Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

3Δ Ανακατασκευή Προσώπων από Εικόνες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ηλία Μήτσουρα

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής ΕΜΠ

Αθήνα, Μάρτιος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

3Δ Ανακατασκευή Προσώπων από Εικόνες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ηλία Μήτσουρα

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Μαρτίου 2022.

.....	.....	.....
Στέφανος Κόλλιας Καθηγητής ΕΜΠ	Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής ΕΜΠ	Γεώργιος Στάμου Καθηγητής ΕΜΠ

Αθήνα, Μάρτιος 2022

.....

Ηλίας Μήτσουρας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών ΕΜΠ

Περίληψη

Η 3Δ ανακατασκευή προσώπων από εικόνες αποσκοπεί στον ακριβή προσδιορισμό της 3Δ γεωμετρίας και των χαρακτηριστικών των εικονιζόμενων 2Δ προσώπων, έτσι ώστε να παραχθούν 3Δ μοντέλα όσο το δυνατόν πιο ρεαλιστικά και όμοια με αυτά. Παρ΄ όλη την πρόοδο που έχει σημειωθεί τα τελευταία χρόνια στην επιστήμη της Μηχανικής Μάθησης και της Όρασης Υπολογιστών και παρά τα υψηλά ποσοστά ακρίβειας και πιστότητας στην ανακατασκευή που έχουν επιτευχθεί από μεθόδους, οι οποίες βασίζονται αποκλειστικά στη χρήση βαθέων συνελικτικών νευρωνικών δικτύων, το πρόβλημα απέχει αρκετά από την οριστική επίλυσή του, κυρίως λόγω της πλούσιας δομικής ποικιλομορφίας και του υψηλού επιπέδου λεπτομερειών που παρουσιάζει το ανθρώπινο πρόσωπο.

Στην παρούσα εργασία υλοποιείται ένα σύστημα ανακατασκευής 3Δ προσώπων από 2Δ έγχρωμες εικόνες χαμηλής ανάλυσης, οι οποίες έχουν ληφθεί σε μη δεσμευμένο περιβάλλον, με αποτέλεσμα να παρουσιάζουν ποικίλες διακυμάνσεις ως προς την πόζα και την έκφραση των εικονιζόμενων προσώπων καθώς και τις συνθήκες φωτισμού. Το προτεινόμενο σύστημα ανακατασκευής αποτελείται από ένα κατάλληλα σχεδιασμένο συνελικτικό νευρωνικό δίκτυο, το οποίο και εκπαιδεύεται χωρίς επίβλεψη. Για την 3Δ αναπαράσταση του ανθρώπινου προσώπου χρησιμοποιείται κατάλληλο 3D Morphable Face Model, το οποίο κωδικοποιεί την πληροφορία για το σχήμα, την έκφραση και την υφή ενός προσώπου μέσω αντίστοιχων διανυσμάτων συντελεστών.

Δοθείσης μιας 2Δ εικόνας στην είσοδο, το δίκτυο προσδιορίζει τα διανύσματα συντελεστών σχήματος, έκφρασης και υφής, τα οποία χρησιμοποιούνται στη συνέχεια για το σχηματισμό του 3Δ μοντέλου του προσώπου. Η εκπαίδευση πραγματοποιείται χωρίς επίβλεψη, ελαχιστοποιώντας μια υβριδική συνάρτηση απώλειας, η οποία απαρτίζεται από κατάλληλα επιλεγμένες επιμέρους συναρτήσεις απώλειας, έτσι ώστε αυτές να εξετάζουν το μεγαλύτερο εύρος του φάσματος των χαρακτηριστικών του ανθρώπινου προσώπου.

Τα ποιοτικά και ποσοτικά αποτελέσματα που προκύπτουν έπειτα από εφαρμογή του προτεινόμενου εκπαιδευμένου δικτύου σε πραγματικά δεδομένα, αποδεικνύουν την δυνατότητα σθεναρούς και λεπτομερούς ανακατασκευής 3Δ προσώπων από 2Δ εικόνες, φανερώνοντας ωστόσο την ευαισθησία της όλης διαδικασίας σε έντονες πόζες, εκφράσεις και κακές συνθήκες φωτισμού.

Λέξεις-Κλειδιά: 3Δ ανακατασκευή προσώπων, 3D Morphable Model, Φωτισμός, Βαθιά Μάθηση, Όραση Υπολογιστών.

Abstract

The task of 3D face reconstruction from images aims to obtain the detailed 3D geometry and features of a 2D face of a person, in order to produce a realistic 3D face model that gives strong depiction of the person's identity. Despite the advances of Machine Learning and the high rates of reconstruction accuracy and fidelity achieved by methods based on convolutional neural networks, the problem of 3D face reconstruction is far from being permanently solved, with the existing methods falling short when it comes to the depiction of some of the facial features, mainly due to the rich structural diversity and the high level details that characterize the human face.

In the present thesis a system for reconstructing 3D faces solely from low resolution 2D RGB images is implemented. The 2D images are captured in an unconstrained environment, under an unknown and diverse variation of poses, expressions and illuminations. The proposed reconstruction system consists of a properly designed convolutional neural network, which is trained via unsupervised learning. For the 3D representation of the human face, a suitable 3D Morphable Model is used, which encodes the information about shape, expression and texture of a face using appropriate coefficient vectors.

Given a 2D image as input, the network estimates shape, expression, and texture coefficients for the depicted face, which are then used to form the corresponding 3D face model (3D morph). The training process is implemented in an usupervised manner, by minimizing a hybrid loss function, comprised of carefully selected sub-loss functions, which examine a wide range of facial features.

The qualitative and quantitative results obtained after using the proposed trained network in real data, prove that it is possible to robustly recover the detailed 3D face geometry from 2D facial images, revealing at the same time that the reconstruction process is very sensitive in extreme poses, expressions and illumination variations of the depicted person's face.

Keywords: 3D face reconstruction, 3D Morphable Model, Illumination, Deep Learning, Computer Vision.

Ευχαριστίες

Η παρούσα διπλωματική εργασία σηματοδοτεί το τέλος των προπτυχιακών σπουδών μου στη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, κατά τη διάρκεια των οποίων απέκτησα σημαντικές εμπειρίες και γνώσεις, τόσο σε επιστημονικό όσο και σε ανθρωπιστικό επίπεδο. Τα υλικά και πνευματικά αυτά εφόδια συνέβαλαν καθοριστικά στη διαμόρφωση της προσωπικότητάς μου ως επιστήμονα και κυρίως ως υπεύθυνου ατόμου. Στο πλαίσιο αυτό, θα ήθελα να ευχαριστήσω όλους όσουν με στήριξαν και συνέβαλαν στην επιτυχή περάτωση της πολυετούς αυτής πορείας.

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ.Ανδρέα-Γεώργιο Σταφυλοπάτη, για την ευκαιρία που μου προσέφερε να εργαστώ πάνω στο παρόν θέμα στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης.

Έπειτα, θα ήθελα να ευχαριστήσω ιδιαιτέρως τον κ.Γεώργιο Σιόλα, για την άριστη συνεργασία που είχαμε και την εξαιρετική αμεσότητα και προθυμία του στην επίλυση των όποιων αποριών και ερωτημάτων είχα καθόλη τη διάρκεια της εκπόνησης της διπλωματικής.

Τέλος, ιδιαίτερες ευχαριστίες εκφράζω στους φίλους μου και κυρίως στους γονείς και τα αδέρφια μου, χωρίς τους οποίους όλη αυτή η πορεία δεν θα ήταν εφικτή.

Ηλίας Μήτσουρας

Αθήνα, Μάρτιος 2022

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά

Η ακριβής και λεπτομερής 3Δ ανακατασκευή αντικειμένων από εικόνες αποτελεί ένα μακροχρόνιο πρόβλημα, το οποίο παρουσιάζει σημαντικότατο ενδιαφέρον στους τομείς των γραφικών και της υπολογιστικής όρασης. Σκοπός της ανακατασκευής είναι ο προσδιορισμός της 3Δ γεωμετρίας και των χαρακτηριστικών των εικονιζόμενων αντικειμένων και ο μετέπειτα κατάλληλος συνδυασμός τους για το σχηματισμό αντίστοιχων 3Δ μοντέλων.

Στο πλαίσιο αυτό, μία από τις σημαντικότερες κατηγορίες ανακατασκευής είναι αυτή του ανθρώπινου προσώπου, το οποίο σε αντίθεση με τα κοινά αντικείμενα χαρακτηρίζεται από υψηλό επίπεδο λεπτομερειών και ποικιλομορφίας. Έτσι, δεδομένης μιας 2Δ εικόνας ενός ατόμου, η διαδιακασία της ανακατασκευής στοχεύει στον υπολογισμό της 3Δ γεωμετρίας και της υφής του προσώπου του, έτσι ώστε να παραχθεί ένα ρεαλιστικό 3Δ μοντέλο προσώπου, το οποίο να μοιάζει όσο το δυνατόν περισσότερο με το άτομο αυτό. Η εν λόγω διαδικασία, της ανάκτησης της 3Δ τοπολογίας του προσώπου από 2Δ εικόνες, χρησιμοποιείται σε πληθώρα εφαρμογών, όπως η 3Δ αναγνώριση προσώπων (3D assisted-face recognition) [6, 7], η 3Δ αναγνώριση εκφράσεων (3D expression recognition) [8], η δημιουργία 3Δ avatars για χρήση σε περιβάλλοντα εικονικής και επαυξημένης πραγματικότητας (VR/AR) [1], η σύνθεση εικόνων [2, 3], η δημιουργία ομιλούντων χαρακτήρων μέσω φωνητικής οδήγησης (speech-driven facial animation) [4, 5], κ.ά.

Η πληθώρα των εφαρμογών στις οποίες χρησιμοποιείται, σε συνδυασμό με την επιτακτική ανάγκη για απόκτηση ρεαλιστικών μοντέλων προσώπων, ανεξάρτητων των εκφράσεων και των αντίστοιχων μεταβαλλόμενων χαρακτηριστικών, καθιστούν τη διαδικασία της 3Δ ανακατασκευής προσώπων ιδιαιτέρως σημαντική και απαραίτητη.

Η ραγδαία εξέλιξη του τεχνολογικού υλικού και λογισμικού καθώς και της επιστήμης της Μηχανικής Μάθησης, έχει συμβάλλει τα μέγιστα στην προσέγγιση του προβλήματος της ανακατασκευής προσώπων από εικόνες, αναπτύσσοντας μεθόδους, οι οποίες οδηγούν σε ιδιαιτέρως ικανοποιητικά αποτελέσματα. Ωστόσο, όπως αναφέρθηκε και παραπάνω, το ανθρώπινο προσώπο χαρακτηρίζεται από υψηλή πολυπλοκότητα, ενώ συγχρόνως δεν αποτελεί μηχανικό σώμα (non-rigid body), καθώς οποιαδήποτε εκφραστική διακύμανση ή επίδραση του εξωτερικού περιβάλλοντος επηρεάζει άμεσα τη μορφή του. Το γεγονός αυτό, καθιστά τη διαδικασία της ανακατασκευής ιδιαιτέρως ευαίσθητη σε εξωτερικούς παράγοντες, οδηγώντας αρκετές φορές σε μη ρεαλιστικά ή εσφαλμένα αποτελέσματα.

Στην παρούσα διπλωματική γίνεται μια προσπάθεια προσέγγισης του προβλήματος της 3Δ ανακατασκευής προσώπων από έγχρωμες 2Δ εικόνες, οι οποίες προέρχονται από μη ελεγχόμενα περιβάλλοντα (in-the-wild images). Σκοπός της ανακατασκευής είναι ο προσδιορισμός της 3Δ γεωμετρίας και της υφής ενός προσώπου μιας 2Δ εικόνας, ανεξαρτήτως πόζας, έκφρασης και συνθηκών φωτισμού. Η όλη διαδικασία της ανακατασκευής βασίζεται στη χρήση συνελικτικών νευρωνικών δικτύων, τα οποία και εκπαιδεύονται χωρίς επίβλεψη, ενώ για την αναπαράσταση της 3Δ τοπολογίας του ανθρώπινου προσώπου χρησιμοποιείται το στατιστικό μοντέλο 3D Basel Face Model (ενότητα 2.1.3).

Το δίκτυο ανακατασκευής αποτελείται από ένα κατάλληλα επιλεγμένο συνελικτικό νευρωνικό δίκτυο, το οποίο δοθείσης μιας εικόνας στην είσοδο, παράγει ένα διάνυσμα συντελεστών, το οποίο και κωδικοποιεί όλη την πληροφορία για το σχήμα, την έκφραση και την υφή του εικονιζόμενου προσώπου, τις συνθήκες φωτισμού και την τοποθέτηση της κάμερας. Οι συντελεστές αυτοί αξιοποιούνται στη συνέχεια, έτσι ώστε να σχηματιστεί η 3Δ δομή του προσώπου της εικόνας. Το δίκτυο εκπαιδεύεται χωρίς επίβλεψη σε ένα σύνολο προεπεξεργασμένων εικόνων εισόδου, έτσι ώστε να ελαχιστοποιείται μια κατάλληλα επιλεγμένη υβριδική συνάρτηση απώλειας.

Τα αποτελέσματα που προκύπτουν, αποδεικνύουν την δυνατότητα σθεναρούς και λεπτομερούς ανακατασκευής 3Δ προσώπων από 2Δ εικόνες, αποκαλύπτοντας συγχρόνως την ευαισθησία της όλης διαδιακασίας σε έντονες πόζες, εκφράσεις και μεταβολές στο φωτισμό των εικονιζόμενων προσώπων.

1.2 Δομή της Εργασίας

Η παρούσα διπλωματική διαρθρώνεται σε 5 κεφάλαια, εκ των οποίων το **κεφάλαιο 1** αποτελεί μια πρώτη εισαγωγή και μια γενική περιγραφή του εξεταζόμενου προβλήματος, παρουσιάζοντας συνοπτικά τα βασικά του μέρη καθώς και την προσέγγιση που ακολουθείται για την αντιμετώπισή του.

Τα υπόλοιπα κεφάλαια της διπλωματικής οργανώνονται ως εξής:

- **κεφάλαιο 2**: Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο το οποίο είναι απαραίτητο για την κατανόηση της διαδικασίας της ανακατασκευής. Έτσι, στο πρώτο μέρος του αναλύονται τα στατιστικά μοντέλα που χρησιμοποιούνται για την αναπαράσταση της 3Δ δομής του ανθρώπινου προσώπου, ενώ στο δεύτερο, γίνεται μια εκτενής αναφορά στις υπάρχουσες μεθόδους ανακατασκευής, στα βασικά χαρακτηριστικά και τις διαφορές τους.
- **κεφάλαιο 3**: Στο κεφάλαιο αυτό παρουσιάζεται η αυτή καθαυτή προτεινόμενη μέθοδος ανακατασκευής. Αρχικά περιγράφεται το στάδιο της προεπεξεργασίας των εικόνων του dataset, ενώ στη συνέχεια πραγματοποιείται η παραμετροποίηση του περιβάλλοντος. Τέλος, αναλύονται τα επιμέρους στάδια της ανακατασκευής καθώς και η διαδικασία της εκπαίδευσης του δικτύου.
- **κεφάλαιο 4**: Το κεφάλαιο αυτό αποτελεί το πειραματικό μέρος της εργασίας. Έτσι, προσδιορίζονται αρχικά οι παράμετροι που χρησιμοποιούνται κατά την εκπαίδευση του δικτύου και κατόπιν παρατίθενται και σχολιάζονται τα ποιοτικά και ποσοτικά αποτελέσματα που προκύπτουν έπειτα από εφαρμογή του εκπαιδευμένου δικτύου σε πραγματικά δεδομένα.
- **κεφάλαιο 5**: Στο κεφάλαιο αυτό γίνεται μια συνοπτική ανακεφαλαίωση των όσων αναλύθηκαν και αναπτύχθηκαν στα προηγούμενα κεφάλαια και συνοψίζεται η συμπερασματολογία που προκύπτει από τα αποτελέσματα του κεφαλαίου 4. Τέλος, γίνεται μια σύντομη αναφορά σε τρόπους επέκτασης της λειτουργίας του προτεινόμενου δικτύου καθώς και στις κατευθύνσεις που ακολουθεί η έρευνα στον τομέα της 3Δ ανακατασκευής προσώπων.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 3D Morphable Face Model

2.1.1 Γενική Μορφή Μοντέλου

Το πρόβλημα της μοντελοποίησης του ανθρώπινου προσώπου έχει προσελκύσει από τις απαρχές του το ενδιαφέρον αρκετών ερευνητών στον τομέα της υπολογιστικής όρασης. Κατά τη διάρκεια των τελευταίων χρόνων, έχουν αναπτυχθεί και προταθεί αρκετά μοντέλα για την αναπαράσταση και την απόδοση της 3Δ δομής και της υφής του προσώπου του ανθρώπου, με σημαντικότερο εξ αυτών το *3D Morphable Face Model*.

Με τον όρο 3D Morphable Face Model (3DMM) αναφερόμαστε σε ένα γενετικό (generative) μοντέλο σχημάτων και υφών προσώπου, η λειτουργία του οποίου βασίζεται στις εξής δύο παραδοχές:

- Πρώτον, όλα τα πρόσωπα χαρακτηρίζονται από μία πυκνή σημείου προς σημείο αντιστοιχία. Η ιδιότητα αυτή επιτρέπει την παραγωγή ρεαλιστικών προσώπων (morphs), μέσω κατάλληλων γραμμικών συνδυασμών ήδη υπαρχόντων προσώπων.

- Δεύτερον και προκειμένου να παραχθούν ικανοποιητικά αποτελέσματα, το σχήμα του προσώπου θα πρέπει να διαχωριστεί από το χρώμα και οι παράμετροι αυτές θα πρέπει να μην εξαρτώνται από εξωτερικούς παράγοντες, όπως οι συνθήκες φωτισμού και το είδος της κάμερας.

Καθώς το 3D Morphable Face Model πρόκειται εν γένει για ένα στατιστικό μοντέλο, κρίνεται σκόπιμο να αναλυθεί το μαθηματικό του υπόβαθρο, έτσι ώστε να γίνει κατανοητός ο τρόπος με τον οποίο επιτυγχάνεται η 3Δ αναπαράσταση και απόδοση των προσώπων.

Προκειμένου λοιπόν να αναπαρασταθεί η γεωμετρία ενός προσώπου, ορίζεται ένα διάνυσμα-σχήματος

$$\mathbf{S} = \left[\mathbf{v}_1^T, \ldots, \mathbf{v}_N^T\right]^T = (X_1, Y_1, Z_1, X_2, \ldots, Y_N, Z_N)^T \in \mathbb{R}^{3N}, \tag{2.1}$$

διαστάσεων $[3N, 1]$, το οποίο περιλαμβάνει τις συντεταγμένες X, Y, Z των N σημείων από τα οποία αυτό αποτελείται.

Με αντίστοιχο τρόπο, η υφή του προσώπου αναπαρίσταται με ένα διάνυσμα-υφής

$$\mathbf{T} = \left[\mathbf{r}_1^T, \ldots, \mathbf{r}_n^T\right]^T = (R_1, G_1, B_1, R_2, \ldots, G_N, B_N)^T \in \mathbb{R}^{3N}, \tag{2.2}$$

διαστάσεων επίσης $[3N, 1]$, το οποίο περιλαμβάνει τις R, G, B τιμές των χρωμάτων των N αντίστοιχων σημείων του.

Χρησιμοποιώντας τώρα τις εξισώσεις 2.1 και 2.2, νέα σχήματα $\boldsymbol{S}_{model}$ και υφές $\boldsymbol{T}_{model}$ προσώπων μπορούν να προκύψουν μέσω κατάλληλων γραμμικών συνδυασμών των σχημάτων και των υφών των M αρχικά διαθέσιμων προτύπων προσώπων, ως εξής:

$$\mathbf{S}_{model} = \sum_{i=1}^{M} a_i \mathbf{S}_i, \quad \mathbf{T}_{model} = \sum_{i=1}^{M} b_i \mathbf{T}_i, \quad \sum_{i=1}^{M} a_i = \sum_{i=1}^{M} b_i = 1, \tag{2.3}$$

όπου a_i και b_i οι συντελεστές βαρύτητας του κάθε πρότυπου προσώπου i, με $1 \leq i \leq M$.

Κατά αυτόν τον τρόπο, το 3D Morphable Face Model ορίζεται μαθηματικά ως το σύνολο των προσώπων $(\mathbf{S}_{model}(\boldsymbol{a}), \mathbf{T}_{model}(\boldsymbol{b}))$, με συντελεστές παραμετροποίησης $\boldsymbol{a} = (a_1, a_2, \ldots, a_M)^T$ και $\boldsymbol{b} = (b_1, b_2, \ldots, b_M)^T$, η μεταβολή των οποίων οδηγεί κάθε φορά σε τυχαία συνθετικά πρόσωπα.

Προκειμένου το σύστημα σύνθεσης προσώπων να είναι εύχρηστο και αξιόπιστο, θα πρέπει να υποβληθεί σε στατιστική επεξεργασία, έτσι ώστε να μειωθεί αφενός η διαστατικότητα των δεδομένων του και να εξασφαλιστεί αφετέρου η παραγωγή αληθοφανών και ρεαλιστικών προσώπων. Για λόγους απλότητας και οικονομίας, η ακριβής διαδικασία επεξεργασίας του συνόλου δεδομένων των προσώπων του 3D Morphable Face Model παραλείπεται, καθώς αυτή περιγράφε-

ται λεπτομερώς στο paper των Blanz και Vetter [9], οι οποίοι παρουσίασαν για πρώτη φορά την ιδέα του μοντέλου αυτού, πριν από περίπου 23 χρόνια.

Παρά ταύτα, αξίζει να αναφερθεί πως η βασική στατιστική διαδικασία που χρησιμοποιείται για την συμπίεση της διάστασης των δεδομένων, είναι η *Ανάλυση Κύριων Συνιστωσών* (Principal Component Analysis - PCA), η οποία μετατρέπει μία ομάδα τιμών (παρατηρήσεων) δυνητικά συσχετιζόμενων μεταβλητών, σε μία ομάδα νέων τιμών μη γραμμικά συσχετιζόμενων μεταβλητών οι οποίες καλούνται κύριες συνιστώσες. Στην πράξη, η μετάβαση αυτή γίνεται μέσω ενός ορθογώνιου μετασχηματισμού (αλλαγή βάσης), ο οποίος οδηγεί σε ένα ορθοκανονικό σύστημα συντεταγμένων, του οποίου οι βάσεις (κύριες συνιστώσες) ορίζονται με τέτοιο τρόπο, ώστε η πρώτη συνιστώσα να εξηγεί τη μέγιστη δυνατή διακύμανση που αναπτύσσεται μεταξύ των αρχικών μεταβλητών, η δεύτερη, μη συσχετιζόμενη με την πρώτη, να εξηγεί ένα σημαντικό μέρος αυτής αλλά πάντα μικρότερο της πρώτης κ.ο.κ.

Εφαρμόζοντας λοιπόν τη διαδικασία της Ανάλυσης Κύριων Συνιστωσών στα δεδομένα σχήματος $\boldsymbol{S}_i$ και υφής $\boldsymbol{T}_i$ των διαθέσιμων προσώπων του dataset, οι σχέσεις σχήματος και υφής της εξίσωσης 2.3 των νέων συνθετικών προσώπων τροποποιούνται ως εξής:

$$\mathbf{S}_{model} = \overline{\mathbf{S}} + \sum_{i=1}^{d_s} \boldsymbol{u}_i^{[s]} \sigma_i^{[s]} \alpha_i = \overline{\mathbf{S}} + \mathbf{U}_s \boldsymbol{\alpha}, \quad \text{με} \quad 1 \leq d_s \leq M-1 \tag{2.4}$$

και

$$\mathbf{T}_{model} = \overline{\mathbf{T}} + \sum_{i=1}^{d_t} \mathbf{u}_i^{[t]} \sigma_i^{[t]} \beta_i = \overline{\mathbf{T}} + \mathbf{U}_t \boldsymbol{\beta}, \quad \text{με} \quad 1 \leq d_t \leq M-1, \tag{2.5}$$

όπου $\overline{\mathbf{S}}$, $\overline{\mathbf{T}} \in \mathbb{R}^{3N}$ το μέσο σχήμα και η μέση υφή των προσώπων αντίστοιχα, $\mathbf{U}_s \in \mathbb{R}^{3N \times d_s}$, $\mathbf{U}_t \in \mathbb{R}^{3N \times d_t}$ τα ορθοκανονικά μητρώα βάσεων σχήματος και υφής των δύο μοντέλων PCA, των οποίων οι στήλες περιέχουν τα κανονικοποιημένα ιδιοδιανύσματα σχήματος $\boldsymbol{u}_i^{[s]} \sigma_i^{[s]}$ και υφής $\mathbf{u}_i^{[t]} \sigma_i^{[t]}$ αντίστοιχα, $\sigma_i^{[s]}, \sigma_i^{[t]} \in \mathbb{R}$ οι τυπικές αποκλίσεις για τα διανύσματα βάσης του σχήματος και υφής και $\boldsymbol{\alpha} \in \mathbb{R}^{d_s}$, $\boldsymbol{\beta} \in \mathbb{R}^{d_t}$, τα διανύσματα που περιέχουν τους συντελεστές σχήματος και υφής αντίστοιχα, οι οποίοι σχηματίζουν ένα μοναδικό κάθε φορά μοντέλο ποσώπου. Οι βαθμοί ελευθερίας του τροποποιημένου αυτού μοντέλου (Εξ.2.4-2.5) προσδιορίζονται από το πλήθος των κύριων συνιστωσών d_s και d_t, το οποίο επιλέγεται στη γενική περίπτωση έτσι ώστε να μειώνει τη διαστατικότητα

του αρχικού μοντέλου (Εξ.2.3), χωρίς συγχρόνως να χάνεται σημαντικό μέρος της πληροφορίας του.

Στο σημείο αυτό αξίζει να επισημανθεί, ότι όλες οι ορθοκανονικές βάσεις που προέκυψαν με τη μέθοδο PCA έχουν κλιμακωθεί με τις κατάλληλες τυπικές αποκλίσεις, έτσι ώστε να ισχύουν οι σχέσεις

$$\mathbf{U}_i^T \mathbf{U}_i = diag\left(\sigma_1^{[i]}, \ldots, \sigma_k^{[i]}\right), \tag{2.6}$$

όπου $k = d_s, d_t$ για $i = \{\text{s, t}\}$ αντίστοιχα.

2.1.2 Αποτύπωση Εκφράσεων Προσώπου

Η σχετικά απλή μορφή του 3D Morphable Face Model η οποία αναλύθηκε στην προηγούμενη ενότητα, είναι μεν αρκετά αποτελεσματική στην αναπαράσταση της βασικής γεωμετρίας του ανθρώπινου προσώπου καθώς και της υφής αυτού (χρωματική απόχρωση δέρματος), υστερεί ωστόσο σε ένα πολύ σημαντικό πεδίο, το οποίο μάλιστα είναι άρρηκτα συνυφασμένο με την ανθρώπινη φύση, αυτό των εκφράσεων. Εφόσον λοιπόν οι εκφράσεις του προσώπου ενός ατόμου αποτελούν αναπόσπαστο τμήμα της ταυτότητάς του και καθώς υπάρχουν δυναμικά χαρακτηριστικά του προσώπου τα οποία μεταβάλλονται με αυτές (π.χ. ρυτίδες, λακάκια κ.α.), κρίνεται απαραίτητη η επέκταση του βασικού 3D Morphable Face Model, έτσι ώστε αυτό να μπορεί να αποτυπώνει και τυχόν εκφραστικές διακυμάνσεις.

Προς την κατεύθυνση αυτή και δεδομένου ότι οι εκφράσεις του προσώπου εντάσσονται στη γενική κατηγορία των γεωμετρικών ιδιοτήτων του, γίνεται επέκταση της σχέσης της Εξ.2.4 του αρχικού μοντέλου ως εξής:

$$\mathbf{S}_{model} = \overline{\mathbf{S}} + \sum_{i=1}^{d_s} \boldsymbol{u}_i^{[s]} \sigma_i^{[s]} \alpha_i + \sum_{i=1}^{d_e} \boldsymbol{u}_i^{[e]} \sigma_i^{[e]} \delta_i = \overline{\mathbf{S}} + \mathbf{U}_s \boldsymbol{\alpha} + \mathbf{U}_e \boldsymbol{\delta}, \tag{2.7}$$

όπου $\overline{\mathbf{S}}$ το μέσο σχήμα των προσώπων, $\mathbf{U}_s \in \mathbb{R}^{3N \times d_s}$, $\mathbf{U}_e \in \mathbb{R}^{3N \times d_e}$ τα ορθοκανονικά μητρώα βάσεων σχήματος και έκφρασης των δύο μοντέλων PCA, των οποίων οι στήλες περιέχουν τα κανονικοποιημένα ιδιοδιανύσματα σχήματος $\boldsymbol{u}_i^{[s]} \sigma_i^{[s]}$ και έκφρασης $\mathbf{u}_i^{[e]} \sigma_i^{[e]}$ αντίστοιχα, $\sigma_i^{[s]}, \sigma_i^{[e]} \in \mathbb{R}$ οι τυπικές αποκλίσεις για

τα διανύσματα βάσης του σχήματος και έκφρασης και $\boldsymbol{\alpha} \in \mathbb{R}^{d_s}$, $\boldsymbol{\delta} \in \mathbb{R}^{d_e}$, τα διανύσματα που περιέχουν τους συντελεστές σχήματος και έκφρασης αντίστοιχα.

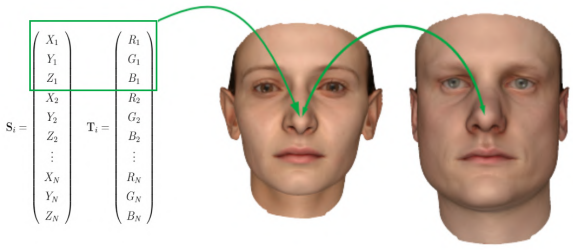
Με την προσθήκη αυτή, συνυπολογίζεται στο γεωμετρικό μέρος του μοντέλου, το οποίο πλέον αναπαρίσταται από το τροποποιημένο διάνυσμα σχήματος-έκφρασης, τόσο το σχήμα όσο και οι εκφράσεις του προσώπου, ενώ η υφή εξακολουθεί να περιγράφεται μέσω της σχέσης της Εξ.2.5. Κατά αυτό τον τρόπο κατασκευάζεται ένα πιο πλήρες και ακριβές μοντέλο αναπαράστασης, το οποίο είναι ικανό να αποτυπώσει με πιο λεπτομερή και αποτελεσματικό τρόπο την 3Δ δομή και τα ιδιαίτερα χαρακτηριστικά του προσώπου ενός ατόμου.

Τέλος και πριν ολοκληρωθεί η ενότητα αυτή, σημειώνεται ότι στην παρούσα εργασία, η 3Δ αναπαράσταση του ανθρώπινου προσώπου γίνεται μέσω του επεκταμένου 3D Morphable Face Model, το οποίο και περιγράφεται από τις σχέσεις των Εξ.2.5 και 2.7.

2.1.3 3D Basel Face Model

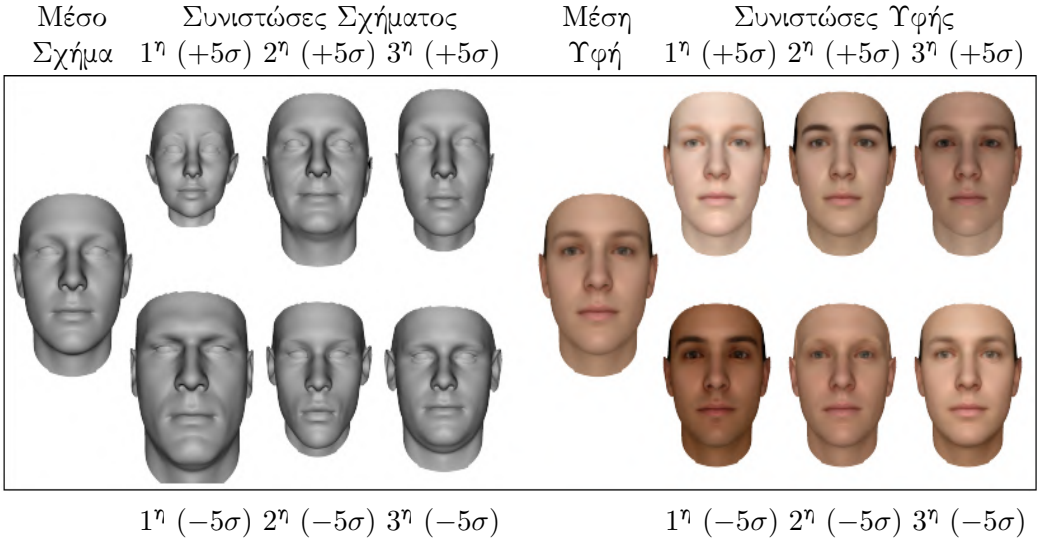
Σύμφωνα με όσα αναφέρθηκαν έως τώρα, η κατασκευή ενός 3D Morphable Face Model απαιτεί την ύπαρξη ενός dataset προσώπων, στο οποίο και εφαρμόζεται η στατιστική επεξεργασία που αναλύθηκε στην ενότητα 2.1.1. Η δημιουργία ενός τέτοιου dataset αποτελεί μια δύσκολη και σχετικά περίπλοκη διαδικασία, καθώς προϋποθέτει την ύπαρξη ενός ακριβούς και γρήγορου 3D face scanner, με τον οποίο γίνεται το σκανάρισμα των προσώπων αρκετών εκατοντάδων ατόμων, καθώς και τον υπολογισμό της πυκνής αντιστοίχισης μεταξύ των προσώπων αυτών.

Προς την κατεύθυνση αυτή, έχουν προταθεί και κατασκευαστεί αρκετά μοντέλα, καθένα εκ των οποίων χρησιμοποιεί μια ελαφρώς τροποποιημένη αναπαράσταση για το πρόσωπο, καθώς και διαφορετικό dataset στο οποίο εφαρμόζεται η στατιστική επεξεργασία. Στην παρούσα εργασία χρησιμοποιείται το *3D Basel Face Model του 2009* [37], σύμφωνα με το οποίο το ανθρώπινο πρόσωπο αναπαρίσταται μέσω ενός τριγωνικού πλέγματος (triangle mesh) με $N = 53490$ *σημεία*, τα οποία σχηματίζουν συνολικά 160470 *τρίγωνα*. Το dataset του εν λόγω μοντέλου αποτελείται από 200 συνολικά face scans, εκ των οποίων τα μισά αντιστοιχούν σε γυναίκες και τα άλλα μισά σε άντρες, ηλικίας μεταξύ 8 και 62 ετών και σωματικού βάρους μεταξύ 40 και 123 κιλών.



Σχήμα 2.1: Αντιστοίχιση διανυσμάτων σχήματος-υφής με σημεία του προσώπου. Κάθε τριάδα συντεταγμένων και χρώματος, αντιστοιχεί στην ίδια πάντα περιοχή των παραγόμενων προσώπων.

Αφού λοιπόν πραγματοποιηθεί κατάλληλη προεπεξεργασία στο dataset, έτσι ώστε να παγιωθούν στο παραμετρικό μοντέλο οι θέσεις των περιοχών του προσώπου (Σχ.2.1) και αφού υπολογισθεί με κατάλληλους αλγορίθμους η σημείου προς σημείο αντιστοιχία μεταξύ όλων των διαθέσιμων face scans, τα επεξεργασμένα πλέον face scans μπορούν να χρησιμοποιηθούν στις σχέσεις των Εξ.2.5 και 2.7, έτσι ώστε να παραχθούν νέα πρόσωπα.



Σχήμα 2.2: Μέσο σχήμα και υφή, μαζί με τις αντίστοιχες τρεις πρώτες κύριες συνιστώσες τους, όπως αυτές υπολογίζονται με τη μέθοδο PCA, για διαφορετικές τιμές τυπικών αποκλίσεων ($\pm\sigma$).

2.2 Υπάρχουσες Μέθοδοι

Όπως αναφέρθηκε και στην εισαγωγή, το πρόβλημα της 3Δ ανακατασκευής προσώπων από 2Δ εικόνες αποτελεί στη γενική του περίπτωση ένα μη καλώς τεθειμένο πρόβλημα (ill posed problem), γεγονός το οποίο σε συνδυασμό με την ευρεία χρήση του σε πληθώρα εφαρμογών, το έχει καθιερώσει ως αντικείμενο ιδιαίτερης μελέτης και έρευνας στους τομείς των γραφικών και της υπολογιστικής όρασης.

Στο πλαίσιο αυτό, έχουν προταθεί διάφορες προσεγγίσεις για την αντιμετώπισή του, ξεκινώντας από κάποιες πρώιμες μεθόδους οι οποίες βασίζονται αποκλειστικά στη χρήση μαθηματικών εργαλείων και καταλήγοντας σε πιο σύγχρονες, οι οποίες εκμεταλλεύονται τις ιδιότητες και τις δυνατότητες των νευρωνικών δικτύων και της μηχανικής μάθησης, έτσι ώστε να επιτύχουν υψηλή ακρίβεια και πιστότητα στην ανακατασκευή. Γενικά, οι υπάρχουσες μέθοδοι για την 3Δ ανακατασκευή προσώπων από εικόνες μπορούν να διακριθούν στις εξής τρεις μεγάλες κατηγορίες:

- *Μέθοδοι βελτιστοποίησης*,
- *Μέθοδοι επιβλεπόμενης μάθησης* και
- *Μέθοδοι μη επιβλεπόμενης μάθησης*

2.2.1 Μέθοδοι Βελτιστοποίησης

Οι μέθοδοι της κατηγορίας αυτής προσεγγίζουν το πρόβλημα της 3Δ ανακατασκευής προσώπων από εικόνες ως ένα πρόβλημα βελτιστοποίησης, επιδιώκοντας αρχικά τη δημιουργία της 3Δ τοπολογίας ενός προσώπου (σχήμα, έκφραση, υφή) και έπειτα την προσαρμογή της τοπολογίας αυτής στο πρόσωπο της 2Δ εικόνας. Για τη δημιουργία της 3Δ τοπολογίας του προσώπου χρησιμοποιείται κάποιο μοντέλο αναπαράστασης όπως το 3D Morphable Face Model, το οποίο και αναλύθηκε στην ενότητα 2.1, ή το Active Appearance Model [11], το οποίο χρησιμοποιεί ένα γραμμικό μοντέλο για την από κοινού κωδικοποίηση της διακύμανσης του σχήματος και της υφής του προσώπου.

Σκοπός της ανακατασκευής είναι η εκτίμηση των βέλτιστων συντελεστών των μοντέλων αναπαράστασης, έτσι ώστε το παραγόμενο 3Δ πρόσωπο να προσεγγίζει κατά το δυνατόν το πρόσωπο της 2Δ εικόνας. Προκειμένου ωστόσο να καταστεί εφικτή η σύγκριση μεταξύ ενός 3Δ μοντέλου και μιας 2Δ εικόνας, θα πρέπει να γίνει rendering του 3Δ μοντέλου, έτσι ώστε αυτό να προβληθεί στην επιφάνεια μιας εικόνας, η οποία στη συνέχεια θα συγκριθεί με την αρχική εικόνα, με σκοπό τον υπολογισμό κάποιων συναρτήσεων απώλειας. Τελικά, η ελαχιστοποίηση των συναρτήσεων αυτών μέσω κατάλληλων μεθόδων μη-γραμμικής βελτιστοποίησης, θα οδηγήσει στην εύρεση των φαινομενικά βέλτιστων συντελεστών.

Συνολικά, η διαδικασία ανακατασκευής με χρήση μεθόδων βελτιστοποίησης φαίνεται στο Σχ.2.3 και περιλαμβάνει τα εξής στάδια:

1. Δημιουργία μιας 3Δ τοπολογίας προσώπου, με χρήση τυχαίων συντελεστών σχήματος, υφής και έκφρασης,

2. Rendering του 3Δ μοντέλου του προσώπου και σχηματισμός 2Δ εικόνας,

3. Υπολογισμός συναρτήσεων απώλειας μέσω σύγκρισης της συνθετικής και της αρχικής 2Δ εικόνας,

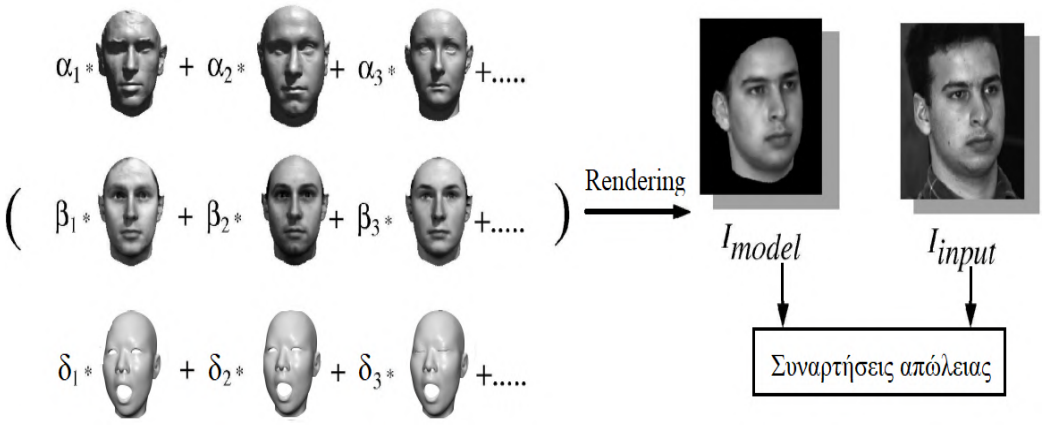
4. Ελαχιστοποίηση των συναρτήσεων απώλειας με χρήση κατάλληλων μαθηματικών μεθόδων.

Οι μέθοδοι βελτιστοποίησης αποτέλεσαν την πρώτη σχετικώς επιτυχημένη προσπάθεια αντιμετώπισης του προβλήματος της 3Δ ανακατασκευής προσώπων από εικόνες, σε μια εποχή όπου οι δυνατότητες του λογισμικού ήταν αρκετά περιορισμένες.

Εντούτοις, είναι ιδιαίτερα ευαίσθητες στις συνθήκες του περιβάλλοντος (φωτισμός, μετατοπίσεις και περιστροφές του προσώπου) και ως εκ τούτου δεν είναι αποτελεσματικές σε in-the-wild εικόνες, όπου το περιβάλλον λήψης είναι μη ελεγχόμενο, οδηγώντας σε μη ρεαλιστικά αποτελέσματα. Το γεγονός αυτό, σε συνδυασμό με την υπολογιστική τους πολυπλοκότητα, λόγω των επαναληπτικών αλγορίθμων που χρησιμοποιούν για τη βελτιστοποίηση, τις καθιστά μη αξιόπιστες ως προς την πιστότητα και την ακρίβεια των παραγόμενων προσώπων.

Την αδυναμία αυτή των μεθόδων βελτιστοποίησης ήρθε να αντιμετωπίσει η εξέλιξη του λογισμικού και κυρίως η θεμελίωση και ανάπτυξη των νευρωνικών δικτύων, η οποία προσέδωσε αξιοπιστία και σθεναρότητα (robustness) στη διαδικασία της ανακατασκευής, απεξαρτητοποιώντας τη από τις συνθήκες του

περιβάλλοντος, αναπτύσσοντας κατά αυτό τον τρόπο δίκτυα ικανά να ανταπεξέλθουν σε πολύ απαιτητικές εφαρμογές.



Σχήμα 2.3: Διαδικασία ανακατασκευής με μεθόδους βελτιστοποίησης (και χρήση στη συγκεκριμένη περίπτωση 3DMM). Σκοπός των μεθόδων είναι η ελαχιστοποίηση των συναρτήσεων απώλειας μεταξύ της αρχικής εικόνας I_{input} και της συνθετικής εικόνας I_{model}, ετσι ώστε να ευρεθούν οι βέλτιστοι συντελεστές σχήματος α_i, υφής β_i και έκφρασης δ_i. Για την ελαχιστοποίηση χρησιμοποιούνται κατάλληλες μαθηματικές μέθοδοι βελτιστοποίησης.

2.2.2 Μέθοδοι Επιβλεπόμενης Μάθησης

Κατά τη διάρκεια των τελευταίων χρόνων και έπειτα από την παρουσίαση του πρώτου μοντέλου νευρωνικού δικτύου και του νευρώνα ως της βασικής του μονάδας, από τους McCulloch και Pitts το 1943 [12], οι εξελίξεις στον τομέα της μηχανικής μάθησης υπήρξαν ραγδαίες, οδηγώντας στη χρήση και καθιέρωση των νευρωνικών δικτύων σε πληθώρα εφαρμογών.

Στο πλαίσιο αυτό, αναπτύχθηκαν διάφορες μέθοδοι για την αντιμετώπιση του προβλήματος της 3Δ ανακατασκευής προσώπων από εικόνες, οι οποίες βασίζονται κυρίως στη Βαθιά Μάθηση (Deep Learning) και εκμεταλλεύονται τις δυνατότητες που προσφέρει η χρήση των *συνελικτικών νευρωνικών δικτύων* (CNNs).

Για την επίλυση του προβλήματος της ανακατασκευής, οι μέθοδοι αυτές αποσκοπούν είτε στον προσδιορισμό των βέλτιστων συντελεστών των μοντέλων αναπαράστασης του προσώπου (3DMM, AAM), είτε στην άμεση απεικόνιση των pixels της 2Δ εικόνας σε σημεία του 3Δ πλέγματος του μοντέλου του προσώπου. Στόχος σε κάθε περίπτωση είναι το παραγόμενο πρόσωπο να ταυτίζεται όσο το δυνατόν περισσότερο με το πρόσωπο της εικόνας.

Η εκπαίδευση των συνελικτικών μοντέλων ανακατασκευής, μπορεί να πραγματοποιηθεί είτε μέσω επιβλεπόμενης είτε μέσω μη επιβλεπόπενης-ασθενώς επιβλεπόμενης μάθησης. Στην παρούσα ενότητα, θα αναλυθούν οι μέθοδοι οι οποίες για την εκπαίδευση των δικτύων τους χρησιμοποιούν τεχνικές επιβλεπόμενης μάθησης, ενώ στην ακόλουθη θα παρουσιαστούν οι μέθοδοι μη επιβλεπόμενης-ασθενώς επιβλεπόμενης μάθησης.

Στην Επιβλεπόμενη Μάθηση (Supervised Learning) η εκπαίδευση του δικτύου προϋποθέτει την ύπαρξη ενός συνόλου δεδομένων εκπαίδευσης (training dataset), τα οποία διαθέτουν ετικέτες (labels), έτσι ώστε για κάθε είσοδο να είναι γνωστή η επιθυμητή έξοδος. Στην περίπτωση του προβλήματος της 3Δ ανακατασκευής προσώπων ωστόσο, η δημιουργία ενός τέτοιου συνόλου δεδομένων είναι ιδιαιτέρως δύσκολη και απαιτητική, καθώς τα δεδομένα εισόδου είναι 2Δ εικόνες προσώπων και οι επιθυμητές έξοδοι οι αντίστοιχες 3Δ τοπολογίες. Τα περισσότερα υπάρχοντα σύνολα δεδομένων 2Δ προσώπων - 3Δ τοπολογιών [13, 14] προέρχονται από μερικές εκατοντάδες άτομα, γεγονός το οποίο τα καθιστά ακατάλληλα για την εκπαίδευση βαθέων συνελικτικών δικτύων. Σε μια ιδεατή περίπτωση, θα μπορούσε κανείς να σκανάρει με χρήση κάποιου 3D scanner τα πρόσωπα πολλών χιλιάδων ατόμων και στη συνέχεια να τα χρησιμοποιήσει για την εκπαίδευση του δικτύου ανακατασκευής. Ωστόσο, η διαδικασία αυτή είναι πρακτικώς αδύνατη, με αποτέλεσμα να απαιτείται ένας άλλος τρόπος για τη δημιουργία του επιθυμητού συνόλου δεδομένων.

Για την αντιμετώπιση του προβλήματος αυτού, έχουν προταθεί αρκετές τεχνικές, οι οποίες μοιράζονται την ιδέα της δημιουργίας ενός συνθετικού συνόλου δεδομένων για την εκπαίδευση των δικτύων ανακατασκευής. Για παράδειγμα, οι Richardson *et al.*[15] προτείνουν τη δημιουργία ενός συνθετικού συνόλου δεδομένων 3Δ προσώπων μέσω της χρήσης του 3DMM, παράγοντας αρχικά τυχαίες γεωμετρίες προσώπων, οι οποίες στη συνέχεια γίνονται rendered με τυχαίες συνθήκες φωτισμού και προβάλλονται πάνω στο επίπεδο της εικόνας. Μέσω της διαδικασίας αυτής δημιουργείται ένα σύνολο 2Δ εικόνων με γνωστές 3Δ γεωμετρίες, το οποίο μπορεί να χρησιμοποιηθεί για την επιβλεπόμενη μάθηση.

Οι Dou *et al.*[16] εκπαιδεύουν το μοντέλο ανακατασκευής, αξιοποιώντας τόσο πραγματικές όσο και συνθετικές 2Δ εικόνες. Οι πραγματικές εικόνες χρησιμοποιούνται για την αρχικοποίηση και τη βασική εκπαίδευση του συνελικτικού δικτύου, ενώ οι συνθετικές, μαζί με τις αντίστοιχες 3Δ γεωμετρίες των προσώπων, οι οποίες παράγονται με την ίδια διαδικασία όπως αυτή των Richardson *et al.*, αξιοποιούνται για τη ρύθμισή του (fine tuning).

Οι Tran *et al.*[17] υπολογίζουν αρχικά τις παραμέτρους του 3DMM για κάθε μία από τις περίπου $500k$ εικόνες του συνόλου δεδομένων *CASIA*. Στη συνέχεια τα παραμετρικά μοντέλα που αφορούν στο ίδιο άτομο συνενώνονται και έτσι σχηματίζεται μία μοναδική αναπαράσταση μέσω του 3DMM για κάθε άτομο. Οι μοναδικές πλέον αυτές 3Δ τοπολογίες μαζί με τις αντίστοιχες 2Δ εικόνες σχηματίζουν το σύνολο των δεδομένων εκπαίδευσης του δικτύου ανακατασκευής.

Οι Kim *et al.*[18] προσαρμόζουν το εκπαιδευμένο με συνθετικά δεδομένα δίκτυο ανακατασκευής τους σε πραγματικά δεδομένα, χρησιμοποιώντας έναν αλγόριθμο εκκίνησης (bootstrapping algorithm), διαδικασία ωστόσο η οποία συντελείται χωρίς επίβλεψη.

Όλες οι προαναφερθείσες μέθοδοι ανακατασκευής χρησιμοποιούν είτε μεικτά είτε αμιγώς συνθετικά σύνολα δεδομένων για την εκπαίδευση των δικτύων τους. Τα σύνολα δεδομένων αυτά ωστόσο περιέχουν αρκετά πρόσωπα τα οποία δεν είναι ρεαλιστικά, κάτι το οποίο οφείλεται στον τεχνητό τρόπο με τον οποίο και δημιουργήθηκαν. Το γεγονός αυτό, σε συνδυασμό με τις περιορισμένες πιθανόν εκφραστικές διακυμάνσεις τις οποίες μπορεί να αποδώσει το εκάστοτε μοντέλο αναπαράστασης του προσώπου, καθιστά τις μεθόδους ανακατασκευής με επιβλεπόμενη μάθηση ιδιαιτέρως ευαίσθητες στα σύνολα δεδομένων εκπαίδευσης που χρησιμοποιούνται και κατ΄ επέκταση λιγότερο ικανές να γενικεύουν σε πραγματικά δεδομένα.

2.2.3 Μέθοδοι Μη Επιβλεπόμενης Μάθησης

Προκειμένου να αντιμετωπιστεί το πρόβλημα της έλλειψης επαρκών διαθέσιμων συνόλων δεδομένων 2Δ εικόνων - 3Δ γεωμετριών προσώπου, οι πιο πρόσφατες έρευνες σχετικά με την επίλυση του προβλήματος της 3Δ ανακατασκευής προσώπων από εικόνες έχουν στραφεί στον τομέα της μη επιβλεπόμενης μάθησης, αποσκοπώντας στην εκπαίδευση των δικτύων ανακατασκευής με εισόδους απλές 2Δ εικόνες χωρίς ετικέτες.

Κατά τη μη επιβλεπόμενη μάθηση δεν παρέχεται στο δίκτυο κάποια πληροφορία σχετικά με το είδος των δεδομένων εισόδου, με αποτέλεσμα το ίδιο να καλείται να αυτο-οργανωθεί, έτσι ώστε να προσδιορίσει σχέσεις και μοτίβα, σύμφωνα με τα οποία θα μπορέσει να αξιολογήσει την ακρίβεια και την αξιοπιστία των εξόδων του. Οι σχέσεις αυτές, καλούνται συναρτήσεις απώλειας (loss functions) και τροποποιούνται κάθε φορά ανάλογα με τη μορφή και τη φύση του συνόλου των δεδομένων εκπαίδευσης. Σκοπός της εκπαίδευσης είναι η ελαχιστοποίηση των συναρτήσεων απώλειας, η οποία εξασφαλίζει την κατά το δυνατόν ομοιότητα των εξόδων του δικτύου με τις αντίστοιχες εισόδους.

Στην περίπτωση του προβλήματος της 3Δ ανακατασκευής προσώπων, ο καθορισμός των συναρτήσεων απώλειας για την εκπαίδευση των δικτύων ανακατασκευής, αποτελεί μια δύσκολη διαδικασία, καθώς οι εικόνες εισόδου αναπαριστούν ανθρώπινα πρόσωπα, τα οποία χαρακτηρίζονται από πληθώρα και υψηλό επίπεδο γεωμετρικών, εκφραστικών και χρωματικών λεπτομερειών. Απαιτείται επομένως ο προσδιορισμός μιας υβριδικής συνάρτησης απώλειας, οι συνιστώσες της οποίας καλύπτουν το μεγαλύτερο εύρος των χαρακτηριστικών αυτών, έτσι ώστε οι 3Δ τοπολογίες προσώπων τις οποίες παράγουν τα δίκτυα να αποτυπώνουν τα ιδιαίτερα χαρακτηριστικά της προσωπικότητας των ατόμων, τα οποία απεικονίζονται στις 2Δ εικόνες εισόδου.

Στο πλαίσιο αυτό, οι Tewari *et al.*[19] αναπτύσσουν το μοντέλο MoFA, το οποίο χρησιμοποιεί ένα δίκτυο Αυτο-κωδικοποιητή (Autoencoder, AE) για τη διαδικασία της ανακατασκευής. Η μοντελοποίηση του ανθρώπινου προσώπου γίνεται με χρήση ενός γενικευμένου 3DMM, συνολικά 257 συντελεστών, το οποίο αποτυπώνει τη γεωμετρία, τις εκφράσεις και την υφή του προσώπου, συνυπολογίζοντας συγχρόνως τις συνθήκες φωτισμού του περιβάλλοντος και την τοποθέτηση της κάμερας (μετατόπιση, περιστροφή). Η εκπαίδευση του δικτύου γίνεται χωρίς επίβλεψη, χρησιμοποιώντας μία τοπική φωτομετρική συνάρτηση απώλειας (pixel-wise photometric loss function), έτσι ώστε να εξασφαλιστεί ότι η παραγόμενη συνθετική εικόνα, η οποία προκύπτει έπειτα από rendering του 3Δ μοντέλου του προσώπου, μοιάζει με την εικόνα εισόδου. Η προσέγγιση αυτή, παρ΄ όλη την αποτελεσματικότητά της, προσδίδει μια τοπικότητα στο δίκτυο ανακατασκευής, καθώς λαμβάνονται υπόψιν μόνο άμεσες pixel προς pixel σχέσεις μεταξύ της παραγόμενης και της αρχικής εικόνας εισόδου, με αποτέλεσμα να παραλείπονται βασικά στοιχεία της ταυτότητας του ατόμου και να συγχέονται χρωματικές αποχρώσεις με συνθήκες φωτισμού του περιβάλλοντος (π.χ. σκούρες αποχρώσεις ερμηνεύονται ως συνθήκες αμυδρού φωτισμού και το αντίστροφο).

Για τη διατήρηση των χαρακτηριστικών της ταυτότητας του ατόμου, οι Genova *et al.*[20] προτείνουν την επέκταση του βασικού δικτύου του αυτοκωδικοποιητή, συμπεριλαμβάνοντας στη συνάρτηση απώλειας και έναν όρο για τα ιδιαίτερα χαρακτηριστικά του προσώπου (identity loss). Τα χαρακτηριστικά προσδιορίζονται τόσο για το πρόσωπο της εικόνας εισόδου όσο και για το πρόσωπο της αντίστοιχης παραγόμενης συνθετικής εικόνας, μέσω ενός προεκπαιδευμένου δικτύου αναγνώρισης προσώπων (FaceNet) και η διαφορά τους χρησιμοποιείται με κατάλληλο τρόπο στη συνάρτηση απώλειας κατά την εκπαίδευση. Το εκτεταμένο αυτό μοντέλο επιτυγχάνει τη διατήρηση της ταυτότητας του προσώπου, αδυνατεί ωστόσο να αποδώσει τα χαμηλού-επιπέδου (low-level) χαρακτηριστικά του, καθώς η λειτουργία του βασίζεται στη σύγκριση χαρακτηριστικών, τα οποία υπολογίζονται με τρόπο έτσι ώστε να είναι ανεξάρτητα συνθηκών φωτισμού και εκφράσεων, κάτι το οποίο δε συμβαίνει στην περίπτωση των πραγματικών εικόνων.

Τέλος, οι Wu *et al.*[21] στο μοντέλο τους MVF-Net, εκπαιδεύουν ένα βαθύ συνελικτικό δίκτυο για τον άμεσο προσδιορισμό των παραμέτρων του 3DMM, χρησιμοποιώντας πολλαπλές εικόνες για το ίδιο πρόσωπο, υπό διαφορετικές οπτικές γωνίες, έτσι ώστε να αποτυπωθούν καλύτερα η γεωμετρία και τα χαρακτηριστικά του. Η υβριδική συνάρτηση απώλειας που χρησιμοποιείται για την εκπαίδευση του δικτύου αποτελείται από μία τροποποιημένη συνιστώσα φωτομετρικής απώλειας και από μία συνιστώσα απώλειας σημείων ενδιαφέροντος του προσώπου (landmark loss) και ευθυγράμμισης (alignment loss). Κάθε συνάρτηση απώλειας υπολογίζεται για όλες τις εικόνες που αντιστοιχούν στο ίδιο άτομο, ενώ για τον εντοπισμό των σημείων ενδιαφέροντος του προσώπου χρησιμοποιείται ένας state-of-the-art ανιχνευτής.

Στο σημείο αυτό αξίζει να σημειωθεί πως πέραν των μεθόδων που αναφέρθηκαν έως τώρα, υπάρχουν και μέθοδοι οι οποίες συνδυάζουν την επιβλεπόμενη και την μη επιβλεπόμενη μάθηση και οι οποίες για την εκπαίδευση των μοντέλων ανακατασκευής, χρησιμοποιούν μεικτά σύνολα δεδομένων, τα οποία αποτελούνται τόσο από απλές 2Δ εικόνες χωρίς ετικέτες, όσο και από συνδυασμούς 2Δ εικόνων - 3Δ τοπολογιών προσώπων. Με τον τρόπο αυτό επιδιώκεται η διεύρυνση της διαδικασίας εκπαίδευσης και γίνεται μια προσπάθεια εξισορρόπησης μεταξύ του μη επιβλεπόμενου και του επιβλεπόμενου μέρους αυτής, έτσι ώστε το δίκτυο να προσαρμόζεται καλύτερα σε πραγματικές συνθήκες.

Για παράδειγμα, οι Zeng *et al.*[22] προτείνουν το μοντέλο DF^2-Net, το οποίο αποτελείται από τρία επιμέρους συνελικτικά δίκτυα, ένα για την ανάκτηση του βασικού μοντέλου του προσώπου (Dense Network) και δύο για την κλιμακωτή αποτύπωση των λεπτομερειών υψηλών συχνοτήτων (Fine Network, Finer

Network). Η εκπαίδευση των τριών δικτύων πραγματοποιείται με διαφορετικό τρόπο και με χρήση διαφορετικών συνόλων δεδομένων. Συνολικά κατασκευάζονται 3 σύνολα δεδομένων εκπαίδευσης: 1) ένα συνθετικό σύνολο 3Δ προσώπων μέσω ενός 3DMM, τα οποία στη συνέχεια γίνονται rendered και προβάλλονται πάνω σε 2Δ εικόνες, 2) ένα συνθετικό σύνολο 2Δ εικόνων, των οποίων υπολογίζονται τα αντίστοιχα 3Δ μοντέλα μέσω ενός υπάρχοντος, προεκπαιδευμένου δικτύου ανακατασκευής και 3) ένα σύνολο 2Δ εικόνων λεπτομερειών υψηλού επιπέδου. Έτσι, το πρώτο δικτύο (Dense Network) εκπαιδεύεται με επίβλεψη και χρήση του συνόλου δεδομένων 1), το δεύτερο (Fine Network) εκπαιδεύεται επίσης με επίβλεψη και χρήση του συνόλου δεδομένων 2) και το τρίτο (Finer Network) εκπαιδεύεται χωρίς επίβλεψη και χρήση του συνόλου δεδομένων 3). Η μεικτή αυτή εκπαίδευση οδηγεί σε αρκετά ικανοποιητικά αποτελέσματα, χαρακτηρίζεται ωστόσο από υψηλά ποσοστά συνθετικών δεδομένων, με αποτέλεσμα το δίκτυο να μη γενικεύει ιδιαιτέρως καλά σε εικόνες μη ελεγχόμενου περιβάλλοντος.

Από όσα αναφέρθηκαν έως τώρα, εύκολα μπορεί κανείς να συμπεράνει ότι οι μέθοδοι μη επιβλεπόμενης μάθησης παρουσιάζουν ξεκάθαρη υπεροχή έναντι των μεθόδων βελτιστοποίησης και επιβλεπόμενης μάθησης και αποτελούν αναμφισβήτητα το μέλλον της έρευνας για το πρόβλημα της 3Δ ανακατασκευής προσώπων από εικόνες, ειδικά στην περίπτωση όπου αυτές προέρχονται από μη ελεγχόμενο περιβάλλον. Καθεμία από τις υπάρχουσες μεθόδους της κατηγορίας αυτής προσπαθεί να εστιάσει σε κάποιο ιδιαίτερο στοιχείο της τοπολογίας του ανθρώπινου προσώπου, είτε γεωμετρικό-εκφραστικό (π.χ. ρυτίδες και ανωμαλείες της επιφάνειας του προσώπου), είτε χρωματικό (π.χ. χρωματικές διακυμάνσεις και ιδιαίτερες αποχρώσεις δέρματος), χωρίς ωστόσο να έχει προταθεί έως τώρα κάποιο μοντέλο ανακατασκευής το οποίο να αποδίδει όλες τις λεπτομερειες αυτές σε πολύ υψηλό επίπεδο, γεγονός αναμενόμενο αν αναλογιστεί κανείς την πολυπλοκότητα και την ιδιαιτερότητα του ανθρώπινου προσώπου.

Κεφάλαιο 3

Προτεινόμενη Μέθοδος

Στο κεφάλαιο αυτό θα παρουσιαστεί και θα αναλυθεί η προτεινόμενη μέθοδος για την αντιμετώπιση του προβλήματος της 3Δ ανακατασκευής προσώπων από εικόνες. Για το σχεδιασμό ενός ολοκληρωμένου και αποτελεσματικού συστήματος ανακατασκευής, πρέπει να ληφθεί υπόψιν ένα μεγάλο εύρος συνιστωσών, οι οποίες εντάσσονται στο επιστημονικό πεδίο των γραφικών και της υπολογιστικής όρασης και οι οποίες αφορούν τόσο στην προεπεξεργασία των δεδομένων εισόδου, όσο και στον τρόπο με τον οποίο μοντελοποιείται ο περιβάλλων χώρος και το σκηνικό.

Προς την κατεύθυνση αυτή, το παρόν κεφάλαιο είναι χωρισμένο σε θεματικές ενότητες, καθεμία εκ των οποίων πραγματεύεται κάποια από τις επιμέρους συνιστώσες οι οποίες είναι απαραίτητες για την υλοποίηση και τη λειτουργία του συστήματος ανακατασκευής. Έτσι, γίνεται αναφορά στον τρόπο δημιουργίας του συνόλου δεδομένων εκπαίδευσης, στη διαδικασία παραμετροποίησης του περιβάλλοντος και τέλος παρουσιάζεται ο αλγόριθμος μάθησης που χρησιμοποιείται για την εκπαίδευση του δικτύου ανακατασκευής.

Η σωστή θεωρητική και μαθηματική θεμελίωση των ανωτέρω συνιστωσών και η μετέπειτα κατάλληλη αξιοποίησή τους στη διαδικασία της εκπαίδευσης, οδηγεί σε αρκετά ικανοποιητικά αποτελέσματα, με τη μορφή των 3Δ παραγόμενων προσώπων να προσεγγίζει σε υψηλό βαθμό τα πρόσωπα των αντίστοιχων 2Δ εικόνων.

3.1 Προεπεξεργασία Εικόνων - Δημιουργία Dataset

Απαραίτητη προϋπόθεση για τη σωστή εκπαίδευση του δικτύου ανακατασκευής αποτελεί η ύπαρξη ενός πλήρους και καλά επεξεργασμένου συνόλου δεδομένων, το οποίο θα πρέπει να περιλαμβάνει εικόνες προσώπων που αντιστοιχούν σε άτομα όλων των ηλικιών, φύλων και δερματικών αποχρώσεων. Επιπλέον, τα πρόσωπα των εικόνων θα πρέπει να παρουσιάζονται σε διαφορετικές θέσεις (περιστροφές και μετατοπίσεις), ενώ θα πρέπει να καλύπτουν όσο το δυνατόν μεγαλύτερο μέρος του φάσματος των εκφραστικών διακυμάνσεων.

Για τη δημιουργία του dataset το οποίο θα χρησιμοποιηθεί για την εκπαίδευση του προτεινόμενου δικτύου ανακατασκευής, επιλέγονται περίπου $100k$ εικόνες από το *CelebA* dataset [23], το οποίο περιλαμβάνει περισσότερες από $200k$ εικόνες προσώπων διάσημων ατόμων, σε διαφορετικές πόζες και εκφράσεις, σε θορυβώδη και μη παρασκήνια (Σχ.3.1).

Σχήμα 3.1: Δείγματα εικόνων από το CelebA dataset

Προκειμένου οι εικόνες αυτές να μπορέσουν να χρησιμοποιηθούν αποτελεσματικά για την εκπαίδευση του δικτύου, θα πρέπει να υποβληθούν σε κατάλληλη προεπεξεργασία, έτσι ώστε να καθιερωθεί μια πρώιμη στοιχειώδης συσχέτιση μεταξύ αυτών και των αντίστοιχων 3Δ μοντέλων.

Στα πλαίσια της παρούσας εργασίας, η διαδικασία προεπεξεργασίας των εικόνων του dataset περιλαμβάνει τα εξής στάδια:

1. *Ευθυγράμμιση και περικοπή προσώπων*,
2. *Εντοπισμός* 68 *σημείων ενδιαφέροντος των προσώπων*,
3. *Κατασκευή μάσκας δέρματος.*

3.1.1 Ευθυγράμμιση και Περικοπή Προσώπων

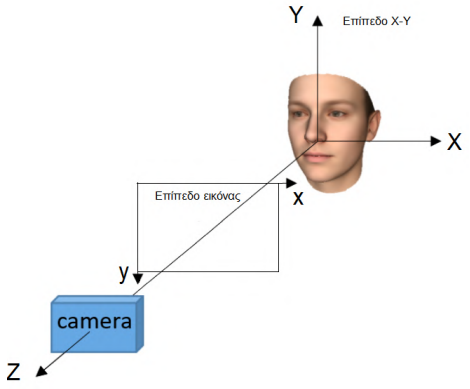
Το πρώτο στάδιο της προεπεξεργασίας των εικόνων του dataset συνίσταται στον εντοπισμό των προσώπων που αυτές απεικονίζουν και στην ευθυγράμμιση αυτών με το 3Δ μοντέλο, το οποίο παράγεται μέσω του 3DMM και συγκεκριμένα του Basel Face Model.

Για τον εντοπισμό των προσώπων χρησιμοποιείται ο *MTCNN* (Multi-task Cascaded Convolutional Networks) [24, 25], ένας off-the-shelf face detector, ο οποίος βασίζεται στη χρήση συνελικτικών νευρωνικών δικτύων για τον εντοπισμό και την εξαγωγή βασικών χαρακτηριστικών προσώπων από εικόνες. Πιο συγκεκριμένα, δοθείσης μιας εικόνας, ο MTCNN εντοπίζει τα πρόσωπα τα οποία παρουσιάζονται σε αυτή και εξάγει για κάθε ένα από αυτά, ένα σύνολο 5 σημείων ενδιαφέροντος (landmarks) και ένα πλαίσιο οριοθέτησης (bounding box). Τα σημεία ενδιαφέροντος αντιστοιχούν στις περιοχές των ματιών, της μύτης και του στόματος και βάσει αυτών υπολογίζεται το πλαίσιο οριοθέτησης, έτσι ώστε να περικλείει ολόκληρο το πρόσωπο, όπως φαίνεται στο Σχ.3.2. Σε περίπτωση όπου το πρόσωπο κάποιας εικόνας δεν αναγνωριστεί επιτυχώς, η εικόνα αυτή παραλείπεται και αφαιρείται από το dataset, έτσι ώστε να διευκολυνθεί η μετέπειτα εκπαίδευση του δικτύου.

Χρησιμοποιώντας λοιπόν τα 5 σημεία ενδιαφέροντος τα οποία εξάγονται από τον MTCNN, η διαδικασία της ευθυγράμμισης του 2Δ με το 3Δ πρόσωπο, ανάγεται στο πρόβλημα της ευθυγράμμισης των αντίστοιχων 2Δ και 3Δ σημείων ενδιαφέροντος, εκ των οποίων τα τελευταία παρέχονται από το Basel Face Model. Σκοπός είναι ο προσδιορισμός ενός μετασχηματισμού ομοιότητας (similarity transformation), έτσι ώστε τα 2Δ και τα 3Δ σημεία ενδιαφέροντος να ευθυγραμμιστούν όσο γίνεται πιο αποτελεσματικά.

Σχήμα 3.2: Εξαγωγή σημείων ενδιαφέροντος (5 σημαντικότερων) και πλαισίων οριοθέτησης προσώπων, με χρήση του MTCNN face detector.

Η διαδικασία αυτή ωστόσο εμπλέκει δεδομένα διαφορετικών διαστάσεων και ως εκ τούτου θα πρέπει παρουσιαστεί αναλυτικά ο τρόπος με τον οποίο υλοποιείται. Στα πλαίσια της παρούσας εργασίας, υιοθετούνται τα ορθοκανονικά συστήματα συντεταγμένων του 3Δ κόσμου (3D world space coordinate system) και των 2Δ εικόνων (2D image coordinate system), τα οποία παρουσιάζονται στο Σχ.3.3. Σύμφωνα με τη σύμβαση αυτή, το 3Δ μοντέλο του προσώπου βρίσκεται στην αρχή (σημείο $(0, 0, 0)$) του ορθοκανονικού συστήματος X, Y, Z κοιτώντας προς τα θετικά του άξονα Z, ενώ η εικόνα τοποθετείται έτσι ώστε το αριστερό επάνω σημείο της να βρίσκεται στο κέντρο (σημείο $(0, 0)$) του συστήματος x, y. Για την απόδοση του σκηνικού χρησιμοποιείται μια κάμερα μικρής οπής (pinhole camera), για την οποία θα γίνει ξεχωριστή αναφορά σε επόμενη ενότητα.



Σχήμα 3.3: Ορθοκανονικά συστήματα συντεταγμένων 3Δ κόσμου και 2Δ επιπέδου της εικόνας.

Έχοντας τώρα ορίσει τα συστήματα συντεταγμένων του 3Δ κόσμου και του 2Δ επιπέδου της εικόνας, η διαδικασία της ευθυγράμμισης του προσώπου μιας εικόνας, βασίζεται στην αντίστοιχη διαδικασία των [29, 30] και έγκειται στον υπολογισμό ενός παράγοντα κλιμάκωσης s και ενός παράγοντα μετατόπισης $\boldsymbol{t} = [t_x, t_y]^T$, έτσι ώστε να ισχύει με όσο το δυνατόν μεγαλύτερη ακρίβεια η σχέση,

$$s \cdot \begin{bmatrix} X_i \\ Y_i \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad \text{με} \quad 1 \leq i \leq 5, \tag{3.1}$$

όπου x_i, y_i οι συντεταγμένες του i-οστού σημείου ενδιαφέροντος του προσώπου της εικόνας, ως προς το 2Δ ορθοκανονικό σύστημα συντεταγμένων αυτής και X_i, Y_i οι συντεταγμένες του αντίστοιχου i-οστού σημείου ενδιαφέροντος του 3Δ προσώπου ως προς το 3Δ σύστημα συντεταγμένων του κόσμου. Στην παραπάνω σχέση, έχει παραλειφθεί η τρίτη διάσταση (άξονας Z) των 3Δ σημείων ενδιαφέροντος, έτσι ώστε να καταστεί δυνατή η εύρεση μιας αντιστοίχισης μεταξύ 2Δ και 3Δ σημείων. Καθώς η λύση της Εξ.3.1 δεν είναι τετριμμένη, χρησιμοποιείται η μέθοδος των Ελαχίστων Τετραγώνων για την εύρεση των παραμέτρων s και $\boldsymbol{t}$, οι οποίες την ικανοποιούν συγχρόνως και με τον καλύτερο τρόπο για κάθε ένα εκ των 5 σημείων ενδιαφέροντος.

Αφού λοιπόν υπολογιστούν οι παράγοντες s και $\boldsymbol{t}$, η αρχική εικόνα, διαστάσεων $w_0 \times h_0$, κλιμακώνεται με τον παράγοντα s και προκύπτει μια εικόνα διαστάσεων $w_s \times h_s = \dfrac{w_0}{s} \times \dfrac{h_0}{s}$. Στη συνέχεια το κέντρο της εικόνας αυτής μετατοπίζεται κατά $\left(t_x - \dfrac{w_0}{2}\right)/s$ και $\left(t_y - \dfrac{h_0}{2}\right)/s$ στις διευθύνσεις των αξόνων x και y αντίστοιχα και τέλος, το εικονιζόμενο πρόσωπο περικόπτεται βάσει του μετατοπισμένου αυτού κέντρου, έτσι ώστε η τελική επεξεργασμένη εικόνα με το περικομένο πρόσωπο να έχει διαστάσεις $w \times h = 224 \times 224$.

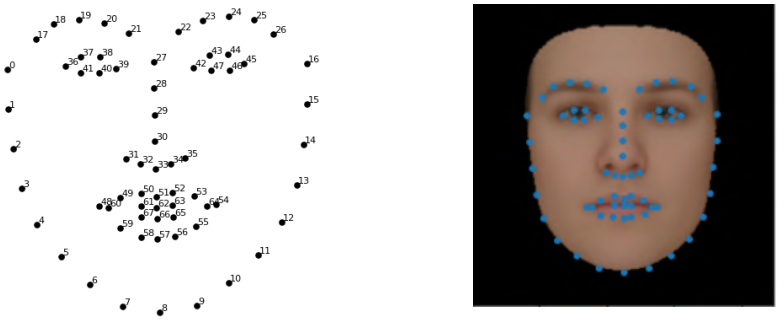
Σχήμα 3.4: Αρχική εικόνα (αριστερά) και επεξεργασμένη εικόνα έπειτα από ευθυγράμμιση και περικοπή (δεξιά).

Στο Σχ.3.4 παρουσιάζονται τα αποτελέσματα της ευθυγράμμισης και περικοπής των εικόνων, εκ των οποίων εύκολα αντιλαμβάνεται κανείς την αναγκαιότητα της διαδικασίας αυτής και τη συμβολή της στη βελτίωση της απόδοσης του δικτύου ανακατασκευής. Η ευθυγράμμιση αυτή εφαρμόζεται σε κάθε μία εκ των επιλεγμένων εικόνων του CelebA dataset και έτσι προκύπτουν ισάριθμες νέες ευθυγραμμισμένες εικόνες προσώπων, οι οποίες θα αποτελέσουν το dataset που θα χρησιμοποιηθεί για την εκπαίδευση του προτεινόμενου δικτύου ανακατασκευής.

3.1.2 Εντοπισμός 68 σημείων ενδιαφέροντος

Το επόμενο στάδιο της προεπεξεργασίας των εικόνων περιλαμβάνει τον εντοπισμό 68 σημείων ενδιαφέροντος σε κάθε ένα από τα πρόσωπα τα οποία απεικονίζουν, έτσι ώστε αυτά να χρησιμοποιηθούν κατά τη διάρκεια της εκπαίδευσης για τον υπολογισμό κατάλληλων συναρτήσεων απώλειας.

Για τον εντοπισμό των σημείων ενδιαφέροντος χρησιμοποιείται ο προεκπαιδευμένος ανιχνευτής της βιβλιοθήκης *dlib* [26, 27, 28], ο οποίος δοθείσης μιας εικόνας στην είσοδο, εντοπίζει το εικονιζόμενο πρόσωπο και επιστρέφει ένα μητρώο διαστάσεων 68×2, το οποίο περιλαμβάνει τις συντεταγμένες (x, y) στο επίπεδο της εικόνας, για κάθε ένα εκ των 68 σημείων ενδιαφέροντος. Το αποτέλεσμα της ανίχνευσης παρουσιάζεται στο Σχ.3.5, μαζί με την αντίστοιχη αρίθμηση η οποία ακολουθείται για τα σημεία ενδιαφέροντος.



Σχήμα 3.5: Αρίθμηση 68 σημείων ενδιαφέροντος (αριστερά) και εντοπισμός τους στο μέσο πρόσωπο του Basel Face Model (δεξιά)

Τα 68 αυτά σημεία προσδιορίζουν πλήρως τόσο την εξωτερική (γραμμή κάτω γνάθου) όσο και την εσωτερική (περιοχές ματιών, μύτης, στόματος και βλεφαρίδων) γεωμετρία του προσώπου και είναι απαραίτητα για τη διαδικασία της ανακατασκευής. Συνολικά, στο στάδιο αυτό, για κάθε μία από τις ευθυγραμμισμένες εικόνες του dataset, οι οποίες προέκυψαν μέσω της διαδικασίας της ενότητας 3.1.1, υπολογίζονται τα 68 σημεία ενδιαφέροντος των εικονιζόμενων προσώπων και αποθηκεύονται σε ένα αρχείο *.txt*, με αποτέλεσμα να προκύπτει ένα ισάριθμο σύνολο αρχείων σημείων ενδιαφέροντος, τα οποία θα ανακτηθούν μετέπειτα από το δίκτυο ανακατασκευής για τον προσδιορισμό της γεωμετρίας των προσώπων.

3.1.3 Κατασκευή Μάσκας Δέρματος

Η απόδοση του προτεινόμενου δικτύου ανακατασκευής επηρεάζεται σε μεγάλο βαθμό από τυχόν αποκρύψεις (occlusions) περιοχών του προσώπου (π.χ λόγω παρουσίας γυαλιών), καθώς και από διάφορα αισθητικά στοιχεία αυτού, όπως το μούσι και το ιδιαίτερα έντονο make-up, τα οποία καλύπτουν μεγάλες περιοχές δέρματος. Αυτό έχει ως αποτέλεσμα να ελαττώνεται σημαντικά η ωφέλιμη πληροφορία σχετικά με την υφή του προσώπου και να συγχέονται περιοχές του παρασκηνίου της εικόνας με χαρακτηριστικά του εικονιζόμενου ατόμου.

Για να αντιμετωπιστεί το πρόβλημα αυτό, θα πρέπει να σχεδιαστεί ένας ταξινομητής δέρματος (skin classifier) [31, 32], ο οποίος θα διακρίνει τα pixels της εικόνας που αντιστοιχούν σε περιοχές δέρματος, σύμφωνα με κάποιο κατάλληλο κανόνα απόφασης (decision rule). Κεντρική ιδέα της διαδικασίας της ταξινόμησης είναι ο υπολογισμός της δεσμευμένης πιθανότητας $P(skin/\boldsymbol{c})$, η οποία σύμφωνα με τον κανόνα του Bayes ισούται με:

$$P(skin/\boldsymbol{c}) = \frac{P(\boldsymbol{c}/skin)P(skin)}{P(\boldsymbol{c}/skin)P(skin) + P(\boldsymbol{c}/\neg skin)P(\neg skin)}, \tag{3.2}$$

όπου $\boldsymbol{c}$ είναι το διάνυσμα χρώματος ενός δεδομένου pixel (εκφρασμένο σε κατάλληλο χρωματικό χώρο), $P(skin)$, $P(\neg skin)$ οι εκ των προτέρων πιθανότητες (prior propabilities) οποιουδήποτε χρώματος να αποτελεί ή να μην αποτελεί αντίστοιχα απόχρωση δέρματος και $P(\boldsymbol{c}/skin)$, $P(\boldsymbol{c}/\neg skin)$ οι δεσμευμένες πιθανότητες επιλογής του συγκεκριμένου διανύσματος χρώματος $\boldsymbol{c}$, δεδομένου ότι αυτό αντιστοιχεί ή δεν αντιστοιχεί σε pixel δέρματος.

Υπολογίζοντας τώρα την πιθανότητα της Εξ.3.2, ένα pixel κατηγοριοποιείται ως pixel δέρματος εάν

$$P(skin/\boldsymbol{c}) \geq \Theta, \tag{3.3}$$

όπου $0 \leq \Theta \leq 1$ κατάλληλα επιλεγμένο κατώφλι απόφασης (decision threshold).

Καθώς η εν λόγω διαδικασία ταξινόμησης αποτελείται από δύο κλάσεις, οι εκ των προτέρων πιθανότητες της Εξ.3.2 ικανοποιούν τη σχέση

$$P(skin) + P(\neg skin) = 1, \tag{3.4}$$

με αποτέλεσμα να απαιτείται ο προσδιορισμός μόνο μίας εκ των δύο.

Δεδομένου ότι το σύνολο των δεδομένων εκπαίδευσης του ταξινομητή αποτελείται από συλλογή pixels, εκ των οποίων κάποια ανήκουν σε περιοχές δέρματος και κάποια όχι, μια λογική επιλογή για την εκ των προτέρων πιθανότητα $P(skin)$ είναι αυτή να ισούται με το λόγο του συνολικού αριθμού των pixels δέρματος προς το συνολικό αριθμό όλων των διαθέσιμων pixels, δηλαδή,

$$P(skin) = \frac{T_s}{T_s + T_n}, \tag{3.5}$$

όπου T_s και T_n ο συνολικός αριθμός των pixels, τα οποία ανήκουν ή δεν ανήκουν αντίστοιχα σε περιοχές δέρματος.

Για τον προσδιορισμό της δεσμευμένης πιθανότητας $P(\boldsymbol{c}/skin)$ και κατ΄ επέκταση της πιθανότητας $P(\boldsymbol{c}/\neg skin)$, θα πρέπει να μοντελοποιηθεί και να προσεγγιστεί η κατανομή των pixels του dataset μέσω κάποιου παραμετρικού στατιστικού μοντέλου.

Η μοντελοποίηση αυτή μπορεί να γίνει με μεγάλη ακρίβεια, χρησιμοποιώντας το *Μοντέλο Μείξης Γκαουσιανών Κατανομών* (Gaussian Mixture Models), ή εν συντομία GMM, το οποίο πρόκειται ουσιαστικά για ένα πιθανοτικό μοντέλο που χρησιμοποιείται για την αναπαράσταση ενός κανονικά κατανεμημένου υποπληθυσμού ενός συνολικού πληθυσμού. Για την αναπαράσταση των υποπλυθησμών, το μοντέλο χρησιμοποιεί ένα πλήθος επιμέρους συνιστωσών, όπου κάθε συνιστώσα ακολουθεί μια κανονική κατανομή. Κατά αυτό τον τρόπο δημιουργούνται διαφορετικές συστάδες (clusters) δεδομένων, με τα στοιχεία κάθε συστάδας να παρουσιάζουν έντονη συσχέτιση μεταξύ τους. Σε αντίθεση με την περίπτωση των απλών Γκαουσιανών κατανομών, το Μοντέλο Μείξης Γκαουσιανών είναι ικανό να αντιμετωπίσει πολυτροπικά (multimodal) δεδομένα, δηλαδή δεδομένα στα οποία υπάρχει έντονη επικάλυψη, καθώς η ομαδοποίησή τους

δεν είναι μονοσήμαντη και προκύπτει από μείξη διαφορετικών συνιστωσών.

Χρησιμοποιώντας λοιπόν το εν λόγω μοντέλο [33], κάθε pixel δέρματος μπορεί να θεωρηθεί ότι προέρχεται από έναν υπερπληθυσμό G, ο οποίος αποτελείται από μία μείξη πεπερασμένου αριθμού g Γκαουσιανών κατανομών $G_1, \ldots, G_g$ με συντελεστές βαρύτητας $\pi_1, \ldots, \pi_g$ αντίστοιχα, όπου

$$\sum_{i=1}^{g} \pi_i = 1 \text{ και } \pi_i \geq 0. \tag{3.6}$$

Στην περίπτωση αυτή, η συνάρτηση πυκνότητας πυθανότητας (p.d.f) ενός τυχαίου pixel χρώματος $\boldsymbol{c}$ (διαστάσεων d) δίνεται από τη σχέση,

$$P(\boldsymbol{c}/skin) = \sum_{i=1}^{g} \pi_i \cdot \frac{1}{(2\pi)^{d/2} \left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\boldsymbol{c}-\boldsymbol{\mu}_i)^T (\boldsymbol{\Sigma}_i)^{-1} (\boldsymbol{c}-\boldsymbol{\mu}_i)}, \tag{3.7}$$

όπου π_i οι συντελεστές μείξης, $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{id}]^T$ το μέσο διάνυσμα και $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ το μητρώο συνδιακύμανσης (covariance matrix) της $i-$οστής Γκαουσιανής κατανομής.

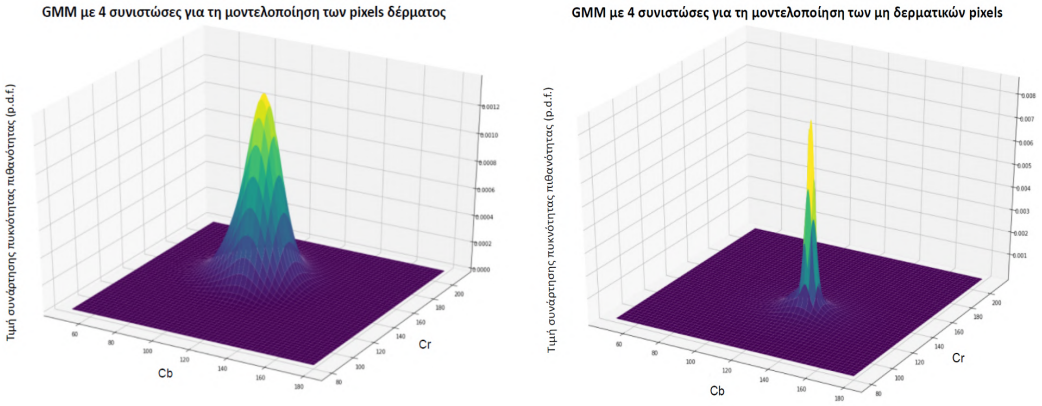
Έπειτα τώρα από τον υπολογισμό της πιθανότητας της Εξ.3.7, η πιθανότητα $P(\boldsymbol{c}/\neg skin)$ υπολογίζεται με τελείως αντίστοιχο τρόπο. Έτσι, καθορίζονται πλήρως όλες οι απαραίτητες ποσότητες της Εξ.3.2 και προσδιορίζεται η ζητούμενη πιθανότητα $P(skin/\boldsymbol{c})$ που απαιτείται για την ταξινόμηση των pixels της εικόνας.

Στα πλαίσια της εργασίας, για την υλοποίηση της παραπάνω διαδικασίας ταξινόμησης, χρησιμοποιείται ένας αφελής Μπεϋζιανός ταξινομητής (naive Bayes classifier), ο οποίος εκπαιδεύεται στο skin image dataset των Jones *et al.*[31].

Για την επεξεργασία των χρωματικών συνιστωσών των pixels επιλέγεται ο χώρος χρώματος $YCbCr$ [36], όπου το Y αντιπροσωπεύει τη φωτεινότητα (luma) και τα Cb και Cr είναι η μπλε και η κόκκινη απόχρωση αντίστοιχα. Η επιλογή του χρωματικού αυτού χώρου έναντι του πρωταρχικού χώρου RGB οφείλεται στο ότι οι συνιστώσες χρώματος Cb και Cr των pixels δέρματος παρουσιάζουν μεγαλύτερη ομοιομορφία ως προς το εύρος των τιμών τους, σε σχέση με τις αντίστοιχες R, G, B χρωματικές συνιστώσες και συγκεντρώνονται σε μια σχετικά περιορισμένη περιοχή.

Για την ομαδοποίηση των δερματικών και μη δερματικών pixels εκφρασμένων στο χρωματικό χώρο YCbCr, χρησιμοποιούνται δύο μοντέλα μείξης Γκαουσιανών κατανομών με $g = 4$ Γκαουσιανές συνιστώσες, τα οποία ακολουθούν τη

συνάρτηση πυκνότητας πιθανότητας της Εξ.3.7, με $\boldsymbol{c} = [Y, Cb, Cr]^T$. Οι στατιστικές τιμές (μέσα διανύσματα, μητρώα συνδιακύμανσης) των μοντέλων αυτών λαμβάνονται από τους Deng *et al.*[30], οι οποίοι επεξεργάστηκαν το προαναφερθέν dataset και υπολόγισαν τις τιμές αυτές μέσω κατάλληλου expectation-maximization αλγορίθμου. Στο Σχ.3.6 παρουσιάζονται τα GMMs που προκύπτουν από τις εν λόγω Γκαουσιανές κατανομές.



Σχήμα 3.6: GMMs τεσσάρων συνιστωσών για τη μοντελοποίηση των δερματικών (αριστερά) και μη δερματικών (δεξιά) pixels του sikn image dataset [31].

Αξιοποιώντας τώρα την παραπάνω μοντελοποίηση, υπολογίζεται για κάθε pixel μιας εικόνας η πιθανότητα της Εξ.3.2 και ανάλογα με αυτή γίνεται η ταξινόμησή του ως pixel δέρματος ή όχι. Ως κατώφλι απόφασης για την ταξινόμηση επιλέγεται η τιμή $\Theta = 0.5$, για την οποία αποδίδεται στο pixel η τιμή 1, ενώ ακολουθώντας τη μέθοδο των [30], pixels των οποίων η πιθανότητα $P(skin/\boldsymbol{c})$ έχει τιμή μικρότερη του 0.5 δεν αντιμετωπίζονται αυτομάτως ως μη δερματικά, αλλά τους αποδίδεται η τιμή της αντίστοιχης υπολογιζόμενης πιθανότητας. Αυτό γίνεται, καθώς όπως αναφέρθηκε και νωρίτερα, οι εικόνες προέρχονται ως επί το πλείστον από μη ελεγχόμενο περιβάλλον, με αποτέλεσμα τα χαρακτηριστικά των προσώπων που απεικονίζουν να μην είναι πάντα ευδιάκριτα. Η άμεση απόρριψη των pixels εκείνων των οποίων η πιθανότητα να ανήκουν σε περιοχές δέρματος δεν είναι αρκετά μεγάλη, θα ήταν λανθασμένη και θα οδηγούσε σε σημαντική ελάττωση της διαθέσιμης πληροφορίας.

Για το λόγο αυτό, σε κάθε pixel i μιας εικόνας προσώπου I αποδίδεται η τιμή,

$$A_i = \begin{cases} 1, & \text{εάν } P_i(skin/\boldsymbol{c}_i) > 0.5 \\ P_i(skin/\boldsymbol{c}_i), & \text{σε οποιαδήποτε άλλη περίπτωση,} \end{cases} \tag{3.8}$$

όπου $\boldsymbol{c}_i$, το διάνυσμα χρώματος του i-οστού pixel. Οι τιμές A_i σχηματίζουν μία καινούρια ασπρόμαυρη εικόνα I_{skin}, η οποία αποτελεί ουσιαστικά μια μάσκα δέρματος του εικονιζόμενου προσώπου.

Η προαναφερθείσα διαδικασία εφαρμόζεται σε κάθε επεξεργασμένη εικόνα προσώπου του dataset εκπαίδευσης και κατά αυτό τον τρόπο προκύπτουν ισάριθμες μάσκες δέρματος, οι οποίες θα χρησιμοποιηθούν μετέπειτα από το δίκτυο ανακατασκευής για την εξαγωγή πληροφοριών σχετικά με το δέρμα των εξεταζόμενων προσώπων.

Στο Σχ.3.7 παρουσιάζονται μερικά παραδείγματα εικόνων, μαζί με τις αντίστοιχες μάσκες δέρματος, έτσι ώστε να γίνει αντιληπτή η ικανότητα αυτών να εντοπίζουν και να διαχωρίζουν το δέρμα από άλλα στοιχεία της εικόνας.

Σχήμα 3.7: Εικόνες προσώπων μαζί με τις αντίστοιχες μάσκες δέρματος

Ανακεφαλαιώνοντας, μετά την εφαρμογή των βημάτων της προεπεξεργασίας (ενότητες 3.1.1, 3.1.2 και 3.1.3) σε κάθε μία από τις επιλεγμένες εικόνες του CelebA datset, προκύπτουν τα εξής:

- Ένα σύνολο περίπου $100k$ ευθυγραμμισμένων εικόνων προσώπων, οι οποίες αποτελούν το dataset για την εκπαίδευση του προτεινόμενο δικτύου ανακατασκευής,
- Ένα σύνολο ισάριθμων αρχείων $.txt$ με τα 68 σημεία ενδιαφέροντος των προσώπων των εικόνων αυτών και
- Ένα σύνολο ισάριθμων ασπρόμαυρων εικόνων, οι οποίες αποτελούν τις μάσκες δέρματος των προσώπων.

3.2 Παραμετροποίηση Περιβάλλοντος

Με την ολοκλήρωση της προεπεξεργασίας των εικόνων και τη δημιουργία του επιθυμητού συνόλου δεδομένων εκπαίδευσης, επόμενο βήμα είναι η παραμετροποίηση του περιβάλλοντος ανακατασκευής. Όπως ήδη έχει αναφερθεί και στην ενότητα 2.1.3, για την αναπαράσταση της 3Δ τοπολογίας ενός προσώπου χρησιμοποιείται το Basel Face Model, το οποίο για τον υπολογισμό του σχήματος, της έκφρασης και της υφής του προσώπου απαιτεί τον προσδιορισμό συνολικά 224 συντελεστών ($\boldsymbol{\alpha}, \boldsymbol{\delta}$ και $\boldsymbol{\beta}$ αντίστοιχα). Το μοντέλο αυτό είναι μεν ικανό να αποδώσει αρκετά ικανοποιητικά τη γεωμετρία και την υφή ενός προσώπου, δεν περιέχει όμως καμία πληροφορία σχετικά με τις συνθήκες φωτισμού του περιβάλλοντος χώρου, με αποτέλεσμα να εκλείπουν φαινόμενα όπως οι ανακλάσεις και οι σκιές, τα οποία είναι παρόντα σε κάθε φωτογραφία που προέρχεται από τον πραγματικό κόσμο.

Προκειμένου λοιπόν να προσομοιωθούν οι συνθήκες φωτισμού του πραγματικού κόσμου, θα πρέπει να υιοθετηθεί κατάλληλο μοντέλο φωτισμού, ενώ για την μετέπειτα μετάβαση από τον 3Δ στο 2Δ χώρο της εικόνας, θα πρέπει να επιλεγεί κάποιο μοντέλο κάμερας. Προφανώς, τόσο το μοντέλο φωτισμού όσο και το μοντέλο της κάμερας εισάγουν επιπλέον συντελεστές προς προσδιορισμό, οι οποίοι προστίθενται στους ήδη υπάρχοντες συντελεστές του Basel Face Model. Το πλήθος και το είδος των επιπρόσθετων αυτών συντελεστών, εξαρτάται από τα μοντέλα φωτισμού και κάμερας που επιλέγονται, τα οποία και θα αναλυθούν στις επόμενες ενότητες.

Στο σημείο αυτό σημειώνεται ότι η παραμετροποίηση του περιβάλλοντος ανακατασκευής, δεν αποτελεί μια μονοσήμαντα ορισμένη διαδικασία. Η επιλογή διαφορετικών μοντέλων αναπαράστασης, φωτισμού και κάμερας οδηγεί σε πολύ διαφορετικά αποτελέσματα και επηρεάζει με άμεσο τρόπο την αποτελεσματικότητα του δικτύου ανακατασκευής.

Στα πλαίσια της παρούσας εργασίας, επιλέγονται μοντέλα τα οποία εξασφαλίζουν το ρεαλισμό και την πιστότητα των παραγόμενων προσώπων, χωρίς ωστόσο αυτό να σημαίνει ότι αποτελούν τα βέλτιστα. Σε κάθε περίπτωση, η μοντελοποίηση των συνθηκών του περιβάλλοντος αποτελεί μια ιδιαίτερα δύσκολη διαδικασία, η οποία θα πρέπει να είναι ευπροσάρμοστη στα δεδομένα και τα ζητούμενα του εκάστοτε προβλήματος ανακατασκευής.

3.2.1 Τροποποιημένο 3D Basel Face Model

Έχοντας ολοκληρώσει τόσο τη θεωρητική όσο και την μαθηματική ανάλυση του 3D Morphable Face Model (ενότητα 2.1.3), είναι απαραίτητο να δοθούν ορισμένες πρακτικές διευκρινίσεις, οι οποίες αφορούν στην εφαρμογή του μοντέλου στο σύστημα ανακατασκευής προσώπων το οποίο μελετάται στην παρούσα εργασία. Στο πλαίσιο λοιπόν της εν λόγω διπλωματικής, για τη μοντελοποίηση και την 3Δ αναπαράσταση του ανθρώπινου προσώπου, χρησιμοποιείται το 3D Basel Face Model (2009) [37] με ορισμένες τροποποιήσεις ως προς τα μέρη του προσώπου τα οποία αυτό αναπαριστά.

Πιο συγκεκριμένα, καθώς το ενδιαφέρον μας έγκειται κυρίως στην ανακατασκευή του κεντρικού τμήματος του προσώπου, το αυθεντικό 3D Basel Face Model τροποποιείται, έτσι ώστε να αφαιρεθούν από αυτό οι περιοχές των αυτιών και του λαιμού [30]. Επιπλέον, επιλέγονται οι πρώτες *80 ορθοκανονικές βάσεις* της μεθόδου PCA τόσο για το σχήμα όσο και για την υφή, ενώ για την αποτύπωση των εκφράσεων χρησιμοποιούνται οι *πρώτες 64 ορθοκανονικές βάσεις* έκφρασης από τους Guo *et al.*[38], οι οποίες κατασκευάζονται βάσει του FaceWarehouse dataset [39].

Συνοψίζοντας, το μοντέλο αναπαράστασης του ανθρώπινου προσώπου που χρησιμοποιείται στο προτεινόμενο δίκτυο ανακατασκευής, βασίζεται στο 3D Basel Face Model και περιγράφεται από τις σχέσεις των Εξ.2.5 και 2.7, με $\boldsymbol{d_s} = \boldsymbol{d_t} = \mathbf{80}$, $\boldsymbol{d_e} = \mathbf{64}$, $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{\mathbf{80}}$ και $\boldsymbol{\delta} \in \mathbb{R}^{\mathbf{64}}$. Η 3Δ δομή του προσώπου αναπαρίσταται μέσω ενός τριγωνικού πλέγματος συνολικά $\boldsymbol{n} = \mathbf{35709}$ ***σημείων***, τα οποία σχηματίζουν **70789** ***τρίγωνα***.

Σχήμα 3.8: 3Δ αναπαράσταση ανθρώπινου προσώπου με χρήση του πλήρους (αριστερά) και του τροποποιημένου (δεξιά) 3D Basel Face Model.

3.2.2 Μοντέλο Φωτισμού

Η έως τώρα ανάλυση επικεντρώθηκε κυρίως στον τρόπο με τον οποίο το ανθρώπινο πρόσωπο μοντελοποιείται και παρουσιάζεται στις 3 διαστάσεις, ως προς τη γεωμετρία, τις εκφράσεις και την υφή του, χωρίς ωστόσο να έχει γίνει κάποια αναφορά στον περιβάλλονται χώρο μέσα στον οποίο αυτό τοποθετείται. Καθώς οι 2Δ εικόνες στις οποίες βασίζεται η διαδικασία της ανακατασκευής προέρχονται από μη ελεγχόμενα περιβάλλοντα, στοιχεία όπως η θέση, το είδος της κάμερας και ο φωτισμός συμβάλλουν καθοριστικά στον τρόπο με τον οποίο το πρόσωπο ενός ατόμου αποτυπώνεται σε αυτές. Ιδιαίτερα ο φωτισμός ενός σκηνικού, επηρεάζει σε πολύ μεγάλο βαθμό τον τρόπο με τον οποίο γίνεται αντιληπτή μια επιφάνεια, δεδομένου ότι διάφορα μοτίβα όπως οι σκιές και οι κατοπτρισμοί (specularities) μεταβάλλονται δυναμικά με αυτόν.

Κατ΄ αντιστοιχία με τον πραγματικό κόσμο, οι συνθήκες του σκηνικού στο οποίο τοποθετείται το 3Δ μοντέλο του ανθρώπινου προσώπου, συμβάλλουν σημαντικά στον τρόπο με τον οποίο αυτό παρουσιάζεται και προβάλλεται εν συνεχεία στο επίπεδο της 2Δ εικόνας. Προκειμένου λοιπόν τα παραγόμενα 3Δ πρόσωπα να είναι ρεαλιστικά, θα πρέπει οι συνθήκες του περιβάλλοντος χώρου τους να μοντελοποιηθούν κατάλληλα, έτσι ώστε να προσομοιάζουν κατά το δυνατόν τις αντίστοιχες συνθήκες του πραγματικού κόσμου.

Στα πλαίσια της παρούσας εργασίας, το πρόσωπο αντιμετωπίζεται ως μία *Λαμπερτιανή επιφάνεια* (Lambertian surface), δηλαδή μια επιφάνεια η οποία ανακλά ομοιόμορφα προς όλες τις κατευθύνσεις την προσπίπτουσα ακτινοβολία, με αποτέλεσμα να εμφανίζεται ομοιόμορφα φωτεινή από κάθε οπτική γωνία (ισότροπη επιφάνεια). Για το φωτισμό του 3Δ προσώπου θεωρείται μια μακρινή πηγή φωτός (distant illumination), ενώ οι συνολικές συνθήκες φωτισμού του σκηνικού περιγράφονται με χρήση *Σφαιρικών Αρμονικών Συναρτήσεων* (Spherical Harmonics) [40, 41, 42, 43].

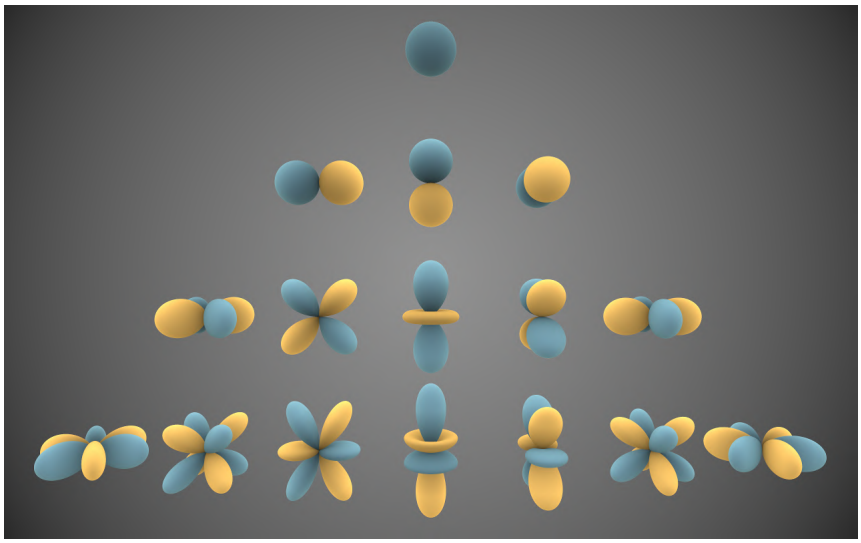
Οι σφαιρικές αρμονικές συναρτήσεις αποτελούν ένα ισχυρό μαθηματικό εργαλείο, το οποίο χρησιμοποιείται σε εφαρμογές γραφικών και υπολογιστικής όρασης για το φωτισμό αντικειμένων σε πραγματικό χρόνο (real-time). Πρόκειται στην ουσία για ένα σύνολο ορθοκανονικών συναρτήσεων βάσης, ανάλογο εκείνων του μετασχηματισμού Fourier, οι οποίες όμως ορίζονται πάνω στην επιφάνεια μιας σφαίρας. Η σφαιρική αρμονική συνιστώσα Y_{lm} δίνεται από την ακόλουθη σχέση [40],

$$Y_{lm}(\theta, \phi) = N_{lm} P_{lm}(\cos\theta) e^{Im\phi}, \tag{3.9}$$

με

$$N_{lm} = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}, \tag{3.10}$$

όπου N_{lm} είναι ένας παράγοντας κανονικοποίησης και P_{lm} το αντίστοιχο πολυώνυμο Legendre. Οι δείκτες l, m ικανοποιούν τις σχέσεις $l \geq 0$ και $-l \leq m \leq l$ αντίστοιχα, με το δείκτη l να εκφράζει ουσιαστικά την τάξη των σφαιρικών αρμονικών. Επομένως, δοθείσης μίας τάξης l υπάρχουν συνολικά $2l+1$ συναρτήσεις βάσης. Οι σφαιρικές αρμονικές συναρτήσεις μπορούν να γραφούν είτε ως τριγωνομετρικές συναρτήσεις σε σφαιρικές συντεταγμένες θ, ϕ, είτε ως πολυώνυμα σε καρτεσιανές συντεταγμένες x, y και z, με $x^2+y^2+z^2=1$. Στη γενική περίπτωση, η σφαιρική συνάρηση Y_{lm} εκφράζει ένα πολυώνυμο βαθμού l.



Σχήμα 3.9: Οι πρώτες 5 ομάδες σφαιρικών αρμονικών ($l = 0, \ldots, 4$). Το μπλε χρώμα αντιστοιχεί σε θετικές και το κίτρινο σε αρνητικές τιμές της συνάρτησης $Y_{lm}(\theta, \phi)$. Η απόσταση της επιφάνειας από την αρχή των νοητών αξόνων αντιστοιχεί στην απόλυτη τιμή της συνάρτησης $Y_{lm}(\theta, \phi)$ σε γωνιακή κατεύθυνση (θ, ϕ).

Οι πρώτες 9 σφαιρικές αρμονικές (με $l \leq 2$) αντιστοιχούν σε πολυώνυμα μηδενικού (σταθερό, $l = 0$), πρώτου (γραμμικό, $l = 1$) και δευτέρου ($l = 2$) βαθμού και σε καρτεσιανές συντεταγμένες (x, y, z) δίνονται αριθμητικά από τις σχέσεις [43]:

$$
\begin{aligned}
(x, y, z) &= (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta) \\
Y_{00}(\theta, \phi) &= 0.282095 \\
(Y_{11}; Y_{10}; Y_{1-1})\,(\theta, \phi) &= 0.488603(x; z; y) \\
(Y_{21}; Y_{2-1}; Y_{2-2})\,(\theta, \phi) &= 1.092548(xz; yz; xy) \\
Y_{20}(\theta, \phi) &= 0.315392\left(3z^2 - 1\right) \\
Y_{22}(\theta, \phi) &= 0.546274\left(x^2 - y^2\right)
\end{aligned}
\tag{3.11}
$$

Αξιοποιώντας τώρα τη θεωρία των σφαιρικών αρμονικών, μια τυχαία συνάρτηση $f(\theta, \phi)$ μπορεί να προσεγγιστεί χρησιμοποιώντας κατάλληλο αριθμό βάσεων $Y_{lm}(\theta, \phi)$ σφαιρικών αρμονικών, αφού υπολογιστούν οι αντίστοιχοι συντελεστές f_{lm}, σύμφωνα με τη σχέση [40],

$$
\begin{aligned}
f(\theta, \phi) &= \sum_{l=0}^{\infty} \sum_{m=-l}^{l} f_{lm} Y_{lm}(\theta, \phi), \\
f_{lm} &= \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} f(\theta, \phi) Y_{lm}^{*}(\theta, \phi) \sin\theta d\theta d\phi.
\end{aligned}
\tag{3.12}
$$

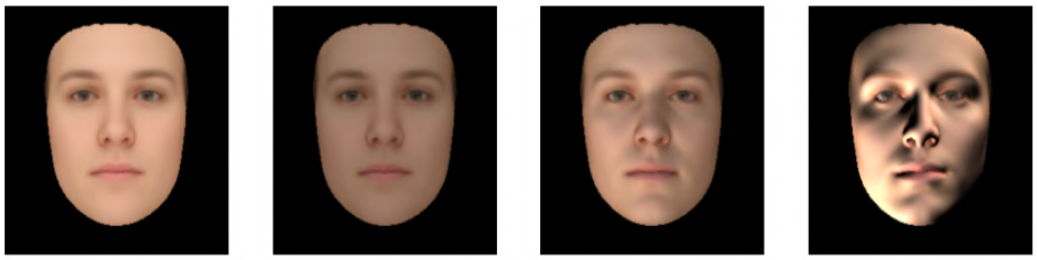
Η τελευταία σχέση αποτελεί τη βάση για τη μοντελοποίηση των συνθηκών φωτισμού ενός σκηνικού. Στόχος είναι η προσέγγιση της συνάρτησης *έντασης ακτινοβολίας* E (irradiance), η οποία εκφράζει την ροή ακτινοβολίας μέσω μιας επιφάνειας S, ανά μονάδα επιφάνειας. Προς την κατεύθυνση αυτή, οι Hanrahan και Ramamoorthi [43] έδειξαν ότι η συνάρτηση *έντασης ακτινοβολίας* E μπορεί να προσεγγιστεί αρκετά ικανοποιητικά χρησιμοποιώντας μόνο *9 συντελεστές σφαιρικών αρμονικών συναρτήσεων*, για $l = 0, 1, 2$ και $m = 0$, $-1 \leq m \leq 1$, $-2 \leq m \leq 2$ αντίστοιχα. Ακολουθώντας λοιπόν τη διαδικασία προσέγγισης με χρήση των σφαιρικών αρμονικών συναρτήσεων, η ακτινοβολία (radiosity) ενός σημείου $\mathbf{v}_i$ της επιφάνειας του 3Δ προσώπου, δίνεται από τη σχέση [19]:

$$
\mathbf{C}\left(\mathbf{r}_i, \mathbf{n}_i, \boldsymbol{\gamma}\right) = \mathbf{r}_i \cdot \sum_{b=1}^{B^2} \boldsymbol{\gamma}_b \mathbf{H}_b\left(\mathbf{n}_i\right), \tag{3.13}
$$

όπου $\boldsymbol{n}_i$ το μοναδιαίο κάθετο διάνυσμα (normal vector) της επιφάνειας στο σημείο $\mathbf{v}_i$, $\boldsymbol{r}_i = [R_i, G_i, B_i]^T$ το διάνυσμα υφής του σημείου $\mathbf{v}_i$, $\boldsymbol{H}_b : \mathbb{R}^3 \to \mathbb{R}$

οι σφαιρικές αρμονικές συναρτήσεις βάσης, $B = 3$ ο αριθμός των ομάδων των συναρτήσεων που χρησιμοποιούνται και $\boldsymbol{\gamma}_b \in \mathbb{R}^3$ οι αντίστοιχοι συντελεστές παραμετροποίησης του έγχρωμου φωτισμού, για κόκκινο, μπλε και πράσινο κανάλι χρώματος. Στο σημείο αυτό επισημαίνεται πως οι συναρτήσεις βάσης $\boldsymbol{H}_b$ αντιστοιχούν μία προς μία στις συναρτήσεις της Εξ.3.11 και συμβολίζονται με τον τρόπο αυτό για να αποφευχθεί η χρήση των δύο δεικτών l και m (κατά αντίστοιχο τρόπο και οι συντελεστές $\boldsymbol{\gamma}_b$).

Δεδομένου ότι οι τιμές των ορθοκανονικών βάσεων $\boldsymbol{H}_b$ μπορούν να υπολογιστούν εύκολα για κάποιο σημείο $\mathbf{v}_i$ γνωρίζοντας το αντίστοιχο διάνυσμα $\boldsymbol{n}_i$ μέσω των εξισώσεων της Σχ.3.11, για τον υπολογισμό της ακτινοβολίας του, $\mathbf{C}$, απαιτείται μόνο ο προσδιορισμός των 27 συντελεστών $\boldsymbol{\gamma}_b$ (9 συντελεστές για κάθε κανάλι χρώματος). Στο προτεινόμενο δίκτυο ανακατασκευής, ο προσδιορισμός των συντελεστών των σφαιρικών αρμονικών πραγματοποιείται μέσω του χρησιμοποιούμενου νευρωνικού δικτύου, το οποίο έως το σημείο αυτό καλείται να υπολογίσει συνολικά $224 + 27 = 251$ *συντελεστές*, εκ των οποίων οι πρώτοι χρησιμοποιούνται από το Basel Face Model για τον προσδιορισμό του σχήματος, της έκφρασης και της υφής του προσώπου και οι τελευταίοι εφαρμόζονται στη Σχ.3.13 για την απόδοση των συνθηκών φωτισμού του σκηνικού.



(i) (ii) (iii) (iv)

Σχήμα 3.10: Μέσο πρόσωπο Basel Face Model υπό διαφορετικές συνθήκες φωτισμού: (i) χωρίς φωτισμό, (ii) με 1 σφαιρική αρμονική, (iii) με 4 σφαιρικές αρμονικές και (iv) με 9 σφαιρικές αρμονικές συναρτήσεις.

Στο Σχ.3.10 παρουσιάζεται το μέσο πρόσωπο του Basel Face Model υπό διαφορετικές συνθήκες φωτισμού. Από τις εικόνες του σχήματος, εύκολα διαπιστώνεται ότι η χρήση σφαιρικών αρμονικών συναρτήσεων για τη μοντελοποίηση του φωτισμού οδηγεί σε πολύ πιο ρεαλιστικά μοντέλα προσώπων, με την αύξηση του αριθμού των συναρτήσεων που χρησιμοποιούνται για την προσέγγιση της ακτινοβολίας των σημείων του προσώπου να βελτιώνει αισθητά την ικανότητα του μοντέλου για αποτύπωση έντονων φωτιστικών διακυμάνσεων.

3.2.3 Μοντέλο Κάμερας

Για την αποτύπωση του σκηνικού και του 3Δ μοντέλου του προσώπου γίνεται χρήση μιας ιδανικής κάμερα μικρής οπής, δηλαδή μιας κάμερας με σημειακό διάφραγμα (apperture), η οποία δεν χρησιμοποιεί φακούς (lenses) για την εστίαση του φωτός. Η θέση και ο προσανατολισμός της κάμερας σε σχέση με το 3Δ σύστημα συντεταγμένων του κόσμου, περιγράφεται μέσω ενός μετασχηματισμού (rigid transformation), ο οποίος παραμετροποιείται βάσει μιας περιστροφής $\boldsymbol{R} \in \boldsymbol{SO}(3)$ και μιας μετατόπισης $\boldsymbol{t} \in \mathbb{R}^3$.

Ως προς την περιστροφή, η κάμερα μικρής οπής διαθέτει 3 βαθμούς ελευθερίας (DoF), περί των αξόνων X (roll), Y (pitch) και Z (yaw). Οι γωνίες αυτές είναι αριστερόστροφες ενώ οι άξονες περιστροφής αντιστοιχούν στο ορθοκανονικό σύστημα συντεταγμένων του κόσμου (WCS). Έτσι, οι περιστροφές της κάμερας κατά γωνίες ϕ, θ και ψ περί των αξόνων X, Y και Z αντίστοιχα, εκφράζονται από τα εξής μητρώα περιστροφής:

$$\begin{aligned}
\boldsymbol{R}_Z(\psi) &= \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
\boldsymbol{R}_Y(\theta) &= \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \\
\boldsymbol{R}_X(\phi) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{bmatrix}
\end{aligned} \tag{3.14}$$

Συνδυάζοντας τώρα κατάλληλα τα μητρώα αυτά, η συνολική περιστροφή της κάμερας περί των τριών αξόνων δίνεται από τον μητρώο

$$\begin{aligned}
\boldsymbol{R}(\phi,\theta,\psi) &= \boldsymbol{R}_Z(\psi)\boldsymbol{R}_Y(\theta)\boldsymbol{R}_X(\phi) \\
&= \begin{bmatrix} c_\psi c_\theta & c_\psi s_\theta s_\phi - s_\psi c_\phi & c_\psi s_\theta c_\phi + s_\psi s_\phi \\ s_\psi c_\theta & s_\psi s_\theta s_\phi + c_\psi c_\phi & s_\psi s_\theta c_\phi - c_\psi s_\phi \\ -s_\theta & c_\theta s_\phi & c_\theta c_\phi \end{bmatrix},
\end{aligned} \tag{3.15}$$

όπου $c_\phi = \cos\phi$ και $s_\phi = \sin\phi$ (ομοίως για τις γωνίες θ και ψ).

Ως προς τη μετατόπιση τώρα, η κάμερα διαθέτει και πάλι 3 βαθμούς ελευθερίας, κατά μήκος των αξόνων X, Y και Z του συστήματος συντεταγμένων του κόσμου και περιγράφεται μέσω του διανύσματος $\boldsymbol{t} = [t_X, t_Y, t_Z]^T$. Συνολικά, η κάμερα μικρής οπής που χρησιμοποιείται στα πλαίσια της προτεινόμενης μεθόδου ανακατασκευής, χαρακτηρίζεται από 6 βαθμούς ελευθερίας και περιγράφεται από ένα διάνυσμα $\boldsymbol{p} = [\phi, \theta, \psi, t_X, t_Y, t_Z]^T \in \mathbb{R}^6$.

Η μοντελοποίηση αυτή της κάμερας εισάγει 6 επιπλέον συντελεστές προς προσδιορισμό, με αποτέλεσμα το νευρωνικό δίκτυο του συστήματος ανακατασκευής να καλείται να προσδιορίσει τελικά ένα σύνολο από $251+6 = 257$ *παραμέτρους* (251 παράμετροι για το σχηματισμό και το φωτισμό του 3Δ μοντέλου και 6 για τη θέση και τον προσανατολισμό της κάμερας).

3.2.4 Προοπτική Προβολή

Το μοντέλο της κάμερας περιγράφει ουσιαστικά τον τρόπο με τον οποίο ένα 3Δ σημείο του κόσμου προβάλλεται σε ένα pixel μιας 2Δ εικόνας, ή σε ορολογία κάμερας, τον τρόπο με τον οποίο το φως που εκπέμπεται από μια επιφάνεια, ταξιδεύει και προσπίπτει στον αισθητήρα CCD. Στην παρούσα εργασία χρησιμοποιείται, όπως αναφέρθηκε και στην ενότητα 3.2.3, μία ιδανική κάμερα μικρής οπής, η οποία υποθέτει ότι το αισθητήριο επίπεδο βρίσκεται πίσω από ένα φράγμα, το οποίο διαθέτει ένα και μοναδικό σημείο εκ του οποίου επιτρέπεται η διέλευση φωτός. Στην περίπτωση του ιδανικού αυτού μοντέλου και όταν η απόσταση του κέντρου της κάμερας (κέντρο προβολής) από το επίπεδο προβολής (2Δ επίπεδο εικόνας) είναι πεπερασμένη, η προβολή ονομάζεται *προοπτική*.

Έστω τώρα ένα σημείο (X, Y, Z) του ορθοκανονικού συστήματος συντεταγμένων του κόσμου (WSC) και μία κάμερα μικρής οπής με άξονες X_c, Υ_c, Z_c (CCS) όπως φαίνεται στο Σχ.3.11. Για την προβολή του σημείου (X, Y, Z) πάνω στο προβολικό επίπεδο (PP) χρησιμοποιείται η σχέση

$$\begin{pmatrix} \widetilde{u} \\ \widetilde{v} \\ \widetilde{w} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & u_0 \\ 0 & 1 & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix}}_{\text{Παράμετροι Κάμερας}} \underbrace{\begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_{1\times 3} & 1 \end{pmatrix}^{-1}}_{\text{Αφινικός Μετασχηματισμός}} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \tag{3.16}$$

όπου $(\widetilde{u}, \widetilde{v}, \widetilde{w})$ το προβαλλόμενο στο επίπεδο της εικόνας σημείο, εκφρασμένο

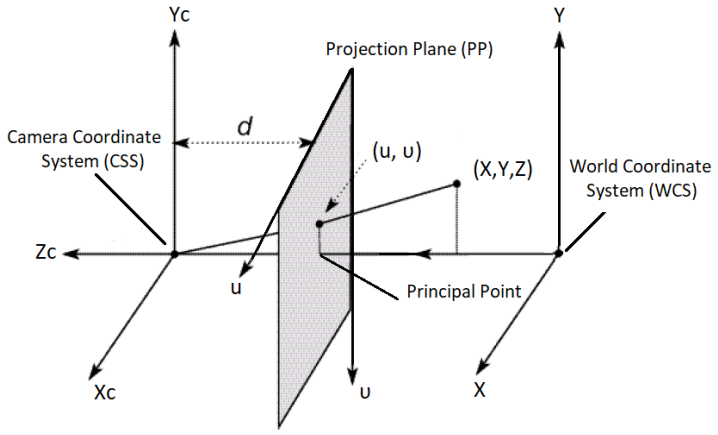
σε ομογενείς συντεταγμένες, f το εστιακό μήκος (focal length) της κάμερας, (u_0, v_0) το πρωτεύον σημείο προβολής (principal point), το οποίο αντιστοιχεί στο κέντρο του προβολικού επιπέδου της εικόνας, $\boldsymbol{R}$ το μητρώο περιστροφής και $\boldsymbol{t}$ το διάνυσμα μετατόπισης του συστήματος συντεταγμένων της κάμερας σε σχέση με το σύστημα συντεταγμένων του κόσμου.

Η διαδικασία επομένως της προοπτικής προβολής αποτελείται από δύο στάδια:

- Εφαρμογή ενός αφινικού μετασχηματισμού που σχηματίζεται από την περιστροφή και τη μετατόπιση της κάμερας και μετάβαση από το σύστημα συντεταγμένων του κόσμου στο αντίστοιχο σύστημα συντεταγμένων της κάμερας (WCS to CCS) και

- Μετάβαση από το σύστημα συντεταγμένων της κάμερας στο προβολικό επίπεδο της εικόνας μέσω πολλαπλασιασμού με τα μητρώα των παραμέτρων της κάμερας.

Όπως φαίνεται και από τη σχέση της Εξ.3.16, για την αναπαράσταση των 3Δ σημείων του κόσμου και των αντίστοιχων προβαλλόμενων σημείων στο επίπεδο της εικόνας, χρησιμοποιούνται ομογενείς συντεταγμένες, λόγω των γεωμετρικών και αριθμητικών ιδιοτήτων τους. Για τον υπολογισμό των καρτεσιανών συντεταγμένων (u, v) του προβαλλόμενου σημείου από τις αντίστοιχες ομογενείς συντεταγμένες $(\widetilde{u}, \widetilde{v}, \widetilde{w})$, χρησιμοποιούνται οι σχέσεις $u = \dfrac{\widetilde{u}}{\widetilde{w}}$ και $v = \dfrac{\widetilde{v}}{\widetilde{w}}$.



Σχήμα 3.11: Προοπτική Προβολή

Στα πλαίσια του προτεινόμενου μοντέλου ανακατασκευής, το εστιακό μήκος της κάμερας μικρής οπής επιλέγεται εμπειρικά περίπου ίσο με $f = 1000mm$, ενώ η κάμερα είναι τοποθετημένη στο σημείο $(0, 0, 10)(dm)$ του άξονα Z κοιτώντας προς την αρνητική κατεύθυνσή του. Τέλος, καθώς οι εικόνες του dataset έχουν διαστάσεις 224×244, το πρωτεύον σημείο της προβολής έχει συντεταγμένες $(u_0, v_0) = (112, 112)$.

3.3 Διαδικασία Ανακατασκευής

Στην παρούσα εργασία επιδιώκεται ο σχεδιασμός και η ανάπτυξη ενός συστήματος 3Δ ανακατασκευής προσώπων από εικόνες, με τρόπο ώστε η συνολική διαδικασία να μην επηρεάζεται κατά το δυνατόν από τις συνθήκες του περιβάλλοντος από το οποίο προέρχονται οι 2Δ εικόνες. Η προσπάθεια αυτή για σθεναρότητα και απεξαρτητοποίηση του συστήματος από περιβαλλοντικούς παράγοντες είναι ιδιαίτερα σημαντική και αναγκαία, καθώς αυτό καλείται να αντιμετωπίσει εικόνες από τον πραγματικό κόσμο, όπου οι μη ελεγχόμενες συνθήκες λήψης, επηρεάζουν άμεσα τα χαρακτηριστικά και τις λεπτομέρειες των εικονιζόμενων προσώπων. Προς την κατεύθυνση αυτή και λαμβάνοντας υπόψιν όσα αναφέρθηκαν στην ενότητα 2.2 σχετικά με τις υπάρχουσες μεθόδους ανακατασκευής, η προσέγγιση που ακολουθείται, βασίζεται στη χρήση συνελικτικών νευρωνικών δικτύων, τα οποία και εκπαιδεύονται χωρίς επίβλεψη.

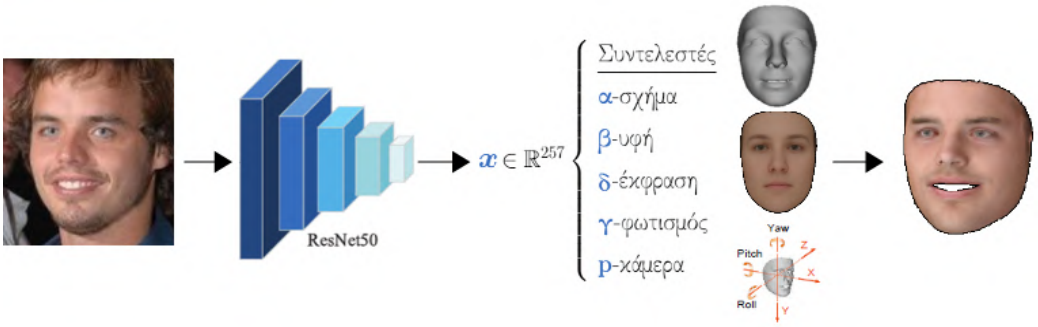
Στην ενότητα 3.2 έγινε η παραμετροποίση του περιβάλλοντος ανακατασκευής, η οποία συνίσταται στην τροποποίηση του Basel Face Model, έτσι ώστε αυτό να χρησιμοποιηθεί για την αναπαράσταση της 3Δ τοπολογίας των προσώπων, στον καθορισμό του είδους της κάμερας και της διαδικασίας της προοπτικής προβολής και τέλος στην μοντελοποίηση φυσικών φαινομένων όπως ο φωτισμός. Μέσω της μοντελοποίησης αυτής, προκύπτει ένα σύνολο άγνωστων συντελεστών, οι οποίοι καθορίζουν πλήρως την 3Δ γεωμετρία του προσώπου, την υφή, τη θέση και τον προσανατολισμό του στο χώρο, καθώς και τα φαινόμενα σκιάσεων και ανακλάσεων που παρουσιάζονται σε αυτό λόγω του φωτισμού. Οι συντελεστές αυτοί αναπαρίστανται μέσω του ακόλουθου διανύσματος:

$$\boldsymbol{x} = [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{p}]^T \in \mathbb{R}^{257}, \tag{3.17}$$

όπου $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ οι συντελεστές σχήματος, υφής και έκφρασης αντίστοιχα, $\boldsymbol{\gamma}$ οι

συντελεστές φωτισμού και $\boldsymbol{p}$ το διάνυσμα θέσης και προσανατολισμού της κάμερας σε σχέση με το σύστημα συντεταγμένων του κόσμου.

Δοθείσης επομένως μιας εικόνας ενός ατόμου, η διαδικασία της ανακατασκευής έγκειται στον προσδιορισμό των βέλτιστων συντελεστών του διανύσματος $\boldsymbol{x}$, έτσι ώστε το παραγόμενο 3Δ πρόσωπο να προσεγγίζει όσο το δυνατόν περισσότερο το πρόσωπο της εικόνας. Στα πλαίσια της προτεινόμενης μεθόδου, ο προσδιορισμός των 257 συνολικά αυτών συντελεστών γίνεται μέσω του συνελικτικού δικτύου *ResNet-50* [44], το οποίο παρουσιάζεται στην επόμενη ενότητα. Συνολικά, η διαδικασία ανακατασκευής παρουσιάζεται στο Σχ.3.12.



Σχήμα 3.12: Διαδικασία 3Δ ανακατασκευής προσώπου από 2Δ εικόνα

3.3.1 Αρχιτεκτονική ResNet-50

Στον πυρήνα της διαδικασίας της ανακατασκευής βρίσκεται το δίκτυο ResNet-50, το οποίο χρησιμοποιείται για την εκτίμηση του διανύσματος συντελεστών $\boldsymbol{x}$ και κατ΄ επέκταση για τον πλήρη προσδιορισμό του σχηματιζόμενου 3Δ προσώπου. Το ResNet-50 πρόκειται ουσιαστικά για ένα βαθύ συνελικτικό δίκτυο με συνολικά 50 επίπεδα, το οποίο χρησιμοποιείται κατά κύριο λόγο για την εξαγωγή χαρακτηριστικών από εικόνες (feature extraction). Το δίκτυο αυτό εισήγαγε για πρώτη φορά ως λογική στην εκπαίδευση ενός δικτύου την μάθηση των Συναρτήσεων Καταλοίπων (Residual Functions) μέσω παραλειπόμενων συνδέσεων (skip connections), αποκλίνοντας από την παραδοσιακή λογική των ακολουθιακών δικτύων (sequential networks). Οι συναρτήσεις καταλοίπων έχουν ως σκοπό το δίκτυο να προσπαθεί να μάθει τα κατάλοιπα των συνελικτι-

κών επιπέδων, δηλαδή των επιπρόσθετων επιπέδων τα οποία δεν χρησιμοποιούνται στην εκπαίδευση. Κατά αυτό τον τρόπο, διευκολύνεται η οπισθοδιάδοση του σφάλματος (backpropagation) καθώς υπάρχουν περισσότερες διαθέσιμες διαδρομές και διατηρείται σε υψηλά επίπεδα η απόδοση του δικτύου ακόμα και για μεγάλο βάθος επιπέδων.

Στη γενική του μορφή το δίκτυο ResNet-50 αποτελείται από 50 επίπεδα εκ των οποίων το τελευταίο πρόκειται για ένα πλήρως συνδεδεμένο επίπεδο (fully-connected layer) 1000 νευρώνων. Στα πλαίσια ωστόσο της προτεινόμενης μεθόδου ανακατασκευής το δίκτυο καλείται να εκτιμήσει ένα διάνυσμα συνολικά 257 συντελεστών. Για το λόγο αυτό το τελευταίο πλήρως συνδεδεμένο επίπεδο των 1000 νευρώνων αντικαθίσταται από ένα επίσης πλήρως συνδεδεμένο επίπεδο 257 νευρώνων. Στους πίνακες 3.1 και 3.2 παρουσιάζεται η δομή και το πλήθος των παραμέτρων του τροποποιημένου δικτύου ResNet-50 αντίστοιχα.

Πίνακας 3.1: Αρχιτεκτονική τροποποιημένου δικτύου ResNet-50.

Layer Name	**Output Size**	**ResNet-50**
conv1	112×112	$7 \times 7, 64$, stride 2
conv2_x	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
conv5_x	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
average pool	1×1	average pool
fully connected	1×257	257 full connections

Πίνακας 3.2: Παράμετροι τροποποιημένου δικτύου ResNet-50

Συνολικός αριθμός παραμέτρων:	24,114,305
Αριθμός εκπαιδεύσιμων παραμέτρων:	24,061,185
Αριθμός μη εκπαιδεύσιμων παραμέτρων:	53,120

3.3.2 Διαδικασία Εκπαίδευσης

Στις προηγούμενες ενότητες παρουσιάστηκε το θεωρητικό υπόβαθρο και επιλέχθηκαν τα κατάλληλα μοντέλα για το σχηματισμό και την παραμετροποίηση του περιβάλλοντος ανακατασκευής. Παρ΄ όλα αυτά, το δίκτυο ανακατασκευής δεν είναι ακόμα λειτουργικό, καθώς το προεκπαιδευμένο μοντέλο ResNet-50 το οποίο και είναι υπεύθυνο για την εκτίμηση του διανύσματος συντελεστών $\boldsymbol{x}$ είναι σχεδιασμένο για να αντιμετωπίζει προβλήματα ταξινόμησης εικόνων και όχι προβλήματα 3Δ ανακατασκευής προσώπων. Έχοντας λοιπόν ως αφετηρία τις προκαθορισμένες τιμές των βαρών του, όπως αυτές προέκυψαν από την εκπαίδευση στο dataset ImageNet [45], το τροποποιημένο ResNet-50 θα πρέπει να συνεχίσει την εκπαίδευσή του στις εικόνες του dataset που έχει δημιουργηθεί στα πλαίσια της παρούσας εργασίας (ενότητα 3.1), έτσι ώστε να προσαρμοστεί στις ανάγκες και τις απαιτήσεις του προβλήματος της 3Δ ανακατασκευής.

Η εκπαίδευση βασίζεται σε τεχνικές μη επιβλεπόμενης μάθησης και συνίσταται στην ελαχιστοποίηση κάποιων συναρτήσεων απώλειας, οι οποίες εκφράζουν την διαφορά μεταξύ του 3Δ ανακατασκευασμένου μοντέλου και του αντίστοιχου προσώπου της 2Δ εικόνας εισόδου. Οι συναρτήσεις αυτές θα πρέπει να επιλεχθούν κατάλληλα έτσι ώστε να εξετάζουν όσο το δυνατόν μεγαλύτερο εύρος του φάσματος των χαρακτηριστικών του ανθρώπινου προσώπου, εξασφαλίζοντας ότι το παραγόμενο συνθετικό πρόσωπο θα είναι ρεαλιστικό και λεπτομερές.

Στα πλαίσια της εργασίας, για την εκπαίδευση του προτεινόμενου δικτύου ανακατασκευής χρησιμοποιείται μια *υβριδική συνάρτηση απώλειας* (hybrid loss function) L, η οποία αποτελείται από τις εξής επιμέρους συναρτήσεις:

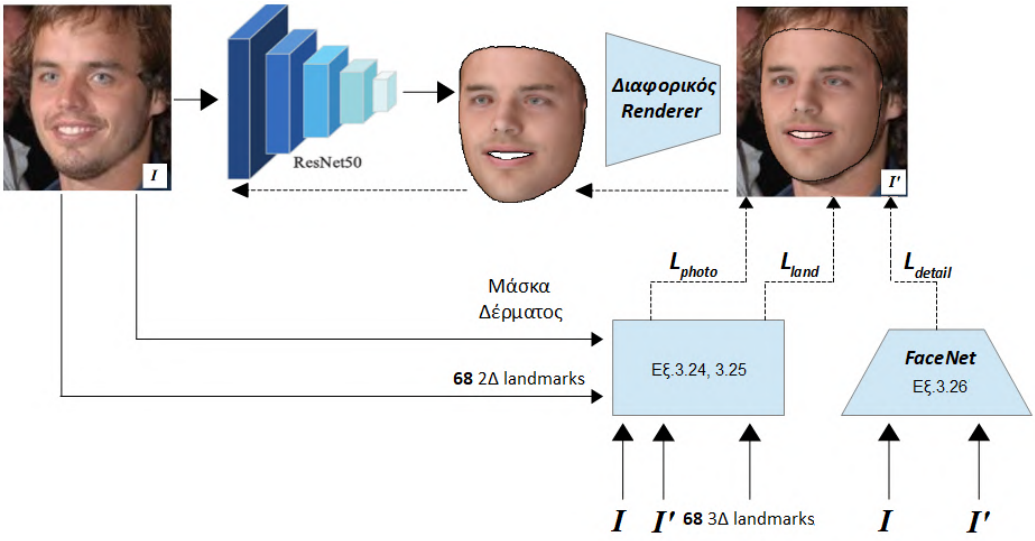
- *Φωτομετρική Συνάρτηση Απώλειας* (Photometric Loss) L_{photo}

- *Συνάρτηση Απώλειας Σημείων Ενδιαφέροντος* (Landmark Loss) L_{land}
- *Συνάρτηση Απώλειας Λεπτομερειών* (Detail Loss) L_{detail}
- *Όρος Κανονικοποίησης* (Regularization Term) L_{reg}

Έτσι, δοθείσης μιας έγχρωμης RGB εικόνας εισόδου I, το τροποποιημένο δίκτυο ResNet-50 εξάγει ένα διάνυσμα συντελεστών $\boldsymbol{x}$ βάσει του οποίου κατασκευάζεται το 3Δ μοντέλο του προσώπου. Εν συνεχεία, το 3Δ αυτό μοντέλο προβάλλεται πάνω σε μια 2Δ εικόνα $I' = I'(\boldsymbol{x})$, η οποία προκύπτει αναλυτικά μέσω της διαδικασίας του rendering που θα αναφερθεί παρακάτω. Τελικά, η συνθετική αυτή εικόνα I' συγκρίνεται με την αρχική εικόνα I και υπολογίζεται η υβριδική συνάρτηση απώλειας

$$L(\boldsymbol{x}) = \omega_{photo}L_{photo}(\boldsymbol{x}) + \omega_{land}L_{land}(\boldsymbol{x}) + \omega_{detail}L_{detail}(\boldsymbol{x}) + \omega_{reg}L_{reg}(\boldsymbol{x}), \quad (3.18)$$

όπου w_k ο συντελεστής βαρύτητας της αντίστοιχης συνάρτησης L_k, με $k = \{photo, land, detail, reg\}$, ο οποίος εκφράζει το ποσοστό συμμετοχής της στην υβριδική συνάρτηση απώλειας.



Σχήμα 3.13: Διαδικασία μη επιβλεπόμενης εκπαίδευσης του προτεινόμενου δικτύου ανακατασκευής. Τα βέλη με τις διακεκομμένες γραμμές εκφράζουν τη διαδρομή του σφάλματος κατά την οπισθοδιάδοση (backpropagation).

Στη συνέχεια δίνεται σε ψευδοκώδικα ο αλγόριθμος που υλοποιείται για τη μη επιβλεπόμενη εκπαίδευση του προτεινόμενου δικτύου ανακατασκευής.

Algorithm 1 Αλγόριθμος εκπαίδευσης του δικτύου ανακατασκευής

Δεδομένα: σύνολο N επεξεργασμένων εικόνων προσώπων I
for $1 \leq \text{iter} \leq \text{max\_iter}$ **do**
  Επιλογή batch k εικόνων
  for $1 \leq j \leq k$ **do**
    Εξαγωγή διανύσματος $\boldsymbol{x_j}$ της εικόνας I_j μέσω του ResNet-50
    Σχηματισμός 3Δ μοντέλου προσώπου βάσει του διανύσματος $\boldsymbol{x}_j$
    Rendering και προβολή του 3Δ μοντέλου σε 2Δ εικόνα $I'(\boldsymbol{x}_j)$
  end for
  Υπολογισμός υβριδικής συνάρτησης απώλειας $L(\boldsymbol{x})$ για όλο το batch
  Ενημέρωση βαρών του ResNet-50
end for

Rendering-Rasterization

Σύμφωνα με τη διαδικασία ανακατασκευής που περιγράφηκε στην αρχή της ενότητας 3.3, δοθέντος ενός διανύσματος συντελεστών $\boldsymbol{x}$ το οποίο έχει προκύψει ως έξοδος του δικτύου ResNet-50 για μια δεδομένη εικόνα εισόδου I, κατασκευάζεται η αντίστοιχη 3Δ τοπολογία προσώπου. Το 3Δ αυτό παραγόμενο μοντέλο του προσώπου θα πρέπει στη συνέχεια να συγκριθεί με την 2Δ εικόνα εισόδου, έτσι ώστε να διαπιστωθεί η ομοιότητα μεταξύ των δύο προσώπων. Για να καταστεί δυνατή η σύγκριση αυτή, το 3Δ πρόσωπο θα πρέπει να προβληθεί με κατάλληλο τρόπο στο επίπεδο της εικόνας, καθώς η σύγκριση μεταξύ 2Δ και 3Δ δεδομένων είναι ιδιαιτέρως δύσκολη και αναποτελεσματική.

Η διαδικασία αυτή, δηλαδή η παραγωγή μιας φωτορεαλιστικής εικόνας $I'(\boldsymbol{x})$ από ένα 3Δ μοντέλο, το οποίο περιγράφεται από το διάνυσμα συντελεστών $\boldsymbol{x}$, αναφέρεται ως rendering και αποτελεί ιδιαίτερο αντικείμενο μελέτης στον τομέα των γραφικών. Για την υλοποίηση της διαδικασίας του rendering έχουν αναπτυχθεί και προταθεί αρκετοί αλγόριθμοι, καθένας εκ των οποίων βασίζεται σε διαφορετικές τεχνικές για τον υπολογισμό της τελικής εικόνας I'. Στόχος σε

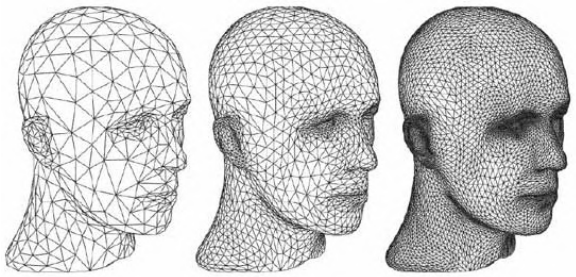
κάθε περίπτωση είναι η αποδοτική και ακριβής μεταφορά του 3Δ μοντέλου και των συνθηκών φωτισμού του περιβάλλοντός του στο επίπεδο της εικόνας.

Μία ιδιαίτερα διαδεδομένη τεχνική rendering είναι αυτή του *Rasterization*, από το raster το οποίο εκφράζει την πλεγματική μορφή μιας εικόνας. Η τεχνική αυτή προσπαθεί ουσιαστικά να επιλύσει *το πρόβλημα της ορατότητας* (visibility problem), δηλαδή να προσδιορίσει τις περιοχές εκείνες μιας 3Δ επιφάνειας, οι οποίες είναι ορατές από την κάμερα, καθώς υπάρχουν και περιοχές, οι οποίες είτε βρίσκονται εκτός του οπτικού πεδίου (field of view) της κάμερας, είτε καλύπτονται από άλλα αντικείμενα. Για την αναπαράσταση της 3Δ δομής μιας επιφάνειας χρησιμοποιείται, όπως ήδη έχει αναφερθεί, το τριγωνικό πλέγμα (triangle mesh), το οποίο όπως φαίνεται και στην εικόνα του Σχ.3.14 αποτελείται από ένα σύνολο 3Δ τριγώνων, τα οποία συνδέονται μέσω των κοινών πλευρών ή κορυφών τους. Ο αριθμός των τριγώνων που χρησιμοποιούνται για την κατασκευή του πλέγματος μπορεί να μεταβάλλεται, ανάλογα με το πόσο πυκνή και λεπτομερής απαιτείται να είναι η αναπαράσταση της επιφάνειας.

Επιστρέφοντας τώρα στη διαδικασία του rasterization μιας επιφάνειας, αυτή μπορεί να αποσυντεθεί χονδρικά σε δύο βασικά στάδια:

1) προσδιορισμός των ορατών περιοχών της επιφάνειας (visibility)

2) σκίαση των προβαλλόμενων στην εικόνα ορατών περιοχών (shading).



Σχήμα 3.14: Αναπαράσταση 3Δ δομής του προσώπου μέσω τριγωνικών πλεγμάτων.

Για τον προσδιορισμό των ορατών περιοχών της επιφάνειας, όλα τα τρίγωνα του πλέγματός της προβάλλονται αρχικά στο επίπεδο της εικόνας μέσω προοπτικής προβολής. Πιο συγκεκριμένα, κάθε σημείο το οποίο αποτελεί κορυφή τριγώνου στο 3Δ πλέγμα προβάλλεται μέσω προοπτικής προβολής στο επίπεδο της εικόνας, με αποτέλεσμα να σχηματίζονται σε αυτή 2Δ τρίγωνα (Σχ.3.15).

Στη συνέχεια, στο στάδιο της σκίασης, εξετάζεται κάθε σημείο της εικόνας έτσι ώστε να διαπιστωθεί εάν αυτό κείται σε κάποιο εκ των 2Δ τριγώνων, τα οποία και αντιστοιχούν σε ορατές περιοχές της επιφάνειας. Για την αναπαράσταση ενός σημείου $\boldsymbol{P}$ συναρτήσει των κορυφών $\boldsymbol{V}_0$, $\boldsymbol{V}_1$ και $\boldsymbol{V}_2$ ενός τριγώνου χρησιμοποιούνται οι βαρυκεντρικές συντεταγμένες $(\lambda_0, \lambda_1, \lambda_2)$ (barycentric coordinates) [46], σύμφωνα με τις οποίες το σημείο P εκφράζεται ως,

$$\boldsymbol{P} = \lambda_0 \boldsymbol{V}_0 + \lambda_1 \boldsymbol{V}_1 + \lambda_2 \boldsymbol{V}_2. \tag{3.19}$$

Στη γενική περίπτωση, οι συντεταγμένες αυτές μπορούν να λάβουν οποιαδήποτε τιμή. Σε περίπτωση ωστόσο που το σημείο P βρίσκεται εντός του τριγώνου με κορυφές τα σημεία $\boldsymbol{V}_0$, $\boldsymbol{V}_1$ και $\boldsymbol{V}_2$ (στο εσωτερικό του ή σε κάποια πλευρά του), περιορίζονται στο εύρος $[0, 1]$ και το άθροισμά τους ισούται με 1 (κανονικοποιημένες). Με άλλα λόγια,

$$\lambda_0 + \lambda_1 + \lambda_2 = 1, \text{ για κάθε } \boldsymbol{P} \in \triangle \boldsymbol{V}_0, \boldsymbol{V}_1, \boldsymbol{V}_2. \tag{3.20}$$

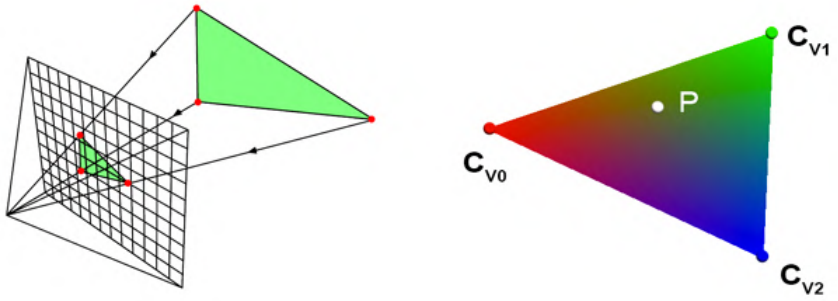
Εάν επομένως για κάποιο σημείο $\boldsymbol{P}$ εκφρασμένο σε βαρυκεντρικές συντεταγμένες ως προς κάποιο τρίγωνο ισχύει η Εξ.3.20, τότε αυτό βρίσκεται εντός του αντίστοιχου τριγώνου και ανήκει σε ορατή περιοχή της επιφάνειας. Για το χρωματισμό του, δεδομένου ότι οι χρωματικές αποχρώσεις των κορυφών του τριγώνου είναι $\boldsymbol{C}_{\boldsymbol{V}_0}$, $\boldsymbol{C}_{\boldsymbol{V}_1}$ και $\boldsymbol{C}_{\boldsymbol{V}_2}$ αντίστοιχα, χρησιμοποιείται η σχέση [47],

$$\boldsymbol{C}_{\boldsymbol{P}} = \lambda_0 \boldsymbol{C}_{\boldsymbol{V}_0} + \lambda_1 \boldsymbol{C}_{\boldsymbol{V}_1} + \lambda_2 \boldsymbol{C}_{\boldsymbol{V}_2}, \tag{3.21}$$

όπου $\boldsymbol{C}_{\boldsymbol{P}}$ το χρωματικό διάνυσμα του σημείου $\boldsymbol{P}$. Έτσι, κάθε ορατό σημείο της εικόνας χρωματίζεται ανάλογα με τις χρωματικές αποχρώσεις των κορυφών του τριγώνου στο οποίο ανήκει, όπως φαίνεται και στο Σχ.3.15.

Τέλος, το σύνολο των ορατών σημείων αντιστοιχίζεται σε pixels της εικόνας σύμφωνα με συγκεκριμένους κανόνες, οι οποίοι έχουν ως σκοπό η τελική εικόνα να είναι ρεαλιστική, χωρίς ασυνέχειες και αναδιπλώσεις χρωμάτων (overlaps).

Στο σημείο αυτό αξίζει να αναφερθεί ότι η διαδικασία που αναλύθηκε στην παρούσα ενότητα, αποτελεί τη βάση της τεχνικής του rasterization, όπως αυτή υλοποιείται από πληθώρα διαθέσιμων αλγορίθμων. Παρά ταύτα, η εξέλιξη της επιστήμης των γραφικών έχει οδηγήσει στην ανάπτυξη και το σχεδιασμό πιο σύνθετων rasterizers, οι οποίοι για την προβολή του 3Δ μοντέλου στην εικόνα χρησιμοποιούν πιο περίπλοκους κανόνες και τεχνικές, έτσι ώστε οι παραγόμενες rendered εικόνες να είναι όσο το δυνατόν πιο ρεαλιστικές.



Σχήμα 3.15: Προβολή τριγώνων 3Δ πλέγματος στην επιφάνεια της εικόνας (αριστερά) και υπολογισμός χρωματικής απόχρωσης σημείου βάσει βαρυκεντρικών συντεταγμένων (δεξιά).

Μοναδιαία Διανύσματα Σημείων Πλέγματος

Πριν την ανάλυση των συναρτήσεων απώλειας οι οποίες χρησιμοποιούνται κατά την εκπαίδευση του δικτύου ανακατασκευής, κρίνεται σκόπιμο να γίνει αναφορά στον τρόπο με τον οποίο υπολογίζονται τα κατευθυντικά μοναδιαία διανύσματα των σημείων της επιφάνειας του προσώπου (vertex normals), καθώς αυτά, όπως αναφέρθηκε και στην ενότητα 3.2.2, κωδικοποιούν χρήσιμη πληροφορία σχετικά με την ορατότητα των περιοχών του προσώπου και τις συνθήκες φωτισμού.

Η διαδικασία αυτή του υπολογισμού των διανυσμάτων των σημείων μιας επιφάνειας δεν είναι μονοσήμαντα ορισμένη, καθώς μπορεί να υλοποιηθεί με διαφορετικούς τρόπους, οι οποίοι εξαρτώνται κυρίως από το είδος της αναπαράστασης που χρησιμοποιείται για την εξεταζόμενη επιφάνεια. Στα πλαίσια της εργασίας, η 3Δ δομή του ανθρώπινου προσώπου αναπαρίσταται μέσω ενός τριγωνικού πλέγματος, το οποίο απαρτίζεται από τρίγωνα συνδεδεμένα μέσω των κοινών πλευρών ή κορυφών τους (Σχ.3.14). Επιπλέον, για κάθε σημείο $\mathbf{v}_i$ του πλέγματος θεωρείται η 1-ring γειτονιά του (1-ring neighborhood), σύμφωνα με την οποία, ως γείτονες θεωρούνται τα σημεία εκείνα τα οποία συνδέονται με το $\mathbf{v}_i$ μέσω μιας ακμής τριγώνου του πλέγματος (Σχ.3.16).

Έχοντας προσδιορίσει λοιπόν την έννοια της γειτονιάς ενός σημείου $\mathbf{v}_i$ του τριγωνικού πλέγματος, η διαδικασία υπολογισμού του αντίστοιχου μοναδιαίου κατευθυντικού διανύσματός του $\boldsymbol{n}_i$ περιλαμβάνει τα εξής βήματα [48]:

Βήμα 1

Για κάθε γειτονικό τρίγωνο $T_{i,j}$ του σημείου $\mathbf{v}_i$ του πλέγματος με $1 \leq j \leq m$, υπολογίζεται το αντίστοιχο μοναδιαίο κάθετο διάνυσμά του $\boldsymbol{n}_{i,j}$ (triangle normal). Ο υπολογισμός αυτός μπορεί να γίνει με σχετικά απλό τρόπο, προσδιορίζοντας αρχικά τα διανύσματα δύο ακμών του $\boldsymbol{e}_{j,1}$ και $\boldsymbol{e}_{j,2}$ αντίστοιχα και λαμβάνοντας στη συνέχεια το εξωτερικό τους γινόμενο $\boldsymbol{e}_{j,1} \times \boldsymbol{e}_{j,2}$. Τελικά, το μοναδιαίο κάθετο διάνυσμα του τριγώνου προκύπτει έπειτα από κανονικοποίηση του εξωτερικού γινομένου. Συνολικά, η προαναφερθείσα διαδικασία εκφράζεται από την ακόλουθη σχέση:

$$\boldsymbol{n}_{i,j} = \frac{\boldsymbol{e_{j,1}} \times \boldsymbol{e_{j,2}}}{||\boldsymbol{e_{j,1}} \times \boldsymbol{e_{j,2}}||}, \text{ για } 1 \leq j \leq m, \tag{3.22}$$

όπου $\boldsymbol{e}_{j,1}$ και $\boldsymbol{e}_{j,2}$ τα διανύσματα δύο ακμών του τριγώνου $T_{i,j}$ και m ο συνολικός αριθμός των γειτονικών τριγώνων του πλέγματος του σημείου $\mathbf{v}_i$.

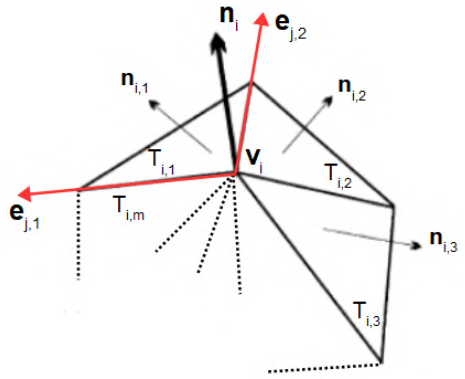
Βήμα 2

Έχοντας υπολογίσει όλα τα μοναδιαία κάθετα διανύσματα $\boldsymbol{n}_{i,j}$ των γειτονικών τριγώνων του σημείου $\mathbf{v}_i$, το μοναδιαίο κατευθυντικό διάνυσμά του $\boldsymbol{n}_i$ μπορεί να προκύψει αθροίζοντας αρχικά τα διανύσματα αυτά και κανονικοποιώντας εν συνεχεία το διανυσματικό τους άθροισμα, σύμφωνα με τη σχέση,

$$\boldsymbol{n}_i = \frac{\sum_{j=1}^{m} \boldsymbol{n}_{i,j}}{\left|\left|\sum_{j=1}^{m} \boldsymbol{n}_{i,j}\right|\right|}. \tag{3.23}$$

Στα πλαίσια της παρούσας εργασίας, για τον υπολογισμό των μοναδιαίων διανυσμάτων των σημείων του τριγωνικού πλέγματος του προσώπου χρησιμοποιείται η σχέση της Εξ.3.23, με το πλήθος των γειτόνων ενός σημείου $\mathbf{v}_i$ στην 1-ring γειτονιά του να ανέρχεται σε $m = 8$.

Τα μοναδιαία αυτά κατευθυντικά διανύσματα είναι ζωτικής σημασίας για την εξακρίβωση της ορατότητας των σημείων του πλέγματος του 3Δ προσώπου και την αποτύπωση των συνθηκών φωτισμού του περιβάλλοντος κατά τη διαδικασία του rendering. Στο Σχ.3.16 παρουσιάζεται η 1-ring γειτονιά ενός σημείου $\mathbf{v}_i$, μαζί με τα διάφορα είδη διανυσμάτων που χρησιμοποιούνται για τον υπολογισμό του επιθυμητού μοναδιαίου κατευθυντικού διανύσματος $\boldsymbol{n}_i$.



Σχήμα 3.16: 1-ring γειτονιά ενός σημείου $\mathbf{v}_i$ του τριγωνικού πλέγματος. Στο σχήμα φαίνονται τα γειτονικά τρίγωνα $T_{i,j}$, τα μοναδιαία κάθετα διανύσματα $\boldsymbol{n}_{i,j}$ των τριγώνων αυτών, τα διανύσματα ακμών $\boldsymbol{e}_{j,1}$ και $\boldsymbol{e}_{j,2}$ και το μοναδιαίο κατευθυντικό διάνυσμα $\boldsymbol{n}_i$ στο σημείο $\mathbf{v}_i$.

Φωτομετρική Συνάρτηση Απώλειας

Απαραίτητη προϋπόθεση για την επιτυχή λειτουργία του δικτύου ανακατασκευής αποτελεί η ακριβής αποτύπωση των χρωματικών διακυμάνσεων του προσώπου της 2Δ εικόνας στο αντίστοιχο 3Δ μοντέλο. Κατά τη διάρκεια της εκπαίδευσης του δικτύου, θα πρέπει να παρακολουθείται με κάποιο τρόπο η χρωματική διαφορά μεταξύ της αρχικής εικόνας εισόδου I και της παραγόμενης εικόνας $I'(\boldsymbol{x})$ που προκύπτει έπειτα από rendering του 3Δ μοντέλου, το οποίο και προσδιορίζεται από το διάνυσμα $\boldsymbol{x}$. Ένας τρόπος για τον προσδιορισμό της διαφοράς αυτής είναι προφανώς η άμεση σύγκριση των χρωματικών τιμών των pixels των δύο εικόνων, σε μία 1-1 αντιστοιχία. Η σύγκριση αυτή ωστόσο αγνοεί την έντονη συσχέτιση που παρουσιάζουν τα γειτονικά pixels της επιφάνειας του προσώπου ενός ατόμου, με αποτέλεσμα να οδηγεί σε αριθμητικώς ακριβείς τιμές, οι οποίες όμως δεν μπορούν να αξιοποιηθούν αποτελεσματικά από το δίκτυο ανακατασκευής.

Για το λόγο αυτό, ακολουθώντας τη μέθοδο των Deng *et al.*[30], για τον προσδιορισμό των χρωματικών διαφορών μεταξύ της αρχικής και της παραγόμενης εικόνας, χρησιμοποιείται η ακόλουθη φωτομετρική συνάρτηση απώλειας,

$$L_{photo}(\boldsymbol{x}) = \frac{\sum_{i \in M} A_i \cdot ||I - I'(\boldsymbol{x})||_2}{\sum_{i \in M} A_i}, \tag{3.24}$$

όπου η μεταβλητή i αντιστοιχεί σε δείκτη pixel, A_i είναι οι τιμές της μάσκας δέρματος της εικόνας I, όπως αυτή αναλύθηκε στην ενότητα 3.1.3 και M είναι η κεντρική περιοχή του προβαλλόμενου προσώπου, η οποία προκύπτει έπειτα από την εφαρμογή κατάλληλης μάσκας στις εικόνες I και I' (Σχ.3.17). Για τον υπολογισμό της άμεσης φωτομετρικής διαφοράς μεταξύ των εικόνων χρησιμοποιείται η L_2-νόρμα ($||\cdot||_2$).

(i) (ii) (iii)

Σχήμα 3.17: (i) Μέσο πρόσωπο Basel Face Model, (ii) Μάσκα για την εξαγωγή της κεντρικής περιοχής του προσώπου, (iii) Κεντρική περιοχή του προσώπου.

Η χρήση της μάσκας δέρματος στον υπολογισμό της φωτομετρικής συνάρτησης απώλειας έχει ως αποτέλεσμα τα pixels τα οποία χαρακτηρίζονται από υψηλό ποσοστό αβεβαιότητας ως προς το είδος τους (δερματικά ή μη) να μην συμμετέχουν με την ίδια βαρύτητα σε σχέση με τα pixels τα οποία είναι βέβαιο ή σχεδόν βέβαιο ότι αντιστοιχούν σε pixels δέρματος. Επιπλέον, η χρήση μάσκας για τη διατήρηση μόνο της κεντρική περιοχής M του προσώπου, εστιάζει τον υπολογισμό της φωτομετρικής διαφοράς στα σημαντικά μέρη του προσώπου, ενισχύοντας κατά αυτό τον τρόπο την ακρίβεια των αποτελεσμάτων.

Συνάρτηση Απώλειας Σημείων Ενδιαφέροντος

Η συνάρτηση απώλειας σημείων ενδιαφέροντος αφορά στη διαφορά μεταξύ των σημείων ενδιαφέροντος του προσώπου της αρχικής 2Δ εικόνας εισόδου και του αντίστοιχου παραγόμενου 3Δ μοντέλου. Κατά τη διάρκεια της προεπεξεργασίας των εικόνων εντοπίστηκαν και αποθηκεύτηκαν σε κατάλληλα αρχεία, 68

σημεία ενδιαφέροντος για κάθε εικόνα. Έτσι, κατά τη διάρκεια της εκπαίδευσης, τα σημεία αυτά μπορούν να χρησιμοποιηθούν μαζί με τα αντίστοιχα 3Δ σημεία ενδιαφέροντος του τροποποιημένου Basel Face Model, έτσι ώστε να εξασφαλιστεί η σωστή γεωμετρική απόδοση των παραγόμενων μοντέλων.

Η σύγκριση ωστόσο των δύο αυτών ειδών σημείων ενδιαφέροντος δεν είναι ιδιαιτέρως εφικτή και αποτελεσματική, καθώς τα μεν πρώτα προσδιορίζονται στις 2 διαστάσεις και τα δε δεύτερα στις 3.

Για να αντιμετωπιστεί το πρόβλημα αυτό, τα σημεία ενδιαφέροντος του 3Δ μοντέλου του προσώπου προβάλλονται μέσω προοπτική προβολής (διαδικασία η οποία αναλύθηκε στην ενότητα 3.2.4) στο επίπεδο της εικόνας και έτσι πλέον είναι δυνατή η σύγκρισή τους με τα αντίστοιχα σημεία της εικόνας εισόδου. Συμβολίζοντας λοιπόν με $\{\boldsymbol{l}_n\}$ τα 68 2Δ σημεία ενδιαφέροντος του προσώπου της εικόνα εισόδου και με $\{\boldsymbol{l'_n}(\boldsymbol{x})\}$ τα αντίστοιχα, προβαλλόμενα στις 2 διαστάσεις, σημεία του 3Δ μοντέλου που προσδιορίζεται από το διάνυσμα συντελεστών $\boldsymbol{x}$, η συνάρτηση απώλειας σημείων ενδιαφέροντος δίνεται από τη σχέση,

$$L_{land}(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N}\omega_n||\boldsymbol{l}_n - \boldsymbol{l}'_n(\boldsymbol{x})||^2, \text{ για } N = 68, \tag{3.25}$$

όπου ω_n το βάρος καθενός εκ των σημείων ενδιαφέροντος (Σχ.3.5), το οποίο καθορίζει το ποσοστό συνεισφοράς του στο σύνολο της συνάρτησης απώλειας.

Συνάρτηση Απώλειας Λεπτομερειών

Η φωτομετρική συνάρτηση απώλειας εξασφαλίζει αρκετά ικανοποιητικά αποτελέσματα ανακατασκευής όσον αφορά στην απόδοση των χρωμάτων της επιφάνειας του προσώπου. Παρ΄ όλα αυτά, όταν χρησιμοποιείται μόνη της, το δίκτυο αδυνατεί να προσδιορίσει λεπτομέρειες της επιφάνειας του προσώπου του ατόμου (π.χ. χρωματικές ανωμαλίες, δομικές ανομοιομορφίες), με αποτέλεσμα να παράγονται 3Δ μοντέλα τα οποία παρουσιάζουν μια σχετική ομαλότητα και μία υπερβολική ομοιομορφία.

Ως εκ τούτου, θα πρέπει να ληφθούν υπόψιν και πιο ουσιαστικά χαρακτηριστικά του προσώπου ενός ατόμου, τα οποία και θα χρησιμοποιηθούν κατά τη διάρκεια της εκπαίδευσης του δικτύου ανακατασκευής. Προς την κατεύθυνση αυτή, στα πλαίσια της προτεινόμενης μεθόδου χρησιμοποιείται το προεκπαι-

δευμένο δίτκυο *FaceNet* [49], το οποίο πρόκειται ουσιαστικά για ένα δίκτυο αναγνώρισης προσώπων (face recognition) που αναπτύχθηκε το 2015 από ερευνητές της Google. Το δίκτυο αυτό, δοθείσης μιας εικόνας στην είσοδο, εντοπίζει το εικονιζόμενο πρόσωπο και εξάγει τα βαθιά χαρακτηριστικά αυτού (deep features) υπό τη μορφή ενός διανύσματος 128 στοιχείων.

Χρησιμοποιώντας λοιπόν το δίκτυο αυτό, γίνεται εξαγωγή των χαρακτηριστικών των προσώπων της αρχικής εικόνας εισόδου I και της αντίστοιχης παραγόμενης εικόνας $I'(\boldsymbol{x})$ και ορίζεται η ακόλουθη συνάρτηση απώλειας λεπτομερειών, σύμφωνα με την απόσταση συνημιτόνου (cosine distancy),

$$L_{detail}(\boldsymbol{x}) = 1 - \underbrace{\frac{\langle f(I), f(I'(\boldsymbol{x}))\rangle}{||f(I)|| \cdot ||f(I'(\boldsymbol{x}))||}}_{\text{ομοιότητα συνημιτόνου}}, \tag{3.26}$$

όπου οι όροι $f(I)$ και $f(I'(\boldsymbol{x}))$ δηλώνουν τα κωδικοποιημένα διανύσματα των χαρακτηριστικών των προσώπων των εικόνων I και $I'(\boldsymbol{x})$ αντίστοιχα και ο τελεστής $\langle \cdot, \cdot \rangle$ εκφράζει το εσωτερικό γινόμενο.

Σύμφωνα λοιπόν με την Εξ.3.26, όταν δύο διανύσματα χαρακτηριστικών διαφέρουν κατά πολύ μεταξύ τους, δηλαδή παρουσιάζουν σημαντική απόκλιση, ο όρος της ομοιότητας συνημιτόνου προσεγγίζει το 0 (τα κάθετα διανύσματα παρουσιάζουν τη μεγαλύτερη απόκλιση) και ως εκ τούτου η απόσταση συνημιτόνου προσεγγίζει το 1. Με αντίστοιχο τρόπο, όταν δύο διανύσματα χαρακτηριστικών μοιάζουν πολύ μεταξύ τους, ο όρος της ομοιότητας συνημιτόνου προσεγγίζει το 1 και ως εκ τούτου η απόσταση συνημιτόνου προσεγγίζει το 0.

Όρος Κανονικοποίησης

Καθώς το δίκτυο ResNet-50 που χρησιμοποιείται στο σύστημα ανακατασκευής πρόκειται για ένα ιδιαίτερα βαθύ συνελικτικό δίκτυο, κατά τη διάρκεια της εκπαίδευσής του εγκυμονεί ο κίνδυνος της υπερπροσαρμογής (overfitting), δηλαδή της εξειδίκευσής του στα δεδομένα του dataset εκπαίδευσης. Αυτό έχει ως αποτέλεσμα το δίκτυο να είναι ιδιαιτέρως αποδοτικό στην ανακατασκευή προσώπων που προέρχονται από εικόνες του dataset που χρησιμοποιήθηκε για την εκπαίδευσή του, χωρίς ωστόσο να παρουσιάζει την ίδια ικανότητα σε τυχαίες εικόνες εισόδου που δεν ανήκουν σε αυτό.

Για να αντιμετωπιστεί λοιπόν το πρόβλημα αυτό, χρησιμοποιείται η τεχνική της κανονικοποίησης, η οποία αποτελεί έναν από τους πιο συνηθισμένους τρόπους αποφυγής της υπερπροσαρμογής κατά την εκπαίδευση ενός δικτύου. Βασική ιδέα της κανονικοποίησης είναι ο έλεγχος των τιμών των συντελεστών τους οποίους προβλέπει το δίκτυο, έτσι ώστε να περιορίζονται οι μεγάλες τιμές, οι οποίες είναι υπεύθυνες κατά κύριο λόγο για την απότομη μεταβολή των βαρών του. Κατά αυτό τον τρόπο, η μεταβολή των βαρών κατά τη διάρκεια της εκπαίδευσης γίνεται με πιο ομαλό τρόπο, εμποδίζοντας το δίκτυο να οδηγηθεί σε αστάθεια και ενισχύοντας συγχρόνως την ικανότητά του να αντιμετωπίζει άγνωστες εισόδους.

Για την υλοποίηση της κανονικοποίησης υπάρχουν διάφορες τεχνικές, οι οποίες αφορούν είτε στην προσθήκη κατάλληλων συνελικτικών επιπέδων στην αρχιτεκτονική του δικτύου (dropout), είτε στην επαύξηση των δεδομένων εκπαίδευσης (data augmentation), είτε στην άμεση προσθήκη ενός όρου τιμωρίας (penaty term) στη συνάρτηση απώλειας που χρησιμοποιείται κατά την εκπαίδευση (L_1, L_2 - Regularization).

Στα πλαίσια της προτεινόμενης μεθόδου, χρησιμοποείται η L_2-κανονικοποίηση, σύμφωνα με την οποία προστίθεται στη συνάρτηση απώλειας του δικτύου ο ακόλουθος όρος τιμωρίας,

$$L_{reg}(\boldsymbol{x}) = \omega_\alpha ||\boldsymbol{\alpha}||^2 + \omega_\beta ||\boldsymbol{\beta}||^2 + \omega_\delta ||\boldsymbol{\delta}||^2, \tag{3.27}$$

όπου $\boldsymbol{x}$ είναι το διάνυσμα συντελεστών που εκτιμάει το δίκτυο και $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ οι συντελεστές σχήματος, υφής και έκφρασης αντίστοιχα. Τα βάρη ω_α, ω_β και ω_δ καθορίζουν το ποσοστό συμμετοχής των συνιστωσών σχήματος, υφής και έκφρασης αντίστοιχα στον όρο τιμωρίας.

Η προσθήκη του όρου αυτού στη συνάρτηση απώλειας έχει ως αποτέλεσμα οι συντελεστές $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ και $\boldsymbol{\delta}$ να ελέγχονται ως προς το εύρος των τιμών το οποίο μπορεί να λάβουν και να περιορίζονται γενικά σε μικρές τιμές, έτσι ώστε η μεταβολή των αντίστοιχων χαρακτηριστικών του παραγόμενου προσώπου να είναι μικρή, αποφεύγοντας έντονες διακυμάνσεις, οι οποίες οδηγούν σε μη ρεαλιστικά αποτελέσματα.

Τέλος, σημειώνεται πως αύξηση των τιμών των βαρών της Εξ.3.27 θέτει πιο στενά περιθώρια μεταβολής για τις τιμές των αντίστοιχων συντελεστών, ενώ μείωσή τους οδηγεί σε χαλάρωση του ελέγχου, επιτρέποντας μεγαλύτερες σχηματικές, εκφραστικές και χρωματικές μεταβολές για το παραγόμενο πρόσωπο.

Κεφάλαιο 4

Πειραματικό Μέρος

Στο κεφάλαιο 3 αναλύθηκε πλήρως η προτεινόμενη διαδικασία ανακατασκευής και περιγράφηκαν τα βήματα που ακολουθούνται για την εκπαίδευση του δικτύου, καθώς και οι αντίστοιχες συναρτήσεις απώλειας, οι οποίες επιλέγονται έτσι ώστε να εξετάζουν όσο το δυνατόν περισσότερες πτυχές των χαρακτηριστικών του ανθρώπινου προσώπου. Έως το σημείο αυτό, η ανάλυση ήταν κυρίως θεωρητική, χωρίς να παρέχονται πληροφορίες σχετικά με την αριθμητική τιμή των παραμέτρων και των συντελεστών που παρουσιάζονται στη διαδικασία της ανακατασκευής.

Στο πλαίσιο αυτό, το παρόν κεφάλαιο είναι αφιερωμένο στο πρακτικό μέρος της προτεινόμενης μεθόδου, δηλαδή στην εκπαίδευση του δικτύου και στην μετέπειτα εφαρμογή του σε πραγματικά δεδομένα. Έτσι, αρχικά προσδιορίζονται οι τιμές των παραμέτρων που χρησιμοποιούνται κατά την εκπαίδευση και εξετάζεται η επίδρασή τους στη συνολική διαδικασία της ανακατασκευής. Εν συνεχεία, γίνεται μια ποιοτική και ποσοτική αξιολόγηση του εκπαιδευμένου πλέον δικτύου, μέσω της εφαρμογής του σε τυχαίες πραγματικές εικόνες προσώπων, οι οποίες προέρχονται από διαφορετικά datasets, έτσι ώστε να καλύπτουν όσο το δυνατόν περισσότερες συνθήκες ανακατασκευής. Τέλος, γίνεται ένας γενικός σχολιασμός των αποτελεσμάτων και εξάγωνται συμπεράσματα για την αποτελεσματικότητα της προτεινόμενης μεθόδου, ενώ επισημαίνονται κάποια σημεία καίριας σημασίας για την απόδοση του δικτύου.

4.1 Παράμετροι Εκπαίδευσης

Στα πλαίσια της διπλωματικής εργασίας, ο κώδικας που αναπτύχθηκε για την προεπεξεργασία των εικόνων του dataset και την υλοποίηση του δικτύου ανακατασκευής, είναι γραμμένος εξ ολοκλήρου σε *python*. Όσον αφορά στο κομμάτι της Μηχανικής Μάθησης, για την τροποποίηση και την εκπαίδευση του συνελικτικού δικτύου ResNet-50 χρησιμοποιήθηκαν οι βιβλιοθήκες *tensorflow* και *keras*, καθώς αυτές συνδυάζουν την εύκολη διαχείριση και τροποποίηση προεκπαιδευμένων δικτύων και την ελευθερία για πλήρη εξατομίκευση της διαδικασίας της εκπαίδευσης. Για την υλοποίηση της διαδικασίας του rendering των 3Δ μοντέλων των προσώπων, αξιοποιήθηκε ο διαφορικός renderer της βιβλιοθήκης γραφικών *tensorflow-graphics*, ο οποίος παράγει εικόνες υψηλής ακρίβειας και πιστότητας.

Όπως αναφέρθηκε και στην ενότητα 3.3.2, η εκπαίδευση του δικτύου ανακατασκευής γίνεται χωρίς επίβλεψη και βασίζεται στην ελαχιστοποίηση κατάλληλα επιλεγμένων συναρτήσεων απώλειας. Προκειμένου ωστόσο η διαδικασία της εκπαίδευσης να ολοκληρωθεί αποτελεσματικά και να οδηγήσει στη σύγκλιση του δικτύου, απαιτείται ο προσδιορισμός μιας πληθώρας παραμέτρων, οι οποίες αφορούν στον αλγόριθμο βελτιστοποίησης που χρησιμοποιείται, στο batch-size, στον ρυθμό μάθησης (learning rate) και στα βάρη των συναρτήσεων απώλειας.

Για την επιλογή των κατάλληλων τιμών των παραμέτρων αυτών, το δίκτυο εκπαιδεύεται αρχικά για ένα μικρό αριθμό επαναλήψεων, χρησιμοποιώντας πολύ λίγες εικόνες ως δεδομένα εκπαίδευσης και η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές με διαφορετικούς συνδυασμούς τιμών παραμέτρων. Ακολουθώντας το παράδειγμα της πλειοψηφίας της υπάρχουσας βιβλιογραφίας, οι τιμές των παραμέτρων εκπαίδευσης που εξετάζονται, κυμαίνονται στα εύρη που φαίνονται στον πίνακα 4.1.

Πίνακας 4.1: Εξεταζόμενα εύρη τιμών παραμέτρων εκπαίδευσης

Παράμετρος	**Εύρος Τιμών**
batch-size	4-10
ω_{photo}, ω_{detail}, ω_{land}, ω_n	0.5-2.5, 0.1-1, $0.001 - 0.05$, 1-20
ω_{reg}	$0.0001 - 0.001$
ω_α, ω_β, ω_δ	$0.5 - 1.5$, $0.001 - 0.1$, $0.5 - 1.5$

Τελικά, λαμβάνοντας υπόψιν την επίδραση των παραμέτρων τόσο στις επιμέρους συναρτήσεις απώλειας, όσο και στην ταχύτητα σύγκλισης, επιλέγονται για την εκπαίδευση του προτεινόμενου δικτύου οι τιμές του πίνακα 4.2.

Πίνακας 4.2: Παράμετροι εκπαίδευσης δικτύου ανακατασκευής.

Παράμετρος	**Τιμή**
Optimizer, learning rate	Adam, $1e-4$
batch-size	8
ω_{photo}, ω_{detail}	2, 0.5
ω_{land}	$2e-3$
ω_n	$\begin{cases} 10, & \text{για } 0 \leq n \leq 16 \\ 20, & \text{για } 28 \leq n \leq 30 \text{ και } 60 \leq n \leq 67 \\ 1, & \text{σε οποιαδήποτε άλλη περίπτωση} \end{cases}$
ω_{reg}	$3e-4$
ω_α, ω_β, ω_δ	1, $1e-2$, 0.8

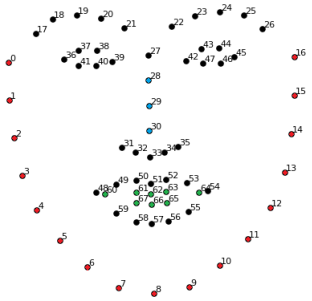
Στο σημείο αυτό υπενθυμίζεται, ότι το βάρος ω_n εκφράζει το ποσοστό συμμετοχής του n-οστού σημείου ενδιαφέροντος του προσώπου στη συνάρτηση απώλειας της Εξ.3.25. Καθώς τα σημεία ενδιαφέροντος του προσώπου δεν είναι όλα εξίσου απαραίτητα για τον προσδιορισμό της γεωμετρίας του, ενισχύεται η συμμετοχή των σημείων εκείνων, τα οποία οριοθετούν περιοχές του, οι οποίες μεταβάλλονται έντονα από άτομο σε άτομο. Έτσι, τα σημεία ενδιαφέροντος της κάτω γνάθου ($0 \leq n \leq 16$) έχουν τιμή βάρους ίση με 10, ενώ τα σημεία της περιοχής της μύτης ($28 \leq n \leq 30$) και του στόματος ($60 \leq n \leq 67$) έχουν τιμή βάρους ίση με 20. Όλα τα υπόλοιπα σημεία συμμετέχουν με το ίδιο ποσοστό στη συνάρτηση απώλειας, με τιμή βάρους ίση με 1 (Σχ.4.1). Κατά αυτό τον τρόπο το δίκτυο δίνει προτεραιότητα και μεγαλύτερη σημασία στην ελαχιστοποίηση της απόστασης μεταξύ των σημείων ενδιαφέροντος της κάτω γνάθου, της μύτης και του στόματος του 3Δ παραγόμενου προσώπου και των αντίστοιχων σημείων του προσώπου της 2Δ εικόνας.

Χρησιμοποιώντας τώρα τις τιμές των παραμέτρων του πίνακα 4.2, το δίκτυο εκπαιδεύεται για περίπου $900k$ επαναλήψεις, με τις τιμές των συναρτήσεων απώλειας L_{photo}, L_{land} και L_{detail} να μειώνονται σταδιακά και να σταθεροποιούνται στις εξής τιμές:

- η L_{photo} περίπου στην τιμή 0.15,
- η L_{land} περίπου στην τιμή 8 και
- η L_{detail} περίπου στην τιμή 0.2.

Αξίζει να σημειωθεί, ότι οι τιμές των παραμέτρων εκπαίδευσης επηρεάζουν σημαντικά την απόδοση του δικτύου ως προς τις επιμέρους συνιστώσες της ανακατασκευής. Έτσι, αυξημένη τιμή του βάρους ω_{photo} οδηγεί σε σημαντική μείωση της τιμής της συνάρτησης απώλειας L_{photo} και ως εκ τούτου σε αποτελέσματα με μεγαλύτερη ακρίβεια ως προς τη συνιστώσα της υφής, τα οποία ωστόσο δεν αποτυπώνουν καλά τη γεωμετρία και τα χαρακτηριστικά των 2Δ προσώπων, καθώς οι συναρτήσεις απώλειας L_{land} και L_{detail} δεν μειώνονται εξίσου με την L_{photo}. Αντίστοιχη είναι και η ερμηνεία στην περίπτωση υψηλών τιμών για τα βάρη ω_{land} και ω_{detail}, όπου το δίκτυο εκπαιδεύεται με έμφαση στη γεωμετρία ή τα χαρακτηριστικά των προσώπων αντίστοιχα, με αποτέλεσμα να υστερεί στις υπόλοιπες συνιστώσες της ανακατασκευής.

Για το λόγο αυτό, οι τιμές του πίνακα 4.2 επιλέγονται, όπως αναφέρθηκε και νωρίτερα, ώστε να υπάρχει μια ισοτιμία και ένας συμβιβασμός μεταξύ των επιμέρους συναρτήσεων απώλειας και να παράγονται κατά αυτό τον τρόπο 3Δ μοντέλα τα οποία αποτυπώνουν με εξίσου ικανοποιητικό τρόπο τη γεωμετρία, την υφή και τα χαρακτηριστικά των αντίστοιχων 2Δ προσώπων.



Σχήμα 4.1: Τα 68 σημεία ενδιαφέροντος του προσώπου. Με κόκκινο (0-16) φαίνονται τα σημεία ενδιαφέροντος της περιοχής της κάτω γνάθου, με μπλέ (28-30) τα σημεία της περιοχής της μύτης και με πράσινο (60-67) τα σημεία της περιοχής του στόματος.

4.2 Αποτελέσματα Ανακατασκευής

Στην ενότητα αυτή θα γίνει παρουσίαση των αποτελεσμάτων που προέκυψαν έπειτα από εφαρμογή του εκπαιδευμένου δικτύου ανακατασκευής σε τυχαία πραγματικά δεδομένα. Τα δεδομένα αυτά πρόκειται ουσιαστικά για έγχρωμες 2Δ εικόνες προσώπων, οι οποίες προέρχονται από τα ακόλουθα datasets:

- **CelebA Dataset** [23]: πρόκειται για το γνωστό dataset (ενότητα 3.1), το οποίο αποτελείται από εικόνες προσώπων διάσημων ατόμων. Από το dataset αυτό, επιλέγονται συνολικά $5,000$ εικόνες για την αξιολόγηση του δικτύου ανακατασκευής. Οι εικόνες αυτές παρουσιάζονται για πρώτη φορά στο δίκτυο και δεν έχουν συμπεριληφθεί στο σύνολο των εικόνων που χρησιμοποιούνται για την εκπαίδευσή του.

- **LFW (Labeled Faces in the Wild) Face Dataset** [50]: το dataset αυτό αποτελείται από ένα σύνολο $13,233$ έγχρωμων 2Δ εικόνων προσώπων, οι οποίες φέρουν ετικέτες με τα ονόματα των αντίστοιχων ατόμων. Ο αριθμός των διαφορετικών εικονιζόμενων ατόμων ανέρχεται σε $5,749$, εκ των οποίων τα $1,680$ παρουσιάζονται περισσότερες από μία φορές, υπό διαφορετικές οπτικές γωνίες και σε μεταβαλλόμενες συνθήκες περιβάλλοντος. Από το dataset αυτό, επιλέγεται για την αξιολόγηση του δικτύου ένα σύνολο $5,000$ εικόνων.

- **300W-LP Dataset** [51]: το εν λόγω dataset αποτελείται συνολικά από $61,225$ έγχρωμες 2Δ εικόνες προσώπων, καθένα εκ των οποίων εμφανίζεται περισσότερες από μία φορές. Ιδιαίτερο χαρακτηριστικό του dataset αυτού αποτελεί το γεγονός ότι είναι σε πολύ μεγάλο βαθμό τεχνητό. Έτσι, έχοντας ως βάση μια αρχική εικόνα ενός ατόμου, παράγεται ένα σύνολο εικόνων, στις οποίες το άτομο παρουσιάζεται σε μεταβαλλόμενες πόζες, οι οποίες καλύπτουν όλο το εύρος των δυνατών θέσεων τοποθέτησης του προσώπου του. Από το εν λόγω dataset επιλέγονται συνολικά $4,500$ εικόνες, οι οποίες χρησιμοποιούνται κυρίως για την αξιολόγηση της απόδοσης του δικτύου ανακατασκευής σε έντοντες πόζες.

- **UTKFace Dataset** [52]: πρόκειται για ένα dataset έγχρωμων 2Δ εικόνων προσώπων, το οποίο περιλαμβάνει περισσότερες από $20,000$ εικόνες, οι οποίες έχουν ληφθεί σε μη ελεγχόμενο περιβάλλον, κάτω από διαφορετικές συνθήκες φωτισμού. Τα εικονιζόμενα άτομα εμφανίζονται σε

διάφορες πόζες, με μεταβαλλόμενες εκφράσεις, ενώ αισθητή είναι και η παρουσία στοιχείων τα οποία προκαλούν αποκρύψεις περιοχών του προσώπου (γυαλιά, καπέλο, μούσι, μαλλιά εντός της περιοχής του προσώπου κ.ά.). Για την αξιολόγηση του δικτύου επιλέγονται συνολικά 3,700 εικόνες από το UTKFace dataset.

- **FFHQ (Flickr-Faces-HQ) Dataset** [53]: το εν λόγω dataset αποτελείται από 70,000 2Δ εικόνες προσώπων υψηλής ανάλυσης, οι οποίες έχουν συλλεχθεί από την ιστοσελίδα Flickr και έχουν υποστεί ευθυγράμμιση και περικοπή με χρήση των εργαλείων της βιβλιοθήκης dlib. Σε αντίθεση με τις προηγούμενες περιπτώσεις, το dataset αυτό αποτελεί το μόνο το οποίο παρέχει εικόνες υψηλής ανάλυσης και ως εκ τούτου κρίνεται σκόπιμο να μελετηθεί, παρόλο που η όλη διαδικασία ανακατασκευής βασίζεται εν γένει σε εικόνες χαμηλής σχετικά ανάλυσης. Από το FFHQ dataset επιλέγονται και χρησιμοποιούνται για την αξιολόγηση της ανακατασκευής συνολικά 2,500 εικόνες.

Στον πίνακα 4.3 συνοψίζονται τα datasets που χρησιμοποιούνται, καθώς και ο αριθμός των εικόνων που επιλέγεται από καθένα εξ αυτών.

Πίνακας 4.3: Datasets εικόνων προσώπου για την αξιολόγηση της απόδοσης του προτεινόμενου δικτύου ανακατασκευής.

Dataset	**Πλήθος Εικόνων**	**Πλήθος Επιλεγμένων Εικόνων**
CelebA	202,599	5,000
LFW	13,233	5,000
300W-LP	61,225	4,500
UTKFace	20,000	3,700
FFHQ	70,000	2,500

Καθώς τα αποτελέσματα της ανακατασκευής είναι σε πολύ μεγάλο βαθμό οπτικά, η αξιολόγηση της απόδοσης του δικτύου συντελείται τόσο ποιοτικά όσο και ποσοτικά. Στην πρώτη περίπτωση, τα παραγόμενα μοντέλα προσώπων συγκρίνονται ποιοτικά με τα αντίστοιχα πρόσωπα των 2Δ εικόνων, ως προς την οπτική τους ομοιότητα, ενώ στη δεύτερη, χρησιμοποιούνται κατάλληλες μετρικές για την αριθμητική διαπίστωση της ομοιότητας αυτής. Οι δύο αυτές προ-

σεγγίσεις της αξιολόγησης παρουσιάζονται στις επόμενες ενότητες, όπου και διευκρινίζεται πλήρως ο τρόπος με τον οποίο πραγματοποιούνται.

Στο σημείο αυτό σημειώνεται, ότι αφού ολοκληρωθεί η εκπαίδευση του δικτύου ανακατασκευής, δεν απαιτείται η εφαρμογή της προεπεργασίας που αναλύθηκε στην ενότητα 3.1 στις εικόνες που δίνονται ως είσοδοι, καθώς τα σημεία ενδιαφέροντος των προσώπων και οι αντίστοιχες μάσκες δέρματος χρησιμοποιούνται μόνο στα πλαίσια της εκπαίδευσης για τον υπολογισμό των συναρτήσεων απώλειας. Παρ΄ όλα αυτά, προκειμένου να διευκολυνθεί η διαδικασία της ανακαστασκευής και καθώς το δίκτυο ResNet-50 που χρησιμοποιείται δέχεται εικόνες διαστάσεων 224×224, κάθε εικόνα, πριν αυτή οδηγηθεί στην είσοδο, υφίσταται τη διαδικασία της ευθυγράμμισης και περικοπής, όπως αυτή περιγράφεται στην ενότητα 3.1.1.

4.2.1 Ποιοτική Αξιολόγηση

Γενικά Αποτελέσματα Ανακατασκευής

Όπως αναφέρθηκε και νωρίτερα, η ποιοτική αξιολόγηση των αποτελεσμάτων της ανακατασκευής συνίσταται στην οπτική σύγκριση του παραγόμενου 3Δ μοντέλου του προσώπου με το αντίστοιχο πρόσωπο της 2Δ εικόνας. Η σύγκριση αυτή, αν και επηρεάζεται σημαντικά από την οπτική του εκάστοτε παρατηρητή, προσφέρει μια αρκετά ικανοποιητική εικόνα για την απόδοση του δικτύου, δεδομένου ότι η οπτική ομοιότητα μεταξύ δύο προσώπων αποτελεί βασική προϋπόθεση και επιδίωξη της ανακατασκευής.

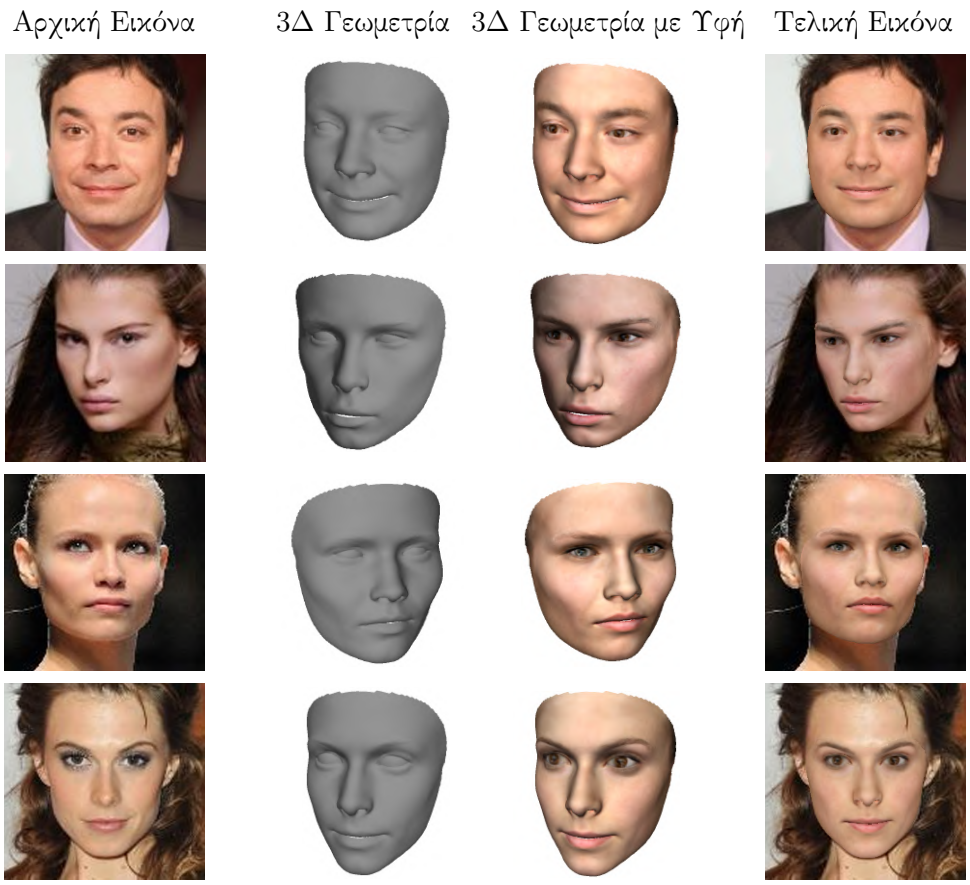
Η παρουσίαση των ποιοτικών αποτελεσμάτων γίνεται ξεχωριστά για κάθε ένα εκ των datasets του πίνακα 4.3, παραθέτοντας κάθε φορά επιλεγμένα δείγματα ανακατασκευής, έτσι ώστε να καλύπτονται διαφορετικές πτυχές του περιβάλλοντος και των χαρακτηριστικών του ανθρώπινου προσώπου. Προκειμένου η μετέπειτα ποιοτική αξιολόγηση να είναι όσο το δυνατόν πιο αντικειμενική και πλήρης, για κάθε εικόνα ενός dataset παρατίθενται τα εξής:

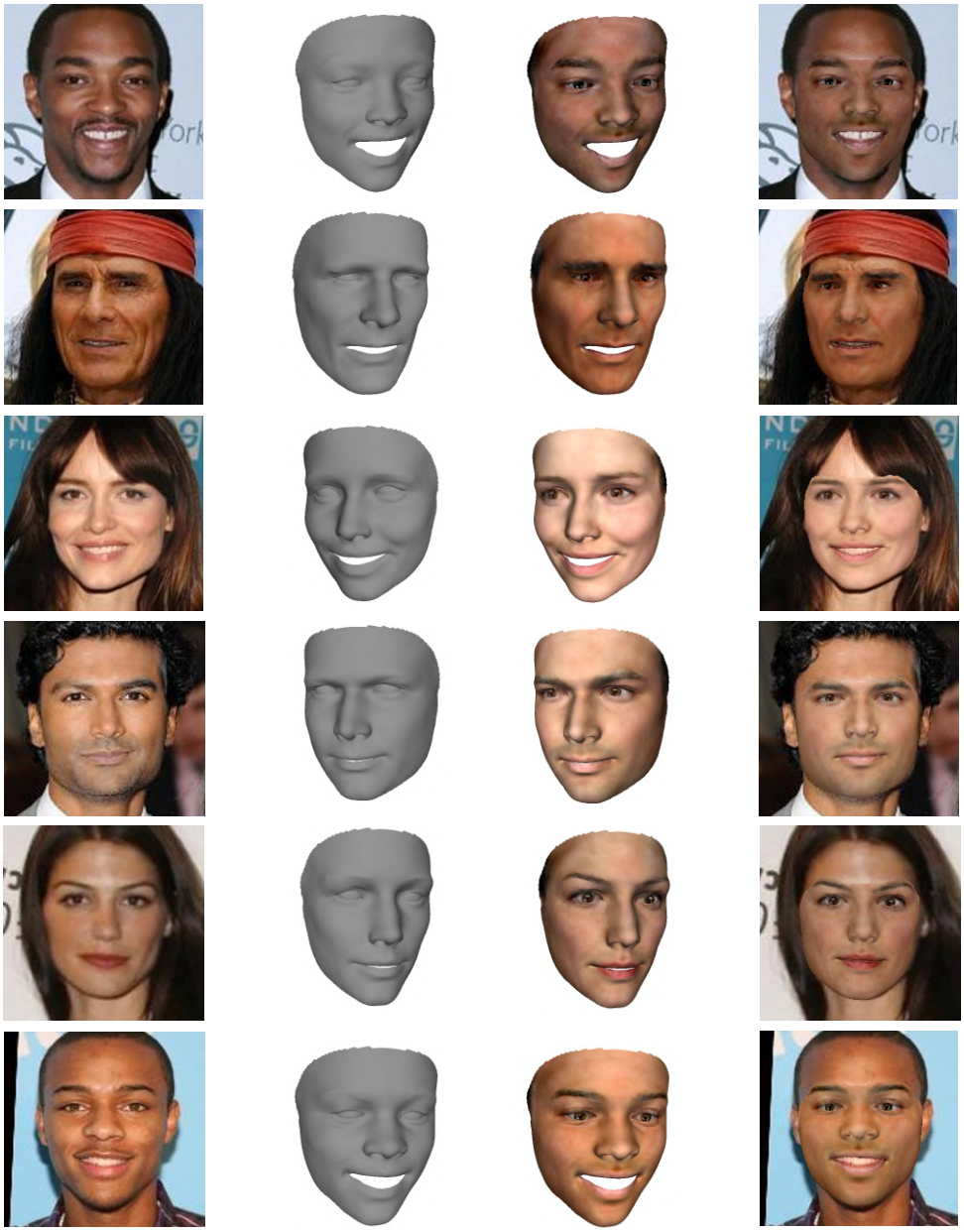
- Η 3Δ γεωμετρία (σχήμα, έκφραση) του ανακατασκευασμένου μοντέλου του προσώπου,
- Το ολοκληρωμένο 3Δ μοντέλο, το οποίο περιλαμβάνει τόσο τη γεωμετρία όσο και την υφή του προσώπου και

- Η αρχική 2Δ εικόνα, στην οποία το πρόσωπο του ατόμου αντικαθίσταται από το αντίστοιχο ανακατασκευασμένο 3Δ πρόσωπο, αφού πρώτα αυτό προβληθεί στο επίπεδό της μέσω της διαδικασίας του rendering.

Κατά αυτό τον τρόπο παρουσιάζονται η αυτή καθαυτή 3Δ γεωμετρία του προσώπου (σχήμα, έκφραση), το ολοκληρωμένο 3Δ μοντέλο με τις συνιστώσες σχήματος, έκφρασης και υφής και η τελική ανακατασκευασμένη εικόνα, στην οποία αποτυπώνεται ο φωτισμός, η πόζα και οι εκφράσεις με τις οποίες το άτομο παρουσιάζεται στη αρχική πρωτότυπη εικόνα. Στη συνέχεια παρατίθενται ορισμένα δείγματα ανακατασκευής που προέκυψαν με χρήση του εκπαιδευμένου προτεινόμενου δικτύου, για κάθε ένα εκ των προαναφερθέντων datasets.

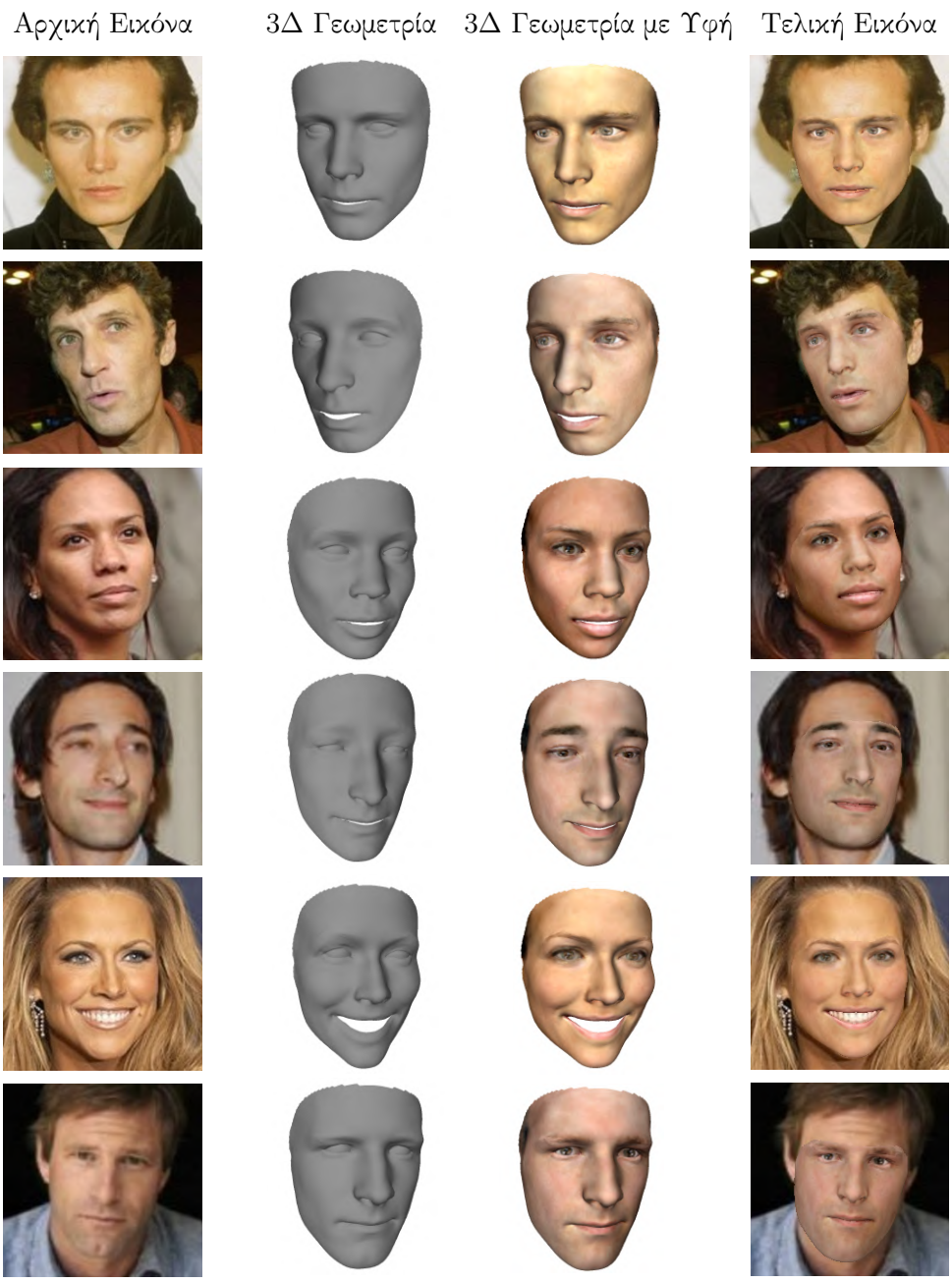
- ***CelebA Dataset***





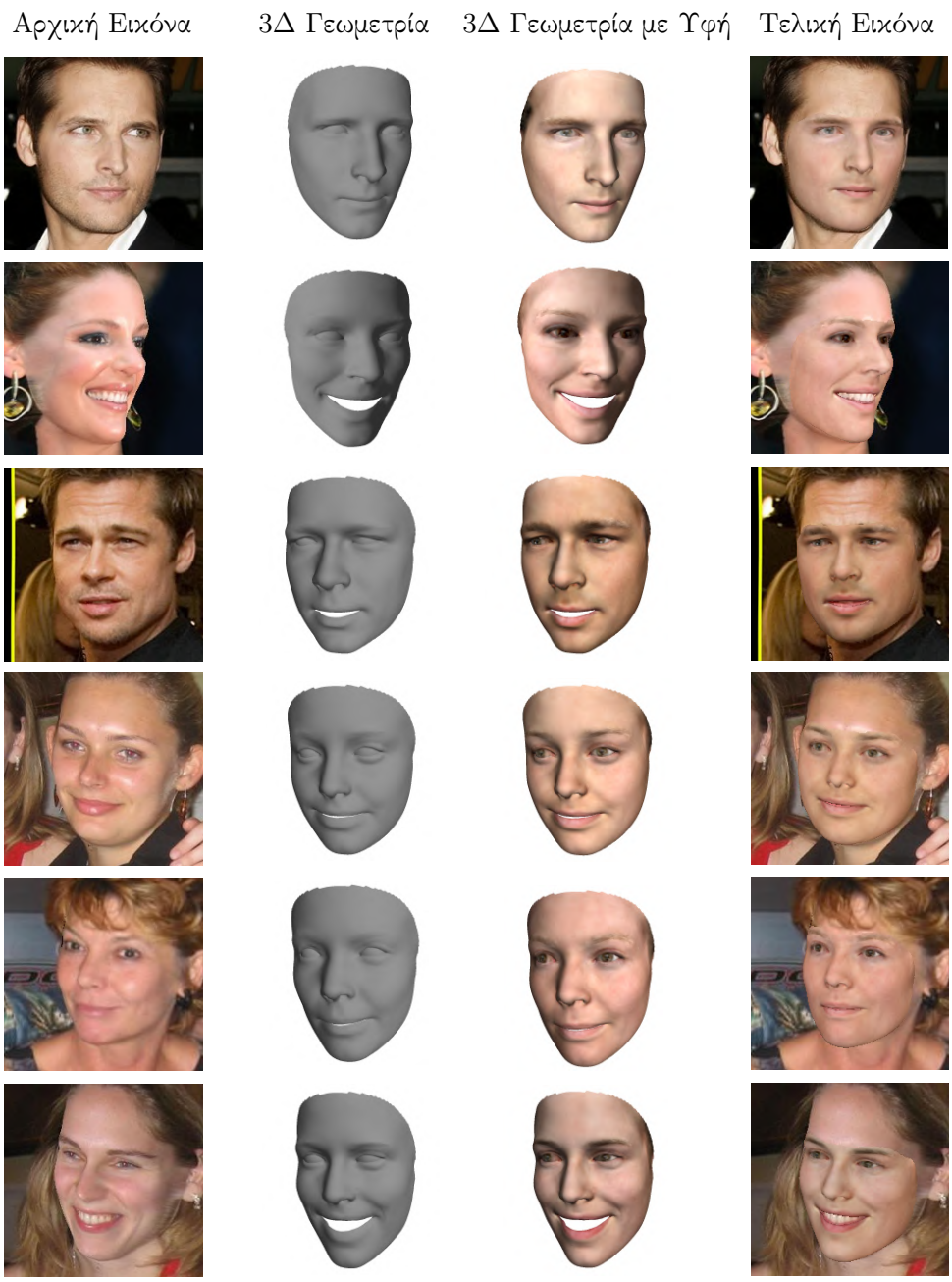
Σχήμα 4.2: Αποτελέσματα ανακατασκευής με χρήση εικόνων του CelebA dataset. Από αριστερά προς τα δεξιά: αρχική εικόνα εισόδου, 3Δ γεωμετρία, πλήρης 3Δ τοπολογία προσώπου, αρχική εικόνα έπειτα από αντικατάσταση της περιοχής του προσώπου με το ανακατασκευασμένο rendered πρόσωπο.

- ***LFW Dataset***



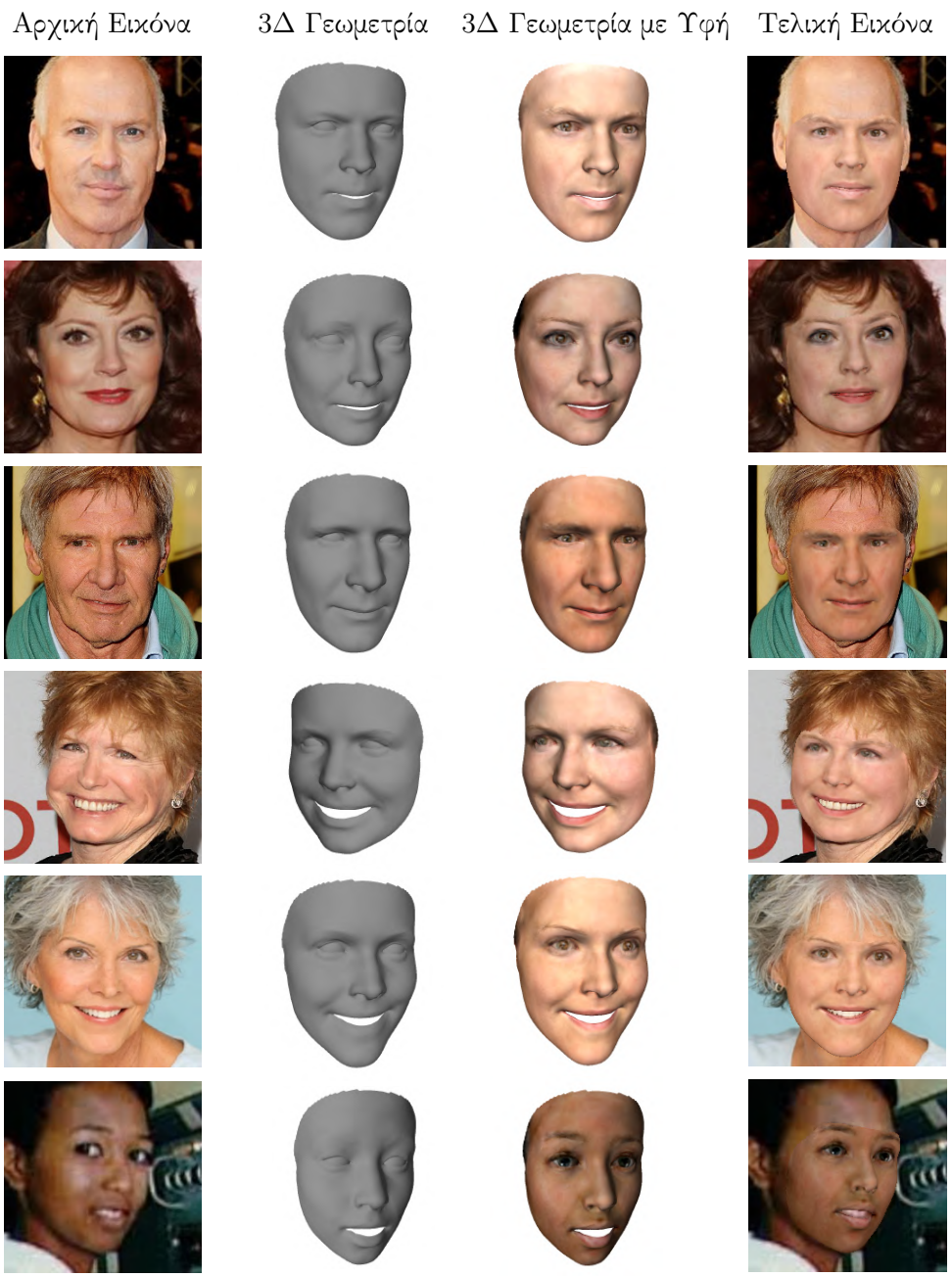
Σχήμα 4.3: Αποτελέσματα ανακατασκευής με χρήση εικόνων του LFW dataset.

- ***300W-LP Dataset***



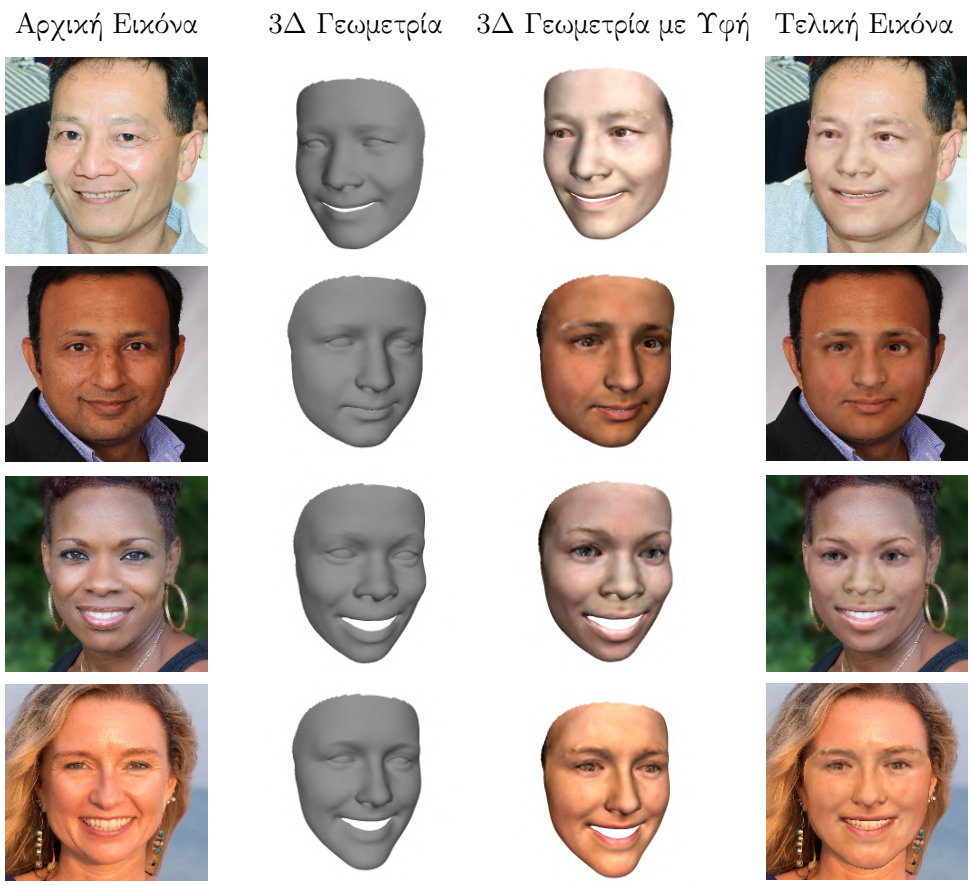
Σχήμα 4.4: Αποτελέσματα ανακατασκευής με χρήση εικόνων του 300W-LP.

- ***UTKFace Dataset***



Σχήμα 4.5: Αποτελέσματα ανακατασκευής με χρήση εικόνων του UTKFace.

- ***FFHQ Dataset***



Σχήμα 4.6: Αποτελέσματα ανακατασκευής με χρήση εικόνων του FFHQ dataset.

Τα αποτελέσματα που παρουσιάζονται στα Σχ.4.2 - 4.6 αποδεικνύουν ότι το προτεινόμενο δίκτυο είναι αρκετά ικανό στον προσδιορισμό της 3Δ τοπολογίας προσώπων από τις αντίστοιχες 2Δ εικόνες, παράγοντας ρεαλιστικά μοντέλα τα οποία αποτυπώνουν τα βασικά χαρακτηριστικά των ατόμων από τα οποία προέρχονται τα πρόσωπα αυτά.

Πιο συγκεκριμένα, με μια πρώτη εποπτική σύγκριση των αρχικών εικόνων και των ανακατασκευασμένων 3Δ γεωμετριών, συμπεραίνεται ότι το δίκτυο επιτυγχάνει το σχηματισμό της βασικής γεωμετρίας του εικονιζόμενου προσώπου, προσαρμόζοντας κάθε φορά κατάλληλα τη μορφή και το σχήμα του παραγόμενου μοντέλου. Κατά αυτό τον τρόπο, παράγονται μοντέλα προσώπων με τετρα-

γωνική, οβάλ, μακρόστενη και κυκλική γεωμετρία, τα οποία συμπίπτουν κατά το δυνατόν με την αντίστοιχη γεωμετρία των αρχικών προσώπων των εικόνων. Το γεγονός αυτό γίνεται έντονα αντιληπτό μέσω των ανακατασκευασμένων εικόνων, στις οποίες τα πρόσωπα των αρχικών εικόνων έχουν αποκοπεί και στη θέση τους έχουν τοποθετηθεί τα αντίστοιχα παραγόμενα 3Δ μοντέλα, αφού πρώτα αυτά έχουν προβληθεί στο 2Δ επίπεδο μέσω της διαδικασίας του rendering. Στις εικόνες αυτές, τα προβαλλόμενα 3Δ μοντέλα καταλαμβάνουν με μεγάλη ακρίβεια τη θέση των αρχικών πρωτότυπων προσώπων, διατηρώντας τη γεωμετρία, την οριοθέτηση και τον προσανατολισμό τους.

Η εξαγωγή της βασικής αυτής 3Δ γεωμετρίας των εικονιζόμενων προσώπων δεν πραγματοποιείται προφανώς με την ίδια αποτελεσματικότητα για κάθε μία εκ των εξεταζόμενων εικόνων και επηρεάζεται ευθέως και σε πολύ μεγάλο βαθμό από την τοποθέτηση των 2Δ προσώπων, από τυχόν αποκρύψεις περιοχών τους και από τις συνθήκες του περιβάλλοντος. Αυτό έχει ως αποτέλεσμα κάποιες γεωμετρίες να προσδιορίζονται με πολύ υψηλή ακρίβεια, όπως για παράδειγμα συμβαίνει στην περίπτωση των εικόνων του CelebA dataset, ενώ κάποιες άλλες να παρουσιάζουν μικρές αποκλίσεις σε σχέση με το αρχικό πρόσωπο, όπως οι τελευταίες εικόνες των datasets 300W-LP, UTKFace και LFW.

Η ικανότητα του δικτύου να προσδιορίζει τις 3Δ γεωμετρίες των προσώπων, όταν αυτά παρουσιάζονται σε έντονες τοποθετήσεις και συνθήκες περιβάλλοντος, είναι πολύ σημαντική και μελετάται ξεχωριστά σε επόμενη ενότητα. Σε κάθε περίπτωση, με μια πρώτη εποπτική αξιολόγηση, διαπιστώνεται ότι το δίκτυο προσδιορίζει με σχετικά μεγάλη ακρίβεια τη γεωμετρία της πλειοψηφίας των προσώπων των εξεταζόμενων εικόνων.

Το δεύτερο σημαντικό στοιχείο της ανακατασκευής, έπειτα από τον προσδιορισμό της 3Δ γεωμετρίας των προσώπων, είναι αυτό της αποτύπωσης των εκφράσεών τους. Ένα επιτυχημένο σύστημα ανακατασκευής θα πρέπει προφανώς να είναι σε θέση να αποδίδει όσο το δυνατόν καλύτερα τυχόν εκφραστικές διακυμάνσεις τις οποίες παρουσιάζουν τα εικονιζόμενα άτομα.

Παρατηρώντας λοιπόν και πάλι τα αποτελέσματα της ανακατασκευής, συμπεραίνεται ότι τα ανακατασκευασμένα 3Δ μοντέλα αποτυπώνουν σε αρκετά ικανοποιητικό βαθμό τις βασικές εκφράσεις των εικονιζόμενων ατόμων, κυρίως σε ό,τι αφορά στο γέλιο και στους μικρούς μορφασμούς. Έτσι, η βασική 3Δ γεωμετρία των μοντέλων εμπλουτίζεται και τροποποείται ελαφρώς, ώστε να αποτυπώνει τα δυναμικά χαρακτηριστικά εκείνα των προσώπων, τα οποία προκαλούνται λόγω των εκφράσεων, όπως για παράδειγμα οι έντονες γωνίες στα ζυγωματικά και στη γραμμή της κάτω σιαγόνας.

Παρ΄ όλη την αποτελεσματικότητα ωστόσο στην αποτύπωση των εκφράσεων, η ανακατασκευή υστερεί σε μια ιδιαίτερη κατηγορία χαρακτηριστικών του προσώπου, αυτή των ρυτίδων, κάτι το οποίο παρατηρείται σε αρκετά από τα ανωτέρω αποτελέσματα.

Οι ρυτίδες αποτελλούν μια πολύ σημαντική κατηγορία τοπικών λεπτομερειών του ανθρώπινου προσώπου και μεταβάλλονται με δυναμικό τρόπο ανάλογα με τις εκάστοτε εκφράσεις του. Η μεταβλητότητα αυτή σε συνδυασμό με τη δομική τους ανομοιομορφία, καθιστά τις ρυτίδες ως μία από τις πιο δύσκολες κατηγορίες χαρακτηριστικών για ανακατασκευή.

Στην περίπτωση του προτεινόμενου δικτύου, τα αποτελέσματα φανερώνουν την αδυναμία του να συλλάβει και να αποτυπώσει τις ρυτιδώσεις του δέρματος στις περιοχές του μετώπου, των ματιών και των ζυγωματικών. Εξαίρεση αποτελούν οι πολύ έντονες και εμφανείς ρυτίδες, κυρίως κατά μήκος της περιοχής μεταξύ των άνω ζυγωματικών και της κάτω σιαγόνας, οι οποίες σε αρκετές περιπτώσεις εντοπίζονται εν μέρει από το δίκτυο και παρουσιάζονται στο ανακατασκευασμένο μοντέλο, όπως για παράδειγμα στην έκτη και έβδομη κατά σειρά εικόνα του CelebaA dataset και στην δεύτερη και τέταρτη κατά σειρά εικόνα του UTKFace dataset. Στις περιπτώσεις αυτές παρουσιάζονται στα 3Δ μοντέλα οι πιο έντονες ρυτίδες των αρχικών προσώπων, χωρίς ωστόσο αυτές να διαθέτουν το περίπλοκο δομικό ανάγλυφο των αντίστοιχων αρχικών ρυτίδων.

Η αδυναμία αυτή του δικτύου να αποτυπώσει τις εκάστοτε ρυτιδώσεις του προσώπου μιας εικόνας, οφείλεται αφενός στο υψηλό επίπεδο λεπτομερειών που αυτές παρουσιάζουν και αφετέρου στη χαμηλή διαστατικότητα του στατιστικού μοντέλου αναπαράστασης 3DMM, το οποίο χρησιμοποιεί συγκεκριμένες συναρτήσεις βάσεις για τη μοντελοποίηση των δεδομένων εισόδου. Αυτή η καθολικότητα στη μοντελοποίηση διευκολύνει μεν τη συνολική διαδικασία υπολογιστικά και θεωρητικά, δυσχεραίνει ωστόσο τον εντοπισμό διαφόρων τοπικών λεπτομερειών της επιφάνειας του προσώπου, μεταξύ των οποίων και των ρυτιδών.

Στο πλαίσιο αυτό έχουν προταθεί διάφορες μέθοδοι για τον εντοπισμό και την ανακατασκευή γεωμετρικών λεπτομερειών υψηλών συχνοτήτων, οι οποίες βασίζονται στην ιδέα του Shape-from-Shading (SFS), η οποία αποσκοπεί στον προσδιορισμό του σχήματος και των λεπτομερειών του προσώπου μέσω παρατηρήσης των διαφόρων σκιάσεων και στο Inverse Rendering, το οποίο μοιάζει με το SFS, με τη διαφορά ότι αποσκοπεί στην εκτίμηση όλων των παραμέτρων που απαιτούνται για την παραγωγή ενός πλήρους 3Δ μοντέλου προσώπου και όχι μόνο στη γεωμετρία.

Συνολικά και παρά την αδυναμία του δικτύου ανακατασκευής να αποτυπώνει τις ρυτίδες και γενικά τις περίπλοκες τοπικές λεπτομέρειες ενός προσώπου, τα παραγόμενα 3Δ μοντέλα παραμένουν ρεαλιστικά και αποδίδουν σε αρκούντως ικανοποιητικό βαθμό τα ιδιαίτερα χαρακτηριστικά των εικονιζόμενων ατόμων, χωρίς να υπάρχει περίπτωση για σύγχυση της ταυτότητας δύο ατόμων λόγω της έλλειψης αυτής.

Τέλος, θα πρέπει να αξιολογηθεί το σημαντικότερο ίσως στοιχείο της διαδικασίας της ανακατασκευής, το οποίο δεν είναι άλλο από την υφή των παραγόμενων 3Δ μοντέλων. Προς την κατεύθυνση αυτή, τα άτομα που παρουσιάζονται στις εικόνες των Σχ.4.2 - 4.6 έχουν επιλεχθεί εσκεμμένα κατά τέτοιο τρόπο, ώστε να καλύπτουν μεγάλο εύρος του φάσματος των χρωματικών αποχρώσεων του ανθρώπινου δέρματος.

Μελετώντας λοιπόν τα πλήρη 3Δ ανακατασκευασμένα μοντέλα (τρίτη στήλη αποτελεσμάτων), τα οποία περιλαμβάνουν τις συνιστώσες της 3Δ γεωμετρίας (σχήμα, έκφραση) και της υφής και συγκρίνοντάς τα με τα αντίστοιχα πρωτότυπα πρόσωπα των 2Δ εικόνων, φαίνεται ότι το προτεινόμενο δίκτυο προσαρμοζεται με αρκετά αποτελεσματικό τρόπο στις ιδιαίτερες χρωματικές αποχρώσεις του δέρματος των προσώπων, καθώς αυτές μεταβάλλονται από ιδιαίτερα ανοιχτές σε ιδιαίτερα σκούρες.

Πιο αναλυτικά, οι μεσαίες δερματικές αποχρώσεις, δηλαδή οι αποχρώσεις που δεν είναι ούτε πολύ ανοιχτές ούτε πολύ σκούρες, εντοπίζονται και αποτυπώνονται καλύτερα από το δίκτυο, οδηγώντας σε 3Δ μοντέλα των οποίων η υφή προσεγγίζει σε πολύ υψηλό βαθμό την υφή των αντίστοιχων αρχικών 2Δ προσώπων, όπως για παράδειγμα παρατηρείται στην πρώτη, δεύτερη, τέταρτη και έβδομη κατά σειρά εικόνα του CelebA dataset, στην τέταρτη και έκτη εικόνα του LFW dataset, στην πρώτη, τρίτη τέταρτη και πέμπτη εικόνα του 300W-LP dataset και στην πρώτη, τρίτη, τέταρτη και πέμπτη εικόνα του UTKFace dataset. Η καλή αυτή αποτύπωση των μεσαίων χρωματικών αποχρώσεων, οφείλεται προφανώς στην ουδετερότητα που αυτές παρουσιάζουν και στη σχετικά περιορισμένη διακύμανση των τιμών των χρωμάτων τους, γεγονός το οποίο διευκολύνει τον προσδιορισμό των παραμέτρων υφής από το δίκτυο.

Όσον αφορά τώρα στις πιο σκούρες αποχρώσεις δέρματος, αυτές μπορεί να ταξινομηθούν σε δύο μεγάλες υποκατηγορίες: στις μεσαίες προς σκούρες (π.χ. έκτη, όγδοη και δέκατη εικόνα του CelebA dataset, τρίτη εικόνα LFW dataset, δεύτερη εικόνα FFHQ dataset) και στις καθαρώς σκούρες (π.χ. πέμπτη εικόνα CelebA dataset, έκτη εικόνα UTKFace dataset, τρίτη εικόνα FFHQ dataset). Από τις δύο υποκατηγορίες αυτές, η πρώτη αντιμετωπίζεται με αρκετά αποτελεσματικό τρόπο από το δίκτυο ανακατασκευής, με την υφή των παραγόμενων

μοντέλων να ανταποκρίνεται σε πολύ καλό βαθμό στην πραγματικότητα, ενώ η δεύτερη, παρόλο που προσεγγίζεται με εξίσου ικανοποιητικό τρόπο, παρουσιάζει κάποιες μικρές ανομοιομορφίες ως προς την υφή των παραγόμενων προσώπων, όπως για παράδειγμα φαίνεται στην τέταρτη εικόνα του FFHQ dataset.

Αυτή η μικρή πτώση της απόδοσης του δικτύου σε σκούρες αποχρώσεις δέρματος, οφείλεται αφενός στη μεγαλύτερη διακύμανση που παρουσιάζουν οι αντίστοιχες τιμές των χρωμάτων τους και αφετέρου σε έναν παράγοντα ο οποίος αφορά στη διαδικασία της εκπαίδευσης. Όπως αναφέρθηκε και στην ενότητα 3.3.2, για τον υπολογισμό της φωτομετρικής συνάρτησης απώλειας, η οποία χρησιμοποιείται για την εκπαίδευση του δικτύου και η οποία εξετάζει το μέρος της ανακατασκευής που σχετίζεται με την υφή των προσώπων, αξιοποιείται μια μάσκα δέρματος (Σχ.3.7), η οποία εκφράζει την πιθανότητα ενός pixel προσώπου να αντιστοιχεί σε pixel δέρματος. Στις ιδιαίτερα σκούρες δερματικές αποχρώσεις οι RGB τιμές κάποιων δερματικών pixels μπορεί λανθασμένα να συγχυστούν με τιμές άλλων στοιχείων του προσώπου (σκούρα μαλλιά, μούσια, γυαλιά), με αποτέλεσμα να τους αποδοθεί μικρή τιμή πιθανότητας να αντιστοιχούν σε pixels δέρματος. Αυτό έχει ως αποτέλεσμα να μη συμμετέχουν με την ίδια βαρύτητα στον υπολογισμό της συνάρτησης απώλειας και ως εκ τούτου να τους αποδίδεται ελαφρώς διαφορετική χρωματική τιμή από αυτή των γειτονικών τους pixels.

Στο σημείο αυτό, αξίζει να σημειωθεί ότι η ποιοτική αξιολόγηση της απόδοσης του δικτύου ανακατασκευής ως προς τη συνιστώσα της υφής δεν είναι μια διαδικασία απλή και μονομερής, καθώς η χρωματική απόδοση των παραγόμενων 3Δ μοντέλων επηρεάζεται σε πολύ μεγάλο βαθμό από την πόζα του προσώπου και τις συνθήκες φωτισμού του περιβάλλοντος στο οποίο έγινε η λήψη της αρχικής εικόνας. Έτσι, ακόμα και στην περίπτωση των μεσαίων δερματικών αποχρώσεων, τα αποτελέσματα της ανακατασκευής μπορεί να μην είναι τα αναμενόμενα, εάν οι συνθήκες φωτισμού του εικονιζόμενου προσώπου είναι κακές, ή σε περίπτωση που υπάρχουν αποκρύψεις της επιφάνειάς του, λόγω αντικειμένων ή έντονου makeup. Χαρακτηριστικό παράδειγμα αποτελεί η πέμπτη κατά σειρά εικόνα του LFW dataset, η έκτη εικόνα του 300W-LP dataset και η τέταρτη εικόνα του FFHQ dataset, όπου το έντονο makeup, η μεγάλη στροφή και ο έντονος φωτισμός των προσώπων αντίστοιχα, έχουν επηρεάσει την απόδοση της υφής τους στα 3Δ μοντέλα.

Συνοψίζοντας, μέσω της εποπτικής παρατήρησης των αποτελεσμάτων της ανακατασκευής, συμπεραίνεται ότι το προτεινόμενο δίκτυο ανταπεξέρχεται αρκετά ικανοποιητικά στις εκάστοτε ιδιαιτερότητες των εικονιζόμενων προσώπων και τις συνθήκες του περιβάλλοντός τους, παράγοντας 3Δ τοπολογίες προσώ-

πων, των οποίων οι συνιστώσες σχήματος, έκφρασης και υφής αποδίδουν με ρεαλισμό και σχετική ακρίβεια την ταυτότητα των εικονιζόμενων ατόμων. Παρά ταύτα, οι 3Δ αυτές τοπολογίες αδυνατούν να αποτυπώσουν κάποιες τοπικές λεπτομέρειες υψηλών συχνοτήτων, όπως οι ρυτίδες. Η απόδοση του δικτύου επηρεάζεται άμεσα από την πόζα, την έκφραση και τη δερματική απόχρωση των εικονιζόμενων προσώπων, καθώς και από τις συνθήκες του περιβάλλοντός τους, κάτι το οποίο αναλύεται πιο διεξοδικά στην επόμενη ενότητα.

Ειδικές Περιπτώσεις Ανακατασκευής

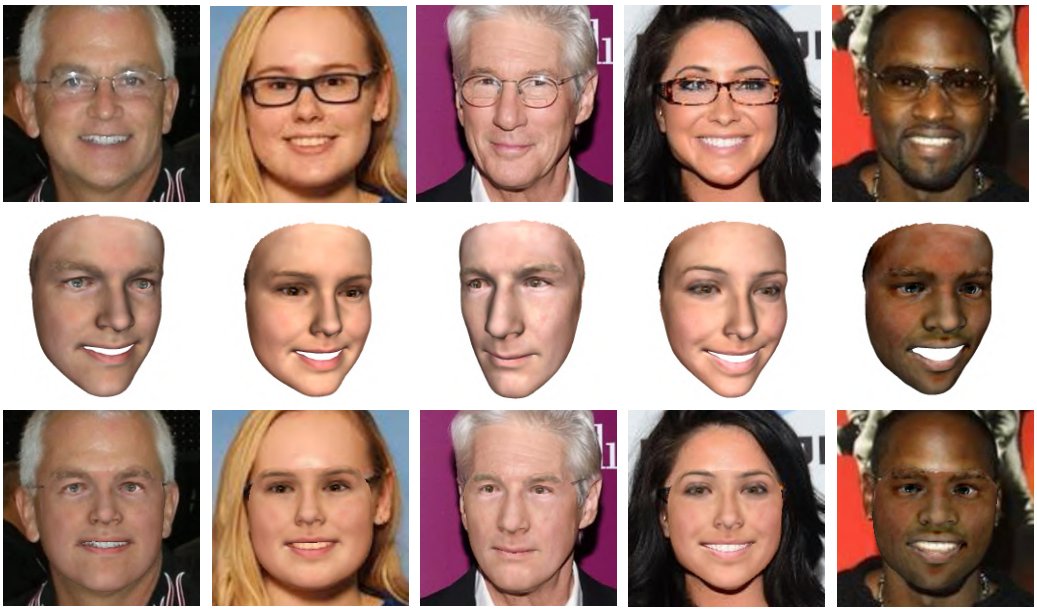
Τα αποτελέσματα της προηγούμενης ενότητας επιβεβαιώνουν την ικανότητα του προτεινόμενου δικτύου να προσδιορίζει σε επαρκώς ικανοποιητικό βαθμό την 3Δ τοπολογία προσώπων, βασιζόμενο αποκλειστικά στη χρήση των αντίστοιχων 2Δ εικόνων στις οποίες αυτά παρουσιάζονται. Τα αποτελέσματα αυτά ωστόσο αφορούν κυρίως σε συμβατικές περιπτώσεις, όπου τα εικονιζόμενα άτομα, αν και παρουσιάζουν μεταβολές ως προς τις δερματικές αποχρώσεις και τις εκφράσεις τους, εμφανίζονται σε σχετικά μέτριες πόζες, δηλαδή σε πόζες εντός ενός μικρού εύρους γωνιών, χωρίς έντονες αποκρύψεις περιοχών του προσώπου τους, ενώ οι συνθήκες του περιβάλλοντος λήψης των αντίστοιχων εικόνων είναι εν γένει ευνοϊκές και καλές.

Προκειμένου η αξιολόγηση της απόδοσης του δικτύου να είναι αμερόληπτη και πλήρης, θα πρέπει να εξεταστούν και κάποιες ειδικές κατηγορίες ανακατασκευής, στις οποίες η όλη διαδικασία επηρεάζεται και δυσχεραίνεται από διάφορα φαινόμενα, τα οποία αφορούν είτε στον τρόπο με τον οποίο παρουσιάζονται τα εικονιζόμενα άτομα, είτε στις περιβαλλοντικές συνθήκες στις οποίες γίνεται η λήψη των αντίστοιχων εικόνων. Κατά αυτό τον τρόπο θα μελετηθεί η σθεναρότητα του δικτύου, η οποία είναι απαραίτητη για την αποδοτική λειτουργία του.

Προς την κατεύθυνση αυτή, στην παρούσα ενότητα παρουσιάζονται δείγματα ανακατασκευής ταξινομημένα σε επιμέρους ενότητες, με την κάθε μία εξ αυτών να εξετάζει μία εκ των ακόλουθων ειδικών περιπτώσεων ανακατασκευής:

- Αποκρύψεις λόγω γυαλιών
- Έντονες εκφράσεις
- Κακές συνθήκες φωτισμού
- Ακραίες πόζες

- **Αποκρύψεις λόγω γυαλιών**



Σχήμα 4.7: Αποτελέσματα ανακατασκευής προσώπων με αποκρύψεις λόγω γυαλιών. Από πάνω προς τα κάτω: αρχικές εικόνες, ανακατασκευασμένες 3Δ τοπολογίες προσώπων, αρχικές εικόνες έπειτα από αντικατάσταση της περιοχής των προσώπων με τα ανακατασκευασμένα rendered πρόσωπα.

Από τα αποτελέσματα του Σχ.4.7 φαίνεται ότι ακόμα και στην περίπτωση αποκρύψεων λόγω της ύπαρξης γυαλιών, το δίκτυο επιτυγχάνει να προσδιορίσει την 3Δ γεωμετρία των εικονιζόμενων προσώπων, με τις περιοχές των ματιών στις οποίες παρατηρείται και το μεγαλύτερο μέρος των αποκρύψεων, να ανακατασκευάζονται ακολουθώντας τη μορφολογία και τη χρωματική απόχρωση των γειτονικών περιοχών του προσώπου, χωρίς να παρατηρείται κάποια έντονη ανομοιομορφία.

Παρ΄ όλα αυτά, ειδικά στην περίπτωση των γυαλιών ηλίου της τελευταίας εικόνας, το δίκτυο αδυνατεί να προσδιορίσει την ακριβή θέση των οφθαλμών του ατόμου, με αποτέλεσμα αυτοί να εμφανίζουν μια στρέψη προς το τμήμα της μύτης, η οποία προφανώς δεν υπάρχει στην αρχική εικόνα. Επιπλέον, σε όλα τα αποτελέσματα παρατηρείται μια λανθασμένη ομοιογένεια ως προς τη δερματική απόχρωση κατά μήκος των περιοχών που καλύπτονται από το σκελετό των γυαλιών, γεγονός αναμενόμενο καθώς το δίκτυο δεν μπορεί να αποκομίσει κάποια

άμεση πληροφορία για την υφή των περιοχών αυτών, με αποτέλεσμα να χρησιμοποιεί παρόμοιες χρωματικές τιμές με αυτές των γειτονικών περιοχών.

Σε κάθε περίπτωση, τα αποτελέσματα της ανακατασκευής είναι αρκετά ικανοποιητικά, δεδομένης της ύπαρξης αποκρύψεων, κάτι το οποίο οφείλεται και στη χρήση της μάσκας δέρματος (Σχ.3.7) που χρησιμοποιήθηκε κατά την εκπαίδευση του δικτύου και η οποία αναλύθηκε στην ενότητα 3.1.3. Το είδος των γυαλιών επηρεάζει άμεσα τη διαδικασία της ανακατασκευής, με τα πιο σκούρα γυαλιά να προκαλούν μεγαλύτερες αποκρύψεις και ως εκ τούτου πτώση της απόδοσης του δικτύου.

- **Έντονες εκφράσεις**

Σχήμα 4.8: Αποτελέσματα ανακατασκευής προσώπων με έντονες εκφράσεις.

Και στην περίπτωση των έντονων εκφράσεων λοιπόν, τα αποτελέσματα του Σχ.4.8 αποδεικνύουν ότι το προτεινόμενο δίκτυο ανταπεξέρχεται με εξίσου αποτελεσματικό τρόπο, αποτυπώνοντας στα παραγόμενα μοντέλα τις εκφράσεις των αρχικών 2Δ προσώπων.

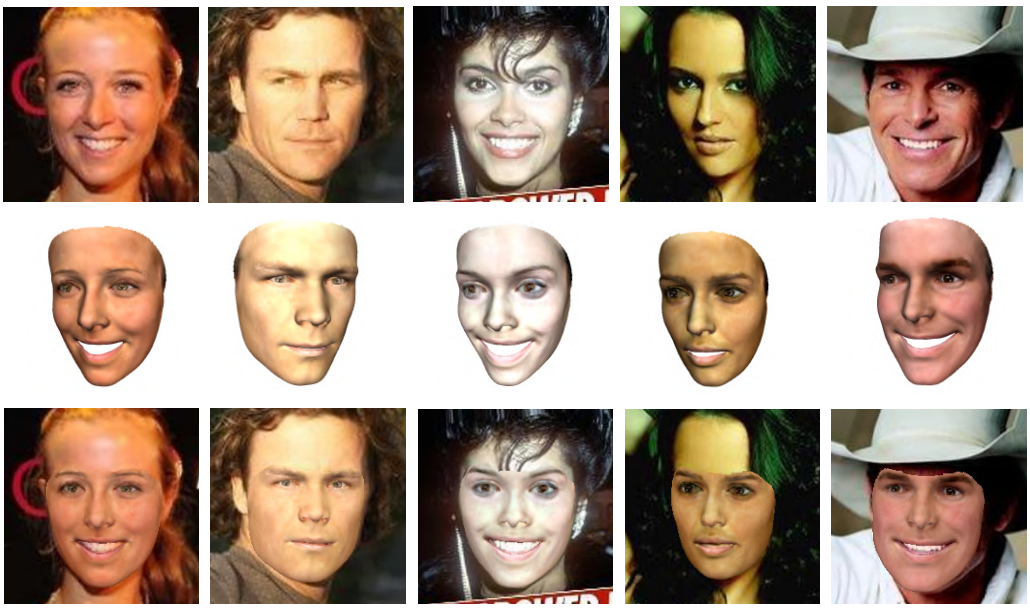
Η απόδοση ωστόσο των εκφράσεων αυτών πραγματοποιείται έως ένα βαθμό εις βάρος του σχήματος των εικονιζόμενων προσώπων. Έτσι, καθώς αυξάνεται

η ένταση των εκφράσεων, το δίκτυο παράγει μοντέλα των οποίων το γεωμετρικό σχήμα παρουσιάζει μια σχετική απόκλιση από το αντίστοιχο σχήμα των πρωτότυπων προσώπων. Η σταδιακή αυτή αύξηση της απόκλισης μεταξύ του σχήματος του αρχικού και του ανακατασκευασμένου προσώπου γίνεται εμφανής στις τρεις τελευταίες εικόνες του Σχ.4.8, όπου τα προβαλλόμενα στις αρχικές εικόνες 3Δ μοντέλα αποκλίνουν γεωμετρικά από τα αρχικά 2Δ πρόσωπα.

Η ικανότητα του δικτύου να αντιμετωπίζει έντονες εκφράσεις εξαρτάται σε μεγάλο βαθμό τόσο από τη συνάρτηση απώλειας που χρησιμοποιείται κατά την εκπαίδευσή του, όσο και από την εκφραστική ποικιλία που προσφέρει το στατιστικό μοντέλο που αξιοποιείται στην Εξ.2.7 και το οποίο, στα πλαίσια της παρούσας άσκησης, κατασκευάζεται από το FaceWarehouse dataset.

Σε γενικές γραμμές, διαπιστώνεται ότι υπάρχει ένας συμβιβασμός μεταξύ της ακριβούς απόδοσης των εκφράσεων και του σχήματος στα παραγόμενα μοντέλα, έτσι ώστε να διατηρείται η ισορροπία και ο ρεαλισμός στα ανακατασκευασμένα πρόσωπα.

- **Κακές συνθήκες φωτισμού**



Σχήμα 4.9: Αποτελέσματα ανακατασκευής σε κακές συνθήκες φωτισμού.

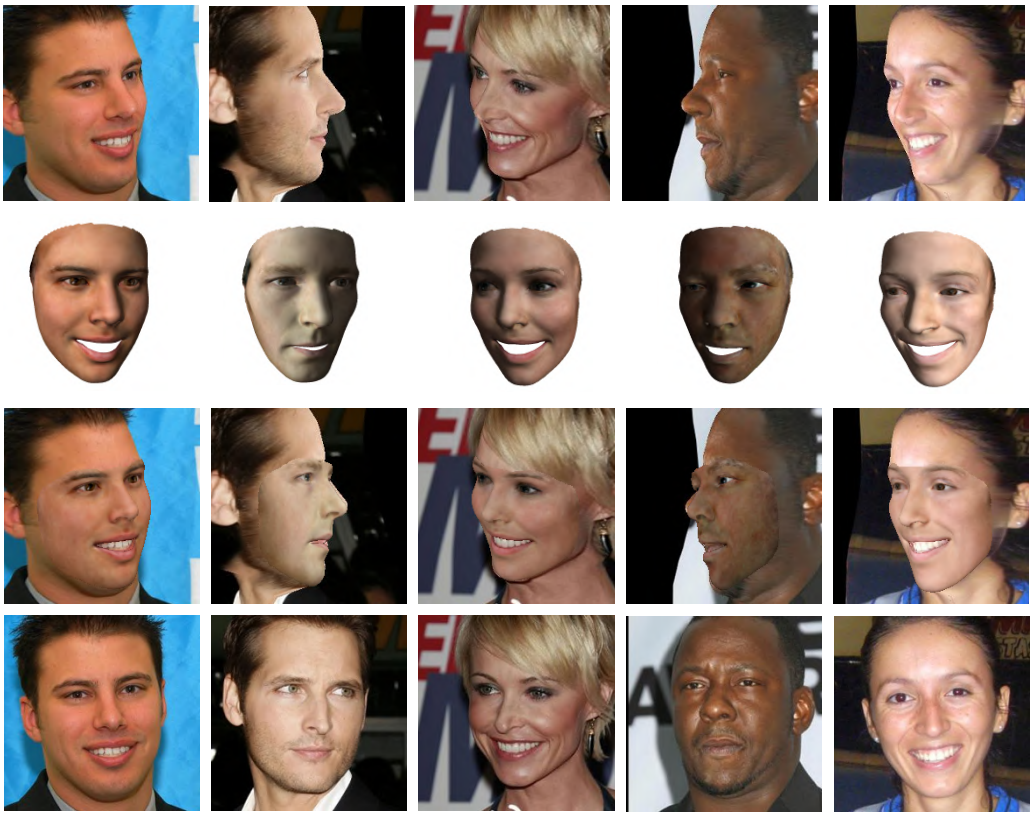
Σε αντίθεση τώρα με τις προηγούμενες περιπτώσεις, οι κακές συνθήκες φωτισμού των εικονιζόμενων προσώπων φαίνεται να επηρεάζουν σε πολύ μεγάλο βαθμό τα αποτελέσματα της ανακατασκευής. Έτσι, παρατηρώντας τα δείγματα του Σχ.4.9 διαπιστώνεται ότι ο κακός φωτισμός, είτε αυτός αφορά σε ακραίες αποχρώσεις του φωτός, είτε σε ανομοιόμορφη κάλυψη της περιοχής του προσώπου οδηγεί σε παραγόμενα μοντέλα των οποίων η υφή παρουσιάζει απόκλιση από την υφή των αντίστοιχων αρχικών προσώπων.

Πιο συγκεκριμένα, στις δύο πρώτες εικόνες όπου και επικρατούν συνθήκες έντονου θερμού φωτός, το δίκτυο επιτυγχάνει να αποτυπώσει έως ένα βαθμό τις βασικές αποχρώσεις του φωτός στα 3Δ πρόσωπα, αδυνατεί ωστόσο να συλλάβει τις έντονες ανακλάσεις και σκιάσεις που παρατηρούνται στην επιφάνεια του δέρματος των αρχικών προσώπων. Πιο αποτελεσματική φαίνεται να είναι η ανακατασκευή στις συνθήκες ψυχρού φωτισμού της τρίτης εικόνας, με το ανακατασκευασμένο μοντέλο να παρουσιάζει μεγαλύτερη ομοιότητα ως προς την υφή με το εικονιζόμενο πρόσωπο. Τέλος, στην τέταρτη και πέμπτη εικόνα όπου υπάρχει απόκρυψη περιοχών του προσώπου λόγω έλλειψης φωτισμού, τα παραγόμενα 3Δ μοντέλα παρουσιάζουν τη μεγαλύτερη απόκλιση από τα αρχικά πρόσωπα, τόσο ως προς την υφή όσο και ως προς το σχήμα τους, γεγονός το οποίο οφείλεται στην αδυναμία του δικτύου να αποκτήσει πληροφορία για τη μορφολογία και την υφή των σκοτεινών περιοχών τους.

Προκειμένου να βελτιωθεί η ικανότητα του δικτύου να συλλαμβάνει έντονες και μη ευνοϊκές συνθήκες φωτισμού και δεδομένου ότι ο φωτισμός του περιβάλλοντος μοντελοποιείται βάσει των σφαιρικών αρμονικών συναρτήσεων (ενότητα 3.2.2), θα μπορούσε να χρησιμοποιηθεί μεγαλύτερος αριθμός συντελεστών $\boldsymbol{\gamma}_b$ και κατά συνέπεια μεγαλύτερος αριθμός αρμονικών συναρτήσεων βάσης $\boldsymbol{H}_b$, έτσι ώστε να προσεγγίζεται με μεγαλύτερη ακρίβεια η ακτινοβολία των σημείων της επιφάνειας των 3Δ προσώπων. Η επέκταση αυτή ωστόσο δεν εγγυάται την αισθητή βελτίωση των αποτελεσμάτων της ανακατασκευής, καθώς η συνάρτηση έντασης ακτινοβολίας, η οποία μοντελοποιείται βάσει των συναρτήσεων αυτών, είναι ιδιαιτέρως σύνθετη, ειδικά σε περιπτώσεις έντονου ή κακού φωτισμού.

Συνολικά, συμπεραίνεται ότι οι κακές συνθήκες φωτισμού έχουν άμεση επίδραση στην απόδοση του δικτύου ανακατασκευής και στην ποιότητα των παραγόμενων 3Δ προσώπων. Το δίκτυο αποδίδει έως ένα βαθμό τον έντονο θερμό και ψυχρό φωτισμό, εις βάρος βέβαια της υφής των 3Δ μοντέλων, ενώ αδυνατεί να αντιμετωπίσει με τον ίδιο τρόπο περιπτώσεις όπου υπάρχουν αποκρύψεις περιοχών του προσώπου λόγω ελλιπούς φωτισμού.

- **Έντονες πόζες**



Σχήμα 4.10: Αποτελέσματα ανακατασκευής προσώπων σε έντονες πόζες. Από πάνω προς τα κάτω: αρχικές εικόνες, ανακατασκευασμένες 3Δ τοπολογίες προσώπων, αρχικές εικόνες έπειτα από αντικατάσταση της περιοχής των προσώπων με τα ανακατασκευασμένα rendered πρόσωπα, εικόνες των ίδιων ατόμων σε φυσιολογικές πόζες.

Όπως και στην περίπτωση των κακών συνθηκών φωτισμού, τα αποτελέσματα του Σχ.4.8 υποδεικνύουν ότι οι έντονες πόζες των εικονιζόμενων προσώπων επηρεάζουν σημαντικά την απόδοση του δικτύου, τόσο ως προς τη γεωμετρία όσο και ως προς την υφή.

Στις περιπτώσεις όπου τα πρόσωπα των εικονιζόμενων ατόμων παρουσιάζονται μετρίως στραμμένα αριστερά ή δεξιά, όπως για παράδειγμα στην πρώτη και την τρίτη εικόνα, τα αποτελέσματα της ανακατασκευής είναι αρκετά ικανοποιητικά, με τη γεωμετρία και την υφή των παραγόμενων 3Δ μοντέλων να προσεγγίζουν την αντίστοιχη γεωμετρία και υφή των αρχικών προσώπων.

Καθώς όμως η γωνία στροφής των εικονιζόμενων προσώπων αυξάνεται και το μεγαλύτερο μέρος της εικόνας καταλαμβάνεται από το προφίλ τους, η ανακατασκευή δυσχεραίνεται όλο και περισσότερο, μιας και χάνεται σημαντικό μέρος της ωφέλιμης πληροφορίας την οποία το δίκτυο μπορεί να εξάγει από αυτά. Αυτό έχει ως αποτέλεσμα να μην γίνεται αποτελεσματικά η εξαγωγή των γεωμετρικών και χρωματικών πληροφοριών και να παράγονται κατά αυτό τον τρόπο 3Δ μοντέλα τα οποία δεν είναι ιδιατέρως ακριβή.

Στο πλαίσιο αυτό, η ανακατασκευή του προσώπου της τελευταίας εικόνας, στην οποία το άτομο παρουσιάζεται σε σχετικά μεγάλη στροφή, επιτυγχάνει να προσδιορίσει σε σημαντικό βαθμό την 3Δ γεωμετρία και τα βασικά χαρακτηριστικά του, υστερώντας ωστόσο στη συνιστώσα της υφής, η οποία δεν συλλαμβάνει εξίσου αποτελεσματικά τις χρωματικές αποχρώσεις της κεντρικής περιοχής του προσώπου του.

Ακόμα πιο ανακριβή είναι τα αποτελέσματα ανακατασκευής των προσώπων της δεύτερης και της τέταρτης εικόνας, τα οποία παρουσιάζονται σε πλήρες προφίλ (γωνία σχεδόν 90° ως προς το νοητό κατακόρυφο άξονα της εικόνας). Στις περιπτώσεις αυτές, το δίκτυο αδυνατεί να προσδιορίσει τόσο τη γεωμετρία όσο και την υφή τους, με τα ανακατασκευασμένα μοντέλα να παρουσιάζουν λανθασμένα ένα σχετικά μακρόστενο οβάλ σχήμα και μια ελαφρώς τροποποιημένη υφή. Επιπλέον, οι περιοχές των αρχικών προσώπων οι οποίες δεν είναι ορατές λόγω της τοποθέτησής τους, αποδίδονται με ιδιαίτερα σκούρες αποχρώσεις στα αντίστοιχα 3Δ μοντέλα, με αποτέλεσμα να μην είναι ευδιάκριτες, γεγονός βέβαια το οποίο είναι αναμενόμενο, καθώς για τις περιοχές αυτές, το δίκτυο δεν έχει κάποιο τρόπο να εξάγει πληροφορίες σχετικά με τη μορφή και την υφή τους.

Συνοψίζοντας, το προτεινόμενο δίκτυο φαίνεται να ανταποκρίνεται έως ένα βαθμό και στην περίπτωση της ανακατασκευής προσώπων σε έντονες πόζες, με την απόδοσή του ωστόσο να επηρεάζεται άμεσα από την στροφή την οποία αυτά παρουσιάζουν. Μεγάλες στροφές των εικονιζόμενων προσώπων προκαλούν πιο εκτεταμένες αποκρύψεις περιοχών τους, ελαττώνοντας τη συνολική ωφέλιμη πληροφορία την οποία μπορεί να αξιοποιήσει το δίκτυο και οδηγώντας κατά αυτό τον τρόπο σε μοντέλα με αποκλίνουσα γεωμετρία και υφή.

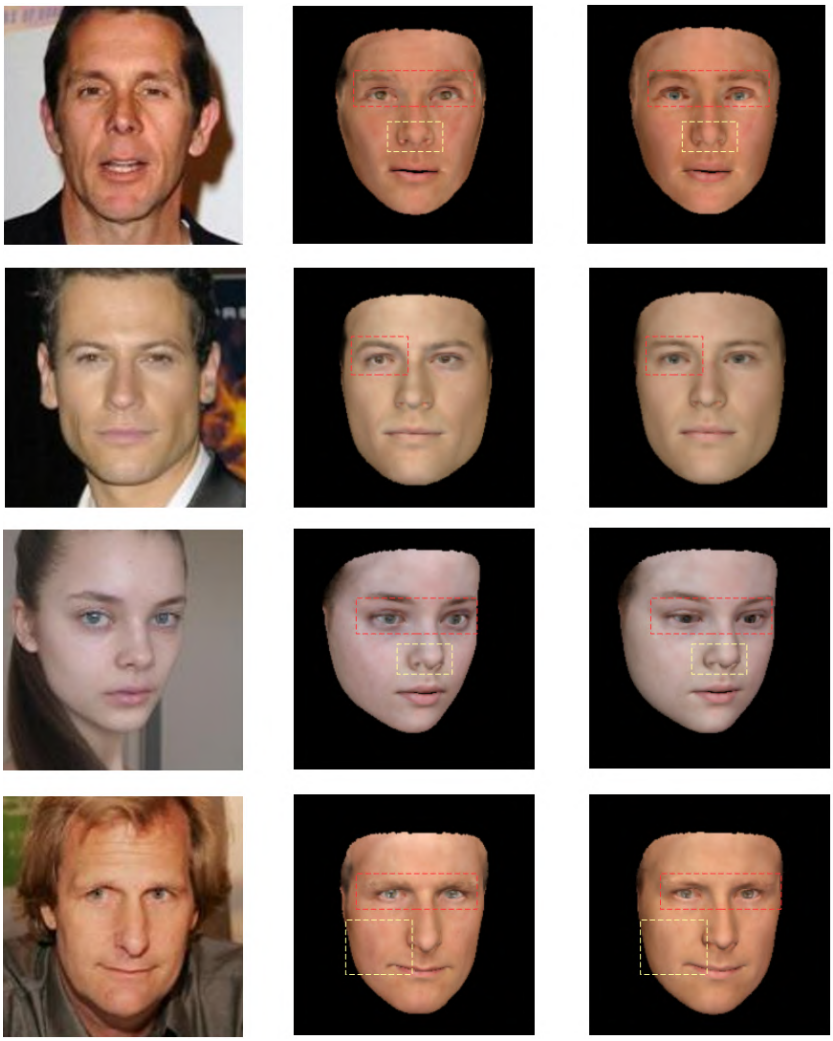
Επίδραση Συνάρτησης Απώλειας Λεπτομερειών

Προκειμένου η διαδικασία της ανακατασκευής να είναι αποτελεσματική και ρεαλιστική, θα πρέπει, όπως ήδη έχει αναφερθεί, τα παραγόμενα 3Δ μοντέλα να προσεγγίζουν όσο το δυνατόν περισσότερο τη γεωμετρία και την υφή των αρχικών εικονιζόμενων ατόμων στα οποία αντιστοιχούν. Η ακριβής ωστόσο απόδοση του σχήματος και των δερματικών αποχρώσεων των εικονιζόμενων ατόμων δεν αποτελεί αυτοσκοπό της όλης διαδικασίας, καθώς ακόμα και στην ιδανική περίπτωση όπου οι δύο συνιστώσες αυτές αποδοθούν με σχεδόν τέλειο τρόπο, οι ανακατασκευασμένες 3Δ τοπολογίες θα εξακολουθούν να διαφέρουν από τα αρχικά 2Δ πρόσωπα και αυτό γιατί δεν θα αποτυπώνουν τα ιδιαίτερα χαρακτηριστικά εκείνα των ατόμων, τα οποία δημιουργούν την έννοια της ταυτότητας.

Για να αντιμετωπιστεί το πρόβλημα αυτό, στα πλαίσια της εργασίας χρησιμοποιείται κατά την εκπαίδευση του δικτύου η συνάρτηση απώλειας λεπτομερειών L_{detail}, η οποία και αναλύθηκε στην ενότητα 3.3.2. Η συνάρτηση αυτή συνίσταται ουσιαστικά στον υπολογισμό της απόστασης συνημιτόνου μεταξύ των 128 χαρακτηριστικών του αρχικού και του ανακατασκευασμένου προσώπου, τα οποία υπολογίζονται μέσω του ανιχνευτή προσώπων FaceNet.

Προκειμένου λοιπόν να γίνει εμφανής η σημασία και η συνεισφορά της συνάρτησης απώλειας λεπτομερειών στην αποτελεσματικότητα της ανακατασκευής, στην παρούσα ενότητα μελετάται ποιοτικά η απόδοση του προτεινόμενου δικτύου, όταν αυτό εκπαιδεύεται με και χωρίς την εν λόγω συνάρτηση. Προς την κατεύθυνση αυτή, στο Σχ.4.11 παρουσιάζονται ορισμένα αποτελέσματα ανακατασκευής, στα οποία το ίδιο πρόσωπο ανακατασκευάζεται αρχικά με και έπειτα χωρίς χρήση της συνάρτησης L_{detail}. Οι κύριες διαφορές μεταξύ των δύο ανακατασκευών σημειώνονται με έγχρωμα πλαίσια.

Από τα αποτελέσματα φαίνεται ότι η ανακατασκευή χωρίς χρήση της συνάρτησης απώλειας L_{detail} οδηγεί σε 3Δ μοντέλα, των οποίων η επιφάνεια παρουσιάζει μια έντονη ομοιογένια, χωρίς να αποτυπώνει τις ιδιαιτερότητες των εικονιζόμενων ατόμων. Ιδιαίτερο πρόβλημα στην περίπτωση αυτή παρατηρείται στην απόδοση του χρώματος και της θέση των ματιών, τα οποία στα περισσότερα εκ των αποτελεσμάτων παρουσιάζονται με λανθασμένη χρωματική απόχρωση ίριδας και σε λάθος τοποθέτηση. Επιπλέον, παραλείπονται και δεν εντοπίζονται μικρές κινήσεις μερών του προσώπου, όπως για παράδειγμα μια ελαφριά ανύψωση των φρυδιών, ενώ οι δερματικές αποχρώσεις δεν προσαρμόζονται το ίδιο καλά στις εκάστοτε συνθήκες φωτισμού.



Σχήμα 4.11: Αποτελέσματα ανακατασκευής με και χωρίς χρήση της συνάρτησης απώλειας λεπτομερειών. Από αριστερά προς τα δεξιά: αρχική εικόνα, ανακατασκευή με χρήση της L_{detail}, ανακατασκευή χωρίς χρήση της L_{detail}.

Συμπεραίνεται λοιπόν, ότι η χρήση της συνάρτησης L_{detail} κατά την εκπαίδευση του δικτύου βελτιώνει αισθητά τα αποτελέσματα της ανακατασκευής, κυρίως σε ό,τι αφορά στην απόδοση των ιδιαίτερων χαρακτηριστικών της ταυτότητας των εικονιζόμενων ατόμων. Η παράλειψή της οδηγεί σε 3Δ μοντέλα ακριβή μεν ως προς το σχήμα και την υφή, τα οποία δε αδυνατούν να αποτυπώσουν την ιδιαιτερότητα των προσώπων των ατόμων αυτών.

4.2.2 Ποσοτική Αξιολόγηση

Σε αντίθεση με την περίπτωση της ποιοτικής αξιολόγησης, η ποσοτική αξιολόγηση των αποτελεσμάτων της ανακατασκευής δεν βασίζεται στη διαίσθηση, αλλά στη μαθηματική μοντελοποίηση του σφάλματος μεταξύ του παραγόμενου 3Δ μοντέλου και του αντίστοιχου προσώπου της 2Δ εικόνας.

Ιδανικά, το σφάλμα ανακατασκευής θα έπρεπε να υπολογιστεί μεταξύ της 3Δ παραγόμενης τοπολογίας και του αντίστοιχου 3Δ face scan του προσώπου της 2Δ εικόνας. Λόγω όμως της περιορισμένης ύπαρξης ζευγαριών 2Δ εικόνων και αντίστοιχων 3Δ face scans, το σφάλμα ανακατασκευής θα υπολογιστεί μεταξύ της αρχικής 2Δ εικόνας και της αντίστοιχης συνθετικής εικόνας που προκύπτει έπειτα από αντικατάσταση της περιοχής του προσώπου με το αντίστοιχο rendered παραγόμενο 3Δ μοντέλο (Σχ.4.12). Για την ποσοτικοποίηση του σφάλματος αυτού, μπορούν να χρησιμοποιηθούν αρκετές μετρικές, κάθε μία από τις οποίες εστιάζει σε διαφορετικες πτυχές της διαφοράς της αρχικής και της ανακατασκευασμένης συνθετικής εικόνας.

Στα πλαίσια της εργασίας, χρησιμοποιούνται οι εξής μετρικές:

- Μέσο Τετραγωνικό Σφάλμα (Mean Square Error, MSE)
- L_1-απόσταση (L_1-distance)
- Peak Signal-to-Noise Ratio (PSNR)
- Δείκτης Δομικής Ομοιότητας (Structural Similarity Index Measure)
- Ομοιότητα Συνημιτόνου (Cosine Similarity)

Μέσο Τετραγωνικό Σφάλμα

Το μέσο τετραγωνικό σφάλμα (MSE) αποτελεί μία από τις πιο απλές και διαδεδομένες μετρικές για τη σύγκριση της πιστότητας μεταξύ μιας αρχικής εικόνας αναφοράς και μιας προσέγγισής της και υπολογίζεται λαμβάνοντας τη μέση τιμή του τετραγώνου της διαφοράς των εντάσεων των pixels των δύο εικόνων.

Έτσι, δοθείσης μιας αρχικής εικόνας $I(i, j)$ και μιας εικόνας προσέγγισης $K(i, j)$, αμφότερες σε γκρίζα κλίμακα και διαστάσεων $M \times N$, το μέσο τετραγωνικό σφάλμα δίνεται από τη σχέση [54],

$$\mathrm{MSE}(I, K) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i,j) - K(i,j)]^2 . \tag{4.1}$$

Για τις ανάγκες της εργασίας, όπως σημειώθηκε και νωρίτερα, ο υπολογισμός του MSE πραγματοποιείται μεταξύ μιας αρχικής εικόνας ενός προσώπου I και της εικόνας $I'(\boldsymbol{x})$, η οποία σχηματίζεται από την εικόνα I έπειτα από αντικατάσταση της περιοχής του προσώπου της με το αντίστοιχο ανακατασκευασμένο rendered 3Δ πρόσωπο (Σχ.4.12). Και οι δύο αυτές εικόνες, είναι διαστάσεων $M = N = 224$ και εκφράζονται στον χρωματικό χώρο RGB, οπότε η σχέση 4.1 δεν επαρκεί ως έχει και θα πρέπει να επεκταθεί καταλλήλως για τον υπολογισμό του μέσου τετραγωνικού σφάλματός τους.

(i) (ii) (iii)

Σχήμα 4.12: (i) Αρχική 2Δ εικόνα, (ii) Rendered ανακατασκευασμένο πρόσωπο, (iii) Αρχική εικόνα έπειτα από αντικατάσταση της περιοχής του προσώπου με το ανακατασκευασμένο rendered πρόσωπο της εικόνας (ii). Οι εικόνες (i) και (iii) χρησιμοποιούνται για την εξαγωγή των μετρικών.

Στο πλαίσιο αυτό, η διαδικασία που ακολουθείται είναι η εξής [55]:

1. Υπολογισμός των μέσων τετραγωνικών σφαλμάτων MSE_R, MSE_G και MSE_B των χρωματικών συνιστωσών R, G και B αντίστοιχα, σύμφωνα με τη σχέση,

$$\mathrm{MSE}_C(I, I') = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I_C(i,j) - I'_C(i,j)]^2 , \text{ με C=\{R,G,B\}}, \tag{4.2}$$

όπου $I_C(i,j)$ και $I'_C(i,j)$ οι τιμές έντασης της εκάστοτε χρωματικής συνιστώσας C, των εικόνων I και I' αντίστοιχα, στη θέση (i,j).

2. Υπολογισμός του ολικού μέσου τετραγωνικού σφάλματος MSE, λαμβάνοντας τον μέσο όρο των μέσων τετραγωνικών σφαλμάτων των τριών χρωματικών συνιστωσών, δηλαδή,

$$\mathrm{MSE}(I, I') = \frac{1}{3}\left(\mathrm{MSE}_R + \mathrm{MSE}_G + \mathrm{MSE}_B\right). \tag{4.3}$$

Το μέσο τετραγωνικό σφάλμα προσφέρει ένα απλό και εύχρηστο τρόπο αξιολόγισης της ομοιότητας μεταξύ της αρχικής και της ανακατασκευασμένης εικόνας. Εξαρτάται ωστόσο σε πολύ μεγάλο βαθμό από την κλίμακα των χρωματικών τιμών των pixels τους. Από το σημείο αυτό και στο εξής, το MSE υπολογίζεται μεταξύ εικόνων, οι οποίες είναι εκφρασμένες στο εύρος $[0, 1]$.

Έτσι, για κάθε εικόνα $I_{d,i}$ ενός dataset d, υπολογίζεται αρχικά το MSE μεταξύ της εικόνας αυτής και της αντίστοιχης ανακατασκευασμένης εικόνας $I'_{d,i}$. Στη συνέχεια λαμβάνεται η μέση τιμή του συνόλου των μέσων τετραγωνικών σφαλμάτων όλων των εικόνων του dataset και προκύπτει κατά αυτό τον τρόπο ένα ολικό μέσο τετραγωνικό σφάλμα για το εν λόγω dataset.

Η διαδικασία αυτή περιγράφεται από την ακόλουθη σχέση,

$$\mathrm{MSE}_d = \frac{1}{N_d}\sum_{i=1}^{N_d}\mathrm{MSE}(I_{d,i}, I'_{d,i}), \tag{4.4}$$

όπου $d = \{\text{CelebA, LFW, 300W-LP, UTKFace, FFHQ}\}$ τα εξεταζόμενα datasets, N_d το πλήθος των εικόνων του dataset d και $\mathrm{MSE}(I_{d,i}, I'_{d,i})$ το μέσο τετραγωνικό σφάλμα των εικόνων $I_{d,i}$ και $I'_{d,i}$ του dataset d.

Στον πίνακα 4.4 παρουσιάζονται οι αριθμητικές τιμές των μέσων τετραγωνικών σφαλμάτων που προέκυψαν για κάθε ένα εκ των datasets.

Πίνακας 4.4: Μέσο τετραγωνικό σφάλμα ανά dataset.

Dataset	**MSE**
CelebA	0.00382118
LFW	0.00337362
300W-LP	0.00347708
UTKFace	0.00378266
FFHQ	0.00450852

L_1-απόσταση

Η L_1-απόσταση ή αλλιώς απόσταση Manhattan πρόκειται ουσιαστικά για την L_1 νόρμα και ισούται με το άθροισμα των απόλυτων διαφορών των αντίστοιχων pixels δύο εικόνων.

Έτσι, δοθείσης μιας αρχικής εικόνας ενός προσώπου I, διαστάσεων $M \times N$ και της ανακατασκευασμένης εικόνας $I'(\boldsymbol{x})$, διαστάσεων επίσης $M \times N$, η L_1 απόσταση, ακολουθώντας αντίστοιχη διαδικασία με αυτή του MSE, υπολογίζεται ως εξής:

1. Υπολογισμός των αποστάσεων $L_{1,R}$, $L_{1,G}$ και $L_{1,B}$ των χρωματικών συνιστωσών R, G και B αντίστοιχα, σύμφωνα με τη σχέση,

$$L_{1,C}(I, I') = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |I_C(i,j) - I'_C(i,j)|\,, \text{ με C=\{R,G,B\}}, \tag{4.5}$$

   οπου, όπως και πριν, $I_C(i,j)$ και $I'_C(i,j)$ οι τιμές έντασης της εκάστοτε χρωματικής συνιστώσας C, των εικόνων I και I' αντίστοιχα, στη θέση (i,j).

2. Υπολογισμός της ολικής L_1-απόστασης, λαμβάνοντας το μέσο όρο των αποστάσεων των τριών χρωματικών συνιστωσών, δηλαδή,

$$L_1(I, I') = \frac{1}{3}\left(\mathrm{L}_{1,R} + \mathrm{L}_{1,G} + \mathrm{L}_{1,B}\right). \tag{4.6}$$

Καθώς η απόσταση L_1 λαμβάνει σχετικά μεγάλες τιμές και εφόσον το προβαλλόμενο ανακατασκευασμένο πρόσωπο παρουσιάζει σε κάποιες περιπτώσεις αναπόφευκτες διαφορές με το αρχικό, όπως για παράδειγμα σε περιπτώσεις ύπαρξης γυαλιών ή αποκρύψεων λόγω άλλων αντικειμένων, κρίνεται σκόπιμο αντί της απόλυτης απόστασης L_1, να χρησιμοποιηθεί η κανονικοποιημένη απόσταση $L_{1,n}$, η οποία προκύπτει από την Εξ.4.6 ως εξής:

$$L_{1,n}(I, I') = \frac{1}{MN} L_1(I, I'), \tag{4.7}$$

όπου, όπως και πριν, $M = N = 224$.

Για την εξαγωγή μιας ολικής L_1-απόστασης ενός dataset, ακολουθείται η ίδια διαδικασία με αυτή του MSE, η οποία περιγράφεται από τη σχέση,

$$L_{1,d} = \frac{1}{N_d} \sum_{i=1}^{N_d} L_{1,n}(I_{d,i}, I'_{d,i}), \tag{4.8}$$

όπου $d = \{\text{CelebA, LFW, 300W-LP, UTKFace, FFHQ}\}$ τα εξεταζόμενα datasets, N_d το πλήθος των εικόνων του dataset d και $L_{1,n}(I_{d,i}, I'_{d,i})$ η κανονικοποιημένη L_1-απόσταση (Εξ.4.7) των εικόνων $I_{d,i}$ και $I'_{d,i}$ του dataset d.

Στον πίνακα 4.5 παρουσιάζονται οι αριθμητικές τιμές των κανονικοποιημένων μετρικών L_1 που προέκυψαν για κάθε ένα εκ των datasets, σύμφωνα με τη σχέση της Εξ.4.8.

Πίνακας 4.5: L_1-απόσταση ανά dataset.

Dataset	$\boldsymbol{L_1}$
CelebA	0.02247077
LFW	0.02290370
300W-LP	0.01902956
UTKFace	0.02389188
FFHQ	0.02519821

Peak Signal-to-Noise Ratio

Η μετρική PSNR υπολογίζει την αναλογία σήμα-προς-θόρυβο κορυφής μεταξύ δύο εικόνων και μετράται σε ντεσιμπέλ (dB). Η αναλογία αυτή χρησιμοποιείται στη συνέχεια ως μέτρηση της ποιότητας μεταξύ της αρχικής και μιας βελτιωμένης εικόνας.

Για τον υπολογισμό του PSNR μεταξύ των εικόνων I και $I'(\boldsymbol{x})$, απαιτείται αρχικά ο προσδιορισμός του μέσου τετραγωνικού σφάλματος (Mean squared error, MSE), το οποίο δίνεται από τη σχέση της Εξ.4.3. Χρησιμοποιώντας λοιπόν τη σχέση αυτή, το PSNR υπολογίζεται ως εξής [54, 55]:

$$\text{PSNR}(I, I') = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}(I, I')} \right) \Rightarrow$$

$$\begin{aligned} \mathrm{PSNR}(I, I') &= 20 \cdot \log_{10} \left(\frac{\mathrm{MAX}_I}{\sqrt{\mathrm{MSE}(I, I')}} \right) \\ &= 20 \cdot \log_{10} \left(\mathrm{MAX}_I \right) - 10 \cdot \log_{10} \left(\mathrm{MSE}(I, I') \right), \end{aligned} \tag{4.9}$$

όπου MAX_I η μέγιστη τιμή έντασης που μπορεί να λάβει κάποιο pixel της εικόνας I.

Από τη σχεση της Εξ.4.9 διαπιστώνεται ότι όσο υψηλότερη είναι η τιμή της μετρικής PSNR, τόσο πιο όμοιες είναι οι εικόνες I και I' και κατά συνέπεια τόσο πιο επιτυχημένη είναι η ανακατασκευή. Κατά αντίστοιχο τρόπο, μικρή τιμή του PSNR υποδηλώνει μεγαλύτερη διαφορά μεταξύ των δύο εικόνων και άρα ανομοιότητα μεταξύ του παραγόμενου και του αρχικού προσώπου.

Όπως και στις προηγούμενες περιπτώσεις, για τον υπολογισμό του ολικού PSNR_d του εκάστοτε dataset d ακολουθείται η σχέση,

$$\mathrm{PSNR}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} \mathrm{PSNR}(I_{d,i}, I'_{d,i}), \tag{4.10}$$

όπου $d = \{\text{CelebA, LFW, 300W-LP, UTKFace, FFHQ}\}$ τα εξεταζόμενα datasets, N_d το πλήθος των εικόνων του dataset d και $\mathrm{PSNR}(I_{d,i}, I'_{d,i})$ το PSNR των εικόνων $I_{d,i}$ και $I'_{d,i}$ του dataset d.

Στον πίνακα 4.6 παρουσιάζονται οι αριθμητικές τιμές των μετρικών PSNR που προέκυψαν για κάθε ένα εκ των datasets, σύμφωνα με τη σχέση της Εξ.4.10.

Πίνακας 4.6: PSNR ανά dataset.

Dataset	**PSNR (dB)**
CelebA	24.886777
LFW	25.144130
300W-LP	25.221276
UTKFace	24.710577
FFHQ	24.146645

Δείκτης Δομικής Ομοιότητας

Αν και οι μετρικές MSE, L_1 και PSNR αποτελούν αξιόπιστα μέτρα σύγκρισης για την ομοιότητα δύο εικόνων, βασίζονται κυρίως στην άμεση μέτρηση της διαφοράς των εντάσεών τους, με αποτέλεσμα να μην συμβαδίζουν με την ανθρώπινη αντίληψη και όραση, η οποία είναι εξοικειωμένη στον εντοπισμό της διαβάθμισης και της μεταβολής της δομικής πληροφορίας σε μία εικόνα. Προς την κατεύθυνση αυτή, οι Z.Wang *et al.*[56], γνωρίζοντας ότι το ανθρώπινο οπτικό σύστημα εξάγει και τις δομικές πληροφορίες από μία οπτική σκηνή δημιούργησαν έναν νέο αλγόριθμο αξιολόγησης, τον δείκτη δομικής ομοιότητας (SSIM).

Ο υπολογισμός του δείκτη αυτού πραγματοποιείται σε τρία ανεξάρτητα μεταξύ τους μέρη, στα οποία υπολογίζονται τρεις συναρτήσεις σύγκρισης: η συνάρτηση φωτεινότητας (l), αντίθεσης (c) και δομής (s) αντίστοιχα. Οι τρεις αυτές συναρτήσεις υπολογίζονται εντός παραθύρων pixels και δίνονται από τις ακόλουθες σχέσεις:

$$l(\boldsymbol{k}, \boldsymbol{y}) = \frac{2\mu_k\mu_y + C_1}{\mu_k^2 + \mu_y^2 + C_1}, \tag{4.11}$$

$$c(\boldsymbol{k}, \boldsymbol{y}) = \frac{2\sigma_k\sigma_y + C_2}{\sigma_k^2 + \sigma_y^2 + C_2}, \tag{4.12}$$

$$s(\boldsymbol{k}, \boldsymbol{y}) = \frac{2\sigma_{ky} + C_3}{\sigma_k\sigma_y + C_3}, \tag{4.13}$$

όπου με $\boldsymbol{k}$ συμβολίζεται το παράθυρο της αρχικής εικόνας (στην περίπτωσή μας I), με $\boldsymbol{y}$ το παράθυρο της εικόνας προς αξιολόγηση (στην περίπτωσή μας $I'(\boldsymbol{x})$), μ_k, μ_y οι μέσες τιμές των παραθύρων $\boldsymbol{k}$ και $\boldsymbol{y}$ αντίστοιχα, σ_k, σ_y οι τυπικές αποκλίσεις των παραθύρων $\boldsymbol{k}$ και $\boldsymbol{y}$ αντίστοιχα και σ_{ky} η συνδιασπορά των δύο παραθύρων. Οι σταθερές C_1, C_2 και C_3 έχουν σταθεροποιητικό ρόλο και προστίθενται για να αποφευχθεί η διαίρεση με το μηδέν, ενώ στη γενική περίπτωση επιλέγονται ίσες με,

$$C_i = (K_i L)^2, \ \textit{για } i = 1, 2, 3, \tag{4.14}$$

όπου L είναι το δυναμικό εύρος των τιμών των pixels (π.χ. 255 για γκρι εικόνες με κωδικοποίηση των 8-bits) και $K_i << 1$ μια μικρή σταθερά.

Συνδυάζοντας τώρα τις τρεις συναρτήσεις σύγκρισης των Εξ.4.11, 4.12 και 4.13, προκύπτει η μετρική SSIM, σύμφωνα με τη σχέση,

$$\mathrm{SSIM}(\boldsymbol{k},\boldsymbol{y}) = \left[l(\boldsymbol{k},\boldsymbol{y})^{\alpha}, c(\boldsymbol{k},\boldsymbol{y})^{\beta}, s(\boldsymbol{k},\boldsymbol{y})^{\gamma}\right], \tag{4.15}$$

όπου $\alpha > 0$, $\beta > 0$ και $\gamma > 0$ παράμετροι που ρυθμίζουν τη σχετική βαρύτητα στον υπολογισμό των τριών συναρτήσεων σύγκρισης. Για λόγους απλότητας, γίνεται η υπόθεση ότι όλες οι συναρτήσεις συμμετέχουν ισόποσα στον υπολογισμό, με $\alpha = \beta = \gamma = 1$ και ως εκ τούτου η έκφραση για τον υπολογισμό του δείκτη δομικής ομοιότητας γίνεται,

$$\mathrm{SSIM}(\boldsymbol{k},\boldsymbol{y}) = \frac{(2\mu_k\mu_y + C_1)(2\sigma_{ky} + C_2)}{(\mu_k^2 + \mu_y^2 + C_1)(\sigma_k^2 + \sigma_y^2 + C_2)}. \tag{4.16}$$

Όπως αναφέρθηκε και παραπάνω, ο δείκτης δομικής ομοιότητας της σχέσης 4.16 υπολογίζεται εντός των κινούμενων παραθύρων $\boldsymbol{k}$ και $\boldsymbol{y}$, των οποίων το είδος και το μέγεθος επιλέγεται κατάλληλα, ανάλογα με τη δομή των εικόνων. Η παραθύρωση αυτή προσδίδει στο δείκτη δομικής ομοιότητας ένα τοπικό χαρακτήρα, ο οποίος σε αρκετές περιπτώσεις είναι αρκετά χρήσιμος, καθώς επιτρέπει την παρακολούθηση των δομικών μεταβολών από περιοχή σε περιοχή της εικόνας.

Παρά ταύτα, τις περισσότερες φορές απαιτείται η εκτίμηση του μέτρου της δομικής ομοιότητας στο σύνολο της εικόνας. Προς την κατεύθυνση αυτή χρησιμοποιείται ο μέσος δείκτης δομικής ομοιότητας δύο εικόνων (Mean Structural Similarity Index Measure, MSSIM), ο οποίος υπολογίζεται ως εξής:

$$\mathrm{MSSIM}(\boldsymbol{K},\boldsymbol{Y}) = \frac{1}{M}\sum_{i=1}^{M}\mathrm{SSIM}(\boldsymbol{k}_i,\boldsymbol{y}_i), \tag{4.17}$$

όπου M είναι το συνολικό πλήθος των παραθύρων.

Από την παραπάνω ανάλυση προκύπτει ότι για το δείκτη και κατ΄ επέκταση για το μέσο δείκτη δομικής ομοιότητας ισχύει ότι,

$$-1 \leq \mathrm{MSSIM} \leq 1, \tag{4.18}$$

όπου όταν MSSIM=1 παρατηρείται πλήρης δομική ταύτιση, ενώ όταν MSSIM $= -1$ παρατηρείται πλήρης δομική απόκλιση των δύο εικόνων.

Στο Σχ.4.13 παρουσιάζεται ένα παράδειγμα ανακατασκευής μαζί με τον αντίστοιχο δείκτη δομικής ομοιότητας. Για την παραθύρωση των εικόνων χρησιμοποιούνται κυλιόμενα Γκαουσιανά παράθυρα, το μέγεθος των οποίων υπολογίζεται κάθε φορά βάσει των αντίστοιχων τυπικών αποκλίσεών τους.

(i) (ii) (iii)

Σχήμα 4.13: (i) Αρχική 2Δ εικόνα, (ii) Rendered εικόνα ανακατασκευασμένου προσώπου, (iii) Δείκτης δομικής ομοιότητας μεταξύ αρχικού και ανακατασκευασμένου προσώπου. Οι τιμές του δείκτη έχουν μεταφερθεί στο εύρος $[0, 1]$ έτσι ώστε αυτός να απεικονιστεί ως γκρι εικόνα. Οι σκούρες περιοχές, με pixels των οποίων οι τιμές προσεγγίζουν το 0, υποδηλώνουν έντονη δομική διαφορά, ενώ οι φωτεινές περιοχές με pixels των οποίων οι τιμές προσεγγίζουν το 1, υποδηλώνουν έντονη δομική ομοιότητα.

Για τον υπολογισμό του ολικού SSIM_d του εκάστοτε dataset d ακολουθείται η σχέση,

$$\text{SSIM}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} \text{MSSIM}(I_{d,i}, I'_{d,i}), \tag{4.19}$$

όπου $d = \{\text{CelebA, LFW, 300W-LP, UTKFace, FFHQ}\}$ τα εξεταζόμενα datasets, N_d το πλήθος των εικόνων του dataset d και $\text{MSSIM}(I_{d,i}, I'_{d,i})$ ο μέσος δείκτης δομικής ομοιότητας των εικόνων $I_{d,i}$ και $I'_{d,i}$ του dataset d.

Στον πίνακα 4.7 παρουσιάζονται οι αριθμητικές τιμές των μετρικών SSIM που προέκυψαν για κάθε ένα εκ των datasets, σύμφωνα με τη σχέση της Εξ.4.19.

Πίνακας 4.7: SSIM ανά dataset.

Dataset	**SSIM**
CelebA	0.85454112
LFW	0.83843956
300W-LP	0.86920302
UTKFace	0.83945890
FFHQ	0.83847975

Ομοιότητα Συνημιτόνου

Η ομοιότητα συνημιτόνου, για την οποία έγινε αναφορά και στην ενότητα 3.3.2, πρόκειται ουσιαστικά για ένα δείκτη ο οποίος εκφράζει την ομοιότητα μεταξύ δύο διανυσμάτων, βασιζόμενος στη γωνία την οποία αυτά σχηματίζουν στο χώρο των χαρακτηριστικών τους. Η τιμή του κυμαίνεται στο διάστημα $[-1, 1]$, με την τιμή -1 να αντιστοιχεί σε διανύσματα διαμετρικά αντίθετα, την τιμή 0 σε κάθετα διανύσματα και την τιμή 1 σε διανύσματα ίδιας φοράς. Παρ΄ όλα αυτά, λόγω της φυσικής του σημασίας, ο δείκτης αυτός χρησιμοποιείται συνήθως στο εύρος $[0, 1]$, με την την τιμή 0 να εκφράζει μέγιστη ανομοιότητα και την τιμή 1 μέγιστη ομοιότητα μεταξύ των δύο συγκρινόμενων διανυσμάτων.

Στα πλαίσια της άσκησης, ο δείκτης ομοιότητας συνημιτόνου χρησιμοποιείται για τη σύγκριση των διανυσμάτων χαρακτηριστικών ενός αρχικού και του αντίστοιχου ανακατασκευασμένου προσώπου, έτσι ώστε να διαπιστωθεί η ομοιότητα σε επίπεδο χαρακτηριστικών του παραγόμενου μοντέλου και του αρχικού 2Δ προσώπου.

Δοθείσης λοιπόν της εξόδου μιας συνάρτησης εξαγωγής χαρακτηριστικών $f(I)$ του προσώπου μιας 2Δ εικόνας I και της εξόδου $f(I'(\boldsymbol{x}))$ των χαρακτηριστικών του αντίστοιχου ανακατασκευασμένου 3Δ προσώπου, το οποίο προσδιορίζεται από το διάνυσμα συντελεστών $\boldsymbol{x}$ και προβάλλεται μέσω rendering στην εικόνα $I'(\boldsymbol{x})$, η ομοιότητα συνημιτόνου υπολογίζεται ως εξής:

$$\text{cosine similarity} = S_C(I, I') = \frac{\langle f(I), f(I'(\boldsymbol{x}))\rangle}{||f(I)|| \cdot ||f(I'(\boldsymbol{x}))||}, \tag{4.20}$$

όπου ο τελεστής $\langle \cdot, \cdot \rangle$ εκφράζει το εσωτερικό γινόμενο.

Για την εξαγωγή ενός αντιπροσωπευτικού δείκτη ομοιότητας συνημιτόνου $S_{C,d}$ για όλο το dataset d, όπως και σε όλες τις προηγούμενες μετρικές, χρησιμοποιείται η σχέση,

$$S_{C,d} = \frac{1}{N_d} \sum_{i=1}^{N_d} S_C(I_{d,i}, I'_{d,i}), \tag{4.21}$$

όπου $d = \{\text{CelebA, LFW, 300W-LP, UTKFace, FFHQ}\}$ τα εξεταζόμενα datasets, N_d το πλήθος των εικόνων του dataset d και $\mathrm{S}_C(I_{d,i}, I'_{d,i})$ ο δείκτης ομοιότητας συνημιτόνου των εικόνων $I_{d,i}$ και $I'_{d,i}$ του dataset d.

Στο σημείο αυτό σημειώνεται πως για λόγους αντικειμενικότητας και αμεροληψίας, για την εξαγωγή των χαρακτηριστικών των προσώπων των εικόνων I και I' δεν χρησιμοποιείται το δίκτυο FaceNet, καθώς αυτό αξιοποιείται κατά την εκπαίδευση του προτεινόμενου δικτύου ανακατασκευής. Αντί για αυτό, χρησιμοποιείται το προεκπαιδευμένο συνελικτικό δίκτυο αναγνώρισης προσώπων *VGG-Face* [57], το οποίο αποτελεί ουσιαστικά μια εφαρμογή της αρχιτεκτονικής ConvNet VGG-16 [58].

Στον πίνακα 4.8 παρουσιάζονται οι αριθμητικές τιμές της ομοιότητας συνημιτόνου S_C που προέκυψαν για κάθε ένα εκ των datasets, σύμφωνα με τη σχέση της Εξ.4.21.

Πίνακας 4.8: Ομοιότητα συνημιτόνου ανά dataset.

Dataset	$\boldsymbol{S_C}$
CelebA	0.84256459
LFW	0.84264438
300W-LP	0.82654271
UTKFace	0.83393454
FFHQ	0.83232659

Αξίζει να αναφερθεί, πως σε αντίθεση με τις υπόλοιπες μετρικές, η ομοιότητα συνημιτόνου επικεντρώνεται μόνο στην καθαυτή περιοχή του προσώπου των συγκρινόμενων εικόνων. Για το λόγο αυτό, πριν γίνει εξαγωγή των διανυσμάτων χαρακτηριστικών $f(I)$ και $f(I')$, οι εικόνες I και I' περικόπτονται έτσι ώστε να περιλαμβάνουν μόνο τις περιοχές των προσώπων. Για την περικοπή αυτή χρησιμοποιείται και πάλι ο ανιχνευτής MTCNN, ο οποίος εντοπίζει τα πρόσωπα της αρχικής και της ανακατσκευασμένης εικόνας και καθορίζει ένα περιβάλλον πλαίσιο σύμφωνα με το οποίο γίνεται η περικοπή.

Καθώς όμως κάποια πρόσωπα εμφανίζονται σε ακραίες τοποθετήσεις (π.χ. 300W-LP dataset), τα αντίστοιχα ανακατασκευασμένα πρόσωπα μπορεί να αποκλίνουν σημαντικά από αυτά ή, ακόμη και αν τα προσεγγίζουν, να μην ανιχνεύονται από τον MTCNN. Για ευνόητους λόγους, οι εικόνες στις οποίες τα πρόσωπα δεν μπορούν να εντοπιστούν από τον MTCNN, αποκλείονται από τον υπολογισμό, με αποτέλεσμα ο αριθμός των εικόνων ενός dataset στα οποία υπολογίζεται η ομοιότητα συνημιτόνου, να είναι ελαφρώς μικρότερος από το συνολικό αριθμό των εικόνων που το απαρτίζουν, χωρίς ωστόσο το γεγονός αυτό να επηρεάζει σημαντικά τα αποτελέσματα.

Ανάλυση Αποτελεσμάτων

Στην ενότητα αυτή θα γίνει ανάλυση και ερμηνεία των αριθμητικών αποτελεσμάτων τα οποία παρατέθηκαν στις προηγούμενες ενότητες, έτσι ώστε να εξαχθούν ποσοτικά συμπεράσματα για την απόδοση του προτεινόμενου δικτύου ανακατασκευής.

Αρχικά και για λόγους ευκολίας, παρατίθενται συγκεντρωτικά στον ακόλουθο πίνακα όλα τα έως τώρα αποτελέσματα των πινάκων 4.4-4.8.

Πίνακας 4.9: Αριθμητικές τιμές Μετρικών ανά dataset.

Dataset	**Μετρικές**				
	MSE	$\boldsymbol{L_1}$	**PSNR (dB)**	**SSIM**	$\boldsymbol{S_C}$
CelebA	0.00382118	0.02247077	24.886777	0.85454112	0.84256459
LFW	0.00337362	0.02290370	25.144130	0.83843956	0.84264438
300W-LP	0.00347708	0.01902956	25.221276	0.86920302	0.82654271
UTKFace	0.00378266	0.02389188	24.710577	0.83945890	0.83393454
FFHQ	0.00450852	0.02519821	24.146645	0.83847975	0.83232659

Μελετώντας λοιπόν τα αριθμητικά αποτελέσματα του πίνακα 4.9, διαπιστώνονται τα εξής:

- Το μέσο τεραγωνικό σφάλμα παραμένει σε αρκετά χαμηλές τιμές και κυμαίνεται στο εύρος $[0.00337362, 0.00450852]$. Η ελάχιστη τιμή του παρατηρείται στο LFW και η μέγιστη στο FFHQ dataset. Αυτή η ελαφρώς αυξημένη τιμή του σφάλματος στο FFHQ dataset οφείλεται κυρίως σε δύο λόγους: 1) στις ιδιαίτερα ευμετάβλητες συνθήκες του περιβάλλοντος των εικονιζόμενων ατόμων και στην έντονη παρουσία αντικειμένων, τα οποία και αποκρύπτουν περιοχές του προσώπου (π.χ. καπέλα, γυαλιά, μαντήλια, καλλοπιστικά αξεσουάρ κ.α.) και 2) στην ύπαρξη πλήθους εικόνων στις οποίες τα εικονιζόμενα άτομα είναι νεαρής και παιδικής ηλικίας, γεγονός το οποίο καθιστά πιο δύσκολη τη διάκριση του φύλλου και των ιδιαίτερων χαρακτηριστικών τους. Παρόλο λοιπόν που το εν λόγω dataset αποτελείται από εικόνες υψηλής ανάλυσης, παρουσιάζει την πιο μεγάλη

τιμή μέσου τετραγωνικού σφάλματος, για τους λόγους που προαναφέρθηκαν. Σε κάθε περίπτωση, φαίρεται ότι το προτεινόμενο δίκτυο ανταποκρίνεται εξίσου αποτελεσματικά σε όλα τα datasets.

- Η κανονικοποιημένη L_1-απόσταση λαμβάνει τιμές, οι οποίες κυμαίνονται στο εύρος $[0.01902956, 0.02519821]$ και είναι όπως αναμενόταν αρκετά μεγαλύτερη από το μέσο τετραγωνικό σφάλμα. Η μετρική αυτή, σε αντίθεση με το MSE εκφράζει την άμεση διαφορά μεταξύ των αρχικών και των ανακατασκευασμένων εικόνων και ως εκ τούτου συμβαδίζει περισσότερο με την ανθρώπινη αντίληψη ως προς την ομοιότητά τους. Η ελάχιστη τιμή της παρατηρείται στο 300W-LP και η μέγιστη, όπως και στην περίπτωση του MSE, στο FFHQ dataset. Παρόλο επομένως που το 300W-LP dataset απαρτίζεται από εικόνες ατόμων σε έντονες πόζες, παρουσιάζει τη μικρότερη L_1 απόσταση, γεγονός το οποίο οφείλεται σε ένα πολύ σημαντικό χαρακτηριστικό του: την επανεμφάνιση των ίδιων ατόμων σε αρκετές φωτογραφίες. Αυτό έχει ως αποτέλεσμα κάθε φορά που ένα πρόσωπο ανακατασκευάζεται με ικανοποιητικό τρόπο, να επανακατασκευάζεται εξίσου αποτελεσματικά και για ένα μεγάλο εύρος τοποθετήσεών του, διατηρώντας κατά αυτό τον τρόπο σε χαμηλά επίπεδα την τιμή του σφάλματος L_1. Συνολικά, και στην περίπτωση της L_1 απόστασης το δίκτυο ανταποκρίνεται αρκετά ικανοποιητικά στις εικόνες όλων των datasets.

- Η μετρική PSNR λαμβάνει τιμές στο εύρος $[24.146645, 25.221276]$, με το FFHQ dataset να παρουσιάζει την ελάχιστη και το 300W-LP τη μέγιστη τιμή. Όπως αναφέρθηκε και στη σχετική ενότητα, μεγαλύτερες τιμές PSNR υποδεικνύουν καλύτερη ποιότητα και ως εκ τούτου ομοιότητα της ανακατασκευασμένης εικόνας με την αρχική. Στη συγκεκριμένη περίπτωση, μεγαλύτερη πιστότητα στην ανακατασκευή, σύμφωνα πάντα με τη μετρική PSNR παρουσιάζουν οι εικόνες του 300W-LP dataset, ενώ τη σχετικά μικρότερη οι εικόνες του FFHQ dataset. Τα αποτελέσματα αυτά είναι γενικώς αναμενόμενα, για τους ίδιους λόγους που αναφέρθηκαν και στην περίπτωση της L_1-απόστασης. Γενικά, φαίρεται ότι οι τιμές της μετρικής PSNR διατηρούνται σχετικά σταθερές σε κάθε dataset και μπορεί να θεωρηθούν αποδεκτές, δεδομένης της φύσης των συγκρινόμενων εικόνων.

- Ο δείκτης δομικής ομοιότητας SSIM κυμαίνεται στο εύρος $[0.83847975, 0.86920302]$ και ακολουθεί ως προς τα άκρα των τιμών του τις μετρικές PSNR και L_1, με το 300W-LP dataset να παρουσιάζει το μέγιστο και το

FFFHQ τον ελάχιστο SSIM. Όπως και στις προηγούμενες περιπτώσεις λοιπόν, αυτή η σχετική υπεροχή του 300W-LP dataset ως προς την τιμή του SSIM οφείλεται στην επαναλαμβανόμενη παρουσία των ίδιων ατόμων σε διαδοχικές εικόνες, γεγονός το οποίο επιφέρει μια αντίστοιχη επανάληψη της δομής, η οποία εφόσον προσδιοριστεί με αποτελεσματικό τρόπο για μία εκ των εικόνων ενός ατόμου, προσδιορίζεται κατόπιν χωρίς ιδιαίτερη δυσκολία για εικόνες του ίδιου ατόμου σε διαφορετικές πόζες. Η απόδοση του δικτύου ως προς την προσέγγιση της δομής των αρχικών προσώπων παρουσιάζει μια γενική σθεναρότητα και παραμένει σε υψηλά επίπεδα για κάθε ένα εκ των εξεταζόμενων datasets.

- Η ομοιότητα συνημιτόνου, λαμβάνει τιμές εντός του εύρους [0.82654271, 0.84264438] και σε αντίθεση με τις προηγούμενες μετρικές, η μέγιστη τιμή της παρατηρείται στην περίπτωση του LFW και η ελάχιστη στην περίπτωση του 300W-LP dataset. Αυτή η σχετική πτώση του 300W-LP dataset στην τελευταία θέση, ως προς την τιμή της ομοιότητας συνημιτόνου, δικαιολογείται πλήρως, αν αναλογιστεί κανείς τις εντονότατες πόζες των προσώπων των εικόνων του (Σχ.4.10), γεγονός το οποίο έχει ως αποτέλεσμα πολλά πρόσωπα να μην εντοπίζονται ή ακόμη και αν εντοπίζονται να βρίσκονται σε τέτοια τοποθέτηση ώστε να γίνεται ελλιπής ή ακόμη και λανθασμένος προσδιορισμός των χαρακτηριστικών τους. Στη γενική περίπτωση, συμπεραίνεται ότι το δίκτυο επιτυγχάνει να αποτυπώσει σε ικανοποιητικό βαθμό τα χαρακτηριστικά των προσώπων των εικονιζόμενων ατόμων, με την ομοιότητα συνημιτόνου να υπερβαίνει την τιμή 0.8 σε όλα τα datasets.

Συνοψίζοντας, τα αριθμητικά αποτελέσματα συμβαδίζουν με τα αντίστοιχα ποιοτικά αποτελέσματα της ενότητας 4.2.1 και αποδεικνύουν την ικανότητα του προτεινόμενου δικτύου να προσδιορίζει αποτελεσματικά τις 3Δ τοπολογίες προσώπων από 2Δ εικόνες, οι οποίες έχουν ληφθεί σε μεταβαλλόμενες συνθήκες περιβάλλοντος. Οι αριθμητικές τιμές των μετρικών σφάλματος αποκαλύπτουν συγχρόνως τη σθεναρότητα της προτεινόμενης μεθόδου, η οποία προσαρμόζεται επιτυχημένα σε εικόνες διαφορετικών datasets, καθώς και στα ιδιαίτερα χαρακτηριστικά των ατόμων που αυτές απεικονίζουν. Παρά ταύτα, τα ίδια αποτελέσματα φανερώνουν και την ευαισθησία της απόδοσης του δικτύου σε έντονες πόζες, εκφράσεις, κακές συνθήκες φωτισμού και σε αντικείμενα τα οποία προκαλούν αποκρύψεις περιοχών των εξεταζόμενων προσώπων, με εντονότερη εμφάνιση των φαινομένων αυτών να συνεπάγεται αυτομάτως μείωση της ακρίβειας στην ανακατασκευή.

Τέλος, αξίζει να επισημανθούν δύο σημεία καίριας σημασίας όσον αφορά στις μετρικές που χρησιμοποιούνται:

1. Για την εξαγωγή των συνολικών μετρικών για κάθε dataset, έχει χρησιμοποιηθεί ο αριθμητικός μέσος όρος των επιμέρους μετρικών των εικόνων που το απαρτίζουν. Αυτή η απλοϊκή μέση τιμή των αριθμητικών τιμών παρέχει μεν μια αρκετά καλή εικόνα σχετικά με τη συνολική απόδοση του δικτύου στο εν λόγω dataset, αποκρύπτει δε κατά κάποιον τρόπο τις κακές περιπτώσεις ανακατασκευής, όπου το δίκτυο αδυνατεί να προσεγγίσει κατά πολύ τα πρόσωπα των εξεταζόμενων εικόνων, για διάφορους από τους λόγους που έχουν προαναφερθεί. Κατά αυτό τον τρόπο, οι αποτυχημένες αυτές περιπτώσεις ανακατασκευής, συμψηφίζονται μαζί με τις αντίστοιχες επιτυχημένες, οι οποίες είναι προφανώς πολύ περισσότερες σε αριθμό, με αποτέλεσμα να μην γίνονται άμεσα αντιληπτές μέσω των μετρικών, παρότι μεταβάλλουν ελαφρώς τις τιμές τους προς το χειρότερο.

2. Οι μετρικές υπολογίζονται μεταξύ των αρχικών και των ανακατασκευασμένων εικόνων και όχι μεταξύ 3Δ τοπολογιών, λόγω έλλειψης σχετικών datasets με 3Δ face scans προσώπων. Αυτή η σύγκριση έχει προφανώς το μειονέκτημα, ότι η 3Δ παραγόμενη τοπολογία του προσώπου, πριν συγκριθεί με το αρχικό πρόσωπο, προβάλλεται μέσω rendering στο 2Δ επίπεδο της εικόνας. Αυτή η μετάβαση από τον 3Δ χώρο στο επίπεδο της εικόνας εξαρτάται σε πάρα πολύ μεγάλο βαθμό από τον renderer ο οποίος χρησιμοποιείται. Έτσι, για το ίδιο 3Δ πρόσωπο, μπορεί να παραχθούν 2Δ εικόνες, οι οποίες παρουσιάζουν απόκλιση μεταξύ τους, λόγω χρήσης διαφορετικών renderers, με αποτέλεσμα η σύγκριση των εικόνων αυτών με την αρχική εικόνα του προσώπου να οδηγεί σε διαφορετικά αποτελέσματα.

Και τα δύο ανωτέρω σημεία είναι πολύ σημαντικά και προσδιορίζουν κατά πόσο οι μετρικές, όταν υπολογίζονται μεταξύ της αρχικής και της ανακατασκευασμένης συνθετικής εικόνας, μπορούν να χρησιμοποιηθούν για την αξιολόγηση της απόδοσης του δικτύου. Καθώς λοιπόν εξετάζεται ένας αρκετά μεγάλος αριθμός εικόνων από διαφορετικά datasets και εφόσον ο διαφορικός renderer της βιβλιοθήκης tensorflow-graphics που χρησιμοποιείται είναι εξαιρετικά ακριβής και καλά σχεδιασμένος, οι μετρικές, παρόλο που δεν υπολογίζονται μεταξύ 3Δ μοντέλων και πραγματικών 3Δ face scans, προσφέρουν μια πολύ καλή εικόνα για την απόδοση του προτεινόμενου δικτύου ανακατασκευής σε πραγματικά δεδομένα.

Κεφάλαιο 5

Συμπεράσματα και μελλοντική Έρευνα

5.1 Ανακεφαλαίωση - Γενικά Συμπεράσματα

Στην παρούσα εργασία προτείνεται μία μέθοδος για την αντιμετώπιση του προβλήματος της 3Δ ανακατασκευής προσώπων από 2Δ έγχρωμες εικόνες χαμηλής ανάλυσης. Στο πλαίσιο αυτό, σχεδιάζεται ένα δίκτυο ανακατασκευής, το οποίο δοθείσης στην είσοδο μιας εικόνας ενός ατόμου, υπολογίζει ένα διάνυσμα 257 συντελεστών, το οποίο κωδικοποιεί όλη την απαραίτητη πληροφορία για το σχήμα, την έκφραση και την υφή του προσώπου του, τις συνθήκες φωτισμού του περιβάλλοντος και την τοποθέτηση της κάμερας στο χώρο.

Για την 3Δ μοντελοποίηση του ανθρώπινου προσώπου χρησιμοποιείται το στατιστικό μοντέλο αναπαράστασης Basel Face Model, ενώ για την απόδοση των συνθηκών φωτισμού αξιοποιείται η θεωρία των σφαιρικών αρμονικών συναρτήσεων. Το δίκτυο ανακατασκευής βασίζεται αποκλειστικά στη χρήση συνελικτικών νευρωνικών δικτύων και συγκεκριμένα στην αρχιτεκτονική του μοντέλου ResNet-50, η οποία τροποποιείται καταλλήλως για να εξυπηρετεί τις ανάγκες της ανακατασκευής. Η εκπαίδευση παραγματοποιείται χωρίς επίβλεψη, ελαχιστοποιώντας μια σύνθετη υβριδική συνάρτηση απώλειας, με χρήση ενός κατάλληλα κατασκευασμένου dataset εικόνων χαμηλής ανάλυσης διάσημων προσώπων.

Τα αποτελέσματα της ανακατασκευής επιβεβαιώνουν την αποτελεσματικότητα της προτεινόμενης μεθόδου και αποδεικνύουν την ικανότητα ακριβούς α-

νάκτησης 3Δ τοπολογιών προσώπων από 2Δ εικόνες. Συγχρόνως, φανερώνουν την άμεση εξάρτηση της απόδοσης του δικτύου από έντονες πόζες, εκφράσεις και μεταβολές του φωτισμού των εικονιζόμενων ατόμων. Τα φαινόμενα αυτά επηρεάζουν σε πολύ μεγάλο βαθμό τα αποτελέσματα της ανακατασκευής, οδηγώντας κάποιες φορές και ανάλογα με την έντασή τους, σε αποκλίνοντα και μη ρεαλιστικά παραγόμενα 3Δ μοντέλα προσώπων.

5.2 Μελλοντική Έρευνα

Παρά τα θετικά και ικανοποιητικά αποτελέσματα τα οποία λαμβάνονται στα πλαίσια της εργασίας, το προτεινόμενο δίκτυο υστερεί σε κάποιες συνιστώσες της ανακατασκευής, όπως αυτές που αφορούν στην απόδοση υψίσυχνων τοπικών λεπτομερειών και έντονων εκφράσεων. Το πρόβλημα αυτό οφείλεται στις σχετικά περιορισμένες δυνατότητες που προσφέρει το 3DMM, λόγω της μειωμένης διαστατικότητάς του.

Για τη βελτίωση της πιστότητας των παραγόμενων 3Δ μοντέλων και την αντιμετώπιση του προβλήματος αυτού, μπορεί να γίνει επέκταση της βασικής διαδικασίας ανακατασκευής με τους εξής τρόπους:

- **Χρήση Γεννητικών Ανταγωνιστικών Δικτύων (GANs)**, στα οποία ο γεννήτορας (generator) εκπαιδεύεται ώστε να παράγει ένα σύνολο δεδομένων, με κατανομή παρόμοια με αυτή των δεδομένων εισόδου, επιδιώκοντας να ξεγελάσει τον διευκρινιστή, έτσι ώστε να μην μπορεί να ξεχωρίσει εάν τα δεδομένα αυτά είναι πραγματικά ή συνθετικά. Από την πλευρά του, ο διευκρινιστής εκπαιδεύεται ώστε να μπορεί συνεχώς να διακρίνει τα πραγματικά δεδομένα από τα συνθετικά του γεννήτορα. Η ιδέα των GANs έχει χρησιμοποιηθεί σε αρκετές μεθόδους 3Δ ανακατασκευής προσώπων από εικόνες, όπου τα γεννητικά αυτά δίκτυα εκπαιδεύονται είτε για να βελτιώσουν τα ήδη υπάρχοντα 3Δ μοντέλα (3DMM) ως προς την ακρίβεια και την πιστότητα [59, 60, 61], είτε για να παρέχουν καλύτερες εκτιμήσεις των παραμέτρων του 3DMM [62, 63]. Επιπλέον, κατά τα τελευταία χρόνια, τα GANs χρησιμοποιούνται για την απευθείας παραγωγή χαρτών υφής ανθρώπινων προσώπων (UV texture maps, Σχ.5.1), οι οποίοι αντικαθιστούν τη συνιστώσα της υφής του βασικού 3DMM, παράγοντας 3Δ μοντέλα με πολύ υψηλό επίπεδο ακρίβειας και λεπτομερειών [64, 65].

Σχήμα 5.1: Ανακατασκευή προσώπων με χρήση GANs. Από αριστερά προς τα δεξιά: 3Δ γεωμετρία προσώπου, UV χάρτης υφής παραγόμενος από κατάλληλα εκπαιδευμένο GAN, τελικό ανακατασκευασμένο πρόσωπο.

- **Χρήση πολλαπλών συνελικτικών δικτύων**, καθένα εκ των οποίων εξετάζει μια διαφορετική συνιστώσα της ανακατασκευής. Τα δίκτυα αυτά μπορεί είτε να συνδέονται και να εκπαιδεύονται με ακολουθιακό τρόπο, σχηματίζοντας ένα πολυσταδιακό σύστημα (multi-stage system) [66, 67], είτε να συνδέονται και να εκπαιδεύονται παράλληλα, σχηματίζοντας μια πολυστρωματική δομή [68, 69, 70]. Αυτή η κατανομή των συνιστωσών σε επιμέρους δίκτυα επιφέρει σημαντική βελτίωση στην ακρίβεια των αποτελεσμάτων. Παρ΄ όλα αυτά είναι ιδιαιτέρως δύσκολη και επίπονη, καθώς απαιτεί τον προσεκτικό συντονισμό της εκπαίδευσης πολλαπλών νευρωνικών δικτύων.

- **Χρήση πολλαπλών εικόνων του ίδιου ατόμου** κατά την εκπαίδευση των μοντέλων ανακατασκευής, όταν και όπου αυτό είναι εφικτό, έτσι ώστε να εξάγωνται πληροφορίες για το ίδιο άτομο από πολλαπλές θέσεις και οπτικές γωνίες, οι οποίες στη συνέχεια συμψηφίζονται για να παραχθεί ένα τελικό 3Δ μοντέλο [30, 71]. Κατά αυτό τον τρόπο αποκτάται μια πιο ολοκληρωμένη εικόνα για τα χαρακτηριστικά του εκάστοτε προσώπου και αντιμετωπίζονται καλύτερα φαινόμενα όπως αποκρύψεις και κακός φωτισμός.

- **Χρήση μιας πιο σύνθετης υβριδικής συνάρτησης απώλειας**, προκειμένου να ελέγχονται πιο αποτελεσματικά οι διάφορες συνιστώσες τις ανακατασκευής κατά την εκπαίδευση των δικτύων. Αυτό επιτυγχάνεται είτε μέσω της προσθήκης επιπλέον επιμέρους συναρτήσεων απώλειας, είτε μέσω της τροποποίησης και επέκτασης των ήδη υπαρχουσών.

Βιβλιογραφία

[1] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.C. Chen, and H. Li. Avatar Digitization from a Single Image for Real-time Rendering. In *ACM Transactions on Graphics (TOG)*, volume 36, issue 6, pages 195:1–195:14, 2017.

[2] P. Ghosh, P.S. Gupta, R. Uziel, A. Ranjan, Michael J. Black, and T. Bolkart. GIF: Generative Interpretable Faces. In *International Conference on 3D Vision (3DV)*, pages 868–878, 2020.

[3] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.P. Seidel, P. Pérez, M. Zollhöfer and C. Theobalt. StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6141–6150. 2020.

[4] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black. Capture, Learning, and Synthesis of 3D Speaking Styles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–11, 2019.

[5] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement, pages 1-8, 2021.

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 25, pages 9:1063–1074, 2003.

[7] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3D reconstruction for Face Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 843–848, 2004.

[8] J. Wang, L. Yin, X. Wei, and Y. Sun. 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. In *IEEE Computer So-*

ciety Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, pages 1399–1406, 2006.

[9] Volker Blanz, Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999.

[10] P. Paysan and R. Knothe and B. Amberg and S. Romdhani and T. Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, pages 1-6, 2009.

[11] T. F. Cootes, G. J. Edwards and C. J. Taylor. Active Appearance Models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 23, pages 6:681-685, 2001.

[12] W. S. McCulloch and W. Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, volume 5, pages 115-133, 1943.

[13] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In *The Joint ACM Workshop on Human Gesture and Behavior Understanding*, pages 79–80, 2011.

[14] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. In *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, volume 20, pages 3:413–425, 2014.

[15] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *International Conference on 3D Vision (3DV)*, pages 460–469, 2016.

[16] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–26, 2017.

[17] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, 2017.

[18] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR), pages 4625–4634, 2018.

[19] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. P´erez, and C. Theobalt. MoFa: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 1274–1283, 2017.

[20] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] F. Wu, L. Bao, Y. Chen, Y. Ling, Y. Song, S. Li, K. Ngi Ngan and W. Liu. MVF-Net: Multi-View 3D Face Morphable Model Regression. In *Conference on Computer Vision and Pattern Recognition (CVRP)*, pages 1-10, 2019.

[22] X. Zeng, X. Peng and Yu Qiao. DF^{2}Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2315-2324, 2019.

[23] Z. Liu, P. Luo, X. Wang and X. Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1-11, 2015.

[24] K. Zhang, Z. Zhang, Z. Li and Y. Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. In *IEEE Signal Processing Letters*, volume 23, pages 10:1499-1503, 2016.

[25] Iván de Paz Centeno. Implementation of the MTCNN face detector for Keras in Python3.4+. https://github.com/ipazc/mtcnn.

[26] D. E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, pages 1755-1758, 2009.

[27] C. Sagonas, E. Antonakos, G, Tzimiropoulos, S. Zafeiriou, M. Pantic. 300 faces In-the-wild challenge: Database and results. *Image and Vision Computing (IMAVIS), Special Issue on Facial Landmark Localisation "In-The-Wild"*, 2016.

[28] D. E. King. Shape predictor for 68 face landmarks. https://github.com/davisking/dlib-models/blob/master/shape_predictor_68_face_landmarks.dat.bz2

[29] D. Chen, G. Hua, F.Wen, and J. Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision (ECCV)*, pages 122–138, 2016.

[30] Yu Deng, J. Yang, S. Xu, D. Chen, Y. Jia and X. Tong. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *IEEE Computer Vision and Pattern Recognition Workshops*, pages 1-11, 2019.

[31] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 274-280, 1999.

[32] X. Y. Cao and H. F. Liu. A Skin Detection Algorithm Based on Bayes Decision in the YCbCr Color Space. *Applied Mechanics and Materials*, 121–126, pages 672–676, 2011.

[33] M.H. Yang and N. Ahuja. Gaussian Mixture Model for Human Skin Color and Its Applications in Image and Video Databases. In textitProceedings of SPIE - The International Society for Optical Engineering 3656(1), 10.1117/12.333865, 1998.

[34] R. Bhatt and A. Dhall. Skin Segmentation Dataset. *UCI Machine Learning Repository.*

[35] L. Donghui and W. Bin. A Skin Detection Method Based on Bayesian Decision. In *Journal of Image and Graphincs*, 11(1), pages 47-52, 2006.

[36] W. Zhongdong, W. Saichao and H. Zichao. A Bayesian approach to skin detection in YCbCr color space. In *International Joint Conference on Awareness Science and Technology and Ubi-Media Computing (iCAST 2013 and UMEDIA 2013)*, pages 606-610, 2013.

[37] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, 2009.

[38] Y. Guo, J. Z. Zhang, J. Cai, B. Jiang, and J. Zheng. Cnn-based real-time dense face reconstruction with inverserendered photo-realistic face images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[39] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. In *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(3), pages 413–425, 2014.

[40] R. Ramamoorthi. Modeling Illumination Variation with Spherical Harmonics. In *Face Processing: Advanced Modeling Methods*, pages 385–424. 2006.

[41] R. Green. Spherical Harmonic Lighting: The Gritty Details, 2003.

[42] Wikipedia contributors. Spherical harmonics. *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Spherical_harmonics&oldid=1073167901

[43] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01)*, pages 497–500, 2001.

[44] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770-778, 2016.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision (IJCV)*, 2015.

[46] Wikipedia contributors. Barycentric coordinate system . *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Barycentric_coordinate_system&oldid=1064042643

[47] Scratchapixel. Rasterization: A practical implementation (the rasterization stage), https://www.scratchapixel.com/lessons/3d-basic-rendering/rasterization-practical-implementation/rasterization-stage

[48] A. G. Belyaev and Y. Ohtake. Nonlinear diffusion of normals for crease enhancement. In *Proceedings of SPIE* 4476, Vision Geometry X, 2001.

[49] F. Schroff, D. Kalenichenko and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815-823, 2015.

[50] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *University of Massachusetts, Amherst*, Technical Report 07-49, 2007.

[51] X. Zhu, Z. Lei, X. Liu, H. Shi and S. Z. Li. Face Alignment Across Large Poses: A 3D Solution. 2015.

[52] Z. Zhang, Y. Song and H. Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[53] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[54] Wikipedia contributors. Peak signal-to-noise ratio. *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Peak_signal-to-noise_ratio&oldid=1062145991

[55] P. Gupta, P. Srivastava, S. Bhardwaj and V. Bhateja. A modified PSNR metric based on HVS for quality assessment of color images. In *International Conference on Communication and Industrial Application*, pages 1-4, 2011.

[56] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. In *IEEE Transactions on Image Processing*, volume 13, number 4, pages 600-612, 2004.

[57] O. M. Parkhi, A. Vedaldi and A. Zisserman. Deep Face Recognition. In *The British Machine Vision Conference (BMVC)*, pages 1-12, 2015.

[58] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Computing Research Repository (CoRR)*, 2015.

[59] L. Galteri, C. Ferrari, G. Lisanti, S. Berretti and A. Del Bimbo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 25-31, 2019.

[60] L. Galteri, C. Ferrari, G. Lisanti, S. Berrettiand A. Del Bimbo. Deep 3D morphable model refinement via progressive growing of conditional generative adversarial networks In *Computer Vision and Image Understanding*, pages 31-42, 2019.

[61] J. Lin, Y. Yuan, T. Shao and K. Zhou. Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5890-5899, 2020.

[62] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, J. Feng. 3D face reconstruction from a single image assisted by 2d face images in the wild. In *IEEE Transactions on Multimedia*, volume 23, pages 1160-1172, 2021.

[63] Z. Gao, J. Zhang, Y. Guo, C. Ma, G. Zhai and X. Yang. Semi-supervised 3D Face Representation Learning from Unconstrained Photo Collections. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1426-1435, 2020.

[64] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1-15, 2021.

[65] J. Deng , S. Cheng, N. Xue, Y. Zhou and S. Zafeiriou. UV-GAN: Adversarial Facial UV Map Completion for Pose-Invariant Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7093-7102, 2018.

[66] S. Sengupta, A. Kanazawa, C. D. Castillo and D. W. Jacobs. Sf-SNet: Learning shape, reflectance and illuminance of faces in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6296-6305, 2018.

[67] C. Bhagavatula, C. Zhu, K. Luu and M. Savvides. Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses. In *International Conference on Computer Vision (ICCV)*, pAGES 3980-3989, 2017.

[68] X. Fan, S. Cheng, K. Huyan, M. Hou, R. Liu and Z. Luo. Dual neural networks coupling data regression with explicit priors for monocular 3D face reconstruction. In *IEEE Transactions on Multimedia*, volume 23, pages 1252-1263, 2021.

[69] E. Richardson, M. Sela, R. Or-El and R. Kimmel. Learning detailed face reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1259-1268, 2017.

[70] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh and S. Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction "in-the-wild". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 757-766, 2020.

[71] S. Sanyal, T. Bolkart, H. Feng and M. J. Black. Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7755-7764, 2019.