loss (WBCE) and the dice coefficient loss to improve the training of imbalanced samples. Finally, in [219], the focus was put on semantic CD and a Siamese framework with a global hierarchical (G-H) sampling mechanism was trained on three datasets with semantic annotated changes [220] [221]. The purpose of the G-H sampling mechanism is the mitigation of the imbalance problem. The authors also used the binary change mask to constrain the semantic CD results.

It is noted that since DL methods capture spatial information, it logically follows that they perform better in misregistration scenarios than pixel-based methods that exploit only spectral information. Recently, to further enhance their spatial context perception, many studies have adopted spatial attention mechanisms because they capture long-range spatial dependencies which leads to the reduction of pseudochanges [64]. Spatial attention highlights meaningful spatial relationships through the reweighting of the feature maps [53]. The authors in [64] implemented dual attentive fully convolutional Siamese networks to examine spatial and spectral long-range dependencies. They also addressed the imbalance sample problem by using the weighted double-margin contrastive loss. The network was trained and evaluated on two datasets, the multisource VHR dataset proposed in [71] and two VHR image scenes with annotated changes of buildings (WHU building dataset) [73]. In [65], a Siamese-based spatial-temporal attention CNN was introduced, along with one of the largest CD datasets in the field (changes related to buildings). In [53], an end-to-end network, called the pyramid feature-based attention-guided Siamese network was proposed. The authors introduced a co-attention mechanism and trained the network on two different building CD datasets: WHU (orthoimagery) and a challenging dataset of satellite images (with displacement). In [222], a dual-task constrained deep Siamese CNN, which contains a CD network and two semantic segmentation networks, was presented along with a dual attention module. It was trained on the WHU building dataset. In [223], deep features were extracted from a fully convolutional two-stream architecture and were fed into a deeply supervised difference discrimination network. Deep features of the raw images were fused with image difference features by attention modules and change map losses were also introduced in the intermediate layers. The CNN was trained on the dataset created in [71] and on a multisource Google Earth dataset. Finally, in [224] a scheme was proposed that contains an efficient convolution module in combination with fusion strategies based on spatial/spectral attention. The network was trained on the dataset proposed in [71] and on a recent version of WHU with semantic changes.

Even though attention mechanisms dominate the current literature on mitigating the effects of co-registration errors on VHR CD, some other approaches have also been proposed. In [225], three encoder-decoder-structured CNNs were designed to yield change maps from RGB satellite images with small color variations and co-registration errors and a large fully-labeled dataset of Google Earth images was constructed. The ensemble of the networks outperformed each individual CNN. In [71] a conditional adversarial network was trained and evaluated on synthetic images with a small relative shift. Finally, in [52] a framework that consists of two parts was proposed by use of the WHU dataset. It involves a building change detection network that takes bi-temporal binary building maps produced from a building extraction network. The authors simulated arbitrary building changes and various building parallaxes in the binary building map to increase robustness to co-registration errors.

3.2 Motivations and objectives

Although the current scientific research concerning DL CD with co-registration noise has shown promising results, it has mostly focused on images with minor co-registration errors. Based on that, the goal of this study is to assess several state-of-the-art DL CD methods on VHR images with severe co-registration noise. The study evaluates the performance of five state-of-the-art deep DL CD methods, two unsupervised and three supervised on four urban study areas of different morphology. The VHR images are selected from various satellites and exhibit high geometric distortions and co-registration errors. The fundamental logic behind the selection of the DL CD methods was the representation of each main category. Another reason was the public availability of the code proposed by

the creators of the methods, to ensure correct implementation. Thus, the first unsupervised method [61] is a pre-trained network that follows a patch-to-pixel approach, while the second unsupervised method, which was developed for the purpose of our study, has an encoder-decoder architecture and was trained on the study data. Concerning the selected supervised methods, the first (FDCNN) [209] avoids the costly annotation of samples by applying transfer learning by training on a small annotated CD sample of multitype changes, while the second (DASNet) [53] and the third (STANet) [65] apply spatial attention mechanisms to capture long-range spatial dependencies. DASNet was trained on the multisource VHR dataset proposed in [71] (multitype changes) and on the WHU building dataset, and STANet on a large dataset with changes related to buildings. It is noted that the supervised networks were implemented by use of the weights provided by the creators of the methods.

Before applying the CD process, four automatic co-registration methods were evaluated since this pre-processing step is extremely important for the success of the CD problem. The selected methods cover a wide range of the existing literature approaches. The first two are Scale Invariant Feature Transform (SIFT) [226] and the Oriented FAST and Rotated BRIEF (ORB) [227] which detect local features and assign descriptors. The third is a CNN approach [228] and the fourth is the Fourier-Mellin Transform (FMT) [229] which is a global method.

3.3 Theoretical background

3.3.1 Co-registration methods

Four methods were tested for the automatic co-registration of the images. These methods were SIFT, ORB, a CNN feature-based approach, and the FMT. The selected methods cover a wide range of the existing literature approaches.

3.3.1.1 SIFT

SIFT locates local features known as "keypoints" that are scale and rotation invariant. The keypoints are detected by creating different scales of the images (application of Gaussian blur) and locating local maxima and minima. Then, their orientation and magnitude are defined by calculating gradients. Thus, a unique fingerprint is created for each point called "descriptor". The method consists of four parts: Scale space extrema detection, accurate keypoint localization, orientation assignment, and keypoint descriptor generation.

3.3.1.2 ORB

ORB is a fusion of FAST (Features from accelerated segment test) [230] keypoint detector and BRIEF (Binary Robust Independent Elementary Features) [231] descriptor with modifications to enhance the performance. FAST is a corner detection method and BRIEF assigns descriptors by selecting a random pair of pixels in the neighborhood of a keypoint from a Gaussian distribution and comparing their brightness. The FAST modifications refer to the use of a multiscale image pyramid and the assignment of orientation, whereas the BRIEF modifications to the inclusion of orientation invariance.

3.3.1.3 Co-registration with a CNN

The CNN feature-based approach uses a CNN to generate multiscale feature descriptors and then the Expectation Maximization (EM) method [232] is applied to gradually increase the selection of inliers. After detecting a feature point set X from the referenced image and a feature point set Y from the sensed image, the transformed locations of Y (Z) are obtained. The multiscale feature descriptors are generated using three pooling layers ($D_1(x)$, $D_2(x)$, $D_3(x)$) from a pretrained VGG-16 network on the Imagenet dataset. After designing a grid, the feature point is determined as the center of each grid cell. Features x and y are matched according to Equation (3.1).

$$d(x,y) = \sqrt{(2)}d_1(x,y) + d_2(x,y) + d_3(x,y) \quad (3.1)$$

where: $d_i(x,y)$: Euclidean distance of $D_i(x)$, $D_i(y)$.

Inlier selection produces an $M \times N$ prior probability matrix using both convolutional feature and structural information which is then taken by a Gaussian mixture model (GMM) based transformation solver. In order to

compute the matrix, at first an integrated cost matrix is computed using an element-wise Hadamard product. Then, the Jonker-Volgelant algorithm [233] is applied to solve the linear assignment on the cost matrix. Assigned point pairs are regarded as putatively corresponding.

Points in set Y are considered as GMM centroids and EM is then applied to find the optimal transformation parameters. The objective of the approach is to minimize the negative log-likelihood function. EM iteratively solves the non-rigid transformation (Equation (3.2)) and the selection of inliers is updated in every k iterations. The process consists of the E-step where the posterior probability matrix is computed from the last iteration and the M-step where the derivatives are solved and the parameters are updated. As a final step, the transformed image is calculated using thin plate spline interpolation.

$$Z = Y + GW \quad (3.2)$$

where: G: the matrix generated by a Gaussian radial basis function (GRBF) and W contains the transformation parameters.

3.3.1.4 FMT

FMT-based image registration is a global method since it uses all the image pixels of both images to denote the transformation parameters [234]. In this method, at first the FFT of the input images is calculated followed by the calculation of the magnitudes.

Then, the magnitudes are transformed to log-polar coordinates. Taking the Fourier transformation of a log-polar map is equivalent to the computation of the Fourier-Mellin Transform (Equation (3.3)) [235].

$$F_M(k_1, k_2) = \int_{-\infty}^{+\infty} \int_0^{2\pi} f(e^r cos\varphi, e^r sin\varphi) e^{j(k_1 r + k_2 r)} d\varphi dr \quad (3.3)$$

where: r, φ: log-polar coordinates and k: scale.

By applying phase correlation, the angle and the scale can be retrieved. After applying rotation and scale, phase correlation can be applied again and the translation can be calculated as the final step of the 2-D image registration.

3.3.2 Change detection methods

3.3.2.1 Unsupervised methods

The first unsupervised method was the patch-to-pixel CNN proposed in [61] (Figure 3.1). For its implementation [236], Tensorflow [170] and Keras [167] functions were applied. The method uses the VGG-19 architecture pre-trained on the Imagenet database. The size of the input image patches was 224 × 224 px and the output size was 112 × 112 px. Firstly, the feature maps are extracted from five convolutional layers ($Conv_1$, $Conv_2$, …, $Conv_n$) to exploit both the spatial (lower level features) and the semantic information (higher-level features). Since these features are not of the same size due to downsampling (pooling) operations, multilevel maps of the same size are concatenated after being resized to the same size (resampling operations), resulting in a higher-dimensional feature map.

The CD is performed using pixel-wise Euclidean distance in a feature space of k - dimension (Equation (3.4)). For the production of the final change map, the optimum threshold is defined by applying the Otsu [237] segmentation method, which detects the minimal intra-class variance of two classes. For the implementation of the first unsupervised method in our study, Otsu segmentation was applied on images of size 1120 × 1120 px, produced by joining 25 output patches (112 × 112 px) after resampling to the input size (224 × 224 px). It is noted that in the original implementation Otsu segmentation was applied on the output patches (size 112 × 112 px).

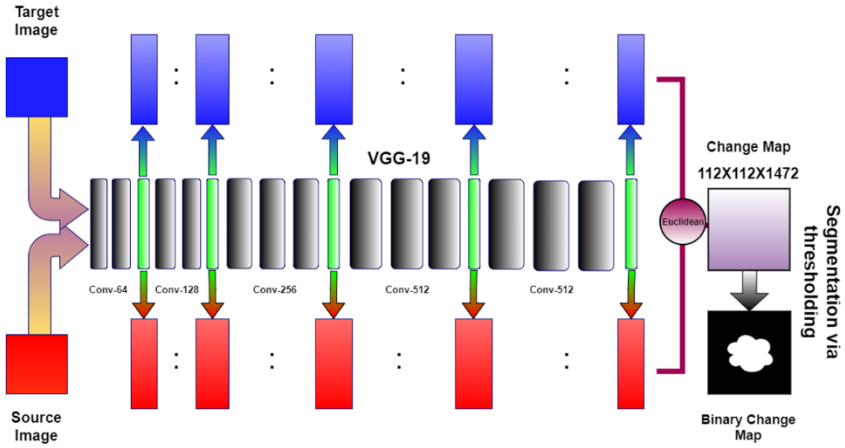


Figure 3.1. Unsupervised change detection approach proposed by El Amin et al. (2016)

$$d_{ij} = \sum_{k=1}^{k} ((\mu_i^k)^2 - (\mu_j^k)^2)^2 \qquad (3.4)$$

where k: feature dimension and μ_i^k and μ_j^k : features values at dimension k_{th} of the positions i and j.

In the second unsupervised method, which was developed for the purpose of our study, an autoencoder CNN with three convolutional layers in the encoder part (64, 32, 16 feature maps) and three convolutional layers in the decoder part (32, 64, 4 feature maps) was implemented by use of Tensorflow and Keras functions. The network was trained on patches of size: 224 × 224 of the images of the first date (four images in total (one per each study area)) and the visible and near-infrared (NIR) bands were used. The input patches were fed to the CNN by a generator function which randomly selected a study area and then a random batch of eight input patches. The model was trained for 400 epochs with 407 train steps on an NVIDIA 1070 Ti Graphical Processing Unit (GPU) for approximately six hours.

Then, similar steps to the first unsupervised method were followed. First, multilevel maps of the same size (128 ×128 px) were created via resampling for the first two and last two convolutional layers, and then the feature maps were combined to create the change map using pixel-wise Euclidean distance and manually applying an Otsu threshold for images of size 1120 × 1120 px.

3.3.2.2 FDCNN

The first supervised method was the feature difference CNN (FDCNN) [209] (Figure 3.2), which uses transfer learning on a CNN (VGG-16) pre-trained on the AID dataset [203] (30 aerial scene types) by training on a small sample of VHR images with annotated changes. For its implementation [238], the Caffe framework [239] was used.

The network consists of three main parts. The first part is a two-channel Sub-VGG-16 with shared weights, composed of the first three scales of VGG-16 with input size 224 × 224 px. The second part is the FD-Net where feature difference maps of three scales are created and normalized (Equation (3.5)). Before computing the feature difference maps, resampling is applied to generate maps of the same size. In addition, the second-period image (X_2) is differentiated from the first period (X_1) image to obtain accurate boundary information on the changes. The third part is the FF-net where the backpropagation of the network is realized by a simple CNN with few training points, which produces the final change magnitude map (CMM).

$$\mathrm{FD}(i) = \frac{\left|F_1^i - F_2^i\right|}{\max\left(\left|F_1^i - F_2^i\right|\right)}, i = 1, \ldots, N \qquad (3.5)$$

where FD: the feature difference map, F_1^i, F_2^i: the feature maps with inputs X_1, X_2, and N: the total number of feature maps.

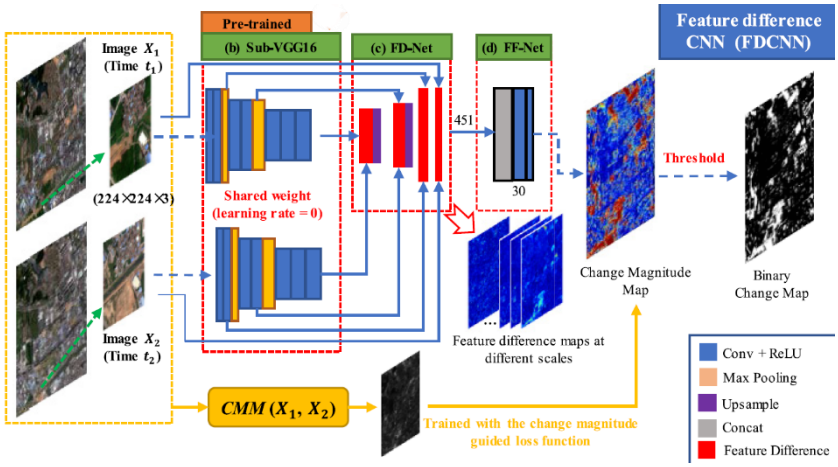


Figure 3.2. Flowchart of FDCNN proposed by Zhang & Shi (2020)

The network implements an improved cross-entropy loss that uses the change magnitude of each pixel as prior knowledge for learning and a weight loss function to alleviate the tendency of the network to no-change miss-detection due to unbalanced training data. CMMs are generated by applying CVA on X_1, X_2. After obtaining the CMM, the final change map is obtained by a threshold.

3.3.2.3 DASNET

The second supervised method was the dual attentive fully convolutional Siamese network (DASNet) [64] (Figure 3.3), which aims at capturing long-range dependencies. The network was trained on two CD datasets. One composed of multisource remote sensing images with multitype annotated changes (spatial resolution of 3 to 100 cm/px) [71] and one composed of two VHR image scenes with annotated changes of buildings (WHU building dataset) [73]. For its implementation [240], the Pytorch library [241] was used.

First, the Siam-Conv module is used to generate local features: F_{t0}, $F_{t1} \in \mathbb{R}^{C\times H\times W}$, and then the dual mechanism is applied to establish the connections between them. The feature F is fed into three convolutional layers to obtain three new features: Fa, Fb, $Fc \in \mathbb{R}^{C\times H\times W}$.

For the spatial attention, Fa, Fb, Fc are reshaped to $\mathbb{R}^{C\times N}$. Then, matrix multiplication is conducted between Fb^T and Fa and a spatial attention map is obtained through a softmax layer (Equation (3.6)), which measures the connection between a feature at position i and a feature at position j. Fc is reshaped to $\mathbb{R}^{C\times N}$ and matrix multiplication with Fs is conducted. Finally, the result is reshaped to $\mathbb{R}^{C\times H\times W}$ and added to the original feature to obtain the final output (Equation (3.7)).

$$Fs_{ji} = \frac{e^{Fa_i \cdot Fb_j}}{\sum_{i=1}^{N} e^{Fa_i \cdot Fb_j}} \qquad (3.6)$$

$$Fsa_j = \eta \sum_{i=1}^{N} \left(Fs_{ji} Fc_j\right) + F_j \qquad (3.7)$$

where Fa, Fb, Fc: features succeeding Siam-Conv, F: original feature, η: scale parameter, and $N = H \times W$.

For the channel attention, F is reshaped to $\mathbb{R}^{C\times N}$ and then matrix multiplication is performed between F^T and F to obtain the channel attention map. Then, similar steps as in spatial attention are followed. Equation (3.6) and Equation (3.7) are used by substituting N with the spectral dimension since it captures the long-range context in the channel dimension. The features obtained through the dual attention mechanism are aggregated.

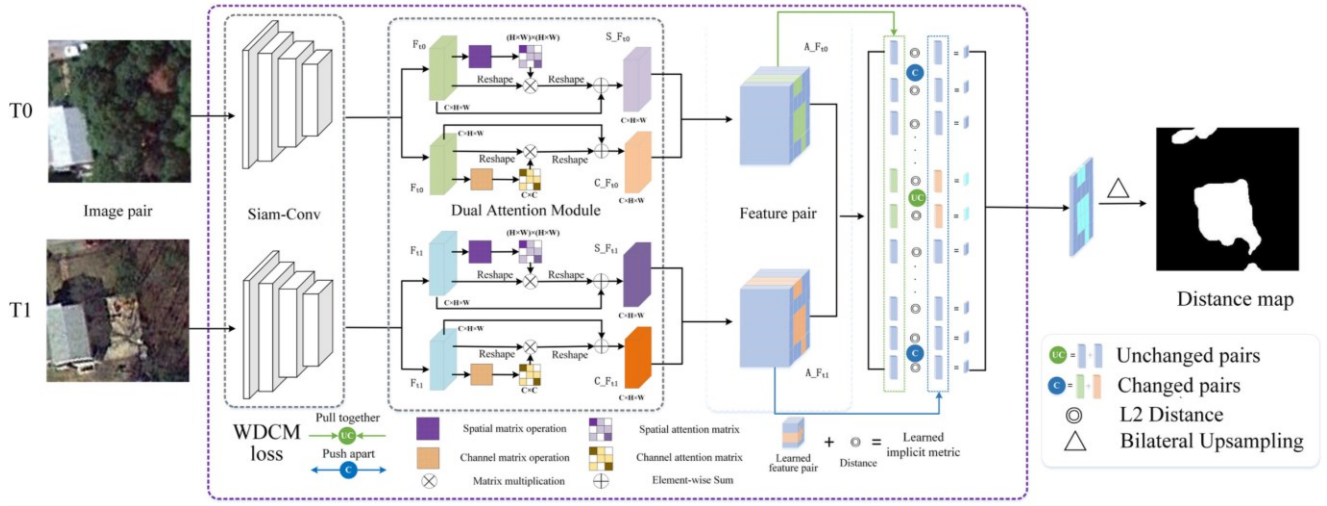


Figure 3.3. Overview of DASNet proposed by Chen et al. (2020)

The weighted double-margin contrastive loss was proposed to address the imbalanced sample problem. It is calculated for the spatial and channel attention modules: L_{sa}, L_{ca}, as well as the final output feature pairs L_e (Equation (3.8)).

$$\text{Loss} = \lambda_1 L_{sa} + \lambda_2 L_{ca} + \lambda_3 \quad (3.8)$$

where λ_i: weight of each loss

The output of DASNet is an RGB image patch of size 256 × 256 px. High red values show a high probability of change. Thus, for the implementation of DASNet in our study, the final binary change map was produced by applying an Otsu threshold in the Red band for images of size 1120 × 1120 px. These images were produced by joining output patches (256 × 256 px) after resampling to the input size, which was 224 × 224 px in the case of our study.

3.3.2.4 STANET

The third supervised method was the spatial-temporal attention-based network (STANet) [65] (Figure 3.4). The authors trained the network on a dataset that they proposed (LEVIR-CD), which contains professionally annotated changes related to buildings (soil/grass/hardened ground building). It was created from 637 VHR Google Earth image pairs (size: 1024 × 1024 px) from Texas, US and represents various types of buildings. For its implementation [242] the Pytorch library was used.

The network has a Siamese structure. First, an FCN (Resnet-18 [243]) is employed to extract the bitemporal image feature maps ($X^{(1)}$, $X^{(2)} \in \mathbb{R}^{C\times H\times W}$). Then, $X^{(1)}$, $X^{(2)}$ are stacked into a feature tensor $X \in \mathbb{R}^{C\times H\times W\times 2}$ and fed to the attention module to create two attention feature maps ($Z^{(1)}$, $Z^{(2)}$) (Equation (3.9)). The self-attention mechanism models attention weights between any two pixels.

$$Z = F(X) + X \quad (3.9)$$

where $Y = F(X)$ is a residual mapping of X to be learned.

Three tensors are introduced to illustrate the basic idea of the self-attention mechanism: query, key, and value, which are obtained from the input feature tensor through three different convolutional layers. The input feature tensor is the concatenation of the bitemporal image feature maps in the temporal dimension. X is firstly transformed into three feature tensors Q, K, $V \in \mathbb{R}^{C\times H\times W\times 2}$ and then Q, K, V are reshaped to the matrices $\bar{K}$, $\bar{Q} \in \mathbb{R}^{C'\times N}$ and $\bar{V} \in \mathbb{R}^{C\times N}$ where $N = H \times W \times 2$ and C' is the feature dimension. Q, K are used in the computation of the attention layer. Then, the spatial-temporal attention map $A \in \mathbb{R}^{N\times N}$ is defined as the similarity matrix (Equation

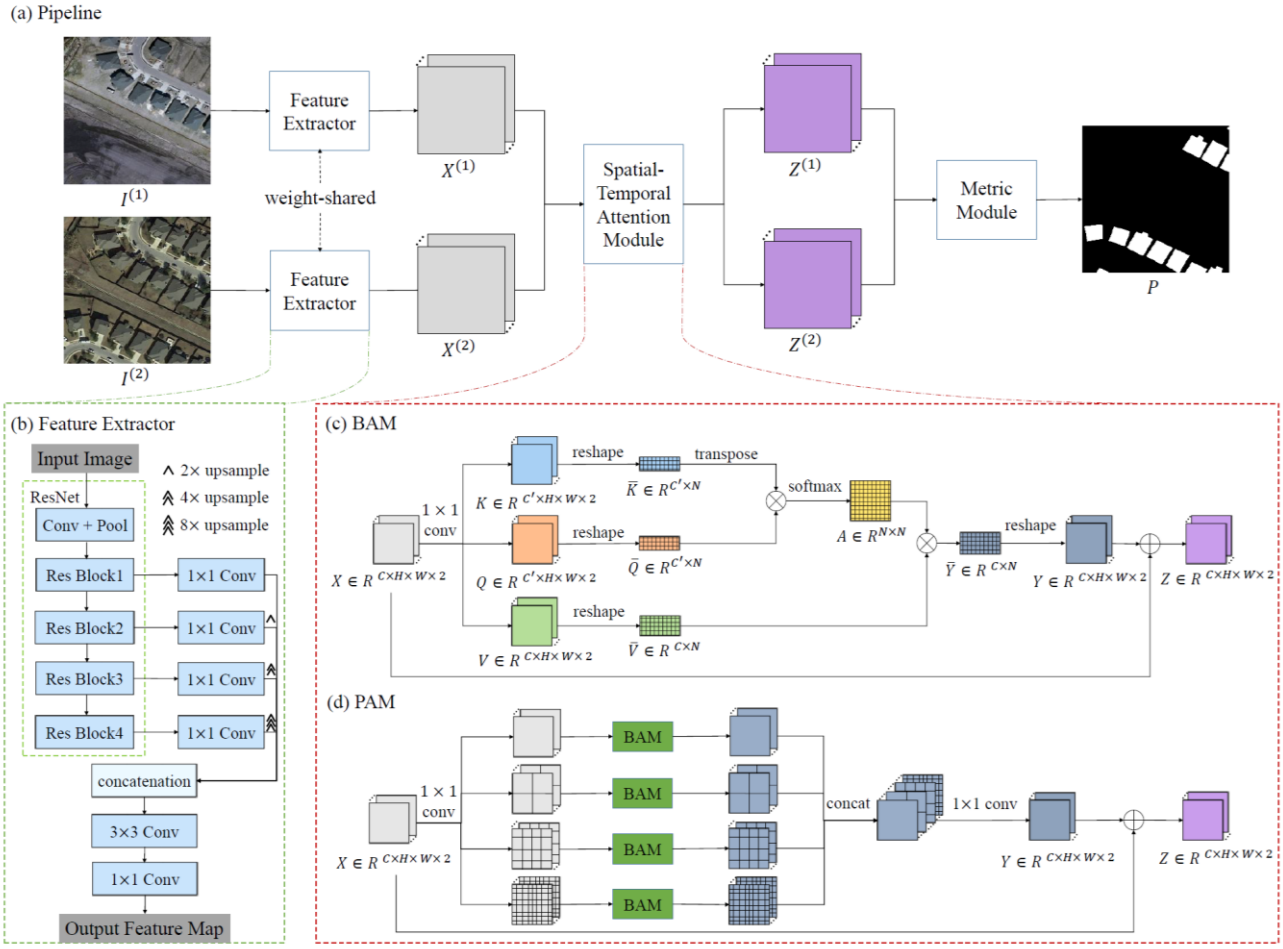


Figure 3.4. The pipeline of STANet proposed by Chen & Shi (2020)

(3.10)). Finally, the output matrix $\bar{Y} \in \mathbb{R}^{C\times N}$ is computed by multiplying $\bar{V}$ and A and then reshaping to $Y \in \mathbb{R}^{C\times H\times W\times 2}$.

$$A = softmax\left(\frac{\bar{K}^T\bar{Q}}{\sqrt{C'}}\right) \quad (3.10)$$

To capture spatial-temporal dependencies in multiple scales and alleviate misregistration issues a pyramid version is implemented, which has four branches of different scale. In each branch, the attention mechanism is applied in subregions and then aggregation is performed. The residual tensor Y and the original tensor X are then added to produce the updated tensor $Z \in \mathbb{R}^{C\times H\times W\times 2}$.

Finally, a distance map D is generated by calculating the distance between each pixel pair in the two feature maps by a residual function. During training, the model is optimized by minimizing the loss calculated by the distance map and the label map. In the testing phase, the label map is calculated by thresholding. The training is performed by a batch-balanced contrastive loss (BCL).

3.4 Data

3.4.1 Description of study areas

The satellite images used in this study were collected from four European areas: Tønsberg (Norway), Granada (Spain), Rhodes (Greece), and Venice (Italy). Tønsberg presents mostly buildings of low height with tiled roofs (gray or red tones). The urban structures are spread among large areas of forests and crops and a river also crosses

Figure 3.5. Locations and thumbnails of the four study areas

Table 3.1. Detailed information of VHR satellite images used for the land cover CD

Area	Collection date	Satellite	Resolution (m)	Size (km^2)
Tønsberg	20/9/2013	WV-2	0.5	25
	12/7/2019	GE01	0.5	
Granada	19/7/2013	GE01	0.5	21
	2/7/2018	WV-3	0.3	
Rhodes	23/4/2013	WV-2	0.5	33
	5/6/2019	WV-3	0.3	
Venice	4/5/2013	GE01	0.5	17
	13/5/2018	WV-3	0.5	

the region. Granada is characterized by a very dense urban fabric, which contains very high buildings with tiled roofs of red tones. The city is also enclosed by steep mountains and a few crops. The city of Rhodes is located on an island and shows a dense urban fabric of medium-height buildings with terraces. The relief is generally flat and there is a moderate quantity of crops. A substantial percentage of the Rhodes images is covered by seawater. Finally, Venice presents very homogeneous buildings with red-tiled roofs at very close distances. As in Rhodes, the Venice images are also surrounded by a high water percentage. The presence of a high amount of ships is also noticeable. The locations and thumbnails for all four study areas are shown in Figure 3.5.

3.4.2 Detailed information on procured images

For the detection of the land cover changes, VHR pan-sharpened images collected from Geoeye-1 (GE01) and Worldview-2/3 (WV-2/3) satellites were used. The images were globally co-registered and contained spectral information in the visual and near-infrared (VNIR) part of the light spectrum. Their time difference varied between five and six years and the area size between 17 and 33 km^2. The spatial resolution for GE01 and WV-2 images was 0.5 m, whereas for WV-3 was 0.3 m. Details about the images are shown in Table 3.1.

(a) (b)

Figure 3.6. Example of visible/non-visible facades in Venice because of the different satellite view angles. (a) Image collected on 13/5/2018 by WV-2. (b) Image collected on 4/5/2013 by GE01.

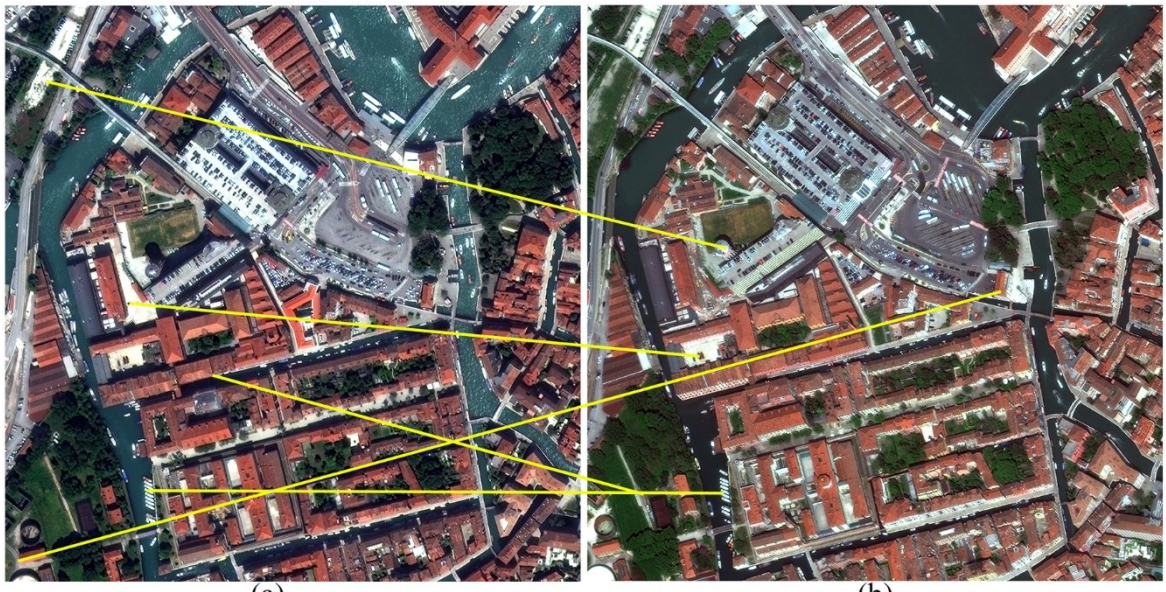
3.5 Results and discussion

Before implementing the CD methodology, the pre-processing steps were applied. These steps included: a) creation of mosaics from the WV-3 images since the area of interest was depicted in multiple tiles, b) resampling of the WV-3 images from 0.3 m to 0.5 m spatial resolution (same as GE01, WV-2), and c) co-registration.

3.5.1 Co-registration

It is important to note that the procured images were not orthorectified, thus the co-registration process was applied locally and not globally. In more detail, SIFT, ORB, and the FMT were tested on samples of size 1120 × 1120 px, whereas the CNN feature-based approach was tested on patches of size 224 × 224 px. The local approach is necessary because of the perspective view geometry that causes a non-uniform scale according to the relief. It should be also noted that because of the different satellite view directions and angles, the images cannot be co-registered with high accuracy (e.g. visible/non-visible facades) (Figure 3.6).

For both SIFT and ORB, the descriptor of one feature in the first set is matched with all other features in the second set using some distance calculation. During the matching process, outliers are excluded by the RANSAC (Random Sample Consensus) [244]. In our case, for both methods, many points were incorrectly matched. An example area in Venice showing incorrectly matched points detected by the SIFT method is shown in Figure 3.7. The image shows that the algorithm fails to generate point descriptors with the adequate information needed to produce correct matches.



(a) (b)

Figure 3.7. Example area in Venice showing incorrectly matched points detected by SIFT. (a) Image collected on 13/5/2018 by WV-2. (b) Image collected on 4/5/2013 by GE01

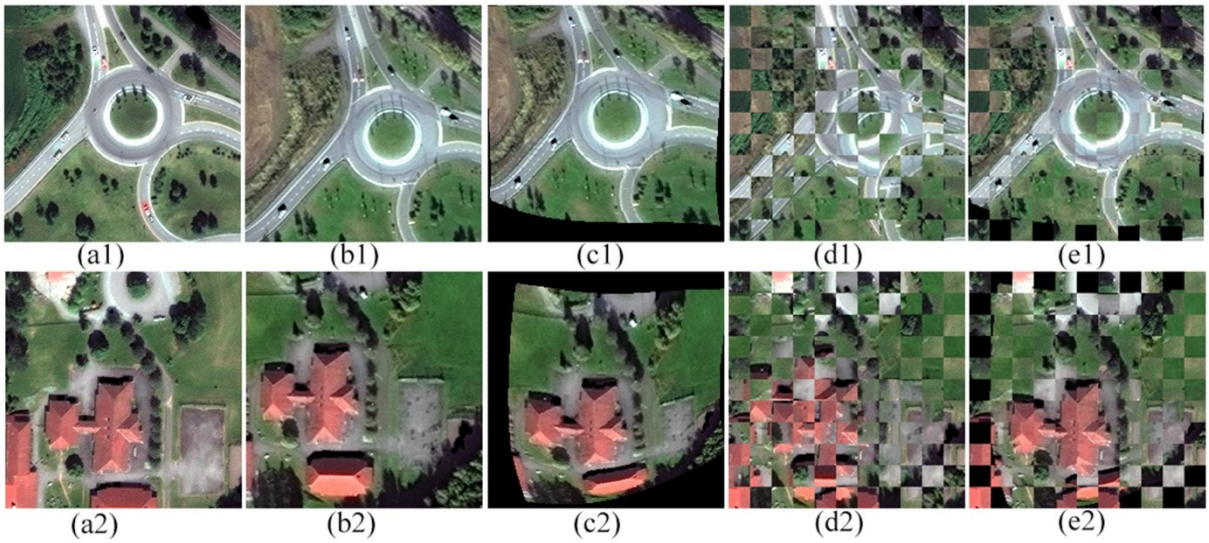


Figure 3.8. Example outputs of the CNN feature-based co-registration (Tonberg). (a1, a2) Image collected on 12/7/2019 by GE01. (b1, b2) Image collected on 20/9/2013 by WV-2. (c1, c2) Co-registered output. (d1, d2) Checkboard display of a1 & b1/ a2 & b2. (e1, e2) Checkboard display of a1 & c1/ a2 & c2

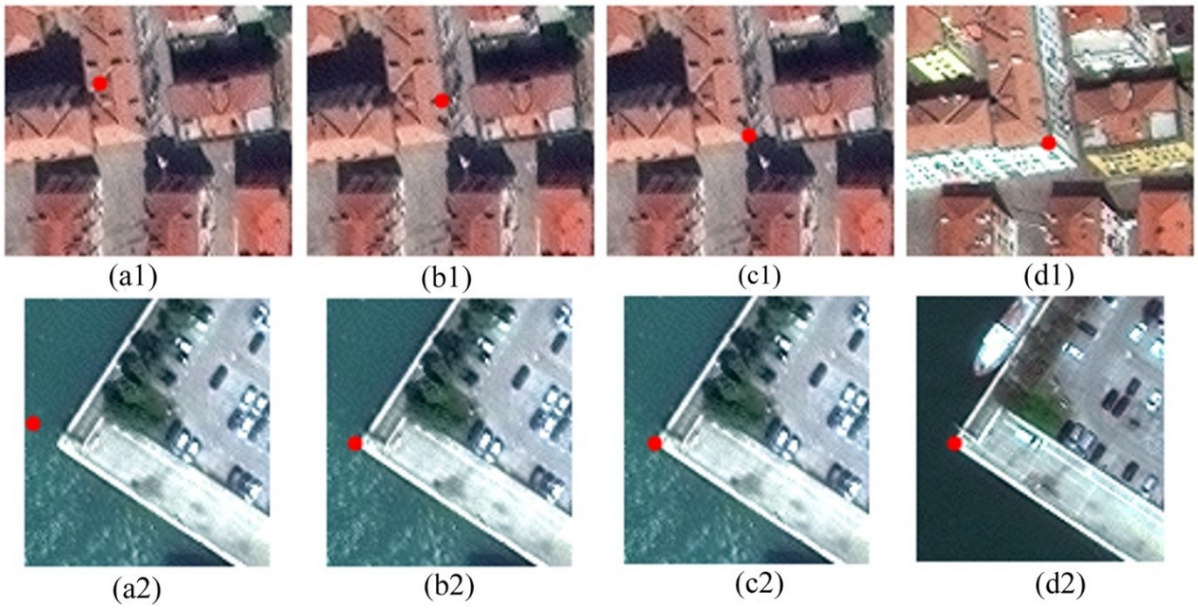


Figure 3.9. Comparison of Fourier-Mellin Transform and manual co-registration (Example outputs in Venice). (a1, a2) Image collected on 13/5/2018 by WV-2. (b1, b2) Co-registered output of Fourier-Mellin Transform. (c1, c2) Manually co-registered output. (d1, d2) Image collected on 4/5/2013 by GE01. The red bullet shows the position for a point.

Concerning the CNN co-registration method, it was observed that the results were inconsistent because they were closely reliant on the objects depicted in the tile. In more detail, the method performed well when a) urban structures with clearly denoted edges (e.g. buildings, roads) were present in the patch and b) the structures were situated in the center of the tile. However, when the patch presented fuzzy objects (e.g. crops), or the pixels were situated close to the borders of the patch, distorted outputs were produced. Figure 3.8 shows two examples of outputs for this method. A checkerboard display is also presented to make the results more easily perceptible.

FMT performed better than the other automatic co-registration methods, but still not as well as the manual approach where matching points are manually collected. Figure 3.9 shows a comparison of two examples of co-registered outputs produced by the Fourier-Mellin Transform and the manual approach. It is shown that the Fourier-Melin Transform shows lower accuracy in areas of variable relief.

Taking into consideration the performance of the four automatic co-registration methods analyzed above, it

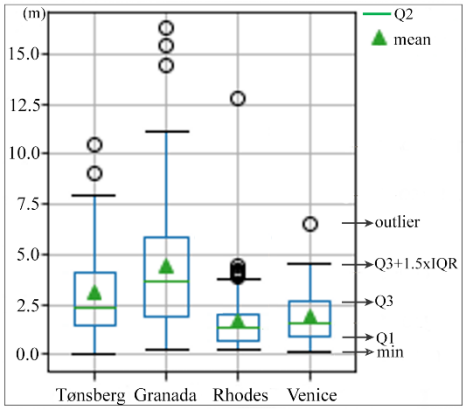


Figure 3.10. Box plots showing the distribution of the co-registration RMSE for the four areas of interest.

was decided to co-register the images manually, so that the co-registration errors are minimized as much as possible, given the case studies. For the implementation of the manual co-registration process, at first a grid with cells of size 1120 × 1120 px was created for each image and matching points were selected manually for 261 grid cells in total (Tønsberg: 84, Granada: 70, Rhodes: 59, Venice: 48). At least four points were selected for each grid cell and then the affine transformation was applied. The selection of the number of points was based on a visual evaluation of the scene height variance and the magnitude of geometric distortions. Thus, the number and height variance of the points increased according to the difficulty of each case.

Figure 3.10 shows the box plots of the RMSE for the four areas of interest. The RMSE was calculated by use of the points that had been selected for the manual co-registration. It can be seen that Granada showed the highest mean RMSE (~4m) followed by Tønsberg (~3m), Venice (~2m), and Rhodes (~ 1.5m). Granada also showed the highest variance as can be seen from the higher distance between the first (Q1) and third quartile (Q3) (~4.5 m) and the values of the outliers (isolated incidents) reaching RMSE values of ~15 m. Lower Q3-Q1 values are presented for Tønsberg (~2.5 m), Venice, and Rhodes (<2 m). The low variance for Rhodes could be explained by similar view directions of WV-2 and WV-3.

3.5.2 Change detection methods

The first unsupervised CD method and the three supervised applied in this study made use of the publicly available code proposed by the creators of each method, to ensure the correct implementation. It is noted that in this study we refer to the DASNet network trained on multitype changes as “DASNetCDD” and to the DASNet network trained on changes of buildings as “DASNetBCDD”. The methods were evaluated both qualitatively by visually observing the outputs of the methods and quantitatively by calculating evaluation metrics.

3.5.2.1 Qualitative evaluation

For the qualitative evaluation, several samples of outputs were observed for all the algorithms. Figure 3.11 shows the results for example areas in Tønsberg and Granada produced by the unsupervised and the supervised methods. Similarly, Figure 3.12 shows the respective results for example areas in Rhodes and Venice. The red square shows the significant changes.

The results of the first unsupervised method show a high commission error caused by different satellite view directions and angles (e.g. visible/non-visible facades), radiometric differences, and insufficient co-registration. It is noted that radiometric differences cause diverse spectral information for the same object and geometric distortions cause object shifts. Similarly to the results of the first unsupervised method, the results of the second unsupervised method show a high commission error caused by the same issues. The second unsupervised method also showed high sensitivity to seasonal changes (e.g. crops).

Since the unsupervised methods are based on comparing the distance of feature maps, it reasonably follows that a large number of pseudochanges will occur in the final result. It should be noted however, that feature maps display the object in various detail levels, thus the output is expected to show lower commission error than directly comparing the original bitemporal images.

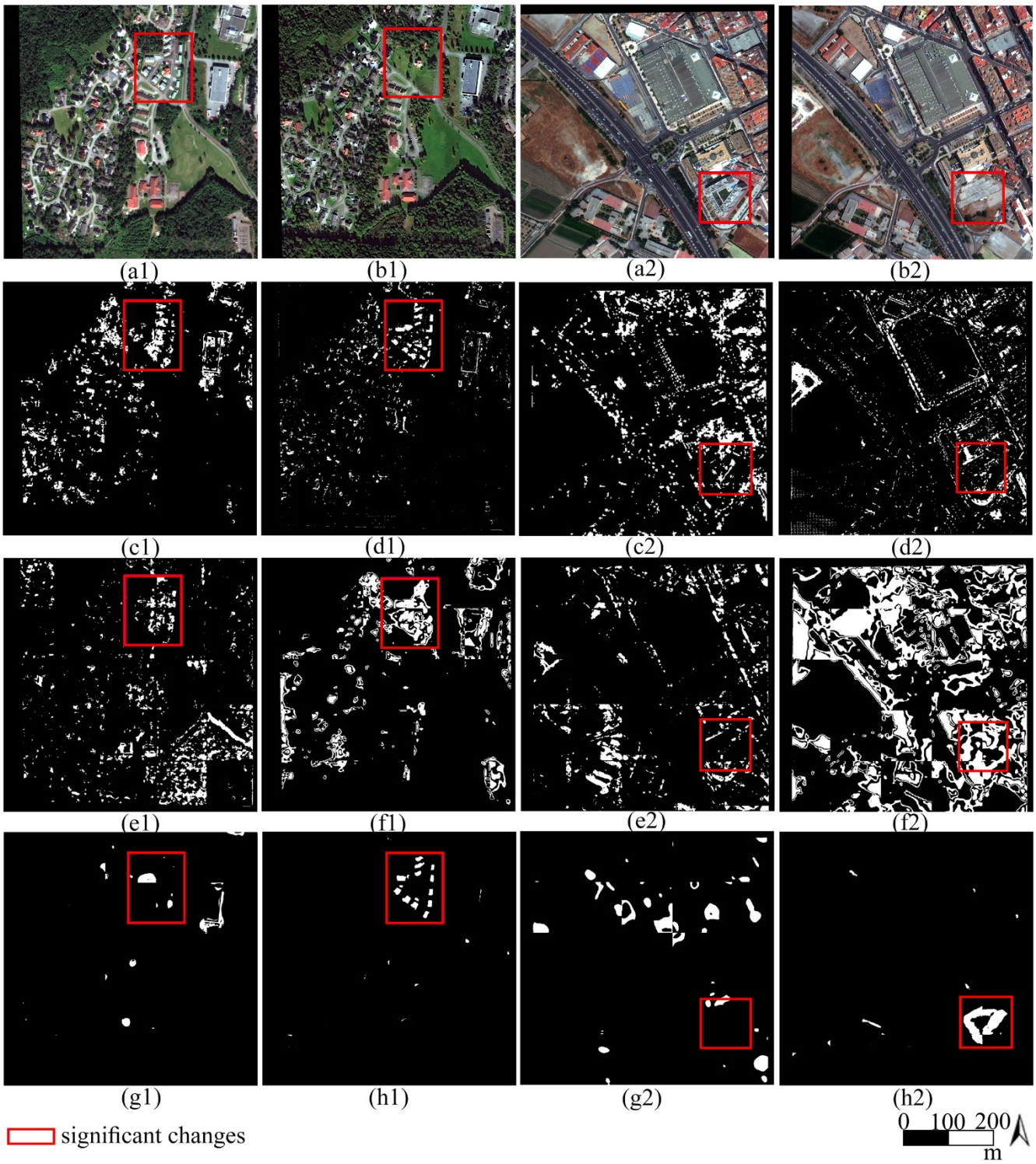


Figure 3.11. Example areas in Tønsberg (1st & 2nd column) and Granada (3rd & 4th column) showing results of the unsupervised and supervised methods. (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet.

Concerning the supervised CD methods, the outputs of FDCNN show a large commission error and it can be observed that even insignificant changes in vegetation scenes are incorrectly detected (mostly pseudochanges in the forest). Large commission error is also produced by DASNetCDD, where high sensitivity for radiometric differences is presented. It can be also observed that there is distortion in the shapes of the objects. It should be noted that the training set of DASNetCDD was dissimilar to our study areas (e.g. it contained images with snow).

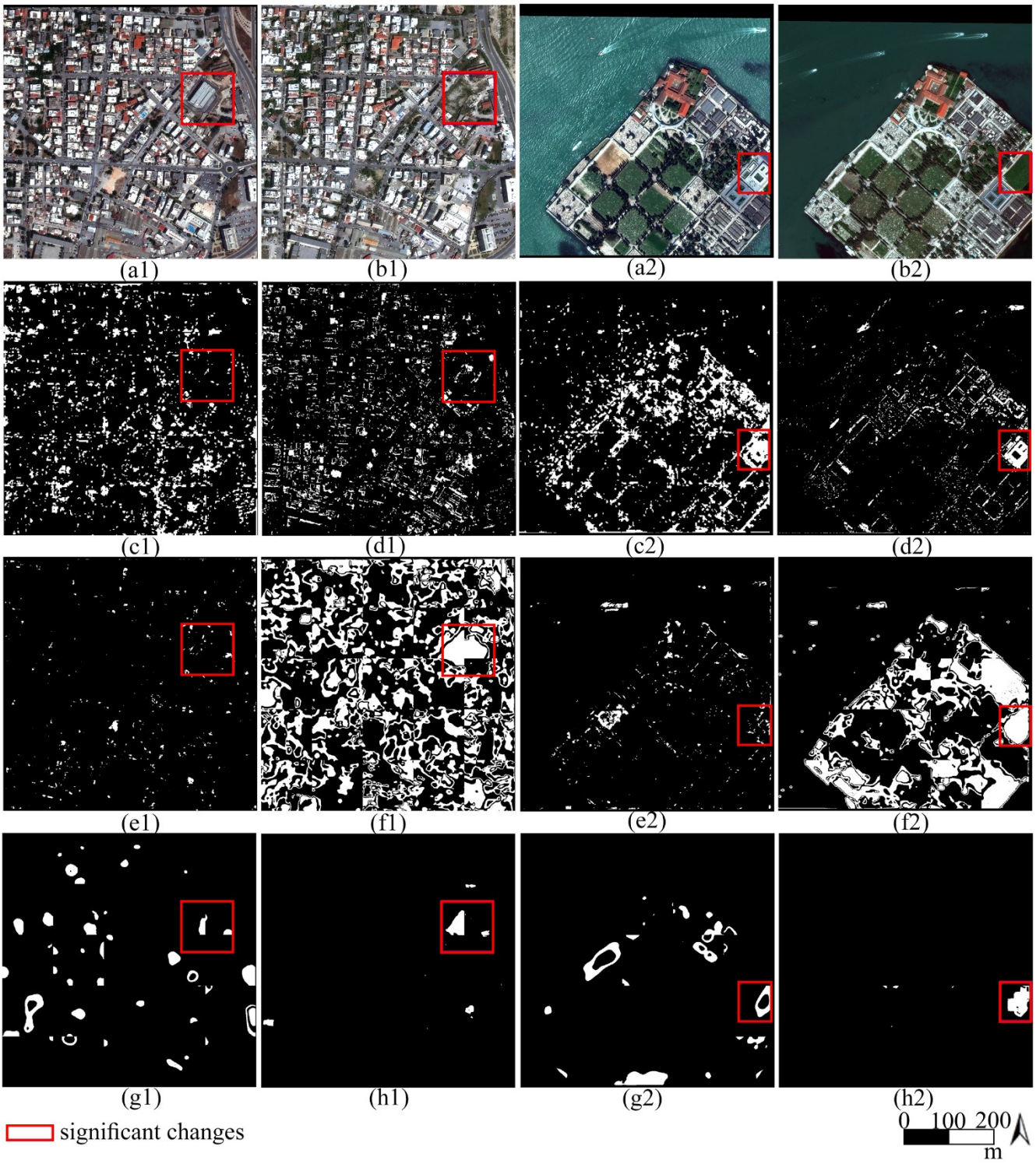


Figure 3.12. Example areas in Rhodes (1st & 2nd columns) and Venice (3rd & 4th columns) showing results of the unsupervised and supervised methods. (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet.

DASNetBCDD also incorrectly detects non-existent changes in buildings while simultaneously showing high omission error. Finally, better results are shown by STANet as it can be seen that changes related to buildings are detected more successfully than in all previously applied unsupervised and supervised methods. It can also be easily seen that the commission error is lower. The good performance of this method can be attributed to the proposed attention mechanism in combination with the large professionally annotated dataset.

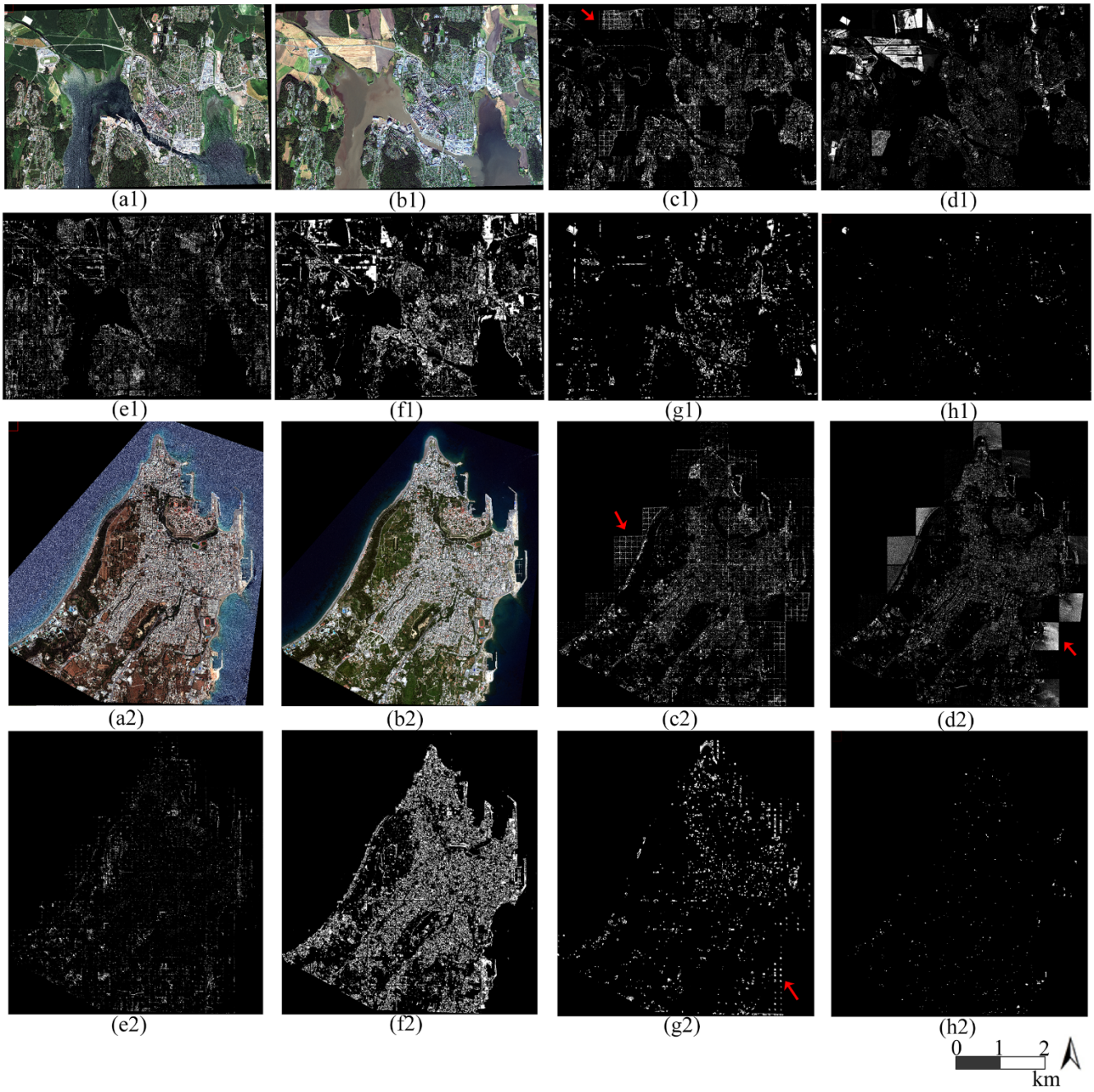


Figure 3.13. Results of the supervised and unsupervised methods for the whole area of Tønsberg (1st & 2nd rows) and Rhodes (3rd & 4th rows). (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet. The red arrows show edge noise or water pseudochanges.

Figure 3.13 shows the results produced by the unsupervised and the supervised methods for the whole study area of Tønsberg and Rhodes, and Figure 3.14 for Granada and Venice respectively. The observation of these figures leads to some further conclusions. In more detail, it can be seen that a) the first unsupervised method sometimes shows noise at the edges of the input CNN patch and b) the second unsupervised method and DASNetBCDD exhibit sensitivity to sunglint/watercolor differences. The above-mentioned issues are indicated by red arrows.

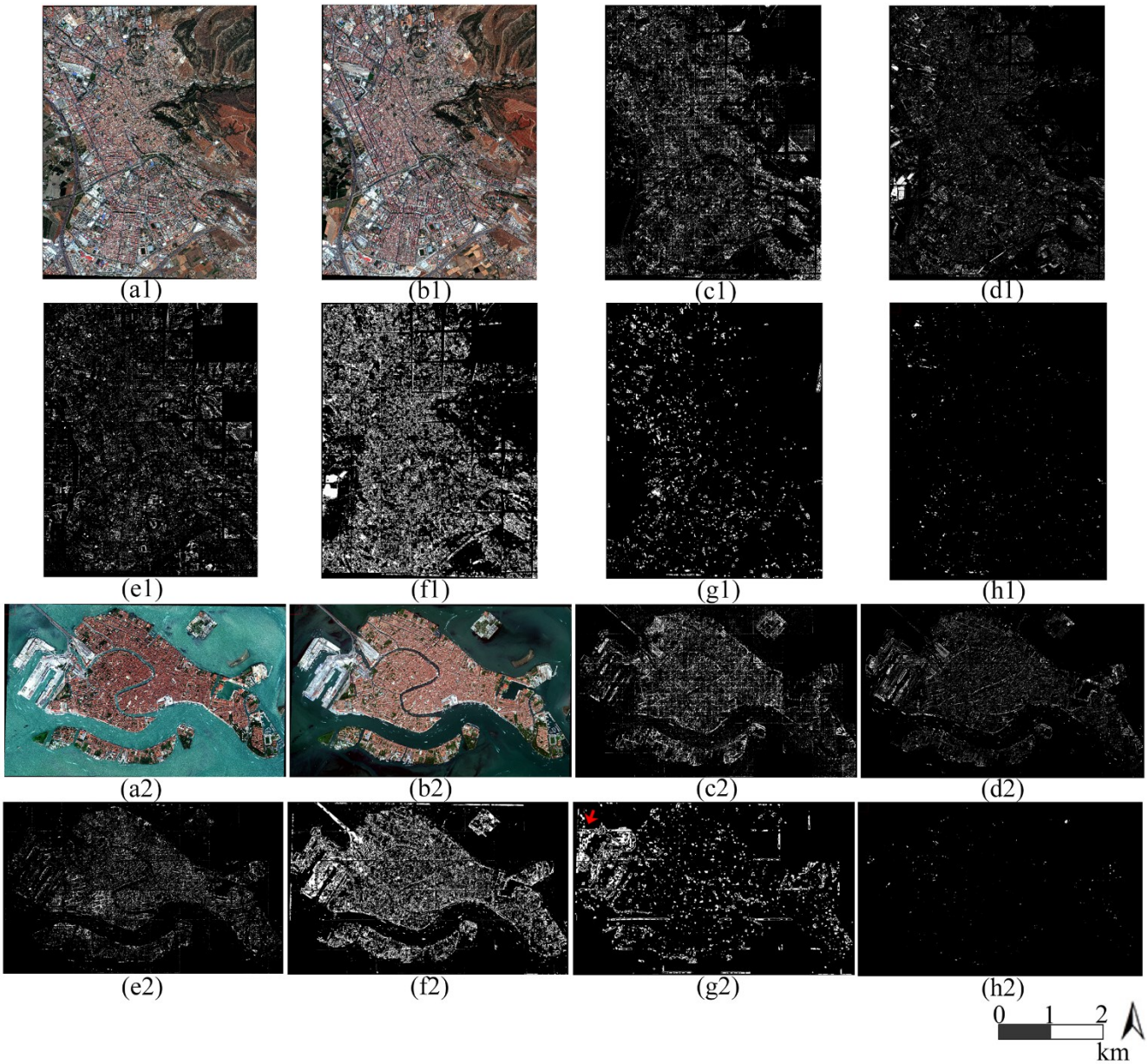


Figure 3.14. Results of the supervised and unsupervised methods for the whole area of Granada (1st & 2nd rows) and Venice (3rd & 4th rows). (a1, a2) Image of the latest date. (b1, b2) Image of the earliest date. (c1, c2) 1st Unsupervised method. (d1, d2) 2nd Unsupervised method. (e1, e2) FDCNN. (f1, f2) DASNetCDD. (g1, g2) DASNetBCDD. (h1, h2) STANet. The red arrow shows water pseudochanges.

3.5.3.2 Quantitative evaluation

Quantitative evaluation was performed by the calculation of metrics. These metrics were recall (Equation (2.16)), which corresponds to omission error, precision (Equation (2.17)), which corresponds to commission error, and Fscore (Equation (2.29)) which combines recall and precision metrics.

It is noted that false negatives were calculated by taking into account only the undetected buildings, whereas true positives by considering detected buildings as well as paving, roofs, and areas of dense tree growth (i.e. soil → forest). False positives were considered changes that are not of interest in this study (i.e. changes related to vehicles and seasonal changes (e.g. agricultural fields)) and pseudochanges. We categorized pseudochanges to those found in forests or in the water (e.g. sunglint) and to those caused by other reasons (e.g. co-registration and radiometric differences).

Table 3.2 Evaluation metrics for the results of STANet

Area	Recall	Precision	Fscore
Tønsberg	0.88	0.61	0.72
Granada	0.90	0.49	0.63
Rhodes	0.93	0.60	0.73
Venice	0.74	0.40	0.51
Training set	0.91	0.84	0.87

A. STANet evaluation for the whole study area

The STANet evaluation metrics for all four study areas and the training set (reported by the creators of STANet) are shown in Table 3.2. By observing the table it can be seen that the omission error is lower than the commission error. The lowest omission error is presented in Rhodes (7%) and the highest in Venice (26%). It is mostly observed in cases not present in the training set. The commission error is higher than ~40% for all study areas and can be attributed mainly to the co-registration errors caused by the different satellite view directions and angles. Radiometric differences were the second reason for the commission error. This error percentage is expected since in much better conditions (training set composed of images from the same satellite with small co-registration errors) the network showed a 16% commission error. Tønsberg and Rhodes present the lowest commission errors (~40%) and the highest Fscores followed by Granada and Venice. The pie charts displayed in Figure 3.15 show the percentages of the types of changes detected by STANet for the four study areas. The highest pseudochanges are presented in Granada and Venice because of the presence of high building blocks and the different view directions and angles of GE01 and WV. Another challenge for Granada was its mountainous terrain because geometric distortions are increased. The lower pseudochanges for Rhodes can be attributed to the similar view direction of WV-2 and WV-3, whereas for Tønsberg to the low building height and higher similarity with the training set. It should be noted that as shown in the box plots of Figure 3.10, Granada presented the highest mean RMSE in the co-registration process, whereas Rhodes the lowest. It is also interesting to notice the high amount of vehicles (ships) that exist in Venice and Rhodes.

B. Evaluation of all methods on the test set

The evaluation metrics (recall, precision, Fscore) for all the methods for a representative sample (test set (~20% of the results)) are shown in Table 3.3 for the unsupervised methods and FDCNN, and in Table 3.4 for DASNetCDD, DASNetBCDD, and STANet. In addition, a new evaluation metric was defined for the needs of the study ("precisionCD" (Equation (3.11)) that associates the commission error with the percentage of the pixels that were classified as change. We believe this index provides a better understanding of the magnitude of the commission error because it directly corresponds to its depiction in the image. The values of precisionCD for the test set are shown in Table 3.5. Finally, the percentages of the types of changes detected by all algorithms on the test set are displayed via pie charts in Figure 3.16 and Figure 3.17.

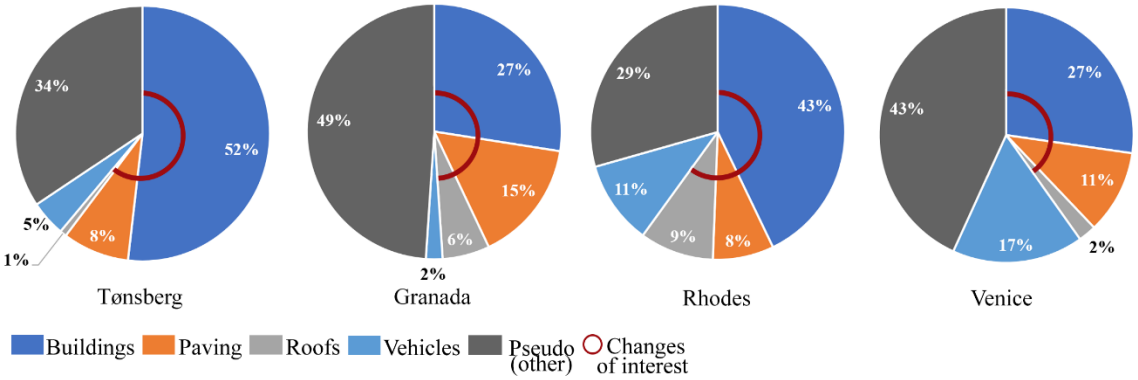


Figure 3.15. Percentages of the types of changes detected by STANet for the whole study area. The "pseudo (other)" category refers to changes caused mostly by co-registration errors and radiometric differences

$$\text{precisionCD} = (1 - \text{precision}) \cdot \%\text{CP} \quad (3.11)$$

where: CP: pixels detected as change.

In Table 3.3 and Table 3.4 it can be observed that DASNetCDD displays the lowest omission error (<9%) followed by the second unsupervised method (<14%). STANet and the first unsupervised method show an average omission error of ~15%, while the lowest performance is exhibited by FDCNN and DASNetBCDD with an average of ~25%. Regarding commission error, STANet shows the best performance (>37%) followed by the second unsupervised method with a minimum difference of 22%. The highest commission errors are shown by the first unsupervised method and DASNetCDD with an average of 80%. Similarly, STANet displays the highest Fscore with an average value of 0.66 followed by the second unsupervised method (0.45). The lowest Fscores are displayed by the first unsupervised method and DASNetCDD (~0.33). Regarding study areas, in general, Tønsberg and Rhodes present the lowest commission errors and the highest Fscores.

In Table 3.5 the values of the precisionCD metric show that when the commission error of STANet is translated into pixels, it is easily understandable that the pixels miss-classified by STANet as change, are 13 times less than DASNetBCCD which also focuses on changes of buildings. In addition, it can be observed that the commission error of DASNetCDD corresponds to the highest number of pixels and that the respective errors of the unsupervised methods, as well as of FDCNN and DASNetBCDD correspond to a similar amount of pixels.
From the pie charts displayed in Figure 3.16 and Figure 3.17 it can be seen that STANet presents the highest percentages of the changes of interest for all four study areas. It is noted that this behavior is expected since the percentage of the changes of interest directly corresponds to the precision values. In addition, STANet presents the lowest percentages of pseudochanges of the "other" category (e.g. co-registration errors, radiometric differences). Further interesting observations are the high sensitivity shown by: a) the second unsupervised method for the detection of seasonal changes followed by DASNetCDD and FDCNN, b) FDCNN for the detection of pseudochanges in the forest, and c) DASNetBCDD for the detection of pseudochanges in water (e.g. sunglint). It is noted that seasonal changes were included in the VHR images used in the training set of FDCNN. In addition, all methods are sensitive to the detection of changes in the presence of vehicles (mostly ships) and that both unsupervised methods show the highest miss-detection on this type of change. STANet shows the lowest percentage of vehicle changes. Finally, regarding study areas, Granada shows the highest percentage of pseudochanges of the "other" category in the results of all the algorithms while Tønsberg the lowest.

Table 3.3. Evaluation metrics on the test set (1st Unsupervised, 2nd Unsupervised, FDCNN)

	1st Unsupervised			2nd Unsupervised			FDCNN		
Area	Recall	Precision	Fscore	Recall	Precision	Fscore	Recall	Precision	Fscore
Tønsberg	0.86	0.27	0.41	0.93	0.37	0.53	0.78	0.28	0.41
Granada	0.76	0.19	0.30	0.86	0.28	0.42	0.74	0.26	0.39
Rhodes	0.88	0.24	0.38	0.96	0.37	0.54	0.73	0.29	0.42
Venice	0.82	0.11	0.19	0.89	0.18	0.3	0.77	0.17	0.27
mean	0.83	0.20	0.32	0.91	0.30	0.45	0.75	0.25	0.37

Table 3.4. Evaluation metrics on the test set (DASNetCDD, DASNetBCDD, STANet)

	DASNetCDD			DASNetBCDD			STANet		
Area	Recall	Precision	Fscore	Recall	Precision	Fscore	Recall	Precision	Fscore
Tønsberg	0.91	0.25	0.39	0.73	0.36	0.49	0.92	0.63	0.75
Granada	0.93	0.16	0.27	0.77	0.28	0.41	0.85	0.52	0.65
Rhodes	0.97	0.25	0.4	0.77	0.26	0.39	0.94	0.59	0.73
Venice	0.95	0.15	0.25	0.77	0.12	0.21	0.69	0.42	0.52
mean	0.94	0.20	0.33	0.76	0.26	0.37	0.85	0.54	0.66

Table 3.5. Calculation of precisionCD on the test set

Area	1st Unsupervised	2nd Unsupervised	FDCNN	DASNetCDD	DASNetBCDD	STANet
Tønsberg	0.0480	0.0552	0.052	0.0951	0.0282	0.0030
Granada	0.0802	0.0518	0.0536	0.2042	0.0313	0.0048
Rhodes	0.0563	0.0347	0.0129	0.1687	0.0188	0.0013
Venice	0.0855	0.0472	0.0378	0.1886	0.0689	0.0023
mean	0.0675	0.0472	0.0391	0.1641	0.0368	0.0028

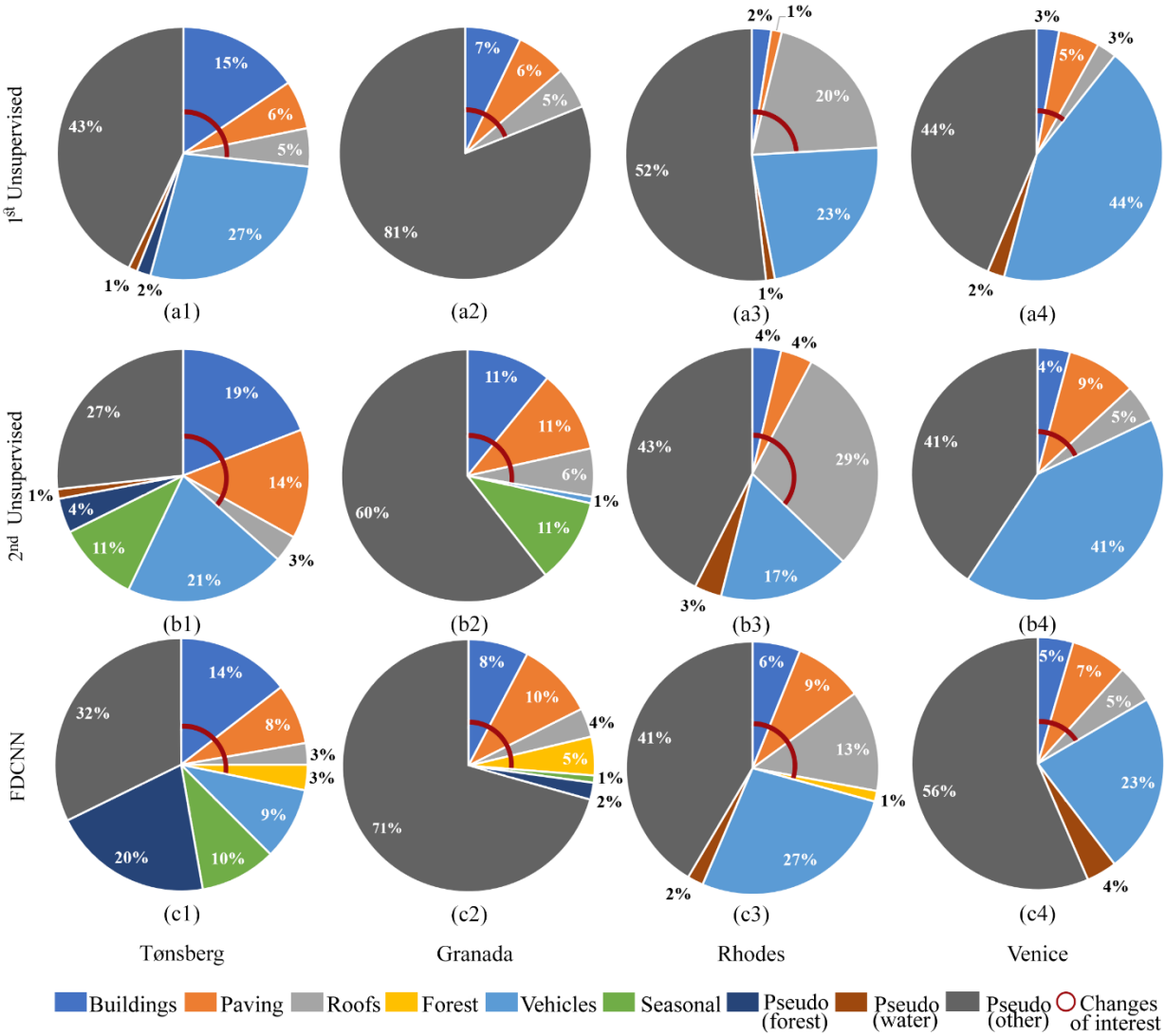


Figure 3.16. Percentages of the types of changes detected on the test set by the 1st unsupervised method (a1-a4), the 2nd Unsupervised method (b1-b4), and FDCNN (c1-c4)

Concerning the need for a human operator, it is not required for the implementation of the unsupervised methods as well as DASNet and STANet. However, for the implementation of FDCNN in our study, a threshold was manually selected. Finally, it should be noted that inference time for the second unsupervised method and STANet was ~ 0.05 sec for an image patch (size: 224 x 224) while for the rest of the methods (first unsupervised, FDCNN, DASNet) was ~ 0.3 sec. The methods were implemented in a machine with an i7-8700K CPU and NVIDIA 1070 Ti GPU.

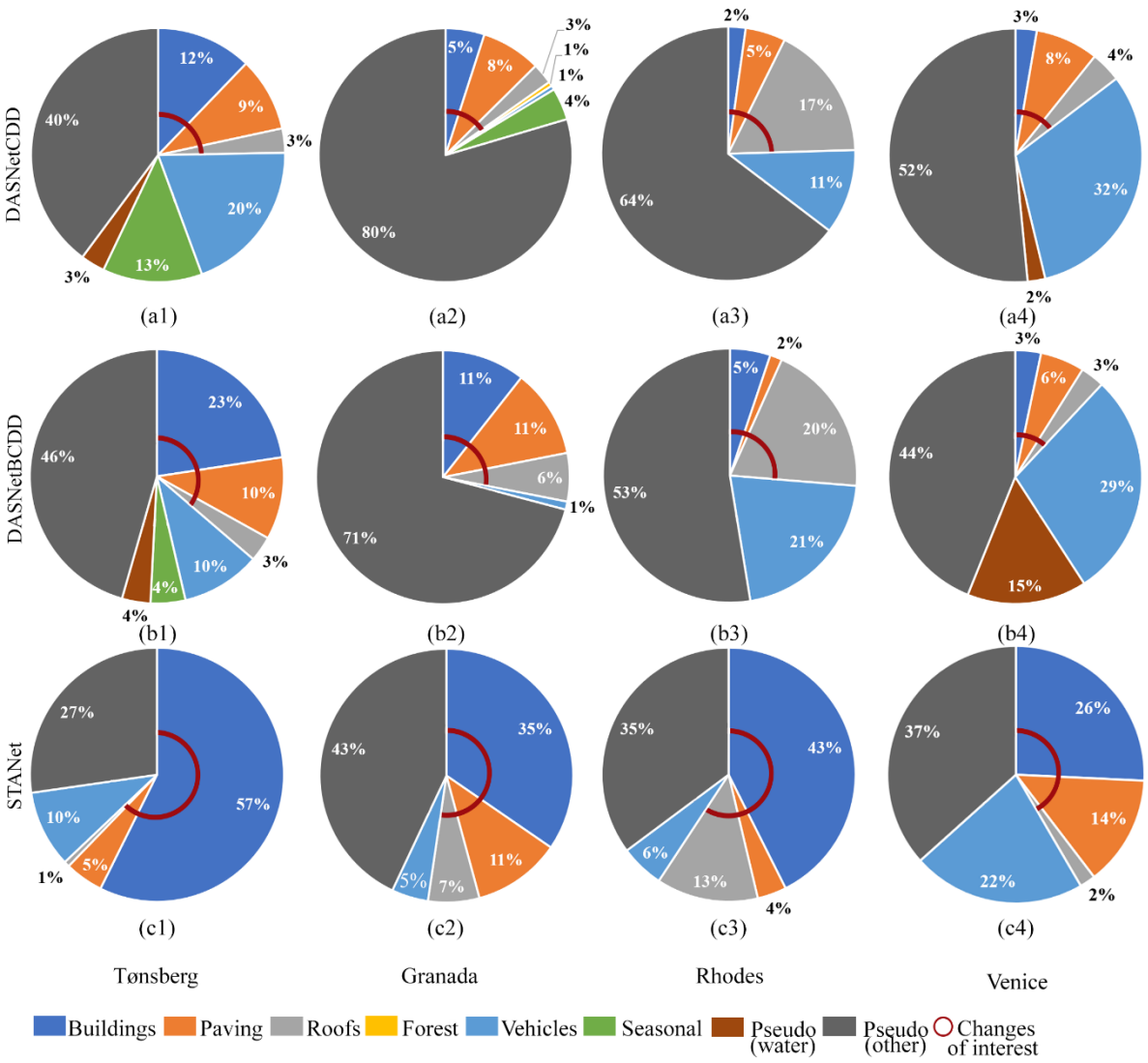


Figure 3.17. Percentages of the types of changes detected on the test set by DASNetCDD (a1-a4), DASNetBCDD (b1-b4), and STANet (c1-c4)

3.6 Conclusions

In this study, five state-of-the-art DL CD methods were evaluated for VHR images with severe co-registration errors. In addition, before applying the CD process, four automatic co-registration methods were evaluated because of the importance of this pre-processing step for the successful output of the CD algorithms. The study was performed on images depicting four European areas with versatile urban patterns.

The implemented co-registration methods covered a wide range of the existing literature approaches. It was observed that SIFT and ORB, as well as a CNN-based method, displayed low performance, while results were more satisfactory for the Fourier-Mellin Transform. However, given the crucial role of co-registration in the final CD result, it was decided to follow the more accurate manual approach, which produced a mean RMSE between 1.5 and 4 m.

Concerning the CD methods, two unsupervised and three supervised were applied. The supervised method called STANet, produced satisfactory results concerning the detection of buildings which are considered the most important indicator for the assessment of urban development. In addition, the commission error for this method was smaller than all other tested methods and was mostly attributed to the remaining co-registration issues. Its success can be attributed to the proposed spatial attention mechanism in combination with a large professionally annotated dataset. The other methods showed a high commission error caused by different satellite view directions and angles that caused geometric distortions, co-registration errors, radiometric differences, seasonal changes, and

changes related to vehicles. Heterogeneity between the training sets and the study data also affected the outputs in the supervised methods.

Besides the evaluation of the co-registration and the CD methods, another contribution of this study was the creation of a novel index called "precisionCD" that associates the commission error with the percentage of the pixels that were classified as change and provides a better understanding of the magnitude of the commission error.

In future work, the creation of annotated datasets with the challenges described in this study (high co-registration errors), would benefit the progress in the CD field.

CHAPTER 4

Marine plastic litter detection through image fusion

In Chapter 4 two studies that were implemented in this PhD thesis concerning the increase of spatial resolution in either the PRISMA or the S2 satellites through image fusion are presented. The ultimate goal of the fusion process is the production of outputs that facilitate the detection and monitoring of marine plastic litter. In the first study, three state-of-the-art deep learning pansharpening methods, originally proposed in the literature for VHR data, were evaluated in PRISMA images. In the second study, three deep learning networks were proposed and compared with relevant state-of-the-art literature networks (originally proposed for pansharpening or spatial super-resolution) for the fusion of S2 and WV-3 data. In both studies, the outputs of the deep learning methods were also compared with the outputs of conventional techniques (implemented by M. Kremezi). In addition, the plastic litter detection was performed by indexes. In section 4.1 the Remote Sensing background on marine plastic litter detection is provided. In sections 4.2 and 4.3 details on the two above-mentioned studies are specified.

4.1 Related work

Marine litter is defined as any persistent, manufactured, or processed solid material discarded, disposed of, or abandoned in the marine and coastal environment [66]. Marine litter may be found originating both on land (i.e. river discharges, flood water events, industrial, and recreational littering, etc.) and at sea (fishing, aquaculture, offshore mining and extraction operations, etc.). Land-based activities account for roughly 80 % of marine litter [67]. Plastics are the most prevalent debris found due to an increase in demand and production of plastic items over the last 70 years, as well as their slow decomposition [245][246]. High concentrations of marine litter endanger marine wildlife through entanglement, colonization of surface areas, or ingestion. The latter negatively impacts human health as well as the marine wildlife which is part of our food chain [68][69][247]. In addition to harming marine life, marine litter has a broad range of negative environmental, socio-economic, and maritime travel safety impact [247][68][69]. Although plastic litter has been reported since the 1960s, it has become a global environmental concern only in more recent years. Scientists estimate that by 2050 the plastic litter mass will outweigh the mass of the fish population [248].

At the global scale, the 2030 Agenda for Sustainable Development, adopted by the United Nations in 2015 [249], calls to conserve and sustainably use the oceans, seas, and marine resources with the Sustainable Development Goal No. 14. Among the SDG 14 targets, the 14.1 calls to prevent and significantly reduce marine pollution of all kinds, in particular from land-based activities, including marine debris and nutrient pollution. From the European perspective, the Marine Strategy Framework Directive (MSFD) requires the EU Member States to ensure that “properties and quantities of marine litter do not cause harm to the coastal and marine environment” [250].

Substantial waste detection, monitoring, and management challenges are faced with regard to plastic litter [69]. Satellite Remote Sensing has been identified as a useful tool for marine debris monitoring as it provides global and continuous temporal coverage [77][78] and has produced encouraging results in initial experiments on the detection of large-sized marine debris. However, it imposes some significant challenges in terms of atmospheric and sea-surface effects, spectral/spatial, and temporal resolutions, and availability of ground-truth data [79]. In [80] a spectral library from laboratory spectroscopic measurements was created and the laboratory analysis of various plastic materials revealed a) distinctive absorption patterns in the SWIR wavelength region, b) peak reflectance in the NIR spectrum, and c) variability in the VIS (visible) spectrum. The authors managed to efficiently detect Expanded Polystyrene (EPS) and Anthropogenic Marine Debris (AMD) mixtures deposited on three beaches of Chiloé island through machine learning techniques when applied to WV-3 images of 0.3 to 1.2 m spatial resolution. However, it was noted that the inclusion of other ranges of wavelengths may enable the detection of plastic objects in adverse weather conditions, as well as the distinction of types of waste in AMD mixtures other than EPS. In [81] spectral at-sensor properties derived from airborne SASI SWIR imagery with

pixel size 0.5×1.2 m^2 were employed for the distinction of ocean plastics from surrounding seawater using the unique absorption features of polymers. The authors used a reference spectral library of several polymer types to identify the plastic type of a large-sized ghost net for which spectral information from 11 SWIR pixels had been previously retrieved. They observed that both wet and dry plastic spectra have absorption features around 1215 and 1732 nm, and the reflectance of wet plastic is lower compared to the one of dry plastic, due to water absorption. Furthermore, the authors highlighted the need to further investigate the size distribution of observed pieces in relation to the pixel size. Lower reflectance for wet plastic spectra was also shown by [251] who verified a theoretical model of plastic reflectance. Their model, however, produced a smaller reduction compared to [81], a fact that could be explained by the differences in the experimental design and the properties of the selected plastic objects.

In [82] the first experiment with artificial floating targets was performed. In more detail, the spectral properties of three artificial floating plastic targets, as well as the surrounding seawater using Sentinel-1 and Sentinel-2 (S2) imagery, were investigated. These floating targets consisted of 10×10 m^2 PET (Polyethylene Terephthalate) -11.5 L water bottles, LDPE (Low Density Polyethylene) plastic bags, and nylon fishing ghost nets. In the optical data, all 10 ×10 m^2 plastic targets were distinguishable from the water due to their higher reflectance. In SAR data though, only the plastic bottles could be detected. The authors asserted that the identification of the plastic types and shapes requires multi- to hyper-spectral imaging. In a follow-up experiment [83], artificial targets in six S2 images in combination with UAV optical data were examined. Pixel coverage of plastic and linear spectral mixture were used to modify the spectra provided by the USGS spectral library. Matched filtering process followed to classify the pixels containing plastics. The methodology revealed promising results. Plastic litter targets were successfully identified when the plastic coverage of the S2 images was larger than 25% of the ground sampling distance (GSD). In [84] the coastal waters of Ghana, North-West America, Vietnam, and the east coast of Scotland were selected as case studies based on persistent or acute incidences of marine plastic litter reported in the scientific literature, popular press, and social media. The authors developed the Floating Debris Index (FDI), which allows detecting materials floating on the ocean surface at sub-pixel scales in S2 images. Then, they applied the Naïve Bayes algorithm on FDI, NDVI, and atmospherically corrected S2 images to compute the probability of a detected pixel belonging to each of the following classes: seaweed, spume, timber, macroplastics, and seawater. The detected pixels were assigned to the class with the highest probability. Candidate plastics were successfully classified as plastics with an accuracy of 86%.

In [252] the authors employed WV-2, ASTER, and SAR satellite datasets for monitoring marine plastic debris events after the great east Japan earthquake in March 2011, when a remarkable amount of >1.5 million tons of debris was generated. They employed satellite imagery to monitor plastic pathways and concluded that high spatial resolution satellite tracking reveals faster floating debris motions than expected within these regions. The same conclusion was drawn in [253] where high-resolution MS satellite imagery was used for the efficient monitoring of marine litter dynamics and the detection of its origin. The study also focused on the detection of the dominant marine plastic pathways. The authors detected and verified multiple floating plastic debris incidents using Planet, S2, and Landsat-8 data by systematically assessing the spectral signatures from pure floating plastics and discriminating them from other floating features on the sea surface such as sargassum, foam, etc. In [254] the Normalized Difference Hydrocarbon Index (NDHI) was developed for the detection of plastic on the shoreline using airborne HS data. This index is based on the Hydrocarbon Index (HI) [255]. The results were promising, indicating that subpixel detection is possible, while further investigation is needed to determine the minimum percentage of coverage that can be detected. Finally, in [256] the potential of supervised (SVR and Semi-supervised Fuzzy c-means (SFCM)) and unsupervised (k-means and Fuzzy c-means (FCM)) classification algorithms to detect floating marine litter was investigated. The authors used S2 images containing floating marine litter targets with sizes 10×10 m^2 and 3×10 m^2, and various combinations of bands and indexes as attribute sets for the classification algorithms. The supervised classification yielded higher accuracy, while the unsupervised algorithms provided many misclassifications. From the above, it was concluded that when artificial large-sized plastic targets are used, S2 spectral resolution (13 bands in the 440–2200 nm part of the spectrum) can detect marine plastic litter.

4.2 Pansharpening PRISMA data for marine plastic litter detection[6] [7] [8] [9]

In section 4.2 the study concerning the pansharpening of PRISMA data for the detection of marine plastic litter is presented. In section 4.2.1 the motivations and objectives of the study are stated and the contributions of the author of this PhD thesis are clarified. In section 4.2.2 the data collection and pre-processing is described. In section 4.2.3 the pansharpening methods that were implemented in this study are analyzed. Architecture and training details are provided for the deep learning methods. In section 4.2.4 three proposed plastic litter indexes are presented. Finally, in section 4.2.5 the results are discussed and in section 4.2.6 the conclusions and contributions are summarized and future work is suggested.

4.2.1 Introduction

Research has indicated that the key requirements needed by Remote Sensing techniques for improving the capability to detect the spectral signature characteristics associated with plastics, and even theoretically being able to discriminate between different polymers, are high spatial and spectral resolutions. So far, due to the technical and physical limitations of satellite sensors, there are critical trade-offs between the spectral and spatial resolution of satellite imagery. Data of high spectral resolution are characterized by low spatial resolution and vice versa. Nevertheless, the plastic crisis stresses the need to increase the current satellite observing systems' potential for marine plastic pollution detection and monitoring. Towards optimizing current observing systems' potentials to detect and identify plastics in marine litter, in this study, several pansharpening methods on the HS data provided by the PRISMA satellite are evaluated and an intersection of the outputs of three indexes is proposed to detect plastic objects efficiently. Medium resolution (30×30 m^2) PRISMA HS images cover a wide spectral range and have a fine spectral resolution (bandwidth ≤12 nm). Pansharpening with PRISMA PAN band could increase the HS data spatial resolution to 5×5 m^2 and their potential for detecting plastic debris in finer scales. The study focuses on the detection of small-sized plastic targets (≤5 m), which makes this research even more challenging. Through controlled experiments with various plastic target sizes, it contributes to investigating the undermost size of the observed targets in relation to the pixel size, as well as the way that the seawater influences the ocean plastic litter spectra. Finally, the study highlights the required pre-processing steps and contributes to evaluating the images provided by the recent HS PRISMA mission for marine litter detection. It is worth noting that not only PRISMA data but also satellite HS data are being evaluated for the first time for their potential to detect plastic litter.

It is noted that the contributions of the author of this PhD in this study refer to the implementation of the state-of-the-art deep learning pansharpening methods and the creation of the proposed plastic litter indexes. The conventional pansharpening methods and the image pre-processing steps were implemented by M. Kremezi. Finally, Assoc. Prof. K. Topouzelis was responsible for the construction of the experimental targets.

---

[6] Kremezi, M., **Kristollari, V.**, Karathanassi, V., Topouzelis, K., Kolokoussis, P., Taggio, N., Aiello, A., Ceriola, G., Barbone, E. and Corradi, P., 2021. Pansharpening PRISMA data for marine plastic litter detection using plastic indexes. IEEE Access, 9, pp.61955-61971. doi: 10.1109/ACCESS.2021.3073903

[7] Kremezi, M., **Kristollari, V.**, Karathanassi, V., Kolokoussis P., 2022, September. Enhancing PRISMA and Sentinel 2 Capabilities for Marine Plastic Litter Detection Using Image Fusion Techniques, Spectral Signature Unmixing and Spectral Indexes. In 41st EARSeL Symposium, Cyprus. (abstract + poster + oral presentation) (peer-reviewed)

[8] Aiello, A., Barbone, E., Ceriola, G., Karathanassi, V., Kolokousis, P., Kremezi, M., **Kristollari, V.**, Taggio, N., 2023, October. Unlocking the Potential of Spectral Signature Unmixing and Machine Learning for Detecting Plastic Marine Litter: Insights from the REACT Project. In ESA Remote Sensing of Marine Litter Workshop 2023 (abstract + oral presentation)

[9] Taggio, N., Aiello, A., Ceriola, G., Kremezi, M., **Kristollari, V.**, Kolokoussis, P., Karathanassi, V. and Barbone, E., 2022. A Combination of machine learning algorithms for marine plastic litter detection exploiting hyperspectral PRISMA data. MDPI Remote Sensing, 14(15), p.3606. doi:10.3390/rs14153606

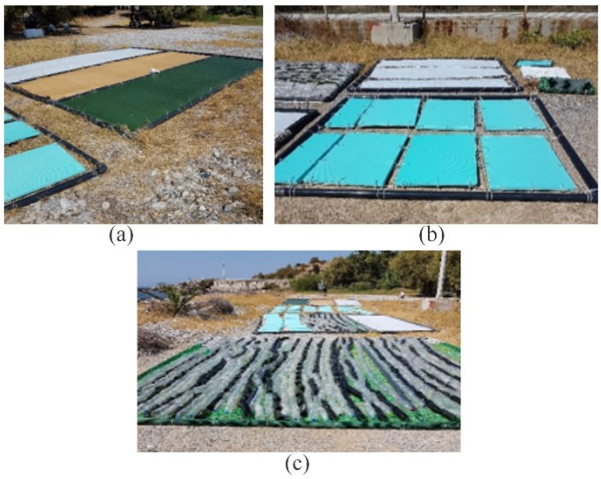


Figure 4.1. The targets. (a) Focus set on HDPE. (b) Focus set on PS. (c) Focus set on PET

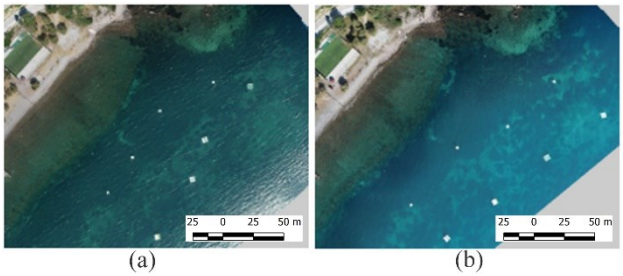


Figure 4.2. Orthophoto image of the targets offshore. (a) Date: 18/09/2020. (b) Date: 22/10/2020

4.2.2 Field campaigns

4.2.2.1 Data acquisition

The controlled experiments took place in the area of Tsamakia beach, in the coastal region of Lesvos island, Greece. The selected area offers plenty of unobstructed space guarantying the construction, deployment, and storing of targets during the experiments. Additionally, Tsamakia beach waters are sufficiently deep and the seabed offers a dark substrate that efficiently simulates deep waters.

For the experiment needs, 12 floating plastic targets were constructed in total. Their size was selected according to the spatial resolution of PRISMA which is expected to be achieved by pansharpening techniques, i.e. 5.1×5.1 m^2 (similar to the resolution of PRISMA fused data), 2.4×2.4 m^2 (nearly half of the resolution of PRISMA fused data), and 0.6×0.6m^2 (about 1/8 of the resolution of PRISMA fused data). For each one of these three different sizes of targets, four types/compositions of plastic materials with various colors were set up (Figure 4.1): a) HDPE (tarps in white, yellow, and green color), b) PET (transparent water bottles and green oil bottles), c) PS (sheets for building insulation in cyan color) and d) all the above materials in equal surface extent. HDPE as well as LDPE and PP are used to make common household items such as plastic bags. These materials have less density than seawater, causing them to float on the sea surface. PET, PVC, and PS are denser than seawater. They are usually observed on beaches and will most likely float on coastal seawater or close to ships before sinking and littering the seabed.

The analysis was performed on two clear sky PRISMA images collected on September 18th, 2020 and October 2^{2nd}, 2020. On the dates that the satellite passed over the test area, offshore deployment of the targets was carried out. A series of steel and cement anchors were used for the offshore deployment of the targets. The anchors were set above dark patches of the seafloor to minimize the reflectance contribution of a bright seafloor. The targets were deployed at a distance of 30 m from each other to minimize the possibility that more than one target would be captured in the same PRISMA pixel. They were set at varying sea depths due to area restrictions. Larger targets were set deeper (~12 m depth) than smaller targets (~2 m depth) (Figure 4.2). GPS instruments were attached to four of the targets used. In addition, on the experiments' dates, close-range RGB images were acquired using the on-board camera of a DJI Phantom 4 Pro V2.0 UAV. These images were orthorectified (Figure 4.2)

using the Agisoft Metashape software [257]. The spatial resolution of the orthophotos was around 2.5 cm depending on the flight height. In this resolution all the targets are well distinguished. However, the four 0.6×0.6 m^2 targets are not distinguishable at the scale of Figure 4.2.

4.2.2.2 PRISMA data pre-processing

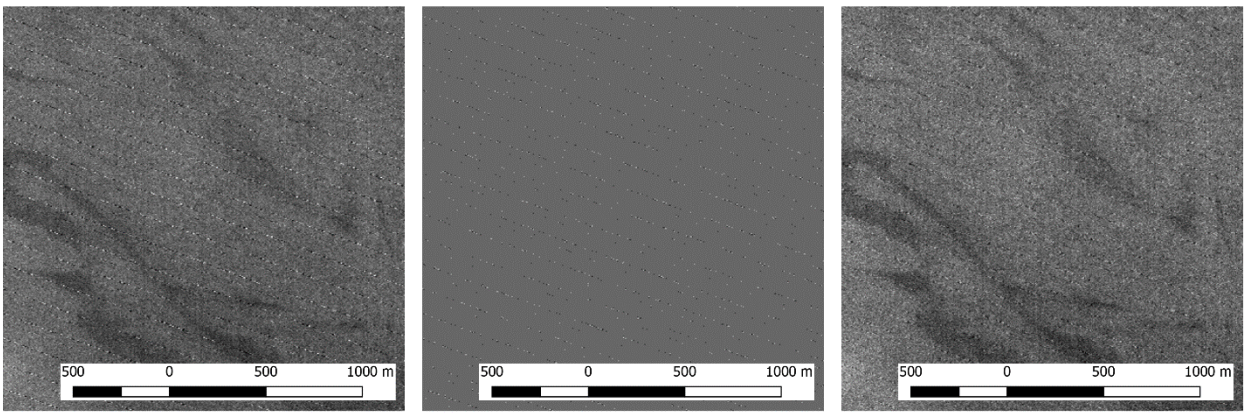
PRISMA HS imagery includes 234 bands (400-2500 nm) at a spatial resolution of 30 m. Additionally, PRISMA PAN imagery (400-700 nm) is provided at a spatial resolution of 5 m. The PAN data is co-registered with the HS data to permit the testing of image fusion techniques. For the study needs, both level 1 (L1) and level 2d (L2D) PRISMA products were analyzed. Because atmospheric correction over water areas affects image radiometry, pansharpening was decided to be carried out using the L1 products. Regarding pre-processing, it was decided to avoid applying an atmospheric correction to the HS image to mitigate errors that could arise from any correction scheme. The available L1 products presented a slight misalignment between the HS and the PAN image; thus, fine co-registration between the two datasets was initially carried out. Finally, in the PAN images, a linear periodic noise was observed in water areas where the radiance values measured by the sensor are considerably low compared to land areas.

Elimination of such noise is usually accomplished by Fourier filtering where the image is decomposed into frequency waves by a 2D Fourier transformation and then filtering of specific frequencies (discrete spikes) takes place on the frequency domain of the magnitude. However, in PRISMA images, such spikes were not observed for two reasons: 1) the linear noise presented in the image contains both high- and low-intensity values and 2) the lines are not continuous and present various spacing among them. Moreover, the spatial frequency of the linear pattern is not constant. Thus, a new method was developed. Firstly, a high- pass Gaussian filter was applied on the PAN image, which amplifies the noise and produces an image with gray pixels (zero value) for the non-noisy pixels of the original PAN image (Figure 4.3) and with bright or dark pixels for the noisy pixels. This process highlights pixels that present different values from their neighbors, including pixels that present plastic targets. Then, the linear noise's inclination and the number of the highlighted pixels that lay on lines having such an inclination are calculated.

If the number of the highlighted pixels exceeds a threshold for each line, then the algorithm assigns the mean value of water pixels to the highlighted pixels. This method does not eliminate the linear noise with 100% accuracy; however, the low number of bright residuals slightly affects the plastic detection process. Although around 10% of the noisy pixels remain, their intensity values are closer to those of the water pixels. A variety of pansharpening methods were then applied to the PRISMA data to procure an HS image with better spatial resolution (section 4.2.4). Before pansharpening, bands with low signal-to-noise-ratio were excluded from the data resulting in an HS image with 175 bands. The bands that were removed were in the intervals 1350-1470 nm and 1800-1950 nm. In these spectral regions, water vapor absorbs much of the incident solar radiation.

4.2.3 Pansharpening methods

Pansharpening of HS images is still an open issue. So far, only a few methods have been presented in the literature



(a) (b) (c)

Figure 4.3. Zoomed in water PAN image. (a) Before the noise reduction. (b) High-pass Gaussian result. (c) After the noise reduction

to address it, the majority of which has been developed in order to fuse PAN and MS data. However, with the increasing availability of HS systems, the pansharpening methods have been extended to the fusion of HS and PAN images. The arisen difficulty consists in defining a fusion model that yields good performance in the part of the HS spectral range that is not covered by PAN data, in which the high spatial resolution information is missing [86]. In the last decades, a variety of pansharpening techniques have been developed. Most of them can be roughly classified into five categories: component substitution (CS), multiresolution analysis (MRA), hybrid, Bayesian, and deep learning (DL) methods.

4.2.3.1 Conventional pansharpening methods

In the CS approach, a component of the HS image is substituted with the PAN image. These methods rely upon the higher spectral resolution image's projection into another space to separate spatial and spectral information. Subsequently, the component that contains the spatial information is substituted with the PAN image and the sharpened data are projected back to the original space [86]. PCA is commonly exploited in CS approaches. Other CS methods are the Gram-Schmidt (GS) and the GS Adaptive (GSA) [258]. In the MRA approach spatial details extracted from the PAN image through a multiresolution analysis are injected into the upsampled HS bands. A well-known method in this category is the Smoothing Filter-based Intensity Modulation (SFIM) algorithm [259]. Local Mean Matching (LMM) and Local Mean and Variance Matching (LMVM) filters [260] also belong to the MRA approach. The hybrid approach uses concepts from the CS and MRA-based methods. Since CS methods are known for preserving spatial information but generating spectral distortion, whereas MRA methods preserve the spectral information but may have some spatial blur, hybrid methods have been created to find a balance between spectral and spatial preservation. Such a method is the Guided Filter PCA (GFPCA) [261]. The Bayesian approach utilizes knowledge modeling through an appropriate distribution to solve the probabilistic framework that results in the pan-sharpened HS image. The main idea of the Bayesian methods is to see the PAN image as the spatial degradation of the result we want to restore and the HS image as its spectral degradation. A good modeling knowledge of those degradations is needed to reverse them to restore the fused image. The Naïve Gaussian prior method (BayesNaive) [262] and the HySure [263] method belong in this category.

4.2.3.2 Deep learning pansharpening methods

Recent research in pansharpening involves deep learning approaches based on CNNs. In this study, three CNNs have been applied. The first two followed a supervised approach and were trained using the Keras library [167] (backend: Tensorflow [170]). The third followed an unsupervised approach and was trained using the Pytorch [241] library.

A. PNN

For the first DL method, the pansharpening Neural Network (PNN) proposed by Masi et al. (2016) [87] was applied. This three-layer architecture was originally proposed for the pansharpening of VHR MS satellite images (Figure 4.4). The first convolutional layer computes 64 feature maps using a 9 px × 9 px receptive field (patch size) and the second computes 32 feature maps with a 5 px × 5 px kernel size. ReLU [165] (Equation (2.11)) was used as an activation function in the hidden layers while the Sigmoid function [166] (Equation (2.12)) was used in the output layer with a 5 px × 5 px kernel size. It is noted that the identity function was proposed in the original implementation for the output layer. The backpropagation process was implemented according to the Adam method [164].

The spatial resolution of the input and output of the network was defined according to Wald's protocol. In more detail, for the study needs the network was trained on an input that resulted from concatenating: i) the PAN image (original spatial resolution: 5 m) downsampled to the spatial resolution of the HS image which for PRISMA corresponds to 30 m and ii) the HS image downsampled by the same ratio, i.e. 1/6 to 180 m and then upsampled to its original size. The original HS image was fed to the network as an output. Thus, the trained CNN is expected to approximate the function that upscales a PRISMA HS image by the ratio mentioned above. During the inference stage, the pan-sharpened image (spatial resolution: 5 m) was created by feeding the network with an input that results from concatenating: i) the original PAN image and ii) the original HS image upsampled to the size of the original PAN.

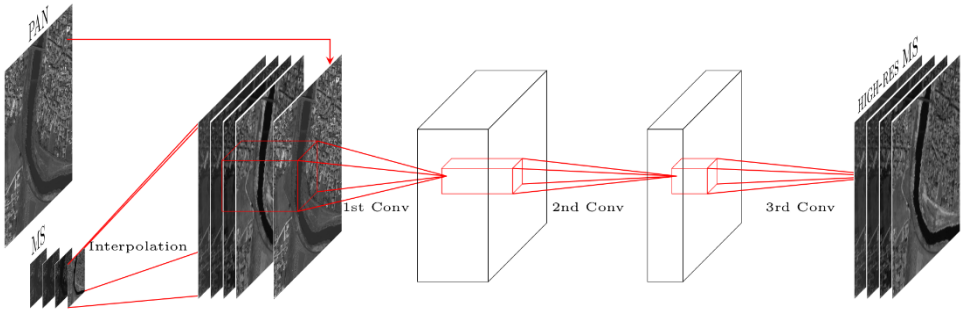


Figure 4.4. CNN architecture proposed by Masi et al. (2016) for pansharpening of VHR multispectral satellite images

The training was performed on the PRISMA image acquired on 18/09/2020 (size: 1000 px × 1000 px) on ~ 60000 patches (size: 9 px × 9 px) and lasted for ~6 hours (160 epochs, batch size:128). For the Pansharpening of any other PRISMA image, a fine-tuning process is required. In this study, the fine-tuning process lasted for ~ 1 hour. Fine-tuning was considered necessary since the radiance values differ between different acquisition dates. It is noted that this CNN, besides being trained on the original bands, was also trained on values produced after clipping 1% of the histogram values for each band (left and right) to prevent lower performance due to sparse extreme values.

B. CAE

The second DL method was based on the convolutional autoencoder (CAE) architecture proposed by Azarang et al. (2019) [264] for the Pansharpening of VHR MS satellite images (Figure 4.5). The autoencoder architecture is composed of an encoder and a symmetric decoder. The encoder consists of three convolutional layers and two pooling layers. The decoder consists of three convolutional layers and two upsampling layers. To enhance performance, for the purpose of the Pansharpening of the PRISMA images, skip connections were added between the encoding and the decoding part. ReLU was used as an activation function in the hidden layers while the Sigmoid function was used in the output layer. The backpropagation process was implemented according to the Adam method.

The spatial resolution of the input and output of the network was defined according to Wald's protocol. For the current study, the network was trained on an input that resulted from downsampling the PAN image from 5 m to 30 m (ratio: 1/6) and then upsampling it to its original size (6000 px × 6000 px). The original PAN image was fed to the network as an output. During the inference stage, the pansharpened image (spatial resolution: 5 m) was created by feeding the network with the upsampled HS bands to the size of the original PAN. The HS bands were fed to the network one by one.

The training was performed on the PAN band of the PRISMA image acquired on 18/09/2020 on ~1.5 million patches (size: 8 px × 8 px) and lasted for ~5 hours (150 epochs, batch size: 128).

C. GDD

For the third DL method, the guided deep decoder (GDD) proposed by Uezato et al. (2020) [265] was applied (Figure 4.6). GDD is composed of an encoder-decoder network with skip connections and a deep decoder network. The encoder-decoder network is similar to the architecture of U-net [204] and produces the features of a guidance image at multiple scales. The network introduces an upsampling refinement unit (URU) and a feature refinement unit (FRU) to promote similar spatial locality and semantic alignment with the features of the guidance image. The proposed loss function is presented below (Equation (4.1))

$$L = \mu\left\|\widetilde{\mathbf{X}_\mathbf{S}} - \widetilde{\mathbf{Y}}\right\|_F^2 + \left|\mathbf{D}\nabla\widetilde{\mathbf{X}} - \nabla\widetilde{\mathbf{G}}\right| \quad (4.1)$$

where $\widetilde{\mathbf{X}}$ is the output pansharpened image, $\widetilde{\mathbf{Y}}$ is the HS input image, $\widetilde{\mathbf{G}}$ is the PAN input image expanded to the same number of bands of $\widetilde{\mathbf{X}}$, $\nabla\widetilde{\mathbf{X}}$ is the image gradient of $\widetilde{\mathbf{X}}$, $\nabla\widetilde{\mathbf{G}}$ is the image gradient of $\widetilde{\mathbf{G}}$, $\widetilde{\mathbf{X}_\mathbf{S}}$ is the spatially downsampled $\widetilde{\mathbf{X}}$, $\mathbf{D}$ is the diagonal matrix to weight each channel of $\nabla\widetilde{\mathbf{X}}$ so that the magnitude of $\widetilde{\mathbf{X}}$ is scaled to that of $\nabla\widetilde{\mathbf{G}}$, μ is a scalar controlling the balance between the two terms, $\|\cdot\|_F$ is the Frobenius norm, and $|\cdot|$ is the l_1 norm.

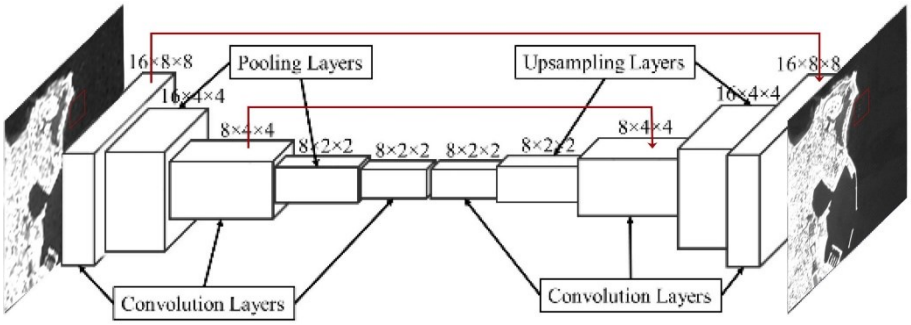


Figure 4.5. CNN architecture proposed by Azarang et al. (2019) for pansharpening of VHR multispectral satellite images. The red arrows show skip connections

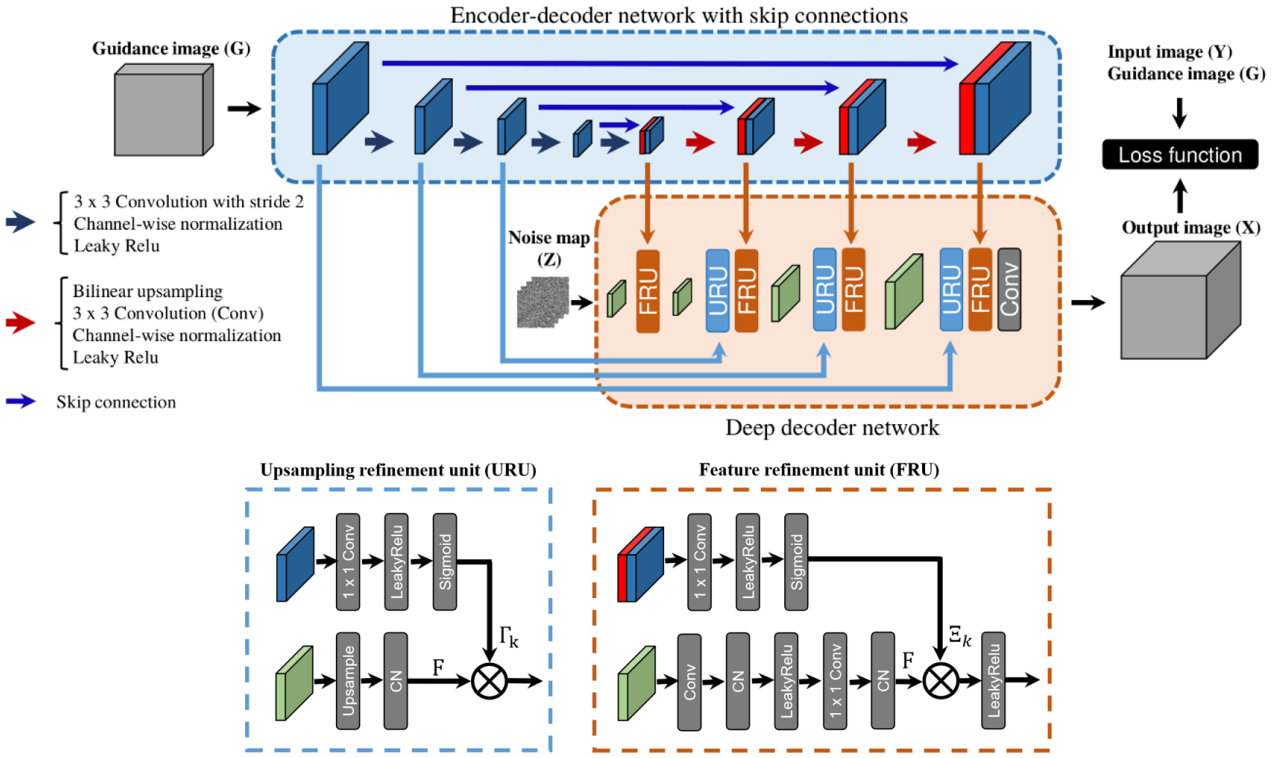


Figure 4.6. CNN architecture proposed by Uezato et al. (2020) for pansharpening of VHR multispectral satellite images. The red arrows show skip connections

This network was tested on the PRISMA image acquired on 18/09/2020. A segment of the PAN band with size 210 px × 200 px represented the guidance image. The input HS segment corresponding to the same region (35 px × 33 px) was fed to the model in seven separate groups because of memory limitation, resulting in seven separate trainings. Each training lasted for ~20 min (6000 iterations). The final pansharpened image was created by concatenating the partial output images. It is noted that the acronyms used for the three DL approaches were acquired from the respective studies.

4.2.4 Pansharpening results and evaluation

Pansharpening methods were initially evaluated for their ability to discriminate the plastic targets from water. In Figure 4.7 and Figure 4.8 the spectral signatures of the various water samples (the same for every image) are shown for the original HS image and each pan-sharpened result in blue color. The various plastic materials are shown in different colors although their identification is not of interest at this point. Water vapor absorption at (720, 820, 940, and 1120) nm and molecular oxygen absorption at 760 - 770 nm are easily observed in all the signatures. The plastic targets and random spectral signatures of water pixels show that the target and water

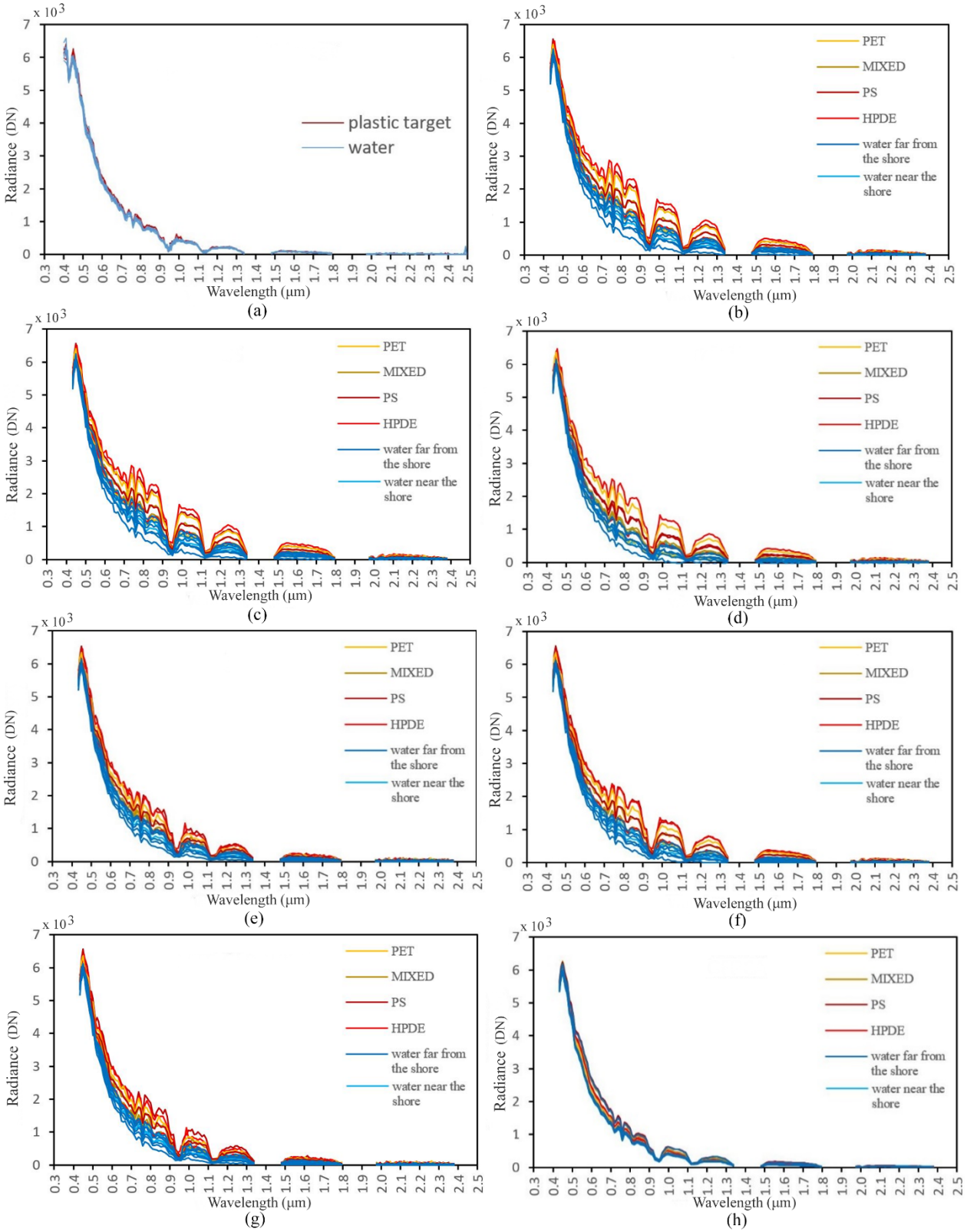


Figure 4.7. Spectral signatures of water and plastic targets. (a) Original HS image (30 m spatial resolution). (b) PCA. (c) GS. (d) GSA. (e) SFIM. (f) MTF-GLP. (g) MTF-GLP-HPM. (h) GFPCA

spectra have a similar shape but the spectra of the targets present higher radiance values, except for a few water spectra corresponding to water pixels near the shore. In Table 4.1 similarity measurements ((SAD) and the

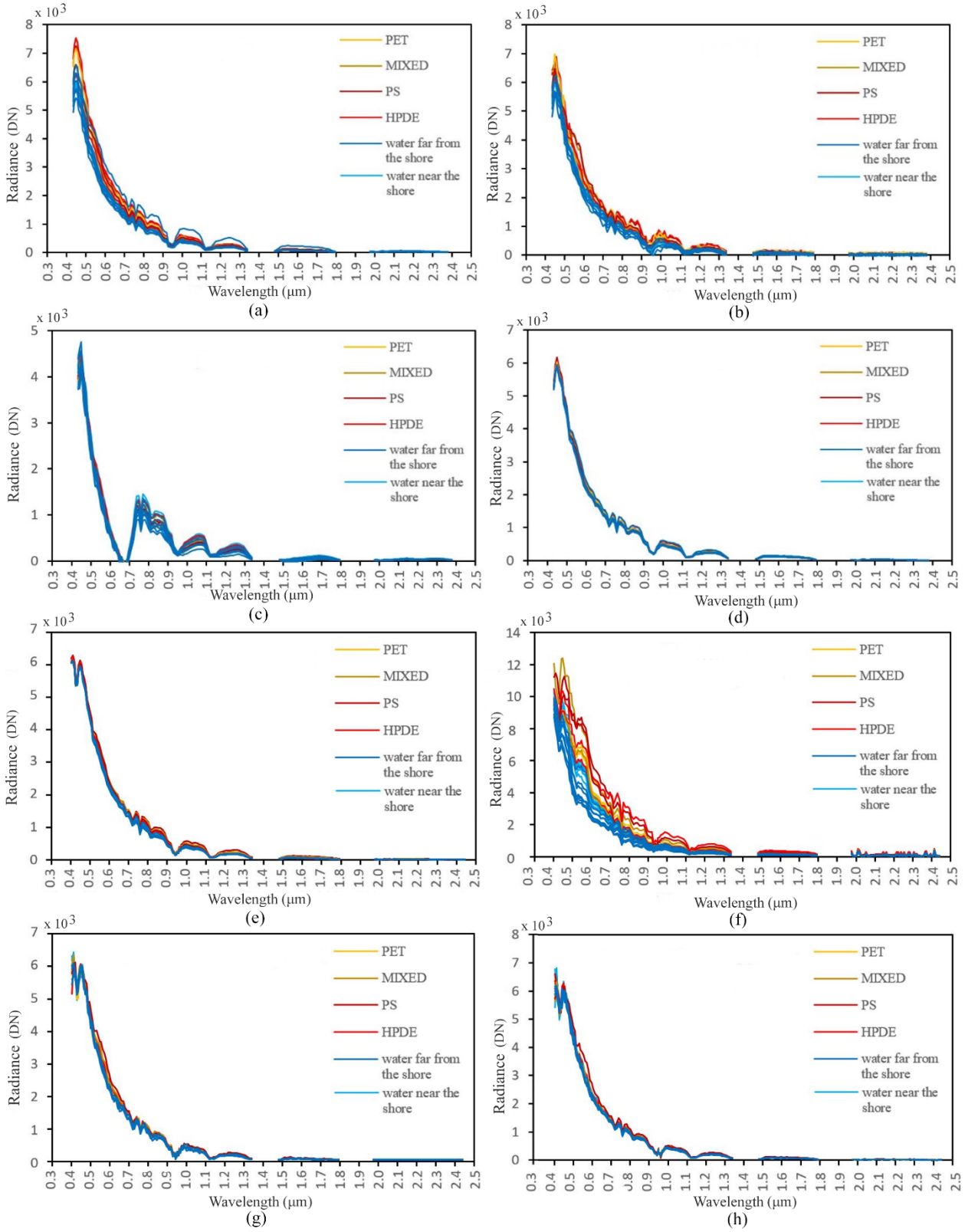


Figure 4.8. Spectral signatures of water and plastic targets. (a) LMM. (b) LMVM. (c) BayesNaive. (d) HySure. (e) PNN. (f) PNN-histogram clipping. (g) CAE. (h) GDD

correlation coefficient (CC)) between water and plastic target spectra are shown. The min, max, and, mean values of SAD and CC measurements between plastic and water spectral signatures are indicated for the original image

Table 4.1. Similarity measurements between water pixels and plastic target spectra

	SAD (rad)			CC		
	min	max	mean	min	max	mean
Original HS image	0.026690	0.052040	0.039209	0.998348	0.999545	0.999036
PCA	0.022155	0.361378	**0.137439**	**0.942857**	0.999646	**0.988574**
GS	0.022306	0.358058	0.136646	0.944418	0.999643	0.988811
GSA	**0.026358**	0.372364	0.131318	0.945684	0.999558	0.990175
SFIM	0.026099	0.245734	0.096315	0.974023	0.999536	0.994593
MTF-GLP	0.024633	0.327536	0.134929	0.956624	0.999588	0.990092
MTF-GLP-HPM	0.026200	0.262613	0.101668	0.970225	0.999531	0.994001
LMM	0.005229	0.041735	0.019314	0.999155	0.999982	0.999750
LMVM	0.010859	0.106932	0.048492	0.996401	0.999929	0.998860
GFPCA	0.003576	0.088835	0.031062	0.996457	0.999991	0.999300
BayesNaive	0.018312	0.130884	0.060868	0.993126	0.999778	0.997962
HySure	0.001423	0.043559	0.014458	0.998976	0.999999	0.999847
PNN	0.005967	0.060582	0.030452	0.998096	0.999982	0.999426
PNN-histogram clipping	0.019818	0.300965	**0.139743**	**0.940634**	0.999745	**0.984474**
CAE	0.020642	0.087132	0.042123	0.995322	0.999715	0.998774
GDD	0.023663	0.081706	0.045095	0.995933	0.999643	0.998632

and the pansharpened results. It is observed that: a) in the original image, water and plastic signatures are significantly correlated and b) in all the pansharpening methods, signatures present low SAD values and high CC values. The latter demonstrates the spectra similarity between water and plastics. Pansharpening methods which exhibit the highest mean SAD values and the lowest mean CC values are the most appropriate for marine plastic discrimination.

Based on Table 4.1 as well as Figure 4.7 and Figure 4.8, it is concluded that the component substitution methods such as PCA, GS, and GSA yield the best results. Plastic spectra present quite higher radiance values than water, while similarity values between water and plastic targets are the smallest. Three MRA methods, SFIM, as well as the (Modulation Transfer Function – Generalized Laplacian Pyramid) MTF-GLP and the MTF-GLP-HPM (High Pass Modulation) [266] also present satisfactory results, whereas Hybrid and Bayesian methods did not achieve sufficient discrimination between plastic target and water spectra. The spectra derived by the Bayesian methods have different shapes compared to the respective original spectra, generating significant spectral distortions. As far as the deep learning methods are concerned, only the PNN trained on values produced after histogram clipping showed a good separation of the random water spectra from the target spectra.

In terms of spatial distortions, only PCA and GS methods produce clear edge results. The results of MRA and hybrid methods seem blurry and duplicate edges are observed along the shoreline and port piers. These drawbacks are caused by the high pass detail injection and may be emphasized by misregistration between HS and PAN data. Bayesian methods produce blurry results with a noise pattern and DL methods present pixelated/blurry outputs.

The less satisfactory results provided by the DL methods can be explained mainly by the large difference between the spatial resolutions of the PAN and the HS bands. Objects depicted in 5m spatial resolution images present much more spatial information (e.g. visible edges) in comparison to what is depicted on a 30 m resolution image. Thus, the problem is much more challenging than e.g. recreating 0.5 m spatial resolution from 2 m, which is the usual case in the majority of the DL pansharpening studies encountered in the scientific literature. Other reasons are the unavailability of HS ground-truth data with 5 m spatial resolution during training and the fact that there is no spectral overlap between the PAN band and the NIR-SWIR bands. It is noted that in the 2022 WHISPERS PRISMA pansharpening challenge [267] (performed later in time than our study), none of the DL competitors managed to outperform the conventional base-line methods. Figure 4.9 shows the outputs for four

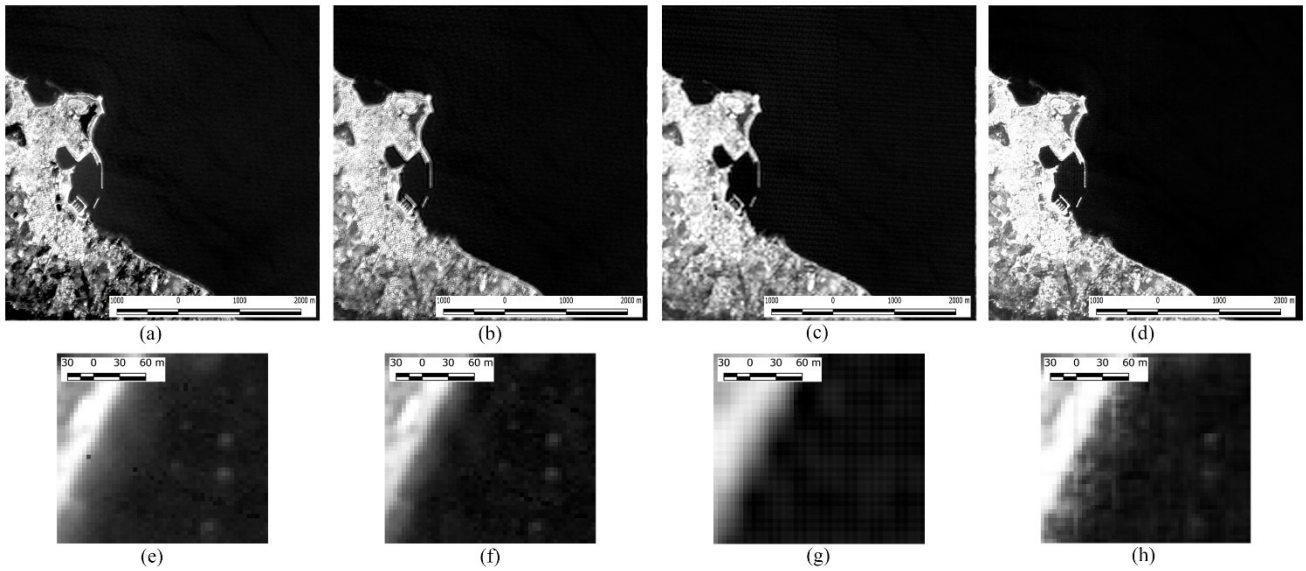


Figure 4.9. Pansharpening results for the PRISMA image acquired on 18/9/2020 (zoomed out and zoomed in view) (670 nm). (a, e) PCA. (b, f) SFIM. (c, g) BayesNaive. (d, h) PNN-histogram clipping

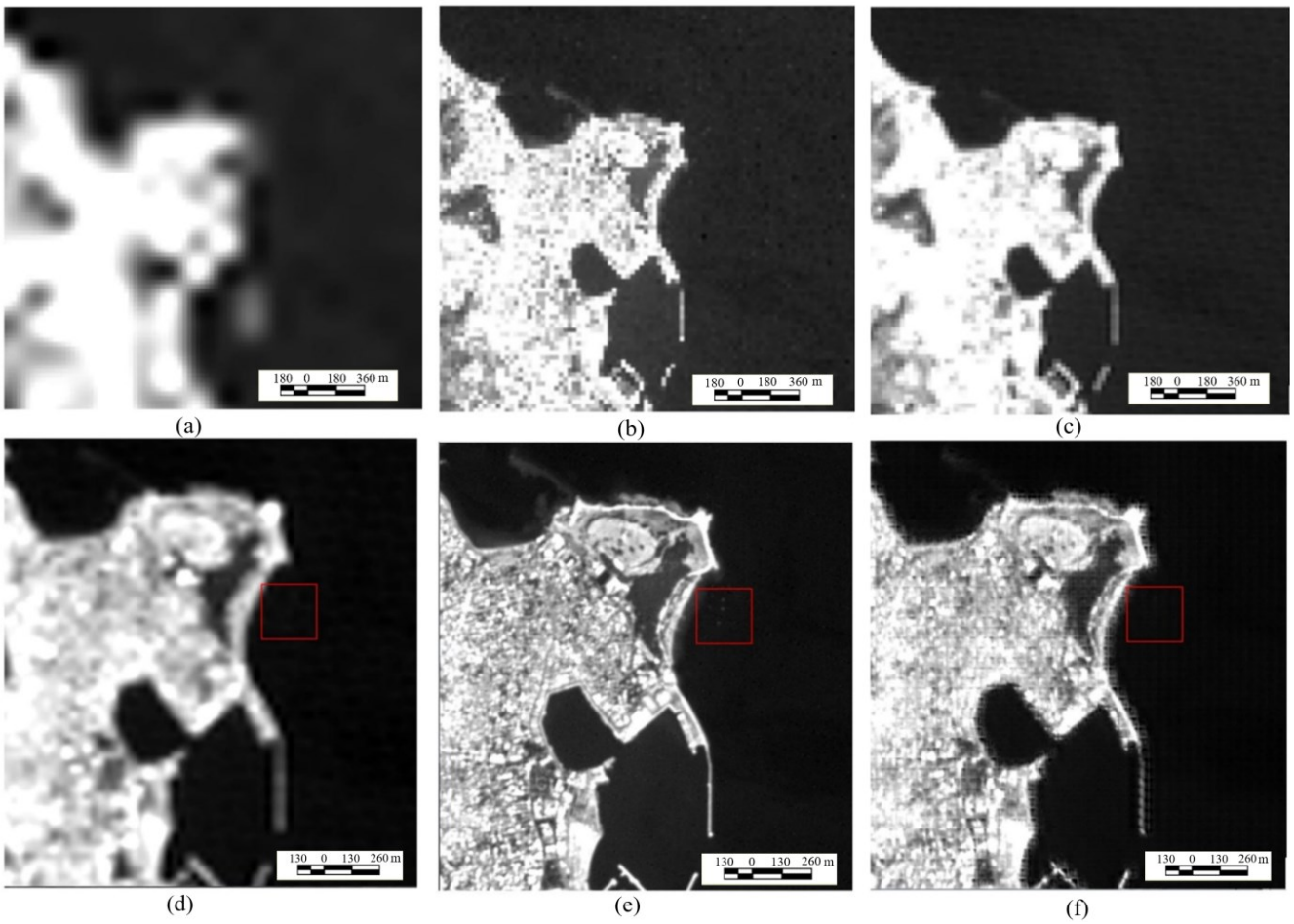


Figure 4.10. 1st row: Segments of the training images of the 1st DL approach: a) HS image used as input, b) PAN image used as input, c) HS predicted image (PNN-histogram_clipping). 2nd row: Segments of the inference images of the 1st DL approach: d) HS image used as input, e) PAN image used as input, f) Pansharpened HS output image (PNN-histogram_clipping/restored to original range of values). The red box contains the area of the experiment (18/09/2020)

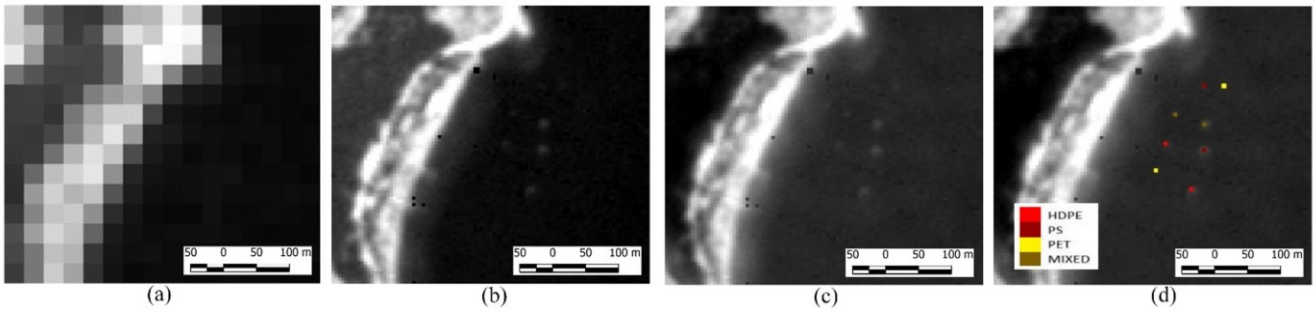


Figure 4.11. Pansharpening of the PRISMA image acquired on 18/9/2020. (a) Original HS image (670nm). (b) PAN image. (c) Pansharpened PCA image (670nm). (d) Pansharpened PCA image (670nm) with plastic target marks

pansharpening methods (PCA, SFIM, BayesNaive, PNN-histogram clipping) for the image acquired on 18/9/2020. Figure 4.10 presents the input and output images during training and inference for the 1st DL method.

Since PCA is the simplest method, it could be selected as the most efficient method for our study. In Figure 4.11, the PCA results of the image acquired on 18/9/2020 are shown along with the PAN band. The targets are highlighted in color. It is observed that all the medium and large-sized targets except for those containing PET material, are easily discriminated in the pansharpened image.

4.2.5 Plastic litter indexes

Marine litter indexes are simple mathematical formulas that rely on discriminative features for detecting marine plastics. The water abundance within the pixel coverage of the HS image as well as the contribution of the neighbor pixels into the radiance registered at the sensor, smooths the discriminative features of the plastic material while similarities to water spectra with crests and troughs in the same wavelengths. However, for plastics, crests present higher radiance values than for water due to the injection of the panchromatic image in the 30 m resolution HS image.

Thus, an intersection of the outputs of three indexes (Equations (4.2), (4.3), (4.4)) was proposed in this study to discriminate plastic targets from water based on radiance differences between spectrum crests and troughs in the VNIR region, since water absorption in the SWIR bands significantly affects the spectra of the plastic objects in the sea. (Index_1: R_i:781 nm, R_j:951 nm, Index_2: R_i:596 nm, R_j:719 nm, Index_3: R_i:492 nm, R_j:719 nm). More details can be found in [268]. In each index output, a threshold is set, enabling the creation of a simplified detection and quantification algorithm. Indicative threshold values are: a) for the first two index images: [mean value of the image] + 2.20 × [standard deviation of the image] and b) for the third index image: [mean value of the image] - 0.60 × [standard deviation of the image].

$$\text{Index}_1 = {R_i}^2 - R_j \quad (4.2) \qquad \text{Index}_2 = {R_i}^2 - {R_j}^2 \quad (4.3) \qquad \text{Index}_3 = R_i - R_j \quad (4.4)$$

In Figure 4.12, can be seen that the plastic targets cannot be detected solely by the panchromatic PRISMA image as they are confused with other materials in the seabed or on the sea surface. It is also shown that even though the target pixels show high concentrations of suspended matter and chlorophyll, using the intersection of the proposed indexes the detection of the plastic targets is quite accurate with only a few pixels (most of them very close to the coast) being erroneously indicated as plastic materials. It should be noted that only a few remaining bright non-plastic pixels are presented in the deeper water areas. These are mainly related to the remaining noise after the filtering of the PAN image.

4.2.6 Conclusions

In this study, an evaluation of the PRISMA imagery potential for marine plastic litter detection was carried out for the first time. To our knowledge, it is also the first attempt to investigate this problem via satellite HS imagery. The study focuses on the detection of small-sized targets (≤ 5 m) which is even more challenging. To this end, the required pre-processing steps, such as fine co-registration of PAN and HS images and elimination of the observed

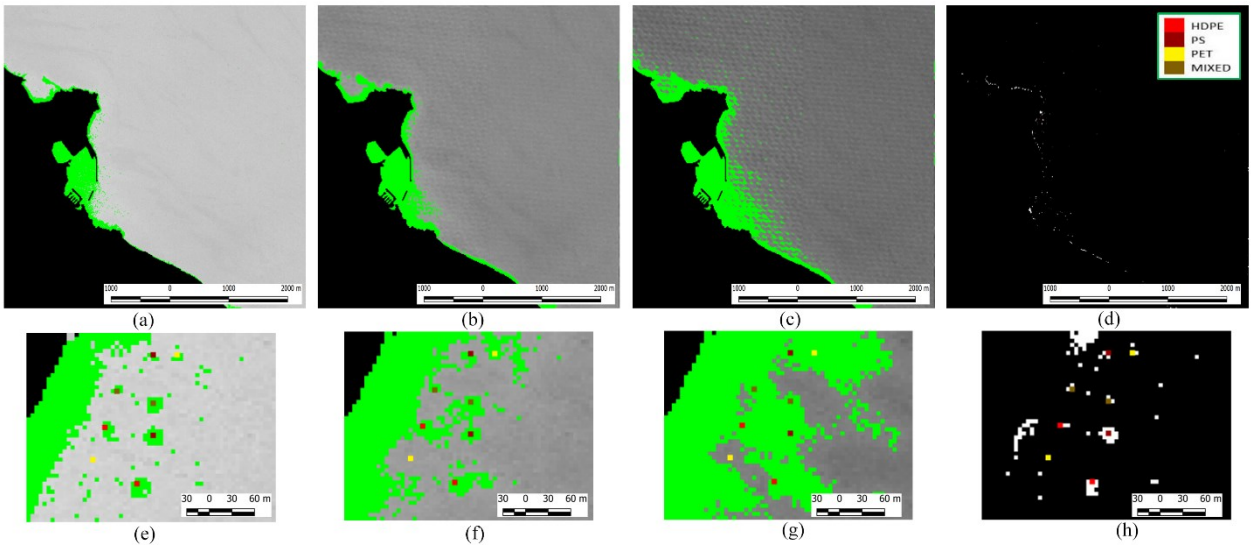


Figure 4.12. Zoomed out and zoomed in view of the area of the experiment of the PRISMA data collected on 18/09/2020. (a, e) Panchromatic image. (b, f) Total suspended matter (TSM) map. (c, g) Chl-a concentration map. (d, h) The intersection of the proposed indexes. Plastic targets are highlighted with colors in images e-h. For the PAN, TSM, and Chl-a images, green color defines the values found in the range of the target values. The land has been masked out

noise in the PAN image have been defined. A new algorithm has been developed to eliminate the periodic noise that is observed in the PRISMA PAN images. Several pansharpening approaches (conventional and deep learning (DL)) have been applied and evaluated for their ability to spectrally discriminate plastics from water as well as for their spatial distortions. Among them, the CS methods yielded the best results. Especially, the simple PCA-based substitution efficiently separates plastic spectra from water without producing blurry and duplicate edges or pixelation in the produced image. In the DL methods, spatial distortions are observed caused by the large difference between the spatial resolutions of the PAN and the HS bands and the unavailability of ground-truth data. However, the importance of histogram clipping as a pre-processing step was established.

In the PCA pansharpened image, plastic targets with sizes 5.1×5.1 m^2 and 2.4×2.4 m^2 are easily detected, while targets with size 0.6×0.6 m^2 cannot be detected. The size of the detectable targets corresponds to 8% pixel coverage of the original HS image. However, it would be interesting to conduct further experiments to see which is the minimum size of the target (or minimum coverage of the PRISMA HS pixel) to allow the acquisition of distinguishable plastic spectral features. This minimum would be important in the context of discrimination versus other non-plastic floating materials.

Among plastic materials, transparent and green PET polymer is the most difficult to detect. Discriminating transparent and green PET polymer is even challenging for targets with 5.1×5.1 m^2 size. In contrast, HDPE and PS polymers as well as the mixed composition of the three materials can be easily detected. Spectra of all plastic materials derived by the pansharpened images present similarities with water spectra. The water abundance within the pixel coverage of the HS image and the contribution of the neighbor pixels into the radiance registered at the sensor, smooth the discriminative features of the plastic material, particularly in the SWIR region where water absorption is very high. Pansharpening injects spatial information from the PAN image into the HS image. However, it cannot enhance the absorption features of the plastic materials. The influence of seawater on ocean plastic spectra is preserved and consequently features observed in the laboratory and airborne-based spectra [269] are not apparent in the derived spectral signatures. However, some spectral characteristics observed in the VNIR region can be exploited for producing marine plastic indexes. These characteristics rely on the magnitude of the radiance differences between crests and troughs along the VNIR region of the spectra that plastic materials present.

The next step is to compare the results with other non-plastic floating materials (e.g. floating vegetation and foam), in view to demonstrate that PRISMA could be used as a stand-alone satellite to detect the likelihood of plastic presence. Targets of vegetation might also be used in future experiments to examine if they are distinguishable from plastic targets.

4.3 Increasing the Sentinel-2 (S2) potential for marine plastic litter monitoring through image fusion [10] [11] [12]

In section 4.3 the study concerning the fusion of S2 and WV-3 data for marine plastic litter monitoring is presented. In section 4.3.1 the motivations and objectives of the study are stated and the contributions of the author of this PhD thesis are clarified. In section 4.3.2 at first a data description is provided and then the background and experimental approach of the data fusion methods are presented. In section 4.3.3 the results are evaluated and in section 4.3.4 the conclusions and contributions are summarized and future work is suggested.

4.3.1 Introduction

Spatial resolution, MS characteristics (namely number, position, and width of the acquisition bands), and SNR are crucial factors in the design of a sensor dedicated to the detection and discrimination of accumulations of marine litter from space. Current orbiting sensors were not designed for such an application. As summarized in section 4.1, a number of studies have exploited the spatial resolution and MS characteristics of S2 to perform detection and discrimination of marine litter accumulations (and targets), however, it is now clear that higher spatial resolution and a greater number of spectral bands, arranged in a dedicated configuration, would significantly improve the detection of marine litter accumulations from orbit. The possibility of fusing images of different orbiting sensors to improve at least one of these aspects would be consequently beneficial.

Thus, in this study, various state-of-the-art image fusion algorithms are evaluated on S2 and WV-3 (4 m spatial resolution) datasets. The methods make use of component substitution, spectral unmixing, and deep learning (DL). The DL literature networks were adjusted to the fusion problem since they originated from either the pansharpening or the single image super-resolution (SISR) domain. In addition, three DL networks were created for the purpose of the study. Finally, experiments with various WV-3 band combinations are conducted in the conventional methods to find the optimal one and various indexes are examined for their capability to detect floating plastic objects on the fused images.

It is noted that the contributions of the author of this PhD in this study refer to the implementation of the deep learning image fusion networks. The conventional image fusion methods, the image pre-processing steps, and the indexes were implemented by M. Kremezi.

4.3.2 Materials and methods

4.3.2.1 Data description

A controlled experiment with artificial plastic targets was conducted. The experiment took place in the Tsamakia beach (Figure 4.13). The location is found in the coastal region of Lesvos Island, Greece, it is protected from any human activities and it provides conditions of both shallow and deep water. The experiment included the processing of WV-3 and S2 data.

A. Field data

In the experiment, three 10×10 m^2 plastic targets were utilized, which had been constructed for the needs of the "Plastic Litter Project 2018" conducted by the Marine Remote Sensing Group of the University of Aegean on June

[10] Kremezi, M., **Kristollari, V.**, Karathanassi, V., Topouzelis, K., Kolokoussis, P., Taggio, N., Aiello, A., Ceriola, G., Barbone, E. and Corradi, P., 2022. Increasing the Sentinel-2 potential for marine plastic litter monitoring through image fusion techniques. Elsevier Marine pollution bulletin, 182, p.113974. doi:10.1016/j.marpolbul.2022.113974

[11] Kremezi, M., **Kristollari, V.**, Karathanassi, V., Kolokoussis P., 2022, September. Enhancing PRISMA and Sentinel 2 Capabilities for Marine Plastic Litter Detection Using Image Fusion Techniques, Spectral Signature Unmixing and Spectral Indexes. In 41st EARSeL Symposium, Cyprus. (abstract + poster + oral presentation) (peer-reviewed)

[12] Aiello, A., Barbone, E., Ceriola, G., Karathanassi, V., Kolokousis, P., Kremezi, M., **Kristollari, V.**, Taggio, N., 2023, October. Unlocking the Potential of Spectral Signature Unmixing and Machine Learning for Detecting Plastic Marine Litter: Insights from the REACT Project. In ESA Remote Sensing of Marine Litter Workshop 2023 (abstract + oral presentation)

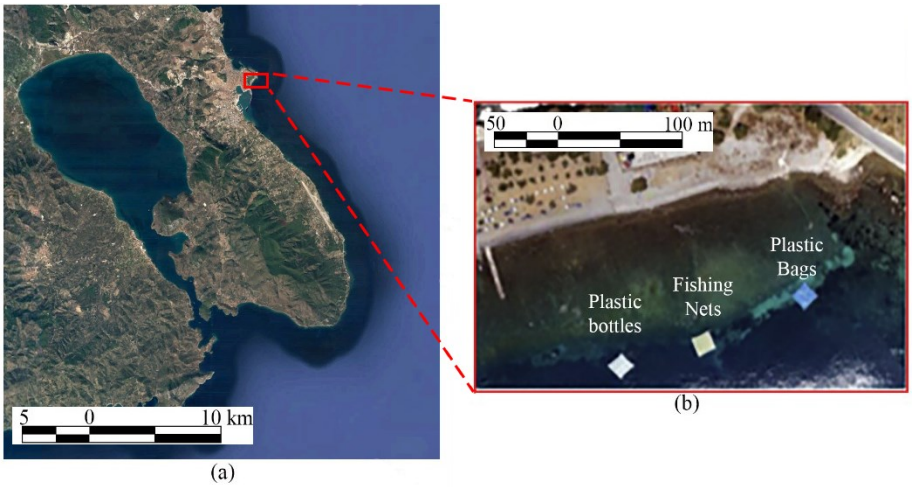


Figure 4.13. (a) Google Earth image with the test area highlighted, (b) UAV photograph of the plastic targets of the experiment (source: Topouzelis et al., 2019)

7th, 2018. Each target contained a different plastic material: PET-1.5 l bottles, LDPE bags, and yellowish nylon fishing nets (Figure 4.13(b)). All the targets are visible in the WV-3 image while only the "plastic bottles" and "fishing nets" are visible in the S2 image (Figure 4.14). All of the constructed targets were anchored in place for both the S2 and WV-3 satellite acquisitions.

B. Satellite data

The image fusion methods were applied to S2 and WV-3 satellite images. As already mentioned in section 2.2.2.1.B S2 carries an optical instrument payload (MSI) that samples 13 spectral bands (442–2200 nm): 4 bands at 10 m, 6 bands at 20 m, and three bands at 60 m spatial resolution. WV-3 provides PAN imagery (450–800 nm) with 0.31 m spatial resolution, 8-band VNIR imagery (400–1040 nm) with 1.24 m resolution, and 8-band SWIR imagery (1195–2365 nm) with 3.7 m resolution. All WV-3 spatial resolutions that have been mentioned above refer to nadir captured imagery and worsen for higher looking angles of the sensor. In this study, we consider S2 images as the images with high spectral resolution and the WV-3 images as the images with high spatial resolution.

C. Pre-processing

The experiment was conducted on June 7th, 2018. An S2 L1C image and a WV-3 image were collected almost synchronously (Figure 4.14). On the date of image acquisition over the test areas, the targets were located offshore and there were clear sky and calm sea conditions. The procured WV-3 image was not atmospherically corrected by the commercial provider but it was converted to Top-Of-Atmosphere (TOA) reflectance to be comparable with the S2 image. This task was carried out by using the instructions from the product provider [270].

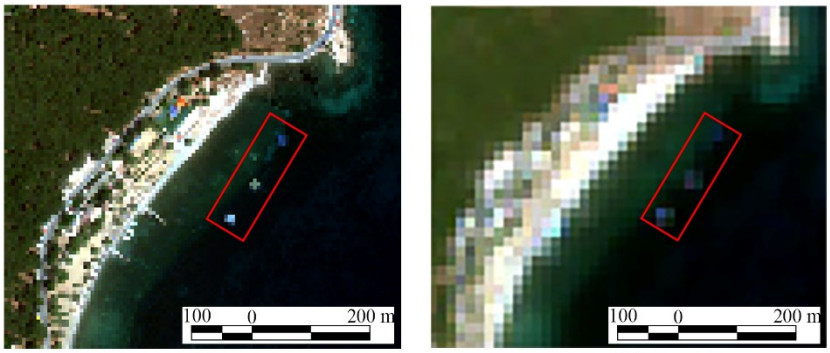


Figure 4.14. (a) Natural colors RGB composite of WV-3 image with 4 m spatial resolution and (b) S2 image with 10 m spatial resolution acquired on 07/06/2018. Red rectangles indicate the area where square plastic targets are located.

Atmospheric corrections were not applied because of a) the clear sky conditions during the experiments, b) concerns expressed in the literature about the possible reduction effects on the spectral signatures [82] [271] which could weaken the plastic signal to undetectable levels, c) the differences in the output between various algorithms [84], and d) the low sensitivity of band subtraction indexes to environmental conditions [272] [273]. In addition, mitigation strategies were not applied for the effects of the different observational geometries (BRDF) since issues that could affect the performance of the fusion process were not observed (e.g. light anisotropies, sunglint) [274].

Fusion approaches utilized all 13 bands of the S2 image, resampled at 20 m spatial resolution. This sampling size was chosen because almost all the bands in NIR and SWIR have 20 m resolution. The WV-3 image was resampled at 4 m spatial resolution. For the DL methods 8 VNIR + 2 SWIR WV-3 bands were used, and for the non-DL, 7 VNIR + 2 SWIR (coinciding WV3 and S2 bands). The georeference of all datasets was checked to ensure alignment between the S2 and the WV-3 image. Further co-registration steps were not considered necessary.

4.3.2.2 Fusion methods

A. Related work

1. Conventional image fusion approaches

Image fusion is the optimal solution to the technological limitations of the spatial and spectral resolutions of a satellite sensor. Image fusion techniques evolved from MS pansharpening [MS + PAN fusion] to MS + MS and MS + HS image fusion [86]. The first approach was developed by adapting various pansharpening/hyper-sharpening techniques to the HS + MS/MS + MS image fusion problem (Component Substitution, Multiresolution Analysis, Sparse representation) [275]. Another more popular category of such methods are algorithms based on spectral unmixing. They exploit all the bands of the common spectral region of the initial images [276][277] [278] [263] and they are ideally used on HS datasets since the high spectral resolution is necessary for decomposing the mixed pixels; however, they have also been proven useful for MS datasets. The spectral unmixing approach has been used in the HS + MS/ MS + MS fusion problem for the last two decades [263][279][280][281][282][283].

2. Single image super-resolution (SISR) with deep learning (DL)

In recent years, several SISR DL methods have been proposed to spatially super-resolve S2 images. In their vast majority, current research has proposed SISR methods based on GAN [284] and ResNet [243] architectures. Although our study uses two sources of data, SISR DL networks can be easily adjusted to fit the fusion problem.

Several methods aim at producing 10 m S2 bands. In [88] two CNNs were trained to spatially super-resolve the S2 20 m and 60 m bands respectively to 10 m. The network was inspired by EDSR [285] which follows the ResNet architecture [243]. A residual design was also used in [89] with the same goal. To spatially super-resolve the 60 m bands their model made use of the 20 m bands in addition to the 10 m bands. In [286] a network was proposed that adds channel attention on residual blocks and increased the resolution of S2 20 m bands to 10 m. In [287] based on PanNet which uses high-pass filtered inputs [288] and a PNN [87] version that uses a residual skip connection at the end of the network [289] S2 20 m bands were super-resolved to 10 m. In [290] a parallel residual network (SPRNet) was proposed and S2 60 m and 20 m bands were super-resolved to 10 m. In [291] a GAN was proposed that uses residual blocks to enhance the spatial resolution of 20 m (×2) and 60 m (×6) S2 bands by injecting information from the 10 m bands. Finally, in [292] a CNN model was proposed that takes the 10 m, 20 m, and 60 m S2 bands as input and produces super-resolved 10 m images for the 20 m and 60 m bands. The model uses Residual-in- Residual Dense Blocks (RRDBs) [293].

Other methods aim at producing higher than 10 m resolutions. In [294] degradation kernel estimation and noise injection were used to construct a dataset of near-natural LR-HR S2 images and then the authors trained a GAN which was composed of an Enhanced Super-resolution (ESR)-GAN-type generator [293], a PatchGAN-type discriminator and a VGG-19-type feature extractor [201]. Their model produces S2 RGB images with 2.5 m spatial resolution by taking as input the respective 10 m images. In addition, by using two sources of data, in [295] a

model based on ESRGAN was trained and 2 m RGB-NIR S2 images were produced. The network was first pre-trained with artificially generated WV LR-HR (10 m - 2 m) image pairs and then fine-tuned with S2-WV image pairs.

B. Image fusion implementation

1. Conventional approaches

Concerning the conventional image fusion approaches that were adapted from pansharpening techniques, in this study, it was decided to implement the PCA method as it outperformed others in the previous study described in section 4.2. Regarding the unmixing-based methods, the Coupled Nonnegative Matrix Factorization (CNMF) [280] and Lanaras' (alternating unmixing approach) [296] as well as the HySure [263] and the FUSE [283] methods (Bayesian approach) have been selected as they have been proven accurate, reliable, and versatile in terms of their adaptiveness for fusing PAN, MS, and HS data [297]. The conventional methods were applied using various WV band combinations (Table 4.2) to determine the optimal number of bands and spectral range for downscaling S2 images.

2. DL approaches

Six DL approaches in total were implemented, among which three literature networks (PNN [87], SRGAN (Super-resolution GAN) [298], RCAN (Residual Channel Attention Network) [299]) and three networks that were created for the purpose of the study (PNN-Siamese, Fusion-ResNet, Fusion-GAN). PNN-Siamese, Fusion-ResNet, and Fusion-GAN were designed based on popular DL concepts (parallel branches, residual blocks, adversarial learning, concatenation). Using a low number of trainable parameters, thus producing lightweight networks was also a key requirement. In more detail: a) PNN-Siamese is a siamese version of PNN, b) Fusion-ResNet is constructed by combining residual blocks and layer concatenation, which are popular concepts proposed in the ResNet and UNet architectures respectively, and c) Fusion-GAN uses a generative adversarial approach and the generator architecture is based on UNet. Fusion-PNN was trained with two different band configurations: a) 8 VNIR and b) 8 VNIR + 2 SWIR, while the rest of the methods were trained only with the first band configuration. The selected SWIR WV-3 bands correspond to 1640–1680 nm and 2185–2225 nm. The selection was based on the fact that these bands are the spectrally closest to the last two S2 SWIR bands (1613 nm, 2202 nm).

Concerning the DL literature networks, they were adjusted to the fusion problem (modification of input layer) since they originated from either the pansharpening (PNN) or the SISR domain (SRGAN, RCAN). PNN was selected because of its simplicity and SRGAN because of its high popularity and the fact that it is the basis of the rest of the spatial super-resolution methods based on GANs. RCAN was implemented because it combines the concept of “attention” with the residual blocks. The inclusion of attention layers in CNNs has shown promising results in many fields.

It is noted that due to the higher computational demand of the DL approaches, the investigation of the optimal band combination was carried out only on the conventional image fusion techniques.

Table 4.2. WV band combinations

Combination	WV Bands
all9	All 9 (VNIR +SWIR)
all7	All7 (VNIR)
234	Blue – Green – Red
2346	Blue – Green – Red – NIR1
2347	Blue – Green – Red – NIR2
1234	Coastal - Blue – Green – Red
2348	Blue – Green – Red – NIR1 – SWIR3
23469	Blue – Green – Red – NIR1 – SWIR6
23478	Blue – Green – Red – NIR2 – SWIR3
23479	Blue – Green – Red – NIR2 – SWIR6

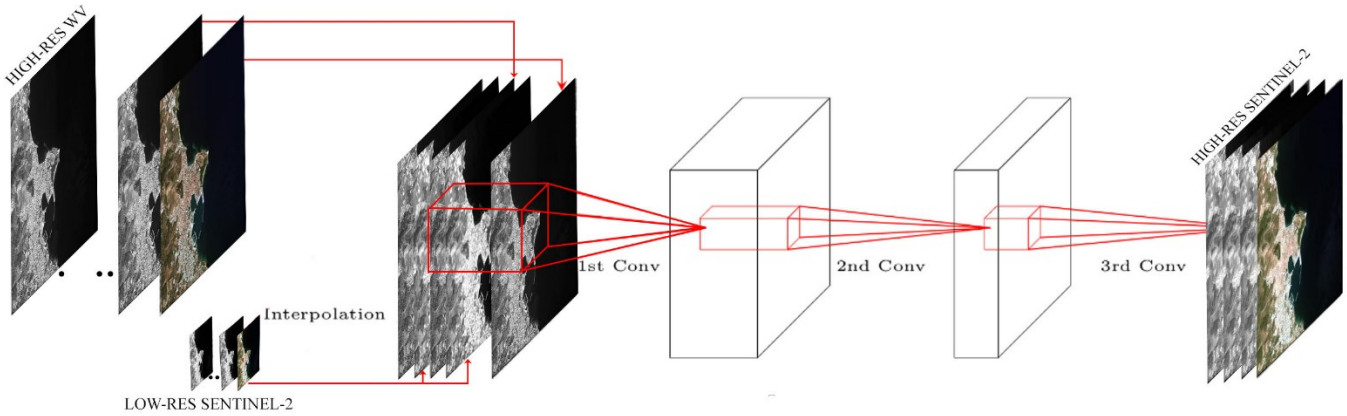


Figure 4.15. Architecture of PNN (Masi et al. (2016)) adjusted to the fusion problem (modification of input layer)

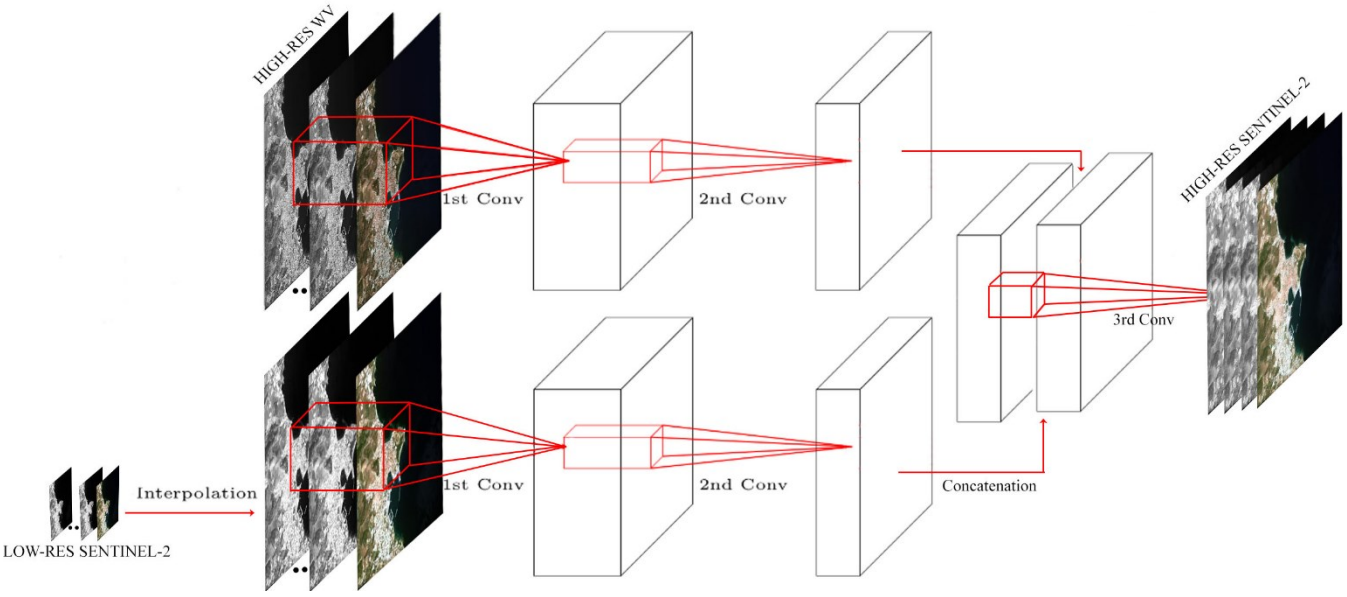


Figure 4.16. Architecture of PNN-Siamese

Architecture – activation functions

All DL approaches except for PNN-Siamese used the Early Fusion (EF) method, i.e. the input of the network was created by concatenating the S2 and WV-3 bands. The sequence of the bands matched the sequence of the corresponding wavelengths. In PNN-Siamese, the S2 and WV-3 bands were fed as input to two different branches. For the creation of the Siamese network, the outputs of the second convolutional layer of the two branches were concatenated and then fed to a third convolutional layer. In both PNN versions (Figure 4.15, Figure 4.16), the suggestions of [87] were followed for the number of feature maps, the activation function, and the size of kernels in the first and second convolutional layers (64/ReLU/9×9, 32/ReLU/5×5), as well as the kernel size of the third convolutional layer (5×5). The Sigmoid activation function [166] (Equation (2.12)) was applied in the output layer. The equation of the ReLU activation function [165] is presented in Equation (2.11).

The architecture of Fusion-ResNet is shown in Figure 4.17. The encoding part of the network was initialized with a convolutional layer and it was also composed of four identity residual blocks and two convolutional residual blocks. The decoding part included two transposed convolutional layers and a convolutional layer towards the end. Skip connections through concatenation were applied between the encoder and the decoder. Further details are shown in Figure 4.17 (e.g. activation-batch normalization layers, convolution hyperparameters, number of feature maps, etc.). It is noted that the input and output images shown in Figure 4.17 correspond to the inference resolution.

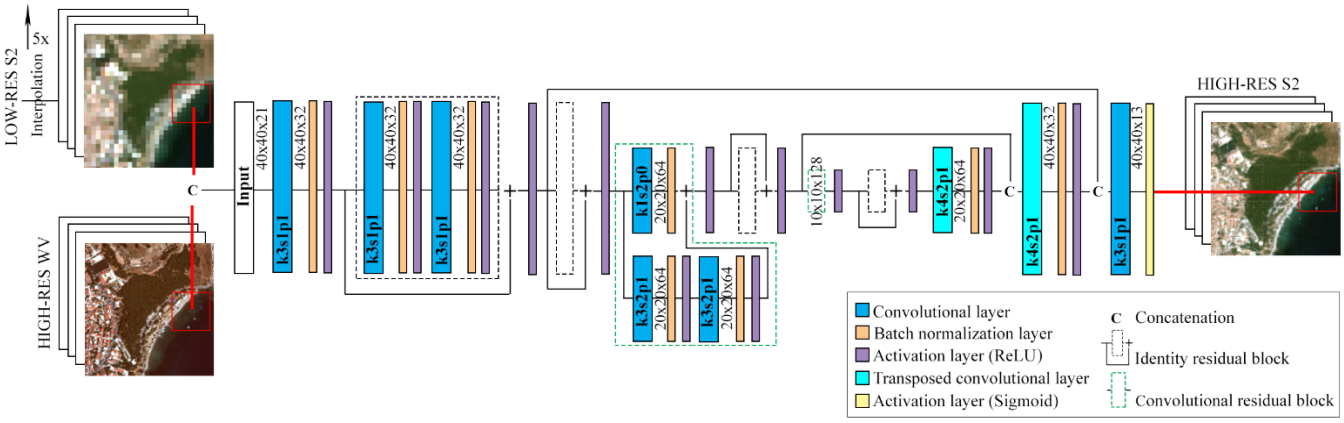


Figure 4.17. Architecture of fusion-ResNet

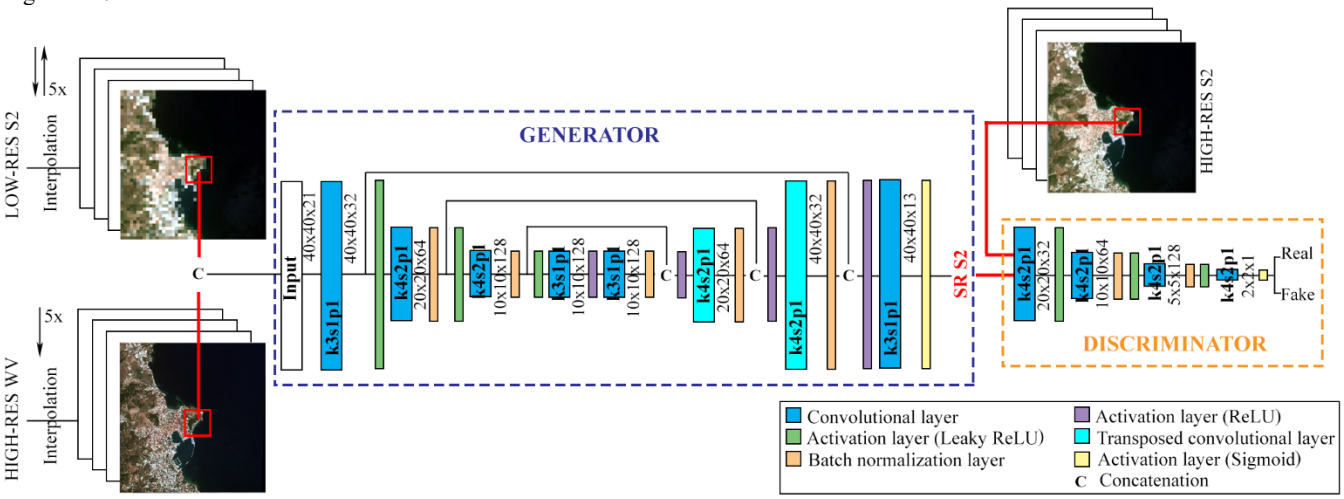


Figure 4.18. Architecture of fusion-GAN

The architecture of Fusion-GAN is shown in Figure 4.18. The network was composed of a generator and a discriminator. The generator encoder consisted of five convolutional layers and the decoder of two transposed convolutional layers and a convolutional layer towards the end. Skip connections through concatenation were applied between the encoder and the decoder. Concerning the discriminator, it consisted of four convolutional layers. Further details are shown in Figure 4.18. It is noted that the input images shown in Figure 4.18 correspond to the training resolution.

SRGAN was implemented with the hyperparameters suggested in [298]. The upsampling layers towards the end of the network were removed because our input x-y size matched the output size. Among others, the generator contained 16 residual blocks and the discriminator eight convolutional layers where the number of feature maps gradually increased (64 → 512). Further details are shown in Figure 4.19 and the respective paper.

Finally, RCAN uses the residual in residual structure and incorporates a channel attention module. The settings proposed in [299] were followed for its implementation with the difference that we selected five residual groups and 10 residual channel attention modules instead of 10 and 20 to ensure faster training time. Further details are shown in Figure 4.20 and the respective paper.

Pre-processing – training - inference

Given that ground-truth data at the high-resolution (4 m S2 image) are not available, the DL approaches were trained based on the assumption that the spatial details are self-similar and scale-invariant as considered in previous works (e.g. [88], [286], [290], [292]). Thus, it was assumed that super-resolving from 20 m to 4 m can be learned from super-resolving at a reduced resolution where ground-truth data are available (Wald's protocol). In more detail, during training the inputs of the CNN were (Figure 4.21): a) the WV-3 bands downsampled to 20 m spatial

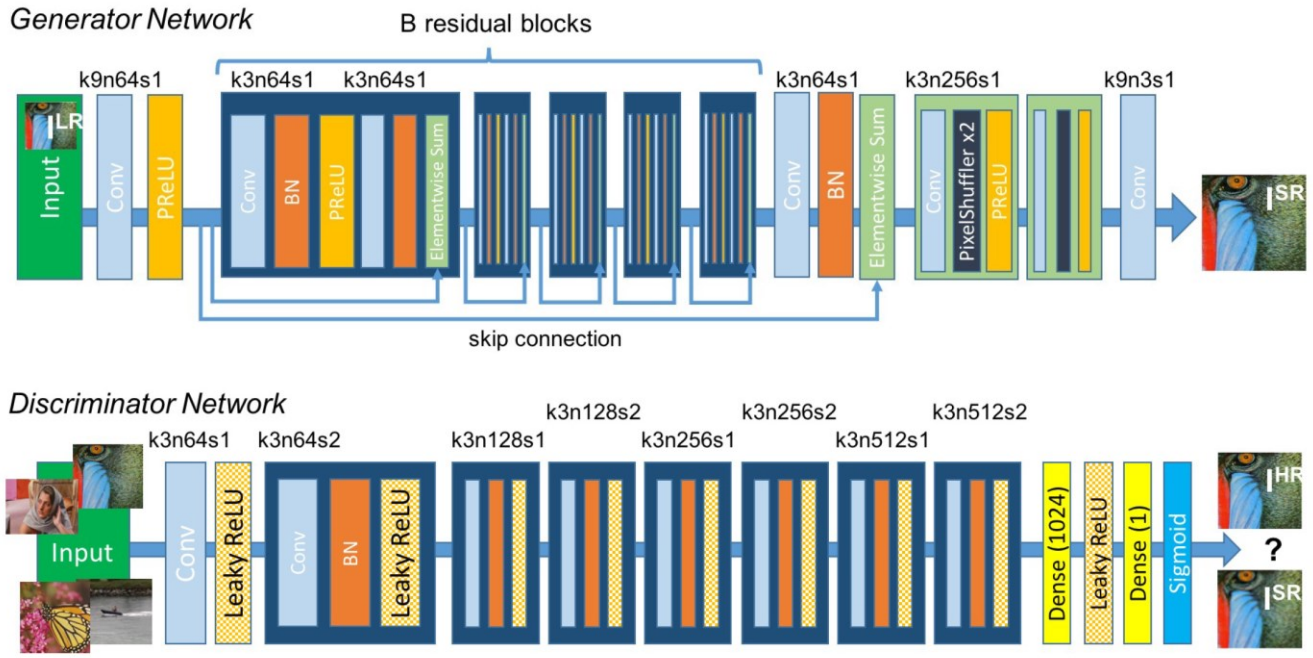


Figure 4.19. Architecture of SRGAN

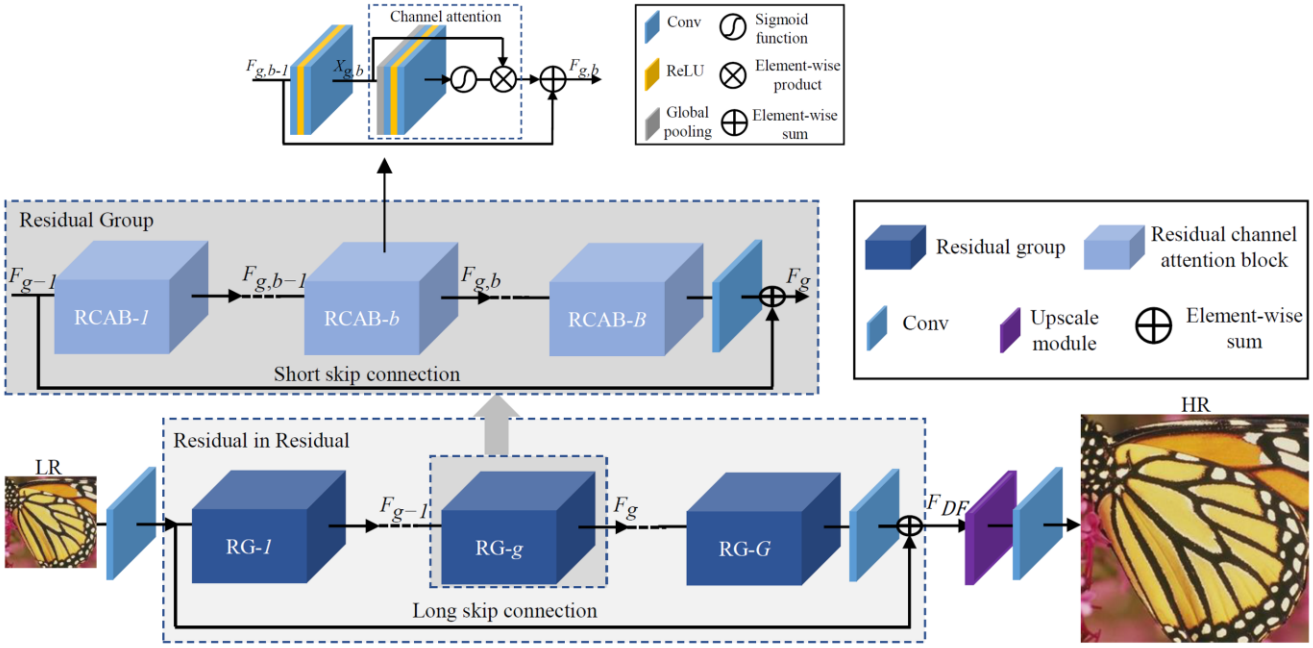


Figure 4.20. Architecture of RCAN

resolution and b) the S2 bands downsampled to 100 m and then upsampled to 20 m. Thus, the spatial resolution ratio between the WV-3 and S2 bands was 1/5. During the inference stage, the fused output image (spatial resolution: 4 m) was created by feeding the network with a) the WV-3 bands with 4 m spatial resolution and b) the S2 bands with 20 m spatial resolution. Nearest neighbor interpolation was used during all resampling operations which led to the same x-y size between the WV-3 and the S2 bands. In addition, for each band, 1% of the histogram values (left and right) were clipped to prevent lower CNN performance due to sparse extreme values.

During training, the Adam method [164] was used to update the weights in the backpropagation process.

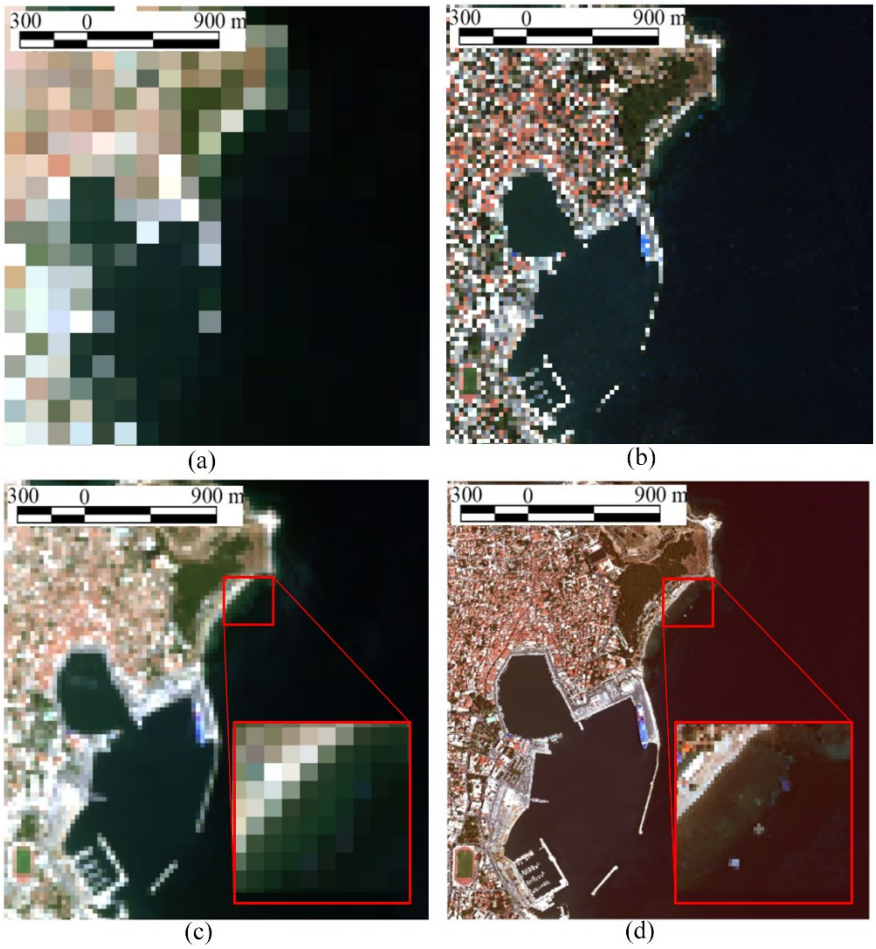


Figure 4.21. (a) S2 image with 100 m spatial resolution, (b) WV-3 image with 20 m spatial resolution, (c) S2 image with 20 m spatial resolution, (d) WV-3 image with 4 m spatial resolution. Collection date: 07/06/2018 (zoomed-in view of the target placement in red window) (natural colors).

Concerning the loss function, MSE was selected as the Loss function for both PNN versions and Fusion-ResNet (Equation (4.5)) and the L1 loss (Equation (4.6)) for RCAN. In the adversarial training (Fusion-GAN, SRGAN), the generator and the discriminator were trained in an alternating way according to the adversarial loss function shown in Equation (4.7) [284]. In addition, since the problem belongs in the super-resolution domain, a content loss was added to the generator loss to significantly increase the performance [298] (Equation (4.8) (Fusion-GAN), Equation (4.9) (SRGAN)). It is noted that the originally proposed VGG loss for SRGAN was substituted with the pixel-wise MSE loss because our output contained 13 bands instead of three, thus the VGG Imagenet [14] pre-trained weights could not be used.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (4.5) \qquad \text{L1} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \quad (4.6)$$

$$min_G max_D \begin{bmatrix} \mathbb{E}_{(HR\sim P(HR))}[\log D(HR)\,] + \\ \mathbb{E}_{(LR\sim P(LR))}[\log\,(1 - D(G(LR)]\end{bmatrix} \quad (4.7)$$

$$L_{G\_Fusion-GAN} = -\log\left(D\big(G(LR)\big)\right) + 10^2 L1 \quad (4.8) \qquad L_{G\_SRGAN} = -10^{-3}\log\left(D\big(G(LR)\big)\right) + MSE \quad (4.9)$$

where Y_i are the observed spectral values, $\hat{Y}_i$ are the predicted spectral values, and n is the number of pixels in a patch.

Table 4.3. Training details

Training details	PNN_VNIR	PNN_VNIR+SWIR	PNN-Siamese	Fusion-ResNet	Fusion-GAN	SRGAN	RCAN
Loss metric	0.0007 (MSE)	0.0006 (MSE)	0.0006 (MSE)	0.0002 (MSE)	0.0101 (L1)	0.0017 (MSE)	0.0014 (L1)
Epochs	4000	4000	4000	800	800	500	500
Batch size	128	128	128	128	128	16	16
Patch size	9x9	9x9	9x9	40x40	40x40	40x40	40x40
Number of patches	48600	48600	48600	6400	6400	6400	6400
Trainable params	170,573	180,941	232,269	906,029	974,542	10,812,831	3,999,957
Library	Keras/TF	Keras/TF	Keras/TF	Pytorch	Pytorch	Pytorch	Pytorch
Training time (h)	2.0	2.2	3.0	1.1	0.9	7.5	10.0

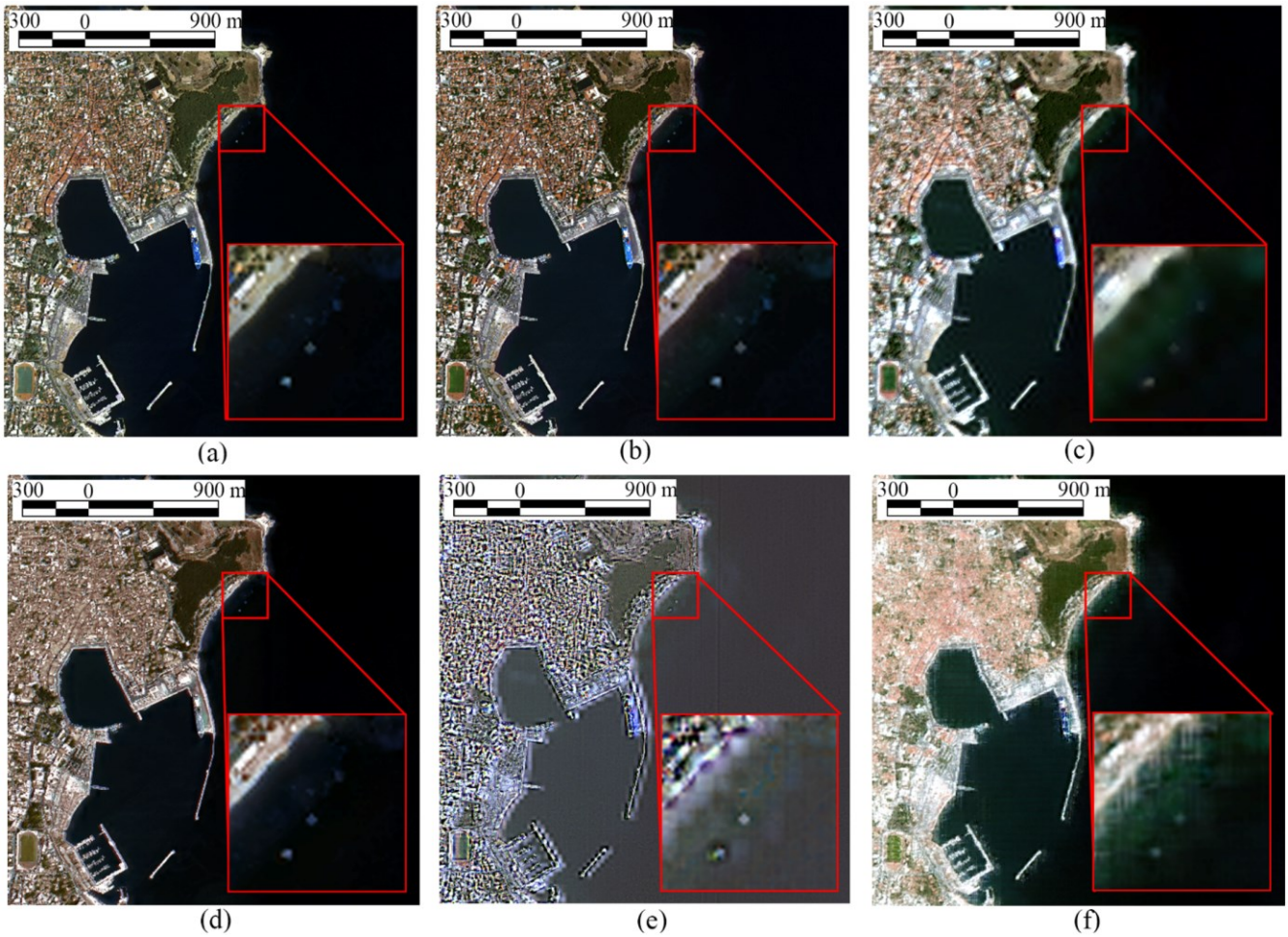


Figure 4.22. Fusion results for the S2 and WV-3 images acquired on 07/06/2018 (zoomed-in view of the target placement in red window) (natural colors). (a) CNMF_2347, b) HySure_all9. (c) PCA_all9. (d) Lanaras'_1234. (e) FUSE_2348. (f) PNN_VNIR

Training details are given in Table 4.3. All models were trained on an NVIDIA 1070 Ti GPU. The x-y size of the images during training was 229 px × 234 px and during inference was 1145 px × 1170 px.

4.3.3. Image fusion results and evaluation

4.3.3.1 Evaluation of spatial information

A CS, four unmixing-based and six DL approaches for image fusion were evaluated for the 07/06/2018 experiment.

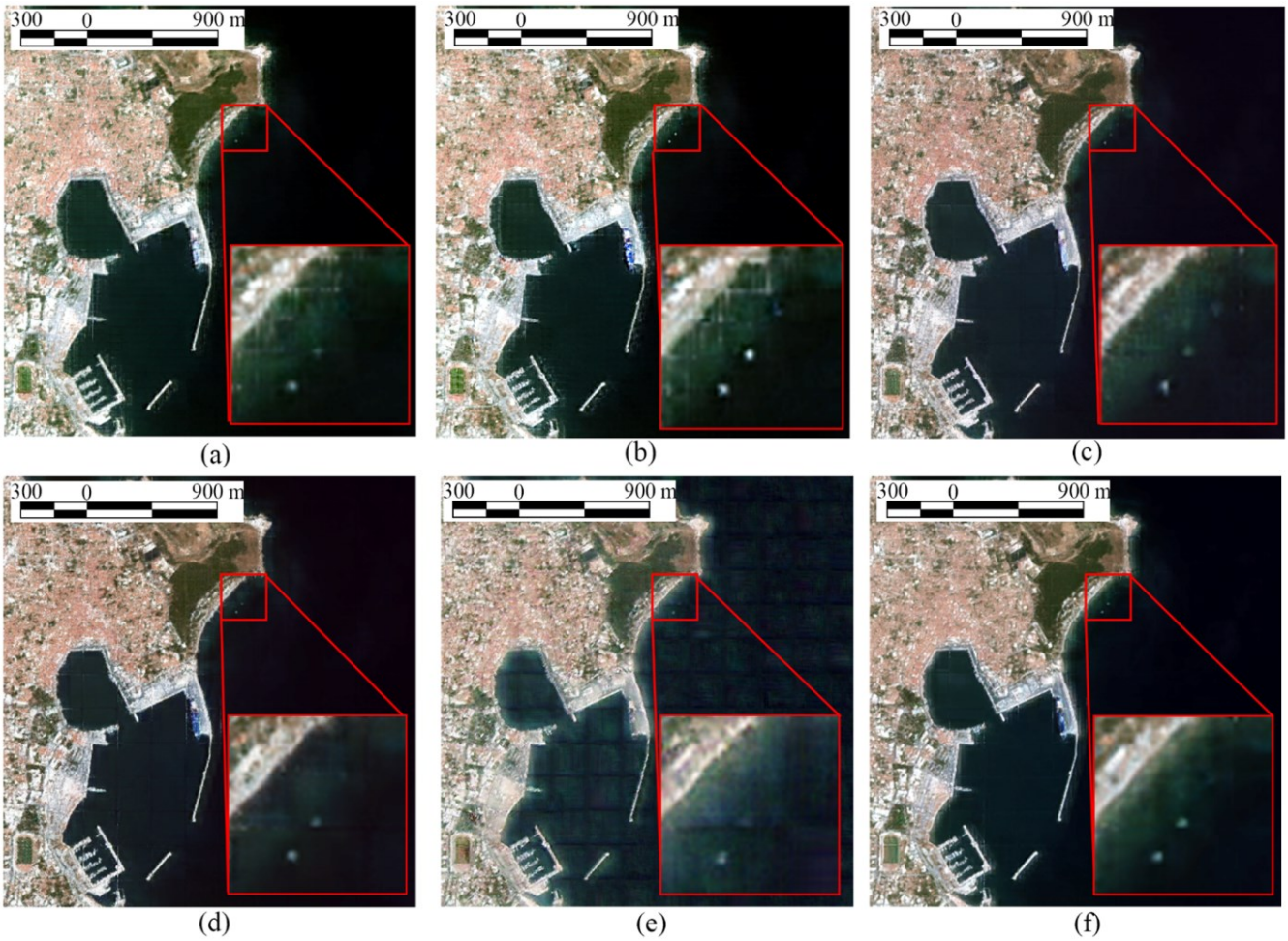


Figure 4.23. Fusion results for the S2 and WV-3 images acquired on 07/06/2018 (zoomed-in view of the target placement in red window) (natural colors). (a) PNN_VNIR+SWIR, (b) PNN_Siamese, (c) Fusion-ResNet, (d) Fusion-GAN, (e) SRGAN, (f) RCAN.

The fused images of the various methods have a 4 m spatial resolution and 13 bands. Natural color composites are shown in Figure 4.22 and Figure 4.23. For the CS and unmixing methods, the composites for the band combination (Table 4.2) that produced the best fusion results are presented. CNMF and HySure outperformed all the other methods in terms of spatial information. Their results present clear edges without any blurring or remaining artifacts and all plastic targets are discernible. Concerning the DL approaches, Fusion-ResNet and Fusion-GAN show less noisy outputs compared to the other DL networks. The lower performance of the DL methods could be explained by training in the reduced scale (lack of relevant high-frequency information) because of the unavailability of ground-truth data and the high ratio (1/5). It is noted that contrary to the lower performance in the high resolution, all DL methods produced high MSE or L1 scores in the reduced resolution (Table 4.3).

The conclusions reached by the visual interpretation of Figure 4.22 and Figure 4.23 are confirmed by image quality metrics (i.e. PSNR, ERGAS, RMSE, and SSIM [300]) and indeed, CNMF_2347 presents the best results according to 3 out of 4 metrics (Table 4.4).

4.3.3.2 Evaluation of spectral information

Spectral signatures were also evaluated (Figure 4.24, Figure 4.25, Figure 4.26, and Figure 4.27). In these Figures, the spectral signatures of the “plastic bottles”, “fishing nets”, “plastic bags” targets and of a random water pixel respectively, are shown for all fused results as well as for the original S2 and WV-3 (resampled to 20 m) images. In Table 4.5, the similarity between these spectra from the fusion results and the reference S2 images is examined using the measure of SAD and CC [300]. It can be seen that all DL methods except for SRGAN outperformed the

Table 4.4. Quality metrics between fusion results and WV reference images

Fusion Method	PSNR	ERGAS	RMSE	SSIM
FUSE_2348	5.89	64.38	331.57	0.192
FUSE_all7	8.71	46.69	239.69	0.067
Lanaras'_1234	9.01	44.35	234.55	0.300
Lanaras'_2347	8.11	49.20	260.48	0.321
CNMF_2347	**27.24**	**5.45**	**29.06**	0.490
CNMF_all7	25.68	6.54	34.94	**0.597**
HySure_all7	19.71	12.98	67.64	0.559
HySure_all9	21.25	10.89	56.65	0.571
PCA_23478	15.24	21.64	114.27	0.028
PCA_all9	15.46	21.10	111.37	0.028
PNN_VNIR+SWIR	4.65	73.30	388.60	0.069
PNN_VNIR	4.62	73.49	389.47	0.068
Fusion-GAN	5.21	68.71	365.10	0.072
SRGAN	4.45	74.97	396.85	0.052
RCAN	5.78	64.50	343.94	0.075
Fusion-ResNet	5.05	69.99	371.39	0.063
PNN-Siamese	4.88	71.39	378.85	0.068

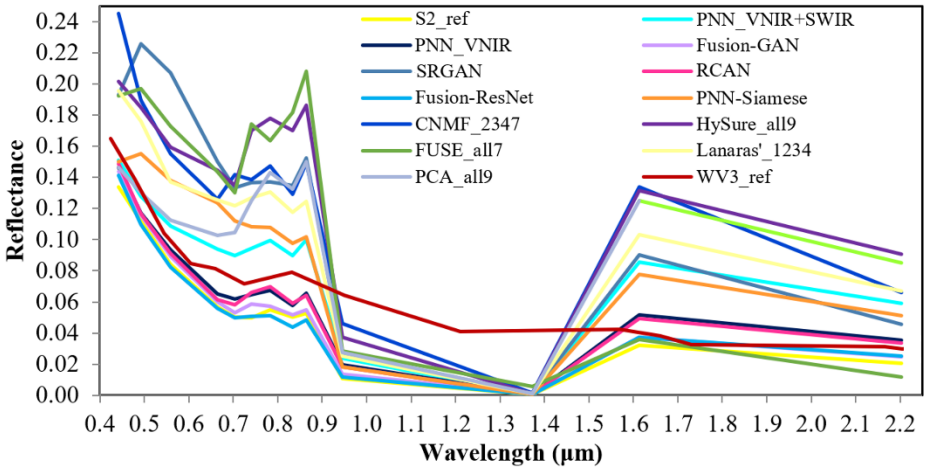


Figure 4.24. Spectral signatures of plastic bottles from all fusion results. S2 and WV-3 reference spectra are also included

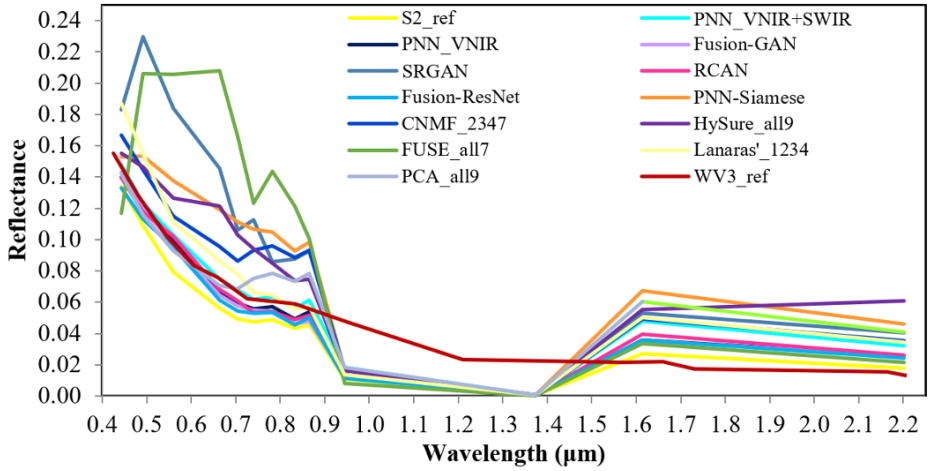


Figure 4.25. Spectral signatures of fishing nets from all fusion results. S2 and WV-3 reference spectra are also included

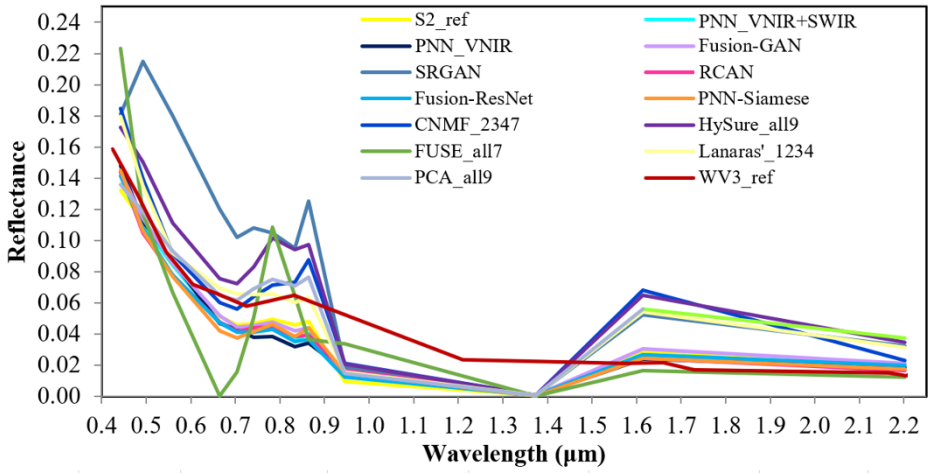


Figure 4.26. Spectral signatures of plastic bags from all fusion results. S2 and WV-3 reference spectra are also included

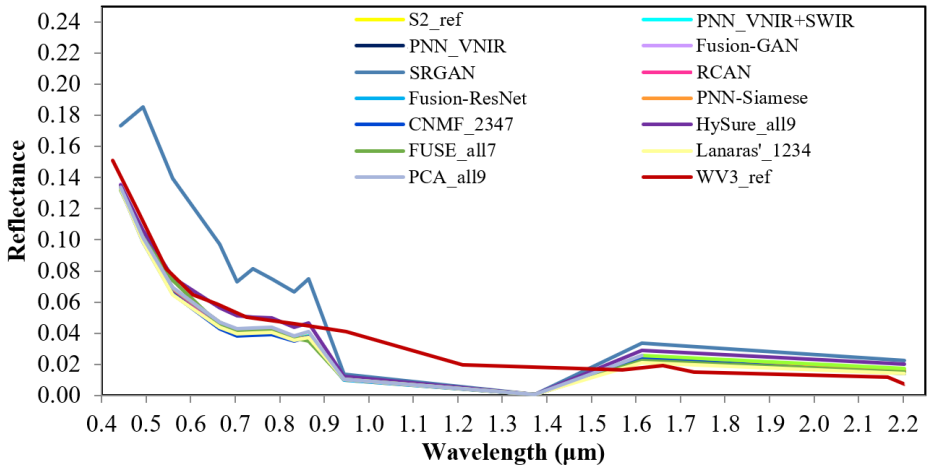


Figure 4.27. Spectral signatures of a random water pixel from all fusion results. S2 and WV-3 reference spectra are also included

Table 4.5. Spectral angle distances and correlation coefficients between fusion results and S2 reference spectra

Fusion Method	SAD						CC					
	Plastic Bottles	Fishing Nets	Plastic Bags	Water Pixel 1	Water Pixel 2	Water Pixel 3	Plastic Bottles	Fishing Nets	Plastic Bags	Water Pixel 1	Water Pixel 2	Water Pixel 3
FUSE_2348	0.37098	0.37687	0.28072	0.09187	0.03267	0.04300	0.76207	0.77718	0.89632	0.99360	0.99874	0.99888
FUSE_all7	0.36224	0.44848	0.41681	0.13655	0.03159	0.36566	0.77258	0.68648	0.83925	0.97591	0.99866	0.84242
Lanaras'_1234	0.24236	0.16581	0.20032	0.03918	0.02284	0.06292	0.92348	0.95368	0.96610	0.99951	0.99944	0.99792
Lanaras'_2347	0.33472	0.19233	0.19853	0.03232	0.01730	0.04230	0.81112	0.95290	0.94393	0.99937	0.99960	0.99896
CNMF_2347	0.24073	0.18939	0.15737	0.03101	0.02657	0.03562	0.92500	0.95776	0.96313	0.99885	0.99919	0.99950
CNMF_all7	0.37138	0.28599	0.14775	0.03040	0.03516	0.01809	0.77054	0.90493	0.97559	0.99910	0.99912	0.99979
HySure_all7	0.35376	0.24772	0.19551	0.08176	0.07707	0.12098	0.79265	0.92077	0.95011	0.99517	0.99613	0.99220
HySure_all9	0.35525	0.23990	0.18414	0.08190	0.07770	0.12228	0.78959	0.92578	0.96016	0.99522	0.99609	0.99202
PCA_23478	0.42454	0.23186	0.17579	0.02572	0.03416	0.08236	0.66803	0.94265	0.97053	0.99953	0.99921	0.99701
PCA_all9	0.41877	0.21243	0.18778	0.02630	0.02649	0.07233	0.67928	0.95398	0.96542	0.99951	0.99950	0.99773
PNN_VNIR +SWIR	0.25883	0.10933	0.08900	0.00928	**0.00909**	0.00962	0.91513	0.99088	0.99113	0.99990	**0.99991**	0.99992
PNN_VNIR	0.09894	0.07254	0.13386	0.00912	0.01276	**0.00511**	0.99165	0.99419	0.98222	0.99989	0.99979	**0.99997**
Fusion-GAN	**0.03372**	**0.05548**	**0.05377**	0.01514	0.01417	0.00944	**0.99835**	**0.99599**	**0.99621**	0.99987	0.99981	0.99994
SRGAN	0.24277	0.21780	0.20430	0.21516	0.18271	0.19169	0.90649	0.92685	0.93819	0.93664	0.95560	0.95782
RCAN	0.09238	0.08684	0.09470	0.01033	0.01118	0.00878	0.99122	0.99066	0.98900	**0.99991**	0.99988	0.99991
Fusion-ResNet	0.05893	0.07196	0.09577	0.00982	0.01300	0.00918	0.99444	0.99382	0.98990	0.99987	0.99979	0.99991
PNN-Siamese	0.25876	0.27071	0.10190	**0.00843**	0.00921	0.00727	0.90122	0.90228	0.98835	0.99990	0.99989	0.99994

non-DL methods in preserving the S2 spectral characteristics. Among the DL methods, the best performance was shown by Fusion-ResNet and Fusion-GAN which are lighter networks than RCAN and SRGAN as can be seen from the number of trainable parameters shown in Table 4.3. CNMF_2347 yields also a good spectral performance.

Overall, water pixels present the lowest SAD values. This is attributed to the fact that the pixel coverage of the targets in the 4 m fused image is higher than the pixel coverage in the original 20 m S2 image. SAD values between water pixels are low because in both images the water pixels are pure. SAD values between plastic pixels are higher because the S2 plastic pixel is mixed with water, while the WV-3 plastic pixel is pure plastic (Figure 4.28). As shown in Figure 4.29, the resulting pure plastic pixels in the fused image enable inter- and intra-class separability.

The WV-3 reference spectra mostly differ from the S2 reference spectra at around 1.0–1.5 μm (Figure 4.24, Figure 4.25, Figure 4.26, and Figure 4.27). This could be explained by observing the lack of overlap in the respective SRFs of the sensors in this spectral region (Figure 4.30). From our experiments, we did not reach safe conclusions about correlations between the performance of the fusion methods and the similarity of SRFs. This would require further analysis which is out of the scope of this study.

Besides the above, several marine litter indexes were also evaluated on the CNMF_2347 fused image and it was shown that FDI [84] which uses SWIR information produced the best results. More information about the FDI performance can be found in [301] where a second relevant experiment is also described.

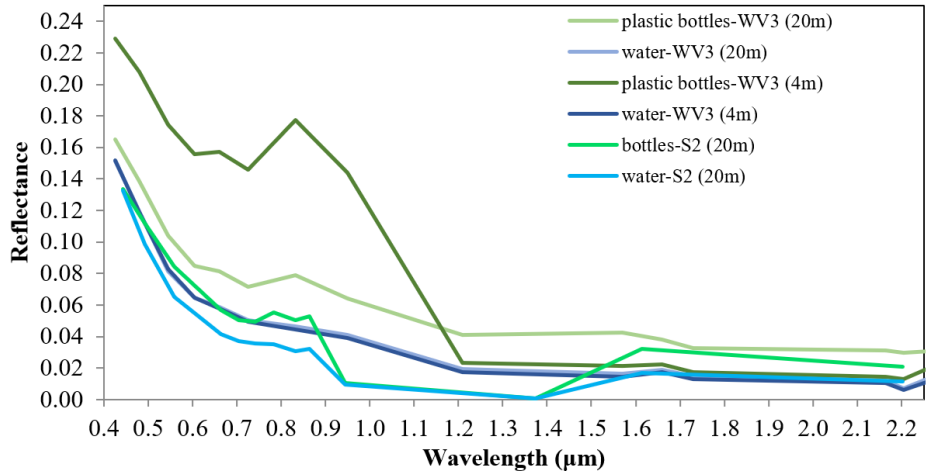


Figure 4.28. Spectral signatures of the plastic bottles target and a water pixel from the S2 and WV-3 reference images.

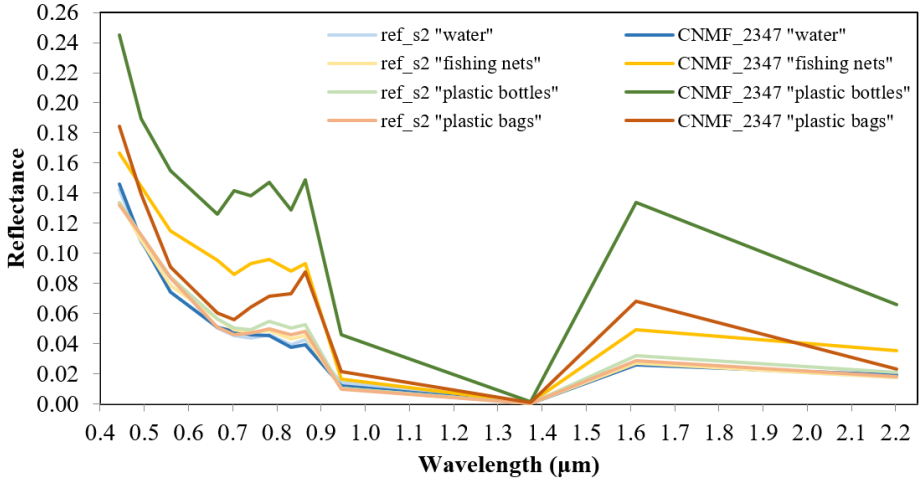


Figure 4.29. Comparison of water and plastic target spectra before and after S2 and WV-3 fusion with CNMF method and 2347 band combination.

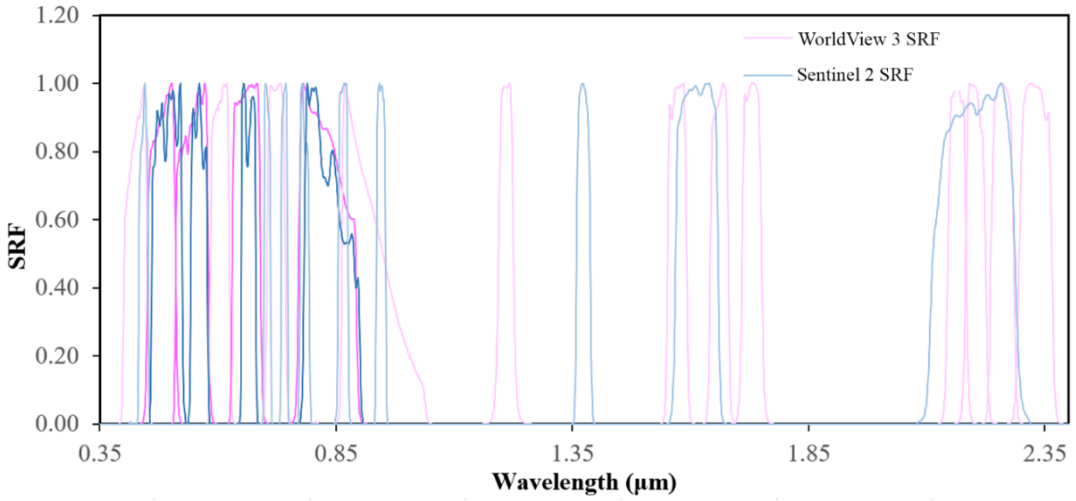


Figure 4.30. Spectral response functions of the S2 and WV-3 sensors.

4.3.4. Conclusions

In this study, S2 and WV-3 images were fused to increase the capability of S2 imagery to detect marine plastic targets. Although the plastics might be visible in the VHR images, the SWIR region of the spectrum, which is captured in S2 imagery, is very important for the detection and identification of plastics. Various image fusion techniques were evaluated on artificial plastic targets for their performance in terms of preserving spectral and spatial information. The CNMF method proved the best for this application as it produces a fused image with clear edges, no blurring, and a relatively favorable impact on the spectral characteristics of the materials. DL methods showed high performance in terms of spectral similarity between the fused and the S2 images. In this regard, Fusion-GAN and Fusion-ResNet (created for the purpose of the study) outperformed all other fusion methods (non-DL and DL).

Furthermore, results showed that the VNIR combination is the most efficient for the image fusion. This is an important finding, because most satellite sensors are sensitive to the VNIR part of the spectrum, so the likelihood of temporally close acquisitions of the same marine litter accumulation increases, which is a critical factor considering that marine litter accumulations move, change, or vanish in a relatively short time. However, exactly for this reason, the described fusion approach might be limited for detecting litter accumulations in the sea, unless suitable constellation-based solutions can offer the possibility of “nearly” simultaneous image acquisitions between different satellites. It is worth mentioning that, still remaining in the frame of monitoring plastic pollution in the environment, the proposed fusion approach could find application in the detection of accumulations of plastic litter on land close to water bodies (i.e. sources of marine litter), which is a more static scenario, although with higher complexity in terms of background.

Concerning the computational needs, the usage of our approach would not be prevented at an operational level because current mainstream industrial and research hardware is capable of maintaining the computational needs for the fusion methods implemented in the study, even for the DL methods. Regarding costs, even though the required processing equipment can be considered inexpensive, procurement of the required commercial VHR data indeed increases the cost, at least for the moment.

Several that detect plastic material using discriminative features in the VIS, NIR, and SWIR parts of the spectrum were applied to the fused CNMF image. In this case, the SWIR bands of the fused image proved to be quite useful and FDI showed the best performance.

Another interesting finding of this research was the observation of dissimilarities in the spectral regions of S2 bands between the signatures of the various plastic materials extracted from the fused images. This topic along with pinpointing the scalability of the proposed methodology could be the subject of future work.

CHAPTER 5

Paired and unpaired GANs for NIR band generation in VHR RGB imagery[13]

In Chapter 5 a methodology that was proposed in this PhD thesis to generate the NIR band in VHR RGB imagery by exploiting paired and unpaired GANs is presented. In section 5.1 a literature review is provided on paired and unpaired image-to-image translation (ITIT) applications in Remote Sensing. In section 5.2 the motivations and objectives of the study are presented. In section 5.3, at first the data are described, and then a theoretical background and implementation details of the methodology are provided. In section 5.4 the results of the methodology are discussed. Finally, in section 5.5 the conclusions and contributions are summarized and future work is suggested.

5.1 Related work

ITIT is an image processing method whose basic idea is to learn the mapping functions between an input and an output image [90]. ITIT can be either performed for paired data (co-registered input and output image) or unpaired data. Multiple Remote Sensing studies have investigated paired ITIT through deep learning (DL) and most often conditional GANs (cGANs) to generate missing information. The applications are various and among others include SAR↔optical, visible (VIS)→map, optical →elevation (digital terrain/surface model), thermal infrared (TIR)↔VIS, VIS-VIS, grayscale→RGB, spectral super-resolution (SSR) (RGB/MS→HS), and RGB→NIR. Concerning unpaired IT, so far, it has been used in Remote Sensing applications exclusively for unsupervised domain adaptation (UDA) purposes as an intermediate step to enhance the semantic segmentation output.

5.1.1 Paired image-to-image translation (ITIT) – Broadly related work

5.1.1.1 SAR↔optical ITIT

The vast majority of the Remote Sensing ITIT research has been focused on SAR↔optical paired deep ITIT because SAR data are unaffected by atmospheric conditions. Due to higher accessibility, most of the studies process high-resolution (>=5 m) data [302][303][91][92]. However, recently a few studies have been conducted on VHR data. In [90], aiming at enhancing change detection performance, NICE-GAN [304], an introspective network based on CycleGAN [305] with multi-scale formulation in the discriminator and residual attention, was used for SAR↔optical ITIT. With the same goal but in a non-adversarial setting, the authors in [306] implemented an optical-SAR domain adaptation-based change detection network where distribution discrepancies in Hilbert space were included. An ITIT adaptation-based change detection technique (based on NICE-GAN) was also proposed in [307] where the features of optical images were transferred to SAR. In [93], the authors took advantage of both Pix2Pix (cGAN) [308] and CycleGAN and performed SAR↔optical mapping by incorporating an additional network called the distortion adaptive module in both directions. Finally, in [309], a Parallel-GAN was proposed for SAR→optical ITIT consisting of a backbone ITIT subnetwork and an adjoint optical image reconstruction subnetwork.

[13] **Kristollari, V.** and Karathanassi, V., 2024. Exploiting Paired and Unpaired Generative Adversarial Networks for NIR Band Generation in VHR RGB Satellite Imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (under review)

5.1.1.2 VHR VIS→map ITIT

Interest has also been shown in the deep ITIT of VHR VIS satellite images to maps because it could prove very beneficial when timely updates are required. Several methods have been applied in a paired setting. In [97], a scale-consistent cGAN was proposed to simultaneously generate multi-level tile maps from multi-scale RS images. In addition, in [98], adversarial deep transfer training schemes were combined with attention-based network designs to generate maps over various regions, and in [310], a level-aware fusion network for multilevel map generation was introduced. Due to the scarcity of paired data, unpaired samples have also been proven to be useful. In [311], the authors designed a semi-supervised learning strategy based on training GANs on rich unpaired samples and then applied fine-tuning on limited paired samples. In addition, in [312], Semi-MapGen was proposed, a network based on semi-supervised GANs, which requires only a small set of accurate and complete matched data and plenty of unpaired data.

5.1.1.3 Optical→elevation ITIT

Mapping optical data to elevation information in a paired fashion is another application that has been studied in the DL literature as an affordable alternative to approaches that require Lidar, InSAR, or stereo pairs. The study described in [99] was among the first to apply a cGAN to translate VHR optical data to elevation. Adversarial learning was also proposed in [313] where Pix2Pix was implemented to map optical to elevation data for S2 and UAV imagery. In a non-adversarial setting, the authors of [314] applied an encoder-decoder model with skip connections [204] and residual blocks [243] for DSM generation from aerial images. Several strategies were investigated and showed that the performance can vary according to the dataset morphology. Similarly, in [315], a U-Net [204] with residual blocks was applied to create elevation information for airborne images.

5.1.1.4 TIR↔VIS ITIT

Other DL studies of paired data have investigated the TIR→VIS ITIT in data collected by geostationary meteorological satellites by use of the Pix2Pix model to enrich the collected information. In [96], Pix2Pix was trained on daytime pairs of the 10.8 μm longwave radiance band and the 0.675 μm visible band of the meteorological imager (MI) onboard COMS, to create the non-existent nighttime visible reflectance band. In a later study for the above-mentioned satellite [316], Pix2Pix was applied for virtual nighttime visible imagery generation using multiband infrared observations and a brightness temperature difference. In a similar concept, in [317], Pix2Pix was trained on thermal band differences of the Advanced MI (AMI) sensor, of the GK-2A Geostationary Korea Multi-Purpose Satellite, to provide virtual RGB bands during day and night. Besides the abovementioned low-resolution meteorological applications, VIS→TIR VHR mapping by Pix2Pix has also been explored [318] as an intermediate step to achieve thermal geolocation in low illumination environments, motivated by the limited availability of satellite thermal data.

5.1.1.5 VIS-VIS ITIT

Data collected by geostationary meteorological satellites have also been used for cross-satellite deep paired VIS-VIS ITIT with the Pix2Pix network to generate missing bands. In [94], Pix2Pix was trained on blue band radiance images of AHI to generate a simulated green band (useful for monitoring water and vegetation) for the GOES-16 ABI sensor. In addition, in [95] Pix2Pix was trained on blue band radiance images of the GK-2A/AMI sensor to generate simulated green and red bands (useful for monitoring atmospheric environments) for GEMS onboard the GK-2B satellite.

5.1.2 Paired image-to-image translation (IT) – Closely related work

5.1.2.1 Spectral super-resolution (SSR)

HS images carry valuable spectral information. However, the fact that until very recently satellites that provide global and publicly available HS data were inexistent and the high cost of airborne sensors hinders the exploitation of HS information. Thus, HS image reconstruction from RGB and MS data (SSR) has attracted recent attention. In the past, the problem was approached in a non-DL manner with linear unmixing [319], regression [320], and methods that exploit sparse representations and dictionary learning [321][322]. More recently, numerous DL

studies have been published. All SSR literature methods so far, have evaluated their paired methodologies in the output as a whole without isolating particular bands (e.g. NIR).

In [102], a variant of the dense CNN called "Tiramisu" [323] was trained on MS data collected from ALI on board the EO-1 satellite to predict the HS Hyperion bands. The qualitative evaluation showed that the predicted output was less noisy than the ground-truth data. In addition, the quantitative evaluation employed abundance estimation. In [103], CNN regression models were investigated to produce the Hyperion HS bands from the Landsat 7/8 MS bands. The authors showed that the CNN regression produced better performance compared to conventional regression methods. The model outputs were also evaluated by classification with SVM over principal components (PCs). In [100], the authors proposed a cGAN with an additional spectral discriminator to map RGB to HS information in GF-5 data. The quantitative spectral and spatial scores showed that the proposed network was more robust than alternative non-adversarial approaches. In [101] an encoder-decoder model with attention to semantic similarity was implemented to spectrally super-resolve RGB to HS images in 1 m spatial resolution. MS→HS SSR (Hyperion images) by employing semantic information was also proposed in [324] in the form of a change detection subnetwork. Finally, an encoder-decoder model was trained in [325] to spectrally super-resolve UAV and GF-5 MS images. Several input band combinations were explored and it was shown that the inclusion of the NIR band can enhance the performance of the final HS output.

In other studies, SSR attempts have been made to enhance the exploitation of both spatial and spectral information in high/medium spatial resolution data. In [326], the authors proposed a progressive spatial-spectral joint network to reconstruct satellite and airborne HS data from MS. In addition, in [327], a spatial-spectral residual attention was exploited for MS→HS mapping, and in [328], a spatial-spectral feature attention module was introduced in a GAN for both synthetic and real data scenarios.

5.1.2.2 RGB-to-NIR ITIT

The Remote Sensing studies published so far in RGB→NIR paired ITIT have focused exclusively on vegetation applications. The significance of the NIR band in providing rich information for the determination of vegetation parameters has since long been established [104][105][329][330]. In the last years, MS sensors onboard UAVs have been extensively used in precision agriculture. However, small producers have difficulty affording the required equipment. Low-cost RGB cameras on board lightweight UAVs are a more affordable option [106]. Thus, RGB→NIR ITIT could prove very useful for vegetation monitoring.

Before the broad application of DL, this problem had been approached by regression analysis on conventional cameras depicting crops [331], where a green-NIR correlation had been indicated. In recent years, several DL studies have been published for the RGB→NIR ITIT in vegetation areas and the vast majority have employed the Pix2Pix model. In [332], Pix2Pix was trained on UAV RGB crop data to generate the NIR band. The L1 loss was replaced with the Charbonnier penalty function and both in- (same crop type) and out-domain (different crop types) experiments were performed. In [107], RGB→NIR ITIT was performed on WV-2 data by Pix2Pix with residual blocks in the generator. The authors focused on forest areas and evaluated the performance in a cross-domain setting (SPOT, Planet with finetuning). The Planet data ITIT showed less satisfactory evaluation scores due to higher heterogeneity compared to the training inputs. It was also stated that adversarial training increased the performance and that the inclusion of NIR in the classification task lowered the size of the needed annotations. One of the research motivations was the fact that publicly available RGB databases (e.g. [333]) do not contain the NIR band. However, ITIT could generate the missing information. In [106], Pix2Pix was trained on low-cost UAV RGB cameras to estimate the NIR band for agricultural purposes. The network outperformed a previously proposed endmember-based method [334]. The authors also investigated combinations of the original L1 loss with the structural similarity index (SSIM) and a perceptual loss to achieve slightly better results. In [335] Pix2Pix was employed for RGB→NIR field imagery ITIT (agricultural areas) with a DenseNet architecture [323] as the generator. Comparison with the original U-Net generator showed slight improvement. Finally, in [108], RGB→NIR ITIT was performed on S2 data collected all year round. It was observed that the model was unaffected by corrupted pixels but did not show satisfactory generalization ability to Landsat-8 data. The failure in the out-domain performance could be attributed to differences in illumination, as well as atmospheric and sensor conditions.

5.1.3 Unsupervised domain adaptation (UDA)

Unpaired ITIT has recently widely been used in the form of UDA in VHR VIS Remote Sensing when annotations are available in the source domain but not in the target domain. The goal is to enhance cross-domain semantic segmentation (CDSS) tasks by decreasing domain shifts caused by the rich structure diversity, the variability in atmospheric/lighting conditions and viewing angles, as well as the different sensor characteristics. In [336], CDSS was performed by a curriculum-style local-to-global cross-domain adaptation framework. The adaptation process was conducted in an easy-to-hard way using an entropy-based score and adversarial learning. In [337], the CDSS task was approached by a bidirectional domain adaptation adversarial network that takes advantage of the information from both domains. In [338], the authors proposed a deep covariance alignment model to align category features. In [339], a two-stage framework was applied, which performs fine-grained local and category-level alignment on top of global alignment. The framework used adversarial learning and knowledge distillation. In [340], a cyclic GAN with residual connections was proposed followed by a semantic segmentation stage. An in-network resizer module was included to address the scale discrepancy. Finally, in [341], the authors implemented a lightweight UDA model relying on latent representation separation and mixing across domains which can be used in a one-shot setting.

5.2 Motivations and Objectives

As mentioned in the above literature review, the SSR-published studies have evaluated their paired data methodologies in the output as a whole without isolating particular bands like the NIR. In addition, to the best of our knowledge, the RGB→NIR literature in total, has exclusively explored only the vegetation category. However, the NIR information has also among others proven useful for general scene recognition [342], mineral mapping [343], plastic litter detection (sections 4.2, 4.3), and nighttime image generation for the monitoring of environmental and socio-economic dynamics [344]. Thus, the generation of the NIR band would be significant to be explored in more detail and for more land cover categories. Such research would also enforce the capability of including the missing NIR band in publicly available RGB databases like Google Earth and the ones presented in [333][72][65].

Concerning the unpaired ITIT, none of the previous RGB→NIR paired ITIT studies has employed UDA in combination with the paired ITIT to improve the NIR prediction. Since UDA has been irrefutably recognized as capable of decreasing the domain discrepancies in the CDSS task, it should be at least logical to test it in the paired ITIT task. Unlike the CDSS task, where spectral fidelity is not required when applying UDA (e.g. applying an SS model trained on green roofs (labeled source data) to classify red→green roofs (UDA) (unlabeled target data)), in the paired cross-domain ITIT, spectral fidelity is significant to predict a reliable NIR band. In this case, when applying UDA, the source and target data should be collected from the same geographic zone and in the same month to avoid seasonal changes.

This study aims at developing a methodology that predicts NIR information in VHR RGB in- and out-domain data (do not belong to the domain of the training set (different date/region/satellite)). The proposed methodology is composed of three stages. The first stage is NIR prediction (paired data – cGANs), the second stage is the implementation of UDA (unpaired data), and the third stage is the enhancement of the NIR predictions of the first stage by applying the pretrained paired cGANs on the UDA produced images. The goal of UDA is to create data that are closer radiometrically.

In summary, the main objectives can be summarized as follows:

a) To investigate the performance of the NIR prediction models in an in- and out-domain (different regions/sensors/dates) setting with heterogeneous bi-temporal data. Through several configurations, we explore the effects of normalization techniques, as well as the inclusion of residual blocks and attention modules.

b) To explore the possibility of UDA through unpaired GANs in improving the NIR prediction on data of independent domains. In three configurations, the effects of batch size and normalization techniques are examined.

c) To propose a three-stage GAN framework in a paired and unpaired setting to generate NIR images.

d) To implement the models on three main land cover thematic categories: a) impervious surfaces/urban fabric (manmade objects), b) vegetation (forest, crops), and c) ground.

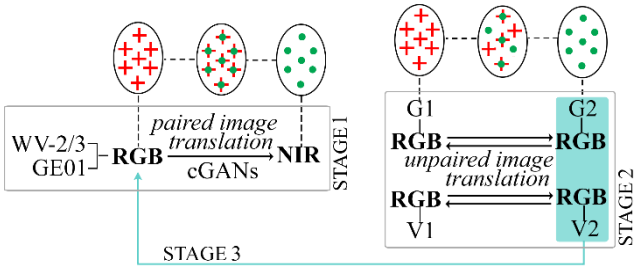


Figure 5.1. Depiction of the three-stage framework.

5.3 Data and methodology

5.3.1 Datasets

For the implementation of the methodology, Geoeye-1 (GE01) and WV-2/3 satellite VHR images were employed. The images contained four bands (RGB-NIR). They were collected from four European areas (Granada/Spain (G), Tønsberg/Norway (T), Rhodes/Greece (R), Venice/Italy (V)) in a bi-temporal fashion. The areas were heterogeneous since the morphology differed (G: dense high urban fabric/red-tiled roofs/steep mountains/few agricultural fields, T: sparse low buildings/grey-tiled roofs/flat terrain/high presence of agriculture and forest, R: dense urban fabric with terraces/few crops, V: very dense homogenous buildings/red-tiled roofs/limited vegetation). More info can be found in section 3.4 where this dataset was used for the first time. In the text "1" refers to the earliest date and "2" to the latest (e.g. G1, T2). Whenever needed, the images were resampled at 0.5 m spatial resolution. The experiments were performed on 8-bit radiometric resolution (often encountered on public datasets). Table 3.1 in Chapter 3 shows details about the images. Water areas were masked and not taken into account because they were out of the focus of this study.

5.3.2 Proposed Method

In this study, a three-stage GAN framework in a paired and unpaired setting is proposed. In the first stage prediction of the NIR band with RGB input was investigated, by employing cGANs (paired data/each pixel in the source data corresponds to a pixel in the target data and vice-versa). In the second stage, a model based on CycleGAN for the G1/G2 and the V1/V2 RGB image pairs was applied, aiming at producing closer radiometrically images through UDA (unpaired data). The above pairs were selected because they were collected from the same geographic region and in the same month to avoid seasonal changes. Finally, in the third stage, the effect of UDA (second stage) on the NIR prediction (first stage) for the G2 and V2 images was explored. The three-stage framework is shown in Figure 5.1.

5.3.2.1 Conditional Generative Adversarial Networks (cGANs)

GANs are generative models that learn a mapping from random noise vector z to output image y, $G: z \rightarrow y$ [284]. GANs consist of a generator G and a discriminator D. In image generation applications, the goal of the generator is to produce synthetic (fake) images that challenge the ability of the discriminator to differentiate them from real images. The training is described by the objective function shown in Equation (5.1) where G aims at minimizing L_{GAN} against an adversarial D that aims at maximizing it.

$$\min_G \max_D L_{\mathrm{GAN}}(G, D) \quad (5.1)$$

$$L_{\mathrm{GAN}}(G, D) = \mathbb{E}_y [\log D(y)] + \mathbb{E}_z [\log (1 - D(G(z))] \quad (5.2)$$

Conditional GANs learn a mapping from the observed image x and random noise vector z, to y, $G: \{x, z\} \rightarrow y$ [308]. In this case, the minmax two-player training is performed on Equation (5.3).

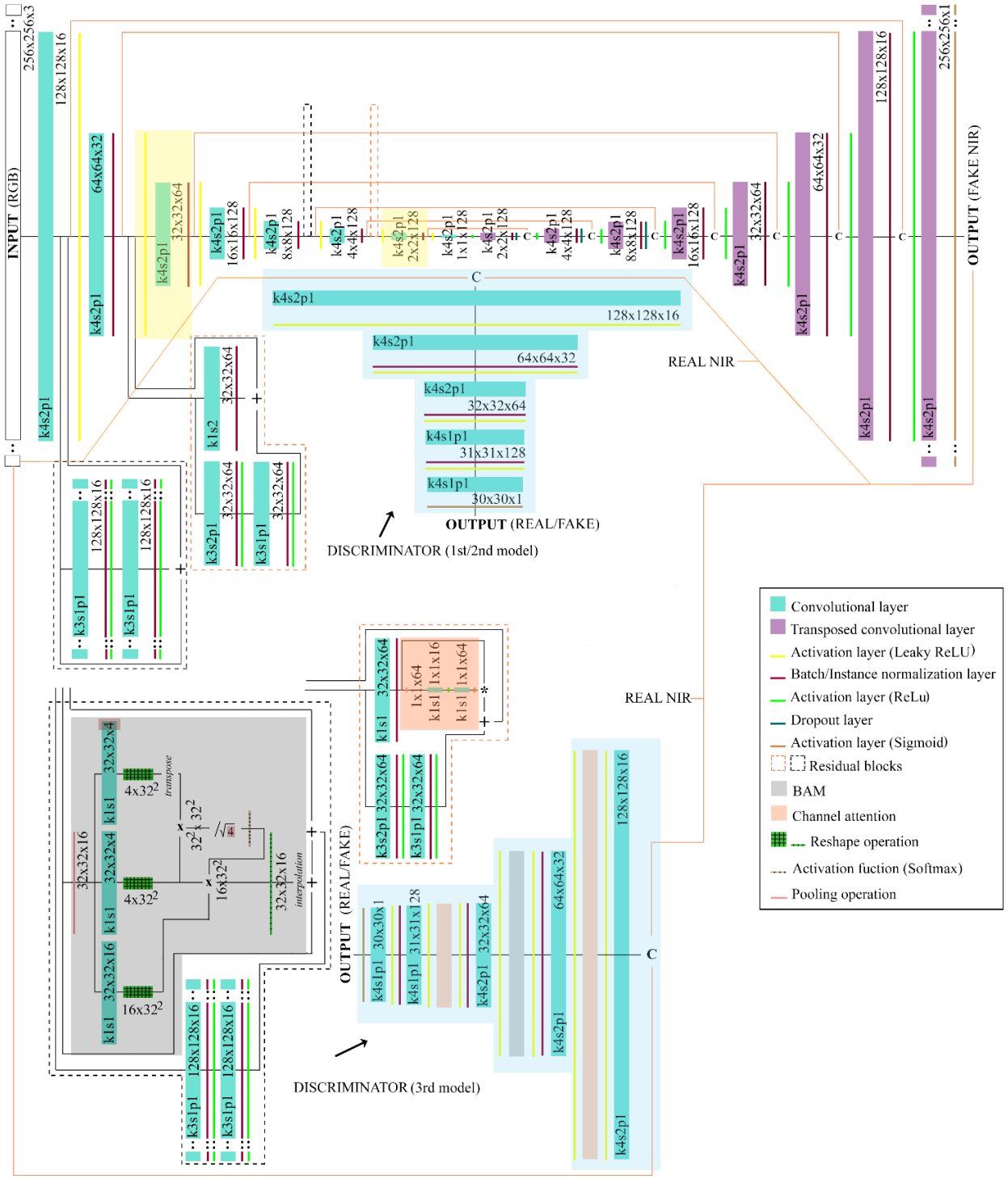


Figure 5.2. Architecture of the three cGANs for the NIR prediction. The yellow highlight shows layers that were not included in the second and third models.

$$L_{\text{cGAN}}(G,D) = \mathbb{E}_{x,y}\left[\log D(x,y)\right] + \mathbb{E}_{x,z}\left[\log\left(1 - D(x,G(x,z))\right)\right] \quad (5.3)$$

In this study, the training of the models was implemented in an alternating way according to Equation (5.4) which includes the addition of the $L1$ loss (Equation 5.5) to Equation (5.3) so that the generator except for antagonizing the discriminator is also tasked to produce outputs similar to the ground-truth (reconstruction loss). In Equation

5.5 λ is a trade-off parameter between $L_{cGAN}(G,D)$ and $L_{L1}(G)$. In our study, the value was set to 100.

$$\min_G \max_D L_{\text{cGAN}}(G,D) + \lambda L_{L1}(G) \quad (5.4)$$

$$L_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1] \quad (5.5)$$

5.3.2.2 cGANs implementation

In the first stage of the proposed methodology three cGANs of different architecture were applied. The first model was inspired by Pix2Pix which is very popular in the paired ITIT literature and was used as a basis for the development of the other two models. As shown by the creators of the Pix2Pix model, the inclusion of noise in the input is redundant in the ITIT application because it is ignored, thus it was not included. However, dropout with value 0.5 (randomly zeroes some of the input tensor) was used in some hidden layers to provide stochasticity both in training and inference.

In all models, the dimensions of the generator input patch were 256x256x3 (RGB) and of the output patch were 256x256x1 (NIR prediction). In addition, the input of the discriminator was the concatenation of the RGB patch with either the NIR ground-truth or the generator prediction. All three architectures are shown in Figure 5.2.

The first model generator consisted of eight convolutional layers in the encoder part and eight transposed convolutional layers in the decoder part. Skip connections through concatenation were applied between the encoder and the decoder layers. Leaky ReLU [345] (Equation (5.6)) was selected in the encoder part as the activation function, while ReLU [165] (Equation (2.11)) was used in the decoder part and Sigmoid [166] (Equation (2.12) in the output layer. Dropout was applied in three decoder layers. Concerning the discriminator, it consisted of five convolutional layers and was in a PatchGAN form [308].

$$\varphi(x) = \max(0.1x, x) \quad (5.6)$$

The second model generator was constructed based on the first model by adding two different types of residual blocks, the identity (enclosed in black dotted line) and the convolutional (enclosed in orange dotted line). The residual blocks were added to the encoder part of the generator. The architecture of the discriminator of the second model is the same as that of the first model. Finally, the third model generator was constructed based on the second model by adding spatial (BAM (basic attention module) [65]) and channel attention [299] modules on the residual blocks. The attention modules were also added to the discriminator. The logic of the architecture design was based on testing state-of-the art ANN concepts (residuals, channel, spatial attention). In addition, the particular concepts were proven successful in previous research of the author (sections 3, 4.3).

The first model was trained on the G1 image (G1BN, G1IN), the T2 image (T2BN), and the whole dataset (allBN, allIN) in two versions, one with batch normalization (BN) and one with instance normalization (IN). The second model was trained with IN on G1 (G1INRB) and the whole dataset (allINRB). Finally, the third model was trained only on the whole dataset (allINRBAt). The learning rate was set to $2\text{x}10^{-4}$.

Training details for the cGANs are shown in Table 5.1. The inference time for a batch of 32 patches was

Table 5.1. cGANs training details

Training details	G1BN	G1IN	G1INRB	T2BN	allBN	allIN	allINRB	allINRBAt
Epochs	50	50	50	50	150	150	150	150
Batch size	32	32	32	32	32	32	32	32
Patch size	256	256	256	256	256	256	256	256
Training steps	1312	1312	1312	1504	1312	1312	1312	1312
Trainable params G	3,404,801	3,404,801	3,776,257	3,404,801	3,404,801	3,404,801	3,776,257	3,811,913
Trainable params D	175,793	175,793	175,793	175,793	175,793	175,793	175,793	179,205
Training time	3h	3h	3h	3h	10h	10h	11h	14h

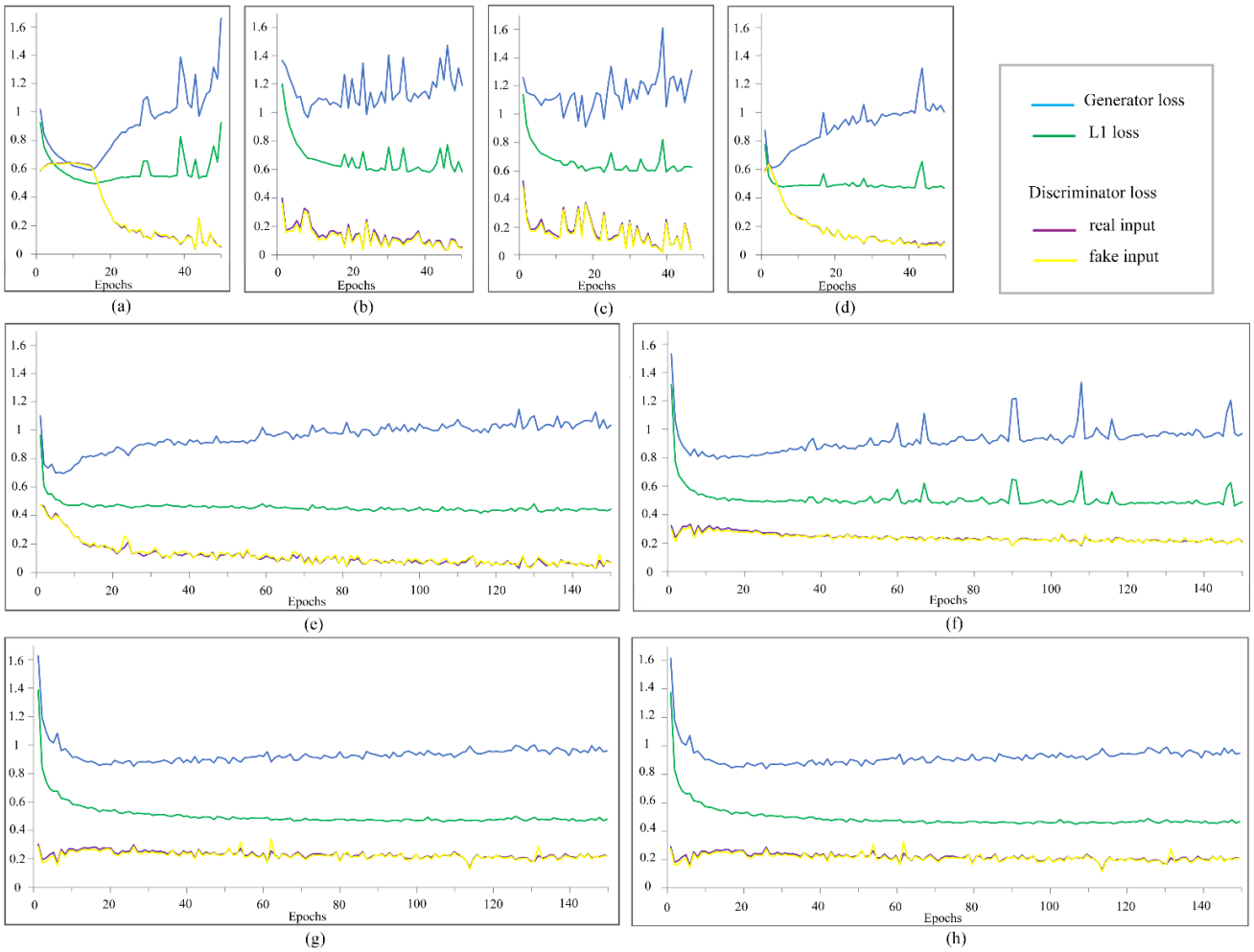


Figure 5.3. cGANs loss function values during training. (a) G1BN, (b) G1IN, (c) G1INRB, (d) T2BN, (e) allBN, (f) allIN, (g) allINRB, (h) allINRBAt

0.029 s for the first model, 0.039 s for the second, and 0.104 s for the third. In Figure 5.3 graphs of the loss functions for the generator and the discriminator are depicted during training. To make the graph more easily perceptible, the generator loss was divided by ten and the $L1$ loss was multiplied by ten. The selection of the final weights for each model was performed by an empirical process based on the lowest $L1$ loss values in combination with the observation of the performance of the predictions on image samples. It is observed that $L1$ converges faster in the BN models than in the IN. Also, smoother lines are observed in G1BN and T2BN compared to G1IN and G1INRB.

5.3.2.3 CycleGAN

CycleGAN was designed to perform unpaired ITIT with adversarial training. Since the $G: x \rightarrow y$ mapping is under-constrained, an inverse mapping $F: y \rightarrow x$ is also employed and a Cycle-consistency loss is introduced to impose $F(G(x)) \approx x$ and $G(F(y)) \approx y$ (Equation (5.7)).

$$L_{\text{cyc}}(G,F) = \mathbb{E}_x\left[\|F(G(x)) - x\|_1\right] + \mathbb{E}_y\left[\|G(F(y)) - y\|_1\right] \quad (5.7)$$

Both mappings are simultaneously trained. The objective function is a combination of the Cycle-consistency and the adversarial losses. Two discriminators (D_x , D_y) are included in CycleGAN, one for each mapping. D_x aims at differentiating x from $F(y)$ and D_y aims at distinguishing y from $G(x)$. The adversarial loss for the $G: x \rightarrow$

y mapping and its discriminator D_y is shown in Equation (5.8). The equivalent loss for the $F: y \rightarrow x$ mapping and its discriminator D_y is shown in Equation 5.11.

$$L_{\text{GAN}}(G, D_y) = \mathbb{E}_y \left[\log D_y(y)\right] + \mathbb{E}_x \left[\log\left(1 - D_y(G(x))\right)\right] \quad (5.8)$$

$$L_{\text{GAN}}(F, D_x) = \mathbb{E}_x \left[\log D_x(x)\right] + \mathbb{E}_y \left[\log\left(1 - D_x(F(y))\right)\right] \quad (5.9)$$

Except for the above losses, as suggested in [305] an identity loss (Equation (5.10)) [346] was additionally utilized in our study to retain color fidelity between the input and the output. Thus, the full objective is expressed in Equation (5.11).

$$L_{\text{identity}}(G, F) = \mathbb{E}_y[\|G(y) - y\|_1] + \mathbb{E}_x[\|F(x) - x\|_1] \quad (5.10)$$

$$L(G, F, D_x, D_y) = L_{\text{GAN}}(G, D_y) + L_{\text{GAN}}(F, D_x) + \lambda_1 L_{\text{cyc}}(G, F) + \lambda_2 L_{\text{identity}}(G, F) \quad (5.11)$$

where λ_1:10 and λ_2:5

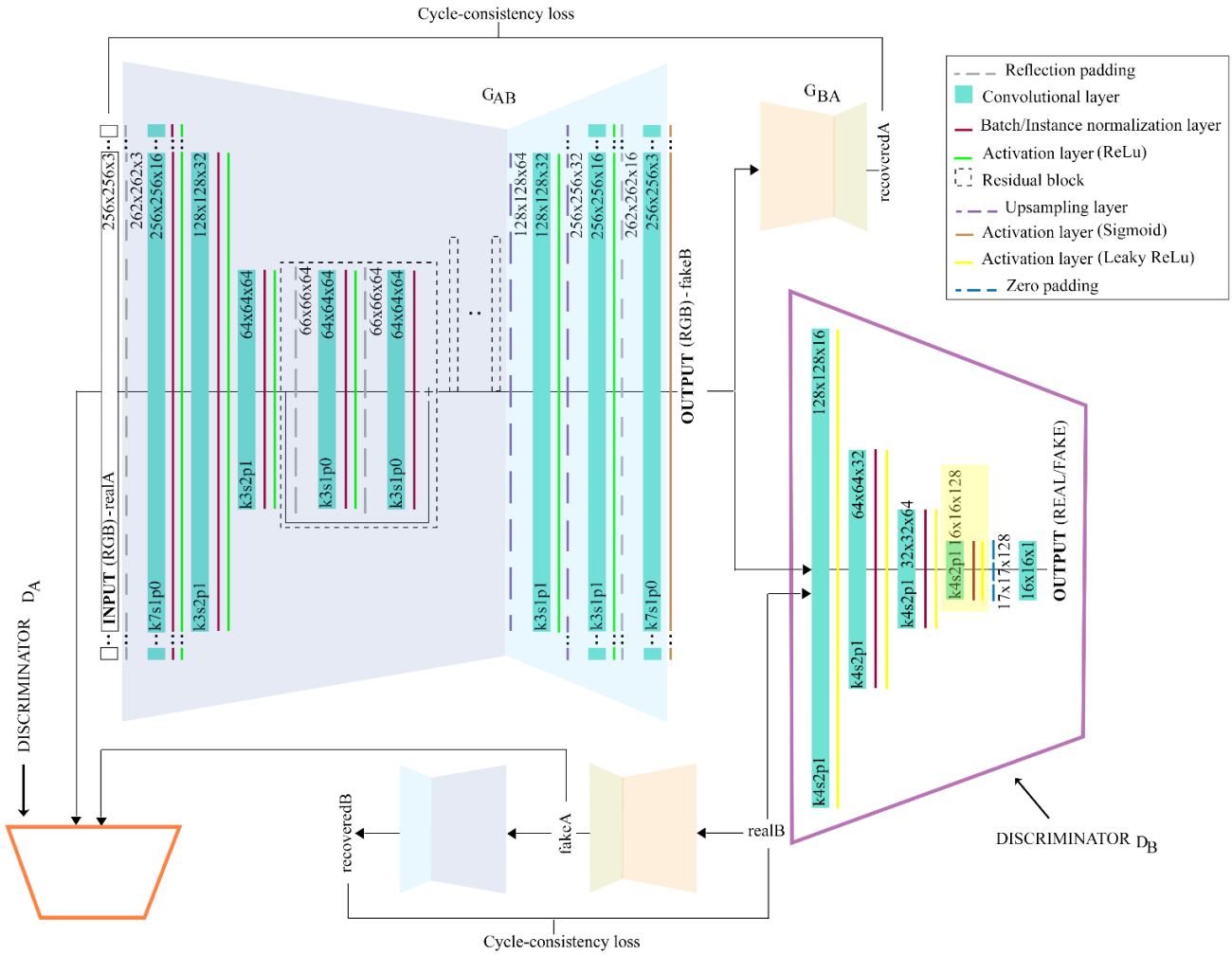


Figure 5.4. Architecture of the CycleGAN-based model. In the first version, nine residual blocks were used in the generator, while in the second and third, three. The yellow highlighted layers in the discriminator were removed in the second and third versions.

Table 5.2. UDA training details

Training details	GIN1	GIN14	GBN14	VIN1	VIN14	VBN14
Epochs	200	80	80	200	80	80
Batch size	1	14	14	1	14	14
Patch size	256	256	256	256	256	256
Training steps	1312	1312	1312	1092	1092	1092
Trainable params G	715,651	272,515	272,515	715,651	272,515	272,515
Trainable params D	175,089	43,057	43,057	175,089	43,057	43,057
Training time	8	26	24	7	21	20

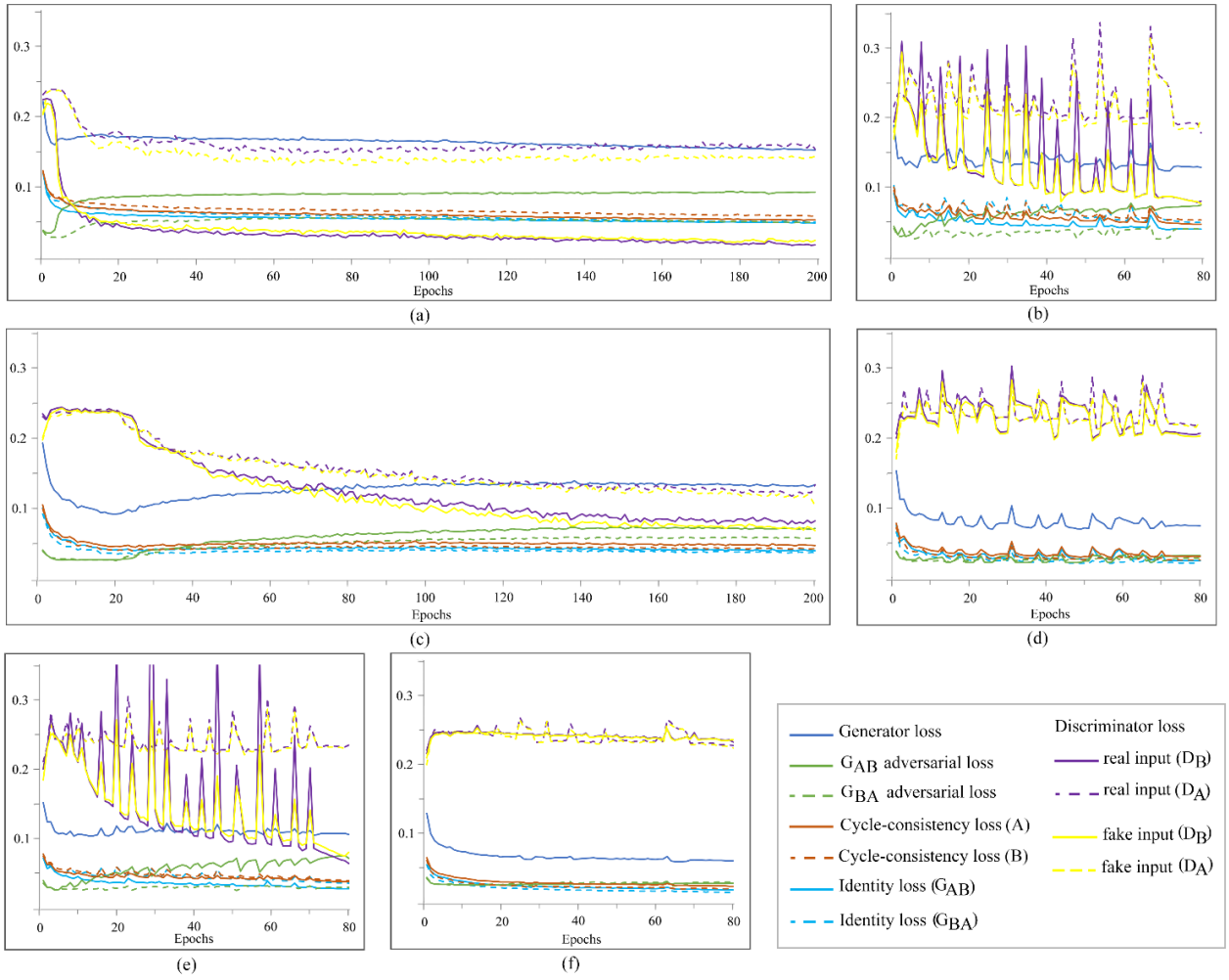


Figure 5.5. Loss function values during training of the CycleGAN-based model. (a) GIN1, (b) GIN14, (c) VIN1, (d) VIN14, (e) GBN14, (f) VBN14

5.3.2.4 CycleGAN-based model implementation

In the second stage of the methodology, UDA was performed by implementing three versions of a model based on CycleGAN which is very popular in unpaired ITIT. The goal of UDA was to produce closer radiometrically images. It was performed on the G1/G2 and V1/V2 pairs because they were collected from the same geographic region and in the same month to avoid seasonal changes. The architecture of the three versions is shown in Figure 5.4.

In the first and second versions, IN was employed with batch sizes 1 and 14 (maximum capacity of the available computer memory) respectively. In the third version, BN was employed with batch size 14. All three versions were trained on the G1/G2 (GIN1, GIN14, GBN14) and the V1/V2 (VIN1, VIN14, VBN14) RGB pair of images.

Three convolutional layers and nine residual blocks formed the first version generator encoder, and two Upsampling and three convolutional layers formed the generator decoder. ReLU was selected as the activation function across the network and Sigmoid as the activation function of the output layer. The PatchGAN discriminator consisted of five convolutional layers and Leaky ReLU was selected as the activation function.

In the second and third versions, the difference compared to the first version architecture was the use of three residual blocks instead of nine in the generator, and the removal of a convolutional layer (shown in yellow highlight) in the discriminator to alleviate the computational load. The learning rate for the first version was set constant to $2x10^{-4}$ for the first 100 epochs and then linear decay to zero was implemented. For the second and third versions, the learning rate decay was implemented for the last ten epochs.

The training details for the three versions are shown in Table 5.2. The inference time for a batch of 32 images for the first version was 0.113 s and for the second and third was 0.073 s. In Figure 5.5 graphs of the loss functions for the generator and the discriminator are depicted during training. To make the graph more easily perceptible, the generator loss and the G_{AB} and G_{BA} adversarial losses were divided by ten. Also, in our case, "A" represents the G1 or the V1 image, and "B" represents the G2 or the V2 image. The selection of the final weights was based on the lowest Cycle-consistency loss along with observing image samples. In the following comments, we refer to the G1/G2 trained generators as GG_{AB} and GG_{BA}, and the V1/V2 trained generators as VG_{AB} and VG_{BA}.

It is observed that the behavior of the identity losses is similar for GG_{AB}, $GG_{BA,}$ and VG_{AB}, VG_{BA} concerning the values, with slightly lower values in the V1/V2 pair. Smoother lines are observed in the BN14 training compared to IN14. Similar are the observations for the Cycle-consistency loss. For the generator adversarial losses, the values in the V1/V2 training are similar to GG_{BA}, and lower in all three models compared to the GG_{AB}. Concerning the discriminator losses, GD_B shows lower values than GD_A, in contrast to the behavior of the generator adversarial losses as expected.

5.3.2.5 Effect of UDA on NIR prediction

In the third stage predictions were made and evaluated by the G1BN model for the outputs of the CycleGAN-based model for G2 and V2.

5.4 Results and discussion

5.4.1 Paired ITIT – First stage

The results of the paired ITIT were evaluated quantitatively and qualitatively. For the quantitative evaluation, the RMSE (Equation (5.12)) and the structural similarity (SSIM) (Equation (5.13)) [347] were calculated between the predicted NIR and the ground-truth with kernel size 7x7. It is noted that RMSE is sensitive to spectral information, while SSIM is sensitive to geometry.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (5.12)$$

$$\text{SSIM} = \frac{(2\mu_y\mu_{\hat{y}} + c_1)(2\sigma_{y\hat{y}} + c_2)}{(\mu_y{}^2 + \mu_{\hat{y}}{}^2 + c_1)(\sigma_y{}^2 + \sigma_{\hat{y}}{}^2 + c_2)} \quad (5.13)$$

where $c_1 = (k_1 L)^3$, $c_2 = (k_2 L)^3$, L: the dynamic range of the pixel values, k_1=0.01, and k_2=0.03.

5.4.1.1 Quantitative evaluation - Mean values

RMSE and SSIM were estimated for the three cGANs (section 5.3.2.2) in eight implementations in total (G1BN, G1IN, G1INRB, T2BN, allBN, allIN, allINRB, allINRBAt) for each of the eight images of the dataset as a whole which included all land cover categories (all) and for three separate categories: impervious (e.g. buildings, roads), vegetation (forest, crops) and ground. The three categories in the images were delineated by masks that were created in a graphics editor [348]. At first, vegetation was detected by the normalized difference vegetation index (NDVI) [349] and then the ground category was manually masked. The remaining area constituted the impervious category. The range of the image values was [0, 1].

The evaluation scores are presented in Table 5.3 and Table 5.4. The mean score values are depicted for each image as well as for the whole dataset (total mean). The bold font indicates the best values for each image. Regarding the evaluation of the dataset as a whole, from the total mean values, it can be noticed that the IN versions trained on the whole dataset (allIN, allINRB, allINRBAt) outperformed the respective BN (allBN) on the total mean RMSE, while slight differences existed in the total mean SSIM. Also, allINRB performed slightly better on

Table 5.3. Evaluation scores in paired ITIT (NIR prediction - first stage) – all/impervious

		all								impervious							
		G1BN	G1IN	G1INRB	T2BN	allBN	allIN	allINRB	allINRBAt	G1BN	G1IN	G1INRB	T2BN	allBN	allIN	allINRB	allINRBAt
RMSE←	G1	**0.065**	0.070	0.067	0.145	0.100	0.089	**0.085**	0.089	**0.056**	0.062	0.060	0.130	0.091	0.079	**0.075**	0.081
	G2	**0.084**	0.091	0.093	0.144	0.099	0.080	**0.076**	0.078	**0.069**	0.082	0.078	0.132	0.095	0.076	**0.071**	0.076
	R1	**0.114**	0.124	0.125	0.152	0.087	0.071	**0.072**	0.072	**0.094**	0.108	0.106	0.146	0.071	**0.069**	0.077	0.078
	R2	0.129	0.115	**0.111**	0.220	0.096	0.068	**0.066**	0.070	0.129	0.111	**0.110**	0.206	0.084	**0.066**	0.067	0.072
	T1	0.186	0.155	0.157	**0.106**	0.139	0.075	**0.072**	0.074	0.180	0.191	0.178	**0.112**	0.201	**0.092**	0.093	0.097
	T2	0.205	0.200	0.183	**0.083**	0.134	0.095	**0.088**	0.092	0.185	0.180	0.163	**0.087**	0.187	0.104	**0.095**	0.102
	V1	**0.096**	0.099	0.114	0.158	**0.067**	0.074	0.073	0.076	**0.089**	0.091	0.108	0.137	**0.059**	0.066	0.068	0.072
	V2	0.133	**0.105**	0.108	0.174	0.076	0.076	**0.075**	0.079	0.114	**0.092**	0.094	0.149	**0.065**	0.069	0.069	0.073
	total mean	0.127	**0.120**	**0.120**	0.148	0.100	0.078	**0.076**	0.079	0.115	0.115	**0.112**	0.137	0.107	0.078	**0.077**	0.081
SSIM→	G1	**0.874**	0.865	0.873	0.716	**0.842**	0.829	0.829	0.824	**0.896**	0.881	0.894	0.730	**0.865**	0.856	0.853	0.842
	G2	**0.850**	0.847	0.847	0.740	**0.877**	0.872	0.873	0.873	**0.883**	0.864	0.876	0.750	**0.881**	**0.881**	**0.881**	0.873
	R1	0.742	0.775	**0.788**	0.731	**0.876**	0.865	0.868	0.869	0.843	0.833	**0.847**	0.762	**0.918**	0.905	0.888	0.880
	R2	0.799	**0.829**	0.823	0.688	0.905	0.900	**0.907**	0.893	0.847	**0.871**	0.864	0.736	**0.933**	0.925	0.928	0.908
	T1	0.674	0.732	0.742	**0.777**	0.846	**0.863**	0.860	0.850	0.641	0.645	0.672	**0.713**	0.726	**0.777**	0.766	0.742
	T2	0.647	0.682	0.686	**0.791**	**0.818**	0.776	0.776	0.785	0.614	0.638	0.662	**0.774**	0.743	0.752	**0.773**	0.742
	V1	0.824	**0.832**	0.830	0.752	**0.911**	0.880	0.879	0.874	0.868	**0.875**	0.865	0.786	**0.940**	0.918	0.911	0.903
	V2	0.759	0.816	**0.821**	0.748	**0.908**	0.878	0.875	0.868	0.821	0.862	**0.868**	0.789	**0.937**	0.915	0.906	0.896
	total mean	0.771	0.797	**0.801**	0.743	**0.873**	0.858	0.858	0.854	0.802	0.809	**0.818**	0.755	**0.868**	0.866	0.863	0.848

Table 5.4. Evaluation scores in paired ITIT (NIR prediction - first stage) – vegetation/ground

		vegetation								ground							
		G1BN	G1IN	G1INRB	T2BN	allBN	allIN	allINRB	allINRBAt	G1BN	G1IN	G1INRB	T2BN	allBN	allIN	allINRB	allINRBAt
RMSE←	G1	0.094	0.095	**0.089**	0.201	0.128	0.123	**0.116**	0.117	**0.042**	0.048	0.049	0.089	0.079	**0.058**	0.061	0.060
	G2	0.125	**0.121**	0.130	0.193	0.114	0.101	0.097	**0.095**	**0.060**	0.062	0.079	0.092	0.083	**0.050**	0.052	**0.050**
	R1	**0.144**	0.149	0.153	0.166	0.110	0.075	0.066	**0.064**	0.096	0.115	0.102	**0.089**	0.071	0.059	0.060	**0.058**
	R2	0.141	0.133	**0.125**	0.272	0.124	0.079	**0.072**	0.077	0.096	0.086	**0.078**	0.142	0.078	0.044	**0.042**	0.046
	T1	0.187	0.132	0.141	**0.097**	0.084	0.077	**0.073**	**0.073**	0.190	0.178	0.178	**0.124**	0.221	0.045	**0.042**	0.047
	T2	0.220	0.214	0.195	**0.083**	0.110	0.092	**0.087**	**0.087**	0.115	0.129	0.134	**0.064**	0.108	**0.069**	0.070	0.089
	V1	**0.131**	0.139	0.142	0.270	0.105	0.115	**0.102**	0.104	No data							
	V2	0.196	**0.151**	0.156	0.266	0.113	0.100	**0.094**	0.097								
	total mean	0.155	**0.142**	**0.142**	0.193	0.111	0.095	**0.088**	0.089	**0.100**	0.103	0.103	**0.100**	0.107	**0.054**	**0.054**	0.059
SSIM→	G1	0.814	0.812	**0.821**	0.677	**0.771**	0.743	0.756	0.762	**0.908**	**0.908**	0.897	0.739	**0.895**	0.894	0.883	0.880
	G2	0.774	**0.790**	0.779	0.713	0.852	0.830	0.842	**0.854**	0.878	**0.899**	0.872	0.753	0.912	**0.922**	0.906	0.912
	R1	0.599	**0.690**	0.703	0.684	0.815	0.806	0.838	**0.853**	0.741	0.782	**0.803**	0.752	**0.902**	0.879	0.880	0.881
	R2	0.702	0.743	**0.744**	0.585	0.845	0.847	**0.867**	0.858	0.834	**0.865**	0.839	0.743	**0.934**	0.921	0.917	0.914
	T1	0.665	0.755	0.765	**0.837**	**0.909**	0.890	0.898	0.898	0.744	**0.777**	0.766	0.680	0.814	**0.895**	0.872	0.852
	T2	0.656	0.697	0.691	**0.801**	**0.849**	0.782	0.772	0.803	0.757	**0.797**	0.786	0.771	**0.885**	0.869	0.867	0.839
	V1	0.605	0.609	**0.653**	0.579	**0.762**	0.687	0.719	0.722	No data							
	V2	0.537	**0.655**	0.658	0.602	**0.803**	0.746	0.766	0.768								
	total mean	0.669	0.719	**0.727**	0.685	**0.826**	0.791	0.807	0.815	0.810	**0.838**	0.827	0.740	0.890	**0.897**	0.887	0.880

the total mean RMSE than allIN. Granada, Rhodes, and Venice showed RMSE <0.13 in the models trained on G1 and <=0.1 in the models trained on the whole dataset. Thus, cGANs are promising not only in in- domain data (training set) but also in out-domain. As expected, the models trained on the whole dataset showed better scores than those trained on G1 or T2 on different geographical regions (i.e. G1: R1, R2, T1, T2, V1, V2/ T2: G1, G2, R1, R2, V1, V2).

In the impervious category, Granada, Rhodes, and Venice showed RMSE <0.13 in the versions trained on the whole dataset and G1. In the G1 training, G1INRB performed slightly better on total mean RMSE than G1IN. The highest RMSE and the lowest SSIM for the versions trained on G1 were produced in T1 and T2 because the Tønsberg's urban fabric was significantly different from Granada's. For the models trained on the whole dataset, it should be noted that only the BN version (allBN) was affected by the domain gap regarding spectral information in Tønsberg compared to Granada, Rhodes, and Venice, because of its low representation in the dataset. However, the spatial performance was similar as shown by the SSIM score in Tønsberg. Concerning the normalization, allIN, allINRB, and allINRBAt performed better in total mean RMSE than allBN.

In the vegetation category, higher total mean RMSE and lower total mean SSIM were displayed compared to the impervious category. Thus, predicting NIR in vegetation was a more challenging problem. Still, RMSE in Granada, Rhodes, and V1 remained $\lesssim 0.15$. The models trained on G1 showed the highest RMSE in T2. In addition, G1BN displayed high RMSE in T1 and V2, contrary to G1IN and G1INRB. Regarding the versions trained on the whole dataset, allBN total mean RMSE values were more resembling to allIN, allINRB, and allINRBAt compared to the impervious category, and allINRB performed better on total mean RMSE than allIN. Finally, T2BN overall spatially performed similarly to G1BN but spectrally worse, as in the impervious category.

Finally, in the ground category, the total mean RMSE values were the lowest of the three categories and the total mean SSIM values were the highest. It is noted that this category was not taken into account in V1 and V2 because there was no data of this kind. All versions showed RMSE <0.12 except for G1BN and allBN on T1. In addition, total mean RMSE values were lower in the IN models that were trained on the whole dataset.

5.4.1.2 Quantitative evaluation - Boxplots

Except for quantitatively assessing the paired ITIT by the mean score values, boxplots were also created (Figure 5.6). In the boxplots, the first (Q1), second (Q2), and third (Q3) quartiles are depicted. The conclusions produced by observing Table 5.3 and Table 5.4 are reinforced by the boxplots. It can be observed that in the evaluation of the dataset as a whole, allIN, allINRB, and allINRBAt showed overall the lowest RMSE values. In addition, allINRB performed slightly better than allIN.

The above-mentioned models were also significantly superior in T1 and T2 in the impervious category in RMSE. As already mentioned, the higher RMSE and lower SSIM values compared to the urban fabric, are easily perceptible in the vegetation boxplots. In addition, the superior performance of G1IN and G1INRB over G1BN on T1 and V2 can be noticed. Concerning the models trained on the whole dataset, allINRB performed slightly better than allIN. Finally, the lower performance of the models trained on G1 and allBN on T1 is visible in the ground category. The lower SSIM scores of G1BN compared to G1IN and G1INRB should also be noted.

5.4.1.3 Qualitative evaluation

The qualitative evaluation was conducted by visual interpretation and it is complementary to the quantitative. Samples of pseudo-color composites of the NIR predictions are displayed in Figure 5.7 and Figure 5.8. These figures show samples for each of the eight images (G1, G2, R1, R2, T1, T2, V1, V2) for all the trained versions (G1BN, G1IN, G1INRB, T2BN, allBN, allIN, allINRB, allINRBAt).

In G1, in the versions trained on the whole dataset, allINRBAt seems to be less affected by the stitching noise between the patches. In the urban structures, there are high similarities for all versions with T2BN being more divergent. In vegetation, G1INRB shows values closer to the ground-truth, while T2BN predicted NIR seems to have the greatest divergence from the true NIR, followed by allBN. In addition, the IN versions show higher vegetation NIR values compared to the BN.

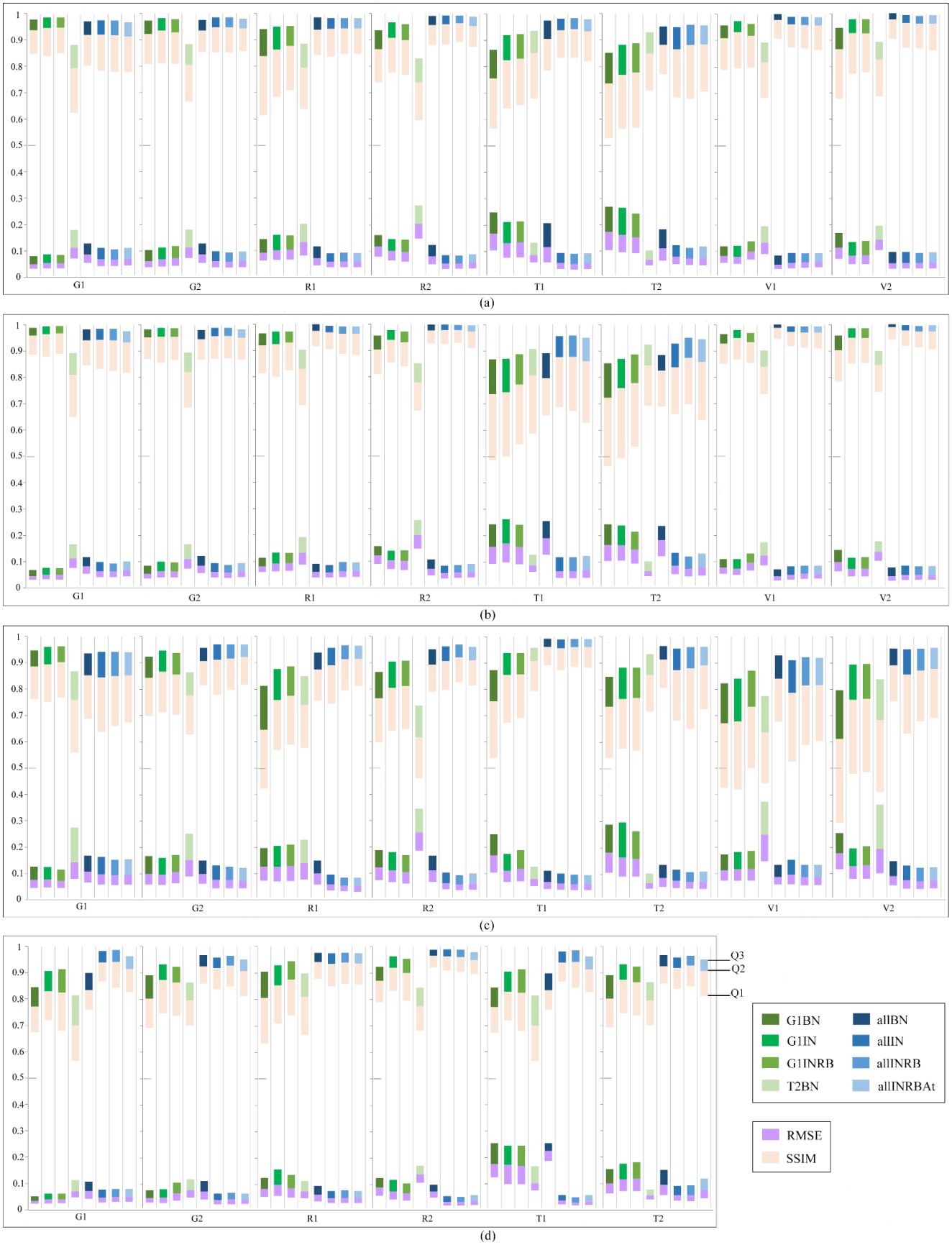


Figure 5.6. Boxplots of the evaluation scores in paired IT. (a) all, (b) impervious, (c) vegetation, (d) ground

Figure 5.7. Samples of pseudo-color composites of the NIR predictions in the paired ITIT (Granada, Rhodes) – first stage. Red color is assigned to the NIR band, green color to the RED band, and blue color to the GREEN band.

Figure 5.8. Samples of pseudo-color composites of the NIR predictions in the paired ITIT (Tønsberg, Venice) – first stage. Red color is assigned to the NIR band, green color to the RED band, and blue color to the GREEN band.

In G2, the conclusions are similar to G1. As in G1, allINRBAt is unaffected by the stitching noise between the patches. In addition, the urban fabric shows a high visual resemblance between all models. Concerning vegetation, G1IN and G1INRB are spectrally closer to the true NIR than G1BN, and the worst performance is shown by the lower NIR values in T2BN. Finally, in the ground category, G1BN and the IN versions trained on the whole dataset have the highest performance.

In R1, all methods show visually similar outputs in the impervious category, and the NIR information is more accurately expressed by the IN versions trained on the whole dataset in the vegetation category. Moreover, G1IN and G1INRB show higher vegetation NIR values compared to the ground-truth, while T2BN NIR values are lower.

In R2, concerning the patch stitching noise, allINRBAt is the most homogenous as also observed in G1 and G2. The urban structures seem to be visually similar as well, with T2BN showing slightly more different values. In vegetation, the highest dissimilarities from the ground-truth are displayed in T2BN where NIR values are much lower. Besides the above, allIN, allINRB, and allINRBAt show the most accurate NIR predictions in the vegetation and ground categories.

In T1, T2BN as well as the IN versions trained on the whole dataset present the best performance in the urban fabric. However, there is spatial noise. In vegetation, the best performance is expressed by allINRBAt and the worst by G1BN. Finally, the IN models trained on the whole dataset display the highest similarity in the ground category followed by T2BN.

In T2, the conclusions are similar to T1 in the impervious category where T2BN, allIN, allINRB, and allINRBAt show the best performance but artifacts are present. Regarding vegetation, the performance of the versions trained on the whole dataset and T2BN seems similar and is better than the rest. Finally, in the ground category, the highest difference from the ground-truth is noticed in G1INRB.

In V1, in the impervious category, all versions resemble each other and T2BN shows slightly higher differences compared to the ground-truth. T2BN also shows the lowest performance in the vegetation areas.

In the final study region (V2), all versions show outputs that look alike regarding the urban structures, while in the vegetation category, T2BN is the most dissimilar compared to the ground-truth followed by G1BN.

5.4.2 Unpaired ITIT - Second stage

The results of UDA on the G1/G2 and V1/V2 image pairs were evaluated quantitatively by calculating the RMSE and the SSIM and qualitatively by visual interpretation. The quantitative scores were based on corresponding samples from the image pairs for the impervious, vegetation, and ground categories. They are presented in Table 5.5 for the three versions trained on the G1/G2 (GIN1, GIN14, GBN14) and V1/V2 (VIN1, VIN14, VBN14) pairs. In addition, output samples of the unpaired ITIT are depicted in Figure 5.9.

Table 5.5. Evaluation scores in unpaired ITIT (second stage)

		G1/G2									V1/V2						
		impervious			vegetation			ground			impervious			vegetation			ground
		B	G	R	B	G	R	B	G	R	B	G	R	B	G	R	
RMSE←	original	0.047	0.034	0.035	**0.022**	**0.016**	0.014	0.023	0.028	0.068	0.038	0.030	0.050	**0.039**	**0.064**	0.060	No data
	IN1	**0.033**	0.056	0.057	**0.019**	0.023	0.021	0.038	**0.028**	**0.038**	0.120	0.124	0.078	**0.019**	**0.047**	**0.022**	
	IN14	**0.038**	0.043	0.041	**0.016**	0.019	0.020	0.032	0.035	**0.043**	0.139	0.155	0.091	**0.018**	**0.029**	**0.021**	
	BN14	**0.035**	**0.029**	**0.034**	**0.022**	**0.015**	**0.012**	0.034	0.039	**0.053**	0.066	0.065	0.056	**0.020**	**0.044**	**0.030**	
	HM	**0.042**	**0.033**	**0.034**	**0.022**	**0.015**	0.017	**0.019**	**0.028**	**0.045**	0.052	0.055	0.052	0.045	**0.034**	**0.026**	
SSIM→	original	0.745	0.881	0.962	**0.681**	**0.911**	0.888	-0.066	0.300	0.790	0.760	0.606	1.009	**0.185**	**0.457**	0.031	No data
	IN1	**0.785**	0.667	0.904	**0.781**	0.875	0.749	**0.525**	**0.783**	**0.928**	0.520	0.499	0.805	**0.354**	0.379	0.423	
	IN14	0.715	0.759	0.945	**0.909**	0.883	0.857	**0.665**	**0.676**	**0.830**	0.571	0.533	0.878	**0.323**	**0.498**	0.266	
	BN14	**0.773**	**0.898**	**0.963**	**0.699**	0.907	**0.928**	-0.314	0.037	0.734	0.656	0.570	**1.200**	**0.333**	**0.545**	0.162	
	HM	**0.753**	**0.885**	**0.965**	**0.695**	**0.916**	0.901	-0.080	**0.305**	0.783	0.720	**0.607**	**1.056**	**0.188**	**0.505**	**0.008**	

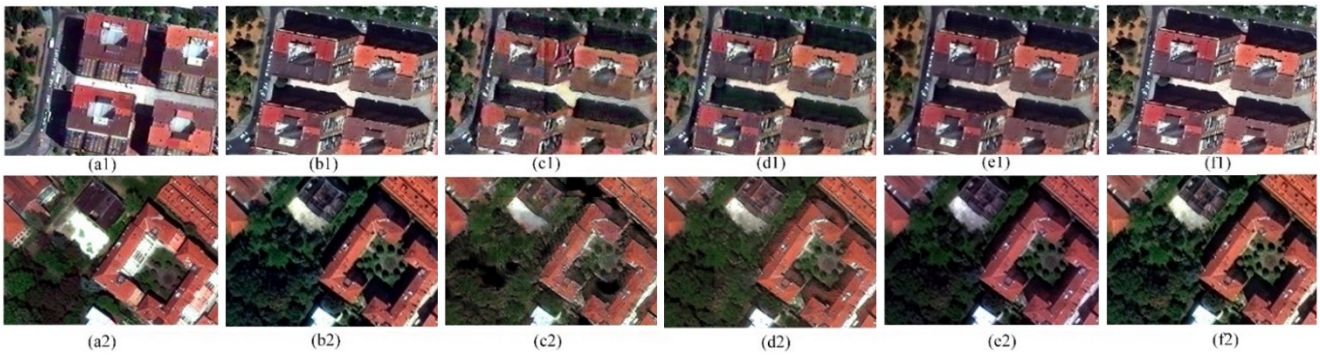


Figure 5.9. Output samples of the unpaired ITIT (natural color composites). The top row shows the Granada images and the bottom row shows the Venice images. (a1-a2) original G1/V1 image, (b1-b2) original G2/V2 image, (c1-c2) IN1, (d1-d2) IN14, (e1-e2) BN14, (f1-f2) HM.

For Granada, the scores were at first calculated between the original G1 and G2 images (Table 5.5 - original), and then between the original G1 and the predicted G2 model outputs (Table 5.5 - IN1, IN14, BN14). A similar calculation was conducted for Venice, where the scores were at first calculated between the original V1 and V2 images, and then between the original V1 and the predicted V2 model outputs.

The goal of this stage was to increase the spectral similarity of G2 in terms of G1 and the spectral similarity of V2 in terms of V1. Then (third stage), the domain-adapted G2 and V2 will act as input on G1BN to test the effect of UDA on the NIR prediction. The logic is based on the fact that G1BN performed better in the first stage on G1 compared to G2, and on V1 compared to V2. It is also noted that G1BN was preferred over G1IN and G1INRB because it showed a higher difference in the performance of V1 compared to V2.

Besides comparing the different GAN versions, the Histogram Matching (HM) method [350] was also implemented and added to the evaluation (Table 5.5 – HM). In Table 5.5 the bold font indicates the scores that outperformed the original G2/V2 images. Gray highlight shows the best score.

By observing the quantitative scores in Table 5.5, it can be noticed that in Granada, BN14 and HM showed slightly lower RMSE in the urban structures in all three bands, and IN14 and IN1 in the blue band. In the visual interpretation (Figure 5.9), the impervious category seems visually similar in all outputs, with IN1 showing the lowest spatial performance (also observed in SSIM – green/red). In Venice, none of the methods produced in the urban fabric lower RMSE than the original. It should be noted, however, that BN14 and HM showed values closer to the original. Thus, a safe conclusion cannot be reached on the UDA of the impervious category from this study alone, and experiments on more data are required for future work.

In the vegetation category, by putting more focus on the green band, similar RMSE/SSIM values were observed among all outputs in Granada with BN14 and HM being closer to the original. Moreover, all outputs are similar in the visual interpretation. Nevertheless, in Venice, all outputs displayed improved RMSE compared to the original, with IN14 and HM showing the best values and also the highest visual similarity with V1. IN1 had the lowest SSIM (lowest spatial performance) and BN14 the highest. Thus, IN14, BN14 networks, and HM could be promising for the UDA of the vegetation category.

Finally, in the ground category, which was investigated in Granada, by putting more focus on the red band, it can be noticed that all methods outperformed the original RMSE, with IN1 showing the highest performance both in RMSE and in SSIM followed by IN14. The visual similarity of IN1 with G1 was also the highest, followed by IN14, HM, and BN14. Thus, IN1 and IN14 seem to be the more promising for the UDA of the ground category.

As a general conclusion, BN14 shows the best performance among the GANs overall because it improves the RMSE in the impervious and ground categories of Granada, as well as in the vegetation category of Venice. Moreover, BN14 shows the closest RMSE in vegetation in Granada and in the impervious surfaces in Venice.

Table 5.6. Evaluation scores of the third stage (NIR prediction)

	RMSE↓								SSIM↑							
	G2				V2				G2				V2			
	all	impervious	veg	ground	all	impervious	veg	ground	all	impervious	veg	ground	all	impervious	veg	ground
1st stage	0.084	0.069	0.125	0.060	0.133	0.114	0.196		0.850	0.883	0.774	0.878	0.759	0.821	0.537	
IN1	0.163	0.147	0.222	0.107	0.171	0.164	0.194	No data	0.416	0.472	0.336	0.361	0.477	0.513	0.357	No data
IN14	0.144	0.122	0.207	0.098	0.146	0.139	**0.169**		0.538	0.598	0.442	0.502	0.616	0.662	0.460	
BN14	0.100	0.089	0.142	0.063	**0.108**	**0.101**	**0.130**		0.700	0.744	0.607	0.715	0.747	0.793	**0.586**	
HM	0.093	0.070	0.152	0.063	**0.123**	0.114	**0.156**		0.842	0.881	0.749	0.879	**0.790**	**0.838**	**0.615**	

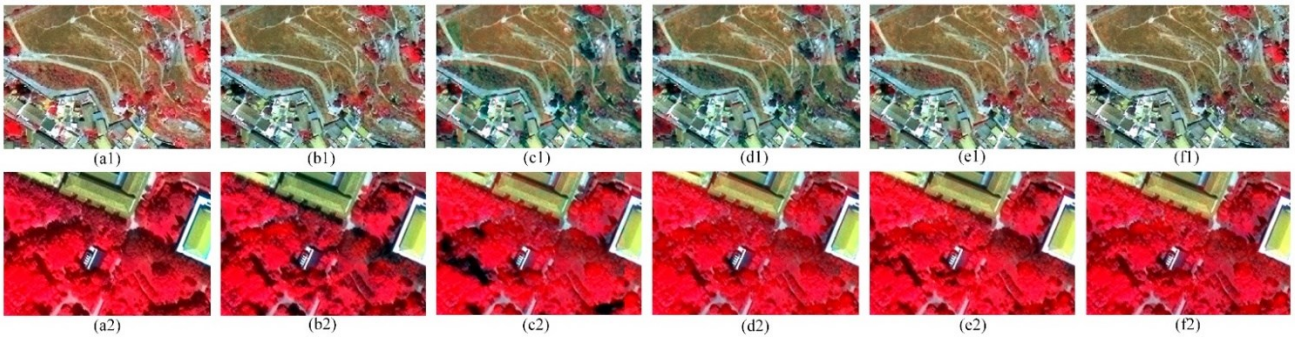


Figure 5.10. Samples of pseudo-color composites of the NIR predictions in the third stage. The top row shows the Granada images and the bottom row shows the Venice images. (a1-a2) original G2/V2 image, (b1-b2) First stage, (c1-c2) IN1, (d1-d2) IN14, (e1-e2) BN14, (f1-f2) HM.

5.4.3 Effect of UDA on NIR prediction – Third stage

In the third and final stage, the effect of UDA on the NIR prediction was investigated for the G2 and V2 images, by re-implementing the G1BN version. G1BN was selected because of the high difference in the RMSE values in the vegetation category between V1 (0.131) and V2 (0.196) (Table 5.4). As in the previous stages, quantitative and qualitative evaluation was performed. The quantitative scores (RMSE, SSIM) are depicted in Table 5.6. In addition, samples of pseudo-color composites of the NIR predictions are displayed in Figure 5.10.

By observing the RMSE/SSIM values in Table 5.6, it can be noticed that in the G2 impervious category, the NIR prediction was not improved. It should be noted however that BN14 and HM (best scores in the second stage) had RMSE/SSIM values closer to the ones of the first stage (before UDA). In the visual interpretation, the urban fabric seems visually similar to IN displaying some patch stitching noise. In V2, BN14 and HM displayed slightly better RMSE and SSIM scores respectively compared to the ones of the first stage. That was also the visual conclusion with BN14 and HM exhibiting slightly higher NIR values than the original V2. As in the second stage, further experiments should be performed in future work, to produce conclusive results regarding the NIR prediction in the urban fabric.

In the vegetation category, in G2 the NIR prediction was not improved in any of the outputs, but BN14 and HM showed scores closer to the ones of the first stage. In V2, there was significant improvement in BN14 followed by HM and IN14. The above conclusions are also visually verified (Figure 5.10). Thus, our study shows that improved scores in the UDA could lead to improved NIR predictions.

Finally, in the ground category, even though in the second stage all outputs outperformed the original RMSE scores, in the third stage there was no improvement. It should be also noted that BN14 and HM showed similar values to the original. This conclusion should be further tested in the ground category in future work.

As a general conclusion, UDA seems to be meaningful for the improvement of the NIR prediction when there are high radiometric/spectral differences between the two RGB domains (e.g. the vegetation category in V1 and V2 where high dissimilarity was caused by shadows). It should be also noted that BN14 which showed the most positive results among the GANs in the second stage, outperformed all methods in the third stage.

5.5 Conclusions

In this study, a three-stage GAN framework to generate NIR information in VHR RGB data that do not belong to the domain of the training set (different date/region/satellite) was proposed. In the developed methodology both paired and unpaired data were exploited. The experiments were evaluated on three main land cover categories (impervious surfaces/urban fabric, vegetation, ground) in heterogeneous bi-temporal VHR data of different regions and sensors.

In the first stage, three cGANs of different architecture in eight implementations in total were trained on paired RGB VHR imagery to predict the NIR band. The Pix2Pix model along with residual mapping, attention modules, and normalization techniques were explored. In the second stage, unpaired image-to-image translation (ITIT) in the framework of unsupervised domain adaptation (UDA) was performed by training three versions of a model based on CycleGAN. For the model training, bi-temporal data that were collected in the same month (to avoid seasonal changes) from two study areas were employed. Using data with the above characteristics promotes the spectral fidelity required to predict adequate NIR information. Moreover, the performance of different batch sizes and normalization methods was analyzed. In the third and final stage, the effect of UDA on the NIR prediction was investigated.

The experiments of the first stage showed that: a) the paired ITIT cGANs produced satisfactory NIR predictions on all main land cover categories not only in the training set (in-domain) but also in data from different regions/dates/sensors (out-domain) provided that the domain gap was not significantly high, b) the instance normalization (IN) technique outperformed batch normalization (BN), especially on data with low representation on the training set, c) employing residual blocks slightly enhanced the quantitative evaluation scores, and d) employing attention modules reduced the patch stitching noise. The experiments of the second stage showed promising results for the UDA of the vegetation and ground categories. BN produced the most positive results overall among the GANs followed by the Histogram Matching method. Finally, in the third stage, it was concluded that the unpaired data experiments were able to enhance the NIR prediction in the high vegetation category when high dissimilarities, caused by different satellite view angles, existed in the respective RGB domains. BN outperformed all other methods in improving the NIR prediction.

For future work, further experiments on broader datasets could be implemented, to produce conclusive results on the effect of UDA on the NIR prediction of the impervious and ground categories. In addition, the prediction of shortwave-infrared (SWIR) information could be explored.

CHAPTER 6

Conclusions and future work

6.1 Overall conclusions

This PhD thesis investigated the capabilities of different types of ANNs in four Remote Sensing applications: cloud masking in Sentinel-2 (S2) data, VHR change detection (CD), marine plastic litter detection through image fusion and RGB-to-NIR image-to-image translation (ITIT).

In the first application (Chapter 2), three studies were performed that focused on separating clouds from challenging non-cloud objects.

In the first cloud masking study, a multi-layer perceptron (MLP) architecture was implemented that outperformed state-of-the-art rule-based and muti-temporal methods in the separation of clouds from deep water spectra with noise and sunglint. The performance on the directional reflectance effects caused by the S2 broad range of viewing geometries was also investigated (periodic noise). A cloud masking spectra dataset that was made available at the time of the experiment was exploited [43]. Since the above-mentioned dataset was considered inadequate to represent deep water spectra with high noise levels and sunglint, a relevant manual dataset was created for the purpose of the study and was made publicly available. In four configurations, various algorithms that prevent overfitting were tested but slightly affected the output. The effect on the MLP output of applying feature scaling by using the parameters of the test set instead of the training set was also investigated and it was shown that it can be positive. The overall accuracy of the proposed MLP was > 0.92 on the test set, it produced robust and time-efficient results unaffected by the challenging cases and the masks retained the natural cloud shapes. Although a small omission error was overall observed, the results were acceptable in all cases. In addition, a strong correlation (Phi coefficient [173]) was observed between the MLP outputs and MAJA [110] masks (0.65) followed by Fmask (0.58) [18][22]. Besides the above, a measure was defined that indicated that the most important bands in mitigating spectra with noise and sunglint are the ones corresponding to 444, 945, 1374, and 1614 nm. The first two bands are typically used in atmospheric correction, the third for cloud separation in turbid waters and the fourth is the cirrus band.

In the second cloud masking study, a novel fine-tuning methodology for self-organizing maps (SOMs) was developed that managed to correct the misclassified predictions of bright non-cloud spectra in land areas. The proposed fine-tuning approach, which is applied to the output of the non-fine-tuned network (SOM1), is task-independent and requires only small amounts of data. In addition, it lacks the need for further training. It is performed by directly locating the neurons that correspond to the incorrectly predicted bright non-cloud objects and altering their labels. This process was chosen over the common practice of further training the network by feeding the labeled data (supervised training) since it was considered faster, simpler, and more efficient. Further training would probably also require more data than those available. A median and a dilation filter were performed as the final step of fine-tuning to compensate both for remaining omission and commission errors caused by the fact that the altered neurons represented also a percentage of cloud pixels. SOM1 was trained on all categories of the spectra dataset proposed in [43] (the largest available at the time of the experiment). The evaluation on the training dataset by several visualization techniques led to the interpretation of the SOM behavior and the evaluation on an independent image dataset [44] (first publicly available) showed that the fine-tuned version produced an average commission error of less than 1% on the cases of bright non-cloud objects. The respective values for the SOM before fine-tuning were ~3%, for Fmask ~4%, and for Sen2Cor [109] ~8%.

In the third cloud masking study, a patch-to-pixel CNN was created that successfully detected semi-transparent clouds and separated bright clouds from bright non-cloud objects. The model was evaluated on the first publicly available annotated cloud masking image dataset [44], thus the opportunity for a robust and objective evaluation was provided. The activation functions. ReLU and Leaky ReLU, as well as batch normalization (BN)

were investigated with the version using the Leaky ReLU activation function showing slightly higher accuracy in the test set than the version that used ReLU. The version that used BN produced the least accurate and most unstable results. It was shown that the CNN produced exceptional results (accuracy ~98%) both in the training and the test set compared to the state-of-the-art threshold-based methods (Fmask, Sen2Cor) (accuracy ~92%) which performed less favorably. In more detail, the CNN managed to detect even clouds of very high transparency and successfully separated clouds from snow as well as bright urban and desert areas. Thus, the study further reinforces the value of CNNs in applications where spatial context is very important and shows that an architecture that makes use of a smaller number of layers and feature maps compared to recent deep learning literature, consequently being simpler and more time-efficient, can produce very satisfactory results in cloud masking.

In the second application (Chapter 3), five state-of-the-art deep learning (DL) CD methods were assessed on VHR images with severe co-registration errors. In addition, four automatic co-registration methods were evaluated, covering a wide range of the existing literature, because co-registration is a very important pre-processing step. The study was performed on images depicting four European areas with versatile urban patterns. The challenges included geometric distortions, radiometric differences, and seasonal and vehicle-related changes. The heterogeneity between the training sets and the study data was also a challenge for the supervised methods. It was observed that SIFT [226], ORB [227], and a CNN-based method [228] displayed low co-registration performance, while the Fourier-Mellin Transform [229] output was more satisfactory. However, to achieve the best possible CD result, the input to the CD methods was manually co-registered with a mean RMSE 1.5 – 4 m. Concerning the CD methods, STANet [65] produced satisfactory results in building-related changes which are considered the most important indicator for the assessment of urban development. Its commission error was smaller (mean Fscore: ~0.7) and mainly attributed to remaining co-registration issues. Its performance can be attributed to the spatial attention mechanism in combination with a large professionally annotated dataset. A final contribution to this application was the creation of a novel score that provides a better understanding of the magnitude of the commission error.

In the third application (Chapter 4), two studies were performed that focused on the increase of spatial resolution in either the PRISMA or the S2 satellites through image fusion to facilitate the detection and monitoring of marine plastic litter.

In the first study, the potential of HS satellite imagery in marine plastic litter detection was investigated for the first time through PRISMA data. The study focused on the detection of small-sized targets (≤ 5 m), constructed for the needs of the experiment, which is even more challenging. After defining the required pre-processing steps (fine HS/PAN co-registration, periodic noise elimination), several literature pansharpening approaches were evaluated for their ability to spectrally discriminate plastics from water as well as for their spatial distortions. For the pansharpening approaches, four main categories of conventional methods (component substitution (CS), multiresolution analysis (MRA), hybrid, Bayesian) and three state-of-the-art DL networks (originally proposed in the literature for VHR data) were applied. The best performance was displayed by the PCA-based substitution which efficiently separated plastic spectra from water without producing blurry/duplicate edges or pixelation on the output image. In the DL methods, spatial distortions were observed caused by the large difference between the spatial resolutions of the PAN and the HS bands and the unavailability of ground-truth data. However, the importance of histogram clipping as a pre-processing step was established since a good separation of the random water spectra from the target spectra was achieved. In later research [267], it was also shown that DL pansharpening methods could not outperform conventional methods. In the PCA pansharpened image, it was proven that it is possible to detect plastic targets with size 5.1×5.1 m^2 and 2.4×2.4 m^2 (8% HS pixel coverage), while targets of size 0.6×0.6 m^2 cannot be detected. PET was the most difficult material to discriminate among HDPE, PS, and mixed targets. In the pansharpened plastic spectra, the influence of seawater was preserved and consequently SWIR features (high water absorption) observed in the laboratory and airborne-based spectra [269] were not apparent in the derived spectral signatures. However, by exploiting spectral VNIR characteristics an intersection of the outputs of three novel marine plastic indexes was proposed. It is noted that the pre-processing steps and the conventional pansharpening methods were carried out by M. Kremezi.

In the second study, S2 and WV-3 images were fused to increase the capability of S2 imagery to detect marine plastic targets. The SWIR information, available in S2 and absent in the WV-3, is valuable for the

identification and distinction of plastics from other materials. Five conventional and six DL image fusion approaches were evaluated on artificial plastic targets for their performance in terms of preserving spectral and spatial information. Among the DL approaches, three were created for the purpose of the study (PNN-Siamese, Fusion-GAN, Fusion-ResNet). CNMF [280] showed the best performance overall because it produced an output with clear edges, no blurring, and a relatively favorable impact on the spectral characteristics of the materials. DL methods showed high performance in terms of spectral similarity between the fused and the S2 images. In this regard, Fusion-GAN and Fusion-ResNet outperformed all other fusion methods (non-DL and DL). An important finding was the fact that the VNIR WV-3 information was adequate to produce the most efficient fused output, increasing the likelihood of temporally close acquisitions. Other interesting results were: a) the reinforcement of the significance of the SWIR information in detecting plastic, shown by the superiority of FDI compared to four other indexes when applied to the CNMF image, and b) the observation of dissimilarities in the spectral regions of S2 bands between the signatures of the various plastic materials. It is noted that the conventional image fusion methods and the indexes were implemented by M. Kremezi.

In the fourth and final application (Chapter 5), a three-stage GAN framework to generate NIR in- and out-domain images was proposed where both paired and unpaired data were exploited. The term "out-domain" refers to data that do not belong to the domain of the training set (different satellite/collection date/region). The experiments were evaluated on three main land cover categories (impervious, vegetation, ground) on bi-temporal VHR data collected by various satellite sensors from four European heterogeneous areas. In the first stage, three models of cGANs were trained on paired RGB images to predict the NIR band. The Pix2Pix model along with residual mapping, attention modules, and normalization techniques were explored. In the second stage, unsupervised domain adaptation (UDA/ unpaired ITIT) was employed in two study areas, by training on data collected in the same month (to avoid seasonal changes) three versions of a CycleGAN-based model. Different batch sizes and normalization methods were analyzed. In the third and final stage, the effect of UDA on the NIR prediction was investigated. In the first stage, it was demonstrated that: a) paired cGANs produced adequate NIR predictions even in out-domain cases when the domain gap was not significantly high, b) IN performed better than BN, especially on data with low representation on the training set, c) residual mapping slightly enhanced the quantitative scores, and d) attention reduced the patch stitching noise. In the second stage, promising results were shown for the UDA of the vegetation and ground categories and the BN-CycleGAN-based model produced the most positive results overall followed by Histogram Matching. Finally, in the third stage, it was concluded that UDA improved the NIR prediction in the high vegetation category when high dissimilarities, caused by different satellite view angles, existed in the respective RGB domains. BN exceeded in performance all other methods in improving the NIR prediction.

6.2 Future work

In this section, suggestions for future work based on the conclusions of this PhD thesis are presented for the four applications that were studied.

In the first application (cloud masking in S2 data/ Chapter 2), three studies were performed in order to mitigate the challenging issues.

In the first cloud masking study, an MLP architecture was implemented to separate clouds from deep water spectra with noise and sunglint. One of the interesting findings of the first study was the positive effect of making predictions using the feature scaling parameters of the test set instead of the training set when the test set is not adequately represented by the training set. The possibility of generalizing this finding in other applications could be investigated in future research.

In the second cloud masking study, a novel fine-tuning methodology for SOMs was developed that achieved the correction of misclassified predictions of bright non-cloud spectra in land areas. The proposed fine-tuning methodology is task-independent, simple, time-efficient, requires only a few input data points, and outperformed rule-based state-of-the-art algorithms. Thus, the potential of the method in different scenarios could be investigated in future work, especially for big data analysis where the processing time is crucial. The method could also be tested in datasets with greater availability of ground-truth data and compared with supervised SOM approaches.

In the third cloud masking study, a patch-to-pixel CNN was created that managed the detection of semi-transparent clouds and the separation of bright clouds from bright non-cloud objects. Besides evaluating the predicted cloud masks, an initial observation of the feature maps of the first convolutional layer was carried out in an effort to extract the weights of the kernels based on the importance of the feature map. The creation of a database formed by such kernels could be subject of future work because it could provide crucial features that would act as input to several algorithms outside of the context of neural networks.

In the second application (VHR change detection (CD)/ Chapter 3), five state-of-the-art deep learning (DL) CD methods were assessed on VHR images with severe co-registration errors. The use of spatial attention and large annotated datasets appeared to increase the performance. However, the commission error needs to be improved. This goal could be achieved in future work by the creation of large annotated datasets that are characterized by high co-registration errors.

In the third application (marine plastic litter detection through image fusion/ Chapter 4), image fusion was employed in two studies to increase the spatial resolution in either the PRISMA or the S2 satellites. The ultimate goal was the detection of marine plastic litter.

In the first study, marine litter detection was explored for the first time via HS satellite imagery (PRISMA data). Several conventional and three state-of-the-art DL (originally proposed for VHR data) pansharpening methods were evaluated and an intersection of three novel marine plastic litter indexes was proposed. The best performance was displayed by a PCA-based substitution method and artificial targets with size equal to 8% HS pixel coverage were detected. In the DL methods, spatial distortions were observed caused by the large difference between the spatial resolutions of the PAN and the HS bands and the unavailability of ground-truth data. However, histogram clipping managed to produce satisfactory water-plastic spectral separation. In future research, it would be interesting to conduct further experiments to detect the minimum detectable plastic target. Comparison with other non-plastic floating materials (e.g. floating vegetation and foam) should also be considered in future experiments.

In the second study, S2 and WV-3 images were fused to create an output with both high spatial and spectral resolution, and subsequently exploit the S2 SWIR information which is valuable for the detection of plastic materials. Several conventional and DL supervised image fusion approaches were evaluated spectrally and spatially. CNMF [280] showed the best performance overall, while a GAN- and a ResNet-based model (created for the purpose of the study) displayed superiority in the spectral criterion. An interesting finding was the adequacy of the VNIR WV-3 information in detecting plastics because it increases the likelihood of temporally close acquisitions, given the higher availability of suitable satellites. In addition, spectral dissimilarities were observed between the various plastic materials in the fused products, which gives rise to future research that will focus on separating different types of plastic. In future work, the scalability of the experiment should also be defined. Concerning future directions for the DL image fusion approaches, unsupervised networks based on unmixing approaches should be tested.

In the fourth and final application (RGB-to-NIR ITIT/ Chapter 5), both paired and unpaired data were exploited through a three-stage GAN framework to predict VHR NIR images from heterogeneous RGB data. It was demonstrated that cGANs produced adequate NIR information even in out-domain cases when the domain gap was not significantly high, and that unpaired ITIT improved predictions in the high vegetation category where high dissimilarities, caused by different satellite view angles, existed. For future work, further experiments on broader datasets could be implemented to produce conclusive results on the effect of UDA on the NIR prediction of the impervious and ground categories. In addition, the prediction of shortwave-infrared (SWIR) information could be explored.

References

[1] R. Kline, “Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence,” *IEEE Ann. Hist. Comput.*, vol. 33, no. 4, pp. 5–16, Apr. 2011.

[2] T. H. Bullock, M. V. L. Bennett, D. Johnston, R. Josephson, E. Marder, and R. D. Fields, “The Neuron Doctrine, Redux,” *Science (80-.).*, vol. 310, no. 5749, pp. 791–793, Nov. 2005.

[3] M. Tudor, L. Tudor, and K. I. Tudor, “Hans Berger (1873-1941)--the history of electroencephalography,” *Acta medica Croat. Cas. Hravatske Akad. Med. Znan.*, vol. 59, no. 4, pp. 307–313, 2005.

[4] A. L. Hodgkin and A. F. Huxley, “A quantitative description of membrane current and its application to conduction and excitation in nerve,” *J. Physiol.*, vol. 117, no. 4, pp. 500–544, Aug. 1952.

[5] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[6] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.,” *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, 1958.

[7] M. Minsky and S. Papert, “An introduction to computational geometry,” *Cambridge tiass., HIT*, vol. 479, no. 480, p. 104, 1969.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[9] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[10] S. Herculano-Houzel, “The human brain in numbers: a linearly scaled-up primate brain,” *Front. Hum. Neurosci.*, p. 31, 2009.

[11] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[13] R. S. Dehal, C. Munjal, A. A. Ansari, and A. S. Kushwaha, “Gpu computing revolution: Cuda,” in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018, pp. 197–201.

[14] L. Fei-Fei, J. Deng, and K. Li, “ImageNet: Constructing a large-scale image database,” *J. Vis.*, vol. 9, no. 8, pp. 1037–1037, Mar. 2010.

[15] A. De Cesarei, S. Cavicchi, G. Cristadoro, and M. Lippi, “Do humans and deep convolutional neural networks use visual information similarly for the categorization of natural s scenes?,” *Cogn. Sci.*, vol. 45, no. 6, p. e13009, Jun. 2021.

[16] C. Kacfah Emani, N. Cullot, and C. Nicolle, “Understandable Big Data: A survey,” *Comput. Sci. Rev.*, vol. 17, pp. 70–81, Aug. 2015.

[17] C. M. Bishop, “Training with Noise is Equivalent to Tikhonov Regularization,” *Neural Comput.*, vol. 7, no. 1, pp. 108–116, Jan. 1995.

[18] Z. Zhu, S. Wang, and C. E. Woodcock, “Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images,” *Remote Sens. Environ.*, vol. 159, pp. 269–277, Mar. 2015.

[19] X.-Y. Zhuge, X. Zou, and Y. Wang, “A fast cloud detection algorithm applicable to monitoring and nowcasting of daytime cloud systems,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6111–6119, Nov. 2017.

[20] S. A. Ackerman, R. E. Holz, R. Frey, E. W. Eloranta, B. C. Maddux, and M. McGill, “Cloud detection with MODIS. Part II: Validation,” *J. Atmos. Ocean. Technol.*, vol. 25, no. 7, pp. 1073–1086, Jul. 2008.

[21] A. C. Banks and F. Mélin, “An assessment of cloud masking schemes for satellite ocean colour data of marine optical extremes,” *Int. J. Remote Sens.*, vol. 36, no. 3, pp. 797–821, Feb. 2015.

[22] Z. Zhu and C. E. Woodcock, “Object-based cloud and cloud shadow detection in Landsat imagery,” *Remote Sens. Environ.*, vol. 118, no. 1–4, pp. 83–94, Mar. 2012.

[23] O. Hagolle, M. Huc, D. V. Pascual, and G. Dedieu, “A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images,” *Remote Sens. Environ.*, vol. 114, no. 8, pp. 1747–1755, Aug. 2010.

[24] G. Mateo-García, L. Gómez-Chova, J. Amorós-López, J. Muñoz-Marí, and G. Camps-Valls, “Multitemporal cloud masking in the Google Earth Engine,” *Remote Sens.*, vol. 10, no. 7, p. 1079, Jul. 2018.

[25] T. Bai, D. Li, K. Sun, Y. Chen, and W. Li, “Cloud detection for high-resolution satellite imagery using machine learning and multi-feature fusion,” *Remote Sens.*, vol. 8, no. 9, p. 715, Aug. 2016.

[26] Y. Yuan and X. Hu, “Bag-of-Words and Object-Based classification for cloud extraction from satellite imagery,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, no. 8, pp. 4197–4205, Aug. 2015.

[27] J. M. Sholar, “Lightweight deconvolutional neural networks for efficient cloud identification in satellite images,” 2017.

[28] M. Wieland, Y. Li, and S. Martinis, “Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network,” *Remote Sens. Environ.*, vol. 230, no. April, 2019.

[29] M. Segal-Rozenhaimer, A. Li, K. Das, and V. Chirayath, “Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN),” *Remote Sens. Environ.*, vol. 237, no. September 2019, p. 111446, Feb. 2020.

[30] W. Zhang, J. Wang, D. Jin, L. Oreopoulos, and Z. Zhang, “A deterministic self-organizing map approach and its application on satellite data based cloud type classification,” in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2027–2034.

[31] F. M. Riese, S. Keller, and S. Hinz, “Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data,” *Remote Sens.*, vol. 12, no. 1, p. 7, Dec. 2019.

[32] R. W. Gould Jr, R. A. Arnone, and M. Sydor, “Absorption, scattering, and, remote-sensing reflectance relationships in coastal

waters: Testing a new inversion algorith," *J. Coast. Res.*, pp. 328–341, 2001.
[33] Z. Lee, Y.-H. Ahn, C. Mobley, and R. Arnone, "Removal of surface-reflected light for the measurement of remote-sensing reflectance from an above-surface platform," *Opt. Express*, vol. 18, no. 25, p. 26313, Dec. 2010.
[34] Z. Zhang, A. Iwasaki, G. Xu, and J. Song, "Small satellite cloud detection based on deep learning and image compression," *Preprints*, no. February, pp. 1–12, 2018.
[35] X. Zhu and E. H. Helmer, "An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions," *Remote Sens. Environ.*, vol. 214, no. September 2017, pp. 135–153, Sep. 2018.
[36] L. Gomez-Chova, G. Mateo-Garcia, J. Munoz-Mari, and G. Camps-Valls, "Cloud detection machine learning algorithms for PROBA-V," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 2251–2254.
[37] M. Le Goff, J.-Y. Tourneret, H. Wendt, M. Ortner, and M. Spigai, "Deep learning for cloud detection," in *8th International Conference of Pattern Recognition Systems (ICPRS 2017)*, 2017, pp. 1–6.
[38] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, no. August, pp. 235–253, Oct. 2018.
[39] A. Fisher, "Cloud and cloud-shadow detection in SPOT5 HRG imagery with automated morphological feature extraction," *Remote Sens.*, vol. 6, no. 1, pp. 776–800, Jan. 2014.
[40] S. Chen and T. Zhang, "An improved cloud masking algorithm for MODIS ocean colour data processing," *Remote Sens. Lett.*, vol. 6, no. 3, pp. 218–227, Mar. 2015.
[41] H. Ishida, Y. Oishi, K. Morita, K. Moriwaki, and T. Y. Nakajima, "Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions," *Remote Sens. Environ.*, vol. 205, no. October 2017, pp. 390–407, Feb. 2018.
[42] T. Y. Nakajima *et al.*, "Theoretical basis of the algorithms and early phase results of the GCOM-C (Shikisai) SGLI cloud products," *Prog. Earth Planet. Sci.*, vol. 6, no. 1, p. 52, Dec. 2019.
[43] A. Hollstein, K. Segl, L. Guanter, M. Brell, and M. Enesco, "Ready-to-Use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI Images," *Remote Sens.*, vol. 8, no. 8, p. 666, Aug. 2016.
[44] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of Copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure," *Remote Sens.*, vol. 11, no. 4, p. 433, Feb. 2019.
[45] A. Francis, J. Mrziglod, P. Sidiropoulos, and J.-P. Muller, "Sentinel-2 cloud mask catalogue." Zenodo, Nov-2020.
[46] J. Li *et al.*, "A lightweight deep learning-based cloud detection method for Sentinel-2A imagery fusing multiscale spectral and spatial features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.
[47] C. Aybar *et al.*, "CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2," *Sci. Data*, vol. 9, no. 1, p. 782, Dec. 2022.
[48] "Sentinel-2 Cloud Cover Segmentation Dataset (Version 1)." Radiant MLHub, 2022.
[49] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, no. May, pp. 247–259, Aug. 2019.
[50] Y. Oishi, H. Ishida, and R. Nakamura, "A new Landsat 8 cloud discrimination algorithm using thresholding tests," *Int. J. Remote Sens.*, vol. 39, no. 23, pp. 9113–9133, Dec. 2018.
[51] M. Xia, W. Liu, B. Shi, L. Weng, and J. Liu, "Cloud/snow recognition for multispectral satellite imagery based on a multidimensional deep residual network," *Int. J. Remote Sens.*, vol. 40, no. 1, pp. 156–170, 2019.
[52] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sens.*, vol. 11, no. 11, p. 1343, Jun. 2019.
[53] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-SiamNet: Pyramid feature-based attention-guided siamese network for Remote Sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, p. 484, Feb. 2020.
[54] L. Su, M. Gong, P. Zhang, M. Zhang, J. Liu, and H. Yang, "Deep learning and mapping based ternary change detection for information unbalanced images," *Pattern Recognit.*, vol. 66, pp. 213–228, Jun. 2017.
[55] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9960–9975, Dec. 2019.
[56] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2007.
[57] N. Falco, P. R. Marpu, and J. A. Benediktsson, "A toolbox for unsupervised change detection analysis," *Int. J. Remote Sens.*, vol. 37, no. 7, pp. 1505–1526, Apr. 2016.
[58] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, Jun. 2013.
[59] P. Gong, E. F. Ledrew, and J. R. Miller, "Registration-noise reduction in difference images for change detection," *Int. J. Remote Sens.*, vol. 13, no. 4, pp. 773–779, Mar. 1992.
[60] J. Theiler and B. Wohlberg, "Local coregistration adjustment for anomalous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3107–3116, Aug. 2012.
[61] A. Mohammed El Amin, Q. Liu, and Y. Wang, "Convolutional neural network features based change detection in satellite images," in *First International Workshop on Pattern Recognition*, 2016, vol. 10011, no. July, p. 100110W.
[62] C. Zhang, L. He, and L. Jiang, "Refined deep features for unsupervised change detection in high resolution remote sensing images," in *2021 9th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 2021, pp. 1–4.
[63] K. L. de Jong and A. Sergeevna Bosman, "Unsupervised change detection in satellite images using convolutional neural networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, vol. 2019-July, pp. 1–8.
[64] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite

images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.
[65] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for Remote Sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.
[66] UNEP, "Marine litter, an analytical overview, The right start to a healthy life : levelling-up the health gradient among children, young people and families in the European Union : what works," 2005.
[67] J. R. Jambeck *et al.*, "Plastic waste inputs from land into the ocean," *Science (80-.).*, vol. 347, no. 6223, pp. 768–771, Feb. 2015.
[68] S. Casabianca *et al.*, "Plastic-associated harmful microalgal assemblages in marine environment," *Environ. Pollut.*, vol. 244, pp. 617–626, Jan. 2019.
[69] GESAMP, *Guidelines for the monitoring and assessment of plastic litter in the ocean*. London, UK, 2019.
[70] C. Benedek and T. Sziranyi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.
[71] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in Remote Sensing images using conditional adversarial networks," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLII–2, no. June 2018, pp. 565–571, May 2018.
[72] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Underst.*, vol. 187, p. 102783, Oct. 2019.
[73] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[74] L. Shen *et al.*, "S2Looking: A satellite side-looking dataset for building change detection," *Remote Sens.*, vol. 13, no. 24, p. 5094, Dec. 2021.
[75] S. Verma, A. Panigrahi, and S. Gupta, "Qfabric: Multi-task change detection dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1052–1061.
[76] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, and M. Pelillo, "Asymmetric Siamese Networks for Semantic Change Detection," pp. 1–15, 2020.
[77] N. Maximenko *et al.*, "Toward the Integrated Marine Debris Observing System," *Front. Mar. Sci.*, vol. 6, no. JUL, Aug. 2019.
[78] U. N. E. P. UNEP, "Marine plastic debris and microplastics – Global lessons and research to inspire action and guide policy change.," 2016.
[79] K. Topouzelis, D. Papageorgiou, G. Suaria, and S. Aliani, "Floating marine litter detection algorithms and techniques using optical remote sensing data: A review," *Mar. Pollut. Bull.*, vol. 170, p. 112675, Sep. 2021.
[80] T. Acuña-Ruz *et al.*, "Anthropogenic marine debris over beaches: Spectral characterization for remote sensing applications," *Remote Sens. Environ.*, vol. 217, pp. 309–322, Nov. 2018.
[81] S. P. Garaba and H. M. Dierssen, "An airborne remote sensing case study of synthetic hydrocarbon detection using short wave infrared absorption features identified from marine-harvested macro- and microplastics," *Remote Sens. Environ.*, vol. 205, pp. 224–235, Feb. 2018.
[82] K. Topouzelis, A. Papakonstantinou, and S. P. Garaba, "Detection of floating plastics from satellite and unmanned aerial systems (Plastic Litter Project 2018)," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 79, pp. 175–183, Jul. 2019.
[83] K. Topouzelis, D. Papageorgiou, A. Karagaitanakis, A. Papakonstantinou, and M. Arias Ballesteros, "Remote Sensing of sea surface artificial floating plastic targets with Sentinel-2 and unmanned aerial systems (Plastic Litter Project 2019)," *Remote Sens.*, vol. 12, no. 12, p. 2013, Jun. 2020.
[84] L. Biermann, D. Clewley, V. Martinez-Vicente, and K. Topouzelis, "Finding plastic patches in coastal waters using optical satellite data," *Sci. Rep.*, vol. 10, no. 1, p. 5364, Apr. 2020.
[85] K. Kikaki, I. Kakogeorgiou, P. Mikeli, D. E. Raitsos, and K. Karantzalos, "MARIDA: A benchmark for Marine Debris detection from Sentinel-2 remote sensing data," *PLoS One*, vol. 17, no. 1, p. e0262247, 2022.
[86] L. Loncan *et al.*, "Hyperspectral Pansharpening: A Review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, Sep. 2015.
[87] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.
[88] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 305–319, Dec. 2018.
[89] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Sentinel-2 image fusion using a deep residual network," *Remote Sens.*, vol. 10, no. 8, 2018.
[90] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, no. July, pp. 14–34, Sep. 2021.
[91] Y. Pan, I. Ahmed Khan, and H. Meng, "SAR-to-optical image translation using multi-stream deep ResCNN of information reconstruction," *Expert Syst. Appl.*, vol. 224, no. November 2022, p. 120040, Aug. 2023.
[92] Q. Luo, H. Li, Z. Chen, and J. Li, "ADD-UNet: An Adjacent Dual-Decoder UNet for SAR-to-Optical Translation," *Remote Sens.*, vol. 15, no. 12, p. 3125, Jun. 2023.
[93] Y. Qing, J. Zhu, H. Feng, W. Liu, and B. Wen, "Two-way generation of high-resolution EO and SAR images via dual distortion-adaptive GANs," *Remote Sens.*, vol. 15, no. 7, 2023.
[94] J.-E. Park, G. Kim, and S. Hong, "Green band generation for Advanced Baseline Imager Sensor using Pix2Pix with Advanced Baseline Imager and Advanced Himawari Imager observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6415–6423, Aug. 2021.
[95] H.-S. Ryu, J.-E. Park, J. Jeong, and S. Hong, "Generation of hypothetical radiances for missing green and red bands in Geostationary Environment Monitoring Spectrometer," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 9025–9037, 2023.

[96] K. Kim *et al.*, "Nighttime reflectance generation in the visible band of satellites," *Remote Sens.*, vol. 11, no. 18, p. 2087, Sep. 2019.
[97] Y. Liu *et al.*, "CscGAN: Conditional scale-consistent generation network for multi-level Remote Sensing image to map translation," *Remote Sens.*, vol. 13, no. 10, p. 1936, May 2021.
[98] J. Song, J. Li, H. Chen, and J. Wu, "RSMT: A Remote Sensing image-to-map translation model via adversarial deep transfer learning," *Remote Sens.*, vol. 14, no. 4, p. 919, Feb. 2022.
[99] P. Ghamisi and N. Yokoya, "IMG2DSM: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 794–798, May 2018.
[100] L. Liu, S. Lei, Z. Shi, N. Zhang, and X. Zhu, "Hyperspectral Remote Sensing imagery generation from RGB images based on Joint Discrimination," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 7624–7636, 2021.
[101] L. Liu, Z. Shi, Y. Zao, and H. Chen, "Hyperspectral image generation from RGB images with semantic and spatial distribution consistency," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, vol. 2022-July, pp. 1804–1807.
[102] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler, "Learned spectral super-resolution," Mar. 2017.
[103] S. Paul and D. Nagesh Kumar, "Transformation of multispectral data to quasi-hyperspectral data using convolutional neural network regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3352–3368, Apr. 2021.
[104] K. Arai, M. Sakashita, O. Shigetomi, and Y. Miura, "Estimation of protein content in rice crop and nitrogen content in rice leaves through regression analysis with NDVI derived from camera mounted radio-control helicopter," *Int. J. Adv. Res. Artif. Intell*, vol. 3, 2014.
[105] P. J. Navarro, F. Pérez, J. Weiss, and M. Egea-Cortines, "Machine learning and computer vision system for phenotype data acquisition and analysis in plants," *Sensors*, vol. 16, no. 5, p. 641, 2016.
[106] D. C. de Lima, D. Saqui, S. A. T. Mpinda, and J. H. Saito, "Pix2Pix network to estimate agricultural near infrared images from RGB Data," *Can. J. Remote Sens.*, vol. 48, no. 2, pp. 299–315, Mar. 2022.
[107] S. Illarionova, D. Shadrin, A. Trekin, V. Ignatiev, and I. Oseledets, "Generation of the NIR spectral band for satellite images with convolutional neural networks," *Sensors*, vol. 21, no. 16, p. 5646, Aug. 2021.
[108] X. Yuan, J. Tian, and P. Reinartz, "Learning-based near-infrared band simulation with applications on large-scale landcover classification," *Sensors*, vol. 23, no. 9, p. 4179, Apr. 2023.
[109] R. Richter, J. Louis, and U. Müller-Wilm, "Sentinel-2 MSI - Level 2A Products Algorithm Theoretical Basis Document," 2012.
[110] O. Hagolle, M. Huc, C. Desjardins, S. Auer, and R. Richter, "MAJA ATBD Algorithm Theoretical Basis Document," 2017.
[111] S. Platnick *et al.*, "The MODIS cloud products: algorithms and examples from terra," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 2, pp. 459–473, Feb. 2003.
[112] D. Shin, J. K. Pollard, and J.-P. Muller, "Cloud detection from thermal infrared images using a segmentation technique," *Int. J. Remote Sens.*, vol. 17, no. 14, pp. 2845–2856, Sep. 1996.
[113] M. J. Wilson and L. Oreopoulos, "Enhancing a simple MODIS cloud mask algorithm for the Landsat Data Continuity Mission," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 723–731, Feb. 2013.
[114] G. J. Jedlovec, S. L. Haines, and F. J. LaFontaine, "Spatial and temporal varying thresholds for cloud detection in GOES imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1705–1717, Jun. 2008.
[115] R. R. Irish, "Landsat 7 automatic cloud cover assessment," in *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI*, 2000, vol. 4049, p. 348.
[116] S. Foga *et al.*, "Cloud detection algorithm comparison and validation for operational Landsat data products," *Remote Sens. Environ.*, vol. 194, pp. 379–390, Jun. 2017.
[117] D. S. Candra, S. Phinn, and P. Scarth, "Cloud and cloud shadow masking using multi-temporal cloud masking algorithm in Tropical Environmental," *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLI-B2, no. July, pp. 95–100, Jun. 2016.
[118] A. Ahmad and S. Quegan, "Multitemporal cloud detection and masking using MODIS data," *Appl. Math. Sci.*, vol. 8, no. 7, pp. 345–353, 2014.
[119] J. Karvonen, "Cloud masking of MODIS imagery based on multitemporal image analysis," *Int. J. Remote Sens.*, vol. 35, no. 23, pp. 8008–8024, Dec. 2014.
[120] C.-H. Lin, B.-Y. Lin, K.-Y. Lee, and Y.-C. Chen, "Radiometric normalization and cloud detection of optical satellite images using invariant pixels," *ISPRS J. Photogramm. Remote Sens.*, vol. 106, pp. 107–117, Aug. 2015.
[121] S. C. Gallegos, J. D. Hawkins, and C. F. Cheng, "A new automated method of cloud masking for advanced very high resolution radiometer full-resolution data over the ocean," *J. Geophys. Res. Ocean.*, vol. 98, no. C5, pp. 8505–8516, May 1993.
[122] S. A. Ackerman, K. I. Strabala, W. P. Menzel, R. A. Frey, C. C. Moeller, and L. E. Gumley, "Discriminating clear sky from clouds with MODIS," *J. Geophys. Res. Atmos.*, vol. 103, no. D24, pp. 32141–32157, Dec. 1998.
[123] J. Hocking, P. N. Francis, and R. Saunders, "Cloud detection in Meteosat Second Generation imagery at the Met Office," *Meteorol. Appl.*, vol. 18, no. 3, pp. 307–323, Sep. 2011.
[124] R. Q. Iannone *et al.*, "Proba-V cloud detection Round Robin: Validation results and recommendations," in *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, 2017, no. June 2017, pp. 1–8.
[125] M. Wang and W. Shi, "Cloud masking for ocean color data processing in the Coastal Regions," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3196–3105, Nov. 2006.
[126] E. Baseski and C. Cenaras, "Texture and color based cloud detection," in *2015 7th International Conference on Recent Advances in Space Technologies (RAST)*, 2015, pp. 311–315.
[127] M. Hughes and D. Hayes, "Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing," *Remote Sens.*, vol. 6, no. 6, pp. 4907–4926, May 2014.
[128] J. Strandgren, L. Bugliaro, F. Sehnke, and L. Schröder, "Cirrus cloud retrieval with MSG/SEVIRI using artificial neural networks," *Atmos. Meas. Tech.*, vol. 10, no. 9, pp. 3547–3573, Sep. 2017.

[129] A. Taravat, S. Peronaci, M. Sist, F. Del Frate, and N. Oppelt, "The combination of band ratioing techniques and neural networks algorithms for MSG SEVIRI and Landsat ETM+ cloud masking," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 2315–2318.
[130] L. Weng, W. Kong, and M. Xia, "Computing cloud cover fraction in satellite images using Deep Extreme Learning machine," *Int. J. Simul. Syst. Sci. Technol.*, vol. 17, no. 48, pp. 37.1-37.9, Jan. 2016.
[131] H. Liu, D. Zeng, and Q. Tian, "Super-pixel cloud detection using Hierarchical Fusion CNN," *2018 IEEE 4th Int. Conf. Multimed. Big Data, BigMM 2018*, pp. 1–6, Oct. 2018.
[132] G. Mateo-Garcia, L. Gomez-Chova, and G. Camps-Valls, "Convolutional neural networks for multispectral image cloud masking," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, vol. 2017-July, pp. 2255–2258.
[133] Z. Li, H. Shen, Y. Wei, Q. Cheng, and Q. Yuan, "Cloud detection by fusing multi-scale convolutional features," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. IV–3, no. 3, pp. 149–152, Apr. 2018.
[134] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 8, pp. 3631–3640, Aug. 2017.
[135] S. Mohajerani, T. A. Krammer, and P. Saeedi, "A Cloud detection algorithm for Remote Sensing images using fully convolutional neural networks," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1–5.
[136] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, "Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.*, vol. 225, no. September 2018, pp. 307–316, May 2019.
[137] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDnet: CNN-based cloud detection for Remote Sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6195–6211, Aug. 2019.
[138] G. Mateo-Garcia and L. Gomez-Chova, "Convolutional Neural Networks for Cloud Screening: Transfer Learning from Landsat-8 to Proba-V," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, vol. 2018-July, no. 30 m, pp. 2103–2106.
[139] A. A. Charantonis, S. Thiria, B. Berthelot, H. Brogniez, and M. Roux, "Cloud detection algorithm development in preparation for the Sentinel-2 mission," École Polytechnique, 2009.
[140] O. Chabiron, "Cloud detection using SOM. Development of a generic method," Institut de Mathématiques de Toulouse, 2013.
[141] S. Angeli, A. Quesney, and L. Gross, "Image simplification using Kohonen maps: Application to satellite data for cloud detection and land cover mapping," in *Applications of Self-Organizing Maps*, vol. i, no. tourism, InTech, 2012, p. 13.
[142] B. Berthelot, "ATBD Cloud Detection for PROBA-V," 2017.
[143] C. Cox and W. Munk, "Measurement of the roughness of the sea surface from photographs of the Sun's Glitter," *J. Opt. Soc. Am.*, vol. 44, no. 11, p. 838, Nov. 1954.
[144] C. D. Mobley, "Estimation of the remote-sensing reflectance from above-surface measurements," *Appl. Opt.*, vol. 38, no. 36, p. 7442, Dec. 1999.
[145] H. Zhang and M. Wang, "Evaluation of sun glint models using MODIS measurements," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 111, no. 3, pp. 492–506, Feb. 2010.
[146] ESA, "MultiSpectral Instrument (MSI) Overview," 2019.
[147] K. Yuan, G. Meng, D. Cheng, J. Bai, S. Xiang, and C. Pan, "Efficient cloud detection in remote sensing images using edge-aware segmentation network and easy-to-hard training strategy," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 61–65.
[148] M. Shi, F. Xie, Y. Zi, and J. Yin, "Cloud detection of remote sensing images by deep learning," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 701–704.
[149] B. Tian, M. A. Shaikh, M. R. Azimi-Sadjadi, T. H. V. Haar, and D. L. Reinke, "A study of cloud classification with neural networks using spectral and textural features," *IEEE Trans. Neural Networks*, vol. 10, no. 1, pp. 138–151, 1999.
[150] C. Huang *et al.*, "Automated masking of cloud and cloud shadow for forest change analysis using Landsat images," *Int. J. Remote Sens.*, 2010.
[151] S. Wu, B. Zhong, W. Li, and Q. Liu, "A new cloud detection method over Tibetan plateau and its surrounding area," in *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, 2013, pp. 550–553.
[152] C. Hu, "An empirical approach to derive MODIS ocean color patterns under severe sun glint," *Geophys. Res. Lett.*, vol. 38, no. 1, 2011.
[153] S. P. Garaba, J. Schulz, M. R. Wernand, and O. Zielinski, "Sunglint detection for unmanned and automated platforms," *Sensors*, vol. 12, no. 9, pp. 12545–12561, Sep. 2012.
[154] J. M. Nicolas, P. Y. Deschamps, H. Loisel, and C. Moulin, "Algorithm Theoretical Basis Document, POLDER-2/Ocean Color/Atmospheric Corrections," 2005.
[155] J. V. Martins, D. Tanré, L. Remer, Y. Kaufman, S. Mattoo, and R. Levy, "MODIS Cloud screening for remote sensing of aerosols over oceans using spatial variability," *Geophys. Res. Lett.*, vol. 29, no. 12, pp. MOD4-1-MOD4-4, Jun. 2002.
[156] N. Roslan, M. N. M. Reba, M. Askari, and M. K. A. Halim, "Linear and non-linear enhancement for sun glint reduction in advanced very high resolution radiometer (AVHRR) image," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 18, no. 1, p. 012041, Feb. 2014.
[157] E. Ricciardelli, F. Romano, and V. Cuomo, "Physical and statistical approaches for cloud identification using Meteosat Second Generation-Spinning Enhanced Visible and Infrared Imager Data," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2741–2760, Jun. 2008.
[158] M. Schröder, R. Bennartz, L. Schüller, R. Preusker, P. Albert, and J. Fischer, "Generating cloudmasks in spatial high-resolution observations of clouds using texture and radiance information," *Int. J. Remote Sens.*, vol. 23, no. 20, pp. 4247–4261, Jan. 2002.
[159] T. Imai and R. Yoshida, "Algorithm Theoretical Basis for Himawari-8 Cloud Product data Cloud Analysis Information derived from Algorithm Theoretical Basis for Himawari-8 Cloud Mask Product," vol. 43, no. 61, pp. 43–51, 2016.
[160] Z. Li, H. Shen, H. Li, G. Xia, P. Gamba, and L. Zhang, "Multi-feature combined cloud and cloud shadow detection in GaoFen-1

wide field of view imagery," *Remote Sens. Environ.*, vol. 191, no. April 2013, pp. 342–358, Mar. 2017.
[161] D. Frantz, E. Haß, A. Uhl, J. Stoffels, and J. Hill, "Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects," *Remote Sens. Environ.*, vol. 215, no. March, pp. 471–481, Sep. 2018.
[162] J. Lu *et al.*, "P_Segnet and NP_Segnet: New neural network architectures for cloud recognition of Remote Sensing images," *IEEE Access*, vol. 7, pp. 87323–87333, 2019.
[163] S. R. Yhann and J. J. Simpson, "Application of neural networks to AVHRR cloud segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 3, pp. 590–604, May 1995.
[164] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," pp. 1–15, Dec. 2014.
[165] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," no. 1, pp. 2–8, Mar. 2018.
[166] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," pp. 1–20, Nov. 2018.
[167] F. Chollet, "Keras: The Python Deep Learning library," *Keras.Io*, 2015. [Online]. Available: https://github.com/fchollet/keras.
[168] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.
[169] D. H. Park, C. M. Ho, and Y. Chang, "Achieving strong regularization for deep neural networks," 2018.
[170] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016*, May 2016.
[171] O. Allouche, A. Tsoar, and R. Kadmon, "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)," *J. Appl. Ecol.*, vol. 43, no. 6, pp. 1223–1232, Dec. 2006.
[172] A. A.-K. H. Abu-Ein, "A novel methodology for digital removal of periodic noise using 2D fast Fourier transforms," *Contemp. Eng. Sci.*, vol. 7, no. 3, pp. 103–116, 2014.
[173] H. Cramér, "Mathematical methods of statistics, 1946," *Dep. Math. SU*, 1946.
[174] STEP, "Sentinel-2 cloud mask with Fmask," 2016.
[175] STEP, "Reducing Sen2Cor processing time," 2016.
[176] CESBIO, "On-demand SENTINEL2 L2A processing with MAJA on PEPS," 2018.
[177] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.
[178] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, Jan. 2013.
[179] S. H. Hsu, J. P. A. Hsieh, T. C. Chih, and K. C. Hsu, "A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7947–7951, 2009.
[180] M. Hagenbuchner and A. C. Tsoi, "A supervised training algorithm for self-organizing maps for structures," *Pattern Recognit. Lett.*, vol. 26, no. 12, pp. 1874–1884, Sep. 2005.
[181] C. Y. Ji and others, "Land-use classification of remotely sensed data using Kohonen self-organizing feature map neural networks," *Photogramm. Eng. Remote Sensing*, vol. 66, no. 12, pp. 1451–1460, 2000.
[182] G. A. Barreto and A. F. R. Araujo, "Identification and control of dynamical systems using the self-organizing map," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1244–1259, Sep. 2004.
[183] H. Matsushita and Y. Nishio, "Self-Organizing Map with Weighted Connections avoiding false-neighbor effects," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 2554–2557.
[184] A. Ultsch, "Kohonen's self organizing feature maps for exploratory data analysis," *INNC'90*, 1990.
[185] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.
[186] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, "Supervised change detection in VHR images using contextual information and support vector machines," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 20, no. 1, pp. 77–85, Feb. 2013.
[187] D. Cerra, S. Plank, V. Lysandrou, and J. Tian, "Cultural heritage sites in danger—Towards automatic damage detection from Space," *Remote Sens.*, vol. 8, no. 9, p. 781, Sep. 2016.
[188] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of Remote Sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.
[189] J. Chen, X. Chen, X. Cui, and J. Chen, "Change vector analysis in posterior probability space: A new method for land cover change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 317–321, Mar. 2011.
[190] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust Change Vector Analysis (RCVA) for multi-sensor very high resolution optical satellite data," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 50, pp. 131–140, Aug. 2016.
[191] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901.
[192] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, Apr. 1998.
[193] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.
[194] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.
[195] B. Thompson, "Canonical correlation analysis.," in *Reading and understanding MORE multivariate statistics.*, Washington, DC, US: American Psychological Association, 2000, pp. 285–316.
[196] J. L. Gil-Yepes, L. A. Ruiz, J. A. Recio, Á. Balaguer-Beser, and T. Hermosilla, "Description and validation of a new set of object-based temporal geostatistical features for land-use/land-cover change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 121, pp. 77–91, Nov. 2016.

[197] G. Chen, K. Zhao, and R. Powers, "Assessment of the image misregistration effects on object-based change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 19–27, Jan. 2014.
[198] L. Bruzzone and S. B. Serpico, "Detection of changes in remotely-sensed images by the selective use of multi-spectral information," *Int. J. Remote Sens.*, vol. 18, no. 18, pp. 3883–3888, Dec. 1997.
[199] D. A. Stow, "Reducing the effects of misregistration on pixel-level change detection," *Int. J. Remote Sens.*, vol. 20, no. 12, pp. 2477–2483, Jan. 1999.
[200] A. M. El Amin, Q. Liu, and Y. Wang, "Zoom out CNNs features for optical remote sensing change detection," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, 2017, pp. 812–817.
[201] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, Sep. 2014.
[202] B. Hou, Y. Wang, and Q. Liu, "Change detection based on deep features and low rank," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2418–2422, Dec. 2017.
[203] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[204] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, pp. 234–241.
[205] "2D Semantic Labeling_Vaihingen Data." [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx.
[206] J. Liu *et al.*, "Convolutional neural network-based transfer learning for optical aerial images change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 127–131, Jan. 2020.
[207] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.
[208] C. Wu, H. Chen, B. Do, and L. Zhang, "Unsupervised change detection in multi-temporal VHR images based on deep kernel PCA convolutional mapping network," pp. 1–17, Dec. 2019.
[209] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, 2020.
[210] S. Ghaffarian, N. Kerle, E. Pasolli, and J. Jokar Arsanjani, "Post-disaster building database updating using automated deep learning: An integration of pre-disaster OpenStreetMap and multi-temporal satellite data," *Remote Sens.*, vol. 11, no. 20, p. 2427, Oct. 2019.
[211] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.
[212] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4063–4067.
[213] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial Remote Sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.
[214] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[215] DeepAI, "PASCAL VOC Dataset," 2012. [Online]. Available: https://deepai.org/dataset/pascal-voc.
[216] R. Liu, Z. Cheng, L. Zhang, and J. Li, "Remote Sensing image change detection based on information transmission and attention mechanism," *IEEE Access*, vol. 7, pp. 156349–156359, 2019.
[217] R. Liu, D. Jiang, L. Zhang, and Z. Zhang, "Deep depthwise separable convolutional network for change detection in optical aerial images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 1109–1118, 2020.
[218] X. Li, M. He, H. Li, and H. Shen, "A combined Loss-based multiscale fully convolutional network for high-resolution Remote Sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[219] Q. Zhu *et al.*, "Land-Use/Land-Cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022.
[220] P. Lv, Y. Zhong, J. Zhao, and L. Zhang, "Unsupervised change detection based on hybrid conditional random field model for high spatial resolution Remote Sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 4002–4015, Jul. 2018.
[221] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, "Land-Use/Land-Cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution Remote Sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[222] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for Remote Sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[223] C. Zhang *et al.*, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, no. May, pp. 183–200, Aug. 2020.
[224] A. Raza, H. Huo, and T. Fang, "EUNet-CD: Efficient UNet++ for change detection of very high-resolution Remote Sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[225] K. Lim, D. Jin, and C.-S. Kim, "Change detection in high resolution satellite images using an ensemble of convolutional neural networks," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, no. November, pp. 509–515.
[226] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[227] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International*

Conference on Computer Vision, 2011, pp. 2564–2571.
[228] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal Remote Sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, no. c, pp. 38544–38555, 2018.
[229] D. Casasent and D. Psaltis, "New optical transforms for pattern recognition," *Proc. IEEE*, vol. 65, no. 1, pp. 77–84, 1977.
[230] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision--ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, 2006, pp. 430–443.
[231] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6314 LNCS, no. PART 4, 2010, pp. 778–792.
[232] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–22, Sep. 1977.
[233] R. Jonker and T. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," in *DGOR/NSOR: Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR*, 1988, pp. 622–622.
[234] X. Guo, Z. Xu, Y. Lu, and Y. Pang, "An application of Fourier-Mellin transform in image registration," *Proc. - Fifth Int. Conf. Comput. Inf. Technol. CIT 2005*, vol. 2005, pp. 619–623, 2005.
[235] D. Gupta and M. K. Patil, "A review on image registration," *Int. J. Eng. Res. Technol.*, vol. 3, no. 2, pp. 2630–2633, 2014.
[236] A. M. E. Amin, Q. Liu, and Y. Wang, "Unstructured-change-detection-using-CNN," 2016. [Online]. Available: https://github.com/vbhavank/Unstructured-change-detection-using-CNN.
[237] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man. Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.
[238] M. Zhang and W. Shi, "FDCNN," 2020. [Online]. Available: https://github.com/MinZHANG-WHU/FDCNN.
[239] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675–678.
[240] J. Chen *et al.*, "DASNet," 2022. [Online]. Available: https://github.com/lehaifeng/DASNet.
[241] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.
[242] H. Chen and Z. Shi, "STANet for Remote Sensing image change detection," 2020. [Online]. Available: https://github.com/justchenhao/STANet.
[243] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 770–778.
[244] M. A. Fischler and R. C. Bolles, "Random sample paradigm for model consensus: Applications to image fitting with analysis and automated cartography," *Graph. Image Process.*, vol. 24, no. 6, pp. 381–395, 1981.
[245] A. Chamas *et al.*, "Degradation rates of plastics in the environment," *ACS Sustain. Chem. Eng.*, vol. 8, no. 9, pp. 3494–3511, Mar. 2020.
[246] E. A. Smail *et al.*, "An introduction to the 'Oceans and Society: Blue Planet' initiative," *J. Oper. Oceanogr.*, vol. 12, no. sup2, pp. S1–S11, Nov. 2019.
[247] L. M. Rios, C. Moore, and P. R. Jones, "Persistent organic pollutants carried by synthetic polymers in the ocean environment," *Mar. Pollut. Bull.*, vol. 54, no. 8, pp. 1230–1237, Aug. 2007.
[248] P. M. Salgado-Hernanz, J. Bauzà, C. Alomar, M. Compa, L. Romero, and S. Deudero, "Assessment of marine litter through remote sensing: recent approaches and future goals," *Mar. Pollut. Bull.*, vol. 168, p. 112347, Jul. 2021.
[249] UN, *United Nations Transforming Our World: the 2030 Agenda for Sustainable Development. A/RES/70/1*. NY, USA, 2015.
[250] E. U. European Parliament Council, *Directive 2008/56/EC of the European Parliament and of the Council*. Strasbourg, France, 2008.
[251] L. Goddijn-Murphy and J. Dufaur, "Proof of concept for a model of light reflectance of plastics floating on natural waters," *Mar. Pollut. Bull.*, vol. 135, no. July, pp. 1145–1157, 2018.
[252] J. P. Matthews, L. Ostrovsky, Y. Yoshikawa, S. Komori, and H. Tamura, "Dynamics and early post-tsunami evolution of floating marine debris near Fukushima Daiichi," *Nat. Geosci.*, vol. 10, no. 8, pp. 598–603, Aug. 2017.
[253] A. Kikaki, K. Karantzalos, C. A. Power, and D. E. Raitsos, "Remotely sensing the source and transport of marine plastic debris in Bay islands of Honduras (Caribbean Sea)," *Remote Sens.*, vol. 12, no. 11, p. 1727, May 2020.
[254] V. Martinez-Vicente, L. Biermann, and A. Mata, "Optical Methods for Marine Litter Detection (OPTIMAL) - Final Report," 2020.
[255] F. Kühn, K. Oppermann, and B. Hörig, "Hydrocarbon Index – an algorithm for hyperspectral detection of hydrocarbons," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2467–2473, Jun. 2004.
[256] B. Basu, S. Sannigrahi, A. Sarkar Basu, and F. Pilla, "Development of novel classification algorithms for detection of floating plastic debris in coastal waterbodies using multispectral Sentinel-2 Remote Sensing imagery," *Remote Sens.*, vol. 13, no. 8, p. 1598, Apr. 2021.
[257] "Agisoft Metashape." Agisoft LLC, Saint Petersburg, Russia, 2015.
[258] B. Aiazzi, S. Baronti, and M. Selva, "Improving Component Substitution Pansharpening Through Multivariate Regression of MS+Pan Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.
[259] J. G. Liu, "Smoothing Filter-based Intensity Modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, Jan. 2000.
[260] V. Karathanassi, P. Kolokousis, and S. Ioannidou, "A comparison study on fusion methods using evaluation indicators," *Int. J. Remote Sens.*, vol. 28, no. 10, pp. 2309–2341, May 2007.
[261] W. Liao *et al.*, "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS Data

Fusion Contest," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, no. 6, pp. 2984–2996, Jun. 2015.
[262] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of multispectral and hyperspectral images with unknown sensor spectral response," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 698–702.
[263] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.
[264] A. Azarang, H. E. Manoochehri, and N. Kehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35673–35683, 2019.
[265] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, pp. 87–102.
[266] G. Vivone, R. Restaino, M. Dalla Mura, G. Licciardi, and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 930–934, May 2014.
[267] G. Vivone, A. Garzelli, Y. Xu, W. Liao, and J. Chanussot, "Panchromatic and Hyperspectral Image Fusion: Outcome of the 2022 WHISPERS Hyperspectral Pansharpening Challenge," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 166–179, 2023.
[268] M. Kremezi *et al.*, "Pansharpening PRISMA data for marine plastic litter detection using plastic i indexes," *IEEE Access*, vol. 9, pp. 61955–61971, 2021.
[269] S. P. Garaba *et al.*, "Sensing ocean plastics with an airborne hyperspectral shortwave infrared imager," *Environ. Sci. Technol.*, 2018.
[270] M. Kuester, "Radiometric Use of WorldView-3 Imagery," *Tech. Note*, 2016.
[271] D. Papageorgiou, "Floating plastic detection using Sentinel-2 imagery," National Technical University of Athens, Athens, Greece, 2019.
[272] C. Hu, "A novel ocean color index to detect floating algae in the global oceans," *Remote Sens. Environ.*, vol. 113, no. 10, pp. 2118–2129, Oct. 2009.
[273] C. Hu, Z. Lee, and B. Franz, "Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference," *J. Geophys. Res. Ocean.*, vol. 117, no. C1, Jan. 2012.
[274] M. Kremezi and V. Karathanassi, "Correcting the BRDF effects on Sentinel-2 ocean images," in *Seventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2019)*, 2019, p. 45.
[275] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geoscience and Remote Sensing Magazine*. 2017.
[276] J. Bieniarz, D. Cerra, J. Avbelj, P. Reinartz, and R. Müller, "Hyperspectral image resolution enhancement based on spectral unmixing and information fusion," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XXXVIII-4/, no. 1, pp. 33–37, Sep. 2012.
[277] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution with spectral unmixing constraints," *Remote Sens.*, vol. 9, no. 11, p. 1196, Nov. 2017.
[278] Q. Wei, J. Bioucas-Dias, N. Dobigeon, J.-Y. Tourneret, M. Chen, and S. Godsill, "Multiband image fusion based on spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7236–7249, Dec. 2016.
[279] O. Berne, A. Helens, P. Pilleri, and C. Joblin, "Non-negative matrix factorization pansharpening of hyperspectral data: An application to mid-infrared astronomy," in *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2010, pp. 1–4.
[280] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled non-negative matrix factorization (CNMF) for hyperspectral and multispectral data fusion: Application to pasture classification," in *2011 IEEE International Geoscience and Remote Sensing Symposium*, 2011, pp. 1779–1782.
[281] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3631–3640.
[282] Q. Wei, N. Dobigeon, and J. Y. Tourneret, "Bayesian fusion of multi-band images," *IEEE J. Sel. Top. Signal Process.*, 2015.
[283] Q. Wei, N. Dobigeon, and J. Y. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Trans. Image Process.*, 2015.
[284] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2014.
[285] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image Super-Resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, vol. 2017-July, pp. 1132–1140.
[286] X. Zhu, Y. Xu, and Z. Wei, "Super-resolution of Sentinel-2 images based on deep channel-attention residual network," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 628–631.
[287] M. Gargiulo, A. Mazza, R. Gaetano, G. Ruello, and G. Scarpa, "Fast super-resolution of 20 m Sentinel-2 bands using convolutional neural networks," *Remote Sens.*, vol. 11, no. 22, p. 2635, Nov. 2019.
[288] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for Pan-Sharpening," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 1753–1761, 2017.
[289] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-Adaptive CNN-based Pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.
[290] J. Wu, Z. He, and J. Hu, "Sentinel-2 Sharpening via parallel residual network," *Remote Sens.*, vol. 12, no. 2, p. 279, Jan. 2020.
[291] K. Zhang, G. Sumbul, and B. Demir, "An approach to super-resolution of Sentinel-2 images based on generative adversarial networks," in *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, 2020, pp. 69–72.
[292] L. Salgueiro, J. Marcello, and V. Vilaplana, "Single-image super-resolution of Sentinel-2 low resolution bands with residual dense convolutional neural networks," *Remote Sens.*, vol. 13, no. 24, p. 5007, Dec. 2021.

[293] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11133 LNCS, 2019, pp. 63–79.
[294] Y. Li and B. Li, "Super-resolution of Sentinel-2 images at 10m resolution without reference images," no. April, pp. 1–22, 2021.
[295] L. S. Romero, J. Marcello, and V. Vilaplana, "Super-resolution of Sentinel-2 imagery using generative adversarial networks," *Remote Sens.*, vol. 12, no. 15, pp. 1–25, 2020.
[296] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, pp. 3586–3594.
[297] K. Kotwal and S. Chaudhuri, "A novel approach to quantitative evaluation of hyperspectral image fusion techniques," *Inf. Fusion*, vol. 14, no. 1, pp. 5–18, Jan. 2013.
[298] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 105–114.
[299] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 294–310, 2018.
[300] P. Jagalingam and A. V. Hegde, "A review of quality metrics for fused image," *Aquat. Procedia*, vol. 4, pp. 133–142, 2015.
[301] M. Kremezi *et al.*, "Increasing the Sentinel-2 potential for marine plastic litter monitoring through image fusion techniques," *Mar. Pollut. Bull.*, vol. 182, p. 113974, Sep. 2022.
[302] Z. Guo, H. Guo, X. Liu, W. Zhou, Y. Wang, and Y. Fan, "Sar2color: Learning imaging characteristics of SAR images for SAR-to-Optical transformation," *Remote Sens.*, vol. 14, no. 15, p. 3740, Aug. 2022.
[303] J. Wei *et al.*, "CFRWD-GAN for SAR-to-Optical Image Translation," *Remote Sens.*, vol. 15, no. 10, p. 2547, May 2023.
[304] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8165–8174, Feb. 2020.
[305] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 2242–2251, Mar. 2017.
[306] C. Zhang *et al.*, "A domain adaptation neural network for change detection with heterogeneous optical and SAR remote sensing images," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 109, no. April, p. 102769, May 2022.
[307] A. Manocha and Y. Afaq, "Optical and SAR images-based image translation for change detection using generative adversarial network (GAN)," *Multimed. Tools Appl.*, vol. 82, no. 17, pp. 26289–26315, Jul. 2023.
[308] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
[309] H. Wang, Z. Zhang, Z. Hu, and Q. Dong, "SAR-to-Optical image translation with hierarchical latent features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
[310] Y. Fu, Z. Fang, L. Chen, T. Song, and D. Lin, "Level-aware consistent multilevel map translation from satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. c, pp. 1–14, 2023.
[311] X. Chen *et al.*, "SMAPGAN: Generative adversarial network-based semisupervised styled map tile generation method," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4388–4406, May 2021.
[312] J. Song, H. Chen, C. Du, and J. Li, "Semi-MapGen: Translation of Remote Sensing image into map via semisupervised adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–19, 2023.
[313] E. Panagiotou, G. Chochlakis, L. Grammatikopoulos, and E. Charou, "Generating elevation surface from a single RGB remotely sensed image using deep learning," *Remote Sens.*, vol. 12, no. 12, p. 2002, Jun. 2020.
[314] H. A. Amirkolaee and H. Arefi, "Height estimation from single aerial images using a deep convolutional encoder-decoder network," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, no. July 2018, pp. 50–66, Mar. 2019.
[315] S. Karatsiolis, A. Kamilaris, and I. Cole, "IMG2nDSM: Height estimation from single airborne RGB images with deep learning," *Remote Sens.*, vol. 13, no. 12, p. 2417, Jun. 2021.
[316] J. Kim, S. Ryu, J. Jeong, D. So, H. Ban, and S. Hong, "Impact of satellite sounding data on virtual visible imagery generation using conditional generative adversarial network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 4532–4541, 2020.
[317] K.-H. Han, J.-C. Jang, S. Ryu, E.-H. Sohn, and S. Hong, "Hypothetical visible bands of Advanced Meteorological Imager Onboard the Geostationary Korea Multi-Purpose Satellite -2A using data-to-data translation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, no. L, pp. 8378–8388, 2022.
[318] J. Xiao, D. Tortei, E. Roura, and G. Loianno, "Long-range UAV thermal geo-localization with satellite imagery," Jun. 2023.
[319] V. Tiwari, V. Kumar, K. Pandey, R. Ranade, and S. Agrawal, "Simulation of the hyperspectral data using multispectral data," *Int. Geosci. Remote Sens. Symp.*, vol. 2016-Novem, pp. 6157–6160, 2016.
[320] N. T. Hoang and K. Koike, "Transformation of Landsat imagery into pseudo-hyperspectral imagery by a multiple regression-based model with application to metal deposit-related minerals mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 133, pp. 157–173, Nov. 2017.
[321] X. Han, J. Yu, J. Luo, and W. Sun, "Reconstruction from multispectral to hyperspectral image using spectral library-based dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1325–1335, Mar. 2019.
[322] K. Fotiadou, G. Tsagkatakis, and P. Tsakalides, "Spectral super resolution of hyperspectral images via coupled dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2777–2797, May 2019.
[323] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers Tiramisu: Fully convolutional denseNets for semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, vol. 2017-July, pp. 1175–1183.
[324] S. Liu, H. Li, G. Zhang, B. Hu, and J. Chen, "Using hyperspectral reconstruction for multispectral images change detection," in

2022 7th International Conference on Image, Vision and Computing (ICIVC), 2022, pp. 183–188.
[325] L. Deng *et al.*, "M2H-Net: A reconstruction method for hyperspectral remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, no. February, pp. 323–348, Mar. 2021.
[326] T. Li and Y. Gu, "Progressive spatial–spectral joint network for hyperspectral image reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[327] X. Zheng, W. Chen, and X. Lu, "Spectral super-resolution of multispectral images using spatial–spectral residual attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[328] Z. Liu, H. Zhu, and Z. Chen, "Adversarial spectral super-resolution for multispectral imagery using spatial spectral feature attention module," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 1550–1562, 2023.
[329] W. Li, R. Dong, H. Fu, and L. Yu, "Large-scale oil palm tree detection from high-resolution satellite images using two-stage convolutional neural networks," *Remote Sens.*, vol. 11, no. 1, p. 11, 2018.
[330] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.
[331] K. Arai, K. Gondoh, O. Shigetomi, and - Yuko, "Method for NIR reflectance estimation with visible camera data based on regression for NDVI estimation and its application for insect damage detection of rice paddy fields," *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 11, pp. 17–22, 2016.
[332] M. Aslahishahri *et al.*, "From RGB to NIR: Predicting of near infrared reflectance from visible spectrum aerial images of crops," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, vol. 2021-Octob, pp. 1312–1322.
[333] MAXAR, "Satellite imagery for natural disasters." [Online]. Available: https://www.maxar.com/open-data. [Accessed: 17-Jul-2023].
[334] D. C. de Lima, D. Saqui, S. Ataky, L. A. de C. Jorge, E. J. Ferreira, and J. H. Saito, "Estimating agriculture NIR images from Aerial RGB data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11536 LNCS, 2019, pp. 562–574.
[335] A. Picon *et al.*, "Deep convolutional neural network for damaged vegetation segmentation from RGB images based on virtual NIR-channel estimation," *Artif. Intell. Agric.*, vol. 6, pp. 199–210, 2022.
[336] B. Zhang, T. Chen, and B. Wang, "Curriculum-style local-to-global adaptation for cross-domain Remote Sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
[337] Y. Cai *et al.*, "BiFDANet: Unsupervised bidirectional domain adaptation for semantic segmentation of Remote Sensing images," *Remote Sens.*, vol. 14, no. 1, p. 190, Jan. 2022.
[338] L. Wu, M. Lu, and L. Fang, "Deep covariance alignment for domain adaptive Remote Sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
[339] L. Wang, P. Xiao, X. Zhang, and X. Chen, "A fine-grained unsupervised domain adaptation framework for semantic segmentation of Remote Sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 4109–4121, 2023.
[340] Y. Zhao, P. Guo, Z. Sun, X. Chen, and H. Gao, "ResiDualGAN: Resize-residual DualGAN for cross-domain Remote Sensing images semantic segmentation," *Remote Sens.*, vol. 15, no. 5, p. 1428, Mar. 2023.
[341] S. F. Ismael, K. Kayabol, and E. Aptoula, "Unsupervised Domain Adaptation for the Semantic Segmentation of Remote Sensing Images via One-Shot Image-to-Image Translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[342] J. Fan, T. Chen, and S. Lu, "Unsupervised feature learning for land-use scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2250–2261, Apr. 2017.
[343] R. Jain and R. U. Sharma, "Airborne hyperspectral data for mineral mapping in Southeastern Rajasthan, India," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 81, no. January, pp. 137–145, Sep. 2019.
[344] X. Huang, D. Xu, Z. Li, and C. Wang, "Translating multispectral imagery to nightime imagery via conditional generative adversarial networks Department of Geography , University of South Carolina , Columbia , SC , U . S . A School of Geography Science , East China Normal University , Shanghai," *Igarss 2020*, no. January, pp. 6758–6761, 2020.
[345] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *ICML Work. Deep Learn. Audio, Speech Lang. Process.*, vol. 28, 2013.
[346] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image Generation," *arXiv Prepr. arXiv1611.02200*, Nov. 2016.
[347] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[348] Adobe Inc., "Adobe Photoshop." .
[349] D. W. Rouse, J.W. , Haas, R.H., SChell, J.A., Deering, "Monitoring vegetation systems in the great plains with ERTS," in *Third Earth Resources Technology Satellite-1 Symposium*, 1973, vol. 1, pp. 309–317.
[350] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. USA: Prentice-Hall, Inc., 2006.

List of Figures

List of Tables