Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρονικής και Υπολογιστών

Αναγνώριση deepfakes μέσω αποτύπωσης των 3D βιομετρικών χαρακτηριστικών του προσώπου

Διπλωματική Εργασία
του
Βέλλιου Γεώργιου-Σεραφείμ

Επιβλέποντες:

Καθ. Αναστάσιος Ντελόπουλος

Υπ. Δρ. Αντώνης Καρακώττας

ΘΕΣΣΑΛΟΝΙΚΗ 2022

Aristotle University of Thessaloniki
Faculty of Engineering

School of Electrical and Computer Engineering
Department of Electronics and Computer Engineering

Detecting deepfakes by imprinting face's 3D biometric characteristics

Thesis
of
Vellios Georgios-Serafeim

Supervisors:

Prof. Anastasios Delopoulos

Phd Cand. Antonis Karakottas

THESSALONIKI 2022

Περίληψη

Το τελευταίο διάστημα οι μέθοδοι παραγωγής παραποιημένων προσώπων, γνωστά και ως deepfakes, έχουν εξελιχθεί ραγδαία, σε βαθμό που πολλές φορές οι διαφορές από τα πραγματικά πρόσωπα δεν είναι ορατές με γυμνό μάτι. Στο πλαίσιο αυτό έχουν προταθεί διάφορες τεχνικές αναγνώρισης παραποιημένων προσώπων. Στην παρούσα εργασία μελετήθηκαν μέθοδοι αναγνώρισης deepfakes που αποτυπώνουν τα 3D βιομετρικά χαρακτηριστικά του προσώπου. Για να μην χρησιμοποιηθούν αυθεντικά 3D δεδομένα, έγινε χρήση 3DMMs. Επίσης στις μεθόδους λήφθηκε υπόψη η μεταβολή στο χρόνο οπότε και αυτές λειτουργούν πάνω σε βίντεο, αφού πρώτα τα καρέ των βίντεο μετατραπούν σε παραμέτρους που εκφράζουν τα 3DMMs. Ακόμη εξάγονται χαρακτηριστικά από τα ορόσημα του προσώπου που βοηθούν στην αναγνώριση των deepfakes, όπως πληροφορίες για το ανοιγοκλείσιμο του στόματος και των ματιών και συγκεκριμένες γωνίες του προσώπου.

Οι μέθοδοι που μελετήθηκαν χωρίζονται σε 2 κατηγορίες, στους One Class Classifiers και στους Binary Classifiers. Στην πρώτη περίπτωση μελετήθηκαν μοντέλα VAE και GAN, 2 για κάθε κατηγορία, και στην δεύτερη ένα βαθύ συνελικτικό δίκτυο, το EfficientNet. Παρατηρήθηκε ότι τα GAN παρουσιάζουν σχετικά καλά αποτελέσματα στην αναγνώρισή των deepfakes, καλύτερα από αυτά των VAE. Ωστόσο ο Binary Classifier παρουσίασε τα καλύτερα αποτελέσματα. Επίσης τα χαρακτηριστικά που προτάθηκαν στα πλαίσια της εργασίας βελτίωσαν την απόδοση της μεθόδου του Binary Classification και θεωρούνται, επομένως, ικανά να βοηθήσουν στην αναγνώρισή των παραποιημένων προσώπων.

Abstract

Recently, the methods of producing modified faces, also known as deepfakes, have evolved rapidly, to the extent that many times the differences from real faces are not visible to the naked eye. In this context, various techniques for identifying modified faces have been proposed. In this work, methods for deepfakes recognition that use 3D biometric characteristics of the face were studied. In order not to use original 3D data, 3DMMs were used. In the studied methods temporal features were taken into account. To achieve this, the frames of the videos that are used on the methods are firstly converted into parameters that express the 3DMMs. Features that help identify deepfakes are also extracted from facial landmarks. Such features are information about how open or close the eyes and the mouth are, as well as specific angles of the face.

The studied methods are divided into two categories, One Class Classifiers and Binary Classifiers. In the first case, VAE and GAN models were studied, two for each category, and in the second case a deep convolutional network , EfficientNet, was studied. It was observed that GANs show relatively good results in recognizing deepfakes, results that outperform those of VAE. However, the Binary Classifier presented the best results. Also, the features proposed in this work improved the performance of the Binary Classifier and are, therefore, considered capable of helping to identify deepfakes.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1 Εισαγωγή

1.1 Διατύπωση του Προβλήματος

Βρισκόμαστε σε μια εποχή ραγδαίων τεχνολογικών εξελίξεων. Η τεχνολογική πρόοδος έχει οδηγήσει σε βελτίωση της καθημερινής μας ζωής καθώς και στην υλοποίηση πολλών πραγμάτων που κάποτε φαίνονταν αδύνατα. Ωστόσο, παρά τον θετικό της αντίκτυπο, έχει οδηγήσει και σε αύξηση ήδη υπαρχόντων προβλημάτων, μεταξύ των οποίων της πλαστογραφίας και της πλαστοπροσωπίας. Στα ήδη υπάρχοντα πλαστά έγγραφα (ταυτότητες, διαβατήρια, κ.τ.λ.), πλαστές υπογραφές, προσποίηση άλλων προσώπων και φωνών, προστίθενται τώρα και οι πλαστές ή παραποιημένες φωτογραφίες και βίντεο που δημιουργούνται μέσω της τεχνολογίας των deepfakes.

Ως deepfakes θεωρείται η τεχνολογία η οποία, χρησιμοποιώντας αλγορίθμους βαθιάς μάθησης (**Deep Learning Algorithms),** δύναται να παράξει συνθετικές εικόνες και βίντεο, κυρίως της περιοχής του προσώπου. Οι μέθοδοι παραποίησης του προσώπου χωρίζονται κυρίως σε δύο κατηγορίες: i) παραποίηση της ταυτότητας (face swapping) και ii) παραποίηση της έκφρασης του προσώπου (face re-enactment). Η μέθοδος βασίζεται στην εύρεση των οροσήμων του προσώπου (facial landmarks, κεφ. 2.5) και στην σύνθεση νέων εικόνων όπου συνδυάζονται τόσο το αληθινό πρόσωπο όσο και το πρόσωπο με το οποίο θα γίνει η παραποίηση, ώστε τα ορόσημα των δύο προσώπων να βρίσκονται όσο το δυνατόν πιο κοντά.

Εικόνα 1: Παράδειγμα μιας deepfake εικόνας που παράχθηκε με την μέθοδο face-swapping

Η τεχνολογία των deepfakes συνεχώς εξελίσσεται. Τα deepfakes χρησιμοποιούνται σε αρκετούς τομείς, μεταξύ των οποίων η ιατρική και η εκπαίδευση. Στην ιατρική τα deepfakes μπορούν να χρησιμοποιηθούν για να μειώσουν τον χρόνο ανάκαμψης για άτομα που πάσχουν από παράλυση ή προβλήματα εθισμού. Για παράδειγμα, το βίντεο ενός αθλητή που έχει παραποιηθεί με το πρόσωπο του ασθενή μπορεί να εμψυχώσει τον ασθενή και να συνεργαστεί για πιο γρήγορη ανάρρωση. Στον τομέα της εκπαίδευσης, η χρήση των

deepfakes για δημιουργία συνθετικών βίντεο ή παραποιημένων βίντεο από ιστορικά πρόσωπα θα κάνει σίγουρα την εκπαιδευτική διαδικασία πιο γοητευτική και θα επιφέρει καλύτερα αποτελέσματα.

Ωστόσο, η τεχνολογία των deepfakes μπορεί να χρησιμοποιηθεί και για κακόβουλους σκοπούς. Ένας από αυτούς είναι η δημιουργία βίντεο σεξουαλικού περιεχομένου ατόμων χωρίς την συγκατάθεσή τους. Θύματα αυτής της κακόβουλης χρήσης είναι κυρίως διάσημες γυναίκες καθώς και ανήλικα κορίτσια. Επίσης τα deepfakes χρησιμοποιούνται για την διασπορά ψευδών ειδήσεων και απατών. Σε μία πρόσφατη απάτη ένα deepfake βίντεο του Elon Musk προέτρεπε τους θεατές να επενδύσουν σε ένα κρυπτονόμισμα «απάτη»[49]. Ακόμη υπάρχει πλήθος deepfake βίντεο που δείχνει πρόσωπα ενδιαφέροντος, όπως πολιτικούς αρχηγούς (Barak Obama, Hillary Clinton, κ.τ.λ.), να λένε πράγματα που έχουν σκοπό την ψυχαγωγία των θεατών. Όμως τέτοια βίντεο θα μπορούσαν να στοχεύουν στην διάδοση ρητορικής μίσους ή στην παραπληροφόρηση. Για να τονίσει τον παραπάνω κίνδυνο, ένα βίντεο δημοσιεύτηκε στην πλατφόρμα YouTube που δείχνει τον Barak Obama, πρώην πρόεδρο των Η.Π.Α. να λέει, μεταξύ άλλων φρασεολογίων που δεν ταιριάζουν σε ένα υψηλόβαθμο στέλεχος, την φράση: '*Μπαίνουμε σε μία εποχή στην οποία οι εχθροί μας μπορούν να κάνουν τον οποιοδήποτε να πει το οτιδήποτε οποιαδήποτε στιγμή*.' Στο τέλος του βίντεο αναφέρεται ότι είναι προϊόν deepfake και έχει ενημερωτικό σκοπό για το πως η τεχνολογία αυτή μπορεί να γίνει επικίνδυνη αν χρησιμοποιηθεί για τους λάθους σκοπούς.

Προκειμένου να αντιμετωπιστούν οι αρνητικές επιπτώσεις της χρήσης των deepfakes, η ερευνητική κοινότητα έχει στραφεί στην ανάπτυξη μεθόδων για την αναγνώριση των deepfakes. Οι περισσότερες τεχνικές στηρίζονται στην αναγνώριση ατελειών σε επίπεδο pixels που μπορεί να παραχθούν κατά την διαδικασία της σύνθεσης**.** Όμως, όσο εξελίσσεται η τεχνολογία των deepfakes, τόσο λιγότερες είναι οι 'ατέλειες' που αφήνουν οι μέθοδοι αυτοί πάνω στις εικόνες και, συνεπώς, τόσο λιγότερο αποτελεσματική είναι η χρήση των παραπάνω τεχνικών για τον εντοπισμό τους. Επίσης οι τεχνικές αυτές τις περισσότερες φορές λειτουργούν λαμβάνοντας ως είσοδο παραποιημένες εικόνες χωρίς να λαμβάνουν υπόψη τις όποιες συνδέσεις μεταξύ διαδοχικών εικόνων (βίντεο) που μπορεί να οδηγήσουν στην αναγνώριση των deepfakes. Τέλος οι περισσότερες τεχνικές εκπαιδεύονται σε datasets που έχουν παραχθεί από μία ή μερικές συγκεκριμένες μεθόδους δημιουργίας deepfakes, και έτσι αντιμετωπίζουν δυσκολίες στην γενίκευση.

1.2 Βιβλιογραφική Επισκόπηση

Στα πλαίσια της ψηφιακής εγκληματολογίας έχει προταθεί πλήθος μεθόδων για την αναγνώριση των παραποιημένων προσώπων (**deepfakes**). Θα μπορούσε να ειπωθεί ότι οι τεχνικές αναγνώρισης των deepfakes χωρίζονται σε δύο βασικές κατηγορίες, τεχνικές με προσέγγιση χαμηλού και υψηλού επιπέδου.

1.2.1 Τεχνικές χαμηλού επιπέδου

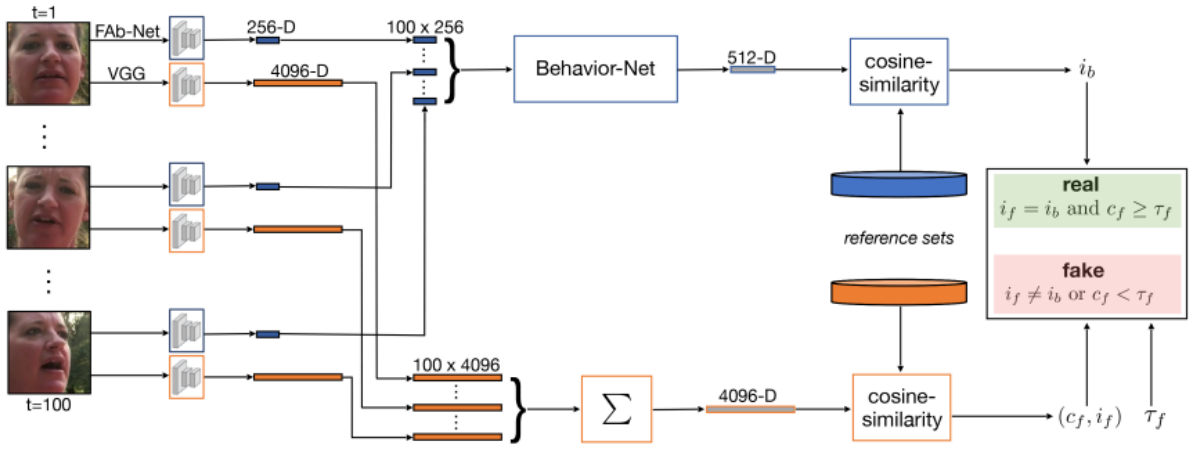
Οι **τεχνικές χαμηλού επιπέδου** στηρίζονται στην αναγνώριση ατελειών σε επίπεδο pixels που μπορεί να παραχθούν κατά την διαδικασία της σύνθεσης. Οι Li et al.[26] χρησιμοποιούν ένα συνελικτικό νευρωνικό δίκτυο (CNN) για να εντοπίσουν τέτοιες ατέλειες που οφείλονται στη διαδικασία 'διπλώματος' (wrapping) ενός προσώπου πάνω σε ένα άλλο. Οι Zhou et al.[27] δημιούργησαν δύο δίκτυα που συνδυάζονται για τον εντοπισμό των deepfakes, με το πρώτο να κατηγοριοποιεί ολόκληρο το πρόσωπο ως παραποιημένο ή όχι και το δεύτερο να ελέγχει μικρότερα υποσύνολα του προσώπου εκμεταλλευόμενο στεγανογραφικά χαρακτηριστικά (steganography features) για ασυνέχειες μεταξύ τους. Στο [28] προτείνεται το δίκτυο MesoNet, ένα CNN με μικρό αριθμό επιπέδων που εκπαιδεύεται να αναγνωρίζει μικροσκοπικές ατέλειες, για τον εντοπισμό των deepfakes. Οι συγγραφείς του [29] απέδειξαν ότι οι εικόνες που παράγονται από GAN[2,8] αφήνουν συγκεκριμένα αποτυπώματα που μπορούν να χρησιμοποιηθούν για τον εντοπισμό εικόνων που έχουν δημιουργηθεί με GAN, στα οποία στηρίζονται πολλές μέθοδοι παραγωγής deepfakes.

Οι τεχνικές χαμηλού επιπέδου, παρότι εμφανίζουν πολύ καλά αποτελέσματα στην αναγνώριση deepfakes, έχουν πρόβλημα στην γενίκευση καθώς έχουν εκπαιδευτεί πάνω σε κάποιες μόνο τεχνικές παραγωγής deepfakes και δεν μπορούν να γενικευτούν σε καινούργιες ή μελλοντικές τεχνικές. Επίσης είναι λιγότερο αποτελεσματικές σε εικόνες στις οποίες έχουν μεταβληθεί το μέγεθος και η κωδικοποίηση. Οι **τεχνικές με προσέγγιση υψηλού επιπέδου** δύναται να λύσουν τα παραπάνω μειονεκτήματα μιας και στηρίζονται σε πιο σημαντικά σημασιολογικά χαρακτηριστικά. Αρκετές ερευνητικές εργασίες [1,5,10,21,13,15,25,17,20,23,18] που στηρίζονται σε τεχνικές υψηλού επιπέδου παρουσιάζονται παρακάτω.

1.2.2 Τεχνικές υψηλού επιπέδου

Στην έρευνα των Cozzolino et al.[1] η αναγνώριση των deepfakes βασίζεται στα βιομετρικά χαρακτηριστικά των προσώπων ώστε να εντοπιστούν πιθανές παραποιήσεις των βίντεο, καθώς στα παραποιημένα βίντεο η ταυτότητα του παραποιημένου προσώπου δεν ταιριάζει με τα βιομετρικά χαρακτηριστικά του. Το προτεινόμενο μοντέλο είναι ένα GAN στο οποίο η εκπαίδευση γίνεται με βάση χαρακτηριστικών του προσώπου που εξάγονται από τρισδιάστατα μορφοποιήσημα μοντέλα (3DMMs). Τα χαρακτηριστικά αυτά εξάγονται για κάθε καρέ του βίντεο και μειώνεται η διάστασή τους ώστε τελικά να αποτελούνται από 62 συντελεστές (40 για το σχήμα, 12 για την πόζα και 10 για την έκφραση). Η εξαγωγή των συντελεστών βασίστηκε στο Regression μοντέλο που προτάθηκε από τον Guo et al.[3] Τόσο το παραγωγικό όσο και το διαχωριστικό δίκτυο αποτελούνται κυρίως από υπολειμματικά επαναληπτικά στρώματα (Residual Blocks)[4] (συλλογή από επίπεδα όπου η έξοδος του ενός προστίθεται σε ένα μεταγενέστερο επίπεδο).

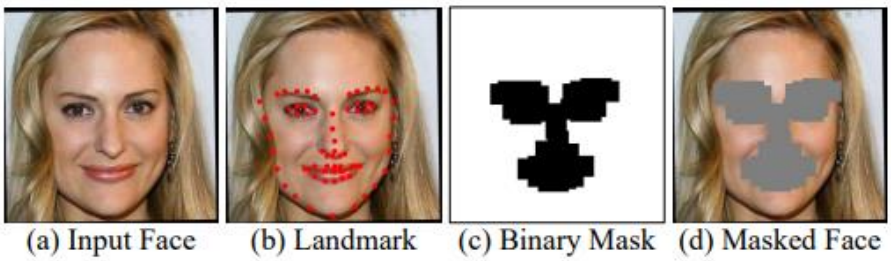
Στο [5] οι συγγραφείς προσεγγίζουν το πρόβλημα χρησιμοποιώντας δύο δίκτυα, ένα δίκτυο αναγνώρισης συμπεριφοράς (Behavior-Net) που στηρίζεται στην κίνηση του κεφαλιού και στις εκφράσεις του προσώπου και ένα δίκτυο αναγνώρισης της ταυτότητας του προσώπου (VGG)[6] το οποίο χρησιμοποιεί συνηθισμένες τεχνικές αναγνώρισης προσώπων. Όπως και παραπάνω, η βασική ιδέα πίσω από τα δύο δίκτυα είναι ότι στα παραποιημένα βίντεο η συμπεριφορά του προσώπου είναι η ίδια με του αρχικού βίντεο και συνεπώς η ταυτότητα που βασίζεται στην συμπεριφορά είναι η ίδια αλλά η ταυτότητα του προσώπου που βασίζεται στα χαρακτηριστικά του είναι διαφορετική. Συγκρίνοντας επομένως τις δύο ταυτότητες οι συγγράφεις μπορούν να καταλήξουν εάν ένα βίντεο έχει υποστεί τροποποίηση ή όχι. Και τα δύο δίκτυα πρόκειται για CNN, όπου στο Behavior-Net η είσοδος είναι διανύσματα της μορφής $X \in R^{d*t}$, όπου t ο αριθμός των καρέ του βίντεο και d διάνυσμα 256 διατάσεων που αναπαριστά την πόζα, την έκφραση και τα ορόσημα του προσώπου όπως αυτό παράγεται από το δίκτυο κωδικοποιητή Fab-Net (Facial Attributes-Net)[7].



Εικόνα 2: Επισκόπηση του τρόπου λειτουργίας του δικτύου στο [5]

Στις δύο προηγούμενες ερευνητικές εργασίες η εκπαίδευση των μοντέλων πραγματοποιήθηκε στο σετ δεδομένων VoxCeleb2[9] , το οποίο περιλαμβάνει περισσότερα από 150.000 βίντεο από συνολικά 5994 διαφορετικά πρόσωπα. Αντίθετα στην έρευνα των Shruti et al.[10] το προτεινόμενο μοντέλο για την αναγνώριση τροποποιημένων βίντεο εκπαιδεύτηκε κάθε φορά για ένα μόνο πρόσωπο ενδιαφέροντος (Person of Interest – POI) με ένα μεγάλο αριθμό βίντεο 10 δευτερολέπτων. Πρόσωπα ενδιαφέροντος ήταν παγκόσμιοι ηγέτες, μεταξύ των οποίων οι Barack Obama, Hillary Clinton, Bernie Sanders, κ.λπ. καθώς η παραποίηση των βίντεο τέτοιων προσώπων δύναται να έχει τεράστιες συνέπειες. Στο πλαίσιο αυτό εκπαιδεύθηκε ένα GAN που χρησιμοποιεί Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs)[11] με χαρακτηριστικά του συντελεστές συσχέτισης Pearson μεταξύ 20 μονάδων δράσης προσώπου (Facial Action Units - AU)[12]. Ως AU θεωρούνται οι κωδικοποιήσεις των κινήσεων των μυών του προσώπου, όπως εσωτερική ανύψωση φρυδιού, εξωτερική ανύψωση φρυδιού, ανοιγόκλεισμα ματιών κ.λπ. Από τους 190 συντελεστές για τα ζευγάρια των χαρακτηριστικών ($\frac{20*19}{2} = 190$), μόνο τα 29 χρησιμοποιήθηκαν καθώς η καλύτερη απόδοση του μοντέλου παρατηρήθηκε για τον αριθμό αυτό.

Οι Dong et al.[21] στηρίζονται επίσης στην ταυτότητα των προσώπων για την αναγνώριση των deepfakes. Στις σύγχρονες μεθόδους παραγωγής deepfakes μόνο το εσωτερικό του προσώπου παραποιείται ενώ το εξωτερικό παραμένει αυτό του πραγματικού προσώπου. Βασισμένοι σε αυτό, οι συγγραφείς εντοπίζουν τα ορόσημα του προσώπου και με μία μάσκα αφαιρούν το εσωτερικό του προσώπου αφήνοντας μόνο το εξωτερικό (παρακάτω εικόνα). Ακολούθως εκπαιδεύουν ένα δίκτυο αναγνώρισης ταυτότητας που δημιουργεί ένα διάνυσμα ταυτότητας (512 διαστάσεων) που στηρίζεται στην αρχιτεκτονική MobileNet[22] με τις κομμένες εικόνες μόνο από αληθινά πρόσωπα. Για την κατηγοριοποίηση μιας εικόνας ως παραποιημένης η όχι, δημιουργείται το διάνυσμα ταυτότητας μιας εικόνας αναφοράς και της υπό εξέτασης εικόνας και ελέγχεται η μεταξύ τους διαφορά. Παρατηρήθηκε ότι η απόδοση του δικτύου που εκπαιδεύτηκε με τις κομμένες εικόνες ήταν αρκετά καλύτερη από όταν εκπαιδεύτηκε με τις αρχικές εικόνες.

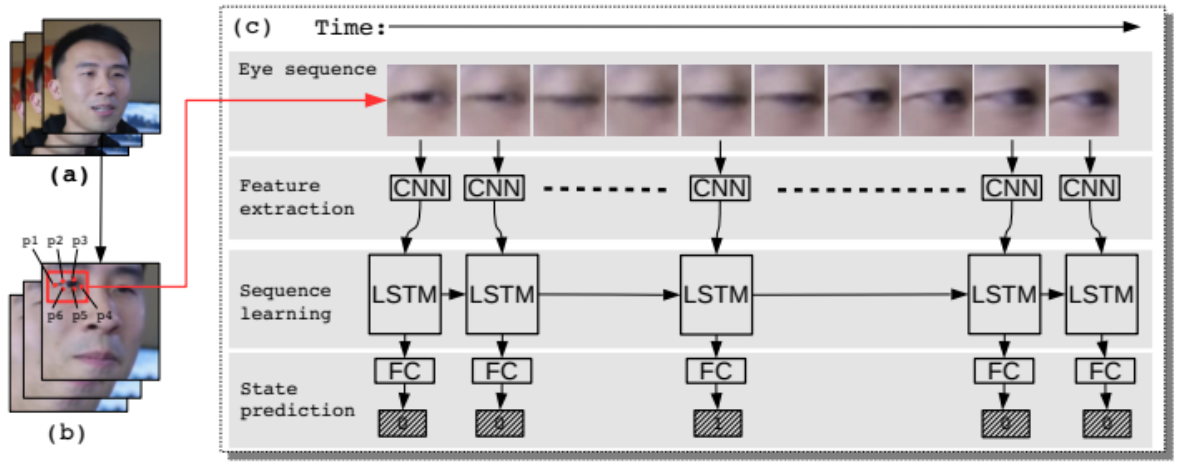


Εικόνα 3: Διαδικασία δημιουργίας των εικόνων που χρησιμοποιούνται ως είσοδοι στο [21]

Σε μία πιο διαφορετική προσέγγιση [13], οι συγγραφείς, για να εντοπίσουν πότε έχει παραποιηθεί κάποιο βίντεο, στηρίζονται στην παρατήρηση ότι για να παραχθούν οι ήχοι που αντιστοιχούν στους φθόνους μι (m), πι (p) και μπι (b) το στόμα του ομιλητή θα πρέπει να είναι τελείως κλειστό. Έτσι, εξάγοντας σε κείμενο τα λεγόμενα του ομιλητή και εντοπίζοντας τις χρονικές στιγμές που εμφανίζονται οι παραπάνω ήχοι, αναλύονται έξι καρέ κοντά στην εμφάνιση του φαινομένου για να διαπιστωθεί αν το στόμα του ομιλητή είναι πράγματι

κλειστό. Τρεις μέθοδοι χρησιμοποιούνται για την αναγνώριση, χειροκίνητη αναγνώριση όπου αναλυτές κατηγοριοποιούν τα καρέ ως κλειστό ή ανοιχτό στόμα και 2 μέθοδοι αυτόματης αναγνώρισης. Στην πρώτη χρησιμοποιούνται τα ορόσημα του προσώπου για να βρεθεί η θέση των χειλιών και με ένα προφίλ έντασης στην μέση των χειλιών αναγνωρίζεται αν το στόμα είναι ανοιχτό ή όχι. Στην δεύτερη χρησιμοποιείται ένα CNN δίκτυο με την αρχιτεκτονική Xception[14] όπου με χειροκίνητα κατηγοριοποιημένες εικόνες με ανοιχτό και κλειστό στόμα του Barack Obama εκπαιδεύεται ένας κατηγοροποιητής που θα αναγνωρίζει αν το στόμα σε μια φωτογραφία είναι ανοιχτό (κλάση 0) ή κλειστό (κλάση 1).

Ένα ακόμη χαρακτηριστικό που μπορεί να βοηθήσει στην αναγνώριση των deepfakes είναι το ανοιγόκλεισιμα των ματιών και, πιο συγκεκριμένα, η συχνότητά τους. Δεδομένου ότι τα σετ δεδομένων με τα οποία εκπαιδεύονται οι μέθοδοι deepfake έχουν ένα μικρό αριθμό εικόνων με κλειστά τα μάτια, είναι αναμενόμενο η συχνότητα που ανοιγοκλείνουν τα μάτια σε deepfake βίντεο να είναι μικρότερη από αυτή των αυθεντικών βίντεο. Για αυτό το λόγο στην έρευνα των Li et al.[15] προτείνεται ένα μοντέλο αναγνώρισης του ανοιγοκλεισίματος των ματιών που βασίζεται στην μακροπρόθεσμη επαναληπτικότητα. Αρχικά εξάγεται μόνο η περιοχή των ματιών από τα αυθεντικά βίντεο και εκπαιδεύεται ένα CNN ώστε να αναγνωρίζει πότε τα μάτια είναι κλειστά και πότε ανοικτά. Από το CNN αφαιρούνται τα δύο τελευταία στρώματα και στην συνέχεια το δίκτυο χρησιμοποιείται ως είσοδος σε ένα επαναληπτικό νευρωνικό δίκτυο (Recursive Neural Network-RNN) με επίπεδα μακράς βραχύχρονης μνήμης (Long Short Term Memory-LSTM)[16]. Η επιλογή του δικτύου αυτού οφείλεται στο γεγονός ότι το ανοιγοκλείσιμο των ματιών παρουσιάζει μεγάλη χρονική εξάρτηση. Συνεπώς στην παραπάνω αρχιτεκτονική το δίκτυο CNN χρησιμοποιείται για την επιλογή των χαρακτηριστικών (feature selection) και το classification πραγματοποιείται στο LSTM-RNN δίκτυο. Η αρχιτεκτονική που χρησιμοποιήθηκε παρουσιάζεται στην παρακάτω εικόνα.

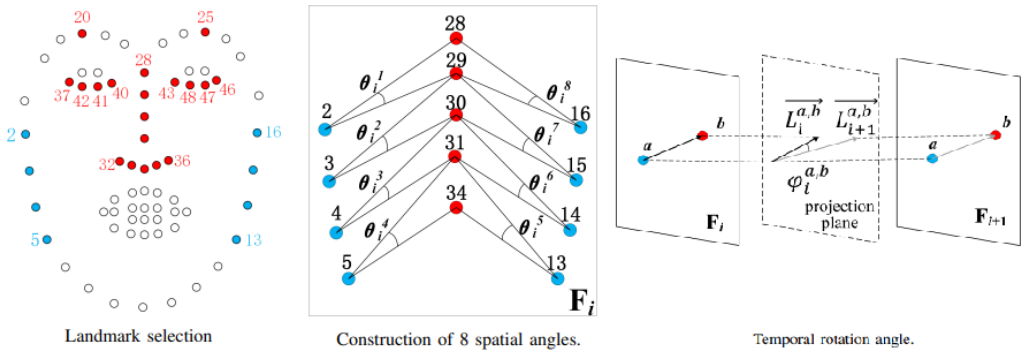


Εικόνα 4: Αρχιτεκτονική του Δικτύου Αναγνώρισης Ανοιγοκλεισίματος ματιών

Οι Nguyen et al.[25] κατηγοριοποιούν τα βίντεο ως αυθεντικά ή παραποιημένα με την βοήθεια βιομετρικών χαρακτηριστικών και, πιο συγκεκριμένα, των φρυδιών. Στο πλαίσιο αυτό εκπαιδεύουν μοντέλα διαφόρων αρχιτεκτονικών πάνω σε εικόνες φρυδιών από καρέ αυθεντικών βίντεο και βασισμένοι στην ομοιότητα συνημίτονου μεταξύ των διανυσμάτων των χαρακτηριστικών της εικόνας αναφοράς και της υπό εξέτασης εικόνας, κατηγοριοποιούν τα πρόσωπα ως παραποιημένα ή όχι.

Στο [17] χρησιμοποιούνται οι ανακρίβειες στην πόζα των προσώπων για την αναγνώριση των deepfakes. Κατά την δημιουργία των deepfakes υπολογίζονται τα ορόσημα προσώπου τόσο του αυθεντικού προσώπου όσο και αυτού με το οποίο θα παραποιηθεί και στην συνέχεια τοποθετείται το ένα πρόσωπο πάνω στο άλλο ελαχιστοποιώντας την απόσταση μεταξύ των οροσήμων του κέντρου του προσώπου. Ωστόσο τα ορόσημα του εξωτερικού προσώπου παραμένουν αναλλοίωτα με αποτέλεσμα η εκτίμηση της πόζας του προσώπου που γίνεται με βάση τα ορόσημα όλου του προσώπου να διαφέρει αρκετά από την εκτίμηση που βασίζεται στα ορόσημα του κέντρου του προσώπου. Την διαφορά αυτή από τις δύο εκτιμώμενες πόζες χρησιμοποιούν οι συγγραφείς για να εκπαιδεύσουν ένα δίκτυο SVM για την κατηγοριοποίηση αυθεντικών εικόνων και deepfakes.

Ωστόσο στο [17] η τρισδιάστατη πόζα εκτιμάται από δισδιάστατα πρόσωπα οπότε παρατηρούνται ανακρίβειες και επίσης αξιοποιείται μόνο η χωρική πληροφορία και όχι η χρονική αφού κάθε καρέ μελετάται ξεχωριστά. Λύση στα προβλήματα του [17] προσπαθούν να δώσουν οι συγγραφείς του [20] οι οποίοι χρησιμοποιούν τα ορόσημα του προσώπου τόσο χωρικά όσο και χρονικά. Αρχικά εντοπίζουν τα ορόσημα που έχουν την μικρότερη μετακίνηση κατά την διάρκεια των βίντεο (τονίζονται ως μπλε και κόκκινα στο παρακάτω σχήμα). Από αυτά τα μπλε χρησιμοποιούνται ως σημεία αναφοράς. Ως χωρικά χαρακτηριστικά χρησιμοποιούνται είναι οι γωνίες που σχηματίζονται μεταξύ δύο γειτονικών διανυσμάτων με αρχή το ίδιο σημείο αναφοράς (θ_i^a), όπως φαίνεται στο 2ο σχήμα στην παρακάτω εικόνα. Χρονικά χαρακτηριστικά θεωρούνται οι γωνίες που σχηματίζονται μεταξύ των ίδιων διανυσμάτων σε διαδοχικά καρέ ($\varphi_i^{a,b}$) (3ο σχήμα). Τελικά εκπαιδεύεται ένα SVM με χαρακτηριστικά τη μέση τιμή, την διακύμανση, την ασσυμετρία, την κύρτωση και τον συντελεστή αυτοσυσχέτισης πρώτου βαθμού των χωρικών και χρονικών χαρακτηριστικών που περιεγράφηκαν παραπάνω υπολογισμένα για κάθε βίντεο.



Εικόνα 5: Τα ορόσημα του προσώπου, οι χωρικές και οι χρονικές γωνίες που χρησιμοποιούνται στο [20]

Επίσης βασισμένοι στην παρατήρηση ότι στις μεθόδους παραγωγής deepfakes μόνο το εσωτερικό του προσώπου παραποιείται, οι Nirkin et al.[23] προσπαθούν να βελτιώσουν την απόδοση των ήδη υπάρχον δικτύων αναγνώρισης deepfakes προσθέτοντας μια αρχιτεκτονική αναγνώρισης της ταυτότητας. Η αρχιτεκτονική αυτή αποτελείται από δύο δίκτυα τύπου Xception που εκπαιδεύονται για να αναγνωρίζουν την ταυτότητα 8631 ατόμων από το σετ δεδομένων VGGFace2[24], με το ένα δίκτυο να αναγνωρίζει την ταυτότητα έχοντας ως είσοδο μόνο το κεντρικό μέρος του προσώπου και το άλλο δίκτυο έχοντας ως είσοδο το γύρω μέρος του προσώπου με κομμένο το κεντρικό μέρος. Ταυτόχρονα εκπαιδεύονται άλλα δύο δίκτυα, ένα για να κάνει κατηγοριοποίηση μεταξύ αυθεντικών και παραποιημένων εικόνων που έχουν δημιουργηθεί με την μέθοδο εναλλαγής προσώπων και ένα για να κάνει κατηγοριοποίηση εικόνων που δεν έχει μεταβληθεί η ταυτότητα του προσώπου αλλά η πόζα και η έκφραση του. Οι έξοδοι όλων των δικτύων τροφοδοτούν ένα πλήρως συνδεδεμένο δίκτυο το οποίο χρησιμοποιείται για την τελική κατηγοριοποίηση των εικόνων. Το τροποποιημένο αυτό δίκτυο Xception παρουσιάζει καλύτερη απόδοση από το αρχικό.

Μια ακόμη προσέγγιση για την επίλυση του προβλήματος βασίζεται στον καρδιακό παλμό. Κάνοντας χρήση της εξ αποστάσεως οπτικής φωτοπληθυσμογραφίας (PPG) καθίσταται δυνατό να εντοπίζεται ο καρδιακός παλμός μέσω μικρών διαφορών στο χρώμα του προσώπου σε ένα βίντεο. Ο καρδιακός παλμός δεν θα είναι φυσιολογικός ή δεν θα υπάρχει καθόλου σε βίντεο που έχουν υποστεί παραποίηση. Στηριζόμενοι στην παρατήρηση αυτή, οι Qi et al.[18] ανέπτυξαν το δίκτυο DeepRythm το οποίο στηρίζεται στην χωροχρονική αναπαράσταση των καρδιακών παλμών (Spatial-Temporal Representation-STR)[19] για την κατηγοριοποίηση των βίντεο. Εξάγοντας το πρόσωπο από τα βίντεο χρησιμοποιώντας τα ορόσημα του προσώπου, γίνεται μεγέθυνση της κίνησης του προσώπου (τονίζονται σε κάθε καρέ τα σημεία του προσώπου στα οποία παρατηρούνται μικροκινήσεις) και δημιουργείται η χωροχρονική αναπαράσταση των βίντεο. Στο χωρικό κομμάτι του δικτύου, ένα CNN εκπαιδεύεται για να ξεχωρίζει τις χωρικές διαφορές του προσώπου που οφείλονται π.χ. σε διαφορετικό φωτισμό. Στο χρονικό κομμάτι εκπαιδεύεται ένα LSTM-RNN δίκτυο που εντοπίζει ποια καρέ του βίντεο οφείλονται κυρίως για την κατηγοριοποίησή του ως τροποποιημένο. Τα δύο δίκτυα εκπαιδεύονται ταυτόχρονα με ένα ακόμη CNN δίκτυο και μαζί αποτελούν το DeepRythm δίκτυο.

1.3 Σκοπός της Διπλωματικής

Σκοπός της παρούσας διπλωματικής είναι η ανάπτυξη μιας μεθόδου αναγνώρισης deepfakes προσώπων μέσω αποτύπωσης των 3D βιομετρικών χαρακτηριστικών του προσώπου, όπως το γεωμετρικό σχήμα, ο τρόπος έκφρασης, η πόζα, το χρώμα κ.τ.λ. Για να επιτευχθεί η αναγνώριση στον 3D χώρο χωρίς χρήση αυθεντικών 3D δεδομένων, θα χρησιμοποιηθούν 3D Μορφοποιήσημα μοντέλα (3D Morphable Models – 3DMMs).

Η μέθοδος αυτή θα λειτουργεί έχοντας ως είσοδο ένα βίντεο και θα το κατηγοριοποιεί ως αυθεντικό η παραποιημένο. Η μέθοδος θα λαμβάνει ακόμη υπόψη και τις συνδέσεις μεταξύ διαδοχικών καρέ ενός βίντεο καθώς οι περισσότερες μέθοδοι δημιουργίας deepfakes λειτουργούν σε ένα μόνο καρέ την φορά και δεν λαμβάνουν υπόψη την χρονική συνέχεια σε ένα βίντεο, οπότε πιθανόν να υπάρχουν ασυνέχειες μεταξύ των καρέ ενός βίντεο που θα οδηγήσουν στην κατηγοριοποίησή του.

Τελικός σκοπός της εργασίας είναι η ανάπτυξη μιας εύρωστης και με καλή γενίκευση, όσον αφορά τις διάφορες τεχνικές παραποίησης της εικόνας του προσώπου, μεθόδου αναγνώρισης deepfakes που θα στηρίζεται τόσο σε χωρικά όσο και σε χρονικά χαρακτηριστικά και θα χρησιμοποιεί 3DMM και όχι απλές εικόνες ή βίντεο για την αναγνώριση των παραποιημένων προσώπων.

1.4 Δομή της Εργασίας

Η παρούσα εργασία οργανώνεται στα ακόλουθα κεφάλαια:

- Κεφάλαιο 1: Στο παρόν κεφάλαιο πραγματοποιήθηκε η εισαγωγή στο πρόβλημα που καλείται να λύσει η εργασία και παρουσιάστηκαν μέθοδοι που έχουν προταθεί στην βιβλιογραφία για την επίλυση του προβλήματος.
- Κεφάλαιο 2: Παρουσιάζονται οι βασικές θεωρητικές γνώσεις που είναι απαραίτητες για την κατανόηση του κειμένου.
- Κεφάλαιο 3: Περιγράφονται οι μέθοδοι που θα χρησιμοποιηθούν στην προσπάθεια επίλυσης του προβλήματος. Επίσης περιγράφεται η αρχιτεκτονική των μεθόδων καθώς και ο τρόπος εκπαίδευσής τους.
- Κεφάλαιο 4: Παρουσιάζονται τα αποτελέσματα των μεθόδων καθώς και το σύνολο πειραμάτων που πραγματοποιήθηκαν για την τελευταία μέθοδο.
- Κεφάλαιο 5: Γίνεται σύνοψη των ευρημάτων της εργασίας και παρατίθενται προτάσεις για μελλοντική έρευνα.

Κεφάλαιο 2 Θεωρητικό Υπόβαθρο

2.1 Αναπαράσταση των προσώπων ως διανύσματα

Η παραμετροποίηση του ανθρώπινου προσώπου αποτελεί ένα απαιτητικό πρόβλημα που συναντάται από τις απαρχές της Γραφικής με υπολογιστές (Computer Graphics). Μία από τις διάφορες μεθόδους που έχει προταθεί και είναι πλέον αρκετά διαδεδομένη είναι τα 3D Μορφοποιήσημα Μοντέλα (**3D morphable models – 3DMMs**).

2.1.1 3D morphable models (3DMMs) – Αρχική Προσέγγιση

Τα 3DMMs αρχικά προτάθηκαν από τους *Blanz et al.* [30]. Τα μορφοποιήσημα μοντέλα βασίστηκαν σε ένα σετ από 3D δεδομένα σαρωμένων προσώπων. Η γεωμετρία του προσώπου παριστάνεται με το διάνυσμα $\boldsymbol{S} = (X_1, Y_1, Z_1, \dots, X_n, Y_n, Z_n) \in \mathcal{R}^{3n}$ όπου X, Y, Z οι συντεταγμένες των n κορυφών. Το χρώμα του προσώπου παριστάνεται από το διάνυσμα $\boldsymbol{T} = (R_1, G_1, B_1, \dots, R_n, B_n, G_n) \in \mathcal{R}^{3n}$ όπου R, G, B οι τιμές των χρωμάτων των n αντίστοιχων κορυφών. Βρίσκοντας τα διανύσματα $\boldsymbol{S_i}, \boldsymbol{T_i}$ για καθένα από τα m σκαναρισμένα πρόσωπα, δημιουργήθηκε ένα μορφοποιήσημο μοντέλο προσώπων όπου κάθε καινούργιο πρόσωπο μπορεί να αναπαρασταθεί από των γραμμικό συνδυασμό των διανυσμάτων $\boldsymbol{S_i}, \boldsymbol{T_i}$, δηλαδή ένα οποιοδήποτε πρόσωπο μπορεί να αναπαρασταθεί ως:

$$(\boldsymbol{S}_{mod}(\vec{\boldsymbol{a}}), \boldsymbol{T}_{mod}(\vec{\boldsymbol{b}})), \quad \text{όπου}$$

$$\boldsymbol{S}_{mod} = \sum_{i=1}^{m} a_i S_i, \qquad \boldsymbol{T}_{mod} = \sum_{i=1}^{m} b_i T_i, \qquad \sum_{i=1}^{m} a_i = \sum_{i=1}^{m} b_i = 1 \text{ και}$$

$$\vec{\boldsymbol{a}} = (\alpha_1, \alpha_2, \dots, \alpha_m)^T, \ \vec{\boldsymbol{b}} = (b_1, b_2, \dots, b_m)^T$$

οι μεταβλητές που ελέγχουν το σχήμα και το χρώμα αντίστοιχα.

Τα διανύσματα $\vec{a}$, $\vec{b}$ δεν έχουν κάποια φυσική σημασία ως προς τα χαρακτηριστικά του προσώπου. Για να είναι δυνατή η έκφραση των χαρακτηριστικών του προσώπου δημιουργήθηκαν τα διανύσματα σχήματος ΔS και χρώματος ΔT τα οποία επηρεάζουν ένα συγκεκριμένο χαρακτηριστικό του προσώπου αν προστεθούν ή αφαιρεθούν από τα διανύσματα S και T αντίστοιχα.

Τα διανύσματα υπολογίστηκαν συγκρίνοντας τα διανύσματα ενός προσώπου σε ουδέτερη πόζα και του ίδιου προσώπου σε συγκεκριμένη έκφραση, δηλαδή:

$$\Delta S = S_{expression} - S_{neutral}, \qquad \Delta T = T_{expression} - T_{neutral}$$

Συνεπώς με την πρόσθεση, την αφαίρεση ή τον γραμμικό συνδυασμό αυτών το διανυσμάτων στα υπάρχοντα του προσώπου μπορούν να αλλάξουν συγκεκριμένες εκφράσεις ή χαρακτηριστικά του προσώπου.

Εικόνα 6: Από αριστερά προς τα δεξιά: Το πρόσωπο που αντιστοιχεί σε μία ταυτότητα με ουδέτερη έκφραση, με χαμόγελο, συνοφρυωμένο, με πιο αρρενωπή εμφάνιση

Οι συντελεστές που χρησιμοποιούνται, ωστόσο, από αυτό το μοντέλο δεν έχουν ιδιαίτερη φυσική σημασία. Επίσης η έκφραση του προσώπου δεν περιγράφεται από τους συντελεστές παρά από πρόσθεση και αφαίρεση των διανυσμάτων ΔS και ΔT. Έτσι έχουν προταθεί και άλλες μέθοδοι αναπαράστασης του προσώπου ως διανύσματα. Ένα από αυτά τα μοντέλα, που διαχωρίζει τα διανύσματα ώστε να αναπαριστούν ταυτότητα, πόζα και έκφραση του προσώπου είναι το **FLAME**[31].

2.1.2 FLAME - Faces Learned with an Articulated Model and Expressions

Το FLAME είναι ένα μοντέλο που συνδυάζει διανύσματα ταυτότητας, πόζας και έκφρασης με χρήση linear blend skinning (**LBS**, μέθοδος μετατροπής των κορυφών μέσα σε ένα πλέγμα μέσω πολλαπλών μετατροπών) και διορθωτικές μίξεις (**blendshapes**) για την πόζα ώστε να αποτυπώσει το σαγόνι, το λαιμό και τα μάτια. Δοσμένου των παραμέτρων της ταυτότητας του προσώπου $\boldsymbol{\beta} \in \mathcal{R}^{|\boldsymbol{\beta}|}$, της πόζας $\boldsymbol{\theta} \in \mathcal{R}^{3k+3}$ (όπου $k = 4$ αρθρώσεις για το σαγόνι, το λαιμό και τα μάτια) και των παραμέτρων της έκφρασης $\boldsymbol{\psi} \in \mathcal{R}^{|\boldsymbol{\psi}|}$, το FLAME δημιουργεί ένα πλέγμα (**mesh**) από $n = 5023$ κορυφές. Το μοντέλο περιγράφεται από την εξίσωση :

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = W(T_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}), \boldsymbol{J}(\boldsymbol{\beta}), \boldsymbol{\theta}, \boldsymbol{w})$$

όπου $W(\boldsymbol{T}, \boldsymbol{J}, \boldsymbol{\theta}, \boldsymbol{w})$ η συνάρτηση blend skinning που περιστρέφει τις κορυφές στο $\boldsymbol{T} \in \mathcal{R}^{3n}$ γύρω από τις αρθρώσεις του σκελετού $\boldsymbol{J} \in \mathcal{R}^{3n}$ και $\boldsymbol{w} \in \mathcal{R}^{kxn}$ τα **linear skinning weights**(καθορίζουν ποιες αρθρώσεις του σκελετού μπορούν να επηρεάσουν μία κορυφή) που ομαλοποιούν γραμμικά την έξοδο του μοντέλου. Οι θέσεις των αρθρώσεων ορίζονται ως συνάρτηση της ταυτότητας β. Το πρότυπο mesh T στην αρχική πόζα είναι το εξής:

$$T_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \boldsymbol{T} + B_s(\boldsymbol{\beta}; \boldsymbol{S}) + B_p(\boldsymbol{\theta}; \boldsymbol{P}) + B_E(\boldsymbol{\psi}; \mathcal{E})$$

όπου έχουν προστεθεί τα blendshapes για το σχήμα $B_s(\boldsymbol{\beta}; \boldsymbol{S}) : \mathcal{R}^{|\boldsymbol{\beta}|} \rightarrow \mathcal{R}^{3n}$, για την διόρθωσή πόζας $B_p(\boldsymbol{\theta}; \boldsymbol{P})$: $\mathcal{R}^{3\kappa+3} \rightarrow \mathcal{R}^{3n}$ και blendshapes για την έκφραση $B_E(\boldsymbol{\psi}; \mathcal{E})$: $\mathcal{R}^{|\boldsymbol{\psi}|} \rightarrow \mathcal{R}^{3n}$. Πιο αναλυτικά:

Shape Blendshapes:

Οι μεταβολές στο σχήμα του προσώπου που οφείλονται σε διαφορετικές ταυτότητες μοντελοποιούνται από γραμμικά blendshapes ως εξής:

$$B_s(\boldsymbol{\beta}; \boldsymbol{S}) = \sum_{n=1}^{|\boldsymbol{\beta}|} \beta_n \boldsymbol{S_n},$$

όπου $\boldsymbol{\beta} = \left[\beta_1, \beta_2, \dots, \beta_{|\boldsymbol{\beta}|}\right]^T$ οι συντελεστές για το σχήμα και $\boldsymbol{S} = \left[S_1, \dots, S_{|\boldsymbol{S}|}\right]^{\boldsymbol{T}}$ η ορθοκανονική βάση για το σχήμα που δημιουργείται κατά την εκπαίδευση του μοντέλου.

Pose Blendshapes:

Η συνάρτηση των blendshapes της πόζας ορίζεται ως:

$$B_p(\boldsymbol{\theta}; \boldsymbol{P}) = \sum_{n=1}^{9k} \left(R_n(\boldsymbol{\theta}) - R_n(\boldsymbol{\theta}^\star)\right) \boldsymbol{P_n},$$

όπου $R(\boldsymbol{\theta})$: $R^{|\boldsymbol{\theta}|} \rightarrow \mathcal{R}^{9k}$ η συνάρτηση που μετατρέπει το διάνυσμα θ σε ένα διάνυσμα που περιέχει συναθροισμένα τα στοιχεία από όλους τους πίνακες περιστροφής, $\boldsymbol{P_n}$ το διάνυσμα που περιγράφει την διαφορά στις τιμές από αυτές που υπολογίστηκαν από τα R_n και $\boldsymbol{P} = [P_1, \dots, P_{9k}] \epsilon R^{3nx9k}$ είναι ο πίνακας που περιέχει όλα τα blendshapes πόζας.

Expression Blendshapes:

Η συνάρτηση των blendshapes της έκφρασης ορίζεται ως:

$$B_E(\boldsymbol{\psi}; \mathcal{E}) = \sum_{n=1}^{|\boldsymbol{\psi}|} \psi_n \boldsymbol{E_n},$$

όπου $\boldsymbol{\psi} = \left[\psi_1, \psi_2, \dots, \psi_{|\boldsymbol{\psi}|}\right]^T$ οι συντελεστές για την έκφραση του προσώπου και $\mathcal{E} = \left[\boldsymbol{E_1}, \boldsymbol{E_2}, \dots, \boldsymbol{E}_{|\boldsymbol{\psi}|}\right] \boldsymbol{\epsilon}\, R^{3nx|\boldsymbol{\psi}|}$ η ορθοκανονική βάση για την έκφραση.

Πριν από την εκπαίδευση του μοντέλου πρέπει να γίνει αρχικοποίηση των τιμών του. Έτσι δημιουργείται ένα αρχικό μοντέλο όπου οι παράμετροι για το σχήμα, την πόζα και την έκφραση αρχικοποιούνται είτε με τη βοήθεια καλλιτεχνών είτε από προηγούμενές ερευνητικές εργασίες. Επίσης τα δεδομένα εκπαίδευσης θα πρέπει να βρίσκονται σε πλήρη αντιστοιχία κορυφών (**vertex correspondence)**. Στο πλαίσιο αυτό, με τη μέθοδο της χρονικής καταχώρησης (**Temporal Registration),** για κάθε αλληλουχία από τρισδιάστατες σαρώσεις προσώπων, υπολογίζεται ένα ευθυγραμμισμένο πρότυπο $\boldsymbol{T_i} \in R^{3N}$, όπου i η κάθε σάρωση. Για την καταχώρηση ενός απλού καρέ (**single-frame registration**) αρχικά υπολογίζονται οι συντελεστές του μοντέλου που εξηγούν καλύτερα το σαρωμένο πρόσωπο μέσω βελτιστοποίησης της συνάρτησης:

$$E(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = E_D + \lambda_L E_L + E_P, \quad \text{όπου}$$

$$E_D = \lambda_D \sum_{V_s} p\left(\min\left||v_s - v_m|\right|\right)$$

ο ορός που μετράει την απόσταση από τις σκαναρισμένες κορυφές v_s στο πλησιέστερο σημείο της επιφάνειας του μοντέλου, λ_D το βάρος που ελέγχει την επίδραση και ρ η συνάρτηση ποινής German-McClure[32]. Ο όρος E_D μετράει την $l2$ κανονικοποιημένη απόσταση μεταξύ των ορόσημων της εικόνας και των αντίστοιχων κορυφών στο πρότυπο μοντέλο και ο όρος E_P κανονικοποιεί τους συντελεστές πόζας, σχήματος και έκφρασης. Επίσης βελτιστοποιείται και το φωτομετρικό σφάλμα ανάμεσα στην πραγματική εικόνα και στην εικόνα που δημιουργείται από το μοντέλο μετά από render του μοντέλου μαζί με το χρώμα (texture).

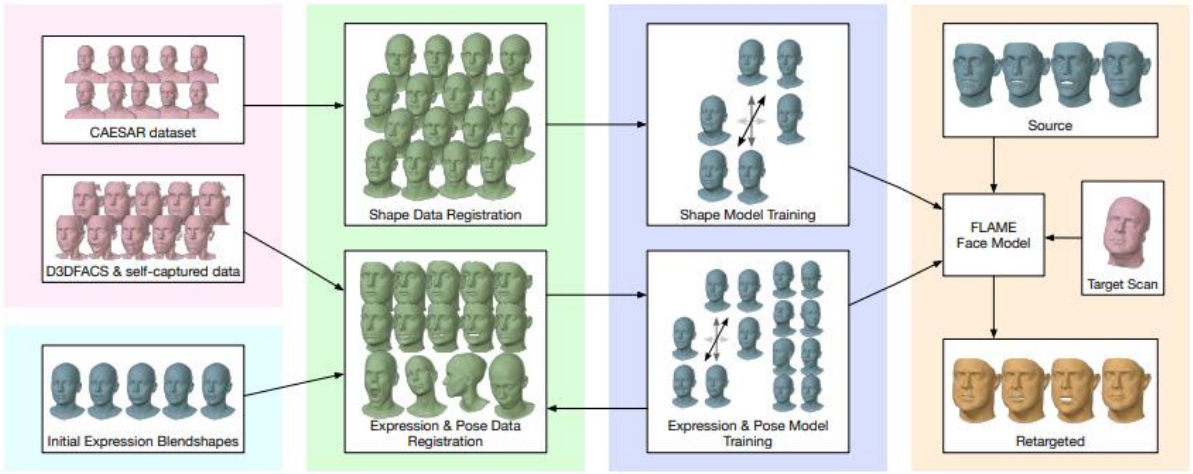
Για να γίνει τελικά η χρονική καταχώρηση, για κάθε ταυτότητα που υπάρχει στη βάση δεδομένων δημιουργείται ένα εξατομικευμένο πρότυπο. Έτσι το γενικό πρότυπο του μοντέλου αντικαθίσταται από το εξατομικευμένο πρότυπο το οποίο και χρησιμοποιείται για την εκπαίδευση με μια αλληλουχία από μια συγκεκριμένη ταυτότητα.

Κατά τη διαδικασία της εκπαίδευσης οι παράμετροι της πόζας $\{\boldsymbol{P}, \boldsymbol{w}, \boldsymbol{J}\}$, της έκφρασης $\mathcal{E}$ και του σχήματος $\{\boldsymbol{T}, \boldsymbol{S}\}$ βελτιστοποιούνται διαδοχικά, ένας κάθε φορά διατηρώντας τους άλλους δύο σταθερούς, με μια επαναληπτική διαδικασία που ελαχιστοποιεί το σφάλμα ανακατασκευής (reconstruction error).

Υπάρχουν 2 είδη παραμέτρων πόζας, παράμετροι που είναι συγκεκριμένοι για κάθε ταυτότητα $\{\boldsymbol{T_i^p}, \boldsymbol{J_i^p}\}$ και παράμετροι που υπάρχουν σε όλες τις ταυτότητες $\{\boldsymbol{P}, \boldsymbol{w}, \boldsymbol{J}\}$**.** Η εκπαίδευση για τις παραμέτρους πόζας $\boldsymbol{\vartheta}$ για κάθε καταχώρηση *j* γίνεται εναλλάσσοντας ανάμεσα στην εκπαίδευση για τις εξατομικευμένες παραμέτρους $\{\boldsymbol{T_i^p}, \boldsymbol{J_i^p}\}$ και τις γενικές παραμέτρους $\{\boldsymbol{P}, \boldsymbol{w}, \boldsymbol{J}\}$**.** Προκειμένου οι εξατομικευμένοι παράμετροι να μην επηρεάζονται από έντονες εκφράσεις του προσώπου, αυτές απομακρύνονται λύνοντας ταυτόχρονα για τις παραμέτρους πόζας και έκφρασης και στη συνέχεια αφαιρώντας τα blendshapes της έκφρασης και λύνοντας μόνο για τις παραμέτρους εξατομικευμένης πόζας.

Για να εκπαιδευτεί η ορθοκανονική βάση Ε για την έκφραση απαιτείται η έκφραση να διαχωριστεί από τις μεταβολές στην πόζα και στο σχήμα. Αυτό επιτυγχάνεται λύνοντας αρχικά για τις παραμέτρους πόζας και στη συνέχεια στην αφαίρεσή τους χρησιμοποιώντας τον αντίστροφο μετασχηματισμό.

Τέλος, για την εκπαίδευση των παραμέτρων του σχήματος του προσώπου, που αποτελούνται από το πρότυπο $\boldsymbol{T}$ και τα blendshapes του σχήματος $\boldsymbol{S}$, αφαιρούνται ανάλογα οι επιδράσεις των παραμέτρων της πόζας και της έκφρασης. Η διαδικασίες της χρονικής καταχώρησης και της εκπαίδευσης του μοντέλου παρουσιάζονται στην παρακάτω εικόνα.



Εικόνα 7: Οι μέθοδοι της χρονικής καταχώρησης και εκπαίδευσης του μοντέλου FLAME. Τα σετ δεδομένων τρισδιάστατων σαρώσεων και τα αρχικά blendshapes για την έκφραση (1ο στάδιο) χρησιμοποιούνται προκειμένου να γίνει η χρονική καταχώριση τόσο των δεδομένων σχήματος όσο και των δεδομένων έκφρασης και πόζας (2ο στάδιο). Κατά την διαδικασία της εκπαίδευσης (3ο στάδιο) εκπαιδεύονται διαδοχικά οι παράμετροι της έκφρασης, της πόζας και του σχήματος, αφαιρώντας κάθε φορά οι επιδράσεις των άλλων παραμέτρων. Τέλος με χρήση των εκπαιδευμένων παραμέτρων πραγματοποιείται μεταφορά έκφρασης στο target scan (4ο στάδιο).

2.2 Μετατροπή εικόνων σε διανύσματα

Στην παρούσα εργασία και για τον σκοπό της αναγνώρισης των παραποιημένων προσώπων μέσω 3D χαρακτηριστικών τους, χρησιμοποιούμε τις παραμέτρους ελέγχου του FLAME, των οποίων η βάση έχει σημασιολογική ερμηνεία καθώς εκφράζουν την 3D γεωμετρία του προσώπου μέσω του σχήματος (ταυτότητας) και της έκφρασης, καθώς και την υφή του προσώπου.

Το framework που επιλέχθηκε για την μετατροπή των εικόνων σε 3DMM είναι το DECA (Detailed Expression Capture and Animation)[33]. Το DECA μαθαίνει ένα μοντέλο μετατόπισης από εικόνες χωρίς επιτήρηση 2D-to-3D. Έτσι μαθαίνει να εξάγει από μία εικόνα λεπτομέρειες για το πρόσωπο, για την έκφραση, την πόζα, το σχήμα, την υφή και τον φωτισμό. Για την γεωμετρία του προσώπου χρησιμοποιείται το μοντέλο FLAME.

Οι παράμετροι από τις οποίες δημιουργείται το 3DMM είναι αυτές του FLAME, και πιο συγκεκριμένα οι παράμετροι της ταυτότητας του προσώπου $\boldsymbol{\beta} \in \mathcal{R}^{|\boldsymbol{\beta}|}$, της πόζας $\boldsymbol{\theta} \in \mathcal{R}^{3k+3}$ (όπου $k = 4$ αρθρώσεις για το σαγόνι, το λαιμό και τα μάτια) και οι παράμετροι της έκφρασης $\boldsymbol{\psi} \in \mathcal{R}^{|\boldsymbol{\psi}|}$. Για το σχήμα του προσώπου το DECA χρησιμοποιεί 100 παραμέτρους, για την έκφραση 50 και για την πόζα 6, δηλαδή $|\beta| = 100$ και $|\psi| = 50$.

Επειδή το FLAME δεν διαθέτει δικό του μοντέλο για την υφή του προσώπου, χρησιμοποιείται αυτό του Basel Face Model (BFM)[34], το οποίο χρησιμοποιεί χάρτες υφής (texture maps) ώστε να περιγράφει η υφή του προσώπου σε δισδιάστατη εικόνα. Στη εικόνα αυτή οι UV συντεταγμένες αντιστοιχίζουν την υφή στο συγκεκριμένο σημείο σε μία 3D κορυφή. Με αυτό τον τρόπο κάθε 3D κορυφή αντιστοιχεί σε κάποια UV συντεταγμένη. Έτσι ο γραμμικός υποχώρος υφής του BFM (linear albedo subspace) μετατρέπεται σε διάταξη UV συμβατή με το mesh του FLAME και δημιουργεί έναν UV χάρτη υφής (albedo map):

$$A(\boldsymbol{a}) \epsilon\, R^{dxdx3},\ \boldsymbol{a} \epsilon\, R^{50}.$$

Το DECA χρησιμοποιεί επίσης ένα ορθοκανονικό μοντέλο κάμερας για να προβάλει το τρισδιάστατο πλέγμα (3D mesh) στην εικόνα. Για τον φωτισμό των προσώπων το DECA χρησιμοποιεί σφαιρικές αρμονικές[35]. Όλα τα παραπάνω χρησιμοποιούνται για να ορίσουν την σκηνή (γεωμετρία, πόζα, υφή του προσώπου, φωτισμό και θέση της κάμερας) και να την αποτυπώσουν σε εικόνα.

Η εκπαίδευση του μοντέλου περιλαμβάνει τη κωδικοποίηση της εικόνας ενός προσώπου σε λανθάνουσες μεταβλητές (**latent variables**), την αποκωδικοποίηση αυτών σε εικόνα και την ελαχιστοποίηση των διαφορών μεταξύ των 2 εικόνων (της αρχικής και της ανακατασκευασμένης). Οι λανθάνουσες μεταβλητές περιλαμβάνουν τις παραμέτρους $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}$ του FLAME, τους συντελεστές υφής $\boldsymbol{\alpha}$, της κάμερας $\boldsymbol{c}$ και του φωτισμού $\boldsymbol{l}$. Το σφάλμα ανακατασκευής που καλείται να βελτιστοποιήσει το μοντέλο είναι το εξής:

$$L = L_{lmk} + L_{eye} + L_{pho} + L_{id} + L_{sc} + L_{reg},$$

όπου L_{lmk} το σφάλμα οροσήμου (landmark re-projection loss) που μετράει τη διαφορά μεταξύ των πραγματικών 2D ορόσημων του προσώπου της εικόνας και των αντίστοιχων οροσήμων που παράγει το FLAME και προβάλλονται στην εικόνα μέσω του μοντέλου της κάμερας.

L_{eye} το σφάλμα από το κλείσιμο των ματιών που υπολογίζει τη σχετική απόσταση των ορόσημων μεταξύ του πάνω και του κάτω μέρους του ματιού και συγκρίνει τις αποστάσεις ανάμεσα στην πραγματική εικόνα και τη προβολή του FLAME πάνω στην εικόνα μέσω του μοντέλου της κάμερας.

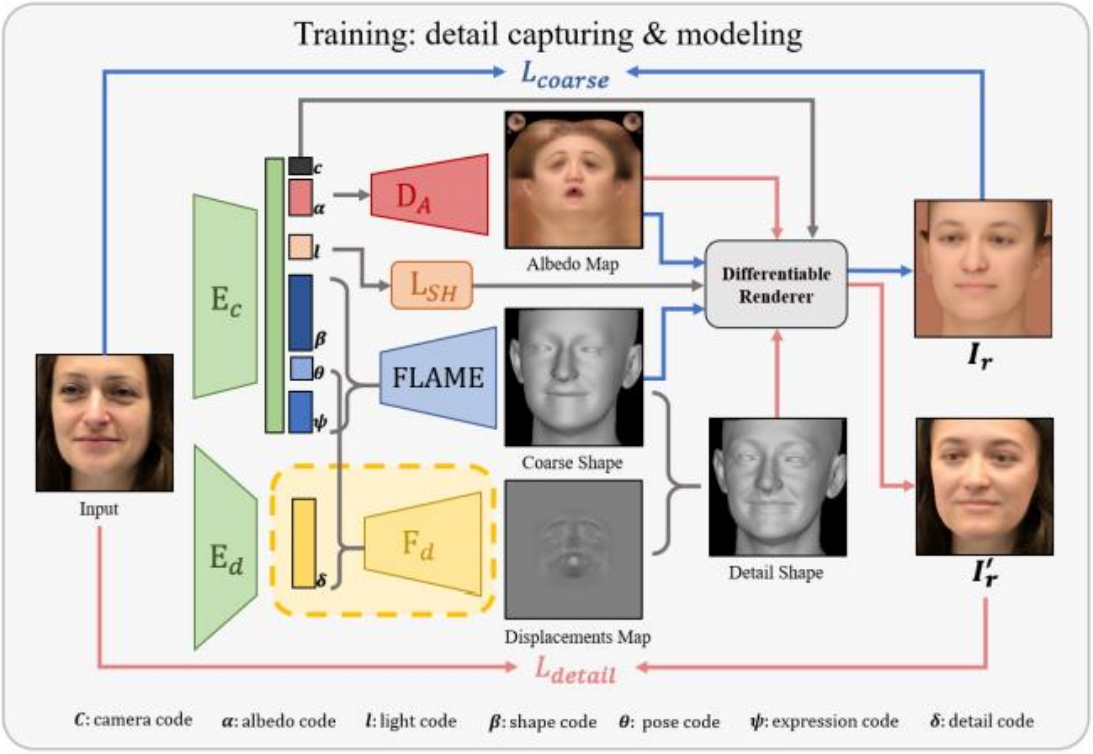
L_{pho} είναι το φωτομετρικό σφάλμα ανάμεσα στην πραγματική εικόνα και την εικόνα που παράγεται με ανακατασκευή από το μοντέλο.

L_{id} είναι το σφάλμα ταυτότητας όπου χρησιμοποιείται ένα μοντέλο αναγνώρισης προσώπου, το οποίο τροφοδοτείται με την κανονική και την ανακατασκευασμένη εικόνα και ελαχιστοποιείται η διαφορά συνημίτονου μεταξύ τους ώστε τα πρόσωπα στις δύο εικόνες να έχουν την ίδια ταυτότητα.

L_{sc} είναι το σφάλμα συνοχής ταυτότητας, όπου το μοντέλο δοσμένου δύο εικόνων που δείχνουν το ίδιο άτομο θα πρέπει να έχει τις ίδιες παραμέτρους σχήματος ($\boldsymbol{\beta_i} = \boldsymbol{\beta_j}$)

Τέλος L_{reg} είναι το σφάλμα κανονικοποίησης, που έχει ως σκοπό να κανονικοποιήσει το σχήμα $E_\beta = \|\boldsymbol{\beta}\|_2^2$, την έκφραση $E_\psi = \|\boldsymbol{\psi}\|_2^2$ και την λευκαύγεια $E_\alpha = \|\boldsymbol{\alpha}\|_2^2$.

Το DECA κάνει πιο ακριβή αναπαράσταση του προσώπου από ότι θα έκανε αν χρησιμοποιούσε μόνο τις παραμέτρους του FLAME μέσω ενός χάρτη αντιστοιχίας μετατόπισης (UV displacement map) $D \epsilon [-0.01, 0.01]^{dxd}$ στον οποίο οι UV συντεταγμένες αντιστοιχίζουν το διάνυσμα μετατόπισης που απεικονίζεται στην 2D εικόνα σε 3D κορυφή . Για την δημιουργία του εκπαιδεύεται ένας ακόμη κωδικοποιητής που μετατρέπει την εικόνα εισόδου σε λανθάνουσες μεταβλητές $\boldsymbol{\delta} \boldsymbol{\epsilon} R^{128}$ που αναπαριστούν ιδιαιτερότητες του προσώπου. Οι λανθάνουσες μεταβλητές $\boldsymbol{\delta}$ μαζί με τις παραμέτρους έκφρασης ψ και την πόζα του σαγονιού $\boldsymbol{\theta_{jaw}}$ ενώνονται στην συνέχεια και τροφοδοτούν ένα αποκωδικοποιητή που μέσω της συνάρτησης F_d δημιουργεί τον UV displacement map D. Η εκπαίδευση του μοντέλου DECA παρουσιάζεται στην παρακάτω εικόνα.



Εικόνα 8: Εκπαίδευση του μοντέλου DECA. Το πρόσωπο της εικόνας εισόδου μετατρέπεται σε latent variables που εκφράζουν την υφή (c,a) ,την γεωμετρία του προσώπου(β,θ,ψ) και τον φωτισμό (l). Επίσης ένας κωδικοποιητής αναπαριστά τις λεπτομέρειες του προσώπου σε 128 latent variables (δ) που, μέσω ενός αποκωδικοποιητή, δημιουργούν έναν displacement map και μαζί με το mesh που δημιουργεί το FLAME από τις παραμέτρους β,θ,ψ, δημιουργείται ένα πιο λεπτομερές σχήμα του προσώπου. Το σχήμα του προσώπου μαζί με τον χάρτη υφής, που δημιουργείται από την αποκωδικοποίηση των παραμέτρων υφής, και τον φωτισμό χρησιμοποιούνται για την ανακατασκευή του προσώπου της εικόνας εισόδου . Σκοπός του μοντέλου είναι η ελαχιστοποίηση του σφάλματος ανακατασκευής.

2.3 Autoencoders

Ο Autoencoder[38] είναι ένας συγκεκριμένους τύπος νευρωνικού δικτύου[36] που είναι σχεδιασμένος έτσι ώστε να κωδικοποιεί την είσοδο σε μια συμπιεσμένη και με νόημα αναπαράσταση και στη συνέχεια να την αποκωδικοποιεί με τέτοιο τρόπο έτσι ώστε η ανακατασκευασμένη είσοδος να μοιάζει όσο το δυνατόν περισσότερο με την αρχική. Το πρόβλημα που καλείται να λύσει ο autoencoder είναι να μάθει τις συναρτήσεις $A: R^n \rightarrow R^p$ (κωδικοποιητής - encoder) και $B: R^p \rightarrow R^n$ (αποκωδικοποιητής – decoder) που να ικανοποιούν την εξίσωση:

$$\arg\min_{A,B} E[\Delta(x, B\ o\ A(x)],$$

όπου E η αναμενόμενη τιμή της κατανομής x και Δ το σφάλμα ανακατασκευής, που υπολογίζει την απόσταση μεταξύ της εξόδου του αποκωδικοποιητή και εισόδου του μοντέλου. Η συνάρτηση που χρησιμοποιείται συνήθως για το σφάλμα ανακατασκευής είναι το μέσο τετραγωνικό σφάλμα (**Mean Squared Error – MSE**) που δίνεται από τον τύπο:

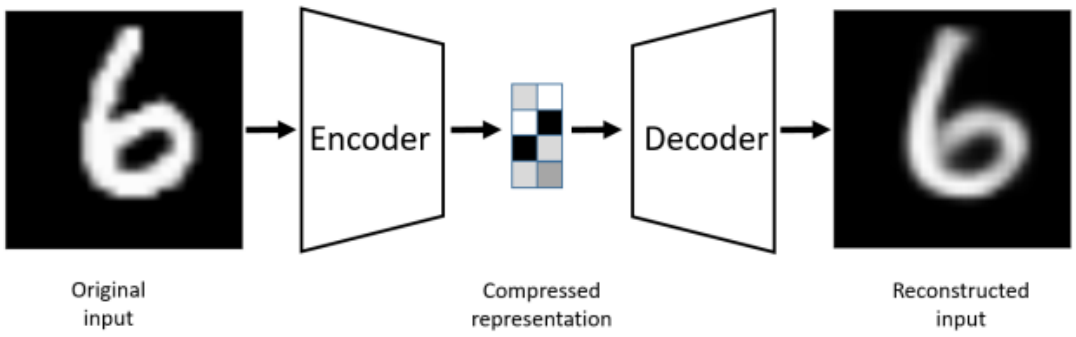
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2,$$

όπου Y θεωρείται η είσοδος του δικτύου και $\hat{Y}$ Η ανακατασκευασμένη είσοδος.

Συνεπώς, ο Autoencoder αποτελείται από 4 κυρίως μέρη:

- *Κωδικοποιητής (Encoder):* Στο επίπεδο αυτό το μοντέλο μαθαίνει πώς να μειώσει τις διαστάσεις της εισόδου και να συμπιέσει την είσοδο σε μια συμπιεσμένη αναπαράσταση.
- *Επίπεδο συμφόρησης (Bottleneck):* είναι το επίπεδο που περιέχει την συμπιεσμένη αναπαράσταση της εισόδου.
- *Αποκωδικοποιητής (Decoder):* Σε αυτό το επίπεδο το μοντέλο εκπαιδεύεται να ανακατασκευάζει την είσοδο από τη συμπιεσμένη αναπαράσταση.
- *Σφάλμα Ανακατασκευής (Reconstruction Loss):* Πρόκειται για την μέθοδο που μετράει πόσο καλά λειτουργεί ο αποκωδικοποιητής και πόσο κοντά μοιάζει η έξοδος του μοντέλου στην αρχική είσοδο.

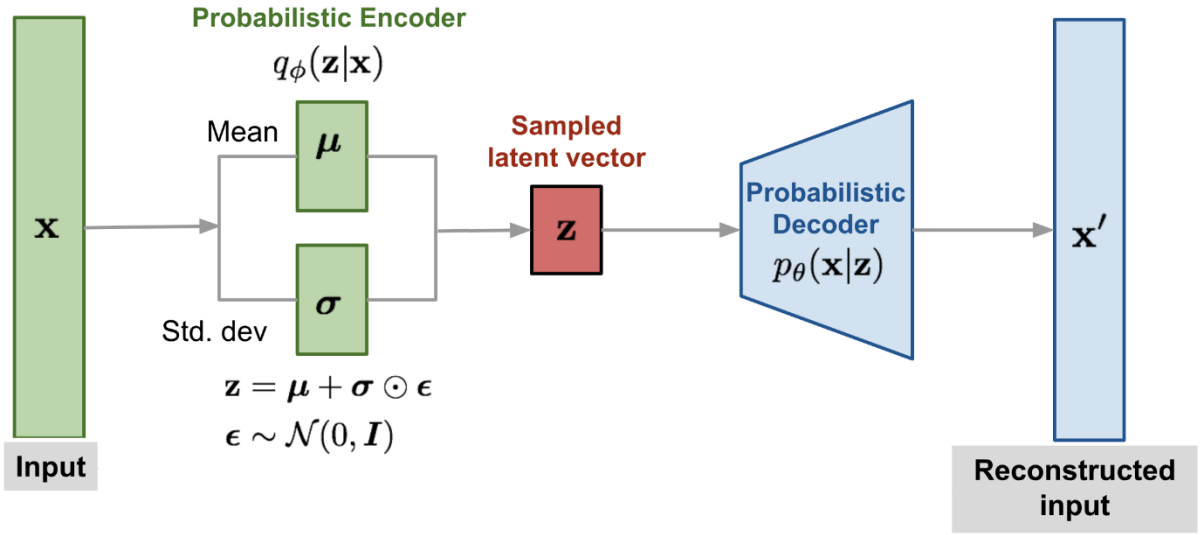
Μια αναπαράσταση του Autoencoder παρουσιάζεται στην παρακάτω εικόνα.



Εικόνα 9: Autoencoder. Η εικόνα εσόδου εισέρχεται σε έναν κωδικοποιητή και μειώνονται οι διαστάσεις σε μία συμπιεσμένη αναπαράσταση. Η αναπαράσταση αυτή εισέρχεται στον αποκωδικοποιητή και δημιουργείται η ανακατασκευασμένη εικόνα.

2.4 Variational Autoencoders

Η ικανότητα αναπαράστασης των Autoencoders βελτιώνεται σε μεγάλο βαθμό με τους Variational Autoencoders (**VAE**)[39]. Τα VAE είναι παραγωγικά μοντέλα που προσπαθούν να περιγράψουν την παραγωγή δεδομένων μέσα από μια πιθανολογική κατανομή. Η δομή ενός VAE παρουσιάζεται στην παρακάτω εικόνα.



Εικόνα 10: Variational Autoencoder. Η είσοδος X εισέρχεται σε έναν πιθανολογικό κωδικοποιητή. Ο χώρος τον λανθάνων μεταβλητών δειγματολειπτείται με χρήση του reparameterization trick. Οι λανθάνουσες μεταβλητές εισέρχονται στον πιθανολογικό αποκωδικοποιητή και παράγεται η ανακατασκευασμένη είσοδος.

Δοσμένου ενός σετ δεδομένων $\boldsymbol{X} = \{\boldsymbol{x_i}\}_{i=1}^{N}$, τα VAE περιλαμβάνουν ένα παραγωγικό μοντέλο για κάθε $\boldsymbol{x_i}$ που εξαρτάται από μια τυχαία λανθάνουσα μεταβλητή $\boldsymbol{z_i}$, όπου ϑ οι παράμετροι από τους οποίους εξαρτάται η πιθανολογική κατανομή ($p_\theta(\boldsymbol{x}|\boldsymbol{z})$). Το μοντέλο αυτό είναι ένας πιθανολογικός αποκωδικοποιητής (Probabilistic Decoder). Συμμετρικά, τα VAE περιλαμβάνουν ένα πιθανολογικό κωδικοποιητή (Probabilistic Encoder) στον οποίο θεωρείται ότι υπάρχει μία μεταγενέστερη πιθανότητα στις λανθάνουσες μεταβλητές $\boldsymbol{z_i}$ δοσμένου του $\boldsymbol{x_i}$ ($q_\varphi(\boldsymbol{z}|\boldsymbol{x})$). Οι παράμετροι ϑ και φ είναι άγνωστοι και πρέπει να υπολογιστούν από τα δεδομένα. Επίσης θεωρούμε μια a priori κατανομή για τις λανθάνουσες μεταβλητές $\boldsymbol{z_i}$ που δηλώνεται από την $p_\theta(\boldsymbol{z_i})$.

Η οριακή log-πιθανοφάνεια (marginal log-likelihood) εκφράζεται ως άθροισμα των μεμονωμένων σημείων $\log p_\theta(\boldsymbol{x_1}, \boldsymbol{x_2}, \ \ldots, \boldsymbol{x_N}) = \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x_i})$ Και κάθε σημείο μπορεί να γραφτεί ως:

$$\log p_\theta(\boldsymbol{x_i}) = D_{KL}(q_\varphi(\boldsymbol{z}|\boldsymbol{x_i}) \,\|(p_\theta(\boldsymbol{x}|\boldsymbol{z_i})) + L(\theta, \varphi; \boldsymbol{x_i}), \qquad (\alpha)$$

Όπου D_{KL} η απόκλιση Kullback-Leibler (μέτρο διαφοράς ανάμεσα σε δύο κατανομές πιθανότητας) και ο δεύτερος όρος L ονομάζεται μεταβλητό κατώτερο όριο επειδή, καθώς η απόκλιση Kullback-Leibler είναι μη αρνητική, ο όρος L είναι το κατώτερο όριο του marginal log-likelihood και μεγιστοποιώντας το κατώτερο όριο βελτιώνεται η προσέγγιση της posterior κατανομής. Το μεταβλητό κατώτερο όριο μπορεί να γραφτεί και ως:

$$L(\theta, \varphi;\ \boldsymbol{x_i}) \triangleq E_{q_\varphi(\boldsymbol{z}|\boldsymbol{x_i})}\big[-\log q_\varphi(\boldsymbol{z}|\boldsymbol{x}) + \log p_\theta(\boldsymbol{x}|\boldsymbol{z})\big]$$

Το μεταβλητό κατώτερο όριο περιέχει την μεταβλητή $q_\varphi(\boldsymbol{z}|\boldsymbol{x})$ που είναι τυχαία και επομένως δεν μπορεί να εφαρμοστεί ο αλγόριθμος της οπισθοδιάδοσης διότι απαιτείται οι παράγωγοι των συναρτήσεων να είναι γνωστοί. Η δειγματοληψία θα πρέπει να εκφραστεί με τέτοιο τρόπο που να επιτρέπεται η οπισθοδιάδοση εντός του δικτύου. Αυτό επιτυγχάνεται με έναν απλό τρόπο που ονομάζεται τρικ επαναπαραμετροποίησης (reparametrisation trick) και βασίζεται στο γεγονός ότι αν η μεταβλητή z είναι μια τυχαία μεταβλητή που ακολουθεί Γκαουσιανή κατανομή με μέση τιμή $g(x)$ και διακύμανση $h(x)$ τότε μπορεί να εκφραστεί ως $z = h(x)\zeta + g(x),\ \zeta \sim N(0, I)$. Έτσι το μεταβλητό κατώτερο όριο μπορεί να διατυπωθεί ως:

$$L(\theta, \varphi;\ \boldsymbol{x_i}) \approx L^{\sim}(\theta, \varphi;\ \boldsymbol{x_i}) = \frac{1}{L}\sum_{i=1}^{L} \log p_\theta(\boldsymbol{x_i}, \boldsymbol{z}_{i,t}) - \log q_\varphi(\boldsymbol{z}_{i,t|\boldsymbol{x_i}})$$

Παίρνουμε έτσι την διαφοροποίηση συνάρτηση:

$$\hat{L}^M(\theta, \varphi; \boldsymbol{X}) = \frac{N}{M}\sum_{i=1}^{M} L^{\sim}(\theta, \varphi;\ \boldsymbol{x_i})$$

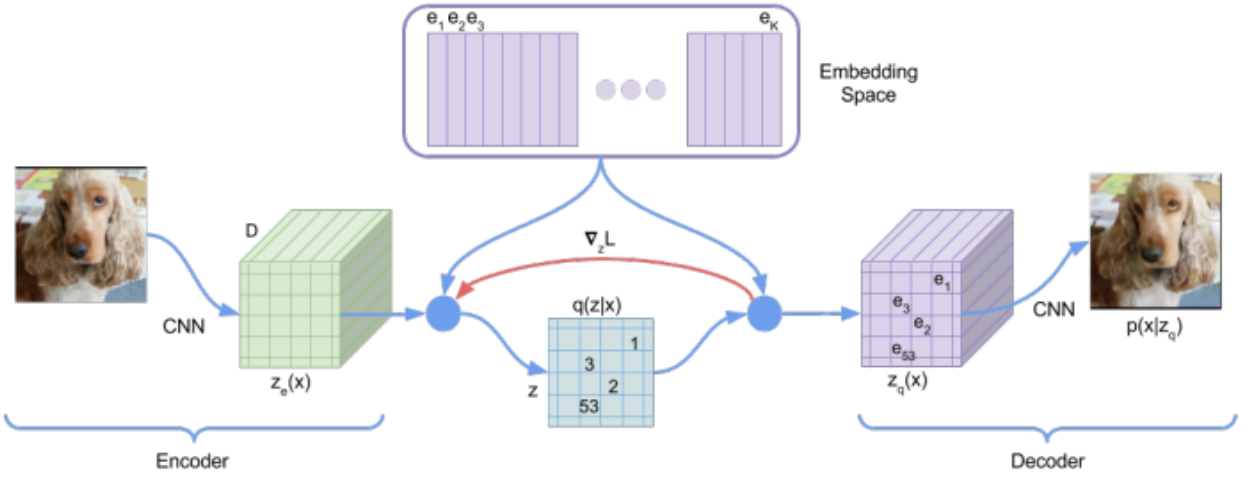
που μπορεί να διαφοροποιηθεί για τα θ, φ και να χρησιμοποιηθούν σε αυτή οι τεχνικές οπισθοδιάδοσης. Παρακάτω παρουσιάζεται ο αλγόριθμος εκπαίδευσης ενός VAE.

Algorithm Pseudo-code for VAE

$(\theta, \phi) \leftarrow$ Initialize Parameter
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints
  $\epsilon \leftarrow L$ random samples of $p(\epsilon)$
  $\mathbf{g} \leftarrow \nabla_{(\theta,\phi)} \hat{\mathcal{L}}^{\mathbf{M}}(\theta, \phi; \mathbf{X})$ {Gradients of Equation $\hat{\mathcal{L}}^{\mathbf{M}}$}
  $(\theta, \phi) \leftarrow$ Update parameters based on $\mathbf{g}$ {e.g., update with SGD or Adagrad}
until Convergenge of (θ, ϕ)
return (θ, ϕ)

2.5 Vector Quantization Variational Autoencoders

Ο Vector Quantization Variational Autoencoder (VQ-VAE)[42] είναι ένα μοντέλο ανακατασκευής της εισόδου που βελτιώνει το απλό VAE όσο αφορά την ποιότητα ανακατασκευής. Μοιάζει σε μεγάλο βαθμό με τα VAE καθώς αποτελείται επίσης από έναν κωδικοποιητή, έναν αποκωδικοποιητή και μία a priori κατανομή p(z). Η διαφορά έγκειται στο ότι στα VAE οι πιθανότητες θεωρείται ότι έχουν κανονική κατανομή ενώ στο VQ-VAE οι πιθανότητες είναι κατηγορικές (categorical distribution) και, επίσης, στο VQ-VAE οι λανθάνουσες μεταβλητές είναι διακριτές και χρησιμοποιείται μία άλλη μέθοδος εκπαίδευσης που είναι εμπνευσμένη από το διανυσματικό κβαντισμό (Vector Quantization). Η δομή του δικτύου VQ-VAE φαίνεται στο παρακάτω σχήμα.



Εικόνα 11: VQ-VAE. Η εικόνα εισόδου εισέρχεται στον CNN κωδικοποιητή που μειώνει τις διαστάσεις τις. Στην συνέχεια βρίσκεται ο χώρος των embeddings μέσω του αλγορίθμου Vector Quantization και τα embeddings, μέσω του CNN αποκωδικοποιητή παράγουν την ανακατασκευασμένη είσοδο.

Το μοντέλο παίρνει την είσοδο x και την περνάει μέσα από τον κωδικοποιητή που παράγει την έξοδο $z_e(x)$. Στη συνέχεια υπολογίζονται οι διακριτές λανθάνουσες μεταβλητές z με μία αναζήτηση nearest-neighbour χρησιμοποιώντας το embedding χώρο $e \in R^{KxD}$ όπου K το μέγεθος των λανθάνων μεταβλητών και D οι διαστάσεις των embedding διανυσμάτων των λανθάνων μεταβλητών e_i. Έπειτα υπολογίζονται οι κατηγορικές μεταγενέστερες κατανομές $q(z|x)$ που ορίζονται ως εξής:

$$q(z = k|x) = \begin{cases} 1 \; \gamma\iota\alpha \; k = \arg\min_j \left\| z_e(x) - e_j \right\|_2 \\ 0 \qquad \alpha\lambda\lambda\omega\acute{\upsilon} \end{cases}$$

Η είσοδος του αποκωδικοποιητή $z_q(x)$ ισούται με το embedding διάνυσμα e_k, $k = \arg\min_j \left|\left|z_e(x) - e_j\right|\right|_2$. Το μοντέλο θεωρείται ως VAE όπου γίνεται να βρεθεί ελάχιστο κατώτερο φράγμα για την $\log p(x)$. Η κατανομή $q(z|x)$ είναι ντετερμινιστική και συνεπώς μπορεί να οριστεί η KL divergent που ισούται με $\log K$.

Καθώς δεν υπάρχει πραγματική κλίση για το embedding διάνυσμα e_k, γίνεται μια εκτίμηση της κλίσης μέσω της αντιγραφή της κλίσης της εισόδου του αποκωδικοποιητή απευθείας στην έξοδο του κωδικοποιητή. Έτσι κατά το backpropagation η κλίση $\nabla_z L$ περνιέται χωρίς αλλαγή στον κωδικοποιητή.

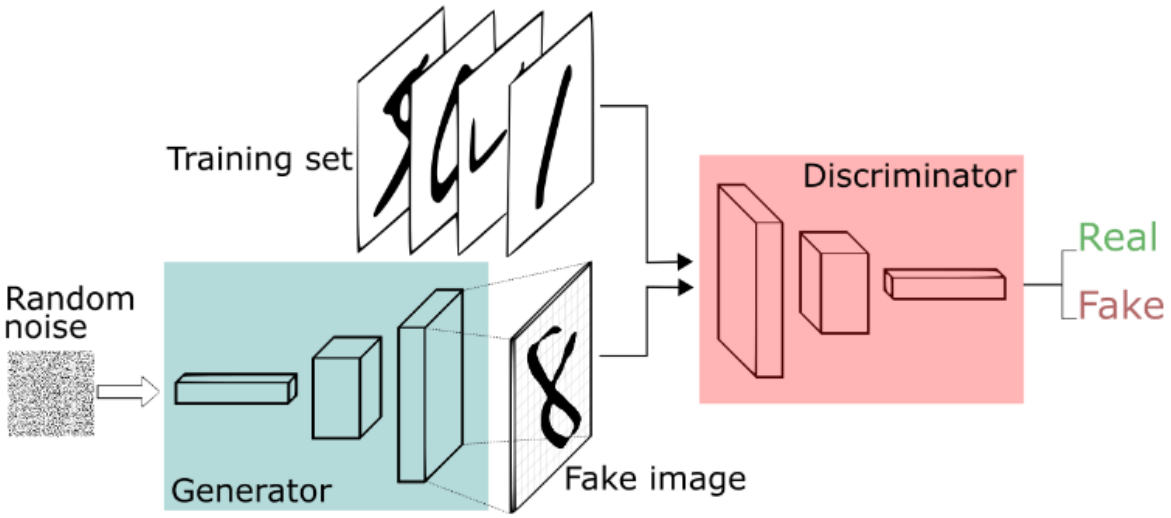
Η συνάρτηση κόστους του μοντέλου αποτελείται από τρεις όρους. Ο πρώτος όρος είναι το σφάλμα ανακατασκευής ($\log p(x|z_q(x))$) που βελτιστοποιεί τον κωδικοποιητή και τον αποκωδικοποιητή. Λόγω της εκτίμησης κλίσης που αναφέρθηκε παραπάνω, τα embeddngs e_i δεν δέχονται κλίση από το σφάλμα ανακατασκευής. Προκειμένου να εκπαιδευτεί, λοιπόν, ο χώρος των embeddings χρησιμοποιείται ο αλγόριθμος Vector Quantization. Σκοπός του αλγορίθμου είναι η ελαχιστοποίηση του MSE για να μετακινήσει τα embedding διανύσματα e_i πιο κοντά στις εξόδους του κωδικοποιητή $z_e(x)$. Ωστόσο, επειδή ο όγκος του χώρου των embeddings είναι αδιάστατος, δύναται να μεγαλώσει αυθαίρετα αν η εκπαίδευση των embeddings είναι πιο αργή από αυτή του κωδικοποιητή. Για την αποφυγή αυτής της αύξησης της εξόδου προστίθεται στην συνάρτηση κόστους και ένα σφάλμα δέσμευσης (commitment loss). Η συνολική συνάρτηση κόστους είναι η:

$$L = \log p(x|z_q(x)) + \ ||sg[z_e(x)] - e||_2^2 + \beta||z_e(x) - sg[e]||_2^2,$$

όπου sg ένας operator που κατά την οπισθοδιάδοση αποκόπτει την είσοδο της κλίσης.

2.6 Generative Adversarial Networks

Τα Generative Adversarial Networks (GANs)[2] είναι μια αρχιτεκτονική μοντέλων μηχανικής μάθησης που περιέχουν 2 νευρωνικά δίκτυα τα οποία ανταγωνίζονται μεταξύ τους έτσι ώστε να γίνουν πιο ακριβείς οι προβλέψεις τους. Τα 2 νευρωνικά δίκτυα που αποτελούν το GAN είναι γνωστά ως Generator και Discriminator. Ο Generator είναι ένα συνελικτικό νευρωνικό δίκτυο που σκοπός του είναι να παράξει εξόδους που θα μπορούσαν εύκολα να θεωρηθούν πραγματικά δεδομένα. Ο Discriminator είναι επίσης ένα συνελικτικό νευρωνικό δίκτυο που σκοπός του είναι να αναγνωρίζει ποιες από τις εξόδους που δέχεται είναι πραγματικές και ποιες όχι. Έτσι τα GAN παράγουν δεδομένα που βοηθούν στην εκπαίδευσή τους και, καθώς εκπαιδεύεται το δίκτυο, ο Generator θα παράγει καλύτερες εξόδους και ο Discriminator θα μάθει να αναγνωρίζει καλύτερα τα αληθινά δεδομένα. Συνεπώς τα δύο δίκτυα ανταγωνίζονται το ένα το άλλο ώστε να βελτιωθεί η απόδοσή και των δύο.



Εικόνα 12: Τυπική δομή ενός Generative Adversarial Network. Ο Generator, δοσμένης της πραγματικής εισόδου μαζί με θόρυβο, παράγει μία ψεύτικη εικόνα με σκοπό να μπερδέψει τον Discriminator. Ο Discriminator δέχεται εικόνες από το σετ δεδομένων και αυτές που παράγει ο Generator και προσπαθεί να καταλάβει ποιες είναι αυθεντικές και ποιες όχι. Τα δύο δίκτυα παίζουν ένα παιχνίδι μεγίστου-ελαχίστου που οδηγεί στην βελτίωση και των δύο δικτύων.

Ο Discriminator έχει ως έξοδο του την τιμή $D(x)\epsilon[0,1]$ που ισούται με την πιθανότητα η είσοδος x να είναι πραγματική (όπου $D(x) = 1$ για την πραγματική είσοδο). Στόχος της εκπαίδευσης του Discriminator είναι η μεγιστοποίηση της πιθανότητας κατηγοριοποίησης πραγματικών εισόδων ως αληθινές και εισόδων παραγμένων από τον Generator ως ψεύτικες. Για αυτό το λόγο ως συνάρτηση κόστους χρησιμοποιείται η cross-entropy και ο στόχος της εκπαίδευσης του Discriminator δίνεται από την εξίσωση:

$$\max_{\mathrm{D}} V(D) = E_{\boldsymbol{x}\sim \boldsymbol{p_{data}}(x)}[\log D(\boldsymbol{x})] + E_{\boldsymbol{z}\sim \boldsymbol{p_z}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))],$$

όπου $p_z(\boldsymbol{z})$ ο θόρυβος εισόδου.

Από την μεριά του Generator σκοπός του είναι να παράξει εξόδους όσο πιο ρεαλιστικά γίνεται και να μπερδέψει τον Discriminator ώστε να νομίζει πως είναι πραγματικά δεδομένα, να έχει δηλαδή μεγάλη τιμή η πιθανότητα D(x). Επομένως στόχος του Generator είναι ο εξής:

$$\min_{\mathrm{G}} V(G) = E_{\boldsymbol{z}\sim \boldsymbol{p_z(z)}}[\log(1 - D\big(G(\boldsymbol{z})\big))]$$

Επειδή στόχος των GANs είναι η εκπαίδευση και των δύο δικτύων, ένα GAN μπορεί να θεωρηθεί ως ένα παιχνίδι μεγίστου-ελαχίστου όπου ο Generator προσπαθεί να ελαχιστοποιήσει το V και ο Discriminator να το μεγιστοποιήσει, ή με μορφή εξίσωσης:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} V(D,G) = E_{\boldsymbol{x}\sim \boldsymbol{p_{data}}(x)}[\log D(\boldsymbol{x})] + E_{\boldsymbol{z}\sim \boldsymbol{p_z}(\boldsymbol{z})}[\log(1 - D\big(G(\boldsymbol{z})\big))]$$

Τα δύο μοντέλα εκπαιδεύονται εναλλασσόμενα. Έτσι στην μία επανάληψη κρατιούνται σταθερές οι παράμετροι του Generator και μέσω του gradient descent εκπαιδεύεται ο Discriminator και στην επόμενη επανάληψη συμβαίνει το αντίθετο. Η εκπαίδευση σταματάει όταν ο Generator αρχίσει να παράγει εξόδους με μεγάλη ακρίβεια. Ο αλγόριθμος εκπαίδευσης των GANs παρουσιάζεται παρακάτω.

Algorithm Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

  for k steps **do**

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
  - Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

  end for

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

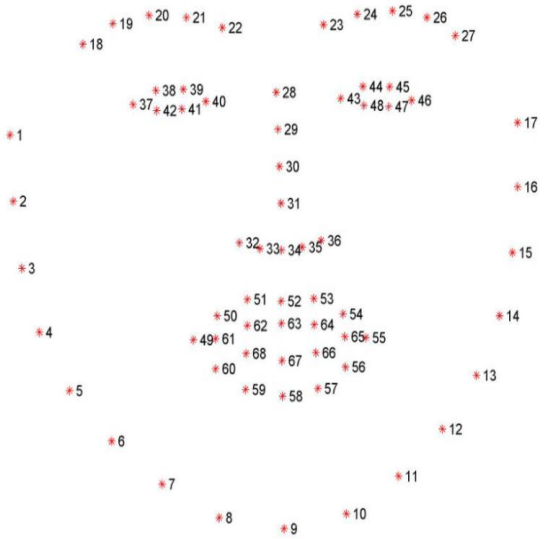
The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

2.7 Αποτύπωση των χαρακτηριστικών του προσώπου

Στα πλαίσια της παρούσας εργασίας είναι απαραίτητο να βρεθεί ένας τρόπος να εκφραστούν τα χαρακτηριστικά του προσώπου καθώς, με την βοήθεια αυτών, δύναται να εξαχθούν χαρακτηριστικά χρήσιμα για την αναγνώριση των deepfakes. Η αποτύπωση αυτή των χαρακτηριστικών του προσώπου θα γίνει μέσω των οροσήμων του προσώπου (facial landmarks).

Τα ορόσημα του προσώπου ορίζονται ως ο εντοπισμός και η τοπικοποίηση ορισμένων χαρακτηριστικών σημείων του προσώπου. Ορόσημα που χρησιμοποιούνται συχνά είναι οι γωνίες του ματιού, η άκρη της μύτης, οι γωνίες του ρουθουνιού, του στόματος, τα σημεία τέλους των φρυδιών, του αυτιού, τον λοβών, του πηγουνιού κτλ. Ορόσημα όπως οι γωνίες του ματιού ή η άκρη της μύτης είναι γνωστό ότι επηρεάζονται ελάχιστα από τις εκφράσεις του προσώπου. Είναι, επομένως, πιο αξιόπιστα και για αυτό είναι γνωστά ως πιστά ορόσημα (fiducial landmarks).

Τα ορόσημα του προσώπου κατατάσσονται σε δύο κατηγορίες, τα πρωτεύοντα και τα δευτερεύοντα ορόσημα. Τα πρωτεύοντα ορόσημα είναι τα άμεσα εντοπιζόμενα ορόσημα, δηλαδή τα πιστά ορόσημα. Έχουν πιο σημαντικό ρόλο στην αναγνώριση της ταυτότητας του προσώπου. Τέτοια ορόσημα όπως οι γωνίες του στόματος, των ματιών, οι άκρες της μύτης και των φρυδιών μπορούν να εντοπιστούν με σχετική ευκολία χρησιμοποιώντας χαρακτηριστικά εικόνας χαμηλού επιπέδου (low level features). Τα δευτερεύοντα ορόσημα είναι το πηγούνι, το περίγραμμα των μάγουλων, τα μεσαία σημεία των φρυδιών και των χειλιών, τα σημεία που δεν είναι άκρες και τα ρουθούνια. Αυτά τα ορόσημα έχουν πιο σημαντικό ρόλο στην αναγνώριση της έκφρασης και την παρακολούθηση της κίνησης του προσώπου.



Εικόνα 13: Τα 68 ορόσημα του προσώπου που χρησιμοποιούνται πιο συχνά στην αποτύπωση των χαρακτηριστικών του προσώπου

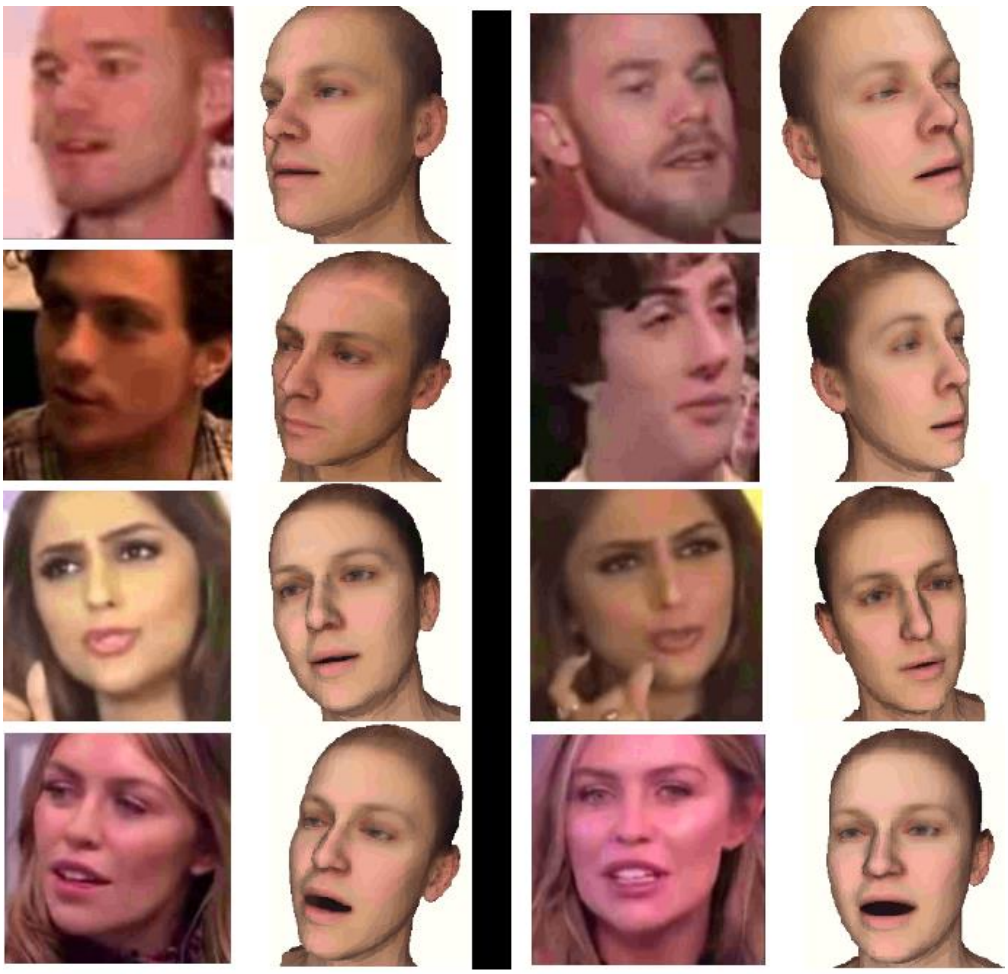
Κεφάλαιο 3 Μεθοδολογία

Για την εύρεση του μοντέλου αναγνώρισης παραποιημένων βίντεο χρησιμοποιήθηκαν διάφορες μέθοδοι και προσεγγίσεις. Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε είναι η Python και το framework που χρησιμοποιήθηκε για την δημιουργία και την εκπαίδευση των νευρωνικών δικτύων είναι το PyTorch[50]. Άλλες βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι scikit-learn[51] για τις μετρικές αξιολόγησης, os για την χρήση των λειτουργιών που εξαρτώνται από το λειτουργικό σύστημα, numpy[52] για την διαχείριση των πολυδιάστατων δομών δεδομένων, OpenCV για την διαχείριση των εικόνων και matplotlib[53] για τα σχεδιαγράμματα.

3.1 Προεπεξεργασία των δεδομένων

Οι μέθοδοι που θα χρησιμοποιηθούν ανήκουν στην κατηγορία των τεχνικών υψηλού επιπέδου, δηλαδή δεν βασίζονται σε artifacts στις εικόνες ή στα βίντεο προκειμένου να αναγνωρίσουν τα παραποιημένα βίντεο αλλά σε χαρακτηριστικά πιο υψηλού επίπεδού που, στην περίπτωσή μας, είναι οι συντελεστές των 3DMMs (κεφ. 2.1.1) και, πιο συγκεκριμένα, οι συντελεστές του μοντέλου FLAME (κεφ. 2.1.2). Έτσι χρησιμοποιούνται 100 παράμετροι για να περιγράψουν το σχήμα (ταυτότητα) του προσώπου, 50 παράμετροι που περιγράφουν την έκφραση του προσώπου και 6 παράμετροι οι οποίοι χρησιμοποιούνται για την πόζα (μέσω μετατροπής τους έτσι ώστε να αντιπροσωπεύουν όλους τους πίνακες περιστροφής). Ωστόσο όλα τα σετ δεδομένων που υπάρχουν είναι είτε με μορφή εικόνων είτε με μορφή βίντεο. Συνεπώς χρειάζεται ένα framework που να δέχεται τις εικόνες ή αλληλουχία εικόνων και θα επιστρέφει τους συντελεστές που χρησιμοποιεί το FLAME. Το framework αυτό είναι το DECA (κεφ. 2.1.3).

Το DECA εξάγει από τα πρόσωπα που ανιχνεύει στις εικόνες τους συντελεστές που απαιτούνται από το FLAME για την αναπαράσταση των προσώπων. Επειδή μας ενδιαφέρει και η χρονική αλληλουχία των προσώπων, οι μέθοδοί που θα μελετηθούν χρησιμοποιούν αλληλουχίες καρέ από βίντεο. Για αυτό το λόγο εξάγονται από τα βίντεο τα διαδοχικά καρέ και στην συνέχεια με την βοήθεια του DECA εξάγονται οι παράμετροι του FLAME. Επίσης εξάγονται 50 παράμετροι για την υφή του προσώπου που χρησιμοποιούνται για να δημιουργηθεί ο UV χάρτης υφής συμβατός με το FLAME και 68 ορόσημα του προσώπου. Παραδείγματα μετατροπής προσώπων από εικόνες σε 3DMM παρουσιάζονται παρακάτω.



Εικόνα 14: Αρχικές φωτογραφίες και τα 3DMM που παράγει το DECA

Από τα ορόσημα του προσώπου που υπολογίζονται από το FLAME μπορούν να εξαχθούν κάποια επιπλέον χαρακτηριστικά που δύναται να βοηθήσουν στην αναγνώριση των deepfakes. Ένα από αυτά είναι ο ρυθμός κίνησης του στόματος. Καθώς τα deepfakes επηρεάζουν τον τρόπο με τον οποίο κινείται το στόμα, θα μπορούσαν να υπάρχουν χρονικές ασυνέχειες οι οποίες θα υποδεικνύουν την παραποίηση του προσώπου. Για να μετρηθεί κατά πόσο ανοιχτό ή κλειστό είναι το στόμα, χρησιμοποιείται το Σκορ Ανοιχτού Κλειστού Στόματος (Mouth Open Closed Score - MOCS)[43] που ορίζεται ως:

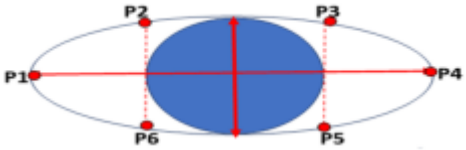
$$0 \leq MOCS = \frac{\sqrt{(c_x - d_x)^2 + \left(c_y - d_y\right)^2 + (c_z - d_z)^2}}{\sqrt{(a_x - b_x)^2 + \left(a_y - b_y\right)^2 + (a_z - b_z)^2}} \leq 1,$$

όπου $a = [a_x, a_y, a_z]^T, b = [b_x, b_y, b_z]^T$ είναι τα ορόσημα που αντιστοιχούν στο πάνω και στο κάτω εξωτερικό μέρος των χειλιών αντίστοιχα (ορόσημα 52 και 58) και $c = [c_x, c_y, c_z]^T, d = [d_x, d_y, d_z]^T$ τα ορόσημα του πάνω και κάτω εσωτερικού μέρους των χειλιών. Με την χρήση του MOCS μετράται η απόσταση μεταξύ του εσωτερικού των χειλιών και του εξωτερικού τους και επιστρέφεται ένας αριθμός $0 \leq MOCS \leq 1$ που όσο μικρότερος είναι τόσο περισσότερο κλειστό είναι το στόμα. Όσο περισσότερο ανοίγει το στόμα τόσο περισσότερο αυξάνονται οι αποστάσεις μεταξύ του εσωτερικού και εξωτερικού μέρους των χειλιών και, συνεπώς, αυξάνεται και το MOCS.

Ένα ακόμη χαρακτηριστικό που μπορεί να είναι χρήσιμο στην αναγνώριση των deepfakes είναι ο ρυθμός ανοιγοκλεσίματος των ματιών. Δεδομένου ότι τα σετ δεδομένων με τα οποία εκπαιδεύονται οι μέθοδοι deepfake έχουν ένα μικρό αριθμό εικόνων με κλειστά τα μάτια, είναι αναμενόμενο ο ρυθμός ανοιγοκλεσίματος των ματιών σε deepfake βίντεο να είναι μικρότερος από αυτό των αυθεντικών βίντεο. Μια μετρική που μπορεί να εκφράσει την κατάσταση του ματιού είναι η EAR (Eye Aspect Ratio)[44] που ορίζεται ως εξής:

$$EAR = \frac{\|P_2 - P_6\| + \|P_3 - P_5\|}{2\|P_1 - P_4\|},$$

όπου $P_{1,2,\ \ldots,\ 6}$ τα ορόσημα του ματιού όπως φαίνονται στην παρακάτω εικόνα και $\|P_a - P_b\| = \sqrt{(P_{a,x} - P_{b,x})^2 + (P_{a,y} - P_{b,y})^2 + (P_{a,z} - P_{b,z})^2}$.

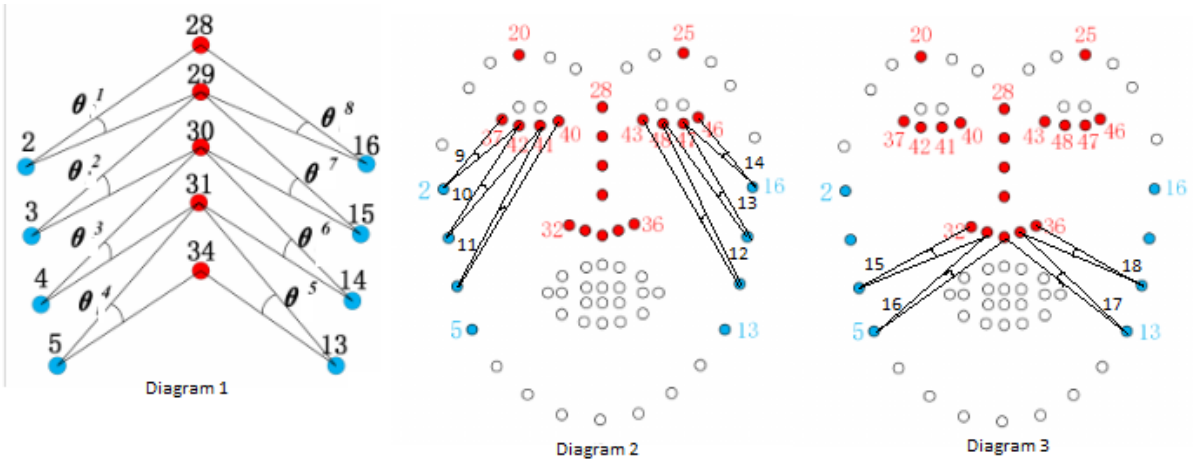


Εικόνα 15: Τα 6 σημεία που χρησιμοποιεί το EAR

Η μετρική EAR παίρνει μικρότερες τιμές όσο πιο κλειστό είναι το μάτι. Επιλέχθηκε να χρησιμοποιηθεί η μετρική EAR τόσο για το δεξί όσο και για το αριστερό μάτι, οπότε έχουμε τις μετρικές EAR_{right}, EAR_{left} οι οποίες χρησιμοποιούν για τα σημεία $P_{1,2,\ \ldots,\ 6}$ τα ορόσημα του προσώπου του αριστερού και του δεξιού ματιού όπως παρουσιάζονται παρακάτω.

Εικόνα 16: Τα ορόσημα που χρησιμοποιούνται στις μετρικές EAR_{right} και EAR_{left}

Τέλος, η μεταβολή των γωνιών ανάμεσα στα ορόσημα του προσώπου θα μπορούσε να αποτελέσει μία καλή ένδειξη για την παραποίηση του προσώπου. Κατά την δημιουργία των deepfakes υπολογίζονται τα ορόσημα προσώπου τόσο του αυθεντικού προσώπου όσο και αυτού με το οποίο θα παραποιηθεί και στην συνέχεια τοποθετείται το ένα πρόσωπο πάνω στο άλλο ελαχιστοποιώντας την απόσταση μεταξύ των οροσήμων του κέντρου του προσώπου. Με αυτόν τον τρόπο τα ορόσημα του εξωτερικού προσώπου είναι αυτά της αυθεντικής ταυτότητας αλλά αυτά του εσωτερικού είναι της νέας ταυτότητας. Παρατηρώντας, λοιπόν, την μεταβολή των γωνιών που έχουν ως σημείο αναφοράς τα ορόσημα του εξωτερικού προσώπου και σχηματίζονται μεταξύ δύο διανυσμάτων με αρχή τα σημεία αναφοράς και τέλος τα ορόσημα του εσωτερικού του προσώπου, δύναται να βρεθούν ασυνέχειες που δεν θα υπήρχαν σε μη παραποιημένα βίντεο. Επιλέχθηκε να χρησιμοποιηθούν 18 γωνιές οι οποίες και παρουσιάζονται στο παρακάτω σχήμα.



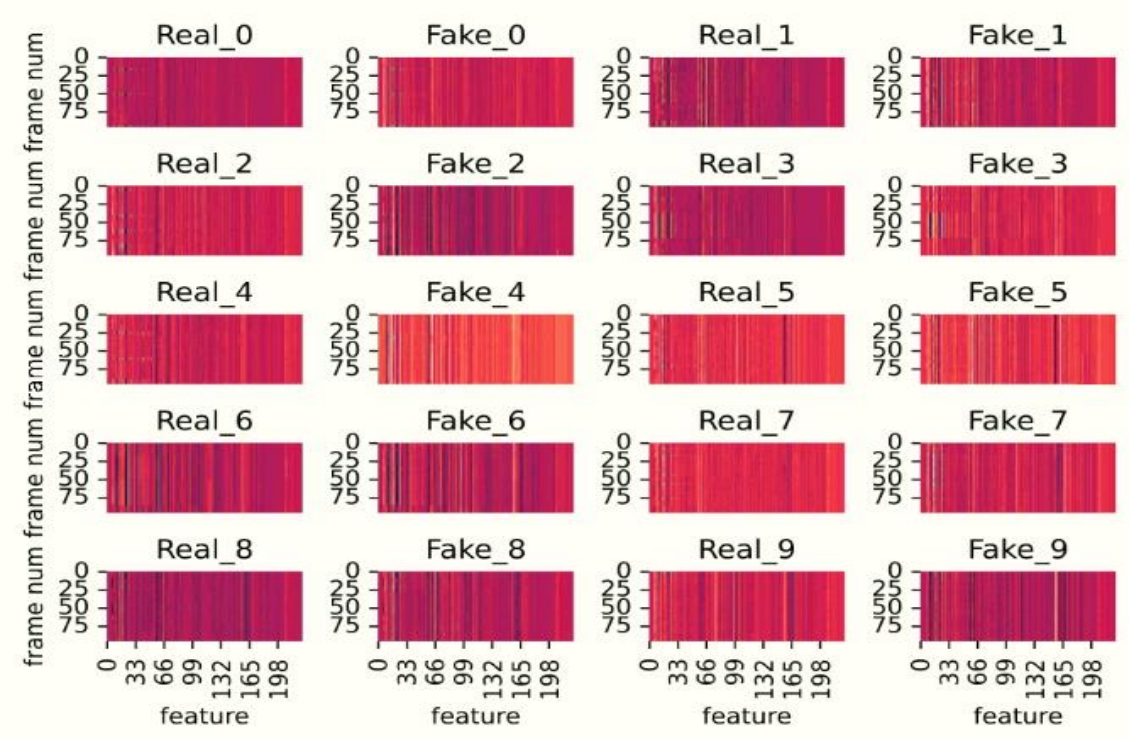
Εικόνα 17: Οι 18 γωνίες με σημείο αναφοράς τα ορόσημα του εξωτερικού προσώπου μεταξύ δύο διανυσμάτων με αρχή το σημείο αναφοράς και πέρας τα ορόσημα του εσωτερικού προσώπου

Συνοψίζοντας, κάθε πρόσωπο μετατρέπεται σε ένα διάνυσμα $f \epsilon R^{227}$, όπου 100 παράμετροι χρησιμοποιούνται για να εκφράσουν το σχήμα (ταυτότητα) του προσώπου, 50 για την έκφραση, 6 για την πόζα, 50 για την υφή (texture) του προσώπου, 1 για το MOCS, 2 για το EAR και 18 για τις γωνίες μεταξύ των οροσήμων του προσώπου. Επειδή μας ενδιαφέρει και η μεταβολή στον χρόνο αυτών των παραμέτρων, μετατρέπουμε *N* διαδοχικά καρέ σε ένα βίντεο και έτσι σχηματίζεται ένας δισδιάστατος διανυσματικός χώρος $v \epsilon R^{Nx227}$, όπου οι γραμμές αντιστοιχούν στις παραμέτρους του προσώπου για μία χρονική στιγμή και οι στήλες στα διαδοχικά καρέ του βίντεο. Με αυτό τον τρόπο κάθε στήλη δείχνει την μεταβολή μιας συγκεκριμένης παραμέτρου στον χρόνο και, ακόμη, ένα βίντεο μπορεί να αναπαρασταθεί σαν μία δισδιάστατη εικόνα και να χρησιμοποιηθούν νευρωνικά δίκτυα που έχουν αναπτυχθεί για χρήση με εικόνες.

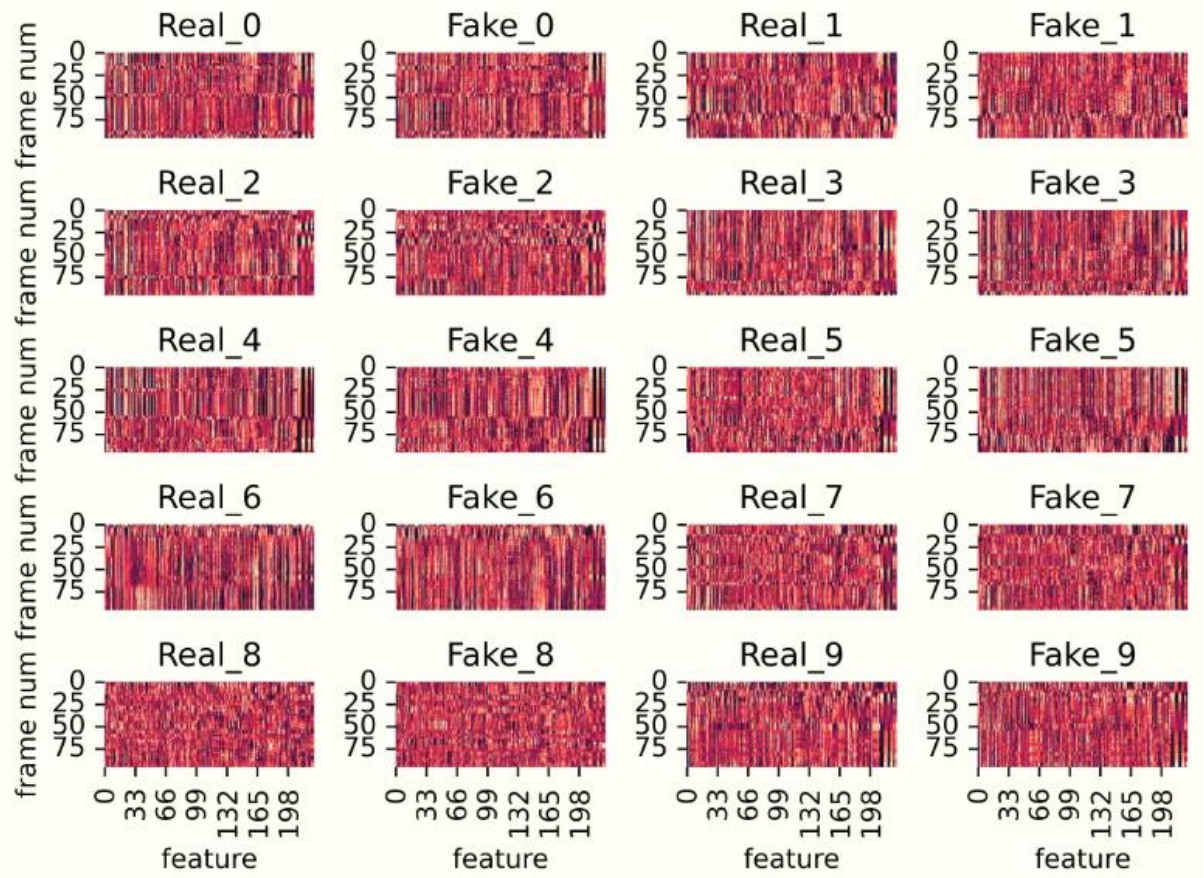
Οι δισδιάστατοι πίνακες των διανυσμάτων υφίστανται κανονικοποίηση μετά την δημιουργία τους. Έτσι όλα τα δεδομένα κανονικοποιούνται ώστε οι τιμές τους να βρίσκονται στο εύρος [0,1]. Δοκιμάστηκαν δύο προσεγγίσεις για την κανονικοποίηση:

- Κανονικοποίηση για όλο δισδιάστατο διανυσματικό χώρο ‘εικόνα’
- Κανονικοποίηση του δισδιάστατου διανυσματικού χώρου για κάθε χαρακτηριστικό στον άξονα του χρόνου

Για να αξιολογηθεί ποια μέθοδος επιφέρει πιο ευδιάκριτα αποτελέσματα έγινε αναπαράσταση των heatmaps των δισδιάστατων διανυσματικών χώρων για 200 βίντεο, 100 αυθεντικά βίντεο διάρκειας δύο δευτερολέπτων (96 καρέ) και τα 100 αντίστοιχα παραποιημένα βίντεο. Για την κανονικοποίηση χρησιμοποιήθηκε ο MinMaxScaler της βιβλιοθήκης scikit-learn. Τα 20 πρώτα ζεύγη βίντεο για τις 2 προσεγγίσεις της κανονικοποίησησς παρουσιάζονται παρακάτω.



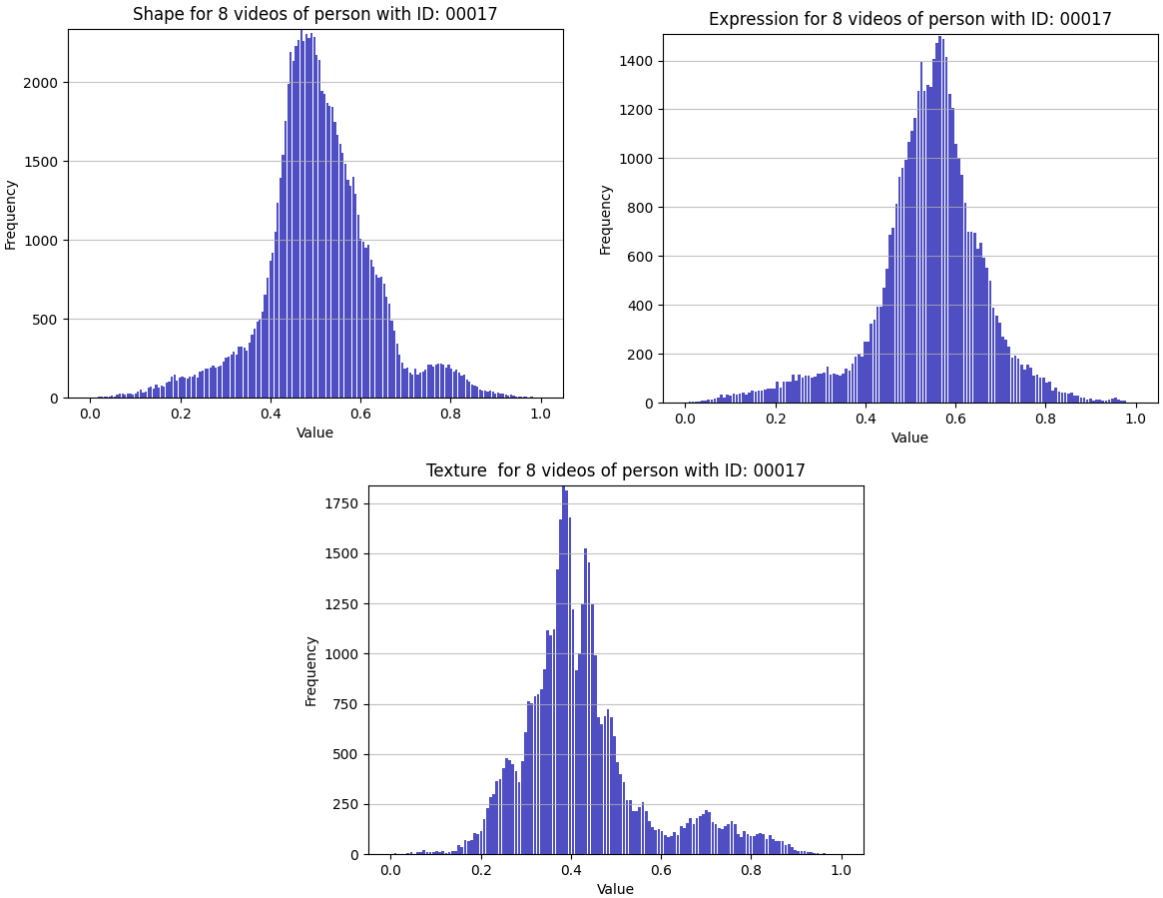
Εικόνα 18: Heatmaps για 10 αυθεντικά και τα αντίστοιχα παραποιημένα βίντεο με κανονικοποίηση σε όλη την 'εικόνα'. Όσο πιο ανοιχτό είναι το χρώμα ενός σημείου της ‘εικόνας’ τόσο πιο κοντά είναι η τιμή του στο 1 ενώ όσο πιο σκούρο τόσο πιο κοντά είναι στο 0. Οι γραμμές αντιστοιχούν στα χαρακτηριστικά και οι στήλες στα διαδοχικά καρέ του βίντεο.



Εικόνα 19: Heatmaps για 10 αυθεντικά και τα αντίστοιχα παραποιημένα βίντεο με κανονικοποίηση των χαρακτηριστικών στον άξονα του χρόνου. Όσο πιο ανοιχτό είναι το χρώμα ενός σημείου της 'εικόνας' τόσο πιο κοντά είναι η τιμή του στο 1 ενώ όσο πιο σκούρο τόσο πιο κοντά είναι στο 0. Οι γραμμές αντιστοιχούν στα χαρακτηριστικά και οι στήλες στα διαδοχικά καρέ του βίντεο.

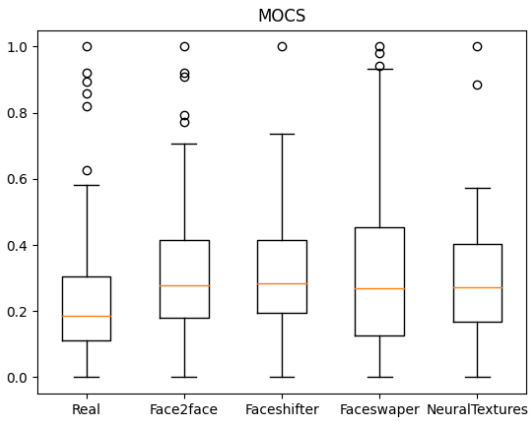
Συγκρίνοντας τα ζεύγη βίντεο για τις 2 προσεγγίσεις παρατηρήθηκε ποιοτικά ότι στην $2^{η}$ προσέγγιση είναι πιο ορατές οι διαφορές ανάμεσα στα αληθινά και τα παραποιημένα βίντεο οπότε επιλέχθηκε να χρησιμοποιηθεί η μέθοδος κανονικοποίησης του δισδιάστατου διανυσματικού χώρου για κάθε χαρακτηριστικό στον άξονα του χρόνου για όλα τα δεδομένα.

Προκειμένου να διαπιστωθεί αν η διαδικασία μετατροπής των προσώπων σε διανύσματα λειτουργεί κανονικά δημιουργήθηκαν τα ιστογράμματα για τους συντελεστές της έκφρασης, της πόζας και του χρώματος για 8 τυχαία επιλεγμένα βίντεο κάθε φορά που όλα απεικονίζουν την ίδια ταυτότητα. Παρατηρήθηκε ότι τα διαγράμματα ακολουθούν σχεδόν κανονική κατανομή και δεν εμφανίζουν ασυνέχειες και, επομένως, η διαδικασία μετατροπής δεν φαίνεται να παρουσιάζει κάποιο πρόβλημα. Παράλληλα έγινε και απεικόνιση εικόνων τυχαίων προσώπων και των αντίστοιχων προσώπων σε 3DMMs και δεν βρέθηκαν παραδείγματα στα οποία η μέθοδος αποτύγχανε κατά μετατροπή των προσώπων. Τα ιστογράμματα για τους συντελεστές μίας από τις ταυτότητες που δοκιμάστηκαν φαίνονται στα παρακάτω διαγράμματα.

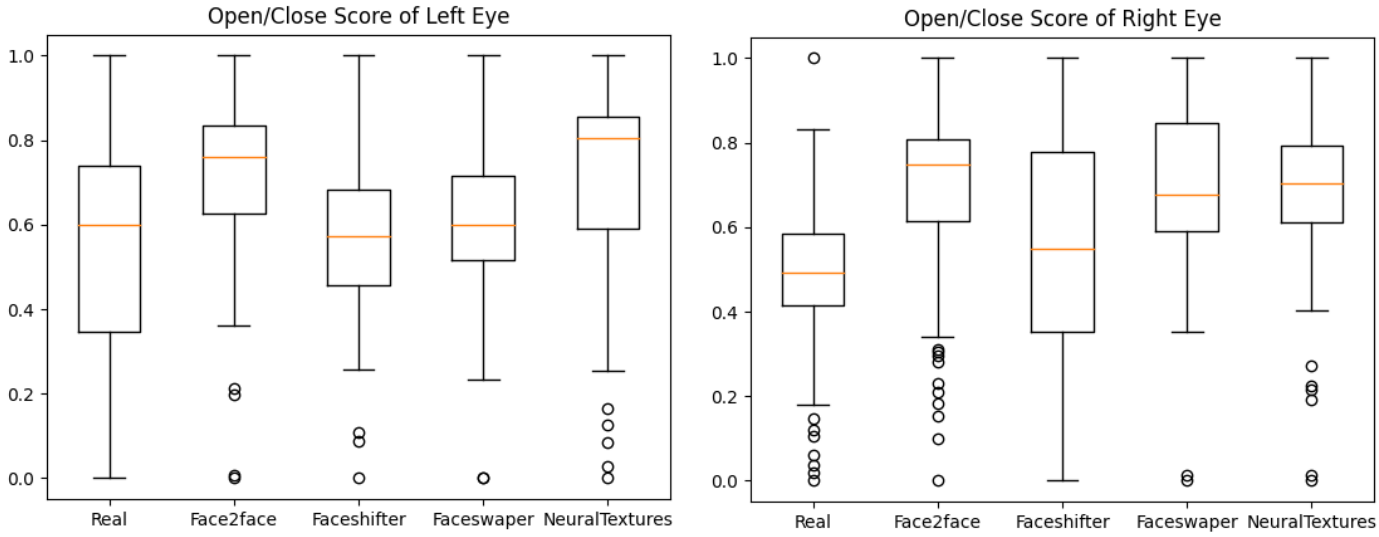


Εικόνα 20: *Ιστογράμματα για τους συντελεστές του σχήματος, της έκφρασης και του χρώματος του προσώπου για 8 βίντεο του ίδιου προσώπου*

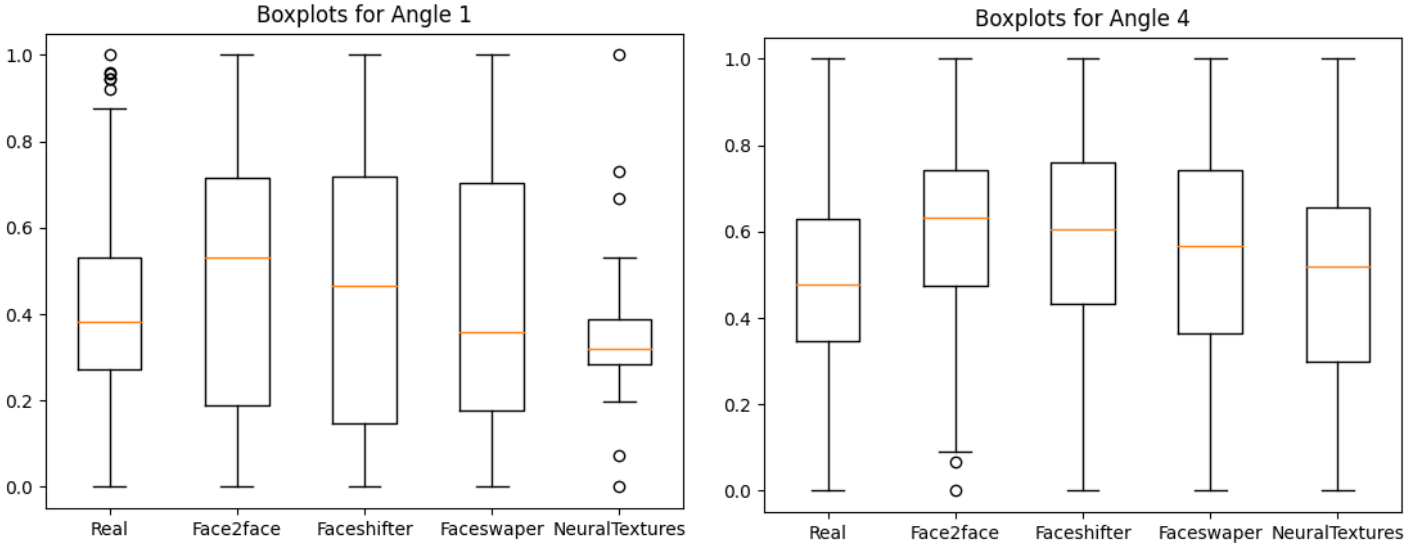
Τέλος, για να διαπιστωθεί αν τα χαρακτηριστικά που παράγονται από τα ορόσημα του προσώπου έχουν πράγματι κάποια διακριτική ικανότητα, υπολογίστηκαν τα χαρακτηριστικά αυτά για ένα αυθεντικό βίντεο και για το ίδιο βίντεο παραποιημένο με τέσσερις διαφορετικές μεθόδους παραγωγής deepfakes (Face2Face, FaceShifter, FaceSwap, NeuralTextures). Τα θηκογράμματα για τα χαρακτηριστικά MOCS, EAR για το αριστερό και το δεξί μάτι και για κάποιες από τις γωνίες που φαίνονται στην Εικόνα 23 παρουσιάζονται στην συνέχεια.



Εικόνα 21: Θηκόγραμμα για το χαρακτηριστικό MOCS



Εικόνα 23: Θηκογράμματα για χαρακτηριστικά EAR_{left}, EAR_{right}



Εικόνα 22: Θηκογράμματα για τις γωνίες 1 και 4 της Εικόνας 23

Παρατηρούμε ότι οι κατανομές των χαρακτηριστικών για τα αυθεντικά βίντεο διαφέρουν από αυτές για τα παραποιημένα και, επομένως, τα χαρακτηριστικά που υπολογίστηκαν από τα ορόσημα του προσώπου (MOCS, EAR και οι γωνίες με σημείο αναφοράς τα ορόσημα του εξωτερικού προσώπου μεταξύ δύο διανυσμάτων με αρχή το σημείο αναφοράς και πέρας τα ορόσημα του εσωτερικού προσώπου) μπορούν να συνεισφέρουν στην αναγνώριση των παραποιημένων βίντεο.

3.2 Παραδοχή για την εκπαίδευση των Μεθόδων

Προκειμένου να καθιερωθεί ένα baseline για όλες τις μεθόδους που θα δοκιμαστούν και να μπορούν να συγκριθούν εύκολα, τα δεδομένα δεν θα αποτελούνται από όλα τα χαρακτηριστικά που αναφέρθηκαν στο παραπάνω κεφάλαιο αλλά μόνο από τους συντελεστές του σχήματος, της έκφρασης και της πόζας που χρησιμοποιεί ο FLAME. Επίσης σε κάθε εποχή δεν θα εκπαιδεύεται το δίκτυο για όλο το βίντεο αλλά για 96 διαδοχικά καρέ από κάθε βίντεο με τυχαία αρχή κάθε φορά. Τα 96 καρέ επιλέχθηκαν διότι αυτός είναι ο ελάχιστος αριθμός καρέ κάθε βίντεο που υπάρχει στο σετ δεδομένων. Επίσης σε κάθε εποχή επιλέχθηκε το μοντέλο να εκπαιδεύεται για διαφορετικά καρέ ώστε να υπάρξει data augmentation και να μειωθεί ο χρόνος εκπαίδευσης. Συνεπώς η είσοδος των μοντέλων, εκτός από την περίπτωση που αναφερθεί διαφορετικά, είναι ένα διάνυσμα που ανήκει στον δισδιάστατο διανυσματικό χώρο $v \epsilon R^{96x156}$ όπου 96 είναι ο αριθμός των διαδοχικών καρέ και 156 είναι ο αριθμός των παραμέτρων του FLAME (100 για το σχήμα, 50 για την έκφραση και 6 για την πόζα).

3.3 1η Μέθοδος – DenseVAE

Η 1η μέθοδος που δοκιμάστηκε χρησιμοποιεί ένα VAE, (κεφ.2.4) για την αναγνώριση των deepfakes. Το VAE δέχεται ως είσοδο μία εικόνα στην οποία, μέσω ενός encoder που αποτελείται από convolution layers μειώνει τις διαστάσεις τις και, στην συνέχεια, χρησιμοποιεί την έξοδο του κωδικοποιητή για να ανακατασκευάσει την εικόνα μέσω ενός αποκωδικοποιητή. Σκοπός του VAE είναι να ελαχιστοποιήσει το σφάλμα MSE ανάμεσα στην εικόνα εισόδου και την ανακατασκευασμένη εικόνα καθώς και να ελαχιστοποιήσει την απόσταση ανάμεσα στις κατανομές πιθανότητας $p_\theta(x|z)$ και $q_\varphi(z|x)$ χρησιμοποιώντας ως σφάλμα απόστασης την KL divergence.

Το VAE στην συγκεκριμένη μέθοδο χρησιμοποιείται ως One Class Classifier ή Outlier Detector όπου θεωρούμε δύο κλάσεις, την κανονική και την όχι κανονική και όλα τα δεδομένα που διαθέτουμε για την εκπαίδευση του μοντέλου ανήκουν αποκλειστικά ή σχεδόν αποκλειστικά στην κανονική κλάση. Σκοπός είναι το μοντέλο να εκπαιδευτεί να αναγνωρίζει τα δείγματα της κανονικής κλάσης και δοσμένου ενός καινούργιου δείγματος , το μοντέλο να μπορεί να αναγνωρίσει αν το δείγμα ανήκει ή όχι στην κανονική κλάση. Στην περίπτωση μας ως κανονική κλάση θεωρούμε τα αυθεντικά βίντεο (που δεν έχουν υποστεί παραποίηση με κάποια μέθοδο deepfake).

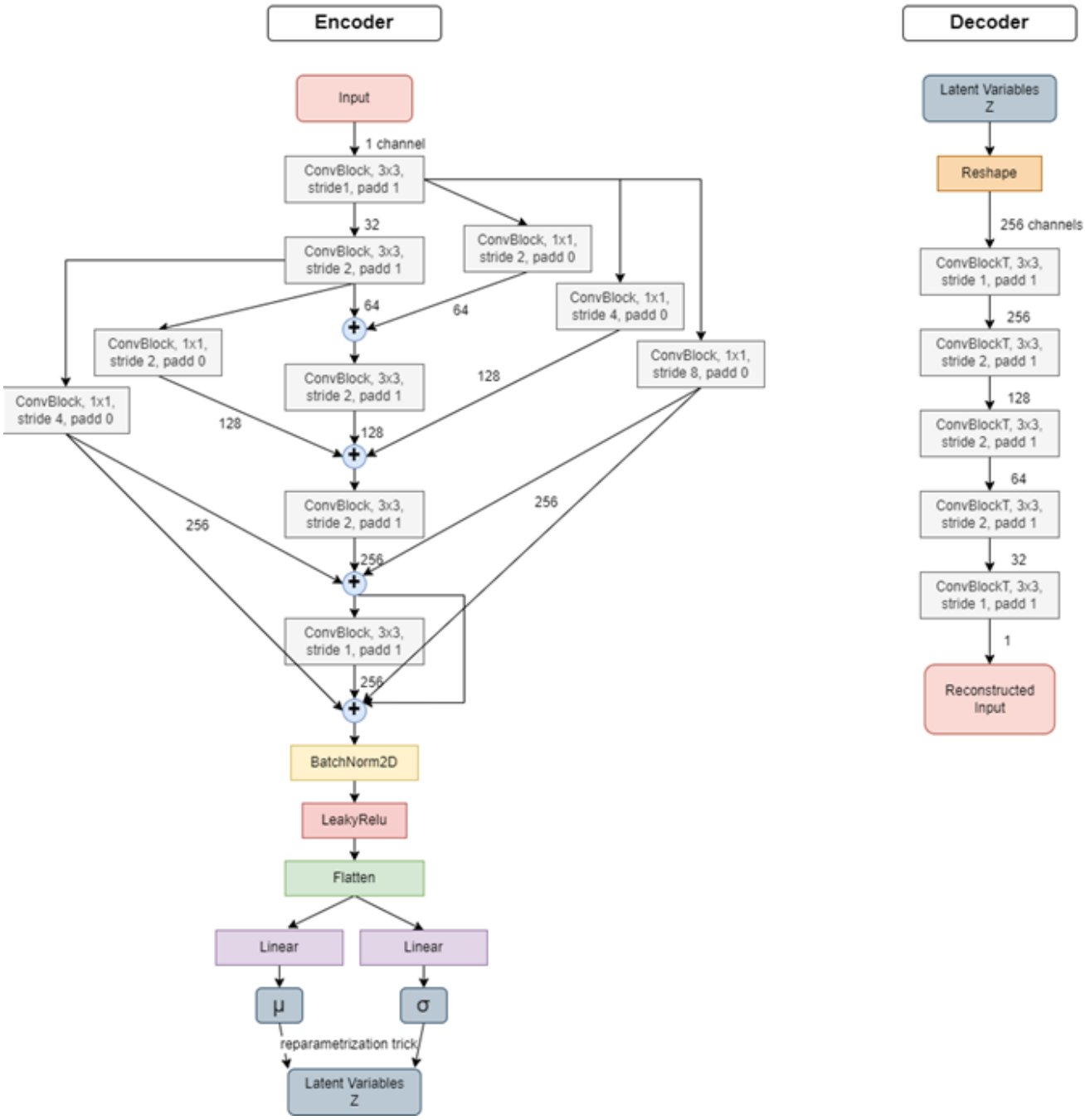
Το VAE μπορεί να χρησιμοποιηθεί ως One Class Classifier αν εκπαιδευτεί να ανακατασκευάζει εικόνες έχοντας ως δεδομένα εκπαίδευσης μόνο αυθεντικά βίντεο (με την μορφή δισδιάστων πινάκων). Με αυτό τον τρόπο μαθαίνει να ανακατασκευάζει με μικρό σφάλμα ανακατασκευής μόνο αυθεντικά βίντεο ενώ θα παρουσιάζει θεωρητικά μεγαλύτερο σφάλμα ανακατασκευής για deepfake βίντεο καθώς δεν έχει εκπαιδευτεί για αυτά. Οπότε θα μπορούσε να βρεθεί ένα όριο MSE ανάμεσα στην είσοδο και την ανακατασκευή της κάτω από το οποίο θα θεωρούμε ότι η είσοδος είναι αυθεντικό βίντεο και πάνω από το οποίο θα θεωρούμε ότι είναι παραποιημένη.

3.3.1 Αρχιτεκτονική

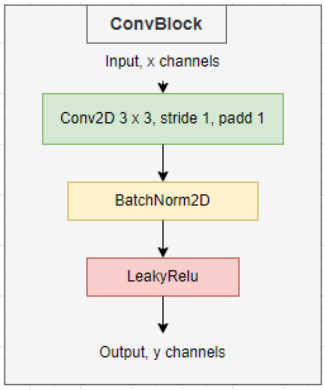
Ύστερα από δοκιμές διαπιστώθηκε ότι η αρχιτεκτονική που επιφέρει το μικρότερο MSE αποτελείται από 5 2D convolutional layers ακολουθούμενα από Batch Normalization[54] και την συνάρτηση ενεργοποίησης Leaky Relu τα καθένα. Στον συνδυασμό των $3^{ων}$ αυτών επιπέδων θα δοθεί η ονομασία *ConvBlock* ως προς διευκόλυνση καθώς θα χρησιμοποιηθεί αρκετές φορές. Ο συνδυασμός αυτός θα ονομάζεται *ConvBlockT* σε περίπτωση που αντί για convolution layer χρησιμοποιείται transpose convolution layer (το οποίο κάνει την αντίθετη πράξη του convolution layer και μπορεί να θεωρηθεί ως η κλίση του 2D convolution layer).

Έτσι ο encoder του VAE θα αποτελείται από 5 ConvBlocks και ο decoder από 5 ConvBlocksT. Επίσης στον encoder όλα τα ConvBlocks θα συνδέονται με την έξοδο όλων των προηγούμενων επιπέδων έτσι ώστε χαρακτηριστικά που έχουν εξαχθεί από προηγούμενα επίπεδα να είναι είσοδοι σε όλα τα επόμενα επίπεδα. Η αρχιτεκτονική αυτή είναι γνωστή ως

DenseNet και εξ αυτής το μοντέλο ονομάζεται *DenseVae***.** Οι συνδέσεις αυτές δεν θα χρησιμοποιηθούν στον decoder καθώς πειραματικά διαπιστώθηκε ότι, ενώ μειώνουν το σφάλμα ανακατασκευής αν χρησιμοποιηθούν στον encoder, δεν συμβάλλουν στην περαιτέρω μείωσή του αν χρησιμοποιηθούν και στον decoder. Το ConvBlock και οι αρχιτεκτονικές του Encoder και του Decoder του VAE παρουσιάζονται παρακάτω.



Εικόνα 24: DenseVae. Αριστερά ο κωδικοποιητής και το στάδιο του Reparameterization Trick. Δεξιά ο αποκωδικοποιητής.



Εικόνα 25: Η δομή ενός *ConvBlock* που χρησιμοποιείται στο DenseVae

3.3.2 Εκπαίδευση

Τελικός στόχος εκπαίδευσης του VAE είναι η ελαχιστοποίηση του κόστους (loss) που αποτελείται από 2 παράγοντες, το σφάλμα MSE ανάμεσα στην είσοδο και την ανακατασκευή της και του όρου KL divergence όπως αυτός περιγράφεται στο κεφ. 2.4. Από το διαθέσιμο σετ δεδομένων, που περιλαμβάνει μόνο βίντεο από αυθεντικά πρόσωπα, το 90% χρησιμοποιείται ως σετ εκπαίδευσης και το 10% ως σετ επικύρωσης (validation set). Ως testing set χρησιμοποιήθηκε ένα σετ δεδομένων που περιέχει τόσο αυθεντικά όσο και παραποιημένα βίντεο. Το μοντέλο εκπαιδεύεται για 300 εποχές χρησιμοποιώντας early stopping. Έτσι μετά την εποχή 100 ελέγχεται το loss στο validation set και, αν δεν βελτιωθεί για 20 εποχές, τότε η εκπαίδευση σταματάει πιο νωρίς από τις 300 εποχές και σώζεται το μοντέλο. Ως optimizer χρησιμοποιήθηκε ο Adam[40] με ρυθμό εκμάθησης 0.002. Ως μετρικές χρησιμοποιήθηκε το σφάλμα ανακατασκευής για την εικόνα μαζί με το KL divergence για τα training και validation sets και η ακρίβεια, το f1 score, το roc auc score, μήτρα σύγχυσης (confusion matrix) και ROC-AUC curve για το testing set.

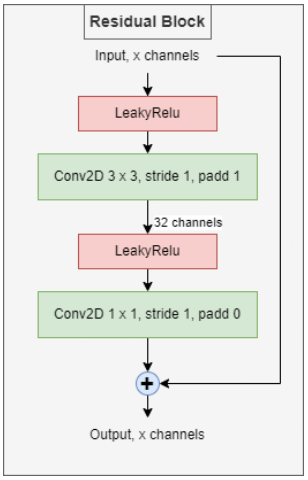
Ο αριθμός των λανθάνων μεταβλητών στο bottleneck layer επιλέγεται να είναι 2000 (13.3% του συνολικού αριθμού των παραμέτρων) ώστε να εξαχθούν χαρακτηριστικά αντιπροσωπευτικά για το βίντεο.

3.4 2η Μέθοδος – VQ VAE

Η 2η μέθοδος που δοκιμάστηκε, η οποία θεωρητικά έχει καλύτερο σφάλμα ανακατασκευής από τα VAE, είναι το VQ-VAE (κεφ. 2.5). Στα VQ-VAE οι πιθανότητες έχουν κατηγορικές κατανομές ενώ στα VAE έχουμε κανονικές και επίσης αντί για λανθάνουσες μεταβλητές χρησιμοποιείται ένας χώρος ενσωμάτωσης (embedding space) που εκπαιδεύεται μέσω Vector Quantization και σκοπός του είναι να φέρει τα embeddings όσο πιο κοντά γίνεται στις εξόδους του Encoder. Όπως και τα VAE, το VQ-VAE δέχεται ως είσοδο μία εικόνα, μειώνει τις διαστάσεις της μέσω ενός κωδικοποιητή, τις μετατρέπει σε μεταβλητές embedding space και με είσοδο τα embeddings ένας αποκωδικοποιητής ανακατασκευάζει την είσοδο. Και σε αυτή τη περίπτωση το μοντέλο λειτουργεί ως One Class Classifier, δηλαδή εκπαιδεύεται να ανακατασκευάζει μόνο αυθεντικά δεδομένα και εντοπίζεται ένα όριο του MSE ανάμεσα στην είσοδο και την ανακατασκευή της πάνω από το οποίο θεωρούμε τα δεδομένα ως παραποιημένα.

3.4.1 Αρχιτεκτονική

Τα VQ-VAE χρειάζονται πιο πολλά convolution layers από τα VAE για να αποφέρουν καλύτερα αποτελέσματα. Για αυτό το λόγο επιλέγεται να χρησιμοποιηθούν 15 convolution layers ακολουθούμενα από την συνάρτηση ενεργοποίησης Leaky Relu για τον κωδικοποιητή και τα ίδια layers αλλά αντικατοπτρισμένα για τον αποκωδικοποιητή. Για να μην εμφανιστεί το πρόβλημα του vanishing gradient (το gradient είναι τόσο μικρό που κατά την διαδικασία της οπισθοδιάδοσης αποτρέπει τα βάρη του δικτύου να αλλάξουν τιμή) και για να μην υπάρξει μείωση της απόδοσης του δικτύου χρησιμοποιούνται **Residual Blocks.** Στα Residual Blocks η είσοδος, αφού περάσει από 2 convolution layers και τα απαραίτητα activation layers, επαναπροστίθεται στην έξοδό τους. Στο δίκτυο δεν χρησιμοποιούνται Batch Normalization Layers καθότι ούτε οι δημιουργοί του μοντέλου έχουν χρησιμοποιήσει. Έτσι τελικά χρησιμοποιούνται 3 convolution layers και 6 Residual Blocks στον κωδικοποιητή και στον αποκωδικοποιητή αντίστοιχα. Η δομή ενός Residual Block και η αρχιτεκτονική του VQ-VAE παρουσιάζεται στις παρακάτω εικόνες.



Εικόνα 26: Το Residual Block που χρησιμοποιείται στο VQ-VAE

Εικόνα 27: Η αρχιτεκτονική του VQ-VAE. Αριστερά ο κωδικοποιητής που αποτελείται και από Residual Blocks. Στο κέντρο ο Vector Quantizer. Δεξιά ο αποκωδικοποιητής.

3.4.2 Εκπαίδευση

Στόχος εκπαίδευσης του VQ-VAE είναι η ελαχιστοποίηση του κόστους το οποίο αποτελείται από 3 παράγοντες, το σφάλμα MSE ανάμεσα στην είσοδο και την ανακατασκευή, το MSE loss ανάμεσα στα embeddings και την έξοδο του Encoder που έχει ως σκοπό την εκπαίδευση του embedding space και ένα commitment loss που στοχεύει στην αποφυγή της υπερβολικής αύξησής των διαστάσεων του embedding space. Και εδώ χρησιμοποιούνται μόνο αυθεντικά βίντεο αφού τελικός σκοπός του μοντέλου είναι να έχει μικρότερο σφάλμα ανακατασκευής για τα αυθεντικά δεδομένα. Από αυτά το 90% χρησιμοποιείται ως σετ εκπαίδευσης και το 10% ως σετ επικύρωσης (validation set). Το testing set θα περιέχει αυθεντικά και τροποποιημένα βίντεο. Το μοντέλο εκπαιδεύτηκε με μέγιστο αριθμό εποχών 300 και early stopping μετά την εποχή 50 με υπομονή 30 εποχές. Εδώ ο αριθμός εποχών είναι μικρότερος από το VAE διότι παρατηρήθηκε ότι το μοντέλο συγκλίνει πιο γρήγορα. Ως αριθμός embeddings επιλέχθηκε το 512 και embedding dimensions 64 που χρησιμοποιούνται στο paper των δημιουργών. Ως optimizer χρησιμοποιήθηκε ο Adam με ρυθμό εκμάθησης 0.0002. Οι μετρικές που χρησιμοποιήθηκαν ήταν το κόστος των 3 παραγόντων για τα training και validation sets και η ακρίβεια, το f1 score, το roc auc score, μήτρα σύγχυσης και ROC-AUC curve για το testing set.

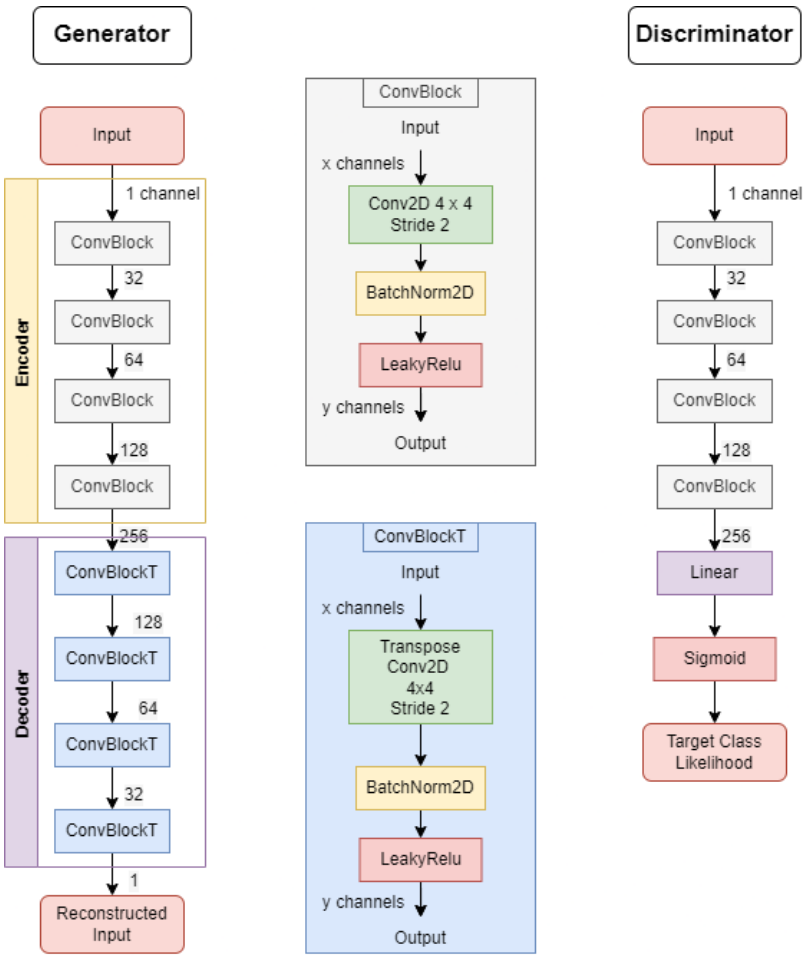
3.5 3η Μέθοδος - GAN

Ως τρίτη μέθοδος επιλέχθηκε να χρησιμοποιηθεί ένα GAN (κεφ. 2.6). Το GAN θα αποτελείται από έναν Generator και έναν Discriminator. Ο Generator είναι ένας Autoencoder που δέχεται στην είσοδό του το αυθεντικό βίντεο και προσπαθεί να το ανακατασκευάσει ώστε να ελαχιστοποιηθεί το σφάλμα ανακατασκευής. Ο Discriminator είναι ένα CNN το οποίο, έχοντας ως είσοδο το δισδιάστο χώρο που αντιπροσωπεύει το βίντεο, προσπαθεί να το κατηγοριοποιήσει ως αυθεντικό η παραποιημένο. Καθώς το σετ δεδομένων περιέχει μόνο αυθεντικά βίντεο, ως παραποιημένα θεωρούνται αυτά που παράγονται από τον Generator. Έτσι ο Generator παράγει εικόνες με τελικό σκοπό να μπερδέψει τον Discriminator και να της κατηγοριοποιήσει ως αυθεντικές, και ο Discriminator εκπαιδεύεται έτσι ώστε να μπορεί να αναγνωρίσει ποια δεδομένα παράγει ο Generator και ποια δεδομένα είναι τα αυθεντικά.

Από τον τρόπο που λειτουργεί το GAN βλέπουμε ότι έχει την ικανότητα να παράξει παραποιημένα δεδομένα χωρίς αυτά να χρειάζεται να υπάρχουν στα δεδομένα εκπαίδευσης. Για αυτόν ακριβώς το λόγο και επιλέχθηκε το GAN ως μία μέθοδος που μπορεί να επιφέρει καλά αποτελέσματα.

3.5.1 Αρχιτεκτονική

Όπως αναφέρθηκε προηγουμένως, το GAN αποτελείται από έναν Generator που παράγει παραποιημένα δεδομένα έχοντας ως είσοδο αυθεντικά και έναν Discriminator που μπορεί να αναγνωρίσει ποια είναι αυθεντικά και ποια όχι. Ο Generator είναι ένας Autoencoder. Για λόγους μείωσης του συνολικού χρόνου εκπαίδευσης και ύστερα από δοκιμές, ο κωδικοποιητής επιλέχθηκε να αποτελείται από 4 convolution layers, το καθένα ακολουθούμενο από Batch Normalization και Leaky ReLu activation layers. Για τον αποκωδικοποιητή χρησιμοποιείται η αντικατοπτρισμένη αρχιτεκτονική και αντί για convolution layers χρησιμοποιούνται transpose convolution layers. Ο Discriminator έχει την ίδια αρχιτεκτονική με τον κωδικοποιητή του Autoencoder με την διαφορά ότι στο τέλος του δικτύου προστίθεται ένα πλήρως συνδεδεμένο επίπεδο ακολουθούμενο από την σιγμοειδή συνάρτηση ενεργοποίησης έτσι ώστε το δίκτυο να μπορεί να χρησιμοποιηθεί για δυαδική ταξινόμηση (binary classification). Οι αρχιτεκτονικές του Generator και του Discriminator του GAN παρουσιάζονται παρακάτω.



Εικόνα 28: Οι αρχιτεκτονικές των Generator και Discriminator του GAN. Αριστερά η αρχιτεκτονική του Generator. Στο κέντρο τα ConvBlock και ConvBlockT που χρησιμοποιούνται στο GAN. Αριστερά ο Discriminator.

3.5.2 Εκπαίδευση

Κατά την εκπαίδευση του GAN εκπαιδεύονται και τα δύο δίκτυα, ο Generator και ο Discriminator. Στόχος της εκπαίδευσης του Discriminator είναι η μεγιστοποίηση της πιθανότητας κατηγοριοποίησης πραγματικών εισόδων ως αληθινές και εισόδων παραγμένων από τον Generator ως ψεύτικες. Από την μεριά του Generator σκοπός του είναι να παράξει εξόδους όσο πιο ρεαλιστικά γίνεται και να μπερδέψει τον Discriminator ώστε να νομίζει πως είναι πραγματικά δεδομένα. Στόχος εκπαίδευσης του GAN είναι αυτός που περιγράφεται στο [2] δηλαδή:

$$L_{G,D} = \min_{\mathrm{G}} \max_{\mathrm{D}} V(D,G) = E_{\boldsymbol{x}\sim \boldsymbol{p_{data}}(x)}[\log D(\boldsymbol{x})] + E_{\boldsymbol{z}\sim \boldsymbol{p_z}(\boldsymbol{z})+\boldsymbol{\eta}}[\log(1 - D\big(G(\boldsymbol{z})\big))]$$

όπου ο Generator προσπαθεί να ελαχιστοποιήσει το V και ο Discriminator να το μεγιστοποιήσει. Η διαφορά από τον στόχο εκπαίδευσης του αρχικού paper είναι ότι στην είσοδο του Generator προστίθεται και θόρυβος από την κανονική κατανομή $\eta \sim N(0, \sigma^2 \boldsymbol{I})$ έτσι ώστε η μέθοδος να είναι πιο ανθεκτική σε θόρυβο και παραμορφώσεις στην είσοδο. Επίσης θέλουμε η έξοδος του Generator να μοιάζει με τα δεδομένα εισόδου οπότε προστίθεται στο συνολικό σφάλμα και το MSE της ανακατασκευασμένης με την αρχική είσοδο του Generator L_G. Έτσι συνολικός στόχος του GAN είναι να ελαχιστοποιήσει το σφάλμα:

$$L = L_{G,D} + \lambda L_G$$

όπου λ>0 παράμετρος που ελέγχει την σημασία των δύο όρων.

Σε κάθε εποχή εκπαιδεύονται διαδοχικά ο Discriminator και ο Generator. Αρχικά, κρατώντας σταθερές τις παραμέτρους του Generator, ο Generator παράγει την παραποιημένη είσοδο από την αυθεντική μαζί με τον θόρυβο. Οι δύο 'εικόνες' δίνονται στον Discriminator και υπολογίζεται η cross-entropy ανάμεσα στις ετικέτες των εισόδων (0 για την έξοδο του Generator, 1 για τα αυθεντικά δεδομένα) και τις προβλέψεις του Discriminator. Στόχος είναι η μεγιστοποίηση του:

$$E_{\boldsymbol{x}\sim \boldsymbol{p_{data}}(x)}[\log D(\boldsymbol{x})] + E_{\boldsymbol{z}\sim \boldsymbol{p_z}(\boldsymbol{z})}[\log(1 - D\big(G(\boldsymbol{z})\big))]$$

Μετά τον υπολογισμό του loss αυτό περνάει μέσω backpropagation στο δίκτυο και ανανεώνονται τα βάρη.

Στην συνέχεια διατηρούνται σταθερές οι παράμετροι του Discriminator. Η έξοδος του Generator εισέρχεται στον Discriminator και υπολογίζεται η cross-entropy ανάμεσα στην ετικέτα της παραποιημένης 'εικόνας' και της πρόβλεψης του Discriminator. Επίσης υπολογίζεται το MSE ανάμεσα στην είσοδο του Generator και της ανακατασκευασμένης εισόδου. Στόχος εκπαίδευσης του Generator είναι η ελαχιστοποίηση του:

$$E_{\boldsymbol{z}\sim \boldsymbol{p_z}(\boldsymbol{z})}[\log(1 - D\big(G(\boldsymbol{z})\big))] + \lambda \|X - X'\|^2$$

όπου ο 2ος όρος είναι το MSE.

Η παραπάνω διαδικασία επαναλαμβάνεται για το πολύ 300 εποχές και μπορεί να σταματήσει σε λιγότερές εποχές αν το σφάλμα ανακατασκευής γίνει μικρότερο από 0.005. Η παράμετρος λ τίθεται ίση με 0.2. Ως optimizer και για τα 2 μοντέλα επιλέχθηκε ο RMSProp[45] με παραμέτρους α=0.9, weight decay=09 και learning rate=0.0002. Το 90% του σετ δεδομένων χρησιμοποιήθηκε ως σετ εκπαίδευσης και το 10% ως σετ επικύρωσης (validation set). Οι μετρικές που χρησιμοποιήθηκαν ήταν το generator loss, discriminator loss και recreation loss για τα training και validation sets όπως αυτά περιεγράφηκαν παραπάνω και η ακρίβεια, το f1 score, το roc auc score, μήτρα σύγχυσης και ROC-AUC curve για το testing set.

3.6 4η Μέθοδος – OCGAN

Ως τέταρτη μέθοδος επιλέχθηκε να χρησιμοποιηθεί ένα πιο σύνθετο GAN, το OCGAN[46]. Στο OCGAN τελικός σκοπός είναι η εκπαίδευση ενός Autoencoder, που παίζει τον ρόλο του Generator, ο οποίος θα λειτουργεί ως One Class Classifier και θα έχει λανθάνουσες μεταβλητές που θα είναι περιορισμένες έτσι ώστε να μπορούν να αντιπροσωπεύουν μόνο βίντεο που θα είναι αυθεντικά. Με αυτό τον τρόπο τα παραποιημένα βίντεο που θα εισέρχονται στον Generator θα έχουν πολύ μεγάλο σφάλμα ανακατασκευής συγκριτικά με τα αυθεντικά.

Το OCGAN αποτελείται από 4 μέρη. Τον Generator που, όπως αναφέρθηκε προηγουμένως, είναι ένας Autoencoder ο οποίος τροφοδοτείται με την αυθεντική ‘εικόνα’ μαζί με Γκαουσιανό θόρυβο (η προσθήκη θορύβου βοηθάει στην καλύτερη γενίκευση του Autoencoder) και ανακατασκευάζει την εικόνα χωρίς θόρυβο. Οι λανθάνουσες μεταβλητές του Autoencoder φράσσονται μέσω της συνάρτησης ενεργοποίησης tanh στο εύρος (-1,1) ώστε να γίνεται καλύτερη δειγματοληψία. Το GAN δεν περιέχει έναν Discriminator αλλά δύο, τον Latent Discriminator και τον Visual Discriminator.

Καθώς σκοπός της μεθόδου είναι η δημιουργία ενός χώρου λανθάνων μεταβλητών όπου όλες οι μεταβλητές θα αντιπροσωπεύουν τα αυθεντικά βίντεο, χρησιμοποιείται ο Latent Discriminator, ο οποίος δέχεται ως είσοδο τις λανθάνουσες μεταβλητές της εξόδου του Encoder από τον Generator και επιστρέφει την πιθανότητα να ανήκουν σε αυθεντικό βίντεο. Επίσης προκειμένου ο Generator να παράγει μόνο αυθεντικά βίντεο χρησιμοποιείται ο Visual Discriminator που λαμβάνει ως είσοδο την ανακατασκευασμένη εικόνα από τον Generator και επιστρέφει την πιθανότητα να αντιπροσωπεύει αυθεντικό βίντεο.

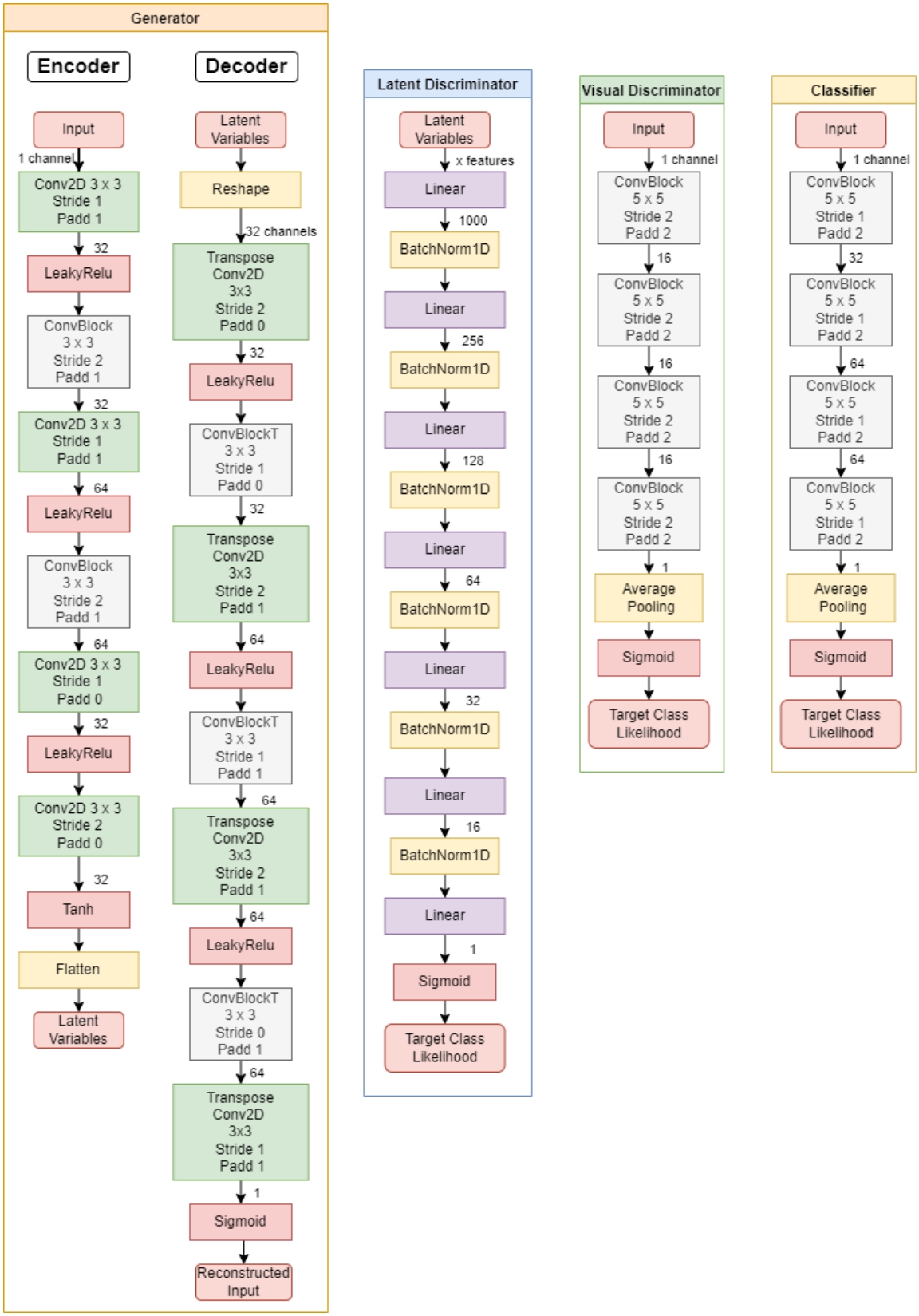
Τέλος το OCGAN περιλαμβάνει και έναν Classifier. Κατά την διαδικασία της εκπαίδευσης δεν μπορούν να ελεγχθούν όλοι οι συνδυασμοί των λανθάνων μεταβλητών έτσι ώστε να επιβεβαιωθεί πως παράγουν αυθεντικά δεδομένα. Κάτι τέτοιο θα μπορούσε να συμβεί αν μειωνόταν ο αριθμός των λανθάνων μεταβλητών. Τότε όμως θα μειωνόταν η απόδοση του μοντέλου. Για αυτό το λόγο χρησιμοποιείται ένας ισχυρός (συγκριτικά με τους Discriminators) Classifier που δεν συμμετέχει στην συνεργατική εκπαίδευση του Generator

και των Discriminators. Ο Classifier δέχεται ως είσοδο ‘εικόνες’ που έχουν ανακατασκευαστεί μέσω του Generator από αυθεντικά δεδομένα και ‘εικόνες’ που παρήχθησαν από συνδυασμό τυχαίων λανθάνων μεταβλητών. Στόχος του είναι να κατατάξει τις πρώτες εικόνες ως αυθεντικές και αυτές που παράγονται τυχαία ως ψεύτικες.

3.6.1 Αρχιτεκτονική

Το OCGAN είναι σχεδιασμένο έτσι ώστε να λειτουργεί με τετράγωνες εικόνες ως εισόδους. Ωστόσο τα δεδομένα για τα οποία θέλουμε να εκπαιδεύσουμε το δίκτυο δεν έχουν τις ίδιες διαστάσεις και στους δύο άξονες. Για αυτό το λόγο πραγματοποιήθηκαν τροποποιήσεις στα δίκτυα του Generator και του Latent Discriminator. Ο Generator είναι ένας Autoencoder, του οποίου ο κωδικοποιητής αποτελείται από 6 convolution layers ακολουθούμενα ανά 2 από Batch Normalization και Leaky ReLu Activation Layers. Στο τέλος του κωδικοποιητή προστίθεται ένα Tanh Activation Layer ώστε οι λανθάνουσες μεταβλητές να είναι στο εύρος (-1, 1). Ο αποκωδικοποιητής αποτελεί αντικατοπτρισμό του κωδικοποιητή χωρίς το τελευταίο επίπεδο και με αντικατεστημένα τα convolution layers από transpose convolution layers. Στο τέλος του προστίθεται άλλον ένα transpose convolution layer και, επίσης, για να προσφέρει μη γραμμικότητα στο δίκτυο προστίθεται ένα sigmoid layer.

Ο Latent Discriminator αποτελείται από 7 πλήρως συνδεδεμένα επίπεδα ακολουθούμενα από Batch Normalization Layers και καταλήγει σε ένα sigmoid layer καθώς σκοπός του είναι να προβλέψει την πιθανότητα οι λανθάνουσες μεταβλητές να ανήκουν σε αυθεντικό βίντεο. Ο Visual Discriminator αποτελείται από 4 convolution layers ακολουθούμενα από Batch Normalization Layers και Leaky ReLu Activation. Στο τέλος προτίθεται ένα Average Pooling Layer και ένα Sigmoid Activation Layer. Τέλος ο Classifier έχει την ίδια δομή με τον Visual Discriminator αλλά με διαφορετικές επιλογές στις παραμέτρους. Οι αρχιτεκτονικές των δικτύων του OCGAN παρουσιάζονται παρακάτω. Στα σχήματα χρησιμοποιούνται τα blocks ConvBlock και ConvBlockT τα οποία φαίνονται στην παραπάνω εικόνα.



Εικόνα 29: Οι αρχιτεκτονικές των δικτύων του OCGAN. Από αριστερά προς τα δεξιά: Ο κωδικοποιητής και ο αποκωδικοποιητής που περιλαμβάνει ο Generator, ο Latent Discriminator, ο Visual Discriminator και ο Classifier.

3.6.2 Εκπαίδευση

Αρχικά εκπαιδεύεται μόνο ο Generator του GAN για τις πρώτες 20 εποχές έτσι ώστε να μάθει να ανακατασκευάζει τις εικόνες σε ικανοποιητικό βαθμό πριν αρχίσει να εκπαιδεύεται μαζί με τους Discriminators. Ο Generator εκπαιδεύεται να ελαχιστοποιεί το MSE ανάμεσα στην είσοδο και την ανακατασκευή, δηλαδή:

$$\min l_{MSE} = \left\| x - D\big(E(x+n)\big) \right\|_2^2,$$

όπου D και E ο Decoder και ο Encoder του Generator αντίστοιχα, χ η αυθεντική 'εικόνα' και n ο θόρυβος που προστίθεται στον Generator.

Αφού περάσουν οι 20 εποχές χρησιμοποιούνται και τα άλλα δίκτυα στην εκπαίδευση. Αρχικά οι παράμετροι όλων των δικτύων εκτός του Classifier διατηρούνται σταθεροί. Ο Classifier τροφοδοτείται με ανακατασκευασμένα αυθεντικά δεδομένα από τον Generator που παράγονται από αυθεντικές 'εικόνες' και από ανακατασκευασμένα δεδομένα του Generator που προήλθαν από τυχαίες λανθάνουσες μεταβλητές στην είσοδο του Decoder του Generator. Στόχος της εκπαίδευσης του Classifier είναι να κατατάξει τις αυθεντικές ανακατασκευασμένες 'εικόνες' ως πραγματικά δεδομένα και τις 'εικόνες' από τυχαίες μεταβλητές ως ψεύτικες. Αυτό επιτυγχάνεται με μεγιστοποίηση της μεταξύ τους εντροπίας. Έτσι τα βάρη του Classifier ενημερώνονται μέσω backpropagation.

Στην συνέχεια εκπαιδεύονται διαδοχικά οι Discriminators και ο Generator. Οι παράμετροι του Generator και του Classifier διατηρούνται σταθεροί. Στόχος του Latent Discriminator είναι να μπορεί να αναγνωρίσει ποιες λανθάνουσες μεταβλητές του Generator έχουν προέλθει από κωδικοποίηση αυθεντικών δεδομένων και ποιες προέρχονται απλώς από τυχαία κατανομή. Στο πλαίσιο αυτό στόχος του είναι να ελαχιστοποιηθεί το :

$$l_{latent} = -(E_{s\sim U(-1,1)}[\log D_l(S)] + E_{x\sim p_x}[\log(1 - D_l\big(En(x+n)\big))]),$$

όπου U(-1, 1) τα τυχαία δείγματα από τις λανθάνουσες μεταβλητές και D_l ο Latent Discriminator. Παρόμοιος είναι και ο στόχος του Visual Discriminator (D_v) ο οποίος δέχεται ως είσοδο 'εικόνες' που δημιουργήθηκαν από ανακατασκευή αυθεντικών δεδομένων από τον Generator και εικόνες που παρήχθησαν από τυχαίες λανθάνουσες μεταβλητές που τροφοδότησαν τον Decoder του Generator. Στόχος του είναι να κατηγοριοποιήσει τις πρώτες ως αυθεντικές και τις δεύτερες ως ψεύτικες, οπότε προσπαθεί να ελαχιστοποιήσει το:

$$l_{visual} = -(E_{s\sim U(-1,1)}[\log D_v(De(S))] + E_{x\sim p_x}[\log(1 - D_v(x))])$$

Έτσι συνολικός στόχος εκπαίδευσης των Discriminators είναι η ελαχιστοποίηση του όρου $l_{latent} + l_{visual}$ ενώ οι παράμετροι του Generator διατηρούνται σταθεροί.

Μετά την εκπαίδευση των Discriminators, οι παράμετροι όλων των δικτύων διατηρούνται σταθεροί. Οι 'εικόνες' που παράγονται από τυχαίες λανθάνουσες μεταβλητές $l_2 = U(-1, 1)$ του Generator τροφοδοτούνται 5 φορές στον Classifier και γίνονται backpropagate μέσω του δικτύου ώστε να βελτιωθεί η πιθανότητα που έχουν οι l_2 να μπερδέψουν τον Classifier ώστε να κατηγοριοποιήσει τις εικόνες που παράγονται από αυτές ως αυθεντικές.

Στην συνέχεια οι παράμετροι όλων των δικτύων εκτός του Generator διατηρούνται σταθεροί. Πραγματοποιείται ξανά ο υπολογισμός του όρου $l_{latent} + l_{visual}$ για τις λανθάνουσες μεταβλητές l_2 που υπολογίστηκαν παραπάνω. Επίσης υπολογίζεται και το l_{MSE} ανάμεσα στα αυθεντικά δεδομένα που χρησιμοποιήθηκαν παραπάνω και την ανακατασκευή τους. Τελικά το συνολικό σφάλμα που υπολογίστηκε $l_{latent} + l_{visual} + \lambda l_{MSE}$ γίνεται backpropagate και ενημερώνονται οι παράμετροι του Generator.

Το OCGAN εκπαιδεύεται για μέγιστο 150 εποχές αλλά η εκπαίδευση μπορεί να σταματήσει πιο νωρίς αν το validation loss δεν βελτιώνεται ύστερα από αρκετές εποχές. Το learning rate μειώνεται στο μισό κάθε 20 εποχές. Το μοντέλο εκπαιδεύεται μόνο για αυθεντικά δεδομένα. Ως validation και testing set χρησιμοποιείται ένα σετ δεδομένων που περιέχει τόσο αυθεντικά όσο και παραποιημένα βίντεο.

Ως optimizer για όλα τα δίκτυα χρησιμοποιείται ο Adam με αρχικό learning rate 0.001 που υποδιπλασιάζεται κάθε 20 εποχές. Το κόστος που παρακολουθείται κατά την εκπαίδευση είναι το $l_{latent} + l_{visual} + 10l_{MSE}$ που χρησιμοποιείται για την εκπαίδευση του Generator. Η μετρική που παρακολουθείται κατά το validation είναι το Roc Auc Score και η ακρίβεια, το f1 score, το roc auc score, μήτρα σύγχυσης και ROC-AUC curve για το testing set.

3.7 5η Μέθοδος – Binary Classification

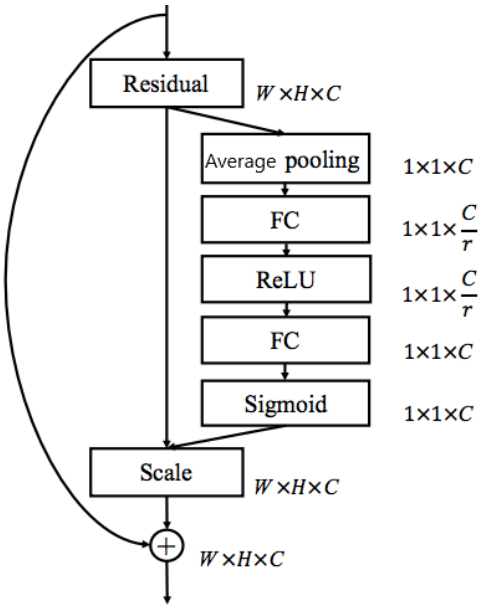
Στην 5η και τελευταίο μέθοδο που χρησιμοποιήθηκε ακολουθήσαμε μία πιο παραδοσιακή προσέγγιση στον χώρο των νευρωνικών δικτύων, αυτή του Binary Classification. Πλέον το πρόβλημα δεν είναι να εκπαιδευτεί ένα μοντέλο που να μπορεί να αναγνωρίσει πότε ένα καινούργιο δείγμα ανήκει ή όχι στην μία κλάση στην οποία έχει εκπαιδευτεί. Αντίθετα το μοντέλο εκπαιδεύεται με δεδομένα που ανήκουν σε δύο κλάσεις (στην περίπτωση μας βίντεο που ανήκουν σε αυθεντικά πρόσωπα και βίντεο προσώπων που έχουν υποστεί παραποίηση) και σκοπός του είναι, δοσμένου ενός καινούργιου δείγματος, να το κατηγοριοποιήσει στην σωστή κλάση, δηλαδή να επιστρέφει την πιθανότητα να ανήκει ή όχι στην κλάση των αυθεντικών βίντεο.

Στο πλαίσιο αυτό δεν θα χρησιμοποιηθεί το σετ δεδομένων που αποτελείται μόνο από αυθεντικά βίντεο αλλά θα συγχωνευτεί με ένα ακόμη σετ δεδομένων που περιλαμβάνει τόσο αυθεντικά όσο και παραποιημένα βίντεο και το μοντέλο θα εκπαιδευτεί στο καινούργιο σετ δεδομένων.

3.7.1 Αρχιτεκτονική

Το δίκτυο που επιλέχθηκε να χρησιμοποιηθεί για το Binary Classification είναι ένα Βαθύ Συνελικτικό Δίκτυο και, συγκεκριμένα, το δίκτυο EfficientNetv2_s από την οικογένεια δικτύων EfficientNet[47]. Το EfficientNet ονομάστηκε έτσι ακριβώς για την αποδοτικότητά του, καθώς εκπαιδεύεται αρκετά πιο γρήγορα από άλλα state of the art μοντέλα και ταυτόχρονα είναι έως και 7 φορές μικρότερο. Πρόκειται για ένα βαθύ συνελικτικό νευρωνικό δίκτυο του οποίου κύρια δομικά χαρακτηριστικά είναι τα MBConv, Fused-MBConv και SE blocks.

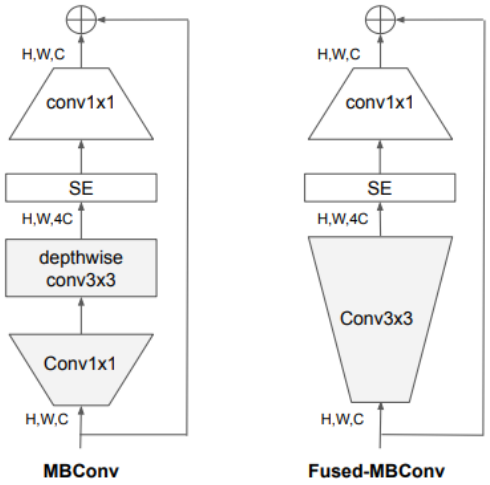
- ***Squeeze-Excitation Block*** *(SE Block)*: Τα CNN, κατά την δημιουργία του χάρτη χαρακτηριστικών της εξόδου χρησιμοποιούν όλα τα κανάλια ισάξια. Ωστόσο η πληροφορία που βρίσκεται σε κάποια από τα κανάλια μπορεί να είναι πιο σημαντική από τα υπόλοιπα κανάλια. Για αυτό το λόγο τα SE Blocks προσθέτουν έναν μηχανισμό στα CNN έτσι ώστε τα βάρη για το κάθε κανάλι να προσαρμόζονται κάθε φορά με βάση το περιεχόμενο του καναλιού. Έτσι τα SE Bocks αποτελούνται από ένα επίπεδο Average Pooling που συμπιέζει το feature map κάθε καναλιού σε ένα μόνο αριθμό για κάθε κανάλι και από 2 πλήρως συνδεδεμένα επίπεδα τα οποία επιστρέφουν τα βάρη που θα χρησιμοποιηθούν για κάθε κανάλι, ανάλογα με την σημασία του περιεχομένου του καναλιού. Η δομή ενός SE Block που χρησιμοποιείται σε ένα Residual Block παρουσιάζεται παρακάτω.



Εικόνα 30: Η δομή ενός SE Block

- ***MBConv****:* Πρόκειται για ένα Inverse Residual Block που αρχικά προτάθηκε στο MobileNet. Τα Residual Blocks αποτελούνται συνήθως από 3 convolution layers. Το πρώτο μειώνει τον αρχικό αριθμό των καναλιών, το δεύτερο πραγματοποιεί ένα 3x3 convolution και το 3ο επαναφέρει τον αριθμό των καναλιών στον αρχικό έτσι ώστε να μπορέσει να προστεθεί η είσοδος στην έξοδο του block. Η χρήση της τεχνικής αυτής βοηθάει σημαντικά στην εκπαίδευση convolutional δικτύων καθώς επιτρέπει στο gradient να διαδοθεί μέσω πολλών επιπέδων. Αντίθετα, στο Inverse Residual Block το πρώτο convolution layer αυξάνει τον αριθμό των καναλιών της εισόδου και το τρίτο το επαναφέρει στον αρχικό αριθμό ώστε να μπορεί να προστεθεί η είσοδος. Επίσης στο δεύτερο επίπεδο αντί για ένα απλό convolution layer χρησιμοποιείται ένα Depth-Wise convolution layer στο οποίο κάθε φίλτρο χρησιμοποιείται μόνο σε ένα κανάλι εισόδου και όχι σε όλα όπως συμβαίνει στο απλό convolution. Με το Depth-Wise convolution μειώνεται ο αριθμός των παραμέτρων του μοντέλου. Τέλος μετά το Depth-Wise convolution επίπεδο προστίθεται ένα SE Block ώστε να αλλάξουν τα βάρη για το κάθε κανάλι του επιπέδου.
- ***Fused-MBConv***: Πρόκειται για ένα MBConv Block όπου το 1ο convolution επίπεδο αύξησης καναλιών και το 2ο επίπεδο 2D convolution έχουν ενωθεί σε ένα μόνο επίπεδο. Αυτό συμβαίνει για να αυξηθεί η ταχύτητα εκπαίδευσης του δικτύου.

Τα MBConv και Fused-MBConv Blocks που χρησιμοποιούνται στο EfficientNet παρουσιάζονται παρακάτω.



Εικόνα 31: MBConv και Fused-MBConv Blocks

Το EfficientNetv2 προσπαθεί να βελτιώσει το EfficientNet που είχε προταθεί νωρίτερα. Προβλήματα που αντιμετώπιζε το αρχικό δίκτυο είναι η χρήση εικόνων μεγάλων διαστάσεων που κάνουν την εκπαίδευση αργή και τη χρήση Depthwise convolution, η οποία είναι αργή στα αρχικά επίπεδα αλλά αποτελεσματική στα μετέπειτα επίπεδα. Έτσι οι δημιουργοί του δικτύου χρησιμοποιούν έναν NAS (Neural Architecture Search) για να βρουν τον συνδυασμό των παραμέτρων (χρήση MBConv ή Fused-MBConv, αριθμός επιπέδων, μέγεθος πυρήνα, αναλογία επέκτασης μέσα στο MBConv) που βελτιστοποιεί τόσο την απόδοση του δικτύου όσο και την ταχύτητα εκπαίδευσης. Έτσι δημιουργήθηκε το EfficientNetv2_s που αποτελείται από 22M παραμέτρους και του οποία η αρχιτεκτονική φαίνεται στον παρακάτω πίνακα.

Επίπεδο	**Stride**	**Αρ. Καναλιών**	**Αρ. Επιπέδων**
2DConv 3x3	2	24	1
Fused-MBConv 3χ3 Αναλογία Επέκτασης 1	1	24	2
Fused-MBConv 3χ3 Αναλογία Επέκτασης 4	2	48	4
Fused-MBConv 3χ3 Αναλογία Επέκτασης 4	2	64	4
MBConv 3x3 with SE Αναλογία Επέκτασης 4	2	128	6
MBConv 3x3 with SE Αναλογία Επέκτασης 6	1	160	9
MBConv 3x3 with SE Αναλογία Επέκτασης 6	2	256	15
2DConv 1x1 & Pooling & Fully Connected Layer	-	128	1

Πίνακας 1: Αρχιτεκτονική του Δικτύου EfficientNetv2_s

3.7.2 Εκπαίδευση

Το μοντέλο εκπαιδεύεται σε ένα σετ δεδομένων που περιέχει τόσο αυθεντικά όσο και παραποιημένα βίντεο. Τελικός στόχος εκπαίδευσης του δικτύου είναι να μπορεί να κατηγοριοποιήσει νέα δεδομένα εκτός του σετ εκπαίδευσης ως αυθεντικά ή παραποιημένα, να προβλέψει, δηλαδή, την κλάση στην οποία ανήκουν. Έτσι το loss που χρησιμοποιείται κατά την εκπαίδευση είναι η Binary Cross-Entropy. Το 80% των δεδομένων χρησιμοποιούνται ως δεδομένα εκπαίδευσης και το 20% ως validation και testing sets. Το μοντέλο εκπαιδεύεται για 300 το πολύ εποχές και χρησιμοποιείται early stopping μετά τις 100 εποχές ώστε να σταματάει η εκπαίδευση του μοντέλου αν δεν βελτιώνεται η απόδοση του για 20 διαδοχικές εποχές. Επειδή στο σετ δεδομένων που χρησιμοποιείται ο αριθμός των αυθεντικών δεδομένων είναι πολύ μεγαλύτερος από αυτόν των παραποιημένων, χρησιμοποιείται ο WeightedRandomSampler της PyTorch έτσι ώστε σε κάθε batch κατά την εκπαίδευση να υπάρχει περίπου ο ίδιος αριθμός βίντεο από τις δύο κλάσεις. Ως optimizer χρησιμοποιείται ο Adam με learning rate 0.0005. Ως μετρικές χρησιμοποιήθηκε το κόστος cross-entropy ανάμεσα στις πραγματικές και προβλεπόμενες ετικέτες για τα training και validation sets και η ακρίβεια, το f1 score, το roc auc score, μήτρα σύγχυσης και ROC-AUC curve για το testing set.

Επίσης σε αυτό το μοντέλο, καθώς παρουσιάζει την καλύτερη ακρίβεια συγκριτικά με τα υπόλοιπα, πραγματοποιήθηκαν πειράματα ως προς τον αριθμό των καρέ των βίντεο που χρησιμοποιούνται για την εκπαίδευση και ως προς το ποια χαρακτηριστικά χρησιμοποιούνται για να βρεθεί ο συνδυασμός που επιφέρει την καλύτερη απόδοση.

Κεφάλαιο 4 Πειράματα και Αποτελέσματα

4.1 Σύνολο Δεδομένων (Datasets)

Για την εκπαίδευση και για το testing των Μεθόδων που αναλύθηκαν στο προηγούμενο κεφάλαιο χρησιμοποιήθηκαν δύο Σύνολα Δεδομένων. Το πρώτο dataset είναι το **VoxCeleb2**[9] το οποίο αποτελείται από 145.569 βίντεο από 5.994 διαφορετικές ταυτότητες. Τα βίντεο έχουν επεξεργαστεί ώστε το πρόσωπο του ομιλητή να βρίσκεται στο κέντρο του βίντεο. Όλα τα βίντεο έχουν ρυθμό ανανέωσης 25 fps και η ελάχιστη διάρκεια βίντεο που υπάρχει στο Dataset είναι 96 καρέ. Το δεύτερο Dataset που χρησιμοποιείται είναι το **FaceForensics++**[48] που αποτελείται από 1.000 αυθεντικά βίντεο από το YouTube και που το καθένα από αυτά έχει παραποιηθεί από 4 αυτόματες μεθόδους χειραγώγησης προσώπου: Deepfakes[55], Face2Face[56], FaceSwap[57] και NeuralTextures[58]. Επομένως το Dataset αποτελείται συνολικά από 1.000 αυθεντικά και 4.000 παραποιημένα βίντεο.

Τα Datasets χρησιμοποιήθηκαν με διάφορους τρόπους ανάλογα με την εκάστοτε μέθοδο υπό εξέταση. Στις πρώτες τέσσερις μεθόδους που λειτουργούν ως One Class Classifiers και εκπαιδεύονται μόνο με αυθεντικά δεδομένα, μόνο το VoxCeleb2 χρησιμοποιήθηκε για την εκπαίδευση. Το 90% των δεδομένων του VoxCeleb2 χρησιμοποιήθηκαν για την εκπαίδευση και το 10% για το validation. Για το testing χρησιμοποιήθηκαν τα 1000 αυθεντικά βίντεο από το FaceForensics++ καθώς και τα 1000 παραποιημένα βίντεο που παρήχθησαν με την μέθοδο Deepfakes. Αυτή η επιλογή έγινε έτσι ώστε ο αριθμός των βίντεο που υπάρχουν στις δύο κλάσεις να είναι ο ίδιος ώστε να εξαχθούν πιο σωστά αποτελέσματα και προκειμένου να μπορούν να συγκριθούν εύκολα οι μέθοδοι μεταξύ τους.

Αντίθετα για την τελευταία μέθοδο του Binary Classification χρειάζονται τόσο αυθεντικά όσο και παραποιημένα δεδομένα για την εκπαίδευση της μεθόδου. Ως εκ τούτου τα δύο σετ δεδομένων, το VoxCeleb2 και το Faceforensics++ ενώθηκαν για να δημιουργηθεί ένα καινούργιο σετ δεδομένων που θα περιέχει συνολικά 146.569 αυθεντικά βίντεο και 4.000 παραποιημένα. Το 80% των δεδομένων του καινούργιου σετ δεδομένων χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου και το 20% ως validation και testing sets.

4.2 Μετρικές Αξιολόγησης

Προκειμένου να αξιολογηθεί η απόδοση των μεθόδων στα testing δεδομένα, χρησιμοποιούνται οι παρακάτω μετρικές αξιολόγησης, οι οποίες είναι κατάλληλες για προβλήματα δυαδικής ταξινόμησης όπως στην περίπτωσή μας (αυθεντικά ή παραποιημένα πρόσωπα). Ως θετική κλάση θεωρούμε τα πραγματικά πρόσωπα και αρνητική τα παραποιημένα πρόσωπα. Οι τιμές TP, TN, FP, FN για την δυαδική ταξινόμηση ορίζονται ως:

- True Positives (TP): η πραγματική κλάση είναι θετική και το μοντέλο την προέβλεψε θετικά
- True Negatives (TN): η πραγματική κλάση είναι αρνητική και το μοντέλο την προέβλεψε αρνητικά
- False Positives (FP): η πραγματική κλάση είναι αρνητική αλλά το μοντέλο την προέβλεψε θετικά
- False Negatives (FN): η πραγματική κλάση είναι θετική αλλά το μοντέλο την προέβλεψε αρνητικά

Ακρίβεια (Accuracy): Πρόκειται για το ποσοστό των σωστών προβλέψεων (tp) προς το συνολικό αριθμό των δειγμάτων που δόθηκαν στην είσοδο ενός δικτύου.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Αξιοπιστία (Precision): Πόση αποτελεσματικότητα έχει το μοντέλο να προβλέπει την θετική κλάση.

$$Precision = \frac{TP}{TP + FP}$$

Ανάκληση ή True Positive Rate (Recall ,TPR): Απαντάει στο ερώτημα για κάθε δεδομένο που ταξινομείται σε μια κλάση, πόσες φορές το μοντέλο έχει προβλέψει τη σωστή κλάση

$$Recall = \frac{TP}{TP + FN}$$

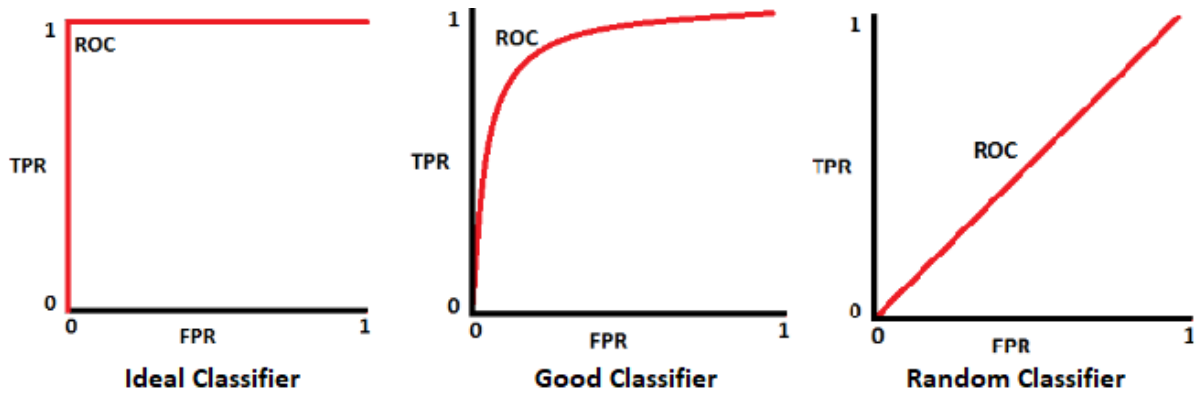
F1 Score: Πρόκειται για τον αρμονικό μέσο ανάμεσα στην Ανάκληση και την Αξιοπιστία και συγκλίνει προς αυτή με την μικρότερη τιμή

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

False Positive Rate (FPR): Πρόκειται για την αναλογία αρνητικών δεδομένων τα οποία εσφαλμένα κατηγοριοποιήθηκαν ως θετικά.

$$FPR = \frac{FP}{FP + TN}$$

Area Under the Curve of Receiver Operating Characteristics Curve (AUC-ROC curve): Πρόκειται για μία μετρική που φανερώνει πόσο καλά το μοντέλο αναγνωρίζει τις κλάσεις. ROC είναι η καμπύλη πιθανότητας και η AUC αναπαριστά την διαχωριστική ικανότητα. Όσο μεγαλύτερη η περιοχή κάτω από την καμπύλη (AUC) τόσο καλύτερη είναι η ικανότητα του μοντέλου να προβλέπει σωστά τις κλάσεις. Στην καμπύλη ROC ο άξονας y αναπαριστά το Recall ή TPR και ο άξονας x το FPR. Οι καμπύλες ROC για έναν ιδανικό classifier, για έναν καλό classifier και ένα random classifier φαίνονται παρακάτω.



Εικόνα 32: Οι καμπύλες AUC-ROC για ιδανικό, καλό και random classifier

AUC-ROC score: Η τιμή της AUC για την καμπύλη ROC. Όσο μεγαλύτερη είναι τόσο καλύτερη είναι η απόδοση του μοντέλου. Έτσι η τιμή 1 αντιστοιχεί σε ιδανικό Classifier, η τιμή 0.7 σε καλό Classifier, 0.5 για τον Random Classifier και μικρότερη από 0.5 σε κακό Classifier.

Μήτρα Σύγχυσης (Confusion Matrix): Ένας πίνακας που περιγράφει την απόδοση του μοντέλου. Οι γραμμές του πίνακα αντιπροσωπεύουν τις πραγματικές κλάσεις των δεδομένων ενώ οι στήλες αντιπροσωπεύουν τις ετικέτες που προβλέφθηκαν από το μοντέλο.

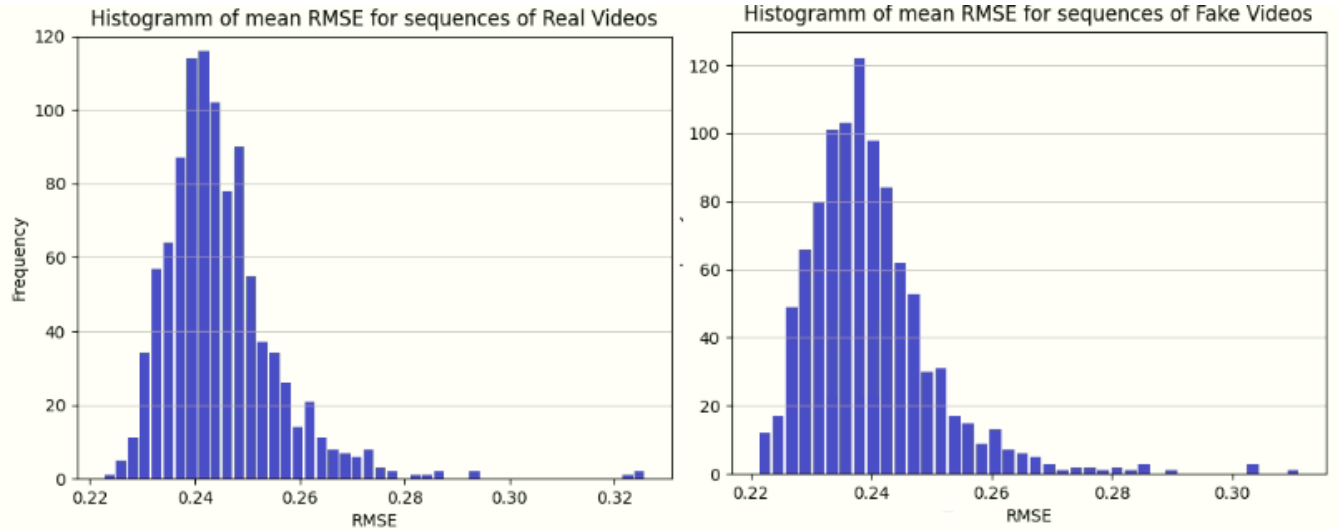
4.3 Αποτελέσματα

Όπως περιγράφεται στο κεφ. 3.2, όλες οι μέθοδοι αρχικά εκπαιδεύτηκαν για είσοδο της μορφής $v \epsilon R^{96x156}$ όπου 96 είναι ο αριθμός των διαδοχικών frames και 156 είναι ο αριθμός των παραμέτρων του FLAME (100 για το σχήμα, 50 για την έκφραση και 6 για την πόζα). Κάθε βίντεο από το testing set αποτελείται από διαφορετικό αριθμό καρέ ο οποίος συνήθως είναι μεγαλύτερος του 96 που δέχονται ως είσοδο τα μοντέλα. Για αυτό το λόγο κάθε βίντεο χωρίζεται σε διαδοχικές ακολουθίες των 96 καρέ (τυχόν καρέ στο τέλος του βίντεο που είναι λιγότερα από 96 αγνοούνται). Για να δοκιμαστούν τα μοντέλα στα testing δεδομένα εφαρμόστηκα 3 μέθοδοι για την μέτρηση της απόδοσης των μοντέλων. Η μήτρα σύγχυσης και η καμπύλη AUC-ROC παρουσιάζονται μόνο για την δεύτερη μέθοδο.

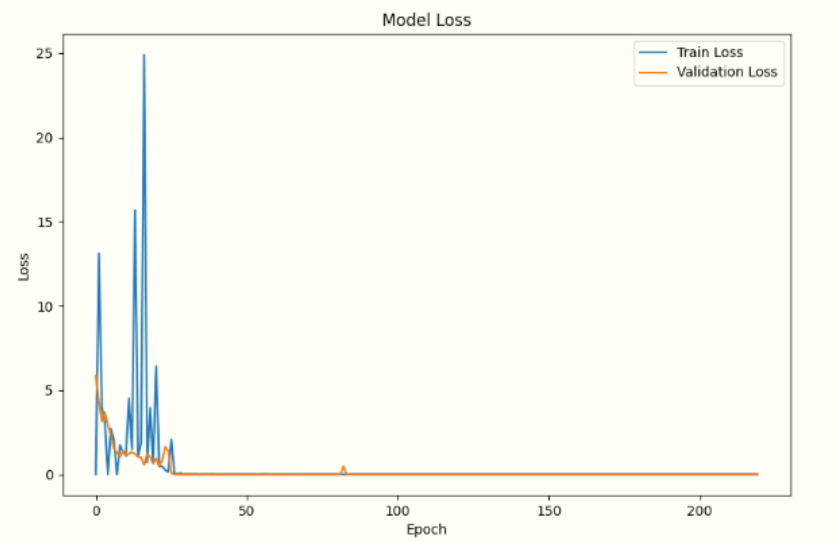
- *1η Μέθοδος:* Κατηγοριοποίηση της κάθε αλληλουχίας ως αυθεντική ή παραποιημένη με βάση την έξοδο των μοντέλων και ,στην συνέχεια, κατηγοριοποίηση του βίντεο από όπου προέρχονται οι αλληλουχίες ως αυθεντικό αν η πλειοψηφία των αλληλουχιών θεωρήθηκαν θετικές ή παραποιημένο αν η πλειοψηφία τους κατηγοριοποιήθηκε ως ψεύτικες.
- *2η Μέθοδος:* Εύρεση της μέσης τιμής της εξόδου των μοντέλων (RMSE για τους One Class Classifiers και πιθανότητας να ανήκει στην αυθεντική κλάση για τον Binary Classifier) για κάθε αλληλουχία καρέ που ανήκουν στο ίδιο βίντεο και εφαρμογή των μετρικών στην μέση τιμή για κάθε βίντεο.
- *3η Μέθοδος*: Θεωρούμε κάθε αλληλουχία καρέ που εξάγεται από τα βίντεο ως ανεξάρτητο βίντεο που έχει την ίδια ετικέτα με το βίντεο από το οποίο προέρχεται και εφαρμογή μετρικών σε όλες τις αλληλουχίες καρέ.

4.3.1 Αποτελέσματα για την 1η μέθοδο- DenseVAE

Η εκπαίδευση του μοντέλου σταμάτησε στην εποχή 220 από το early stopping με μετρική παρακολούθησης το άθροισμα του σφάλματος ανακατασκευής της εικόνας εισόδου και του KL divergence. Ύστερα από δοκιμή για 10 πραγματικά και 10 παραποιημένα βίντεο, η τιμή του κατωφλιού για το **RMSE** (Root Mean Square Error) ανάμεσα στα αρχικά και τα ανακατασκευασμένα δεδομένα πάνω από την οποία οι αλληλουχίες καρέ βίντεο θα θεωρούνται παραποιημένες και κάτω από αυτή αυθεντικές, επιλέχθηκε ίση με 0.2772. Η τιμή επιλέχθηκε τέτοια ώστε να μεγιστοποιείται η απόδοση του μοντέλου. Οι συναρτήσεις κόστους (άθροισμα του σφάλματος ανακατασκευής της εικόνας εισόδου και του KL divergence) για τα σετ εκπαίδευσης και validation , τα ιστογράμματα για τις μέσες τιμές RMSE από τις αλληλουχίες καρέ από τα αυθεντικά και τα παραποιημένα βίντεο του testing set και οι μετρικές αξιολόγησης για το testing set παρουσιάζονται παρακάτω.



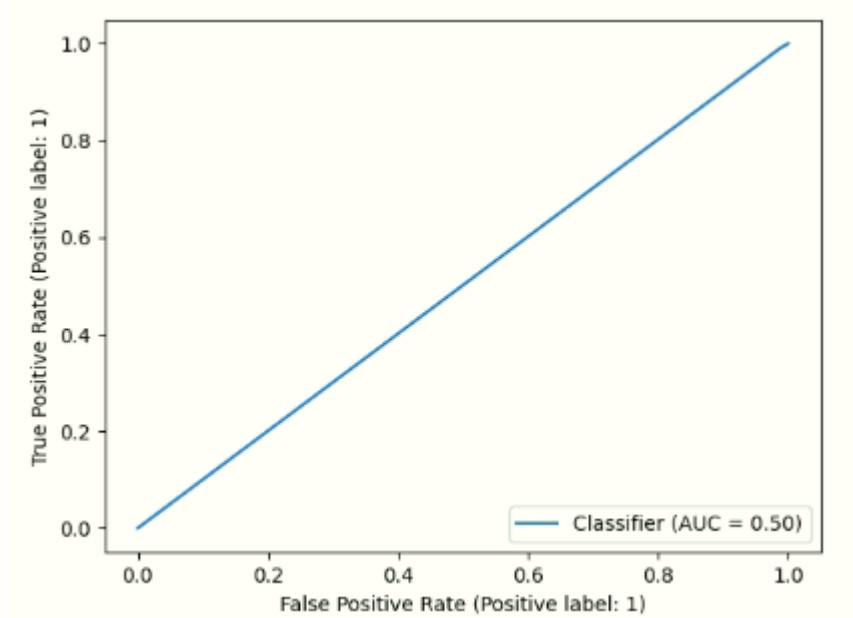
Εικόνα 33: Ιστογράμματα για τις μέσες τιμές RMSE για τις αλληλουχίες καρέ για κάθε αυθεντικό και παραποιημένο βίντεο αντίστοιχα για το DenseVAE. Παρατηρούμε ότι οι δύο κατανομές είναι σχεδόν ίδιες χωρίς να μπορεί να διαπιστωθεί ποια ανήκει στα αυθεντικά και ποια στα παραποιημένα βίντεο.



Εικόνα 34: Losses για το DenseVAE

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	49.9	0.4995	0.9830	0.6624	0.499
2η	**50.15**	**0.5008**	**0.9910**	**0.6653**	**0.5015**
3η	49.81	0.4990	0.9763	0.6605	0.4981

Πίνακας 2: Μετρικές Αξιολόγησης για το DenseVAE. Παρατηρούμε ότι η 2η μέθοδος αξιολόγησης παρουσιάζει τα καλύτερα αποτελέσματα.



Εικόνα 35: AUC-ROC Curve για το DenseVAE (2η Μέθοδος Αξιολόγησης)

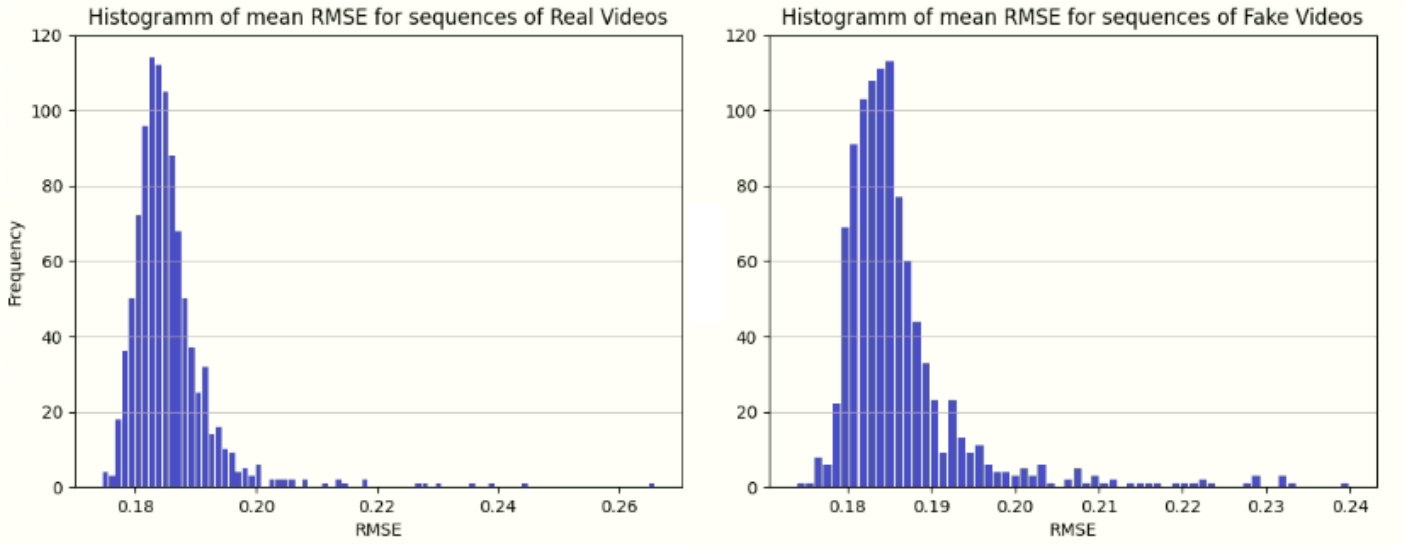
Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**12**	988
1	9	**991**

Πίνακας 3: Μήτρα Σύγχυσης για το DenseVAE (2η Μέθοδος Αξιολόγησης). Παρατηρούμε ότι το μοντέλο κατατάσσει σχεδόν όλα τα δεδομένα ως αυθεντικά ανεξάρτητα την κλάση στην οποία ανήκουν.

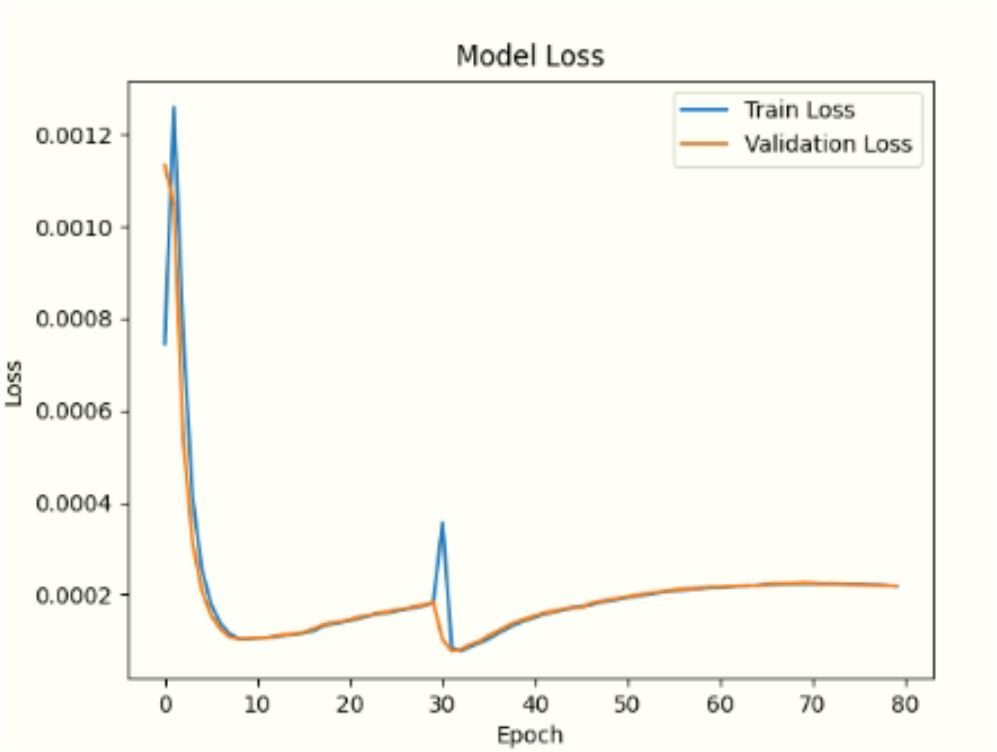
Παρατηρώντας τα RMSE από τα Ιστογράμματα δεν φαίνεται κάποια διαφορά ανάμεσα στην κατανομή των αυθεντικών και των παραποιημένων βίντεο. Αυτό επιβεβαιώνεται και από τις μετρικές αξιολόγησης όπου φαίνεται ότι το DenseVAE λειτουργεί ως Random Classifier (Classifier χωρίς πραγματική ικανότητα κατηγοριοποίησης στις σωστές κλάσεις). Για αυτό το λόγο στην συνέχεια επιλέγεται να χρησιμοποιηθεί ένα μοντέλο ανακατασκευής εικόνων με καλύτερη ικανότητα αναπαράστασης, το VQ-VAE.

4.3.2 Αποτελέσματα για τη 2η μέθοδο – VQ VAE

Η εκπαίδευση του μοντέλου σταμάτησε στην εποχή 80 από το Early Stopping με μετρική παρακολούθησης τη συνάρτηση κόστους που περιγράφεται στο κεφ. 2.5 για το validation set. Παρατηρούμε ότι το VQ-VAE εκπαιδεύεται αρκετά πιο γρήγορα από το DenseVAE. Ύστερα από δοκιμή για 10 πραγματικά και 10 παραποιημένα βίντεο, η τιμή του κατωφλιού για το RMSE ανάμεσα στα αρχικά και τα ανακατασκευασμένα δεδομένα πάνω από την οποία οι αλληλουχίες καρέ βίντεο θα θεωρούνται παραποιημένες και κάτω από αυτή αυθεντικές, επιλέχθηκε ίση με 0.1789. Οι συναρτήσεις κόστους (όπως περιγράφονται στο κεφ. 2.5) για τα δεδομένα εκπαίδευσης και validation, τα ιστογράμματα για τις μέσες τιμές RMSE από τις αλληλουχίες καρέ από τα αυθεντικά και τα παραποιημένα βίντεο του testing set και οι μετρικές αξιολόγησης για το testing set παρουσιάζονται παρακάτω.



Εικόνα 36: Ιστογράμματα για τις μέσες τιμές RMSE για τις αλληλουχίες καρέ για κάθε αυθεντικό και παραποιημένο βίντεο αντίστοιχα για το VQ-VAE. Παρατηρούμε ότι οι δύο κατανομές είναι σχεδόν ίδιες χωρίς να μπορεί να διαπιστωθεί ποια ανήκει στα αυθεντικά και ποια στα παραποιημένα βίντεο. Ωστόσο τα RMSE είναι μικρότερα από αυτά του DenseVAE



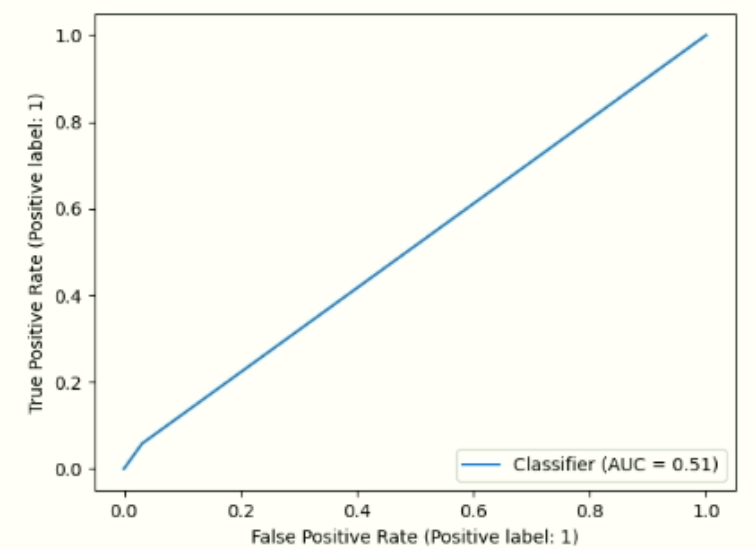
Εικόνα 37: Losses για το VQ-VAE

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	51.25	0.6147	**0.670**	0.1208	0.5125
2η	**51.35**	**0.6517**	0.580	0.1065	**0.5135**
3η	51.15	0.5434	0.144	**0.2282**	0.5115

Πίνακας 4: Μετρικές αξιολόγησης για το VQ-VAE. Παρατηρούμε ότι οι μετρικές της ακρίβειας και του AUC-ROC score είναι καλύτερες για την 2η μέθοδο αξιολόγησης. Το f1-score είναι καλύτερο για την 3η μέθοδο.

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**969**	31
1	942	**58**

Πίνακας 5: Μήτρα Σύγχυσης για το DenseVAE (2η Μέθοδος Αξιολόγησης). Παρατηρούμε ότι το μοντέλο κατατάσσει σχεδόν όλα τα δεδομένα ως παραποιημένα ανεξάρτητα την κλάση στην οποία ανήκουν.

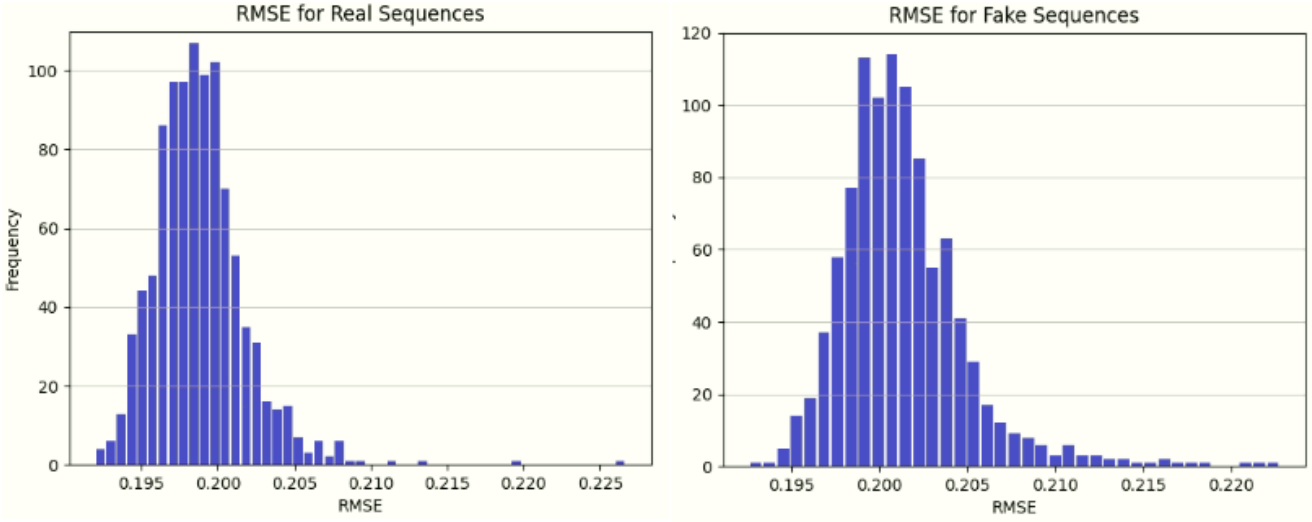


Εικόνα 38: AUC-ROC Curve για το VQ-VAE (2η Μέθοδος Αξιολόγησης)

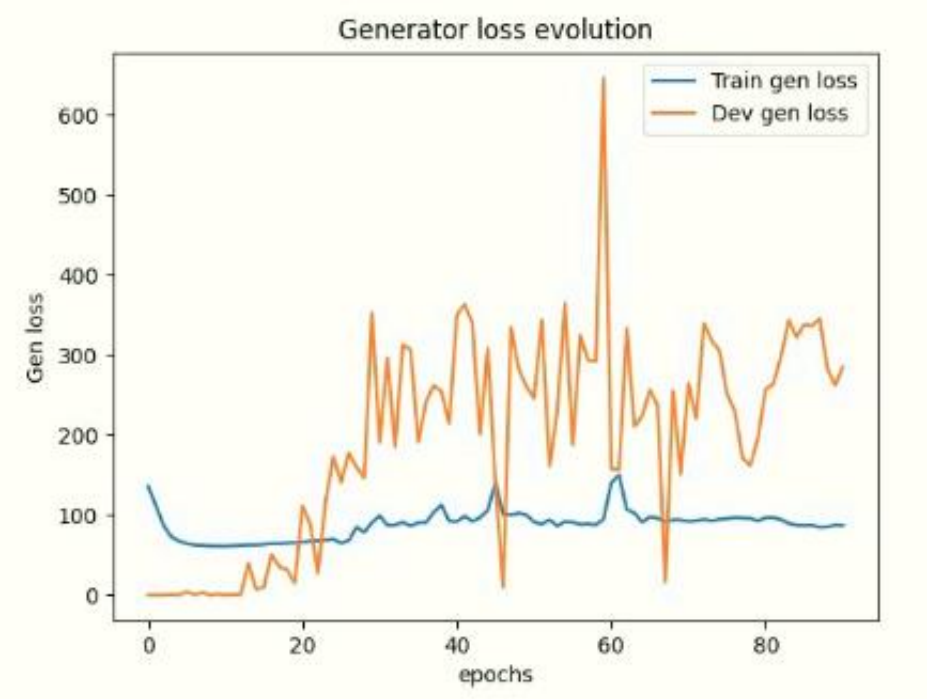
Παρατηρούμε ότι οι μέσες τιμές του RMSE τόσο για τα αυθεντικά όσο και για τα παραποιημένα δεδομένα είναι μικρότερες από τις αντίστοιχες τους DenseVAE που σημαίνει ότι, πράγματι, το VQ-VAE παρουσιάζει μικρότερο σφάλμα ανακατασκευής. Επίσης έχει πιο ομαλή εκπαίδευση από το DenseVAE και από τις μετρικές παρατηρούμε ότι έχει καλύτερη απόδοση και στις 3 μεθόδους αξιολόγησης. Ωστόσο, αν και καλύτερο από το DenseVAE, είναι ελαφρώς καλύτερο από έναν Random Classifier και, επομένως, μη αποτελεσματικός ως Classifier. Για αυτό το λόγο επιλέγεται να δημιουργηθεί ένα μοντέλο GAN που θα δημιουργεί από μόνο του παραποιημένα δεδομένα και τα οποία εικάζουμε ότι θα βοηθήσουν στην δημιουργία ενός πιο αποτελεσματικού μοντέλου.

4.3.3 Αποτελέσματα για την 3η μέθοδο – GAN

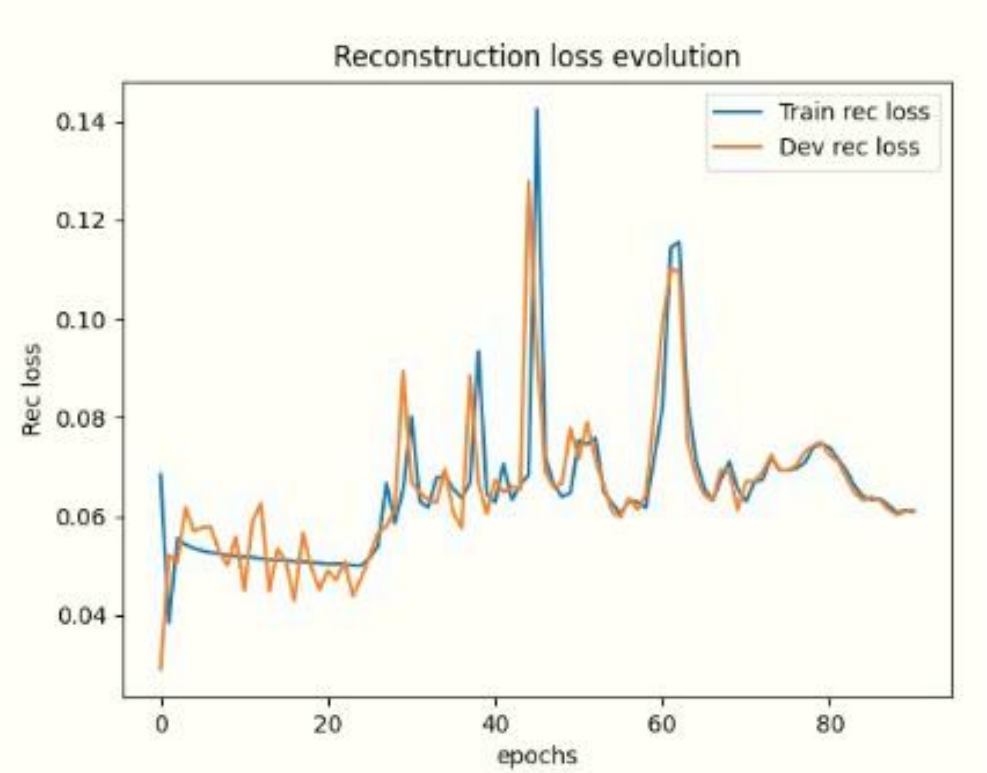
Αρχικά το μοντέλο εκπαιδεύτηκε με σταθερό learning rate αλλά παρατηρήθηκε ότι γινόταν ασταθές μετά την εποχή 20. Για αυτό το λόγο επιλέχθηκε να μειώνεται το learning rate κάθε 25 εποχές με λόγο 0.2 . Η εκπαίδευση τερματίστηκε μετά από 90 εποχές διότι δεν παρατηρήθηκε περαιτέρω βελτίωση για το σφάλμα ανακατασκευής για το validation set μετά την 85. Ύστερα από δοκιμή για 10 πραγματικά και 10 παραποιημένα βίντεο, η τιμή του κατωφλιού για το RMSE ανάμεσα στα αρχικά και τα ανακατασκευασμένα δεδομένα πάνω από την οποία οι αλληλουχίες καρέ βίντεο θα θεωρούνται παραποιημένες και κάτω από αυτή αυθεντικές, επιλέχθηκε ίση με 0.197. Τα Generator, Discriminator και reconstruction losses (όπως αυτά περιγράφονται στο κεφ. 3.5.2) για τα σετ εκπαίδευσης και validation, τα ιστογράμματα για τις μέσες τιμές RMSE από τις αλληλουχίες καρέ από τα αυθεντικά και τα παραποιημένα βίντεο του testing set και οι μετρικές αξιολόγησης για το testing set παρουσιάζονται παρακάτω.



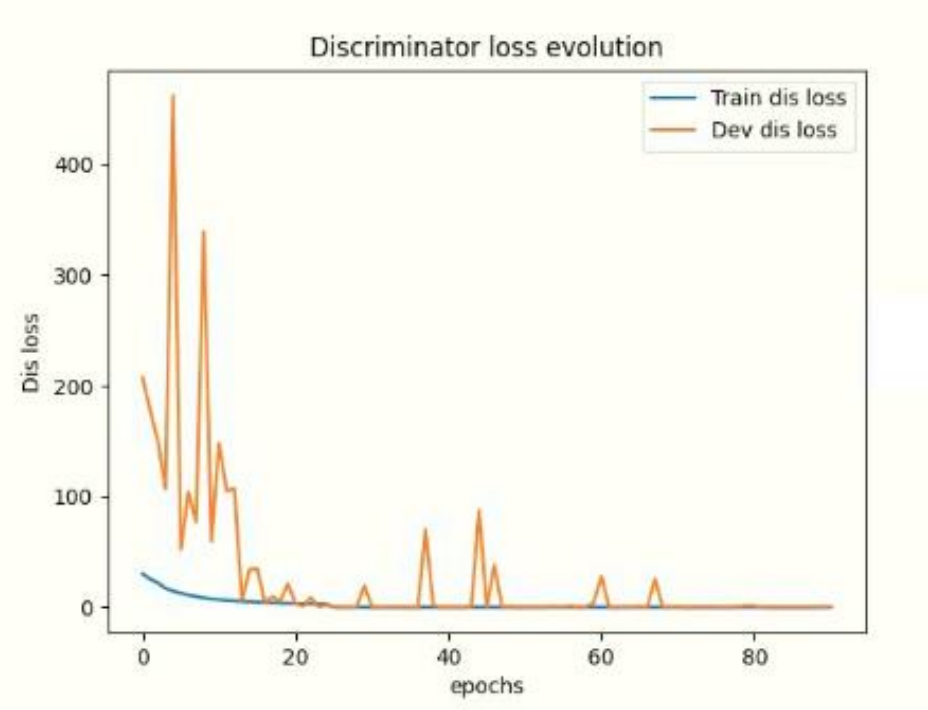
Εικόνα 39: Ιστογράμματα για τις μέσες τιμές RMSE για τις αλληλουχίες καρέ για κάθε αυθεντικό και παραποιημένο βίντεο αντίστοιχα για το GAN. Παρατηρούμε μικρές διαφορές στις κατανομές για τα αυθεντικά και παραποιημένα βίντεο.



Εικόνα 40: Το Loss του Generator του GAN κατά την εκπαίδευση και το validation



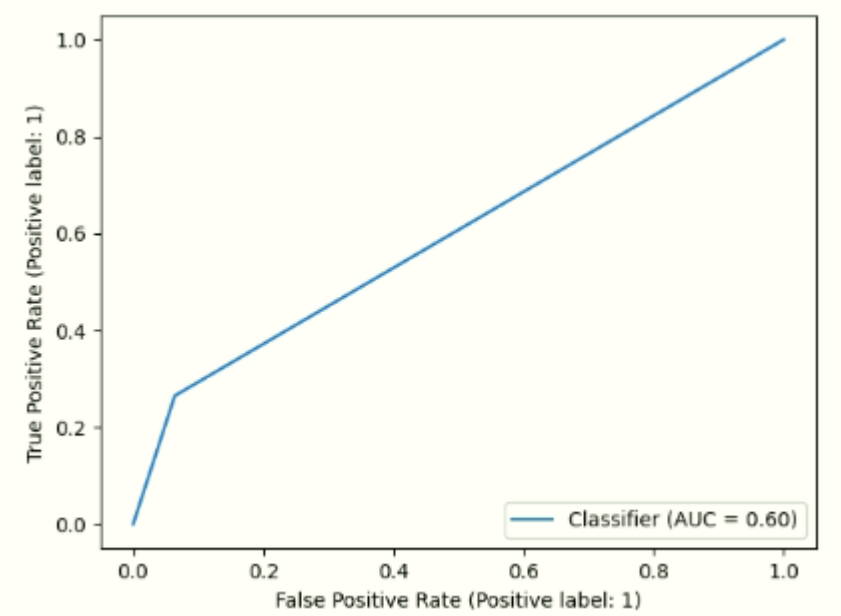
Εικόνα 41: Το σφάλμα ανακατασκευής του Generator του GAN κατά την εκπαίδευση και το validation



Εικόνα 42: Το Loss του Discriminator του GAN κατά την εκπαίδευση και το validation

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	59.55	0.7784	0.2670	0.3976	0.5954
2η	**60.05**	**0.8055**	0.2650	0.3988	**0.6005**
3η	59.25	0.6759	**0.3553**	**0.4657**	0.5925

Πίνακας 6: Μετρικές αξιολόγησης για το GAN. Παρατηρούμε ότι οι μετρικές της ακρίβειας και του AUC-ROC score είναι καλύτερες για την 2η μέθοδο αξιολόγησης. Το f1-score είναι καλύτερο για την 3η μέθοδο.



Εικόνα 43: AUC-ROC Curve για το GAN (2η Μέθοδος Αξιολόγησης)

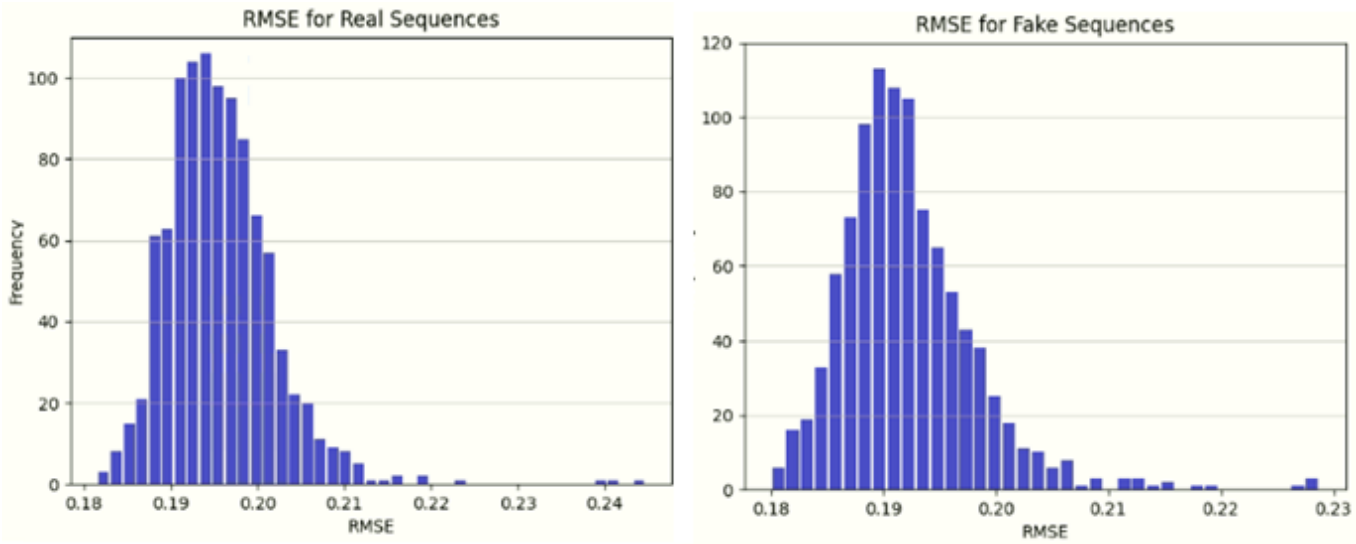
Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**936**	64
1	735	**265**

Πίνακας 7: Μήτρα Σύγχυσης για το GAN (2η Μέθοδος Αξιολόγησης). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί σχεδόν όλα τα παραποιημένα δεδομένα σωστά και κατηγοριοποιεί το 25% περίπου των αυθεντικών δεδομένων ως αυθεντικά.

Παρατηρούμε ότι το μοντέλο δεν έχει τόσο καλή ικανότητα ανακατασκευής όσο το VQ-VAE. Επίσης παρότι η αστάθεια έχει βελτιωθεί συγκριτικά με το 1^{ο} μοντέλο, συνεχίζει να παρατηρείται. Παρόλα αυτά το μοντέλο παρουσιάζει καλύτερη απόδοση στην κατηγοριοποίηση των δεδομένων, με τα καλύτερα αποτελέσματα να παρουσιάζονται για την 2^{η} μέθοδο αξιολόγησης, που έχει 60.05% ακρίβεια και 0.6005 Roc Auc Score. Έτσι το GAN παρουσιάζει καλύτερα αποτελέσματα από τα μοντέλα των 2 πρώτων μεθόδων αφού κατηγοριοποιεί περισσότερα βίντεο στην σωστή κλάση όπως φαίνεται και στην μήτρα σύγχυσης. Δοκιμάζεται, επίσης, ένα πιο σύνθετο GAN, το OCGAN, που πέρα από έναν generator και έναν discriminator που περιέχει το απλό GAN, περιέχει ένα ακόμη discriminator και έναν classifier.

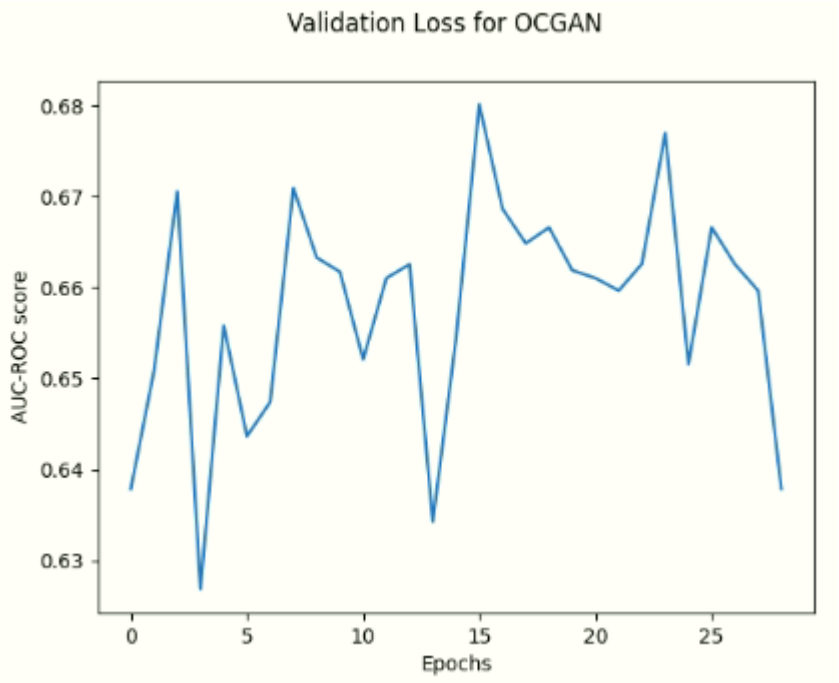
4.3.4 Αποτελέσματα για την 4^{η} μέθοδο – OCGAN

Στις πρώτες 20 εποχές εκπαιδεύεται μόνο ο Generator και, στην συνέχεια, εκπαιδεύονται όλα τα δίκτυα για άλλες 30 εποχές οπότε και τερματίζεται η εκπαίδευση καθώς το roc auc score για το validation set δεν βελτιώνεται για αρκετές εποχές. Όπως και στις προηγούμενες μεθόδους η τιμή του κατωφλιού για το RMSE ανάμεσα στα αρχικά και τα ανακατασκευασμένα δεδομένα πάνω από την οποία οι αλληλουχίες καρέ βίντεο θα θεωρούνται παραποιημένες και κάτω από αυτή αυθεντικές, επιλέχθηκε ίση με 0.192. Τα κόστη (όπως περιγράφονται στο κεφ. 3.6.2) για τα δεδομένα εκπαίδευσης και validation , τα ιστογράμματα για τις μέσες τιμές RMSE από τις αλληλουχίες καρέ από τα αυθεντικά και τα παραποιημένα βίντεο του testing set και οι μετρικές αξιολόγησης για το testing set παρουσιάζονται παρακάτω. Τα κόστη παρουσιάζονται για τις τελευταίες 30 εποχές.



Εικόνα 44: Ιστογράμματα για τις μέσες τιμές RMSE για τις αλληλουχίες καρέ για κάθε αυθεντικό και παραποιημένο βίντεο αντίστοιχα για το OCGAN. Παρατηρούμε διαφορές στις κατανομές για τα αυθεντικά και παραποιημένα βίντεο, μεγαλύτερες από αυτές στο απλό GAN.

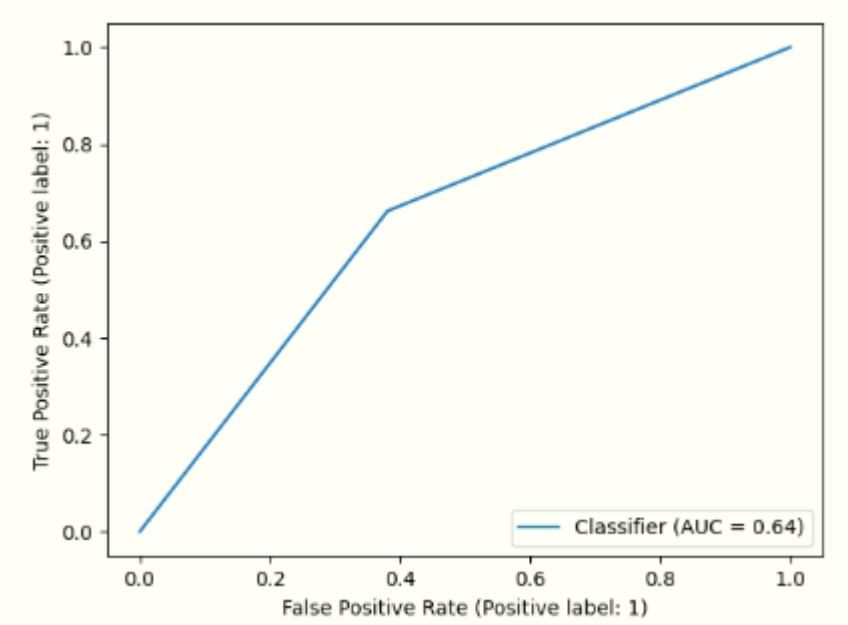
Εικόνα 45: Train Loss για το OCGAN



Εικόνα 46: AUC-ROC score για το validation set για το OCGAN

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	63.05	0.6339	0.6070	0.6126	0.6305
2η	**64.05**	**0.6347**	**0.6628**	**0.6481**	**0.6405**
3η	60.7	0.6060	0.6117	0.6088	0.6070

Πίνακας 8: Μετρικές αξιολόγησης για το OCGAN. Παρατηρούμε ότι η 2η μέθοδος αξιολόγησης παρουσιάζει τα καλύτερα αποτελέσματα.



Εικόνα 47: AUC-ROC Curve για το OCGAN (2η Μέθοδος Αξιολόγησης)

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**619**	381
1	338	**662**

Πίνακας 9: Confusion Matrix για το OCGAN (2η μέθοδος αξιολόγησης). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί το 66% περίπου των δεδομένων στην σωστή κλάση.

Παρατηρούμε ότι το OCGAN παρουσιάζει την καλύτερη απόδοση από όλα τα προηγούμενα μοντέλα, έχοντας ακρίβεια 64.05% και ROC AUC score 0.6405 για τις μέση τιμή RMSE για όλες τις διαδοχικές αλληλουχίες ενός βίντεο. Ωστόσο το train loss (όπως αυτό περιγράφεται στο κεφ. 3.6.2) του μοντέλου αυξάνεται και διατηρείται σταθερό στην συνέχεια αντί να μειώνεται κάτι που ,μαζί με τα κόστη κατά την εκπαίδευση στο απλό GAN (κεφ. 4.3.3), φανερώνουν τα προβλήματα αστάθειας που παρατηρούνται γενικά στα GAN. Τέλος επιλέχθηκε να εκπαιδευτεί ένα μοντέλο που ακολουθεί την πιο παραδοσιακή προσέγγιση του binary classification.

4.3.5 Αποτελέσματα για τη 5η Μέθοδο – Binary Classification με το Efficient Net v2

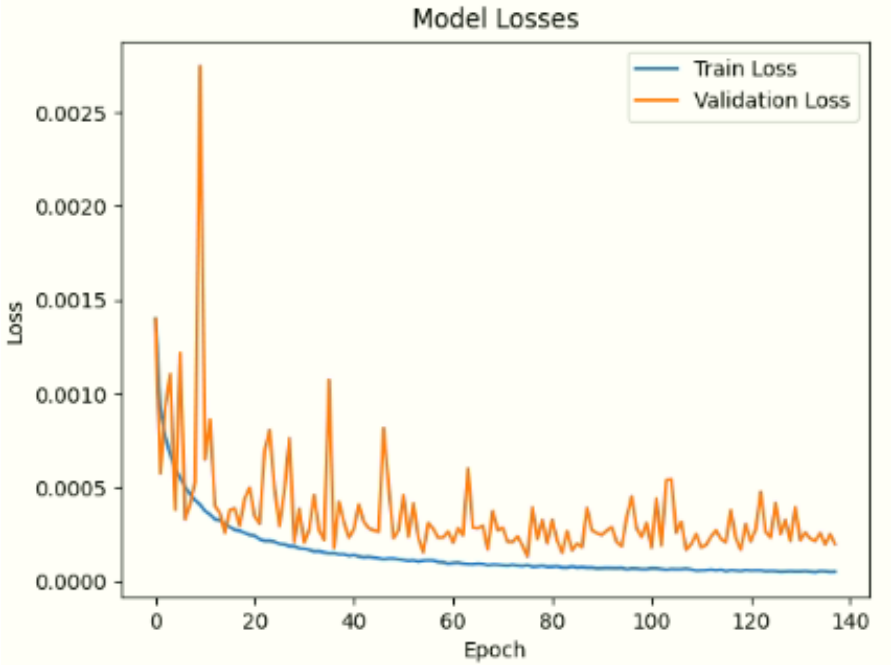
Η τελευταία μέθοδος που δοκιμάστηκε ήταν αυτή της δυαδικής κατηγοριοποίησης με το Βαθύ Συνελικτικό Δίκτυο Efficient Net v2. Η μέθοδος αυτή παρουσιάζει αρκετά καλύτερα αποτελέσματα στο testing set (αποτελείται από 20% των βίντεο του VoxCeleb2 και 20% του FaceForensics++ με τέτοιο τρόπο που να περιλαμβάνονται παραποιημένα βίντεο και από τις τέσσερις μεθόδους παραποίησης) και, για αυτό το λόγο, πέρα από την εκπαίδευση έχοντας ως είσοδο την αλληλουχία χαρακτηριστικών που εξάγονται από διαδοχικά καρέ και χρησιμοποιήθηκε σε όλους τις προηγούμενες μεθόδους, εκπαιδεύτηκε και για διαφορετικές εισόδους. Οι είσοδοι αυτοί διαφέρουν τόσο ως προς των αριθμό των διαδοχικών καρέ όσο και στο ποια χαρακτηριστικά χρησιμοποιούνται. Τα πειράματα αυτά πραγματοποιήθηκαν για να διαπιστωθεί η επίδραση του αριθμού των καρέ και των χαρακτηριστικών στην απόδοση του μοντέλου. Έτσι ο αριθμός των διαδοχικών καρέ επιλέχθηκε να είναι 50 ή 96 και για τα χαρακτηριστικά χρησιμοποιήθηκαν αρχικά μόνο οι παράμετροι του σχήματος, της έκφρασης και της πόζας που χρησιμοποιεί το FLAME, προστέθηκαν στην συνέχεια οι παράμετροι του χρώματος του προσώπου (texture) και τέλος προστέθηκαν επίσης και τα χαρακτηριστικά που εξάχθηκαν από τα ορόσημα του προσώπου τα οποία περιγράφονται στο κεφ. 3.1 . Τα διάφορα πειράματα που πραγματοποιήθηκαν περιγράφονται συνοπτικά στον παρακάτω πίνακα.

Α/Α Πειράματος	***Αριθμός καρέ***	***Χαρακτηριστικά***
1	50	Shape, Expression & Pose parameters
2	96	Shape, Expression & Pose parameters
3	50	Shape, Expression, Pose & Texture Parameters
4	96	Shape, Expression, Pose & Texture Parameters
5	50	Shape, Expression, Pose, Texture Parameters & Landmark Based Features
6	96	Shape, Expression, Pose, Texture Parameters & Landmark Based Features

Πίνακας 10: Περιγραφή των πειραμάτων που πραγματοποιήθηκαν για το Binary Classification

4.3.5.1 Αποτελέσματα για το 1ο πείραμα

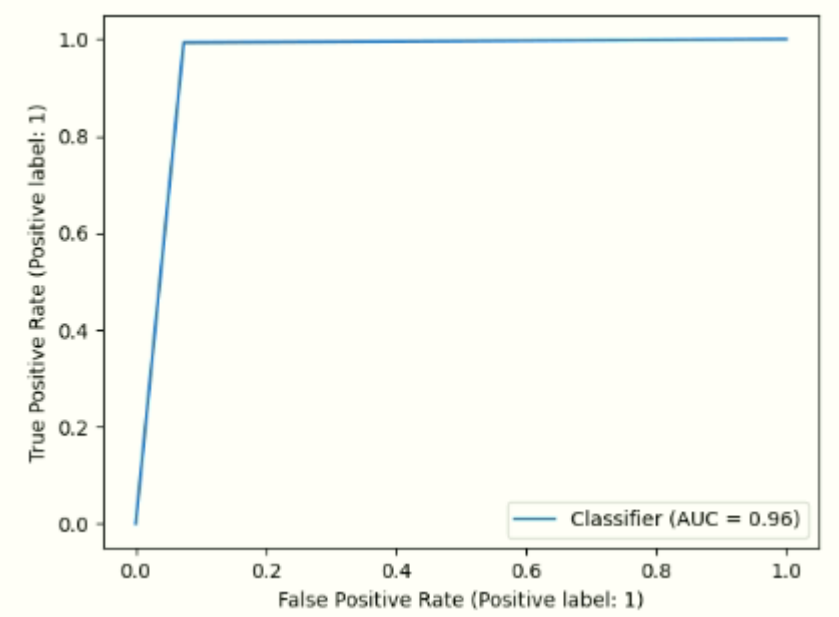
Για το πρώτο πείραμα χρησιμοποιούνται τα χαρακτηριστικά με τα οποία εκπαιδεύτηκαν και όλες οι προηγούμενοι μέθοδοι, δηλαδή 100 παράμετροι για το σχήμα, 50 για την έκφραση και 6 για την πόζα, αλλά ο αριθμός της αλληλουχίας διαδοχικών καρέ είναι 50 αντί για 96. Έτσι είναι ένα διάνυσμα που ανήκει στον δισδιάστατο διανυσματικό χώρο $v \epsilon R^{50x156}$. Το μοντέλο έχει πολύ καλή απόδοση με ακρίβεια 99.12% (για την 2η μέθοδο αξιολόγησης). Το μοντέλο εκπαιδεύεται για 138 εποχές οπότε και τερματίζεται η εκπαίδευση . Παρατηρούμε ότι τα καλύτερα αποτελέσματα παρουσιάζονται για την 2η μέθοδο αξιολόγησης.



Εικόνα 48: Τα losses κατά την εκπαίδευση και το validation για το Binary Classification (1ο πείραμα)

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	98.68	0.9980	0.9884	0.9932	0.9580
2η	**99.12**	0.9980	**0.9929**	**0.9955**	**0.9596**
3η	97.69	**0.9900**	0.9852	0.9876	0.9218

Πίνακας 11: Μετρικές Αξιολόγησης για το Binary Classification (1ο πείραμα). Παρατηρούμε ότι η 2η μέθοδος αξιολόγησης παρουσιάζει τα καλύτερα αποτελέσματα.



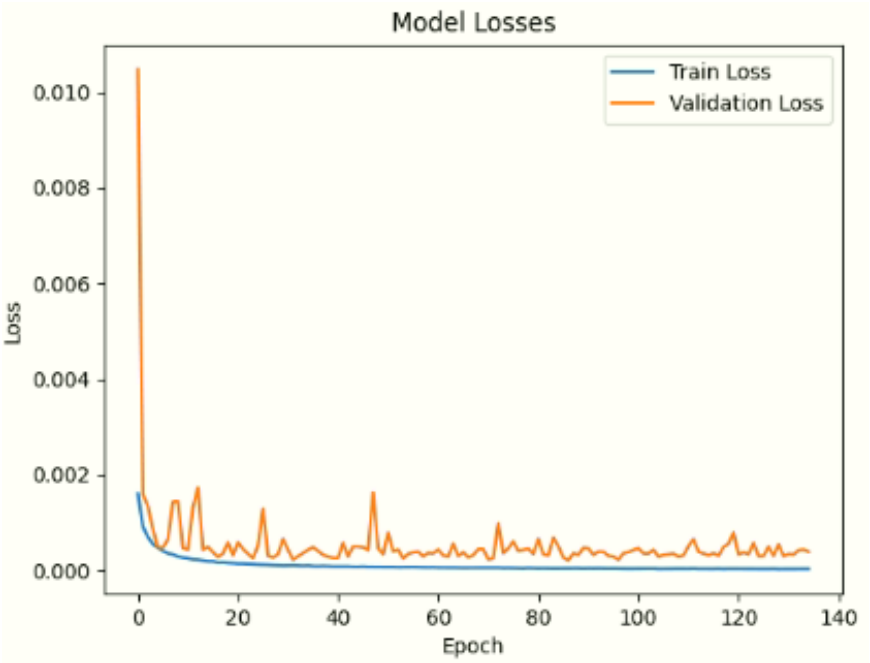
Εικόνα 49: AUC-ROC Curve για το Binary Classification (2η Μέθοδος Αξιολόγησης, 1ο Πείραμα)

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**741**	59
1	211	**29620**

Πίνακας 12: Μήτρα Σύγχυσης για το Binary Classification (1ο Πείραμα). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί σωστά την πλειοψηφία των δεδομένων.

4.3.5.2 Αποτελέσματα για το 2ο πείραμα

Για το 2ο πείραμα η είσοδος είναι της μορφής $v \epsilon R^{96x156}$ όπου 96 είναι ο αριθμός των διαδοχικών frames και 156 είναι ο αριθμός των παραμέτρων του FLAME (100 για το σχήμα, 50 για την έκφραση και 6 για την πόζα), δηλαδή η είσοδος για την οποία εκπαιδεύτηκαν οι μέθοδοι 1 έως 4. Το μοντέλο εκπαιδεύεται για 134 εποχές οπότε και τερματίζεται η εκπαίδευση από το early stopping. Παρατηρούμε ότι το κόστος του cross-entropy για το validation set είναι πιο μικρό από αυτό του 1ου πειράματος και παρουσιάζει μικρότερες διακυμάνσεις.



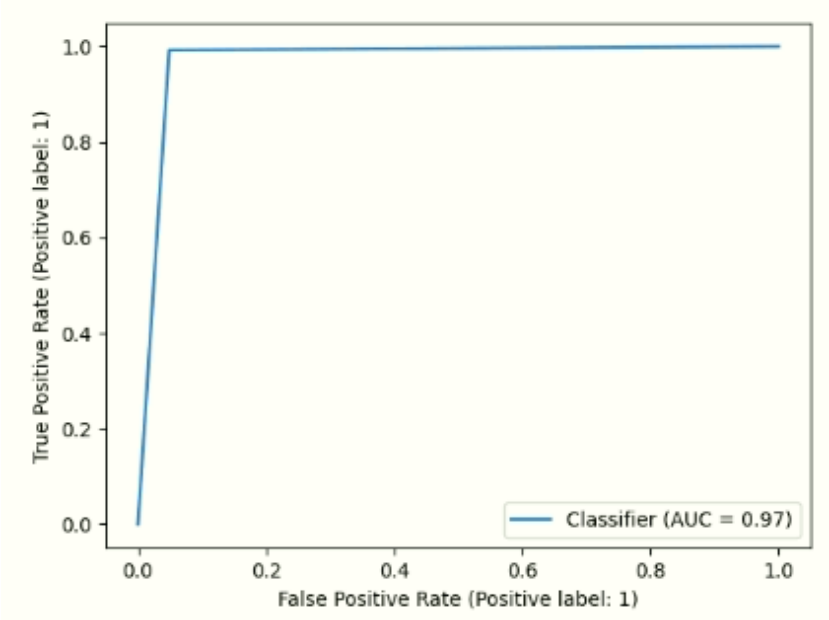
Εικόνα 50: Τα losses κατά την εκπαίδευση και το validation για το Binary Classification (2ο πείραμα)

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	99.02	0.9989	0.9918	0.9949	0.9755
2η	**99.13**	**0.9987**	**0.9924**	**0.9955**	**0.9718**
3η	97.91	0.9933	0.9841	0.9887	0.9484

Πίνακας 13: Μετρικές Αξιολόγησης για το Binary Classification (2ο πείραμα). Παρατηρούμε ότι η 2η μέθοδος αξιολόγησης παρουσιάζει τα καλύτερα αποτελέσματα.

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**761**	39
1	227	**29604**

Πίνακας 14 : Μήτρα Σύγχυσης για το Binary Classification (2ο Πείραμα). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί σωστά την πλειοψηφία των δεδομένων.



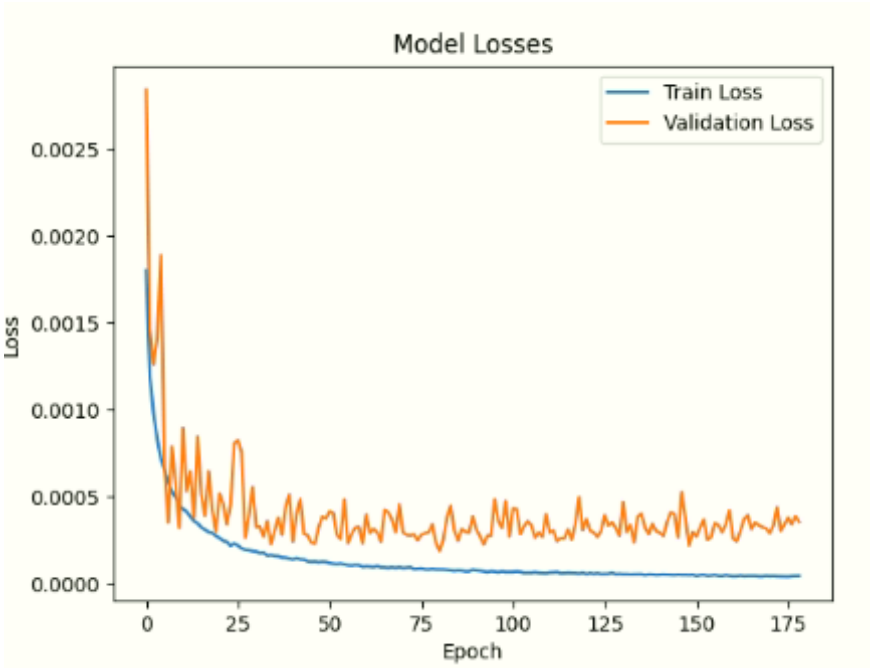
Εικόνα 51: AUC-ROC Curve για το Binary Classification (2η Μέθοδος Αξιολόγησης, 2ο Πείραμα)

4.3.5.3 Αποτελέσματα για το 3ο πείραμα

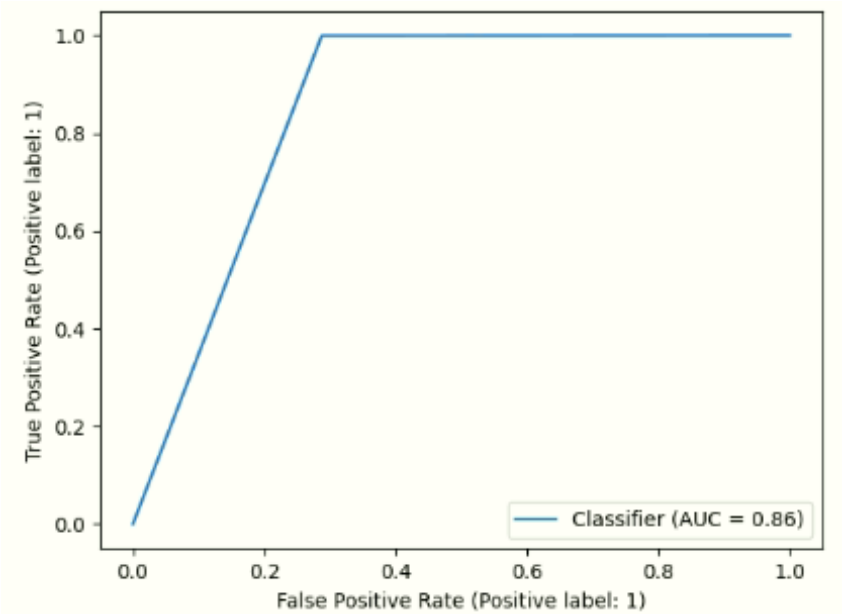
Αυτή τη φορά πέρα από τις παραμέτρους για το σχήμα, την έκφραση και την πόζα του προσώπου, χρησιμοποιούνται 50 παράμετροι και για το χρώμα του προσώπου (texture). Ο αριθμός καρέ του διανύσματος εισόδου είναι 50, συνεπώς η είσοδος του μοντέλου είναι της μορφής $v \epsilon R^{50x206}$. Το μοντέλο εκπαιδεύεται για 175 εποχές οπότε και τερματίζεται η εκπαίδευση. Η εκπαίδευση βλέπουμε ότι διαρκεί περισσότερο από τα 2 πρώτα πειράματα που έχουν μικρότερες διαστάσεις εισόδου.

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	99.17	0.9913	0.9989	0.9959	**0.8738**
2η	**99.19**	**0.9923**	**0.9994**	**0.9961**	0.8559
3η	97.69	0.9783	0.9974	0.9878	0.8403

Πίνακας 15: Μετρικές Αξιολόγησης για το Binary Classification (3ο πείραμα). Παρατηρούμε ότι σχεδόν όλες οι μετρικές παρουσιάζουν καλύτερα αποτελέσματα για τη 2η μέθοδο αξιολόγησης. Το AUC-ROC score είναι καλύτερο για την 1η.



Εικόνα 52: Τα losses κατά την εκπαίδευση και το validation για το Binary Classification (3ο πείραμα)



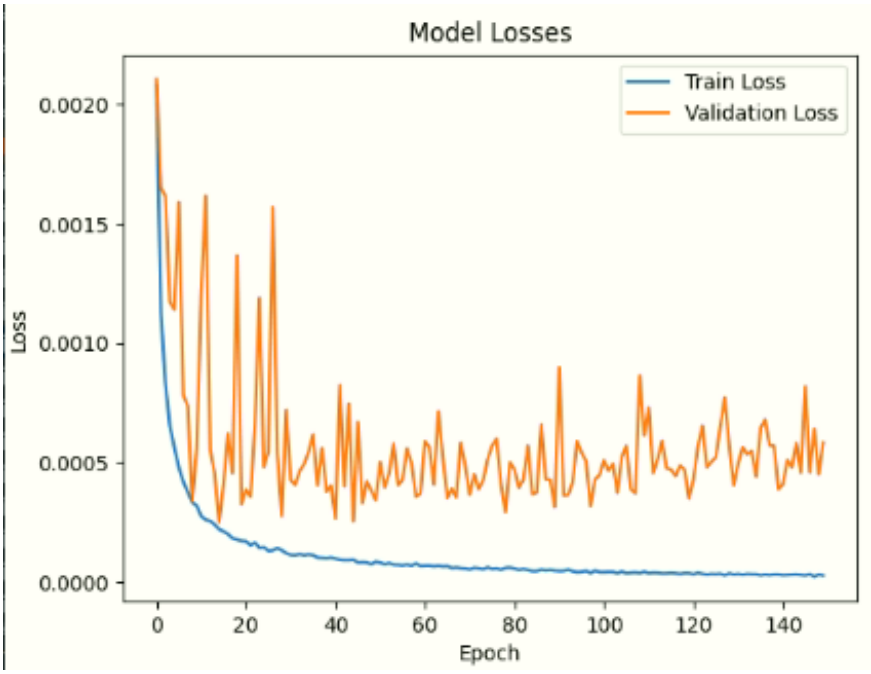
Εικόνα 53: AUC-ROC Curve για το Binary Classification (2η Μέθοδος Αξιολόγησης, 3ο Πείραμα)

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**570**	230
1	18	**29813**

Πίνακας 16: Μήτρα Σύγχυσης για το Binary Classification (3ο Πείραμα). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί σωστά την πλειοψηφία των δεδομένων. Ωστόσο κατηγοριοποιεί λιγότερα παραποιημένα δεδομένα σωστά συγκριτικά με τα 2 προηγούμενα πειράματα.

4.3.5.4 Αποτελέσματα για το 4ο πείραμα

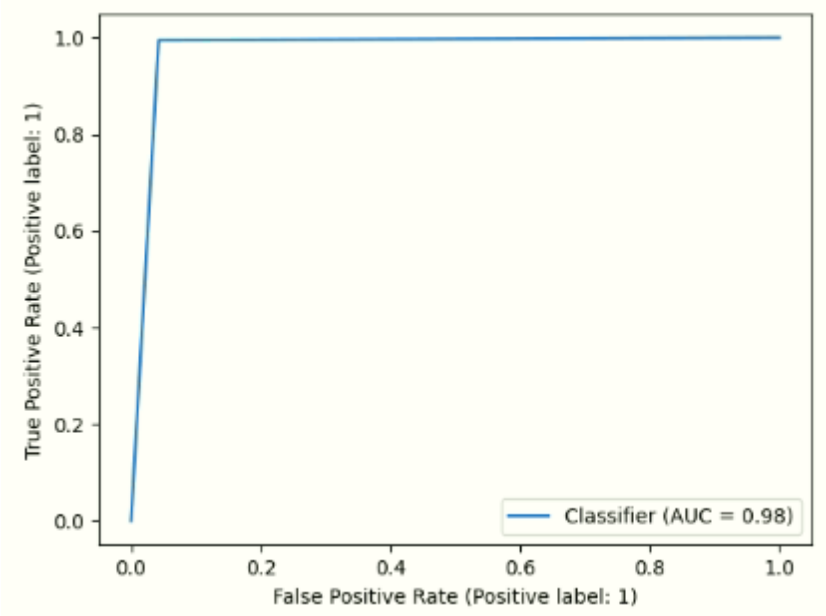
Όπως και στο 3ο πείραμα, και εδώ χρησιμοποιούνται οι παράμετροι για την πόζα, την έκφραση, το σχήμα και το texture του προσώπου με την διαφορά ότι τα καρέ του διανύσματος εισόδου είναι 96, δηλαδή της μορφής $v \epsilon R^{96x206}$. Το μοντέλο εκπαιδεύεται για 150 εποχές.



Εικόνα 54: Τα losses κατά την εκπαίδευση και το validation για το Binary Classification (4ο πείραμα)

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	99.24	0.9989	0.9934	0.9961	0.9754
2η	**99.37**	**0.9989**	**0.9947**	**0.9968**	**0.9761**
3η	98.36	0.9944	0.9879	0.9912	0.9576

Πίνακας 17: Μετρικές Αξιολόγησης για το Binary Classification (4ο πείραμα). Παρατηρούμε ότι η 2η μέθοδος αξιολόγησης παρουσιάζει τα καλύτερα αποτελέσματα.



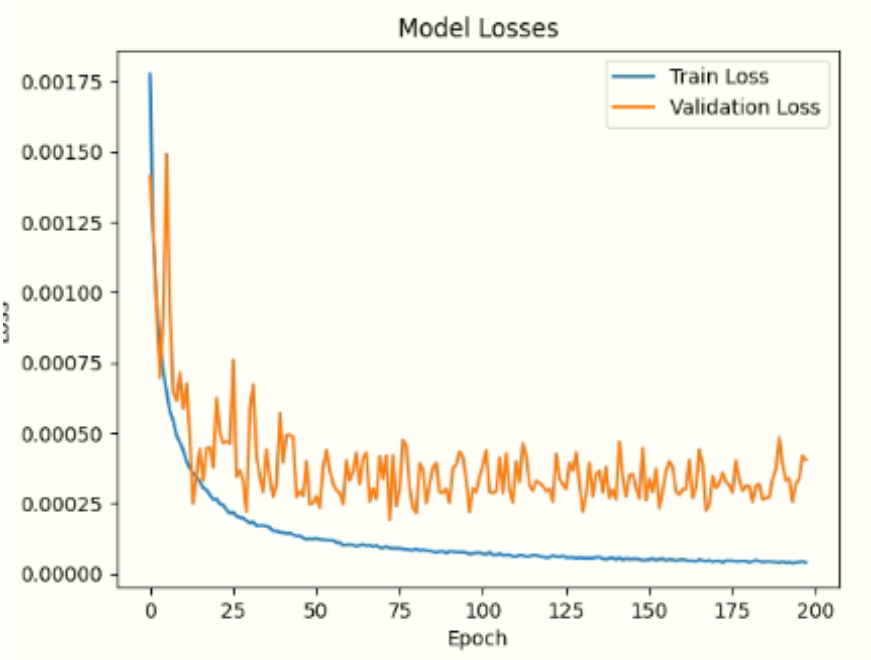
Εικόνα 55: AUC-ROC Curve για το Binary Classification (2η Μέθοδος Αξιολόγησης, 4ο Πείραμα)

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**766**	34
1	158	**29673**

Πίνακας 18: Μήτρα Σύγχυσης για το Binary Classification (4ο Πείραμα). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί σωστά την πλειοψηφία των δεδομένων.

4.3.5.5 Αποτελέσματα για το 5ο πείραμα

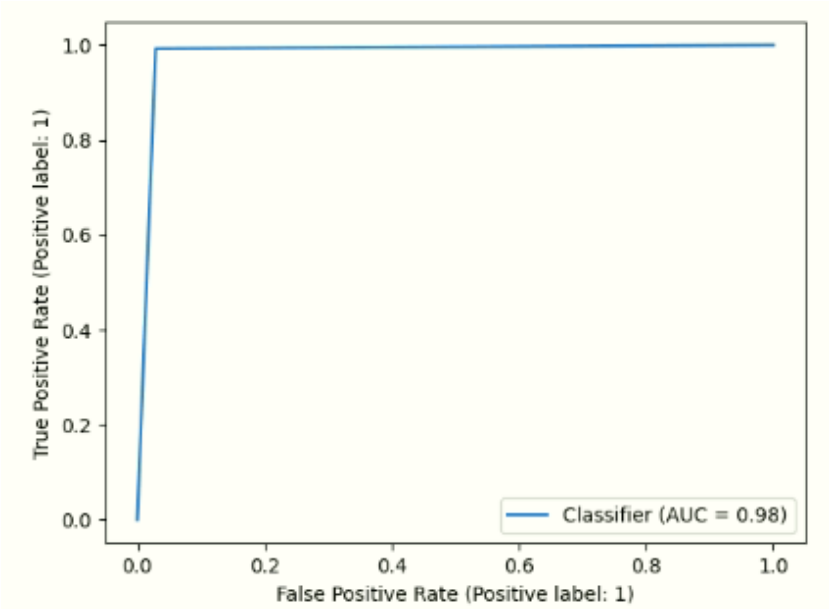
Στο 5ο πείραμα, πέρα από όλες τις παραμέτρους που χρησιμοποιούνται στο 4ο πείραμα, προστίθεται και τα 21 χαρακτηριστικά που εξάγονται από τα ορόσημα του προσώπου (MOCS, EAR και οι γωνίες με σημείο αναφοράς τα ορόσημα του εξωτερικού προσώπου μεταξύ δύο διανυσμάτων με αρχή το σημείο αναφοράς και πέρας τα ορόσημα του εσωτερικού προσώπου, κεφ. 3.1). Ο αριθμός της αλληλουχίας καρέ που επιλέγεται είναι 50. Συνολικά η είσοδος του μοντέλου είναι της μορφής $v \epsilon R^{50x227}$. Το μοντέλο εκπαιδεύεται για 200 εποχές, περισσότερες από τα προηγούμενα πειράματα λόγο του μεγαλύτερου μεγέθους της εισόδου.



Εικόνα 56: Τα losses κατά την εκπαίδευση και το validation για το Binary Classification (5ο πείραμα)

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	98.88	0.9993	0.9892	0.9942	0.9816
2η	**99.2**	**0.9992**	**0.9925**	**0.9959**	**0.9819**
3η	90.02	0.9949	0.9838	0.9893	0.9559

Πίνακας 19: Μετρικές Αξιολόγησης για το Binary Classification (5ο πείραμα). Παρατηρούμε ότι όλες οι μετρικές παρουσιάζουν καλύτερα αποτελέσματα για τη 2η μέθοδο αξιολόγησης.



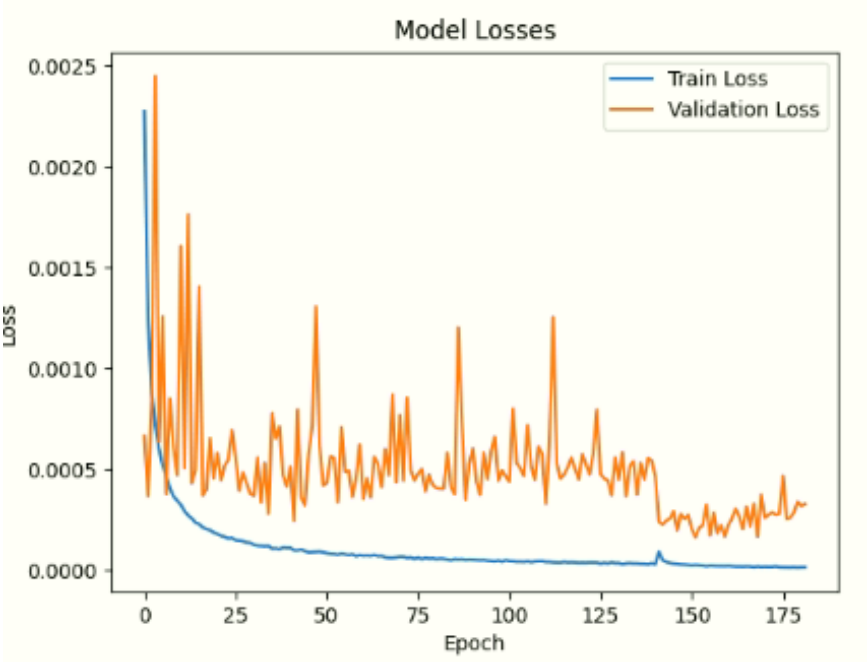
Εικόνα 57: AUC-ROC Curve για το Binary Classification (2η Μέθοδος Αξιολόγησης, 5ο Πείραμα)

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**777**	23
1	223	**29608**

Πίνακας 20: Μήτρα Σύγχυσης για το Binary Classification (5ο Πείραμα). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί σωστά την πλειοψηφία των δεδομένων.

4.3.5.6 Αποτελέσματα για το 6ο πείραμα

Για το 6ο και τελευταίο πείραμα χρησιμοποιήθηκαν τα ίδια χαρακτηριστικά με το 5ο πείραμα αλλά τα καρέ από 50 αυξήθηκαν σε 96. Συνολικά η είσοδος του μοντέλου είναι της μορφής $v \epsilon R^{96x227}$. Το μοντέλο εκπαιδεύεται για 181 εποχές.



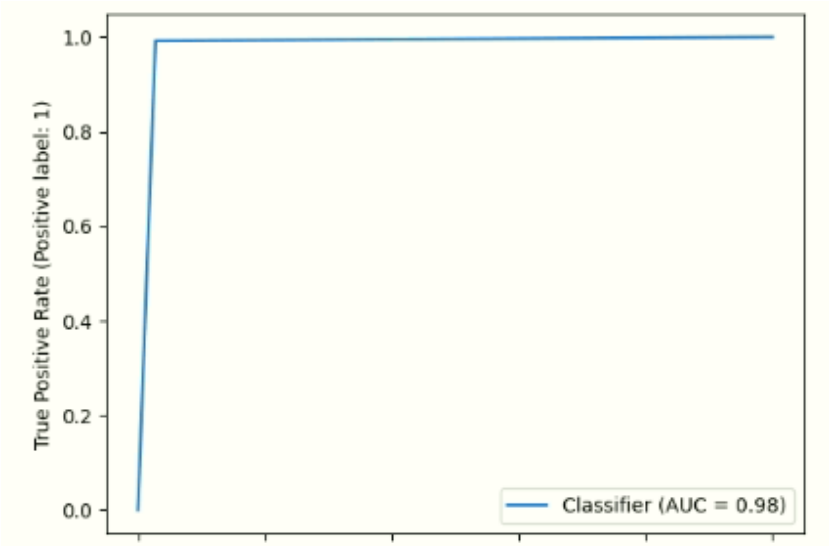
Εικόνα 58: Τα losses κατά την εκπαίδευση και το validation για το Binary Classification (6ο πείραμα)

Μετρική Αξιολόγησης / Μέθοδος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1η	99.09	0.9990	0.9917	0.9953	0.9765
2η	**99.16**	**0.9993**	**0.9921**	**0.9957**	**0.9823**
3η	98.16	0.9940	0.9862	0.9900	0.9539

Πίνακας 21: Μετρικές Αξιολόγησης για το Binary Classification (6ο πείραμα). Παρατηρούμε ότι όλες οι μετρικές παρουσιάζουν καλύτερα αποτελέσματα για τη 2η μέθοδο αξιολόγησης.

Πραγματική Κλάση / Κλάση πρόβλεψης	*0*	*1*
0	**778**	22
1	236	**29595**

Πίνακας 22: Μήτρα Σύγχυσης για το Binary Classification (6ο Πείραμα). Παρατηρούμε ότι το μοντέλο κατηγοριοποιεί σωστά την πλειοψηφία των δεδομένων.



Εικόνα 59: AUC-ROC Curve για το Binary Classification (2η Μέθοδος Αξιολόγησης, 6ο Πείραμα)

4.3.6 Συγκριτικά Αποτελέσματα για το Binary Classification

Παρατηρούμε ότι το EfficientNet v2 παρουσιάζει καλύτερα αποτελέσματα σε όλα τα πειράματα από τις 4 πρώτες μεθόδους. Επίσης και στις 5 μεθόδους παρατηρούμε ότι η 2η μέθοδος αξιολόγησης του testing set (αξιολόγηση της μέσης τιμής της εξόδου των μοντέλων (RMSE για τους One Class Classifiers και πιθανότητας να ανήκει στην αυθεντική κλάση για τον Binary Classifier) για κάθε αλληλουχία καρέ που ανήκουν στο ίδιο βίντεο) παρουσιάζει τα καλύτερα αποτελέσματα συγκριτικά με τις άλλες 2 μεθόδους αξιολόγησης. Οπότε και αυτή χρησιμοποιείται για την σύγκριση των 6 πειραμάτων για το Binary Classification.

Α/Α Πειράματος	***Αριθμός καρέ***	***Χαρακτηριστικά***
1	50	Shape, Expression & Pose parameters
2	96	Shape, Expression & Pose parameters
3	50	Shape, Expression, Pose & Texture Parameters
4	96	Shape, Expression, Pose & Texture Parameters
5	50	Shape, Expression, Pose, Texture Parameters & Landmark Based Features
6	96	Shape, Expression, Pose, Texture Parameters & Landmark Based Features

Πίνακας 23: Περιγραφή των πειραμάτων που πραγματοποιήθηκαν για το Binary Classification

Μετρικές Α/Α Πειράματος	*Ακρίβεια (Accuracy) %*	*Αξιοπιστία (Precision)*	*Ανάκληση (Recall)*	*F1-score*	*AUC ROC score*
1ο	99.12	0.9980	0.9929	0.9955	0.9596
2ο	99.13	0.9987	0.9924	0.9955	0.9718
3ο	99.19	0.9923	0.9994	0.9961	0.8559
4ο	**99.37**	0.9989	**0.9947**	**0.9968**	0.9761
5ο	99.2	0.9992	0.9925	0.9959	0.9819
6ο	99.16	**0.9993**	0.9921	0.9957	**0.9823**

Πίνακας 24: Σύγκριση των Μετρικών Αξιολόγησης για τα πειράματα του Binary Classification (2η Μέθοδος Αξιολόγησης)

Παρατηρούμε ότι για τα πειράματα που χρησιμοποιούσαν τα ίδια χαρακτηριστικά, αυτά στα οποία η είσοδος ήταν 96 καρέ έχουν καλύτερη απόδοση από αυτά όπου η είσοδος είναι 50 καρέ. Επίσης, για τον ίδιο αριθμό καρέ, τα μοντέλα που περιλαμβάνουν και τις παραμέτρους του texture έχουν καλύτερη απόδοση από αυτά που περιλαμβάνουν μόνο τις παραμέτρους για το σχήμα, την έκφραση και την πόζα. Την καλύτερη απόδοση για τον ίδιο αριθμό καρέ παρουσιάζουν τα μοντέλα που περιλαμβάνουν και τα χαρακτηριστικά τα οποία εξήχθησαν από τα ορόσημα του προσώπου. Έτσι για 50 καρέ το μοντέλου του πειράματος 5 έχει ακρίβεια 99.2 και AUC-ROC score 0.9819 που είναι μεγαλύτερα από τα αντίστοιχα μοντέλα για 50 καρέ. Επίσης για 96 καρέ το μοντέλο του πειράματος 6 έχει AUC-ROC score 0.9823 που είναι το καλύτερο AUC-ROC score από όλα τα πειράματα. Το 6ο μοντέλο δεν έχει την καλύτερη ακρίβεια (μικρότερη ακρίβεια από αυτή του 4ου πειράματος) αλλά λόγω της μεγάλης διαφοράς στον αριθμό των δειγμάτων από τις 2 κλάσεις, το AUC-ROC score θεωρείται πιο σημαντική μετρική από την ακρίβεια.

Συνεπώς παρατηρούμε ότι τα χαρακτηριστικά που δημιουργήθηκαν βοηθούν στην βελτίωση των μοντέλων. Επίσης τα μοντέλα με 96 καρέ έχουν καλύτερη απόδοση από αυτά με 50 καρέ. Το μοντέλο με το καλύτερο AUC-ROC score είναι αυτό που έχει είσοδο 96 καρέ με όλες τις παραμέτρους + τα χαρακτηριστικά από τα landmarks και το μοντέλο με την καλύτερη ακρίβεια έχει είσοδο 96 καρέ και μόνο τις παραμέτρους του FLAME.

Κεφάλαιο 5 Συμπεράσματα και Μελλοντική Επέκταση

Στο κεφάλαιο αυτό συνοψίζονται τα ευρήματα της παρούσας εργασίας και προτείνονται εναλλακτικές προσεγγίσεις οι οποίες δύναται να αξιοποιηθούν για την επίλυση του υπό εξέταση προβλήματος.

5.1 Συμπεράσματα

Η παρούσα εργασία κλήθηκε να αντιμετωπίσει το πρόβλημα της αναγνώρισης των deepfakes ή των παραποιημένων προσώπων γενικότερα. Ακολουθήθηκε μια προσέγγιση υψηλού επιπέδου καθώς τα χαρακτηριστικά που χρησιμοποιηθήκαν βασίστηκαν στα 3D βιομετρικά χαρακτηριστικά του προσώπου. Η προσέγγιση που προτάθηκε δεν λαμβάνει υπόψη μόνο τα χωρικά χαρακτηριστικά του προσώπου αλλά και την χρονική συνοχή της μεταβολής του προσώπου στην διάρκεια ενός βίντεο. Προτάθηκαν, επίσης, χαρακτηριστικά που εξάγονται από τα ορόσημα του προσώπου και που δύναται να βοηθήσουν στην αναγνώριση των deepfakes.

Η αρχική προσέγγιση του προβλήματος ως πρόβλημα κατηγοριοποίησης μιας τάξης δεν επίφερε ικανοποιητικά αποτελέσματα. Διαπιστώθηκε, όμως, ότι τα GAN που χρησιμοποιήθηκαν επέφεραν καλύτερα αποτελέσματα και είχαν περισσότερη διακριτική ικανότητα από τα VAE. Ειδικά το OCGAN επέδειξε ικανοποιητικά αποτελέσματα για το πρόβλημα της αναγνώρισης των deepfakes (65% ακρίβεια και 0.65 AUC-ROC score). Συνεπώς αποδεικνύεται ότι τα χαρακτηριστικά που επιλέχθηκαν μπορούν, πράγματι, να χρησιμοποιηθούν αποτελεσματικά στο υπό εξέταση πρόβλημα αν και η προσέγγιση που ακολουθήθηκε δεν είναι ιδιαίτερα αποτελεσματική.

Ωστόσο η προσέγγιση του προβλήματος ως πρόβλημα κατηγοριοποίησης δύο τάξεων, όπου τόσο αυθεντικά όσο και παραποιημένα βίντεο είναι παρούσα κατά την εκπαίδευση, επίφερε πολύ καλά αποτελέσματα. Το βαθύ συνελικτικό δίκτυο που επιλέχθηκε να χρησιμοποιηθεί (Efficient Net v2) εμφάνισε πολύ καλή διακριτική ικανότητα (99.16% ακρίβεια και 0.98 AUC-ROC score) για την αναγνώριση των αυθεντικών από τα παραποιημένα βίντεο. Καθώς τα παραποιημένα βίντεο προέρχονται από τα 4 διαφορετικές μεθόδους (deepfakes, face2face, FaceSwap, NeuralTextures) παρατηρείται ότι το μοντέλο έχει καλή ικανότητα γενίκευσης και σε μεθόδους εκτός αυτής των deepfakes. Επίσης διαπιστώθηκε ότι βίντεο μεγαλύτερης διάρκειας επιφέρουν καλύτερα αποτελέσματα αναγνώρισης των παραποιημένων βίντεο. Τέλος, παρατηρήθηκε ότι τα χαρακτηριστικά που προτάθηκαν βοηθούν στην περαιτέρω βελτίωση της αποτελεσματικότητας της μεθόδου και θα μπορούσαν, επομένως, να χρησιμοποιηθούν για την βελτίωση και άλλων μεθόδων αναγνώρισης παραποιημένων προσώπων.

5.2 Προτάσεις για Μελλοντική Έρευνα

Η προσέγγιση του προβλήματος ως πρόβλημα κατηγοριοποίησης μιας κλάσης, αν και δεν επίφερε πολύ ικανοποιητικά αποτελέσματα, απέδειξε ότι όντως μπορεί να γίνει αναγνώριση των παραποιημένων βίντεο με αυτό τον τρόπο. Στο πλαίσιο αυτό, ποιο εξελιγμένες αρχιτεκτονικές GAN ή και άλλων τύπων μοντέλων που λειτουργούν ως One Class Classifiers θα γινόταν να χρησιμοποιηθούν για την περαιτέρω βελτίωση της προσέγγισης. Επίσης θα μπορούσε να δημιουργηθεί ένα σετ δεδομένων που θα περιέχει περισσότερα αυθεντικά δεδομένα από αυτά που χρησιμοποιήθηκαν καθώς κάτι τέτοιο θα μπορούσε να βελτιώσει την αποτελεσματικότητα της μεθόδου.

Στα πλαίσια του προβλήματος του binary classification, θα μπορούσε να χρησιμοποιηθεί ένα σετ δεδομένων με βίντεο μεγαλύτερης διάρκειας καθώς αποδείχθηκε ότι κάτι τέτοιο θα βελτίωνε την απόδοση του μοντέλου. Επίσης η χρησιμοποίηση άλλων αρχιτεκτονικών πέρα από αυτή του Efficient Net πιθανόν να επίφερε καλύτερα αποτελέσματα στο πρόβλημα της αναγνώρισης των παραποιημένων προσώπων.

Τέλος, θα μπορούσε να χρησιμοποιηθεί και η προσέγγιση που βασίζεται στην ταυτότητα των προσώπων. Με τα χαρακτηριστικά που παρατηρήθηκε ότι έχουν τα καλύτερα αποτελέσματα για το binary classification, θα γινόταν να εκπαιδευτεί ένα μοντέλο που θα αναγνώριζε την ταυτότητα του εικονιζόμενου προσώπου στο βίντεο. Έτσι, έχοντας ένα reference βίντεο από το οποίο θα εξήγαγε την ταυτότητα του προσώπου, θα μπορούσε να αναγνωρίσει αν άλλα βίντεο που εικάζουν πως δείχνουν το ίδιο πρόσωπο έχουν την ίδια ταυτότητα και να τα κατηγοριοποιούσε, με αυτό τον τρόπο, ως αυθεντικά ή παραποιημένα.

Αναφορές

[1] Cozzolino, D., Rössler, A., Thies, J., Nießner, M., & Verdoliva, L. (2020). ID-Reveal: Identity-aware DeepFake Video Detection (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2012.02512

[2] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1406.2661

[3] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In European Conference on Computer Vision (ECCV), 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016

[5] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–6, 2020.

[6] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)

[7] Wiles, O., Koepke, A.S., Zisserman, A.: Self-supervised learning of a facial attribute embedding from video. arXiv: 1808.06882 (2018)

[8] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)

[9] Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep speaker recognition. arXiv: 1806.05622 (2018)

[10] Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano and Hao Li. “Protecting World Leaders Against Deep Fakes.” CVPR Workshops (2019).

[11] Bernhard Scholkopf, John C. Platt, John C. Shawe-Taylor, ¨ Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. Neural Computation, 13(7):1443–1471.

[12] Paul Ekman and Wallace V Friesen. Measuring facial movement. Environmental psychology and nonverbal behavior, 1(1):56–75, 1976.

[13] Agarwal, Shruti and Farid, Hany and Fried, Ohad and Agrawala, Maneesh. Detecting Deep-Fake Videos From Phoneme-Viseme Mismatches. Agarwal_2020_CVPR_Workshops

[14] Franc¸ois Chollet. Xception: Deep learning with depthwise separable convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[15] Y. Li, M. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630787.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Comput., 9(8):1735–1780, 1997

[17] Yang, X., Li, Y., & Lyu, S. (2018). Exposing Deep Fakes Using Inconsistent Head Poses (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1811.00661

[18] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. 2020. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 4318–4327.

[19] X. Niu, S. Shan, H. Han, and X. Chen. 2020. RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. IEEE Transactions on Image Processing 29 (2020), 2409–2423

[20] M. Li, B. Liu, Y. Hu, L. Zhang and S. Wang, "Deepfake Detection Using Robust Spatial and Temporal Features from Facial Landmarks," 2021 IEEE International Workshop on Biometrics and Forensics (IWBF), 2021, pp. 1-6, doi: 10.1109/IWBF50991.2021.9465076.

[21] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Chen, D., Wen, F., & Guo, B. (2020). Identity-Driven DeepFake Detection (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2012.03930

[22] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[23] Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). DeepFake Detection Based on the Discrepancy Between the Face and its Context (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2008.12262

[24] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in Int. Conf. on Automatic Face and Gesture Recognition. IEEE, 2018, pp. 67–74

[25] H. M. Nguyen and R. Derakhshani, "Eyebrow Recognition for Identifying Deepfake Videos," *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020, pp. 1-5.

[26] Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv: 1811.00656 (2018)

[27] Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)

[28] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In IEEE International Workshop on Information Forensics and Security, pages 1–7, 2018

[29] Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. arxiv: 1907.06515 (2019)

[30] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH '99). ACM Press/Addison-Wesley Publishing Co., USA, 187–194. https://doi.org/10.1145/311535.311556

[31] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. 36, 6, Article 194 (December 2017), 17 pages. https://doi.org/10.1145/3130800.3130813

[32] S. Geman and D. E. McClure. 1987. Statistical methods for tomographic image reconstruction. Proceedings of the 46th Session of the International Statistical Institute,Bulletin of the ISI 52 (1987)

[33] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. ACM Trans. Graph. 40, 4, Article 88 (August 2021), 13 pages. https://doi.org/10.1145/3450626.3459936

[34] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter.2009. A 3D face model for pose and illumination invariant face recognition. International Conference on Advanced Video and Signal Based Surveillance. 296–301.

[35] Ramamoorthi and P. Hanrahan. 2001. An efficient representation for irradiance environment maps. Proceedings of the 28th annual conference on Computer graphics and interactive techniques (2001)

[36] M. S. B. M. M. P. W. “Research Paper on Basic of Artificial Neural Network”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, no. 1, Jan. 2014, pp. 96-100, doi:10.17762/ijritcc.v2i1.2920.

[37] O’Shea, Keiron, and Ryan Nash. 2015. “An Introduction to Convolutional Neural Networks.” arXiv. https://doi.org/10.48550/ARXIV.1511.08458.

[38] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chap. Learning Internal Representations by Error Propagation, pp. 318–362. MIT Press, Cambr.

[39] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR abs/1312.6114 (2013).

[40] Kingma, Diederik & Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

[41] Amari, S. (1993). Backpropagation and stochastic gradient descent method. Neurocomputing, 5, 185-196.

[42] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6309–6318.

[43] M. Nilsson, I. Gertsovich and J. S. Bartůněk, "Mouth open or closed decision for frontal face images with given eye locations," 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2010, pp. 1-6, doi: 10.1109/BTAS.2010.5634522.

[44] Dewi, Christine & Chen, Rung-Ching & Jiang, Xiaoyi & Yu, Hui. (2022). Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks. PeerJ Computer Science. 8. e943. 10.7717/peerj-cs.943.

[45] https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

[46] Perera, Pramuditha & Nallapati, Ramesh & Xiang, Bing. (2019). OCGAN: One-Class Novelty Detection Using GANs With Constrained Latent Representations. 2893-2901. 10.1109/CVPR.2019.00301.

[47] Tan, Mingxing & Le, Quoc. (2021). EfficientNetV2: Smaller Models and Faster Training.

[48] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1-11

[49] https://decrypt.co/101365/deepfake-video-elon-musk-crypto-scam-goes-viral

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: an imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 721, 8026–8037.

[51] Pedregosa, Fabian and Varoquaux, Gaël and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Müller, Andreas and Nothman, Joel and Louppe, Gilles and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and Vanderplas, Jake and Passos, Alexandre and Cournapeau, David and Brucher, Matthieu and Perrot, Matthieu and Duchesnay, Édouard. 2012. Scikit-learn: Machine Learning in Python. https://doi.org/10.48550/arxiv.1201.0490.

[52] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2.

[53] Barrett, Paul & Hunter, J. & Miller, J.T. & Hsu, J.-C & Greenfield, P.. (2005). matplotlib -- A Portable Python Plotting Package.

[54] Ioffe, Sergey and Szegedy, Christian. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://doi.org/10.48550/arxiv.1502.03167.

[55] Thanh Thi Nguyen and Quoc Viet Hung Nguyen and Dung Tien Nguyen and Duc Thanh Nguyen and Thien Huynh-The and Saeid Nahavandi and Thanh Tam Nguyen and Quoc-Viet Pham and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding. 10.1016/j.cviu.2022.103525

[56] Thies, Justus and Zollhöfer, Michael and Stamminger, Marc and Theobalt, Christian and Nießner, Matthias. Face2Face: Real-time Face Capture and Reenactment of RGB Videos, 2020. https://doi.org/10.48550/arxiv.2007.14808

[57] Korshunova, Iryna and Shi, Wenzhe and Dambre, Joni and Theis, Lucas. Fast Face-swap Using Convolutional Neural Networks, 2016. https://doi.org/10.48550/arxiv.1611.09577

[58] Ye, Zipeng and Sun, Zhiyao and Wen, Yu-Hui and Sun, Yanan and Lv, Tian and Yi, Ran and Liu, Yong-Jin. Dynamic Neural Textures: Generating Talking-Face Videos with Continuously Controllable Expressions, 2022. https://doi.org/10.48550/arxiv.2204.06180