ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
Πρόγραμμα Μεταπτυχιακών Σπουδών

«***Προηγμένα Συστήματα Πληροφορικής – Ανάπτυξης Λογισμικού και Τεχνητής Νοημοσύνης***»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Δημιουργία και ανίχνευση επιθέσεων μεταμόρφωσης προσώπου με τη χρήση τεχνικών βαθιάς μάθησης** **Face Morphing Attack Generation and Detection using Deep Learning Techniques**
Ονοματεπώνυμο Φοιτητή	ΙΩΑΝΝΗΣ ΠΡΟΔΡΟΜΟΥ
Πατρώνυμο	ΣΟΦΟΚΛΗΣ
Αριθμός Μητρώου	ΜΠΣΠ2337
Επιβλέπων	ΔΙΟΝΥΣΙΟΣ ΣΩΤΗΡΟΠΟΥΛΟΣ, ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ

Ημερομηνία Παράδοσης Ιούλιος 2026

Τριμελής Εξεταστική Επιτροπή

Διονύσιος Σωτηρόπουλος Αναπληρωτής Καθηγητής	Γεώργιος Τσιχριντζής Καθηγητής	Ευάγγελος Σακκόπουλος Καθηγητής

Δήλωση Πνευματικών Δικαιωμάτων

Δηλώνω υπεύθυνα ότι η παρούσα μεταπτυχιακή διπλωματική εργασία αποτελεί δικό μου πρωτότυπο έργο, το οποίο εκπονήθηκε στα πλαίσια των σπουδών μου στο Τμήμα Πληροφορικής του Πανεπιστημίου Πειραιώς. Οι πηγές που χρησιμοποιήθηκαν αναφέρονται ρητά στο κείμενο, ενώ τα συμπεράσματα και οι κρίσεις είναι προϊόντα προσωπικής μελέτης και ανάλυσης.

Επιτρέπεται η αναπαραγωγή ή αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμημάτων αυτής, για εκπαιδευτικούς ή ερευνητικούς σκοπούς, με την προϋπόθεση της σαφούς αναφοράς στην πηγή και την μη εμπορική χρήση. Ζητήματα που σχετίζονται με χρήση της εργασίας για κερδοσκοπικούς σκοπούς πρέπει να απευθύνονται στον συγγραφέα.

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής μου διπλωματικής εργασίας, θα ήθελα να ευχαριστήσω θερμά όσους με στήριξαν σε αυτή την προσπάθεια. Πρώτα από όλους, εκφράζω την ευγνωμοσύνη μου στον επιβλέποντα καθηγητή μου, [Διονύσιο Σωτηρόπουλο], για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου αυτό το θέμα, για την συνεχή καθοδήγηση και τις πολύτιμες συμβουλές καθ' όλη τη διάρκεια της εκπόνησης. Η μεθοδική του προσέγγιση και η αμέριστη υποστήριξη υπήρξαν καθοριστικές για την ολοκλήρωση της εργασίας. Ευχαριστώ επίσης τα μέλη της τριμελούς εξεταστικής επιτροπής για τον χρόνο και την προσοχή που αφιέρωσαν στην αξιολόγηση της παρούσας εργασίας. Τέλος, ένα μεγάλο ευχαριστώ οφείλω στην οικογένειά μου και στους φίλους μου, οι οποίοι με στήριξαν τόσο ηθικά όσο και πρακτικά σε όλη τη διάρκεια των σπουδών μου. Ιδιαίτερα, θα ήθελα να τους ευχαριστήσω για την κατανόηση, την υπομονή και την ανοχή τους απέναντι στη γκρίνια και τον εκνευρισμό μου, που προέκυπταν από τις απαιτήσεις της εργασίας και των σπουδών παράλληλα. Η παρουσία και η στήριξή τους υπήρξαν πολύτιμες, καθώς μοιράστηκαν μαζί μου τόσο τις επιτυχίες όσο και τις δυσκολίες αυτού του ταξιδιού.

[Ιωάννης Προδρόμου]

Πειραιάς, Ιούλιος 2026

Περίληψη

Τα τελευταία χρόνια, τα συστήματα βιομετρικής αναγνώρισης προσώπου έχουν ενταχθεί σε πολλές εφαρμογές υψηλής ασφάλειας, όπως η έκδοση ηλεκτρονικών διαβατηρίων, η πρόσβαση στα σύνορα λόγω ελέγχου και η είσοδος σε λεπτές υπηρεσίες. Παρά τη καθημερινή εξέλιξη της ορθότητας τους, τα συστήματα αυτά διατηρούν μια ευάλωτη κατάσταση σε μια διαφορετική εξελιγμένη μορφή απάτης: τις επιθέσεις μορφοποίησης προσώπου (face morphing attacks). Πρόκειται για τεχνικές που ενώνουν τα χαρακτηριστικά δύο διαφορετικών προσώπων σε μία και μόνο εικόνα, η οποία μπορεί να ταυτοποιηθεί αποτελεσματικά και ως προς τους δύο αρχικά πρόσωπα.

Η εν λόγω μεταπτυχιακή διπλωματική εργασία ερευνά το πρόβλημα από τις δύο όψεις του: της παραγωγής και του εντοπισμού. Στο πρώτο μέρος, δημιουργούνται τέσσερις διαφορετικές μέθοδοι δημιουργίας morphed εικόνων, οι οποίες εκπροσωπούν τόσο τις παραδοσιακές όσο και τις πιο μοντέρνες τάσεις: διαδικασία βασισμένη σε σημεία αναφοράς (landmark-based), αναβαθμισμένη μορφή με Poisson blending, δημιουργία μέσω StyleGAN2 και τέλος δημιουργία μέσω μοντέλων διάχυσης (DDPM).

Στη συνέχεια, προετοιμάζεται και κρίνεται ένα σύστημα ανίχνευσης (Morphing Attack Detection - MAD) στηριγμένο σε ιδιομορφίες Multi-Block Local Binary Patterns (MB-LBP) και διαχωριστή Support Vector Machine (SVM). Δημιουργείται αναλυτική πειραματική εξέταση με χρήση cross-evaluation, ώστε να αναλυθεί η δεξιότητα γενίκευσης κάθε εκπαιδευμένου μοντέλου έναντι διαφορετικών εκδόσεων επιθέσεων.

Τα δεδομένα αποκαλύπτουν ότι οι διαχωριστές / ταξινομητές που προσαρμόζονται σε έναν ειδικό τύπο morph εμφανίζουν δυνατή πτώση απόδοσης όταν υλοποιούνται σε ξεχωριστούς τύπους, επισημαίνοντας το καθήκον για πιο πλήρης και ισχυρά προγράμματα ανίχνευσης. Εξαιρετικά ενδιαφέρον παρουσιάζει η επισήμανση ότι οι γεννητικές μέθοδοι (StyleGAN2, Diffusion, DDPM) εμφανίζονται με ίδια στατιστικά μοτίβα, όπου έχουν διαφορά σε σχέση με αυτά των landmark-based morphs.

Η εργασία ολοκληρώνεται με σχολιασμό των αποτελεσμάτων, αναφορά στον κώδικα δεοντολογίας της έρευνας και προτάσεις για περαιτέρω έρευνα .

Λέξεις-κλειδιά: face morphing, βιομετρική αναγνώριση, StyleGAN2, μοντέλα διάχυσης, DDPM, ανίχνευση επιθέσεων, MB-LBP, SVM, βαθιά μάθηση.

Abstract

In recent years, biometric facial recognition systems have been integrated into many high-security applications, such as the issuance of electronic passports, border access control, and entry into sensitive facilities. Despite daily improvements in their accuracy, these systems remain vulnerable to a different, sophisticated form of fraud: face morphing attacks. These are techniques that combine the features of two different faces into a single image, which can be effectively identified as belonging to both original faces.This master's thesis investigates the problem from both angles: generation and detection.In the first part, four different methods for creating morphed images are developed, representing both traditional and more modern approaches: a landmark-based process, an enhanced version using Poisson blending, generation via StyleGAN2, and finally generation via diffusion models (DDPM). Next, a detection system (Morphing Attack Detection – MAD) based on Multi-Block Local Binary Patterns (MB-LBP) features and a Support Vector Machine (SVM) classifier is developed and evaluated. A comprehensive experimental evaluation using cross-validation is conducted to analyze the generalization ability of each trained model against different attack variants.The data reveal that classifiers tailored to a specific morph type exhibit a significant drop in performance when applied to other types, highlighting the need for more comprehensive and robust detection systems. Of particular interest is the observation that generative methods (StyleGAN2, Diffusion, DDPM) exhibit the same statistical patterns, whereas they differ from those of landmark-based morphs.The paper concludes with a discussion of the results, a reference to the research code of ethics, and suggestions for further research.

Keywords: face morphing, biometric recognition, StyleGAN2, diffusion models, DDPM, attack detection, MB-LBP, SVM, deep learning.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1: Εισαγωγή

1.1 Το Πρόβλημα της Βιομετρικής Ασφάλειας

Η ραγδαία ανάπτυξη της τεχνητής νοημοσύνης και των τεχνικών διαχείρισης εικόνας έχει προσαρμόσει μια νέα πραγματικότητα στην βιομετρικής αναγνώριση. Τα συστήματα αυτόματης αναγνώρισης προσώπου εφαρμόζονται κάθε μέρα σε πολλά προγράμματα, από το άνοιγμα του τηλεφώνου μας μέχρι στα αεροδρόμια όπου γίνεται επιθεώρηση διαβατηρίων. Η μαζική αποδοχή τους εξηγείται στη συνθετική ικανότητά τους να διαθέτουν αυξημένη ασφάλεια, ταχύτητα και άνετο περιβάλλον για τον χρήστη.

Παρ' όλα αυτά, ταυτόχρονα με την ανάπτυξη αυτή, δημιουργούνται και πιο σύχρονες διαδικασίες προσπέρασης των βιομετρικών συστημάτων. Μία από τις πιο επικίνδυνες απειλές είναι η μορφοποιημένη επίθεση προσώπου, η αλλιώς ως face morphing attack. Είναι μια τεχνική όπου οι εικόνες δύο ξεχωριστών ατόμων ενώνονται μέσω ψηφιακής επεξεργασίας σε μία μοναδική εικόνα, η οποία έχει επαρκή ομοιότητα σχεδόν με τα αρχικά πρόσωπα, ώστε ένα σύστημα αναγνώρισης να την επικυρώνει επιτυχώς και για τους δύο συμμετέχοντες.

Η εφαρμόσιμη κατάσταση της επίθεσης είναι ιδιαίτερα ανησυχητική. Στον πιο απλό τύπο του, ένας υπό αναζήτηση εγκληματίας μπορεί να συνεργαστεί με έναν συνεργό, όπου έχει στην κατοχή του νόμιμα δικαιολογητικά. Κατασκευάζουν μαζί μια morphed εικόνα και ο συνεργός την αναθέτει στην αρμόδια αρχή για την δημιουργία διαβατηρίου ή ταυτότητας. Το διαβατήριο δημιουργείται κανονικά, διότι το σύστημα δεν βρίσκει κάποια διαφορά. Έπειτα όμως, και ο εγκληματίας έχει την ικανότητα να χρησιμοποιήσει το ίδιο έγγραφο για να περάσει σύνορα ή να πάρει πρόσβαση σε οργανισμούς, αφού το σύστημα τον αναγνωρίζει επιτυχώς.

1.2 Κίνητρο και Επικαιρότητα της Έρευνας

Από το 2014, όταν ο Ferrara και οι συνεργάτες του εμφάνισαν για πρώτη φορά την ιδέα του morphing attack ως πραγματική απειλή για παραβίαση ηλεκτρονικών διαβατηρίων, το ερευνητικό δίκτυο έχει μελετήσει έντονα με το θέμα. Παρά τις ενέργειες που έχουν γίνει, η ταχύτητα ανάπτυξης των γεννητικών μοντέλων νευρωνικών δικτύων (Generative Adversarial Networks και πιο πρόσφατα Diffusion Models) έχει φτιάξει μια δυσανάλογη κατάσταση: οι τρόποι δημιουργίας αναβαθμίζονται με πολύ γρήγορο ρυθμό σε αντίθεση με τις μεθόδους ανίχνευσης.

Η πρόσφατη βιβλιογραφία αναφέρει ότι ένα βασικό πρόβλημα των υπάρχων ανιχνευτών είναι η μικρή τους ικανότητα γενίκευσης. Ειδικά, ένας ανιχνευτής εκπαιδευμένος σε morphs που έχουν παραχθεί με συγκεκριμένη μέθοδο, εμφανίζει τακτικά ουσιώδης πτώση απόδοσης όταν καλείται να εντοπίσει morphs που έχουν δημιουργηθεί με ξεχωριστή τεχνική. Αυτό το γεγονός, που στη βιβλιογραφία αναφέρεται ως cross-domain generalization failure, είναι ένας από τα κυριότερα εμπόδια της σύγχρονης τεχνολογίας MAD.

1.3 Στόχοι της Διπλωματικής Εργασίας

Η υφιστάμενη μεταπτυχιακή διπλωματική εργασία προσπαθεί να επιδράσει στη μελέτη του ζητήματος των επιθέσεων μορφοποίησης προσώπου με μια προσέγγιση εκ του συνόλου. Συγκεκριμένα, οι βασικοί στόχοι της εργασίας είναι:

1. Η ανάλυση και δημιουργία τεσσάρων διακριτών διαδικασιών παραγωγής morphed εικόνων, οι οποίες καλύπτουν τη κλίμακα από τις κλασσικές τεχνικές μέχρι τις πλέον σύγχρονες μεθόδους που στηρίζονται σε μοντέλα γεννητικά διάχυσης.
2. Η εφαρμογή και ο σχεδιασμός ενός μηχανισμού παρακολούθησης (MAD) στηριζόμενο σε ιδιότητες Multi-Block Local Binary Patterns (MB-LBP) σε ένωση με ταξινομητή Support Vector Machine (SVM).
3. Η οργάνωση εκτεταμένης εργαστηριακής αξιολόγησης με χρήση cross-evaluation, ώστε να υπολογιστεί αριθμητικά η ικανότητα γενίκευσης κάθε σαρωτή απέναντι σε διαφορετικούς τύπους επιθέσεων.

4. Η αξιολόγηση των ευρημάτων με στόχο την απόσπαση συμπερασμάτων με βάση τα δυνατά σημεία και τις δεσμεύσεις των υπάρχων διαδικασιών.
5. Η εφαρμογή με τις αξίες της βιομετρικής έρευνας, ειδικότερα με βάση την προστασία ανηλίκων από τον διαμοιρασμό των δεδομένων τους σε morphing pipelines.

1.4 Συμβολή της Εργασίας

Η διατριβή διαθέτει πολλές καινοτόμες συνδρομές στο πεδίο των επιθέσεων μορφοποίησης προσώπου. Αρχικά, δημιουργείται ένα πλήρες pipeline τεσσάρων ξεχωριστών τύπων morphing, εφαρμοσμένο εξ ολοκλήρου τοπικά σε τυπικό υλικό καταναλωτή (NVIDIA RTX 4070). Επιπλέον, δημιουργείται μια νέα ποσότητα δεδομένων που περιέχει συνολικά 1.444 morphed εικόνες οργανωμένες σε τέσσερις ξεχωριστές κατηγορίες, μαζί με 1.000 πραγματικές εικόνες αναφοράς.

Τέλος, εφαρμόζεται μια συνεχής παρομοίωση όλων αυτών των τύπων μέσω μιας μοναδικής μεθοδολογίας εντοπισμού, διαθέτοντας ισχυρή ενημέρωση σχετικά με τη μετακίνηση των ανιχνευτών από έναν χαρακτήρα επίθεσης σε άλλον. Η εργασία εμφανίζει επίσης τη σημασία της διάστασης ηθικής στην έρευνα μορφοποίησης, εφαρμόζοντας ένα στάδιο αυτόματης επιλογής που αποκλείει τη λειτουργία βιομετρικών δεδομένων ανηλίκων από το τελικό dataset.

1.5 Διάρθρωση της Εργασίας

Η εργασία διαμορφώνεται σε οκτώ ενότητες. Η εν λόγω ενότητα εμβαθύνει τον αναγνώστη στο ζήτημα και στις προθέσεις της εργασίας. Το μέρος 2 παρουσιάζει την βιβλιογραφική επισκόπηση του πεδίου, επεξεργάζοντας τις βασικότερες προσεγγίσεις τόσο στην δημιουργία όσο και στον εντοπισμό των απειλών.

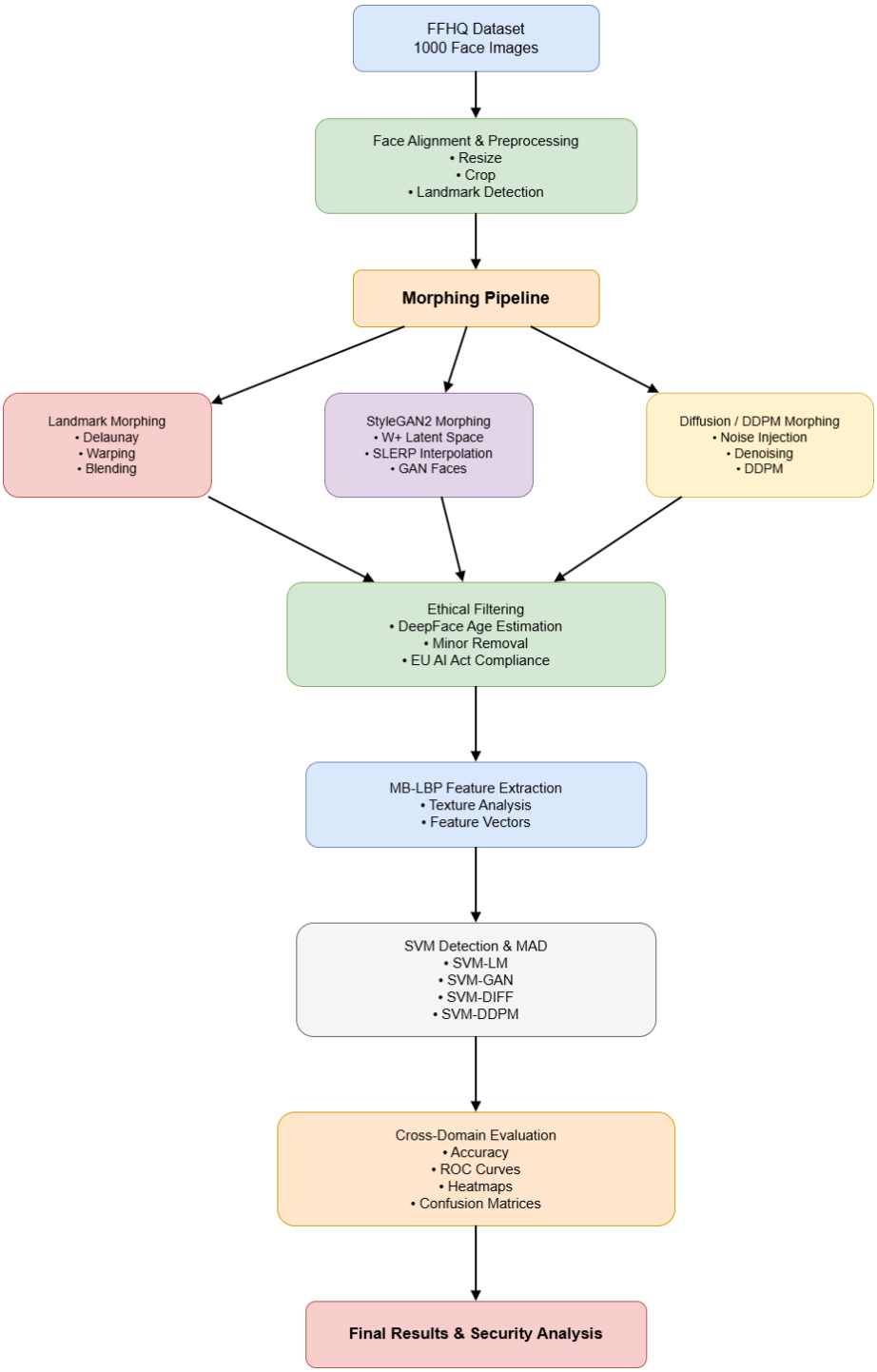
Στην ενότητα 3 προβάλλεται το θεωρητικό υπόβαθρο, με λεπτομερή περιγραφή των αρχιτεκτονικών των γεννητικών μοντέλων που εφαρμόστηκαν (StyleGAN2, DDPM), καθώς και των τεχνικών εξαγωγής χαρακτηριστικών και ταξινόμησης. Η ενότητα 4 αναλύει την μεθοδολογία δημιουργίας κάθε μιας από τις τέσσερις τεχνικές morphing.

Η ενότητα 5 παρουσιάζει την αρχιτεκτονική του συστήματος ανίχνευσης. Στην ενότητα 6 εξετάζονται λεπτομερώς τα δεδομένα της έρευνας, ενώ η ενότητα 7 διατίθεται σε μια αντιπαράθεση επιχειρημάτων και στα θέματα αρχών. Τέλος, η ενότητα 8 συγκεντρώνει τα αποτελέσματα και υποδεικνύει προσεγγίσεις για επόμενα στάδια.

1.6 Μεθοδολογική Προσέγγιση

Η μεθοδολογική προσέγγιση που ενσωματώθηκε στην διπλωματική είναι κατά βάση ερευνητική. Αντί να μείνει σε ανάλυση θεωρητικά, η εργασία βασίζεται στην αληθινή πραγματοποίηση και εφαρμογή όλων των λειτουργιών σε αληθινά δεδομένα, ώστε τα αποτελέσματα να πηγάζουν από εμπειρικές μετρήσεις και όχι από θεωρητικές υποθέσεις.

Η διαδικασία αυτή έχει το πλεονέκτημα ότι προβάλλει ουσιαστικά ερωτήματα που συχνά δεν είναι ορατά στη θεωρία, όπως οι δεσμεύσεις του υλικού, η υπολογιστική πολυπλοκότητα ορισμένων τεχνικών, και τα γεγονότα που παρουσιάζονται κατά το αποτέλεσμα. Ταυτόχρονα, εγγυάται την αναπαραγωγιμότητα: κάθε επακόλουθο έχει την ικανότητα να επιβεβαιωθεί δημιουργώντας εκ νέου τα ανάλογα scripts.



Σχήμα 1.1: Γενική ροή της εργασίας: Dataset -> Morphing -> Filtering -> Detection -> Evaluation

Ένα δεύτερο διαδικαστικό στοιχείο είναι η συγκριτική φύση της μελέτης. Αντί να βασιστεί σε μία μόνο τεχνική, η εργασία συγκρίνει συστηματικά τέσσερις διαφορετικές ικανότητες παραγωγής μέσα από ένα αδιάσπαστο πλαίσιο εντοπισμού, δίνοντας έτσι την εξαγωγή αποτελεσμάτων σχετικά με τις σχέσεις μεταξύ τους.

1.7 Συνεισφορά σε Σχέση με την Υπάρχουσα Βιβλιογραφία

Σε σχέση με την υπάρχουσα βιβλιογραφία, η συγκεκριμένη διπλωματική αλλάζει σε τρία κύρια σημεία. Αρχικά, ενώνει τέσσερις διαφορετικές κατηγορίες χρονολογικά τεχνικών morphing ,από τις κλασικές landmark-based μέχρι τα μοντέρνα μοντέλα διάχυσης σε μία αδιάσπαστη πειραματική μελέτη, κάτι που αραιά και που συναντάται στη βιβλιογραφία όπου οι μελέτες τείνουν να επικεντρώνονται σε μία ή δύο μεθόδους. Επιπλέον, δίνει διαφορετική έμφαση στο ζήτημα της γενίκευσης μέσω της στρατηγικής cross-evaluation, διαθέτοντας έναν ολοκληρωμένο πίνακα μεταφερσιμότητας μεταξύ όλων των τύπων. Τέλος, εφαρμόζει ρητά την ηθική διάσταση μέσω αυτοματοποιημένου περιορισμού ηλικίας, ένα στοιχείο που, παρά τη σημασία του, συχνά δεν εφαρμόζεται από άλλες μελέτες.

ΚΕΦΑΛΑΙΟ 2: Βιβλιογραφική Επισκόπηση

2.1 Εισαγωγή στις Επιθέσεις Μορφοποίησης Προσώπου

Το νόημα της επίθεσης μορφοποίησης προσώπου ως απειλής για βιομετρικά συστήματα επιβεβαιώθηκε επιστημονικά για πρώτη φορά από τους Ferrara, Franco και Maltoni στην ημερίδα με την εργασία τους με τίτλο "The Magic Passport" το 2014 [1]. Στην εργασία αυτή επαληθεύτηκε πρακτικά ότι τα τότε αποδοτικά συστήματα αναγνώρισης προσώπου ήταν επιρρεπής σε morphed εικόνες δημιουργημένες ακόμη και με σχετικά εύκολα τροποποιήσιμα εργαλεία .

Κατά την διάρκεια των χρόνων, το αντικείμενο αναπτύχθηκε με δύο παράλληλες κατευθύνσεις: από τη μία πλευρά, δημιουργήθηκαν όλο και πιο βελτιωμένες τεχνικές παραγωγής morphs, και από την άλλη, τροποποιήθηκαν αντίστοιχοι μηχανισμοί ανίχνευσης, οι οποίοι ονομάζονται Morphing Attack Detection (MAD).

2.2 Παραδοσιακές Μέθοδοι Παραγωγής Morphs

Οι κλασικές μέθοδοι παραγωγής morphs στηρίζονται κατά κύριο λόγο στη εφαρμογή σημείων αναφοράς (landmarks) του προσώπου. Χονδρικά, ο αλγόριθμος ανακαλύπτει πρώτα τα χαρακτηριστικά σημεία και των δύο εικόνων (συνήθως 68 ή 81 σημεία που περικλείουν την περίμετρο του προσώπου, τα μάτια, τη μύτη, το στόμα και τα φρύδια). Έπειτα, εφαρμόζεται ο μέσος όρος των αντίστοιχων σημείων ανά δυάδα, και κάθε πρώτη εικόνα υπάγεται σε μετασχηματιστική γεωμετρία ώστε τα σημεία της να εφάπτονται με τις μέσες θέσεις.

Η βασικότερη φάση είναι ο τελικός σχηματισμός, όπου τα δύο γεωμετρικά μετασχηματισμένα πρόσωπα ενώνονται μέσω ενός γραμμικού συνδυασμού των χρωματικών τους τιμών (alpha blending). Αν και η διαδικασία αυτή είναι μαθηματικά απλή και υπολογιστικά οικονομική, εμφανίζει σοβαρά μειονεκτήματα ποιότητας, όπως εμφανίσιμα artifacts ("ghost effects") στο μέρος των μαλλιών, του φόντου και σε στοιχεία όπως σκουλαρίκια ή γυαλιά. Φρέσκες εξελίξεις των landmark-based τεχνικών συμπεριλαμβάνουν τη χρήση Poisson blending για τον συνδυασμό του morphed προσώπου με το αρχικό υπόβαθρο, όπως συστήνεται από τους Batskos και Spreeuwers (2024) [16]. Η μέθοδος αυτή ελαττώνει βασικά τα ορατά artifacts, αν και δεν τα παραμερίζει εντελώς.

2.3 Παραγωγή με Generative Adversarial Networks

Με την παρουσία των Generative Adversarial Networks (GANs) και συγκεκριμένα των εξελίξεων τους όπως το StyleGAN του Karras και των συνεργατών του (2019) [12], δημιουργήθηκαν νέες δυνατότητες για την κατασκευή υψηλής ποιότητας morphs. Η θεμελιώδης αρχή είναι ότι μια εκπαιδευμένη γεννήτρια (generator) έχει την ικανότητα να ενώσει διανύσματα από έναν λανθάνοντα χώρο (latent space) σε φωτορεαλιστικές εικόνες προσώπων. Η υλοποίηση γραμμικής παρεμβολής (interpolation) ανάμεσα σε latent αναπαραστάσεων δύο πραγματικών προσώπων μπορεί να δημιουργήσει μια μέση εικόνα που κατέχει χαρακτηριστικά και των δύο.

Η εργασία των Venkatesh, Zhang, Ramachandra, Raja, Damer και Busch (2020) [9] παρουσίασε ότι τα GAN-based morphs έχουν την ικανότητα να δημιουργήσουν εξίσου σημαντικό κίνδυνο για τα συστήματα αναγνώρισης προσώπου με τα παραδοσιακά. Η σημαντικότερη πρόκληση είναι η διαδικασία της προβολής (projection), δηλαδή η εύρεση του διανύσματος W+ που αντιστοιχεί σε μια πραγματική εικόνα εισόδου, η οποία είναι υπολογιστικά απαιτητική και απαιτεί iterative optimization.

Η πιο φρέσκια βιβλιογραφία (Sarkar et al., 2022) [10] έχει εστιάσει στη πρόοδο της προβολής με στόχο την δημιουργία morphs που είναι πιο διακριτά ως προς τα πρώτα πρόσωπα, ενισχύοντας έτσι την απειλή προς τα FRS.

Ένα βασικό πλεονέκτημα των GAN-based morphs έναντι των landmark-based είναι η απουσία ορατών artifacts. Επειδή κάθε εικόνα που δημιουργεί η γεννήτρια είναι εξ ορισμού μια “έγκυρη” εικόνα προσώπου από τον καταμερισμό εκπαίδευσης, δεν παρουσιάζονται φαινόμενα ghosting

ή ασυνέχειες. Αυτό ορίζει τα GAN morphs πιο δύσκολα στον εντοπισμό με βάση οπτικά κριτήρια.

Αντίθετα, τα GAN morphs είναι πιθανό να εισάγουν ιδιαιτερότητες artifacts του γεννητικού μοντέλου, όπως έλλειψη τυπικότητας στο φόντο ή στις δεδομένα των μαλλιών και των δοντιών. Αυτά τα “αποτυπώματα” του μοντέλου αποτελούν στόχο για τους ανιχνευτές που επικεντρώνονται σε στατιστικές ανωμαλίες της υφής, όπως θα εμφανιστούν και στα αποτελέσματα του πειράματος της παρούσας εργασίας.

2.4 Νέες Προσεγγίσεις: Diffusion Models

Την διετία (2023-2024) η βιβλιογραφία έχει ξεκινήσει να αναζητά τη εφαρμογή των μοντέλων διάχυσης (Denoising Diffusion Probabilistic Models - DDPMs) για την δημιουργία morphs. Τα DDPMs, που εντάχθηκαν από τους Ho, Jain και Abbeel το 2020 [4], στηρίζονται σε μια διαφορετική διαδικασία αποθορυβοποίησης η οποία σταδιακά μεταμορφώνει τον Gaussian θόρυβο σε αληθοφανής εικόνες.

Το έργο StableMorph (2024) και η σχετική μελέτη του Blasingame και της Liu (2024) [13] οδηγούν ότι τα diffusion-based morphs έχουν τη ικανότητα να υπερκαλύπτουν τόσο τα landmark όσο και τα GAN-based morphs σε όρους ρεαλισμού και συνεπώς απειλής. Η βασική τεχνική που εφαρμόζεται είναι η DDIM inversion σε σύνθεση με ομαλή μετάβαση στον χώρο του θορύβου (SLERP).

Το κύριο προνόμιο των μοντέλων διάχυσης έναντι των GANs είναι η σταθερή εκπαίδευση και ο περιορισμός της κατάστασης mode collapse, όπου η γεννήτρια ενός GAN δημιουργεί μικρή ποσότητα εξόδων. Τα μοντέλα διάχυσης, με βάση την ικανότητα τους στις πιθανότητες, εξυπηρετούν ιδανικότερα την καταμερισμό των δεδομένων, δημιουργώντας πιο πολλές και αληθοφανής εικόνες.

Όμως, τα μοντέλα διάχυσης διαθέτουν και ένα βασικό μειονέκτημα: η διαδικασία γέννησης είναι διαδοχική και χρειάζεται πολλαπλές διελεύσεις από το νευρωνικό δίκτυο (ουσιαστικά εκατοντάδες ή χιλιάδες βήματα στην αρχική μορφή DDPM). Η ενσωμάτωση των DDIM samplers ελάττωσε σημαντικά αυτό το κόστος, δημιουργώντας σημαντικά αποτελέσματα με μόλις 20-50 βήματα.

Λόγω του morphing, το μεγαλύτερο στοίχημα των diffusion μεθόδων είναι η συντήρηση της ταυτότητας και των δύο αρχικών προσώπων. Η απλή παρέμβαση στον χώρο του θορύβου δεν εξασφαλίζει ότι η κατάληξη θα παρουσιάζεται εξίσου και με τα δύο πρόσωπα, και αποτελεί ενεργό πεδίο έρευνας.

2.5 Μέθοδοι Ανίχνευσης (MAD)

Οι μέθοδοι ανίχνευσης διασπόνται σε δύο μεγάλες κατηγορίες: Single-image MAD (S-MAD), όπου αξιοποιείται μόνο η ύποπτη εικόνα, και Differential MAD (D-MAD), όπου επιπλέον παραχωρείται μια αξιόπιστη εικόνα αναφοράς του ατόμου. Η υπάρχουσα διατριβή βασίζεται στις S-MAD μεθόδους.

2.5.1 Μέθοδοι Χειροποίητων Χαρακτηριστικών

Πολλές βιαστικές τακτικές στηρίχθηκαν σε χειροποίητα χαρακτηριστικά εικόνας. Τα Local Binary Patterns (LBP) και οι εκδοχές τους, όπως τα Multi-Block LBP (MB-LBP), όπου χρησιμοποιούνται γενικά παράλληλα με ταξινομητές SVM (Scherhag et al., 2018 [8], 2019 [7]). Άλλες ιδιότητες περιέχουν τα Binarized Statistical Image Features (BSIF), τα Histograms of Oriented Gradients (HOG), και τα Local Phase Quantization (LPQ).

Ταυτόχρονα, έχουν δοκιμαστεί τεχνικές θεμελιωμένες στην ανάλυση του θορύβου του αισθητήρα της κάμερας (Photo Response Non-Uniformity - PRNU), οι οποίες βασίζονται στο γεγονός ότι η διαδικασία morphing παραμορφώνει την ιδιότητα του σχεδίου θορύβου των εικόνων που αναπτύχτηκαν στην αρχή .

2.5.2 Μέθοδοι Βαθιάς Μάθησης

Με την κυριαρχία της βαθιάς μάθησης, οι σύγχρονες μέθοδοι εφαρμόζουν convolutional neural networks (CNNs) ως feature extractors, ή επιπλέον και ως end-to-end ταξινομητές. Δίκτυα όπως το VGG, το ResNet και το EfficientNet έχουν χρησιμοποιηθεί για την εξαγωγή deep embedding στοιχείων, τα οποία στη συνέχεια εφοδιάζουν έναν ταξινομητή.

Τελευταίες μέθοδοι, όπως αυτή του MorDeephy (2023), εστιάζουν στη γενίκευση σε άγνωστα είδη απειλών μέσω συγκεκριμένων προγραμματισμένων αρχιτεκτονικών και loss functions.

2.6 Το Πρόβλημα της Γενίκευσης

Αν και υπήρξε βελτίωση, οι περισσότερες αναρτημένες μέθοδοι MAD εμφανίζουν σημαντική ικανότητα σε intra-dataset σενάρια, αλλά καταρρέουν σε cross-dataset ή cross-method αξιολογήσεις. Η εργασία των Scherhag et al. (2019) [7] στο IEEE Access εμφάνισε ότι ένας ανιχνευτής με Equal Error Rate (EER) 2% σε ένα συγκεκριμένο dataset έχει την ικανότητα να πάει στο 30% όταν χρησιμοποιείται σε ένα διαφορετικό.

Η ανάγκη για robust και transferable MAD συστήματα κατεύθυνε στη προετοιμασία των διαγωνισμών SYN-MAD 2022 [15] και FMADT 2024, οι οποίοι είχαν στόχο στη παραγωγή αντίστοιχων περιστάσεων αξιολόγησης. Η συγκεκριμένη διατριβή κατατάσσεται σε αυτό το πλαίσιο, εκτιμώντας τη γενίκευση ανάμεσα σε τέσσερεις διαφορετικούς τύπους morphs.

2.7 Ηθικές Πτυχές στην Έρευνα Morphing

Μια υπόθεση που έχει εμφανιστεί τώρα τελευταία στη βιβλιογραφία είναι το ηθικό πλαίσιο της δημιουργίας morphs. Η παραγωγή synthetic εικόνων με βάση πραγματικά πρόσωπα φέρνει σοβαρά ζητήματα προστασίας δεδομένων, κυρίως όταν τα δεδομένα πηγάζουν από ανηλίκους.

Ο Κανονισμός 2024/1689 [22] της Ευρωπαϊκής Ένωσης (EU AI Act) και οι σχετικές κατευθυντήριες οδηγίες του ENISA περιέχουν ειδικές αναφορές στη διαμόρφωση βιομετρικών χαρακτηριστικών , ενώ μια αναπτυσσόμενη μερίδα της ερευνητικής κοινότητας υποστηρίζει τη εφαρμογή αποκλειστικά synthetic datasets για την μάθηση MAD συστημάτων, όπως αποκαλύφθηκε στον διαγωνισμό SYN-MAD 2022 [15].

2.8 Σύνολα Δεδομένων στη Βιβλιογραφία

Η αξία και η διαφορετικότητα των συνόλων δεδομένων έχει σημαντικό ρόλο στην φερεγγυότητα των μελετών MAD. Στη πηγή έχουν χρησιμοποιηθεί ποικίλα datasets, με βασικότερα το FERET, το FRGC, το Utrecht ECVP και πιο νεότερα synthetic σύνολα όπως το SMDD (Synthetic Morphing Attack Detection Development).

Το FFHQ, το οποίο εξετάζεται σε αυτή την διατριβή, διαφέρει για την άριστη του ποιότητα και την μεγάλη επάρκεια του. Αν και στην αρχή το φτιάξανε για την εκπαίδευση γεννητικών μοντέλων, η ομοιογένεια της ευθυγράμμισης και η μεγάλη μελέτη του το καθιστούν ιδανικό και για δοκιμές morphing. Ένα θέμα που επιβάλλεται να επισημανθεί είναι ότι ο συσχετισμός μεταξύ μελετών δυσκολεύει από την απουσία κοινών benchmark datasets. Άλλες εργασίες μεταχειρίζονται ξεχωριστά σύνολα, με άλλες μεθόδους παραγωγής morphs, επιπλέον και άλλους ορισμούς των μετρικών, περιορίζοντας τη συγκρισιμότητα.

2.9 Τοποθέτηση της Παρούσας Εργασίας

Σύμφωνα με την γενική εικόνα, η συγκεκριμένη διατριβή βασίζεται σε δύο ερευνητικά ρεύματα : της δημιουργίας morphs με ξεχωριστές τεχνικές και της έρευνας γενίκευσης των ανιχνευτών. Η συνδρομή της συνιστάται στην συγχώνευση αυτών των δύο μεθόδων μέσα από ένα περιορισμένο πειραματικό περιβάλλον, όπου όλες οι παράμετροι πλην του τύπου morph παραμένουν σταθερές. Αυτή η επιτηρούμενη οργάνωση βοηθά τον περιορισμό της επιρροής του τύπου morph στην δυνατότητα ανίχνευσης, διαθέτοντας συμπεράσματα με μεγάλη ευκρίνεια σε σχέση με έρευνες που αλλάζουν ταυτόχρονα αρκετές παράγοντες.

ΚΕΦΑΛΑΙΟ 3: Θεωρητικό Υπόβαθρο

3.1 Βιομετρικά Συστήματα Αναγνώρισης Προσώπου

Ένα κλασσικό βιομετρικό σύστημα αναγνώρισης προσώπου (Face Recognition System - FRS) ενεργεί σε δύο στάδια: το στάδιο εγγραφής και το στάδιο επαλήθευσης. Κατά την διαδικασία εγγραφής, μια εικόνα αναφοράς του χρήστη υπάγεται σε διαμόρφωση, εντοπισμό προσώπου, εξαγωγή χαρακτηριστικών και τελικά αποθηκεύεται ένα συμπαγές διάνυσμα (template) στη βάση δεδομένων του συστήματος. Κατά την διαδικασία της εξακρίβωσης, η εικόνα δέχεται υπό εξέταση την ίδια διαδικασία και το template που δημιουργείται συγκρίνεται με το αποθηκευμένο. Η σύγκριση εμφανίζει ένα σκορ ομοιότητας (similarity score), το οποίο μπαίνει σε σύγκριση με ένα προκαθορισμένο όριο (threshold) ώστε να παρθεί η τελική απόφαση αποδοχής ή απόρριψης. Η βασική επισήμανση είναι ότι το σύστημα δεν βάζει απευθείας εικόνες σε σύγκριση, αλλά την απεικόνιση τους στον χώρο των στοιχείων. Αυτό σημαίνει ότι μια τροποποιημένη εικόνα που βρίσκεται "μεταξύ" δύο πραγματικών στον χώρο αναπαράστασης, θα έχει υψηλή συνάφεια και με τους δύο αρχικούς στόχους, ερμηνεύοντας έτσι γιατί η επίθεση πετυχαίνει.

3.2 Generative Adversarial Networks

Τα GANs καθιερώθηκαν από τον Goodfellow και τους βοηθούς του το 2014 [23] ως ένα νέο πεδίο γεννητικής μοντελοποίησης. Απαρτίζονται από δύο νευρωνικά δίκτυα που εκπαιδεύονται σε αντιπαλοτικό σχήμα (δύο νευρωνικά δίκτυα εκπαιδεύονται μέσα από μια διαδικασία διαρκούς ανταγωνισμού, λειτουργώντας ως «αντίπαλοι» σε ένα παίγνιο μηδενικού αθροίσματος) : μια γεννήτρια G η οποία θέλει να δημιουργήσει παραστατικές εικόνες, και έναν διακριτή D ο οποίος έχει την ικανότητα να διαχωρίσει τις γνήσιες εικόνες από τις δημιουργημένες.

Μαθηματικά, η εκπαίδευση διατυπώνεται ως ένα minimax παιχνίδι:

$$\min_G max_D V(D,G) = E_x[\log D(x)] + E_x[\log(1 - D(G(z)))]$$

όπου x αντιπροσωπεύει πραγματικά δεδομένα και z τυχαίο θόρυβο από κάποια προκαθορισμένη κατανομή. Στην πορεία της εκπαίδευσης, η G εκπαιδεύεται να δημιουργεί ολοένα πιο αληθοφανής εικόνες ώστε να ξεγελά τον D, ενώ ο D βελτιώνεται στο να καταλαβαίνει τις γνήσιες από τις κατασκευασμένες .

3.2.1 Η Αρχιτεκτονική StyleGAN2

Το StyleGAN, και η ανάπτυξη του StyleGAN2 (Karras et al., 2020) [2], καθιέρωσε βασικές καινοτομίες στην δομή των GANs που έφεραν σε άνευ προηγουμένου ποιότητα συνθετικών προσώπων σε ανάλυση 1024×1024 pixels. Η βασική εξέλιξη είναι η εισαγωγή ενός χώρου "styles" W, στον οποίο τροποποιείται ο αρχικός Gaussian θόρυβος Z μέσω ενός δικτύου mapping.

Επίσης, η δομή StyleGAN2-ADA (Karras et al., 2020) [3] εξέλιξε την σταθερότητα της εκπαίδευσης και συμπεριέλαβε μηχανισμό προσαρμοστικού discriminator augmentation, θέτοντας ισχυρή την μάθηση από datasets μικρής χωρητικότητας.

Στην τρέχουσα διατριβή εφαρμόζεται η pre-trained βερσιόν του StyleGAN2-ADA που εκπαιδεύτηκε στο FFHQ dataset. Η εμφάνιση μιας αληθινής εικόνας στον latent χώρο W+ χρειάζεται iterative optimization εμπνευσμένο σε perceptual loss (LPIPS), η οποία είναι υπολογιστικά απαιτητική.

Ο διαχωρισμός μεταξύ των χώρων Z, W και W+ είναι σημαντικός για την κατανόηση της λειτουργίας του StyleGAN2. Ο χώρος Z είναι ο πρωτεύων χώρος του Gaussian θορύβου. Με τη βοήθεια ενός δικτύου mapping οκτώ ολοκληρωτικά συνδεδεμένων στρωμάτων, ο Z τροποποιείται στον ενδιάμεσο χώρο W, ο οποίος είναι σε μικρότερο βαθμό “μπερδεμένος” (disentangled) και δημιουργεί πιο ισχυρό έλεγχο των ιδιοτήτων.

Ο χώρος W+ είναι προέκταση του W, όπου σε κάθε επίστρωση της γεννήτριας συσχετίζεται άλλο διάνυσμα w. Αυτή η επιπλέον ελευθερία δίνει την δυνατότητα την

ακριβέστερη ανασυγκρότηση γνήσιων εικόνων κατά την απεικόνιση, και για τον λόγο αυτό επιλέγεται στα λογισμικά morphing.

Η αρχιτεκτονική StyleGAN2 ενσωματώνει επίσης τη εφαρμογή weight demodulation αντί της αρχικής τεχνικής adaptive instance normalization (AdaIN) του StyleGAN, βελτιώνοντας τα στοιχεία artifacts τύπου “σταγόνας” (droplet artifacts) που παρουσιαζόντουσαν στην προηγούμενη έκδοση.

3.3 Diffusion Probabilistic Models

Τα Denoising Diffusion Probabilistic Models (DDPMs) είναι μια πιο φρέσκια οικογένεια γεννητικών μοντέλων που έχει εμφανιστεί ως επιβλητική εναλλακτική στα GANs. Η κεντρική ιδέα προέρχεται από τη θερμοδυναμική και την στατιστική φυσική.

3.3.1 Forward Diffusion Process

Στην ευθεία διαδικασία διάχυσης, μια καθαρή εικόνα x_0 μετατρέπεται σταδιακά σε θόρυβο μέσω ενός Markov chain T βημάτων. Σε κάθε βήμα t, προστίθεται μικρή ποσότητα Gaussian θορύβου σύμφωνα με τη σχέση:

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{(1-\beta_t)x_{t-1}}, \beta_t I)$$

όπου β_t είναι το variance schedule. Μετά από αρκετά βήματα (συνήθως T=1000), η x_T προσεγγίζει έναν καθαρό Gaussian θόρυβο.

3.3.2 Reverse Diffusion Process

Ο σκοπός είναι η εκπαίδευση της αντίστροφης διαδικασίας p_θ $(x_{t-1} \mid x_t)$, η οποία αρχίζει από καθαρό θόρυβο και σταδιακά μειώνει τον θόρυβο για να δημιουργήσει μια πραγματική εικόνα. Ένα νευρωνικό δίκτυο ε_θ $(x_t$, t) εκπαιδεύεται να μεριμνήσει τον θόρυβο που εντάχθηκε σε κάθε βήμα, μειώνοντας την απόσταση μεταξύ του πραγματικού και του προβλεπόμενου θορύβου.

Η επόμενη κίνηση στην ανάπτυξη των μοντέλων διάχυσης είναι τα Denoising Diffusion Implicit Models (DDIM) των Song, Meng και Ermon (2021) [5], τα οποία εγκρίνουν deterministic sampling σε λιγότερα βήματα, θέτοντας τη γέννηση πιο ικανοποιητική.

3.4 Local Binary Patterns και Παραλλαγές

Τα Local Binary Patterns (LBP), που εντάχθηκαν από τον Ojala και τους βοηθούς του (2002) [6], αποτελούν μια ισχυρή και προγραμματιστικά αποδοτική τεχνική περιγραφής υφής. Για κάθε pixel μιας εικόνας, ο LBP τελεστής εξετάζει την τιμή φωτεινότητάς του με τις τιμές των pixels γειτονικά (συνήθως P γειτόνων σε ακτίνα R), και δημιουργεί έναν δυαδικό κώδικα.

Η βερσιόν Uniform LBP μειώνει τους κώδικες σε αυτούς που διαθέτουν το πολύ δύο αλλαγές 0→1 ή 1→0, ελαττώνοντας τη διάσταση από 2^P σε P×(P-1)+3 patterns. Αυτό αναβαθμίζει τη γενίκευση και την υπολογιστική αποδοτικότητα.

Τα Multi-Block LBP (MB-LBP) διευρύνουν την σημασία θέτοντας σε εφαρμογή τον τελεστή όχι σε ξεχωριστά pixels αλλά σε ορθογώνια blocks της εικόνας. Αυτό προσφέρει robustness σε μικρές μεταθέσεις και τροποποιήσεις κλίμακας. Στην τρέχουσα διατριβή, εφαρμόζεται το Uniform LBP με R=3, P=24, που αναλογεί σε ένα histogram 26 bins.

Η εξήγηση για τη αξιοποίηση των LBP στον εντοπισμό morphing στηρίζεται στο γεγονός ότι η μέθοδος morphing — είτε μέσω alpha blending είτε μέσω νευρωνικών δικτύων — μεταβάλλει τα τοπικά μοτίβα υφής της εικόνας. Μια αυθεντική φωτογραφία περιέχει φυσικό θόρυβο και μικροϋφή που προέρχεται από τον αισθητήρα της κάμερας, ενώ μια morphed εικόνα εμφανίζει επίπεδες ή τεχνητές μικροϋφές.

Τα ιστογράμματα LBP απεικονίζουν αυτές τις διαφορές υφής σε μια πυκνή στατιστική αναπαράσταση, η οποία μετέπειτα έχει την ικανότητα να αποδώσει από έναν ταξινομητή. Παρά την λιτότητα τους, τα LBP έχουν καταδειχθεί εξαιρετικά αποδοτικά σε αφθονία εργασιών ανάλυσης υφής, συνυπολογιζόμενης της ανίχνευσης παρουσίασης .

3.5 Support Vector Machines

Οι Support Vector Machines (SVMs) είναι μια συνήθης αλλά υπερβολικά αποδοτική οικογένεια ταξινομητών. Αν θεωρήσουμε ένα σύνολο εκπαίδευτικών δεδομένων $\{(x_i, y_i)\}$ όπου $x_i \in R^n$και $y_i \in \{-1,+1\}$, ο SVM ζητά να βρει το υπερεπίπεδο που μεγιστοποιεί το περιθώριο (margin) μεταξύ των δύο κλάσεων.

Για μη-γραμμικά διαχωρίσιμα προβλήματα, χρησιμοποιείται ο μηχανισμός των kernels, με πιο συνηθισμένο το Radial Basis Function (RBF) kernel:

$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

Ο SVM είναι εξαιρετικά ιδανικός για ζητήματα δυαδικής ταξινόμησης με μεσαίο μέγεθος δεδομένων και υψηλό μέγεθος χαρακτηριστικών, όπως είναι η κατάσταση του εντοπισμού morphing.

Η μαθηματική διατύπωση του ζητήματος βελτιστοποίησης του SVM, στη soft-margin εκδοχή του, μειώνει τη συνάρτηση:

$$min_{w,b,\xi} \frac{1}{2}||w||^2 + c\sum_{i=1}^{N} \xi_i,$$

$$\text{υπό } y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

όπου w είναι το διάνυσμα βαρών, b η μετατόπιση, ξ_i οι μεταβλητές χαλάρωσης (slack variables) που επιτρέπουν συγκεκριμένα σφάλματα ταξινόμησης, και C η παράμετρος κανονικοποίησης που ελέγχει την μέση λύση μεταξύ του πλάτους του περιθωρίου και των σφαλμάτων εκπαίδευσης.

Η μεταβλητή γάμα του RBF kernel εξετάζει την “εμβέλεια” επίδρασης κάθε δείγματος εκπαίδευσης: μικρές τιμές γ σημαίνουν μεγάλη επιρροή και αβίαστα όρια απόφασης, ενώ μεγάλες τιμές κινούνται σε πιο πολύπλοκα, τοπικά όρια που μπορεί να δημιουργήσουν υπερπροσαρμογή. Σε αυτή την διατριβή εφαρμόστηκε η ρύθμιση gamma='scale', η οποία συμβαδίζει αυτόματα την τιμή με βάση τη διάδοση των ιδιοτήτων .

3.6 Μετρικές Αξιολόγησης MAD Συστημάτων

Η εξέταση ενός MAD συστήματος λειτουργεί με εξειδικευμένες μετρικές που έχουν καταξιωθεί στην βιομετρική κοινότητα. Οι βασικότερες είναι:

- APCER (Attack Presentation Classification Error Rate): το ποσοστό των morphed εικόνων που εσφαλμένα ταξινομούνται ως bonafide. Είναι η σπουδαιότερη μετρική από άποψη ασφαλείας, καθώς εξωτερικεύει την πιθανότητα μιας επίθεσης να περάσει απαρατήρητη.
- BPCER (Bonafide Presentation Classification Error Rate): το ποσοστό των γνήσιων εικόνων που εσφαλμένα απορρίπτονται ως morphed. Συνδέεται με την εμπειρία χρήστη: υψηλό BPCER σημαίνει ότι πολλοί νόμιμοι χρήστες θα αποκλείονται.
- Accuracy: το συνολικό ποσοστό σωστών ταξινομήσεων.
- AUC (Area Under the ROC Curve): μετρική που συνοψίζει την επίδοση του ταξινομητή σε όλα τα δυνατά thresholds, με τιμή 0.5 να αντιστοιχεί σε τυχαία ταξινόμηση και 1.0 σε τέλεια διαχωριστικότητα.
- EER (Equal Error Rate): το σημείο όπου APCER = BPCER, αποτελώντας μια ισορροπημένη μέτρηση της συνολικής επίδοσης.

3.6.1 Σχέση μεταξύ APCER και BPCER

Η σχέση μεταξύ APCER και BPCER είναι ο κεντρικός συμβιβασμός κάθε MAD συστήματος. Ελαττώνοντας το όριο απόφασης ώστε το σύστημα να γίνει πιο προσεκτικό, περιορίζεται το APCER (ελάχιστες επιθέσεις περνούν) αλλά μεγαλώνει το BPCER (περισσότεροι νόμιμοι

χρήστες απορρίπτονται). Αντίθετα, ένα πιο ελαστικό σύστημα λιγοστεύει την ταλαιπωρία των νόμιμων χρηστών αλλά επιτρέπει περισσότερες επιθέσεις να ξεφύγουν.

Σε προγράμματα υψηλής ασφαλείας, όπως η επιθεώρηση διαβατηρίων, παραχωρείται προτίμηση στη μείωση του APCER, επιπλέον και εις βάρος ενός υψηλότερου BPCER. Αυτό γίνεται επειδή τα αποτελέσματα μιας επιτυχημένης επίθεσης (πχ διέλευση εγκληματία) είναι πολύ σοβαρά από την καταπόνηση ενός νόμιμου χρήστη που πρέπει να του γίνει δεύτερος έλεγχος.

3.6.2 Η Καμπύλη ROC

Η καμπύλη Receiver Operating Characteristic (ROC) αναπαριστά την ροή αληθώς θετικών (true positive rate) ως προς την ροή ψευδώς θετικών (false positive rate), για όλους τους δυνατούς περιορισμούς απόφασης. Όσο πιο πλησίον στην επάνω αριστερή γωνία βρίσκεται η καμπύλη, τόσο ακριβέστερη η επίδοση του ταξινομητή. Το εμβαδόν κάτω από την καμπύλη (AUC) περιλαμβάνει την πρόοδο σε έναν μεμονωμένο αριθμό.

3.7 Το Flickr-Faces-HQ Dataset

Το FFHQ συνιστά ένα από τα πιο γνωστά σύνολα δεδομένων προσώπων στον επιστημονικό κόσμο. Φτιάχτηκε από την NVIDIA μαζεύοντας εικόνες από την εφαρμογή Flickr υπό άδειες Creative Commons, και υποβλήθηκε αυτοματοποιημένη ευθυγράμμιση και περιορισμό ώστε όλα τα πρόσωπα να είναι σε κανονικοποιημένη θέση.

Η ποικιλομορφία του dataset ως προς την ηλικία, την εθνικότητα και τις συνθήκες λήψης ορίζεται χαρακτηριστικό πραγματικών συνθηκών. Επίσης όμως, η ύπαρξη εικόνων ανηλίκων αναγκάζει την εφαρμογή κριτηρίων επιλογής, όπως αυτά που αναφέρονται στο Κεφάλαιο 4.

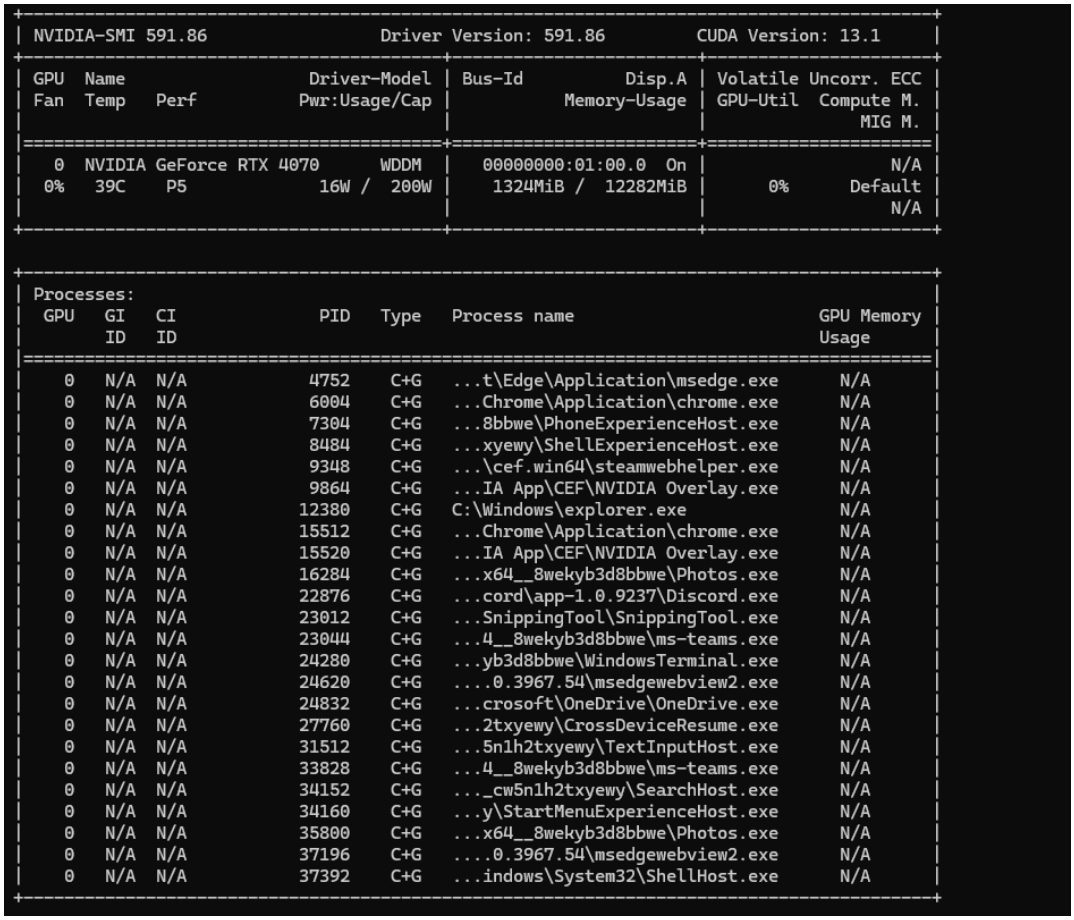
3.8 Σύνοψη Θεωρητικού Υποβάθρου

Στην ενότητα αυτή εμφανίστηκαν οι βασικές ιδέες που χρειάζονται για την αντίληψη της εργασίας: η χρήση των βιομετρικών συστημάτων, οι αρχιτεκτονικές των γεννητικών μοντέλων GAN και DDPM, οι τεχνικές περιγραφής υφής LBP, ο ταξινομητής SVM και οι δείκτες απόδοσης. Οι θεωρητικές αυτές αποδείξεις συγκροτούν το θεμέλιο πάνω στο οποίο οικοδομείται η προσέγγιση των επόμενων κεφαλαίων.

ΚΕΦΑΛΑΙΟ 4: Μεθοδολογία Παραγωγής Morphs

4.1 Επισκόπηση του Pipeline

Σε αυτή την ενότητα αναφέρεται λεπτομερώς η δράση που ακολουθήθηκε για την δημιουργία των morphed εικόνων. Αθροιστικά πραγματοποιήθηκαν τέσσερις ξεχωριστές διαδικασίες, οι οποίες διαλέχτηκαν ώστε να περιλαμβάνουν όλο το φάσμα από τις κλασσικές μέχρι τις πλέον μοντέρνες μεθόδους. Όλη η υλοποίηση εφαρμόστηκε τοπικά, χρησιμοποιώντας μια κάρτα γραφικών NVIDIA GeForce RTX 4070 με 12GB VRAM.



Σχήμα 4.1: Ενδεικτικό screenshot συστήματος ή GPU που χρησιμοποιήθηκε για τα πειράματα.

Πριν από οποιαδήποτε τεχνική morphing, το σύνολο των εικόνων πέρασαν από μια περίοδο προκαταρτικής επεξεργασίας. Η φάση αυτή εμπεριέχει την ανίχνευση του προσώπου, την ευθυγράμμιση (alignment) με βάση τη θέση των οφθαλμών, και τον περιορισμό σε κανονικοποιημένη ανάλυση. Το στάδιο της ευθυγράμμισης είναι βασικό, καθώς εγγυάται ότι τα αντίστοιχα μορφολογικά σημεία βρίσκονται σε ίδιες θέσεις σε όλες τις εικόνες.

4.1.1 Το Σύνολο Δεδομένων Βάσης

Για το αρχικό σημείο δημιουργίας των morphs χρησιμοποιήθηκε το Flickr-Faces-HQ (FFHQ) dataset, το οποίο κατασκευαστικέ από την NVIDIA για την μάθηση του StyleGAN [12]. Το FFHQ περιλαμβάνει 70.000 ποιοτικά υψηλές εικόνες προσώπων σε ανάλυση 1024×1024, με τεράστια συλλογή ως προς την ηλικία, την εθνικότητα, τις εκφράσεις και τις συνθήκες φωτισμού.

Για τις απαιτήσεις της διατριβής, προτιμήθηκε ένα υποσύνολο 1.000 εικόνων, οι οποίες εφαρμόστηκαν τόσο ως αυθεντικές εικόνες αναφοράς (bonafide) όσο και ως πηγή για την δημιουργία των morphs.

Σχήμα 4.2: Παραδείγματα aligned FFHQ εικόνων που χρησιμοποιήθηκαν ως bona fide δείγματα.

4.2 Μέθοδος 1: Landmark-Based Morphing

Η πρωτοεμφανιζόμενη και πιο κλασσική μέθοδος που εφαρμόστηκε στηρίζεται στον ανεύρεση σημείων αναφοράς του προσώπου. Για αυτόν το στόχο εφαρμόστηκε ο προ-εκπαιδευμένος ανιχνευτής 68 σημείων της βιβλιοθήκης dlib [17], ο οποίος ανιχνεύει χαρακτηριστικά σημεία στην περίμετρο του προσώπου, τα φρύδια, τη μύτη, τα μάτια και το στόμα.

4.2.1 Delaunay Triangulation και Warping

Στη περίπτωση που βρεθούν τα σημεία αναφοράς και των δύο εικόνων, υπολογίζεται ο μέσος όρος των θέσεών τους. Επίσης, γίνεται χρήση τριγωνοποίησης Delaunay πάνω στο σύνολο των μέσων σημείων, διαιρώντας το πρόσωπο σε ένα πλέγμα τριγώνων. Οποιοδήποτε τρίγωνο από την αρχική εικόνα επιδέχεται έναν συγγενή μετασχηματισμό ώστε να ενωθεί στο αντίστοιχο τρίγωνο της μέσης γεωμετρίας.

Η τριγωνοποίηση Delaunay επιλέγεται διότι πληθαίνει την ελάχιστη γωνία των τριγώνων, απορρίπτοντας έτσι τη αξιοποίηση πολύ “λεπτών” τριγώνων που θα οδηγούσαν σε ανισότητες κατά την μετατροπή. Στο πρόγραμμα, στο σύνολο των 68 σημείων ενσωματώνονται και τα τέσσερα σημεία των γωνιών της εικόνας επίσης και μεσαία σημεία στις ακμές, με αποτέλεσμα να καλυφτεί ολόκληρη η επιφάνεια.

Έπειτα από τον γεωμετρικό μετασχηματισμό (warping) και των δύο εικόνων στην κοινή μέση γεωμετρία, η τελική morphed εικόνα κατασκευάζεται από τον συνδυασμό γραμμικά των χρωματικών τιμών:

$M(x, y) = \alpha \cdot I_1'(x, y) + (1 - \alpha) \cdot I_2'(x, y)$

όπου I_1' και I_2' είναι οι warped εικόνες, και α ο συντελεστής βάρους, ο οποίος για ισορροπημένα morphs ορίζεται ίσος με 0,5. Η επιλογή α=0,5 επιβεβαιώνει ότι το morph εντοπίζεται ακριβώς

στο μέσο μεταξύ των δύο ταυτοτήτων, μεγαλώνοντας έτσι την πιθανότητα αποτελεσματικής επαλήθευσης και για τους δύο υποψηφίους.

Σχήμα 4.3: Παραδείγματα naive landmark morphs με ορατά artifacts.

Η διαδικασία αυτή δημιούργησε συνολικά 484 morphs. Κατά το αναμενόμενο, στα σημεία εκτός του κύριου προσώπου (μαλλιά, φόντο) αναδείχτηκαν ορατά artifacts σαν ghosting, επειδή ο γραμμικός συνδυασμός σε αυτά τα τμήματα δεν λαμβάνει υπόψη τις ουσιαστικές διαφορές ανάμεσα στις δύο εικόνες.

4.2.2 Βελτίωση με Poisson Blending

Για την ελάττωση των artifacts στις τοπικές περιοχές, δημιουργήθηκε μια δεύτερη, καλύτερη τοποθέτηση της landmark διαδικασίας. Σε αυτήν, το morphed πρόσωπο συνενώνεται με το αρχικό υπόβαθρο μίας εκ των δύο εικόνων μέσω της τεχνικής Poisson blending (seamless cloning).

Η μέθοδος αυτή, αντί να αντιγράφει κατευθειαν τις τιμές των pixels, λύνει μια εξίσωση Poisson ώστε να παραμείνει η κλίση (gradient) του morphed προσώπου ενώ ταυτόχρονα εναρμονίζεται με τις οριακές τιμές του υποβάθρου. Το επακόλουθο είναι μια πιο φυσική αλλαγή χωρίς ορατές ραφές. Με αυτή τη διαδικασία δημιουργήθηκαν επιπλέον 484 masked morphs.

Σχήμα 4.4: Παραδείγματα masked morphs με Poisson blending.

4.3 Μέθοδος 2: StyleGAN2-Based Morphing

Η ακόλουθη οικογένεια μεθόδων εκμεταλλεύεται τη γεννητική ισχύ του StyleGAN2. Σε σχέση με τη landmark μέθοδο που ενεργεί απευθείας στον χώρο των pixels, εδώ η μίξη γίνεται στον λανθάνοντα χώρο (latent space) του δικτύου.

4.3.1 Projection στον Latent Χώρο

Το αρχικό και πιο δύσκολο βήμα είναι η εμφάνιση κάθε πραγματικής εικόνας στον χώρο W+ του StyleGAN2. Αυτό πραγματοποιείται μέσω μιας επαναληπτικής μεθόδους βελτιστοποίησης, κατά την οποία ερευνάται το latent διάνυσμα w το οποίο, όταν επεξεργαστεί από τη γεννήτρια, δημιουργεί εικόνα κατά προσέγγιση κοντά στην αρχική.

Η συνάρτηση κόστους που μειώνεται συνδέει ένα perceptual loss (LPIPS), το οποίο μετρά την αντιληπτική ομοιότητα, με ένα όρο κανονικοποίησης που κρατά το latent διάνυσμα κοντά στην κατανομή του χώρου W.

Παρατηρείται ότι ιδιαίτερα σε αυτό το υλικό, η διαδικασία προβολής εφαρμόστηκε σε λειτουργία fallback (χωρίς τους custom CUDA kernels του StyleGAN2), λόγω ανομοιότητας μεταξύ της έκδοσης του compiler και του CUDA toolkit. Αυτό είχε ως επακόλουθο η προβολή να έχει καθυστέρηση (η εικόνα κοντά 20 λεπτά), χωρίς όμως να αλλάζει η ποιότητα του οριστικού αποτελέσματος.

Σχήμα 4.5: Παραδείγματα αρχικών εικόνων και reconstructed εικόνων από StyleGAN2 projection.

Ολικώς, η εμφάνιση 50 εικόνων επέβαλε προσεγγιστικά 21 ώρες αδιάκοπου υπολογισμού. Παρόλο όλη αυτή επιβάρυνση χρονικά, η αρτιότητα των ανακατασκευασμένων εικόνων ήταν τρομερή, με τις προβαλλόμενες εικόνες να είναι στην πράξη ασαφείς από τις αυθεντικές. Αυτό επιβεβαιώνει ότι ο μηχανισμός fallback, αν και αργός, πραγματοποιεί ολόιδιους υπολογισμούς με την αναβαθμισμένη εκδοχή.

Κάθε απεικόνιση αποθηκεύτηκε ως διάνυσμα latent σε μορφή συμπιεσμένου αρχείου NumPy (.npz), ώστε να είναι άμεσα εκμεταλλεύσιμη για το στάδιο του morphing χωρίς ανάγκη της δαπανηρής διαδικασίας προβολής προς επανάληψη.

4.3.2 Latent Space Interpolation

Εφόσον δύο εικόνες έχουν μεταδοθεί στα αντίστοιχα latent διανύσματα w_1 και w_2, η δημιουργία του morph προκαλείται μέσω παρεμβολής. Αντί για απλή γραμμική παρεμβολή, εφαρμόστηκε σφαιρική γραμμική παρεμβολή, η οποία συμμορφώνεται με την γεωμετρία του latent χώρου:

$$SLERP_{(w1,w2,t)} = \frac{\sin((1-t)\Omega)}{\sin\Omega} w1 + \frac{\sin(t\Omega)}{\sin\Omega} w2$$

όπου Ω είναι η γωνία μεταξύ των δύο διανυσμάτων. Το προκύπτον ενδιάμεσο διάνυσμα υποστηρίζεται στη γεννήτρια, δημιουργώντας το τελικό morph. Η διαδικασία αυτή δημιούργησε 200 morphs εξαιρετικής ποιότητας, χωρίς τα εμφανίσιμα artifacts των landmark μεθόδων.

Σχήμα 4.6: Παραδείγματα GAN morphs από το φάκελο 03_stylegan2/morphs_output.

4.4 Μέθοδος 3: Diffusion-Inspired Morphing

Η τρίτη διαδικασία είναι μια εμπνευσμένη με βάση τα μοντέλα διάχυσης, η οποία δημιουργήθηκε χωρίς την εφαρμογή εξωτερικού προ-εκπαιδευμένου μοντέλου. Σκοπός ήταν η έρευνα της συμπεριφοράς των στατιστικών ιδιοτήτων που εισάγει μια μέθοδος βασισμένη στον θόρυβο.

Ο τρόπος στηρίζεται στην συμπλήρωση ελεγχόμενου Gaussian θορύβου στις δύο εικόνες, τηρούμενη από μια σύνθεση στον χώρο του θορύβου και μια μέθοδο εξομάλυνσης. Με αυτή τη διεργασία δημιουργήθηκαν 176 morphs, τα οποία δοκιμάστηκαν από μέρος φιλτραρίσματος ηλικίας.

Σχήμα 4.7: Παραδείγματα diffusion-inspired morphs από το dataset/morphed_diffusion.

4.5 Μέθοδος 4: DDPM-Based Morphing

Η τέταρτη και πιο μοντέρνα μέθοδος εκμεταλλεύεται ένα ουσιαστικό προ-εκπαιδευμένο μοντέλο διάχυσης. Ειδικότερα, αξιοποιήθηκε το μοντέλο google/ddpm-celebahq-256, το οποίο έχει εξασκηθεί στο CelebA-HQ dataset και δημιουργεί πρόσωπα σε ανάλυση 256×256.

4.5.1 Partial Noise Injection και Interpolation

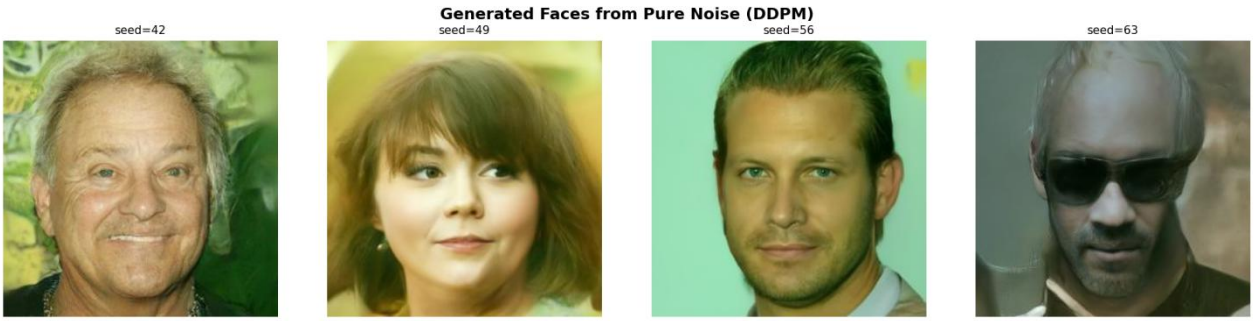
Ξεκινώντας επιχειρήθηκε η εκτίμηση της πλήρους DDIM inversion, κατά την οποία κάθε εικόνα κρυπτογραφείται πλήρως σε αναπαράσταση θορύβου. Όμως, αυτή η διαδικασία τεκμηριώθηκε απρόβλεπτη για το συγκεκριμένο unconditional μοντέλο, δημιουργώντας μη ρεαλιστικά αποτελέσματα όταν πραγματοποιούταν σε εικόνες FFHQ, οι οποίες πηγάζουν από διαφορετική διανομή σε σχέση με το CelebA-HQ.

Εξαιτίας αυτού, εφαρμόστηκε μια μέθοδο περιορισμένης έγχυσης θορύβου (partial noise injection). Στην εν λόγω, αντί για πλήρη inversion, οι δύο εικόνες εκτίθονται σε ευθεία διάχυση έως μια ενδιάμεση χρονική στιγμή (επιλέχθηκε t=700 από τα 1000). Στο ημιθορυβώδες αυτό επίπεδο δημιουργήθηκε γραμμικός συνδυασμός των δύο απεικονίσεων, και έπεται διαφορετική διάχυση από το βήμα αυτό ώστε να φτιαχτεί το τελευταίο morph.

Η μέθοδος αυτή φάνηκε σημαντικά σταθερή και δημιούργησε 100 morphs διακριτής ποιότητας. Το προτέρημα του μοντέλου διάχυσης είναι η ταχύτητα κατά την παραγωγή (περίπου 5 δευτερόλεπτα ανά morph με βάση αυτή την κάρτα γραφικών), λόγω της χρήσης του DDIM sampler που χρειάζεται μόνο 50 βήματα.

Η προτίμηση του χρονικού βήματος t=700 εμφανίστηκε μετά από πειραματισμό. Πιο μικρές τιμές κρατούσαν κατά πολύ τη δομή της μίας εκ των δύο αρχικών εικόνων, δημιουργώντας ασθενή morphs. Μεγαλύτερες τιμές πρόσθεταν ισχυρό θόρυβο, κατευθύνοντας σε αστάθεια και απώλεια ταυτότητας. Η τιμή t=700 είναι το ιδανικό σημείο ισορροπίας ανάμεσα στην συνδυασμό ταυτοτήτων και συντήρηση ρεαλισμού.

Ένα εξειδικευμένο μέρος που άξιζε προσοχής ήταν η εφαρμογή κοινού δείγματος θορύβου και για τις δύο εικόνες κατά την ευθεία διάχυση. Αυτή η επιλογή κατοχυρώνει ότι η διαφορά μεταξύ των δύο ημιθορυβωδών αναπαραστάσεων πηγάζει μόνο στην ουσία των εικόνων και όχι σε τυχαία διαφορά θορύβου, κατευθύνοντας σε πιο περιεκτικά αποτελέσματα κατά τη γραμμική παρεμβολή.



Σχήμα 4.8: Οπτικοποίηση σταδιακής παραγωγής εικόνας από DDPM.

Αξίζει να επισημανθεί ότι οι πρώτες δοκιμές με γραμμική παρεμβολή στον πλήρως αντεστραμμένο χώρο θορύβου (full DDIM inversion με SLERP) ήταν ανεπιτυχής, παράγοντας πολύχρωμα, ανέφικτα αποτελέσματα. Αυτό ερμηνεύτηκε στην έλλειψη σταθερότητας της

αντιστροφής για ένα unconditional μοντέλο εκπαιδευμένο σε ξεχωριστή κατανομή (CelebA-HQ) από αυτή των εικόνων εισόδου (FFHQ), και κατέληξε στην υιοθέτηση της πιο σταθερής μεθόδου μερικής έγχυσης θορύβου.

Σχήμα 4.9: Παραδείγματα DDPM morphs από το dataset/morphed_ddpm_real.

4.6 Ηθικό Φιλτράρισμα Ηλικίας

Ένα αρκετά βασικό μέρος της διαδικασίας, το οποίο συχνά δεν γίνεται στη βιβλιογραφία, είναι η πρόληψη ότι κανένα morph δεν δημιουργείται από εικόνες ανηλίκων. Το FFHQ, ως dataset συγκεντρωμένο από το Flickr, περιέχει και εικόνες παιδιών.

Για την επίλυση αυτού του θέματος, πραγματοποιήθηκαν δύο σταδία σε μια διαδικασία. Στο αρχικό στάδιο, για τα StyleGAN2 morphs, δημιουργήθηκε χειροκίνητη επιθεώρηση (manual review) των εικόνων που είχαν εμφανιστεί, με συνέπεια από τις 50 αρχικές να μείνουν 42 ως αποδεδειγμένα ενήλικες.

Στο επόμενο στάδιο, για τα diffusion morphs, εφαρμόστηκε αυτόματη εκτίμηση ηλικίας μέσω της βιβλιοθήκης DeepFace [21]. Κάθε εικόνα της οποίας η πιθανή ηλικία ήταν μικρότερη από 20 έτη απομακρύνθηκε από τη ανάμειξη της σε morph. Η αντιδραστική αυτή πολιτική (όριο 20 αντί 18) διαλέχτηκε ως έξτρα περιθώριο ασφαλείας με βάση τα σφάλματα του εκτιμητή ηλικίας.

Η μέθοδος αυτή συμβαδίζει με τις αρχές του Κανονισμού 2024/1689 [22] της ΕΕ και με τη γενικότερη τάση της ερευνητικής κοινότητας για υπεύθυνη χρήση βιομετρικών δεδομένων.

Σχήμα 4.10: Ενδεικτικό αποτέλεσμα ή log από το script 16c_auto_filter_diffusion.py.

4.7 Σύνοψη του Παραχθέντος Dataset

Συνολικά, το pipeline δημιούργησε ένα ολοκληρωμένο dataset το οποίο συλλέγεται στον επόμενο πίνακα. Το σύνολο περιέχει 1.000 αυθεντικές εικόνες και 1.444 morphs χωρισμένα σε τέσσερις κατηγορίες.

Τύπος Εικόνας	Μέθοδος	Πλήθος
Bonafide (αυθεντικές)	FFHQ	1.000
Landmark naive morphs	Delaunay	484
Landmark masked morphs	Poisson	484
StyleGAN2 morphs	SLERP	200
Diffusion-inspired morphs	Noise mix	176
DDPM morphs	Partial DDIM	100
ΣΥΝΟΛΟ MORPHS	**—**	**1.444**

Πίνακας 4.1: Σύνθεση του παραχθέντος συνόλου δεδομένων.

ΚΕΦΑΛΑΙΟ 5: Σύστημα Ανίχνευσης

5.1 Αρχιτεκτονική του Συστήματος MAD

Η μέθοδος ανίχνευσης που δημιουργήθηκε ακολουθεί την κοινή τακτική two-stage: ένα στάδιο εξαγωγής στοιχείων συνοδευόμενο από ένα στάδιο ταξινόμησης (classification). Η προτίμηση αυτής της αρχιτεκτονικής, αντί ενός end-to-end deep learning μοντέλου, έγινε για να εκφραστεί και να υπολογιστεί η απόδοση, καθώς και για να είναι άμεσα ισάξια με την κλασική βιβλιογραφία MAD.

Ως παράμετρος ιδιοτήτων επιλέχθηκαν τα Multi-Block Local Binary Patterns (MB-LBP), τα οποία αποδείχτηκαν ιδιαίτερα δραστικά στον εντοπισμό των διακριτικών μοτίβων υφής που ενσωματώνει η μέθοδος morphing. Ως ταξινομητής έγινε χρήση ένα Support Vector Machine με RBF kernel.

5.2 Προεπεξεργασία και Περιοχή Ενδιαφέροντος

Πριν την εξαγωγή στοιχείων, κάθε εικόνα αλλάζει σε grayscale. Έπειτα, περιθωριοποιείται η βασική περιοχή του προσώπου μέσω περικοπής (crop), ώστε να απομονωθούν περιμετρικές περιοχές όπως τα μαλλιά και το φόντο, οι οποίες πιθανολογείται να τοποθετούν θόρυβο στην ανάλυση.

Ιδιαίτερα, πραγματοποιήθηκε αρχική περικοπή (crop) στις σταθερές συντεταγμένες [280:800, 250:770] της αρχικής εικόνας, απομονώνοντας την κεντρική περιοχή του προσώπου σε διαστάσεις 520×520 pixels. Στη συνέχεια, εφαρμόστηκε αλλαγή μεγέθους (resize) με τη χρήση της μεθόδου παρεμβολής cv2.INTER_AREA, με συνέπεια η τελική περιοχή ενδιαφέροντος να κανονικοποιηθεί ομοιόμορφα στις διαστάσεις 300×300 pixels. Αυτή η κανονικοποίηση εγγυάται ότι όλες οι εικόνες συμβάλλουν χαρακτηριστικά από αντίστοιχες ανατομικές περιοχές.

5.3 Εξαγωγή Χαρακτηριστικών MB-LBP

Η εξαγωγή των MB-LBP χαρακτηριστικών υλοποιήθηκε με τη εφαρμογή της υλοποίησης uniform LBP της βιβλιοθήκης scikit-image. Τα κριτήρια που διαλέχτηκαν ήταν ακτίνα R=3 και αριθμός γειτόνων P=24, με τη διαδικασία uniform.

Η προτίμηση uniform patterns οριοθετεί το ιστόγραμμα σε 26 bins, διαθέτοντας μια ομοιογενή αλλά ήπια αναπαράσταση. Το τελικό feature vector κάθε απεικόνισης απαρτίζεται από το κανονικοποιημένο ιστόγραμμα των uniform patterns, διάστασης 26.

Η μικρή διάσταση των στοιχείων (26 τιμές ανά απεικόνιση) συνιστά προτέρημα ως προς την ταχύτητα εκπαίδευσης και την αποφυγή υπερπροσαρμογής (overfitting), αν και μπορεί να απομονώνει την ικανότητα του συστήματος να απεικονίσει πιο σύνθετα μοτίβα.

Σχήμα 4.11: Ενδεικτικό output από την εξαγωγή MB-LBP χαρακτηριστικών.

5.4 Ταξινόμηση με Support Vector Machine

Τα στοιχεία που εξάγονται, εξυπηρετούν έναν ταξινομητή SVM. Πριν την εκπαίδευση, τα στοιχεία κανονικοποιούνται μέσω StandardScaler, ώστε κάθε διάσταση να έχει μηδενική μέση τιμή και μοναδιαία διασπορά. Αυτό είναι σημαντικό για την ακριβή ακολουθία του RBF kernel.

Οι υπερπαράμετροι του SVM ορίστηκαν ως C=1.0 και gamma='scale', τιμές που είναι ένα ισορροπημένο ξεκίνημα. Το μοντέλο μαθαίνει ώστε να διακρίνει δύο κλάσεις: bonafide (κλάση 0) και morph (κλάση 1).

Σχήμα 4.12: Ενδεικτικό output από την εκπαίδευση και αποθήκευση SVM models.

5.5 Στρατηγική Cross-Evaluation

Το πιο πρωτοποριακό τμήμα της δοκιμαστικής διαδικασίας είναι η στρατηγική cross-evaluation. Αντί να μάθει ένα μόνο μοντέλο σε όλα τα είδη morph, εκπαιδεύτηκαν πέντε διαφορετικά μοντέλα SVM:

- SVM-LM: εκπαιδευμένο αποκλειστικά σε landmark morphs.
- SVM-GAN: εκπαιδευμένο αποκλειστικά σε StyleGAN2 morphs.
- SVM-DIFF: εκπαιδευμένο αποκλειστικά σε diffusion-inspired morphs.
- SVM-DDPM: εκπαιδευμένο αποκλειστικά σε DDPM morphs.
- SVM-MIX: εκπαιδευμένο σε συνδυασμό όλων των τύπων.

Κάθε ένα από αυτά τα μοντέλα κρίθηκε όχι μόνο στον τύπο morph στον οποίο μορφωθηκε (in-distribution testing), αλλά και σε όλους τους άλλους τύπους (cross-domain

testing). Έτσι, δημιουργείται ένας πίνακας 5×4 ο οποίος εμφανίζει την δυνατότητα γενίκευσης κάθε μοντέλου.

Αυτός ο δοκιμαστικός σχεδιασμός επικυρώνει την απάντηση σε ένα βασικό ερώτημα: είναι ικανός ένας εκπαιδευμένος ανιχνευτής σε έναν τύπο επίθεσης να προστατεύσει δραστικά απέναντι σε έναν ξένο, διαφορετικό τύπο; Η απάντηση, όπως θα παρουσιαστεί στο επόμενο κεφάλαιο, έχει βασικά πρακτικά αποτελέσματα για τον προγραμματισμό ουσιαστικών συστημάτων ασφαλείας.

5.6 Υλοποίηση και Εργαλεία

Όλη η μεθοδολογία εντοπισμού δημιουργήθηκε σε Python 3.9. Οι βασικές βιβλιοθήκες που εφαρμόστηκαν ήταν η scikit-image για την άντληση των LBP ιδιοτήτων, η scikit-learn για την εκπαίδευση και αξιολόγηση του SVM [20], η OpenCV [18] για την διαμόρφωση εικόνων, και η DeepFace [21] για την υπολογισμό των ετών των ανθρώπων στο στάδιο του φιλτραρίσματος.Τα εκπαιδευμένα μοντέλα αποθηκεύτηκαν σε μορφή joblib για επαναχρησιμοποίηση, και τα επακόλουθα οπτικοποιήθηκαν μέσω των βιβλιοθηκών matplotlib και seaborn.

5.7 Σκεπτικό Σχεδιασμού

Η αρχιτεκτονική του συστήματος αποτυπώνει μια σκόπιμη επιλογή υπέρ της απλότητας και της ερμηνευσιμότητας. Καθώς και τα σύγχρονα deep learning συστήματα πιθανολογείται να φτάσουν υψηλότερες επιδόσεις, η εφαρμογή χειροποίητων ιδιοτήτων MB-LBP σε ανάμειξη με SVM παρέχει μια σαφή και ξεκάθαρη σχέση ανάμεσα στην είσοδο και έξοδο. Αυτή η σαφήνεια είναι εξαιρετικά ανεκτίμητη με αφορμή τη συγκριτική μελέτη γενίκευσης, διότι επιβάλλει την απόδοση των παρατηρούμενων διαφορών επίδοσης στις ιδιότητες των ίδιων των morphs, και όχι σε σύνθετες, δύσκολα κατανοητές διαδράσεις ενός ουσιαστικού νευρωνικού δικτύου. Στην ουσία, η έλλειψη πολυπλοκότητας του ταξινομητή δουλεύει ως ένα “καθαρό” εργαλείο μέτρησης των διαφορών μεταξύ των τύπων morph. Επίσης, το μικρό κόστος υπολογισμού άφησε την επιμόρφωση και έλεγχος πέντε διαφορετικών μοντέλων με αρκετές διασταυρώσεις σε μικρό χρόνο, απλοποιωντας την αναλυτική πειραματική ανίχνευση.

ΚΕΦΑΛΑΙΟ 6: Πειραματική Αξιολόγηση

6.1 Πειραματική Διάταξη

Στην ενότητα αυτή εμφανίζονται και ερευνούνται τα επακόλουθα της πειραματικής αξιολόγησης. Όλες οι δοκιμές πραγματοποιήθηκαν στο ίδιο υλικό (NVIDIA RTX 4070) και με την ίδια τυχαιότητα (random seed = 42), ώστε να επιτευχθεί η δυνατότητα αναπαραγωγής των αποτελεσμάτων.

Για κάθε μοντέλο ξεχωριστά, η συλλογή δεδομένων χωρίστηκε σε 70% εκπαίδευση και 30% έλεγχο, με στρωματοποιημένη δειγματοληψία (stratified sampling) ώστε να παραμείνει η αναλογία των κλάσεων. Οι αυθεντικές απεικονίσεις (bonafide) ήταν συνηθισμένες για όλα τα πειράματα, ενώ άλλαζε μόνο το είδος των morphs που εφαρμόστηκε για την εκπαίδευση.

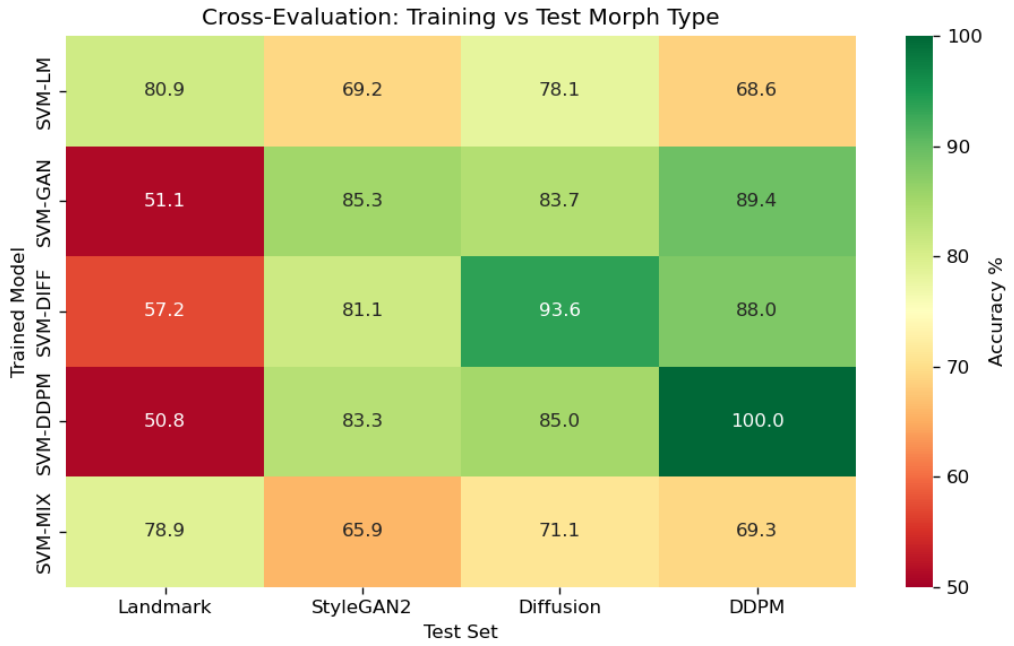
6.2 Ο Πίνακας Cross-Evaluation

Το βασικό αποτέλεσμα της διατριβής εξάγεται στον πίνακα cross-evaluation. Κάθε γραμμή αναλογεί σε ένα εκπαιδευμένο μοντέλο, και κάθε στήλη σε μια κατηγορία morph που εφαρμόστηκε για την εξέταση. Τα διαγώνια δεδομένα εκφράζουν in-distribution επίδοση, ενώ τα άλλα την cross-domain γενίκευση.

Train / Test	Landmark	StyleGAN2	Diffusion	DDPM
SVM-LM	80,9%	69,2%	78,1%	68,6%
SVM-GAN	51,1%	85,3%	83,7%	89,4%
SVM-DIFF	57,2%	81,1%	93,6%	88,0%
SVM-DDPM	50,8%	83,3%	85,0%	100,0%
SVM-MIX	78,9%	65,9%	71,1%	69,3%

Πίνακας 6.1: Πίνακας cross-evaluation (Accuracy %). Πράσινο: in-distribution. Κόκκινο: αποτυχία γενίκευσης.

Εκτός από το πινακάκι, φτιάχθηκε και ένα heatmap το οποίο εμφανίζει τις ίδιες τιμές με κωδικοποίηση με βάση το χρώμα, βοηθώντας να γίνουν πιο κατανοητά τα μοτίβα γενίκευσης.



Σχήμα 6.1: Heatmap του cross-domain evaluation από το φάκελο 04_detection/results.

6.3 Ανάλυση In-Distribution Επίδοσης

Αναλύοντας τα διαγώνια δεδομένα του πίνακα, καταλαβαίνουμε ότι όλα τα μοντέλα έχουν ικανοποιητική έως αρκετά καλό αποτέλεσμα όταν αξιολογούνται στον τύπο morph για τον οποίο προσαρμόστηκαν. Το SVM-DIFF φτάνει το 93,6% και το SVM-DDPM το 100%, ενώ επίσης και το λιγότερο αποτελεσματικό SVM-LM επιτυγχάνει 80,9%.

Η κορυφαία απόδοση του SVM-DDPM (100%) χρήζει προσοχής. Διότι το σύνολο των DDPM morphs ήταν αρκετά μικρό (100 εικόνες, εκ των οποίων μόλις 30 στο σύνολο ελέγχου), η τιμή αυτή λογικά υπερβάλει για την ουσιαστική δυνατότητα του μοντέλου. Αφορά ότι η ένδειξη ότι τα DDPM morphs διαθέτουν σημαντικά διακριτικά χαρακτηριστικά τα οποία είναι εύκολα διακριτά από τις πραγματικές απεικονίσεις με το συγκεκριμένο, μικρό test set.

6.4 Ανάλυση Cross-Domain Γενίκευσης

Το πιο βασικό εύρημα δημιουργείται από τα εκτός διαγωνίου στοιχεία. Διακρίνονται δύο ευθύς μοτίβα τα οποία έχουν άμεσα πρακτικά αποτελέσματα.

6.4.1 Αποτυχία Γενίκευσης από Γεννητικά σε Landmark

Τα μοντέλα που προσαρμόστηκαν σε γεννητικά morphs πέφτουν έξω δραματικά όταν εξετάζονται σε landmark morphs. Ειδικά, το SVM-GAN πραγματοποιεί μόλις 51,1% και το SVM-DDPM 50,8% στα landmark morphs. Τα αποτελέσματα αυτά είναι ίδια με τυχαία ταξινόμηση (50%), δείχνοντας ότι τα μοντέλα αυτά δεν έχουν μάθει απολύτως τίποτα βασικό για τον εντοπισμό landmark morphs.

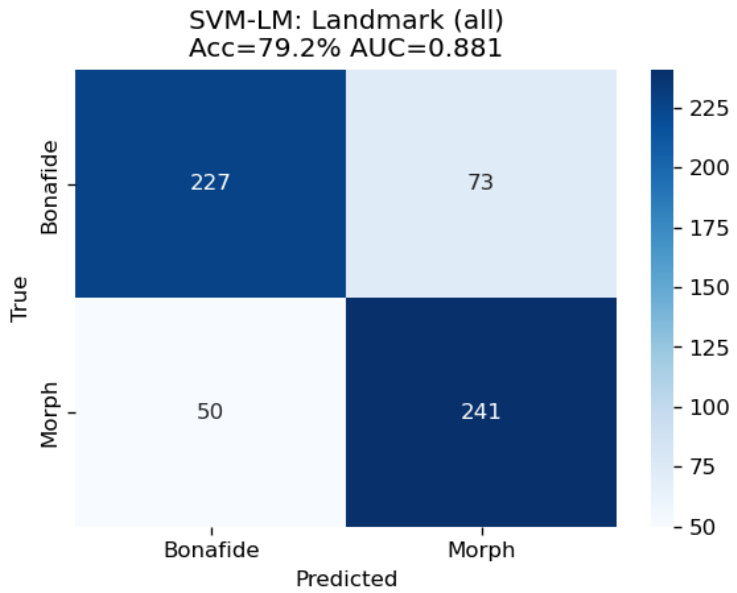
Η ανάλυση είναι ότι τα landmark morphs βάζουν artifacts γεωμετρικής φύσης (ghosting, ασυνέχειες στις ραφές) τα οποία έχουν ουσιώδη διαφορά από τους συσχετισμούς που εισέρχονται από τα γεννητικά μοντέλα στον χώρο της υφής.

6.4.2 Αμοιβαία Μεταφερσιμότητα μεταξύ Γεννητικών Μεθόδων

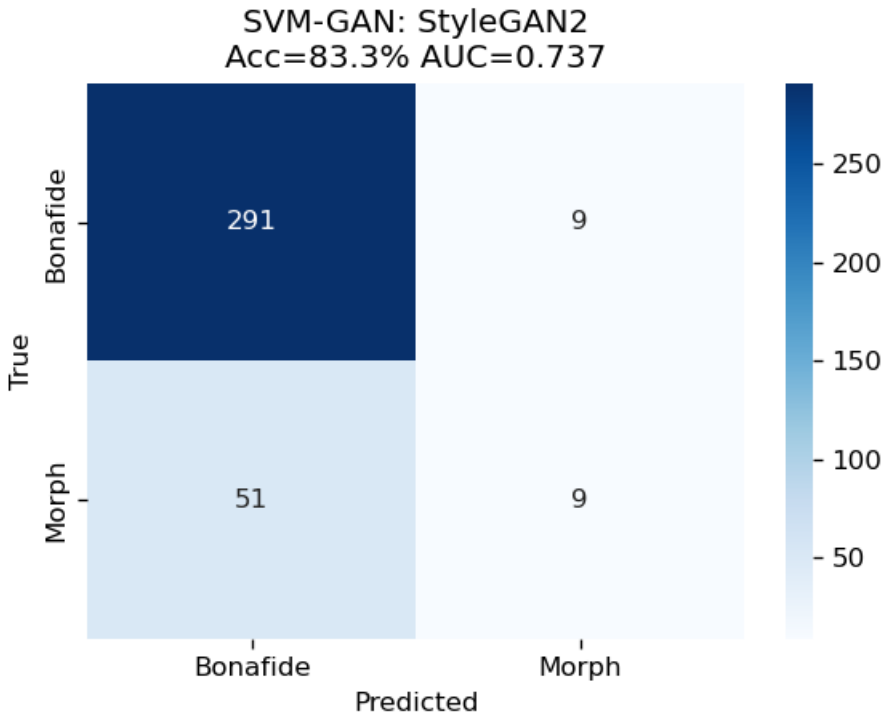
Σε σύγκριση με την τελευταία παρατήρηση, τα γεννητικά μοντέλα εμφανίζουν σημαντική αμφίπλευρη μεταβιβασιμότητα. Το SVM-GAN πραγματοποιεί 89,4% στα DDPM morphs, το

SVM-DDPM 83,3% στα StyleGAN2 morphs, και το SVM-DIFF συντηρεί επιδόσεις άνω του 80% τόσο στα GAN όσο και στα DDPM morphs.

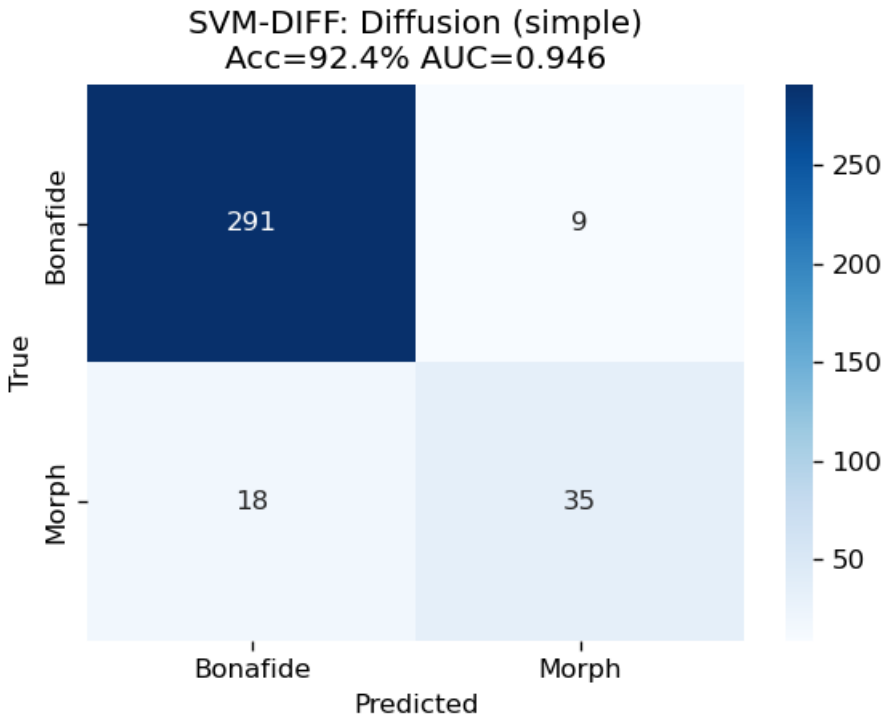
Αυτό δείχνει ότι οι τρεις γεννητικές μέθοδοι (StyleGAN2, diffusion-inspired και DDPM) έχουν στην διάθεση τους κοινά στατιστικά αποτυπώματα, τα οποία είναι διακριτά μέσω των MB-LBP ιδιοτήτων. Ενδεχόμενος, αφορά για ειδικούς τύπους που εισάγουν τα νευρωνικά δίκτυα κατά τη γέννηση, ασχέτως της ειδικής δομής.



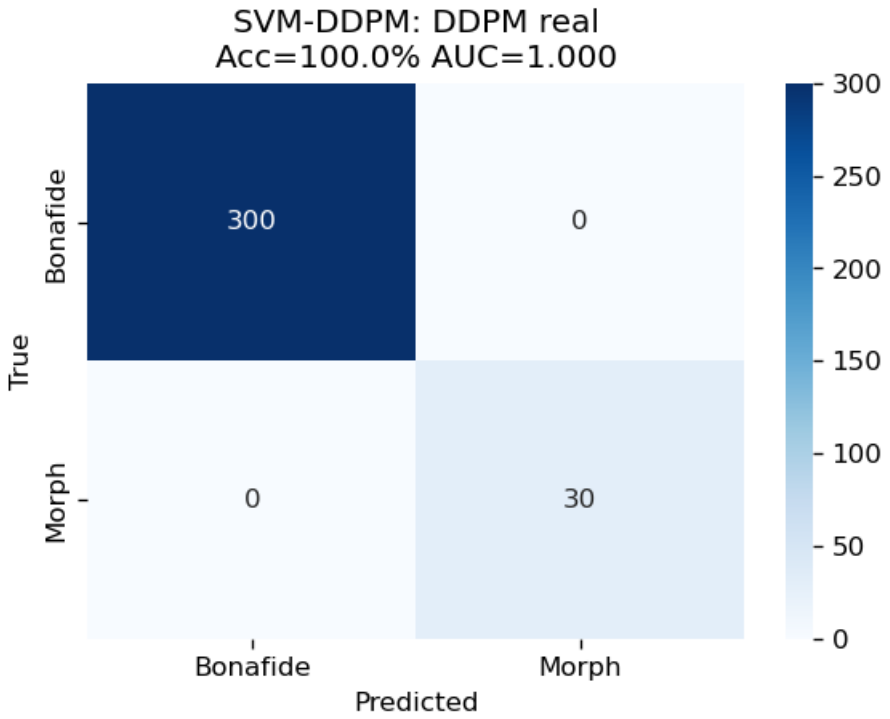
Σχήμα 6.2: Confusion matrix για τον SVM-LM detector.



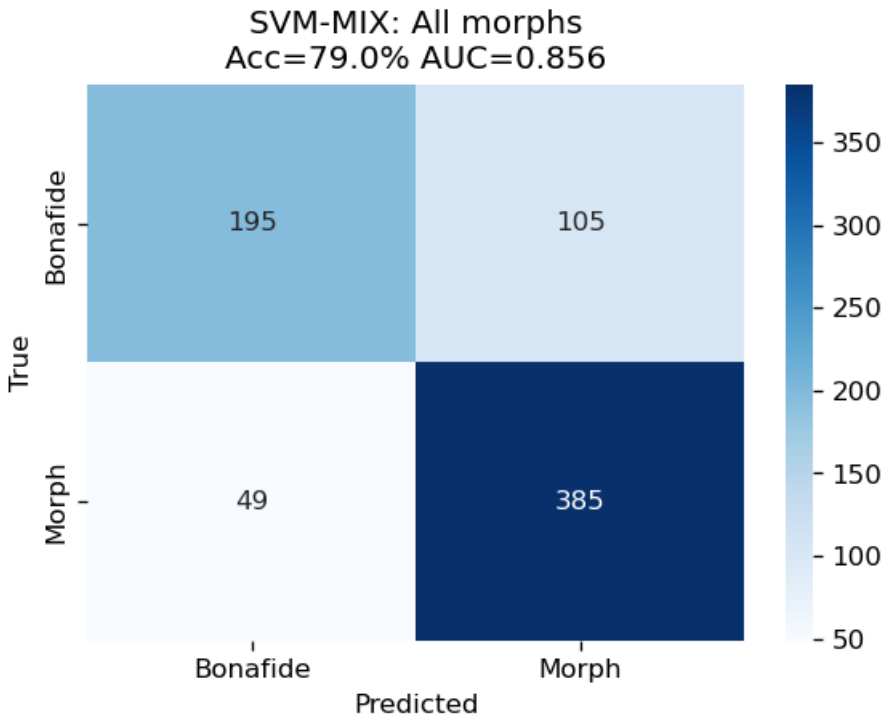
Σχήμα 6.3: Confusion matrix για τον SVM-GAN detector.



Σχήμα 6.4: Confusion matrix για τον SVM-DIFF detector.



Σχήμα 6.5: Confusion matrix για τον SVM-DDPM detector.



Σχήμα 6.6: Confusion matrix για τον SVM-MIX detector.

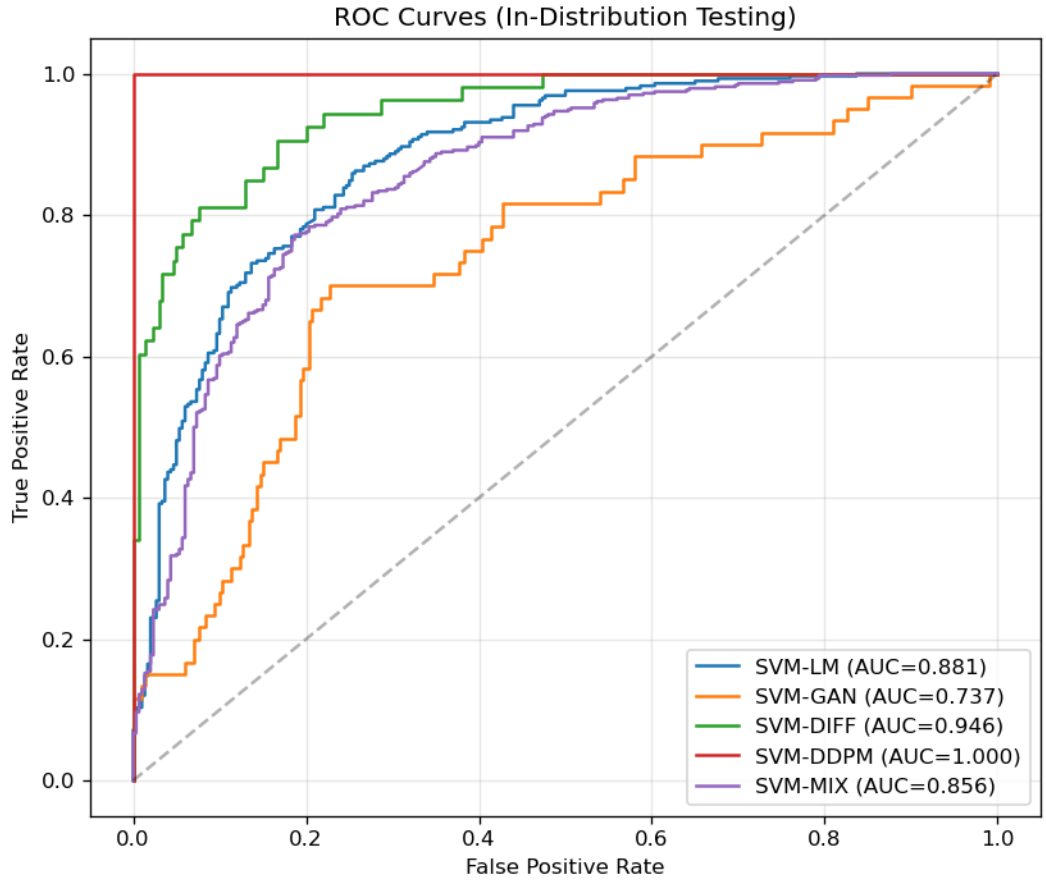
6.5 Η Επίδοση του Μεικτού Μοντέλου

Το SVM-MIX, το οποίο εκπαιδεύτηκε σε όλους τους τύπους morphs ταυτόχρονα, εμφανίζει μια ξεχωριστή συμπεριφορά. Υλοποιεί μέτρια αλλά σίγουρη απόδοση σε όλους τους τύπους (65,9% έως 78,9%), χωρίς όμως να φτάνει την ειδική επίδοση οποιουδήποτε από τα ειδικά μοντέλα στον ανάλογο τομέα τους.

Το συμπέρασμα αυτό δείχνει ότι η απλή ένωση ξεχωριστών τύπων morphs στο σύνολο εκπαίδευσης δεν φτάνει για τη παραγωγή ενός καθολικού ανιχνευτή. Από την άλλη, καταλήγει σε ένα μοντέλο "μέσου όρου" το οποίο συμβιβάζεται μεταξύ των ανταγωνιστικών απαιτήσεων κάθε τύπου. Για την ολοκλήρωση πραγματικών robust ανίχνευσης, χρειάζονται είτε αρκετά πιο μεγάλα αλλά και datasets με μεγαλύτερη ποικιλία, είτε πιο σύγχρονες απεικονίσεις χαρακτηριστικών από τα MB-LBP.

6.6 Καμπύλες ROC και Επιπλέον Μετρικές

Εκτός από την ακρίβεια, εκτιμήθηκαν και οι μετρικές AUC, APCER και BPCER για κάθε μοντέλο στο in-distribution σενάριο. Οι καμπύλες ROC εμφανίζουν τον συμφιλίωση μεταξύ ρυθμού εντοπισμού και λάθος συναγερμών .



Σχήμα 6.7: ROC curves των SVM detectors από το φάκελο 04_detection/results.

Οι επιμέρους πίνακες σύγχυσης για κάθε μοντέλο αναφέρονται στο Παράρτημα Β, διαθέτοντας αναλυτική απεικόνιση των σωστών και λανθασμένων ταξινομήσεων ανά κλάση.

6.6.1 Αναλυτικές Μετρικές ανά Μοντέλο

Το επόμενο πινακάκι μαζεύει τις κύριες μετρικές για κάθε μοντέλο στο in-distribution σενάριο. Επισημαίνεται ότι το APCER εκπροσωπεί το ποσοστό των morphs που απομακρύνθηκαν, ενώ το BPCER το ποσοστό των πραγματικών εικόνων που λανθασμένα απορρίφθηκαν.

Μοντέλο	Accuracy	AUC	APCER	BPCER
SVM-LM	80,9%	0,881	17,2%	24,3%
SVM-GAN	85,3%	0,737	85%	3,0%
SVM-DIFF	93,6%	0,946	34,0%	3,0%
SVM-DDPM	100,0%	1,000	0,0%	0,0%
SVM-MIX	78,3%	0,860	11,3%	35,0%

Πίνακας 6.2: Αναλυτικές μετρικές επίδοσης ανά μοντέλο (in-distribution).

- **SVM-LM: Landmark**
  True Bonafide (TN) = 227
  False Morph (FP) = 73
  False Bonafide (FN) = 50
  True Morph (TP) = 241
  Συνολικά Δείγματα = 227 + 73 + 50 + 241 = 591

  BPCER (Bona Fide Presentation Classification Error Rate):

$$\mathbf{BPCER} = \frac{\boldsymbol{FP}}{\mathbf{TN} + \mathbf{FP}} = \frac{\mathbf{73}}{\mathbf{227} + \mathbf{73}} = \frac{\mathbf{73}}{\mathbf{300}} \approx \mathbf{0,2433} \rightarrow \mathbf{24,3}\ \%$$

  APCER (Attack Presentation Classification Error Rate):

$$\mathbf{APCER} = \frac{\boldsymbol{FN}}{\mathbf{TP} + \mathbf{FN}} = \frac{\mathbf{50}}{\mathbf{241} + \mathbf{50}} = \frac{\mathbf{50}}{\mathbf{291}} \approx \mathbf{0,1718} \rightarrow \mathbf{17,2}\ \%$$

- **SVM-GAN: StyleGAN2**
  True Bonafide (TN) = 291
  False Morph (FP) = 9
  False Bonafide (FN) = 51
  True Morph (TP) = 9
  Συνολικά Δείγματα = 291 + 9 + 51 + 9 = 360

  BPCER (Bona Fide Presentation Classification Error Rate):

$$\mathbf{BPCER} = \frac{\boldsymbol{FP}}{\mathbf{TN} + \mathbf{FP}} = \frac{\mathbf{9}}{\mathbf{291} + \mathbf{9}} = \frac{\mathbf{9}}{\mathbf{300}} \approx \mathbf{0,0300} \rightarrow \mathbf{3,0}\ \%$$

  APCER (Attack Presentation Classification Error Rate):

$$\mathbf{APCER} = \frac{\boldsymbol{FN}}{\mathbf{TP} + \mathbf{FN}} = \frac{\mathbf{51}}{\mathbf{9} + \mathbf{51}} = \frac{\mathbf{51}}{\mathbf{60}} \approx \mathbf{0,8500} \rightarrow \mathbf{85,0}\%$$

- **SVM-DIFF: Diffusion**
  True Bonafide (TN) = 291
  False Morph (FP) = 9
  False Bonafide (FN) = 18
  True Morph (TP) = 35
  Συνολικά Δείγματα = 291 + 9 + 18 + 35 = 353

  BPCER (Bona Fide Presentation Classification Error Rate):

$$\mathbf{BPCER} = \frac{\boldsymbol{FP}}{\mathbf{TN} + \mathbf{FP}} = \frac{\mathbf{9}}{\mathbf{291} + \mathbf{9}} = \frac{\mathbf{9}}{\mathbf{300}} \approx \mathbf{0,0300} \rightarrow \mathbf{3,0}\ \%$$

  APCER (Attack Presentation Classification Error Rate):

$$\mathbf{APCER} = \frac{\boldsymbol{FN}}{\mathbf{TP} + \mathbf{FN}} = \frac{\mathbf{18}}{\mathbf{35} + \mathbf{18}} = \frac{\mathbf{18}}{\mathbf{53}} \approx \mathbf{0,3396} \rightarrow \mathbf{34,00}\%$$

- **SVM-DDPM: DDPM real**
  True Bonafide (TN) = 300
  False Morph (FP) = 0
  False Bonafide (FN) = 0

True Morph (TP) = 30
Συνολικά Δείγματα = 300 + 0 + 0 + 30 = 330

BPCER (Bona Fide Presentation Classification Error Rate):

$$\mathbf{BPCER} = \frac{\boldsymbol{FP}}{\mathbf{TN} + \mathbf{FP}} = \frac{\mathbf{0}}{\mathbf{300} + \mathbf{0}} = \frac{\mathbf{0}}{\mathbf{300}} \approx \mathbf{0,0000} \rightarrow \mathbf{0,0}\ \%$$

APCER (Attack Presentation Classification Error Rate):

$$\mathbf{APCER} = \frac{\boldsymbol{FN}}{\mathbf{TP} + \mathbf{FN}} = \frac{\mathbf{0}}{\mathbf{30} + \mathbf{0}} = \frac{\mathbf{0}}{\mathbf{30}} \approx \mathbf{0,0000} \rightarrow \mathbf{0,0}\ \%$$

- **SVM-MIX: All morphs**

True Bonafide (TN) = 195
False Morph (FP) = 105
False Bonafide (FN) = 49
True Morph (TP) = 385
Συνολικά Δείγματα = 195 + 105 + 49 + 385 = 734

BPCER (Bona Fide Presentation Classification Error Rate):

$$\mathbf{BPCER} = \frac{\boldsymbol{FP}}{\mathbf{TN} + \mathbf{FP}} = \frac{\mathbf{105}}{\mathbf{195} + \mathbf{105}} = \frac{\mathbf{105}}{\mathbf{300}} \approx \mathbf{0,3500} \rightarrow \mathbf{35,0}\ \%$$

APCER (Attack Presentation Classification Error Rate):

$$\mathbf{APCER} = \frac{\boldsymbol{FN}}{\mathbf{TP} + \mathbf{FN}} = \frac{\mathbf{49}}{\mathbf{385} + \mathbf{49}} = \frac{\mathbf{49}}{\mathbf{434}} \approx \mathbf{0,1129} \rightarrow \mathbf{11,3}\ \%$$

6.7 Ανάλυση Ασυμμετρίας της Γενίκευσης

Ένα χαρακτηριστικά σημαντικό φαινόμενο που εμφανίζεται από τον πίνακα cross-evaluation είναι η ασυμμετρία της γενίκευσης. Εξετάζοντας προσεκτικά, επιβεβαιώνεται ότι η μεταφορά γνώσης δεν είναι αμοιβαία με την ίδια ισχύ.

Ειδικότερα, το SVM-LM (εκπαιδευμένο σε landmark) πραγματοποιεί 69,2% στα StyleGAN2 morphs, ενώ το SVM-GAN (εκπαιδευμένο σε StyleGAN2) πραγματοποιεί μόλις 51,1% στα landmark morphs. Αυτή η ανισορροπία δείχνει ότι τα landmark morphs έχουν ένα

μεγαλύτερο φάσμα φανερών χαρακτηριστικών, μερικά από τα οποία συμπίπτουν με αυτά των γεννητικών morphs, ενώ το ανάποδο δεν ευσταθεί στον ίδιο βαθμό.

Μια ενδεχόμενη εξήγηση είναι ότι τα landmark morphs βάζουν τόσο γεωμετρικά artifacts (ghosting) όσο και μεταβολές υφής, ενώ τα γεννητικά morphs εισάγουν κυρίως λεπτές στατιστικές μεταβολές υφής. Έτσι, ένας εντοπιστής εκπαιδευμένος στα πιο “πλούσια” landmark morphs ενημερώνει ένα ευρύτερο σύνολο χαρακτηριστικών.

6.8 Παρατηρήσεις επί της Πρακτικής Εφαρμογής

Από τα αποτελέσματα πηγάζει μια βασική πρακτική παρατήρηση: η διαλογή των δεδομένων εκπαίδευσης ενός συστήματος MAD επηρεάζει σοβαρά την περιοχή αποτελεσματικότητάς του. Ένα σύστημα που εξελίσσεται τώρα και εκπαιδεύεται μόνο σε γνώριμες, τωρινές τεχνικές, μπορεί να μην πετύχει απέναντι σε μια “επαναδραστηριοποίηση” παλαιότερων, λιγότερο σύνθετων τεχνικών επίθεσης.

Αυτό το σχόλιο είναι αντίθετο με την κοινή διαίσθηση ότι τα νεότερα, πιο εξελιγμένα συστήματα είναι πάντοτε ανώτερα. Στο πεδίο της ασφάλειας, η κάλυψη ενός μεγάλου φάσματος τεχνικών επίσης και παρωχημένων φανερώνεται το ίδιο σημαντική με την εμβάθυνση στις πιο τελευταίες απειλές.

6.9 Σύνοψη Ευρημάτων

Συνοψίζοντας, η πειραματική εκτίμηση εμφάνισε τέσσερα βασικά ευρήματα.

1) Η εντός κατανομής επίδοση είναι ικανοποιητική για όλους τους τύπους.

2) Η γενίκευση από γεννητικά σε landmark morphs δεν τα καταφέρνει πλήρως.

3) Υπάρχει ισχυρή αμοιβαία μεταφερσιμότητα μεταξύ των γεννητικών μεθόδων.

4) Το σύνθετο μοντέλο δεν αντιπροσωπεύει επαρκή λύση για το πρόβλημα της συνολικής ανίχνευσης.

Τα αποτελέσματα αυτά αναλύονται αντικειμενικά στο επόμενο κεφάλαιο, μαζί με τα όρια της μελέτης.

ΚΕΦΑΛΑΙΟ 7: Διαβούλευση και Ηθικές Πτυχές

7.1 Ερμηνεία των Αποτελεσμάτων

Τα πειραματικά αποτελέσματα επαληθεύουν και αξιολογούν ένα φαινόμενο που έχει τονιστεί στη βιβλιογραφία: το ζήτημα της γενίκευσης είναι ο κύριος περιορισμός των σύγχρονων συστημάτων MAD. Η σχεδόν απρόβλεπτη επίδοση των γεννητικών μοντέλων στα landmark morphs (περίπου 50%) εμφανίζει ότι μια μέθοδο προστασίας εκπαιδευμένη μόνο σε σύγχρονες απειλές θα ήταν πρακτικά ανυπεράσπιστη απέναντι σε μια απλή, κλασσική επίθεση.

Αυτή η επισήμανση έχει ανάποδη ισχύ. Μια μέθοδος εκπαιδευμένη ειδικά σε landmark morphs (SVM-LM) κρατά μεν κάποια ικανότητα ανίχνευσης γεννητικών morphs (περίπου 68-78%), αλλά σίγουρα μικρότερη από την in-distribution επίδοσή του. Με αποτέλεσμα κανένας ξεχωριστός τύπος εκπαίδευσης δεν διαθέτει ολοκληρωμένη κάλυψη.

7.2 Πρακτικές Συνέπειες για Συστήματα Ασφαλείας

Από οπτική γωνιά, τα αποτελέσματα δείχνουν ότι ένα ουσιαστικό σύστημα ελέγχου διαβατηρίων θα πρέπει να εντάξει διαφορετικούς, ειδικούς ανιχνευτές αντί για έναν μοναδικό καθολικό. Αλλιώς, είναι αναγκαία η τακτική επανεκπαίδευση των μοντέλων καθώς προκύπτουν νέες τεχνικές morphing, σε μια συνεχή κούρσα εξοπλισμών ανάμεσα σε επιτιθέμενους και αμυνόμενους.

Η ισχυρή μεταφερσιμότητα ανάμεσα στις γεννητικές μεθόδους διαθέτει ωστόσο μια ελπίδα αισιοδοξίας: ένας ανιχνευτής εκπαιδευμένος σε morphs από ένα γεννητικό μοντέλο μοιάζει να γενικεύει αρκετά καλά σε άλλα γεννητικά μοντέλα. Αυτό δείχνει ότι, παρά την γρήγορη ύπαρξη νέων αρχιτεκτονικών, οι αμυντικοί μηχανισμοί ίσως δεν έχει νόημα να βελτιώνονται για κάθε νέο μοντέλο ξεχωριστά.

7.3 Περιορισμοί της Μελέτης

Είναι βασικό να εντοπιστούν τα όρια της παρούσας εργασίας. Πρώτον, ο όγκος συγκεκριμένων υποσυνόλων ήταν σχετικά μικρός, συγκεκριμένα για τα DDPM morphs (100 εικόνες). Αυτό χρήζει συγκεκριμένα ευρήματα, όπως το 100% ακρίβειας του SVM-DDPM, λιγότερο φερέγγυα από στατιστική άποψη.

Δεύτερον, η επιλογή των MB-LBP ως αποκλειστικού αφηγητή χαρακτηριστικών, αν και διαθέτει ερμηνευσιμότητα, οριοθετεί την δεξιότητα του συστήματος σε σχέση με μοντέρνες deep learning προσεγγίσεις. Είναι εφικτό ότι ένα CNN-based σύστημα θα εμφάνιζε ξεχωριστά, ίσως καλύτερα, μοτίβα γενίκευσης.

Τρίτον, η εκτίμηση δημιουργήθηκε αποκλειστικά σε εικόνες προκύπτουσες από το FFHQ. Η συμπεριφορά των μοντέλων σε εικόνες από διαφορετικές πηγές, με διαφορετικά χαρακτηριστικά κάμερας και συμπίεσης, θα έχουν διαφορετική αντιμετώπιση.

Τέλος, η μέθοδος εμφάνισης του StyleGAN2 πραγματοποιήθηκε σε λειτουργία fallback λόγω εξειδικευμένων περιορισμών του υλικού, δεδομένο που μεγάλωσε σοβαρά τον χρόνο υπολογισμού, χωρίς όμως να αλλάζει την ποιότητα των τελικών morphs.

7.4 Ηθικές Διαστάσεις

Η μελέτη στο πεδίο του face morphing υψώνει σημαντικά ηθικά προβλήματα. Η ίδια η τεχνολογία που ερευνάται για σκοπούς άμυνας μπορεί να χρησιμοποιηθεί κακόβουλα. Επομένως, η συγκεκριμένη διατριβή εφαρμόζει μια σειρά αρχών υπεύθυνης έρευνας.

Το βασικότερο μέτρο είναι η επιβλητική προστασία των ανηλίκων. Όπως αναφέρθηκε στο κεφάλαιο 4.6, δημιουργήθηκε διπλό στάδιο φιλτραρίσματος ηλικίας ώστε να διασφαλιστεί ότι κανένα morph δεν δημιουργείται από εικόνες που μπορεί να ανήκουν σε ανηλίκους. Η επιλογή ενός συντηρητικού ορίου ηλικίας (20 ετών) αντανακλά την προτίμηση που δίνεται σε αυτό το ζήτημα.

Επίσης, όλες οι εικόνες πηγάζουν από το FFHQ, ένα dataset που μοιράζεται υπό άδεια Creative Commons για ερευνητική χρήση. Τα παραγόμενα morphs εφαρμόζονται ειδικά για

ακαδημαϊκούς σκοπούς και δεν ενδείκνυνται για καμία πρακτική υλοποίηση πέραν της μελέτης εντοπισμού.

Η εργασία ευθυγραμμίζεται επίσης με το πνεύμα του Κανονισμού 2024/1689 [22] της ΕΕ για την τεχνητή νοημοσύνη, ο οποίος τοποθετεί τη βιομετρική ταυτοποίηση στις εφαρμογές υψηλού κινδύνου και χρήζει σημαντική προσοχή στην διαμόρφωση τέτοιων δεδομένων.

ΚΕΦΑΛΑΙΟ 8: Συμπεράσματα και Μελλοντική Εργασία

8.1 Σύνοψη της Εργασίας

Η τρέχουσα μεταπτυχιακή διατριβή ερεύνησε το πρόβλημα των επιθέσεων μορφοποίησης προσώπου από δύο όψεις: την δημιουργία και τον εντοπισμό. Δημιουργήθηκε ένα ολοκληρωμένο pipeline που φτιάχνει morphs με τέσσερις ξεχωριστές τεχνικές, αναπληρώνοντας τη ποικιλία από τις κλασσικές landmark-based μεθόδους έως τα μοντέρνα μοντέλα διάχυσης.

Ταυτόχρονα, φτιάχτηκε ένα σύστημα εντοπισμού βασισμένο σε MB-LBP ιδιότητες και ταξινομητή SVM, το οποίο εκτιμήθηκε μέσω ενός ασυνήθιστου σχεδιασμού cross-evaluation. Ο σχεδιασμός αυτός εμφάνισε σοβαρά μοτίβα με βάση την ικανότητα γενίκευσης των ανιχνευτών.

8.2 Κύρια Συμπεράσματα

Τα βασικά συμπεράσματα της εργασίας συνοψίζονται ως εξής:

1. Οι ανιχνευτές morphing εμφανίζουν φοβερή απόδοση στον τύπο morph για τον οποίο διαπλάθονται, αλλά η γενίκευσή τους σε άλλους τύπους είναι τακτικά ελαττωματική.
2. Η γενίκευση από γεννητικές διαδικασίες προς landmark morphs αστοχεί τελείως, με επιδόσεις ίδιες της τυχαίας ταξινόμησης.
3. Υπάρχει έντονη αμοιβαία μεταφερσιμότητα μεταξύ των πολλών γεννητικών τεχνικών (StyleGAN2, diffusion, DDPM), φανερώνοντας κοινά στατιστικά αποτυπώματα.
4. Ο απλός συνδυασμός τύπων morphs στην εκπαίδευση δεν δημιουργεί έναν δραστικό καθολικό ανιχνευτή, αλλά ένα μοντέλο μέσου όρου.
5. Η συμπερίληψη ηθικού φιλτραρίσματος ηλικίας είναι δυνατή και πρέπει να είναι ένα αναπόσπαστο μέρος κάθε morphing pipeline.

8.3 Προτάσεις για Μελλοντική Εργασία

Η εργασία αυτή ξεκλειδώνει πολλούς προσανατολισμούς για μελλοντική διερεύνηση. Μια εμφανής διεύρυνση είναι η αντικατάσταση των MB-LBP ιδιοτήτων με deep learning απεικονίσεις, όπως embeddings από προ-εκπαιδευμένα CNN δίκτυα, και ο συσχετισμός των μοτίβων γενίκευσης που σχηματίζονται.

Μια δεύτερη κατεύθυνση είναι η επέκταση του dataset, τόσο σε μέγεθος όσο και σε διαφορετικότητα πηγών, ώστε τα πορίσματα να είναι πιο στατιστικά έγκυρα. Η ενσωμάτωση εικόνων από ξεχωριστά datasets θα άφηνε επίσης την εκτίμηση cross-dataset γενίκευσης, πέραν της cross-method που εξετάστηκε εδώ.

Τρίτον, θα ήταν συναρπαστικό η αναζήτηση τεχνικών domain adaptation και ensemble learning, με επιδίωξη τη παραγωγή ενός πραγματικά robust ανιχνευτή ο οποίος θα συσχετίζει τα πλεονεκτήματα των συγκεκριμένων μοντέλων.

Τέλος, η έρευνα πιο εξελιγμένων τεχνικών diffusion morphing, όπως η ολοκληρωμένη DDIM inversion με conditional μοντέλα ή η χρήση Stable Diffusion, θα είχε την ικανότητα να δημιουργεί morphs ακόμη υψηλότερης ποιότητας και να βάλει νέες προκλήσεις στους ανιχνευτές.

8.4 Επίλογος

Η σύγκρουση ανάμεσα στην παραγωγή και την ανίχνευση επιθέσεων μορφοποίησης προσώπου είναι ένας τακτικός ανταγωνισμός, στον οποίο κάθε πρόοδος στο ένα δημιουργεί αντίστοιχη αντίδραση στο άλλο. Η συγκεκριμένη διατριβή βοήθησε σε αυτή την προσπάθεια δημιουργώντας μια συστηματική, ποσοτική μελέτη της επέκτασης των ανιχνευτών, καθώς και ένα ολοκληρωμένο, αναπαρατάξιμο pipeline που έχει την ικανότητα να είναι βάση για περισσότερη διερεύνηση.

Επειδή τα γεννητικά μοντέλα εξακολουθούν να αναπτύσσονται με γρήγορους ρυθμούς, η ανάγκη για αντοχή και επεκτάσιμα συστήματα εντοπισμού γίνεται ολοένα και πιο επιτακτική. Η

προσδοκία είναι ότι έρευνες όπως η παρούσα βοηθούν στην καλύτερη γνώση του προβλήματος και, εν τέλει, στην εξέλιξη ασφαλέστερων βιομετρικών συστημάτων.

Βιβλιογραφία

[1] M. Ferrara, A. Franco, and D. Maltoni, “The Magic Passport,” in IEEE International Joint Conference on Biometrics (IJCB), 2014, pp. 1–7.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.

[3] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training Generative Adversarial Networks with Limited Data,” in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[4] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 6840–6851.

[5] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” in International Conference on Learning Representations (ICLR), 2021.

[6] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[7] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, “Face Recognition Systems Under Morphing Attacks: A Survey,” IEEE Access, vol. 7, pp. 23012–23026, 2019.

[8] U. Scherhag, C. Rathgeb, and C. Busch, “Towards Detection of Morphed Face Images in Electronic Travel Documents,” in 13th IAPR Int. Workshop on Document Analysis Systems (DAS), 2018, pp. 187–192.

[9] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs?,” in Int. Workshop on Biometrics and Forensics (IWBF), 2020.

[10] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, “Are GAN-based Morphs Threatening Face Recognition?,” in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[11] N. Damer, F. Boutros, A. M. Saladie, F. Kirchbuchner, and A. Kuijper, “Realistic Dreams: Cascaded Enhancement of GAN-Generated Images with an Example in Face Morphing Attacks,” in IEEE BTAS, 2019.

[12] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.

[13] Z. Blasingame and C. Liu, “Leveraging Diffusion for Strong and High Quality Face Morphing Attacks,” IEEE Trans. Biometrics, Behavior, and Identity Science, 2024.

[14] N. Damer et al., “Privacy-Friendly Synthetic Data for the Development of Face Morphing Attack Detectors,” in Proc. IEEE/CVF CVPR Workshops (CVPRW), 2022, pp. 1606–1617.

[15] M. Huber et al., “SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data,” in IEEE Int. Joint Conf. on Biometrics (IJCB), 2022.

[16] I. Batskos and L. Spreeuwers, “Improving Landmark-Based Face Morphing with Poisson Blending,” in Int. Workshop on Biometrics and Forensics (IWBF), 2024.

[17] D. E. King, “Dlib-ml: A Machine Learning Toolkit,” Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[18] G. Bradski, “The OpenCV Library,” Dr. Dobb’s Journal of Software Tools, 2000.

[19] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in Advances in Neural Information Processing Systems (NeurIPS), 2019.

[20] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[21] S. I. Serengil and A. Ozpinar, “LightFace: A Hybrid Deep Face Recognition Framework,” in Innovations in Intelligent Systems and Applications Conf. (ASYU), IEEE, 2020.

[22] European Parliament, “Regulation (EU) 2024/1689 on Artificial Intelligence (Artificial Intelligence Act),” Official Journal of the European Union, 2024.

[23] I. Goodfellow et al., “Generative Adversarial Networks,” in Advances in Neural Information Processing Systems (NeurIPS), 2014.

Παράρτημα Α: Δομή Κώδικα και Οδηγίες Εκτέλεσης

Το πλήθος του κώδικα δημιουργήθηκε σε κλιμακωτά αριθμημένα Python scripts, ώστε να αντανακλά τη κυκλοφορία του pipeline. Παρακάτω αναφέρεται περιληπτικά η δομή.

Α.1 Περιβάλλοντα Εκτέλεσης

Εφαρμόστηκαν δύο διαφορετικά conda environments: το “face_morph” (Python 3.9) για τα στάδια προεπεξεργασίας, landmark morphing και ανίχνευσης, και το “stylegan2” για τα στάδια StyleGAN2 και DDPM που απαιτούν PyTorch με υποστήριξη CUDA.

- face_morph: dlib, opencv-python, scikit-learn, scikit-image, deepface, matplotlib, seaborn.
- stylegan2: torch 2.0.1+cu118, diffusers, transformers, accelerate, huggingface-hub.

Α.2 Σειρά Εκτέλεσης Scripts

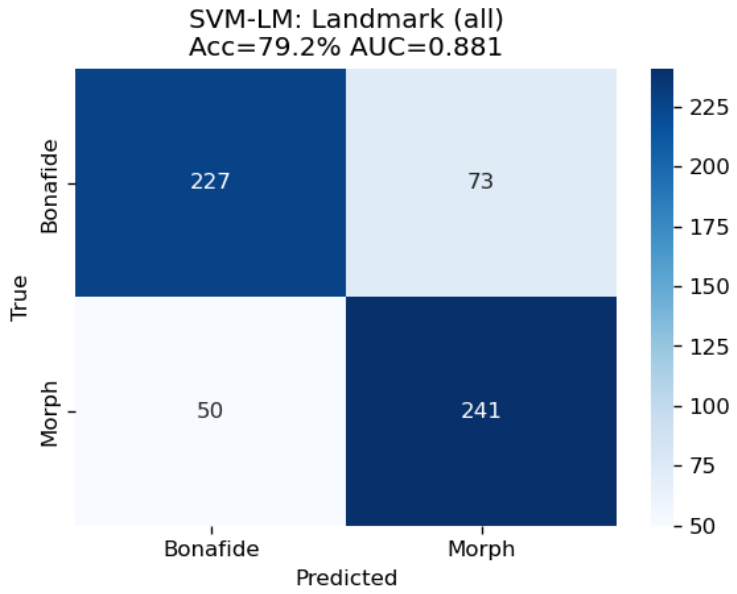
Η υλοποίηση λειτουργεί με τη σειρά: λήψη dataset,

- landmark morphing
- StyleGAN2 projection και morphing
- diffusion/DDPM morphing
- ηθικό φιλτράρισμα
- εξαγωγή χαρακτηριστικών
- τελική αξιολόγηση SVM

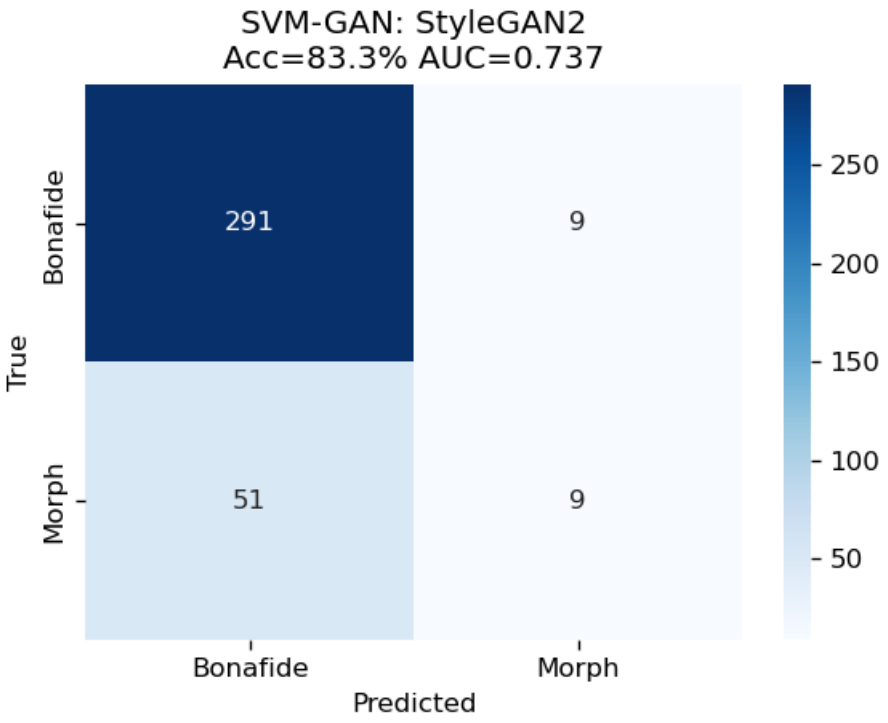
Στον φάκελο της εργασίας υπάρχει και το έγγραφο README.txt με τις αναλυτικές οδηγίες.

Παράρτημα Β: Πίνακες Σύγχυσης και Επιπλέον Σχήματα

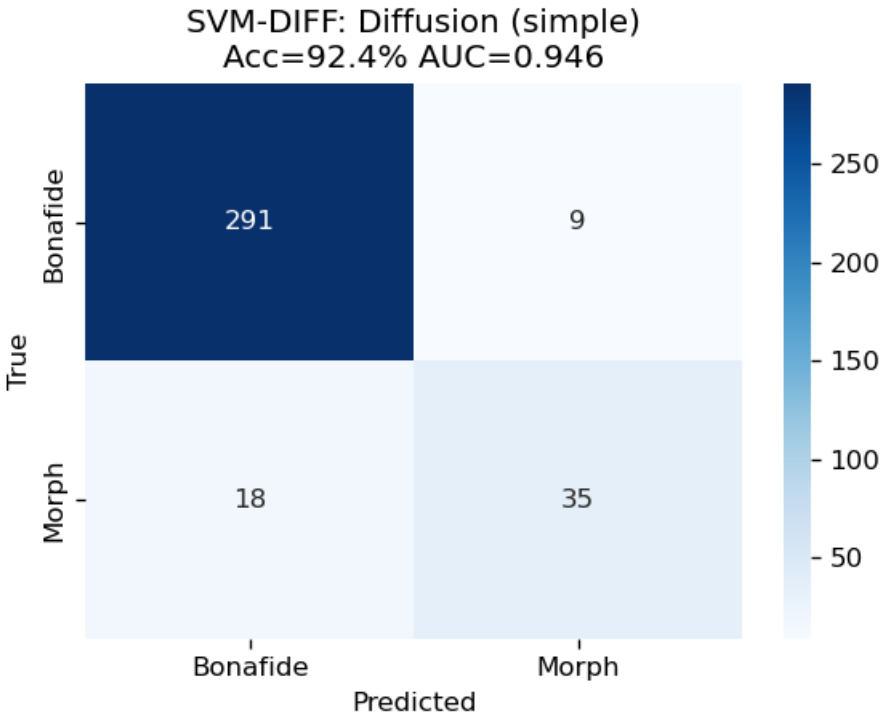
Στο παράρτημα αυτό εμφανίζονται οι πίνακες σύγχυσης (confusion matrices) για κάθε ένα από τα πέντε εκπαιδευμένα μοντέλα, καθώς και τυχόν επιπλέον οπτικοποιήσεις.



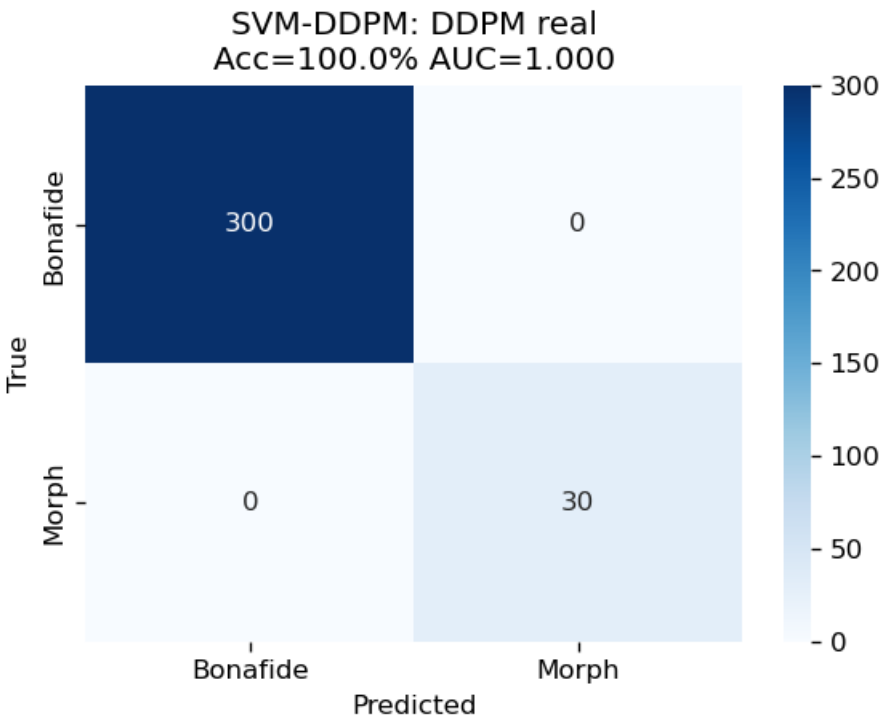
Σχήμα Β.1: Πίνακας σύγχυσης για το μοντέλο SVM-LM.



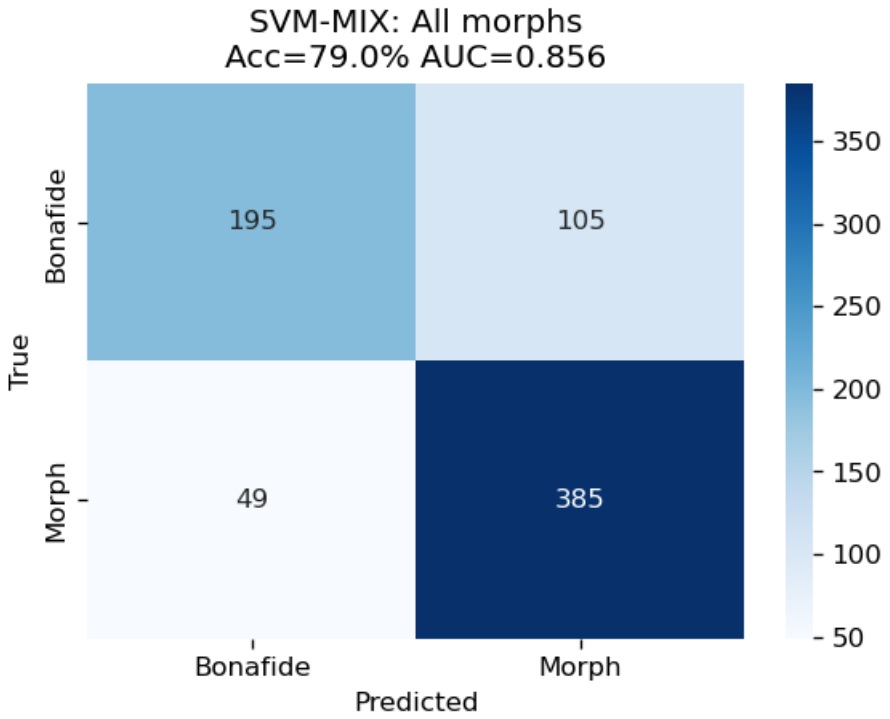
Σχήμα Β.2: Πίνακας σύγχυσης για το μοντέλο SVM-GAN.



Σχήμα Β.3: Πίνακας σύγχυσης για το μοντέλο SVM-DIFF.



Σχήμα Β.4: Πίνακας σύγχυσης για το μοντέλο SVM-DDPM.



Σχήμα Β.5: Πίνακας σύγχυσης για το μοντέλο SVM-MIX.