Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εργαστηριο Συστηματων Τεχνητης Νοημοσυνης και Μαθησης

Generating Realisitc and Sparse Medical Image Counterfactuals using StyleGAN

Diploma Thesis

by

Ioannis Litsos

Επιβλέπων: Αθανάσιος Βουλόδημος
Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Generating Realisitc and Sparse Medical Image Counterfactuals using StyleGAN

Diploma Thesis

by

Ioannis Litsos

Επιβλέπων: Αθανάσιος Βουλόδημος
Επ. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Μαρτίου, 2025.

.....	.....	.....
Αθανάσιος Βουλόδημος	Γεώργιος Στάμου	Α.-Γ. Σταφυλοπάτης
Επ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π

Αθήνα, Μάρτιος 2025

.....

Ιωαννης Λιτσος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © – All rights reserved Ioannis Litsos, 2025.
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Στη μητέρα μου

Περίληψη

Τα μοντέλα Βαθιάς Μάθησης έχουν εφαρμοστεί σε πολύ μεγάλο βαθμό στο πεδίο της ιατρικής απεικόνισης, βελτιώνοντας σημαντικά τις διαγνωστικές δυνατότητες. Ωστόσο, η εγγενής φύση τους ως "μαύρα κουτιά" δημιουργεί ουσιώδη ηθικά και πρακτικά ζητήματα, προκαλώντας ανησυχίες σχετικά με την ευρεία υιοθέτησή τους. Οι αντιπαραδειγματικές εξηγήσεις (counterfactual explanations), οι οποίες παρέχουν ανθρωπίνως κατανοητές πληροφορίες προτείνοντας ελάχιστες, αλλά ουσιώδεις, αλλαγές στα δεδομένα εισόδου για την αλλαγή των προβλέψεων του μοντέλου, αναδύονται ως μια πολλά υποσχόμενη μέθοδο για την επεξήγηση της λειτουργίας αυτών των συστημάτων. Παρ' όλα αυτά, οι υπάρχουσες μέθοδοι στη βιβλιογραφία για την παραγωγή αντιπαραδειγματικών εικόνων στον τομέα της ιατρικής απεικόνισης εμφανίζουν σημαντικούς περιορισμούς, συμπεριλαμβανομένου του μειωμένου ρεαλισμού και της διαγνωστικής λεπτομέρειας των παραγόμενων εικόνων, της ανάγκης κάποιας μορφής εποπτείας κατά την εκπαίδευση του ερμηνευτή (explainer), της υποχρέωσης για επανεκπαίδευση του ερμηνευτή για κάθε ταξινομητή ξεχωριστά, καθώς και της έλλειψης αραιότητας (sparsity) στις προτεινόμενες αλλαγές.

Στην παρούσα διπλωματική εργασία, αντιμετωπίζουμε αυτές τις προκλήσεις δημιουργώντας το SPRUCE (SParse Realistic Uncoupled Counterfactual Explanations), ένα νέο πλαίσιο ειδικά σχεδιασμένο για τη δημιουργία αραιών και ρεαλιστικών αντιπαραδειγματικών εικόνων στον χώρο της ιατρικής απεικόνισης. Το SPRUCE βασίζεται σε μια προσέγγιση με Παραγωγικά Ανταγωνιστικά Δίκτυα (ΠΑΔ) και αξιοποιεί μια εξειδικευμένη συνάρτηση απώλειας, η οποία είναι ειδικά σχεδιασμένη για να διατηρεί τη διαγνωστική συνάφεια και την οπτική πιστότητα των παραγόμενων εικόνων, οδηγώντας παράλληλα σε αραιές (sparse) τροποποιήσεις. Το βασικό πλεονέκτημα του συγκεκριμένου πλαισίου έγκειται στη δυνατότητα αποσύζευξης της εκπαίδευσης του ερμηνευτή από την εκπαίδευση του ταξινομητή, επιτρέποντας έτσι την χρήση ενός ΠΑΔ σε πολλαπλούς ταξινομητές που αναφέρονται σε εικόνες ίδιου τομέα ιατρικής απεικόνισης.

Στην εκτεταμένη πειραματική αξιολόγησή μας, χρησιμοποιήσαμε προηγμένες αρχιτεκτονικές ΠΑΔ,και πιο συγκεκριμένα το StyleGAN2-ADA, σε συνδυασμό με εξελιγμένες μεθόδους αναστροφής ΠΑΔ (GAN inversion), συμπεριλαμβανομένης της αναστροφής μέσω κωδικοποιητή (encoder-based inversion) και της μεθόδου pivotal tuning, ώστε να εξασφαλίσουμε επεξεργάσιμες λανθάνουσες αναπαραστάσεις υψηλής ποιότητας. Εφαρμόσαμε το SPRUCE σε δεδομένα από διάφορους τομείς ιατρικής εικόνας , όπως ακτινογραφίες θώρακος, οπτικές τομογραφίες συνοχής και μαγνητικές τομογραφίες εγκεφάλου,πετυχαίνοντας ιδιαίτερα καλές επιδόσεις σε μετρικές όπως η Fréchet Inception Distance (FID) και η Conditional Maximum Mean Discrepancy (CMMD). Επιπλέον, μέσα από ποιοτική και ποσοτική ανάλυση, διαπιστώσαμε ότι η σημασιολογική συνοχή των αντιπαραδειγματικών τροποποιήσεων συνδέεται με την ευρωστία (robustness) του εκάστοτε ταξινομητή, αναδεικνύοντας έτσι το SPRUCE όχι μόνο ως ένα εργαλείο επεξήγησης αλλά και ως έναν μηχανισμό διαγνωστικής αξιολόγησης και βελτίωσης της ανθεκτικότητας των μοντέλων προτού χρησιμοποιηθούν κλινικά.

Λέξεις-κλειδιά — Αντιπαραδειγματικές Εξηγήσεις, Ερμηνεύσιμη Τεχνητή Νοημοσύνη, Ιατρική Απεικόνιση, Παραγωγικά Ανταγωνιστικά Δίκτυα, Ανταγωνιστική Ανθεκτικότητα, Ταξινόμηση Εικόνας

Abstract

Deep Learning (DL) models have demonstrated great applicability in the field of medical imaging, significantly ameliorating diagnostic capabilities. However, their inherent black-box nature poses substantial ethical and practical challenges, raising concerns about their adoption. Counterfactual explanations, which provide human-interpretable insights by suggesting minimal yet meaningful modifications to input data to alter model predictions, have emerged as a promising approach to illuminate these opaque models.However,current methods for generating medical image counterfactuals exhibit critical limitations, including insufficient realism and diagnostic detail,the requirement of some form of supervision during training ,the necessity to retrain explainers for each classifier independently, and lack of sparsity in generated edits.

In this thesis, we address these challenges,by creating SPRUCE (SParse Realistic and Uncoupled Counterfactual Explanations), a novel framework specifically designed for generating sparse and realistic medical image counterfactuals. SPRUCE introduces a Generative Adversarial Network (GAN)-based approach which utilizes a specialized loss function, explicitly crafted to maintain the diagnostic relevance and visual fidelity of generated images while enforcing sparsity in modifications. The framework's core advantage lies in its ability to decouple the explainer's training from the classifier's training, allowing for the independent training of a generative model that can subsequently be employed across various classifiers within the same medical imaging domain.

In our extensive experimental evaluation, we employed state-of-the-art GAN architectures, particularly StyleGAN2-ADA, coupled with advanced GAN inversion methods, including encoder-based inversion and pivotal tuning, to ensure high-quality, editable latent representations. We validated SPRUCE using multiple medical imaging modalities, such as chest X-rays, OCT scans, and brain MRIs, demonstrating superior performance in terms of Fréchet Inception Distance (FID), Conditional Maximum Mean Discrepancy (CMMD), and other metrics. Furthermore,through our qualitative and quantitative analysis we find that the semantic coherence of counterfactual edits is tied to classifier robustness, positioning SPRUCE not only as an explanatory tool but also as a diagnostic mechanism to assess and improve model robustness prior to clinical deployment.

Keywords — Counterfactual Explanations, XAI, Medical Imaging, Generative Adversarial Networks, Adversarial Robustness, Image Classification

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας μού επέτρεψε να εμβαθύνω σε ένα ευρύ φάσμα θεμάτων που αφορούν την Τεχνητή Νοημοσύνη, τη Βαθιά Μάθηση και την Όραση Υπολογιστών. Μέσα από αυτήν, ήρθα για πρώτη φορά σε ουσιαστική επαφή με την ερευνητική διαδικασία και τις προκλήσεις που τη συνοδεύουν.Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Αθανάσιο Βουλόδημο, για την ευκαιρία που μου προσέφερε να εκπονήσω την εργασία αυτή στο Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης, καθώς και για την εμπιστοσύνη και την πολύτιμη καθοδήγησή του σε κάθε στάδιο της προσπάθειας αυτής.

Επιπλέον, θα ήθελα να ευχαριστήσω από καρδιάς τον Νίκο Σπανό, τον Ιάσονα Λιάρτη και την Παρασκευή Θεοφίλου για την άψογη συνεργασία και τη συνεχή υποστήριξη κατά τη διάρκεια της διπλωματικής μου. Οι συμβουλές τους υπήρξαν καθοριστικές για την επιτυχή ολοκλήρωση της.

Επίσης, θα ήθελα να εκφράσω την εγκάρδια ευγνωμοσύνη μου στους φίλους μου, με τους οποίους μοιραστήκαμε αμέτρητες ώρες μελέτης, συζήτησης, διασκέδασης και ταξιδιών. Οι στιγμές αυτές δημιούργησαν αναμνήσεις που θα με συνοδεύουν για πάντα,ενώ η συντροφικότητα και η υποστήριξή τους με βοήθησαν να παραμείνω προσγειωμένος και αισιόδοξος ακόμα και στις πιο απαιτητικές περιόδους.

Τέλος,ένα ιδιαίτερο ευχαριστώ στην οικογένειά μου, η οποία αποτέλεσε εξαρχής ένα ανεκτίμητο στήριγμα για εμένα. Σε εκείνους οφείλω όλη τη διαδρομή των σπουδών μου μέχρι σήμερα, καθώς η αγάπη, η κατανόηση και η πίστη τους στις δυνατότητές μου υπήρξαν πηγή δύναμης και έμπνευσης.

Ιωάννης Λίτσος, Μάρτιος 2025

Contents

List of Figures

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Εισαγωγή

Τα μοντέλα βαθιάς μάθησης αποτελούν πλέον ένα βασικό εργαλείο στον τομέα της επεξεργασίας εικόνας και είναι ιδιαίτερα διαδεδομένα στην ταξινόμηση ιατρικών εικόνων [74, 8, 92, 56, 133, 101, 87]. Ένας σημαντικός περιορισμός αυτών των μοντέλων είναι ότι λειτουργούν ως "μαύρα κουτιά", δηλαδή η διαδικασία λήψης αποφάσεων παραμένει δυσνόητη στους ανθρώπους, ενώ συχνά βασίζονται σε στατιστικές εξαρτήσεις των δεδομένων εκπαίδευσης και δεν αφομοιώνουν εξειδικευμένη γνώση τομέα [124, 137]. Παρόλο που αυτό μπορεί να είναι αποδεκτό σε άλλους τομείς, στην ιατρική απεικόνιση οι λανθασμένες αποφάσεις ενός ταξινομητή μπορούν να υπονομεύσουν την υγειονομική περίθαλψη και να θέσουν σε κίνδυνο ασθενείς. Το γεγονός αυτό έχει οδηγήσει τόσο την ερευνητική κοινότητα στην ανάπτυξη τεχνικών για Ερμηνεύσιμη Τεχνητή Νοημοσύνη (Explainable Artificial Intelligence) [37, 119], όσο και την αντίστοιχη νομική πλευρά στην εισαγωγή νόμων που καθιερώνουν το "δικαίωμα στην επεξήγηση", δηλαδή το δικαίωμα παροχής αιτιολόγησης για μια απόφαση που λαμβάνεται από ένα σύστημα Τεχνητής Νοημοσύνης το οποίο επηρεάζει σημαντικά ένα άτομο.[1]

Για την επεξήγηση των "μάυρων κουτιών" που χρησιμοποιούνται στην ταξινόμηση εικόνας έχουν αναπτυχθεί διάφορες μέθοδοι. Μία τέτοια προσέγγιση είναι η δημιουργία αντιπαραδειγματικών εικόνων [13, 47, 36, 27]. Οι αντιπαραδειγματικές επεξηγήσεις παρέχουν ένα νέο δεδομένο εισόδου, παρόμοιο με το αρχικό, το οποίο παράγει διαφορετική έξοδο, περιγράφοντας ουσιαστικά τις αναγκαίες αλλαγές στην είσοδο ώστε να αλλάξει η απόφαση του εκάστοτε μοντέλου. Αυτό το είδος επεξήγησης μοιάζει σε μεγάλο βαθμό με την ανθρώπινη διαδικασία λήψης αποφάσεων και έχει δειχθεί ότι είναι εύκολα ερμηνεύσιμο από ανθρώπους [34, 84, 120]. Επιπλέον,η διαφορά σε επίπεδο εικονοστοιχείων (pixels) μεταξύ της αρχικής και της αντιπαραδειγματικής εικόνας μπορεί να λειτουργήσει ως "χάρτης σημαντικότητας" [83](saliency map), με το πρόσθετο πλεονέκτημα ότι η ίδια η αντιπαραδειγματική εικόνα χρησιμεύει ως πιστοποιητικό ορθότητας για τον εν λόγω χάρτη.

Τεχνικά, οποιαδήποτε εικόνα που παράγει διαφορετική έξοδο από την αρχική μπορεί να λειτουργήσει ως αντιπαραδειγματική, αλλά προκειμένου να προσφέρει ουσιαστικές πληροφορίες, η αντιπαραδειγματική εικόνα οφείλει να εμφανίζει μόνο τις απολύτως αναγκαίες αλλαγές στην εικόνα εισόδου για να αναστραφεί η έξοδος του ταξινομητή. Για παράδειγμα, σε έναν ταξινομητή που εντοπίζει πνευμονία σε αξονικές τομογραφίες θώρακα, θα προτιμούσαμε η περιοχή αλλαγής να βρίσκεται αποκλειστικά στους πνεύμονες, και όχι στο υπόβαθρο ή στη μορφή του σκελετού, καθώς αυτό θα μπορούσε να αντιπροσωπεύει έναν εντελώς διαφορετικό ασθενή. Ένας σημαντικός περιορισμός πολλών εργασιών στη σχετική βιβλιογραφία [5, 116, 35, 80, 14] είναι ότι δεν εισάγουν μηχανισμούς στη διαδικασία δημιουργίας αντιπαραδειγματικών επεξηγήσεων που να εξασφαλίζουν ότι η τελική εικόνα θα εμφανίζει αραιές(sparse), αντιληπτά παρόμοιες και οπτικά συνεκτικές αλλαγές. Σε αυτήν την εργασία, χρησιμοποιούμε διάφορους όρους κανονικοποίησης στη διαδικασία δημιουργίας αντιπαραδειγματικών εικόνων που εξασφαλίζουν τις προαναφερθείσες ιδιότητες, συγκεκριμένα την L1 κανονικοποίηση τόσο στον χώρο της εικόνας όσο και στον λανθάνοντα χώρο του Παραγωγικού Ανταγωνιστικού Δικτύου (ΠΑΔ), καθώς και τη συνάρτηση Learned Perceptual Image Patch Similarity [138].

Ένας άλλος σημαντικός περιορισμός σε πολλές εργασίες [86, 5, 61, 116, 105, 80, 14] είναι ότι απαιτούν την από

[1]

κοινού, με κάποιον τρόπο, εκπαίδευση του μοντέλου που παράγει τις επεξηγηματικές εικόνες με την εκπαίδευση του ταξινομητή που εξετάζεται, με αποτέλεσμα να μην μπορούν να εφαρμόσουν το πλαίσιο τους σε οποιοδήποτε προεκπαιδευμένο μοντέλο.

Σε αυτή την εργασία, αξιοποιούμε και επεκτείνουμε τη μέθοδο εξαγωγής αποκλίνοντων χαρακτηριστικών (exceptional feature extraction) από το PIECE [54], η οποία επιτρέπει μια μη εποπτευόμενη προεκπαίδευση του μοντέλου παραγωγής αντιπαραδειγματικών εικόνων και απαιτεί μόνο τον υπολογισμό ορισμένων στατιστικών στοιχείων για τις ενεργοποιήσεις του προτελευταίου επιπέδου του ταξινομητή. Αυτή η τεχνική επίσης δεν θέτει περιορισμούς στον αριθμό κλάσεων εξόδου του ταξινομητή και είναι ικανή να παράγει αντιπαραδειγματικές εικόνες από οποιαδήποτε κλάση σε οποιαδήποτε άλλη, κάτι που δεν ισχύει για πολλές άλλες εργασίες.

Επομένως, στην παρούσα διπλωματική προτείνουμε ένα νέο πλαίσιο, το SPRUCE (SParse Realistic and Uncoupled Counterfactual Explanations). Η συνεισφορά μας συνοψίζεται ως εξής:

- Εισάγουμε το SPRUCE, ένα νέο πλαίσιο για τη δημιουργία αραιών αντιπαραδειγματικών ιατρικών εικόνων.
- Αναπτύσσουμε μια εξειδικευμένη συνάρτηση απώλειας που επιβάλλει τόσο την αραιότητα όσο και την πιστότητα εικόνας στις παραγόμενες επεξηγηματικές, ελαχιστοποιώντας τροποποιήσεις που δεν σχετίζονται με την εκάστοτε πάθηση.
- Χρησιμοποιούμε Παραγωγικά Ανταγωνιστικά Δίκτυα (ΠΑΔ), σε συνδυασμό με τις μεθόδους Encoder for Editing (E4E) [117] και pivotal tuning [96], για να επιτύχουμε ακριβή ανακατασκευή και παραγωγή καθαρών ρεαλιστικών εικόνων, καθώς και για να ενισχύσουμε τον έλεγχο και την δυνατότητα επεξεργασίας στον λανθάνοντα χώρο του ΠΑΔ.
- Αξιοποιούμε μια τεχνική από το PIECE [54] που αποσυνδέει την εκπαίδευση του αντιπαραδειγματικού ερμηνευτή από την εκπαίδευση του ταξινομητή και μας επιτρέπει να χρησιμοποιούμε τον ίδιο επεξηγητή για οποιονδήποτε δυαδικό ή πολλαπλών κλάσεων ταξινομητή που έχει εκπαιδευτεί στην ίδια κατανομή δεδομένων.
- Διαπιστώνουμε ότι η ικανότητα δημιουργίας αραιών, ρεαλιστικών αντιπαραδειγματικών εικόνων συνδέεται άμεσα με την ευρωστία (robustness) του ταξινομητή, καθιστώντας το πλαίσιο μας ένα πολύτιμο εργαλείο για τον εντοπισμό αδυναμιών, αλλά και για την επιθεώρηση μοντέλων.

1.2 Θεωρητικό Υπόβαθρο

1.2.1 Συνελικτικά Νευρωνικά Δίκτυα

Η ενότητα αυτή παρουσιάζει λεπτομερώς τα βασικά στοιχεία των Συνελικτικών Νευρωνικών Δικτύων (CNNs) και τις κύριες διεργασίες που τους επιτρέπουν να μαθαίνουν από μεγάλους όγκους δεδομένων εικόνας.

Ο Νευρώνας και οι Συναρτήσεις Ενεργοποίησης

Στην καρδιά κάθε Νευρωνικού Δικτύου βρίσκεται ο **νευρώνας**, εμπνευσμένος από τους βιολογικούς νευρώνες που επικοινωνούν μεταξύ τους μέσω ηλεκτρικών σημάτων. Κάθε νευρώνας σε ένα δίκτυο:

1. Δέχεται ένα μικρό τμήμα δεδομένων εισόδου,
2. Υπολογίζει το σταθμισμένο άθροισμα συν μία σταθερά (bias),
3. Εφαρμόζει μια *συνάρτηση ενεργοποίησης*,
4. Συμβάλλει στη δημιουργία ενός χάρτη χαρακτηριστικών (feature map).

Όπως φαίνεται στο Σχήμα 1.2.1, η έξοδος του νευρώνα προωθείται στις επόμενες στρώσεις του δικτύου. Βασικό στοιχείο των συναρτήσεων ενεργοποίησης είναι η εισαγωγή μη-γραμμικότητας, ώστε το μοντέλο να μπορεί να μοντελοποιεί σύνθετα πρότυπα. Κάποιες συνηθισμένες συναρτήσεις ενεργοποίησης είναι οι εξής:

- **ReLU:** $f(x) = \max(0, x)$· απλή κι αποτελεσματική, ωστόσο μηδενίζει όλες τις αρνητικές τιμές.
- **Leaky ReLU:** Διατηρεί μια μικρή, μη μηδενική κλίση για αρνητικές εισόδους, μετριάζοντας το πρόβλημα των μη ενεργοποιημένων νευρώνων στην περίπτωση της ReLU.

- **Sigmoid:** Μετατρέπει τις τιμές εισόδου στο εύρος $(0, 1)$, χρήσιμη για πιθανότητες, αλλά μπορεί να κορεστεί επιβραδύνοντας την εκπαίδευση.
- **Tanh:** Παρόμοια με τη sigmoid, αλλά κεντράρεται στο μηδέν στο εύρος $(-1, 1)$· επίσης ευάλωτη σε κορεσμό των gradients.

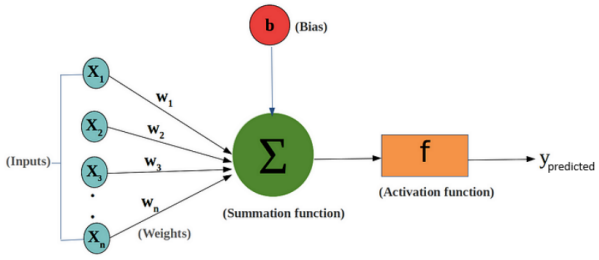


Figure 1.2.1: Σχηματική απεικόνιση της αρχιτεκτονικής ενός νευρώνα, όπου υπολογίζεται σταθμισμένο άθροισμα, προστίθεται bias, εφαρμόζεται συνάρτηση ενεργοποίησης και παράγεται η έξοδος του νευρώνα.

Νευρωνικά Δίκτυα και η Εκπαίδευσή τους

Τα Νευρωνικά Δίκτυα [60] αποτελούνται από πολλαπλές στρώσεις:

- *στρώμα εισόδου* (input layer) για τα αρχικά δεδομένα,
- διάφορα *κρυφά στρώματα* (hidden layers) για την εξαγωγή χαρακτηριστικών,
- ένα *στρώμα εξόδου* (output layer) που παράγει τις τελικές προβλέψεις.

Το σφάλμα ανάμεσα στην πρόβλεψη και την αληθινή τιμή μετριέται από μια **συνάρτηση απώλειας**—συνηθέστερα το μέσο τετραγωνικό σφάλμα (MSE) για παλινδρόμηση ή η συνάρτηση Cross-Entropy για ταξινόμηση—και ελαχιστοποιείται μέσω μιας διαδικασίας **βελτιστοποίησης** γνωστής ως Gradient Descent.

Αλγόριθμος Οπισθοδιάδοσης (Backpropagation) Η εκπαίδευση βασίζεται σε δύο φάσεις: "προώθηση" και "οπισθοδιάδοση". Στην προώθηση, τα δεδομένα ρέουν μέσα από το δίκτυο ώστε να παραχθεί η πρόβλεψη. Ακολουθεί η οπισθοδιάδοση, όπου υπολογίζονται τα gradients της συνάρτησης απώλειας ως προς κάθε παράμετρο (με τον κανόνα αλυσίδας) και ενημερώνονται τα βάρη. Το Σχήμα 1.2.2 αποτυπώνει αυτή τη διαδικασία μέσω της οποίας τα μοντέλα Τεχνητής Νοημοσύνης μπορούν και μαθαίνουν.

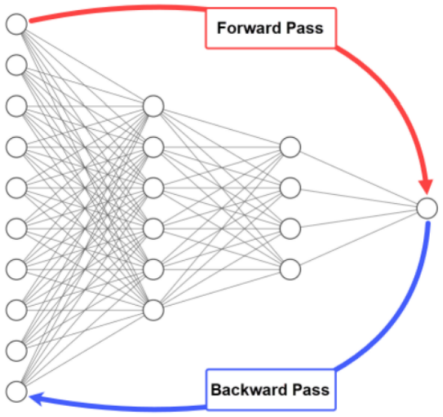


Figure 1.2.2: Φάση προώθησης (υπολογισμός εξόδων) και φάση οπισθοδιάδοσης (υπολογισμός gradients) για την ενημέρωση των βαρών.

Τεχνικές Κανονικοποίησης (Regularization)

Ένα βασικό πρόβλημα στη βαθιά μάθηση είναι η **υπερπροσαρμογή**, όπου το δίκτυο απομνημονεύει λεπτομέρειες του συνόλου εκπαίδευσης χωρίς να γενικεύει σε καινούρια δεδομένα. Για την αντιμετώπισή του, εφαρμόζονται διάφορες τεχνικές κανονικοποίησης:

- **L1/L2 Κανονικοποίηση**: Πρόσθετοι όροι που επιβάλλουν ποινή στα βάρη μεγάλου μεγέθους, αποθαρρύνοντας υπερβολικό μέγεθος ή πλήθος παραμέτρων.
- **Dropout [111]**: Τυχαία "απενεργοποίηση" νευρώνων κατά την εκπαίδευση για να αποφευχθεί η συνεξάρτηση (co-adaptation).
- **Επαύξηση Δεδομένων (Data Augmentation)**: Τεχνική μέσω της οποίας δημιουργούνται επιπλέον παραδείγματα εκπαίδευσης τροποποιώντας τα αρχικά δεδομένα εκπαίδευσης (π.χ., περιστροφή, αναστροφή, κοπή εικόνων).
- **Ρύθμιση Μεγέθους Παρτίδας (Batch Size)**: Υπερπαράμετρος που καθορίζει πόσα δείγματα επεξεργάζονται σε ένα βήμα ενημέρωσης βαρών.

Βασικά Στοιχεία Συνελίξεων (Convolutional Layer Essentials)

Η **συνέλιξη** [31] αποτελεί τον πυρήνα των CNNs, όπου μικροί *πυρήνες (kernels)* σαρώνονται στις χωρικές διαστάσεις της εικόνας για την εξαγωγή τοπικών χαρακτηριστικών. Όπως φαίνεται στο Σχήμα 1.2.3, κάθε πυρήνας έχει *πλάτος*, *ύψος* και *βάθος*, επισημαίνοντας χαρακτηριστικά όπως ακμές ή υφές.

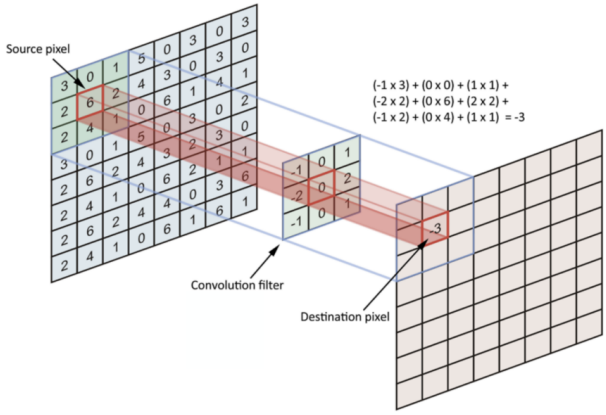


Figure 1.2.3: Ενδεικτική εφαρμογή πυρήνα συνέλιξης σε μια εικόνα (από [10]).

Τα παραδοσιακά CNNs αποτελούνται από ένα συνδυασμό **συνελικτικών φίλτρων**, **στρωμάτων υποδειγματοληψίας (pooling)** και **Πλήρως Συνδεδεμένων Στρωμάτων**, όπως στο Σχήμα 1.2.4.

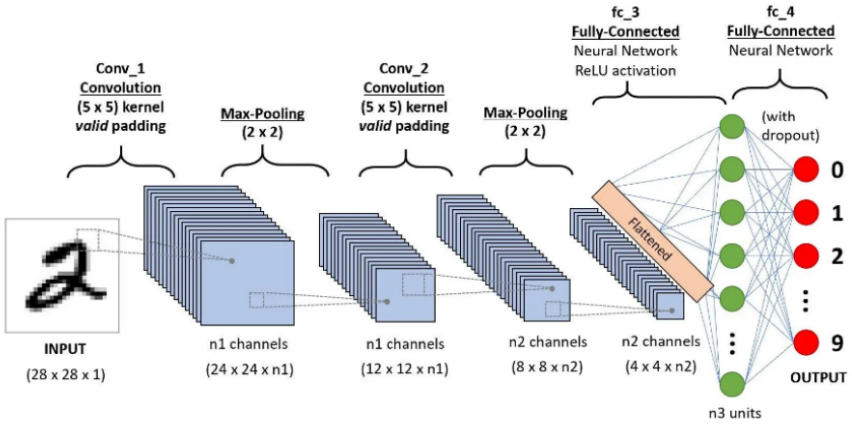


Figure 1.2.4: Τυπική αρχιτεκτονική CNN με διαδοχικές συνελίξεις, υποδειγματοληψία και πλήρως συνδεδεμένες στρώσεις.

Στρώμα Υποδειγματοληψίας (Pooling Layer) Το στρώμα υποδειγματοληψίας (π.χ. max pooling) μειώνει τη χωρική ανάλυση των χαρακτηριστικών, ελαττώνοντας τον υπολογιστικό φόρτο και επιτυγχάνοντας μεταφραστική αμεταβλητότητα (translational invariance). Οι υπερπαράμετροι *stride* και *padding* καθορίζουν τον τρόπο με τον οποίο ο πυρήνας ολισθαίνει και αντιμετωπίζει τα άκρα της εικόνας.

Κανονικοποίηση Κατά Παρτίδες (Batch Normalization) Για τη σταθεροποίηση της εκπαίδευσης, η **Κανονικοποίηση Κατά Παρτίδες** προσαρμόζει στατιστικά τις ενεργοποιήσεις (μηδενική μέση τιμή και μοναδιαία διασπορά) σε κάθε παρτίδα, βελτιώνοντας συχνά τόσο την ταχύτητα σύγκλισης όσο και την τελική ακρίβεια.

Πλήρως Συνδεδεμένο Στρώμα (Fully Connected Layer) Στο πλήρως συνδεδεμένο στρώμα, οι νευρώνες έχουν πλήρη συνδεσιμότητα με όλους τους νευρώνες στο προηγούμενο και το επόμενο στρώμα, όπως φαίνεται στα κανονικά FCNN. Αυτός είναι ο λόγος που μπορεί να υπολογιστεί με τον συνηθισμένο τρόπο μέσω ενός πολλαπλασιασμού μητρών, ακολουθούμενου από μια επίδραση μετατόπισης (bias). Το στρώμα FC βοηθά στην αντιστοίχιση της αναπαράστασης μεταξύ της εισόδου και της εξόδου.

1.2.2 Μοντέλο ConvNeXt

Σε αυτή την ενότητα παρουσιάζεται αναλυτικά η αρχιτεκτονική του ConvNeXt, ένα «εκσυγχρονισμένο» Συνελικτικό Νευρωνικό Δίκτυο (CNN) που αντλεί ιδέες από τους Vision Transformers (ViTs), διατηρώντας ωστόσο τα βασικά γνωρίσματα των παραδοσιακών ταξινομητών τύπου ResNet. Στόχος είναι να επιτευχθεί ή να ξεπεραστεί η απόδοση των Transformers σε μεγάλης κλίμακας ταξινομήσεις εικόνων, με διατήρηση της ερμηνευσιμότητας και της αποδοτικής χρήσης υπολογιστικών πόρων.

Εισαγωγή στο ConvNeXt

Το μοντέλο **ConvNeXt** αποτελεί μια πρόσφατη παραλλαγή CNN, η οποία ενσωματώνει δομικές επιλογές από τους ιεραρχικούς Transformers (π.χ. Swin), χωρίς όμως να βασίζεται αποκλειστικά στον μηχανισμό self-attention. Αντιθέτως, εκσυγχρονίζει τα συνελικτικά στρώματα (convolutional layers) ώστε να επιτυγχάνει συγκρίσιμη επίδοση, συνδυάζοντας την αποδοτικότητα των ConvNets με ορισμένα πλεονεκτήματα των μοντέλων τύπου Transformer.

Κύριες Αρχιτεκτονικές Καινοτομίες

Το ConvNeXt βασίζεται σε μια δομή παρόμοια με αυτή των ResNet, πλαισιωμένη από ιδέες που προέρχονται από Vision Transformers. Οι βελτιώσεις μπορούν να χωριστούν σε:

- **Αλλαγές Macro-επιπέδου (Macro-level):** Υψηλού επιπέδου τροποποιήσεις στη συνολική δομή του δικτύου (π.χ. stem, patchification, multi-stage σχεδίαση).
- **Αλλαγές Μικρο-επιπέδου (Micro-level):** Επεμβάσεις μέσα σε κάθε residual ή bottleneck block (π.χ. ανεστραμμένα bottlenecks, επιλογές συναρτήσεων ενεργοποίησης).

Αλλαγές Μακρο-επιπέδου

(a) Ιεραρχική Πολυσταδιακή Σχεδίαση (Hierarchical Multi-Stage Design) Το ConvNeXt υιοθετεί ιεραρχική δομή παρόμοια με τυπικά CNNs, αλλά αναθεωρεί τον αριθμό των block ανά στάδιο σε $(3, 3, 9, 3)$, μιμούμενο τη διανομή block σε ιεραρχικούς Transformers (π.χ. Swin). Έτσι εξασφαλίζει κλιμακούμενη αρχιτεκτονική, η οποία συλλαμβάνει αποτελεσματικά τόσο χαμηλόσυχνες όσο και υψηλόσυχνες πληροφορίες.

(b) Patchified Convolutional Stem Αντί για το κλασικό `7x7` convolution και max pooling (όπως στα πρώιμα ResNet), το ConvNeXt αντικαθιστά τις αρχικές στρώσεις με μια διεργασία *patchify*—συγκεκριμένα, ένα `4x4` convolution με stride 4, παρόμοιο με το patch tokenization των ViTs. Έτσι περιορίζεται το αρχικό υπολογιστικό κόστος, ενώ παράλληλα δημιουργούνται «patch embeddings» που τροφοδοτούν τα επόμενα στάδια του δικτύου.

(c) Μεγαλύτεροι Πυρήνες Συνελίξεων Το ConvNeXt παύει να περιορίζεται σε `3x3` πυρήνες, επιλέγοντας `7x7` depthwise convolutions. Η διευρυμένη αυτή «περιοχή αντίληψης» (receptive field) μοιάζει περισσότερο με την ευρεία χωρική κάλυψη του self-attention, επιτρέποντας στο δίκτυο να ενσωματώνει χαρακτηριστικά σε μεγαλύτερα τμήματα της εικόνας από νωρίς.

(d) Τροποποιήσεις Εμπνευσμένες από ResNeXt Βασιζόμενο στην ιδέα των grouped convolutions του ResNeXt, το ConvNeXt περνά σε *depthwise* συνελίξεις, όπου κάθε κανάλι επεξεργάζεται ανεξάρτητα. Ο διαχωρισμός χωρικής και καναλικής διάστασης (depthwise + `1x1` pointwise layers) μειώνει τα FLOPs και εναρμονίζεται με τον «διαχωρισμό» καναλιών/χώρου (channel/spatial) των Transformers.

(e) Ανεστραμμένα Bottlenecks (Inverted Bottlenecks) Παρόμοια με το MobileNetV2, το ConvNeXt χρησιμοποιεί ανεστραμμένη δομή bottleneck, όπου η διάσταση των χαρακτηριστικών διευρύνεται κατά 4 φορές πριν από τη συνέλιξη. Αυτή η προσέγγιση θυμίζει τις feed-forward επεκτάσεις στα ViTs και συνιστά πιο εκφραστικό τοπικό μετασχηματισμό.

Αλλαγές Μικρο-επιπέδου

(a) Χρήση GELU αντί ReLU Το ConvNeXt αντικαθιστά τη συνηθισμένη ReLU με την πιο ομαλή **GELU** (Gaussian Error Linear Unit), διαδεδομένη στα ViTs. Η αλλαγή αυτή βελτιώνει τη ροή των gradient (gradient flow), μειώνει την πιθανότητα «νεκρών»(μη ενεργοποιημένων) νευρώνων και μπορεί να αποδώσει καλύτερη απόκριση σε ευαίσθητες μεταβολές χαρακτηριστικών. Στο Σχήμα 1.2.5 αποτυπώνεται αυτή η σχεδιαστική επιλογή.

(b) Λιγότερες Ενεργοποιήσεις και Κανονικοποιήσεις Ενώ σε πολλές CNN αρχιτεκτονικές ακολουθεί μια συνάρτηση ενεργοποίησης μετά από κάθε συνέλιξη, στο ConvNeXt χρησιμοποιείται συχνά μία *GELU* ανά block. Επιπλέον, βελτιστοποιείται η χρήση της Κανονικοποίησης κατά παρτίδες, ενώ συχνά προτιμάται η *Κανονικοποίηση Κατά Στρώμα (Layer Normalization)*.

(c) Κανονικοποίηση Κατά Στρώμα (Layer Normalization) Ως μέρος της προσπάθειας προσέγγισης της ροής εργασιών τύπου Transformer, στο ConvNeXt αντικαθίστανται συχνά η Batch Normalization με **Layer Normalization**, εφαρμόζοντας την κανονικοποίηση σε επίπεδο στρώματος μέσα σε κάθε residual block. Η μέθοδος αυτή στοχεύει στη μείωση της ευαισθησίας ως προς το μέγεθος παρτίδας, διατηρώντας παράλληλα σταθερή εκπαίδευση.

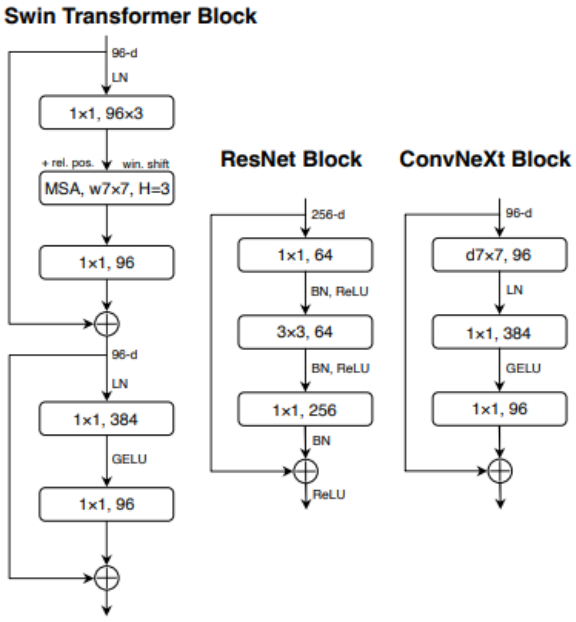


Figure 1.2.5: Σύγκριση σχεδίασης block ανάμεσα σε ResNet, Swin Transformer και ConvNeXt. Διακρίνονται τα ανεστραμμένα bottlenecks και η LayerNorm στο ConvNeXt.

Σύγκριση με Παραδοσιακά CNNs

Στον Πίνακα 1.1 φαίνονται οι κυριότερες διαφορές μεταξύ «κλασικών» δικτύων τύπου ResNet και της εκσυγχρονισμένης προσέγγισης του ConvNeXt.

Χαρακτηριστικό	Παραδοσιακά CNNs	ConvNeXt
Τύπος Συνελίξεων	Standard Convs	Depthwise Separable
Κανονικοποίηση	BatchNorm	LayerNorm
Ενεργοποίηση	ReLU	GELU
Μείωση Διαστάσεων	Max Pooling / Strided Conv	Patchified Convolution
Residual Blocks	Bottleneck Residuals	Ανεστραμμένα Bottlenecks

Table 1.1: Βασικές αντιστοιχίες μεταξύ τυπικών CNN και της εκσυγχρονισμένης αρχιτεκτονικής ConvNeXt.

1.2.3 Παραγωγικά Ανταγωνιστικά Δίκτυα (ΠΑΔ)

Η ενότητα αυτή παρουσιάζει μια ολοκληρωμένη ανασκόπηση των Παραγωγικών Ανταγωνιστικών Δικτύων (ΠΑΔ), εστιάζοντας στην εξέλιξή τους από τις βασικές αρχιτεκτονικές έως το StyleGAN2-ADA, το οποίο αντιμετωπίζει τις προκλήσεις έλλειψης δεδομένων που χαρακτηρίζουν τον χώρο της ιατρικής απεικόνισης. Επιπλέον, καλύπτεται η έννοια της *αναστροφής ΠΑΔ (GAN inversion)*, μια κρίσιμη διαδικασία για την ενσωμάτωση πραγματικών εικόνων στον λανθάνοντα χώρο, επιτρέποντας προηγμένες εφαρμογές όπως η παραγωγή αντιπαραδειγματικών εικόνων και η σημασιολογική επεξεργασία.

Επισκόπηση των Παραγωγικών Ανταγωνιστικών Δικτύων

Τα ΠΑΔ προτάθηκαν από τους Goodfellow et al. [32].Σε ένα **Παραγωγικό Ανταγωνιστικό Δίκτυο (GAN)** εμπλέκονται δύο νευρωνικά δίκτυα—*ένα παραγωγικό (generator)* και ένα *διαχωριστικό (discriminator)*—που εκπαιδεύονται με ανταγωνιστικό τρόπο. Ο *generator* G συνθέτει υποψήφια δείγματα που μοιάζουν ρεαλιστικά, ενώ ο *discriminator* D προσπαθεί να διακρίνει εάν ένα δείγμα είναι πραγματικό ή παραγόμενο. Μαθηματικά, η εκπαίδευση διατυπώνεται ως ένα min-max game:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]. \tag{1.2.1}$$

Τυπική Αρχιτεκτονική (Vanilla). Στο βασικό σενάριο, ο *generator* δέχεται ως είσοδο έναν τυχαίο λανθάνον κώδικα z (συνήθως από κανονική κατανομή) και παράγει συνθετικά δεδομένα $G(z)$. Ο *discriminator* εκτιμά την πιθανότητα ένα δείγμα να είναι πραγματικό. Το Σχήμα 1.2.6 δείχνει αυτή τη θεμελιώδη σχεδίαση.

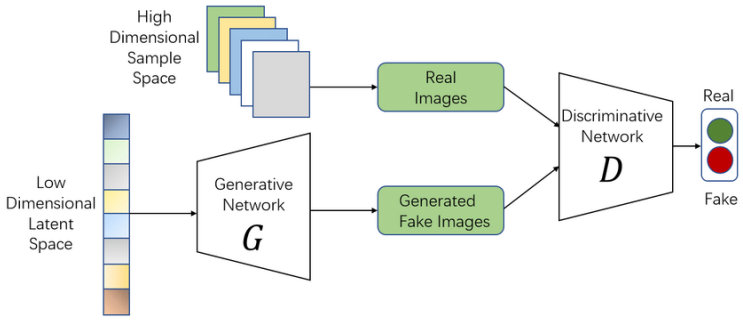


Figure 1.2.6: Βασική αρχιτεκτονική ενός Vanilla GAN, με παραγωγικό δίκτυο G και διαχωριστικό D.

Παραλλαγές ΠΑΔ (GAN Variants)

Από τη σύλληψή τους μέχρι σήμερα, πολλές **παραλλαγές των ΠΑΔ** έχουν προταθεί για την αντιμετώπιση της αστάθειας εκπαίδευσης, του *mode collapse* και της περιορισμένης ποικιλίας δειγμάτων:

- **DCGAN:** Εισήγαγε συνελικτικές στρώσεις τόσο στο G όσο και στο D, βελτιώνοντας την ποιότητα εικόνων [93].
- **WGAN:** Αντικατέστησε την απόκλιση Jensen–Shannon με την απόσταση Wasserstein, σταθεροποιώντας την εκπαίδευση και μειώνοντας το mode collapse [3].
- **CycleGAN:** Πραγματοποίησε αζευγάρωτη μετάφραση εικόνας-προς-εικόνα (unpaired image-to-image translation) μέσω επιβολής *cycle consistency* [143].
- **PGGAN:** Επέτρεψε παραγωγή εικόνων υψηλότερης ανάλυσης προσθέτοντας σταδιακά τα επίπεδα ανάλυσης κατά την εκπαίδευση, βελτιώνοντας τη σταθερότητα και τις λεπτομέρειες [51].
- **StyleGAN / StyleGAN2 / StyleGAN2-ADA:** Εστίασε σε *style-based* σύνθεση για πιο λεπτομερή έλεγχο και καλύτερο αποδιαχωρισμό (disentanglement) του λανθάνοντος χώρου. Το StyleGAN2-ADA πρόσθεσε *adaptive data augmentation* ώστε να υποστηρίζει μικρότερα σύνολα δεδομένων [52, 53].

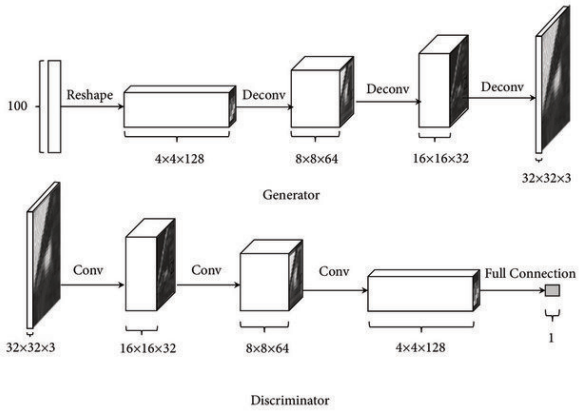


Figure 1.2.7: Παράδειγμα αρχιτεκτονικής DCGAN από [140].

Εφαρμογές ΠΑΔ. Τα ΠΑΔ αξιοποιούνται ευρέως στην ιατρική απεικόνιση (π.χ. επαύξηση δεδομένων, ανίχνευση ανωμαλιών), στην όραση υπολογιστή (π.χ. super-resolution, domain adaptation) και σε πλήθος άλλων περιπτώσεων όπου χρειάζονται συνθετικά και ταυτόχρονα ρεαλιστικά δεδομένα.

StyleGAN και StyleGAN2: Μεθοδολογία Βασισμένη στο "Style"

Το **StyleGAN** [49] εισήγαγε έναν παραγωγικό πυρήνα που επεξεργάζεται έναν ενδιάμεσο λανθάνον κώδικα $\mathbf{w}$ και εφαρμόζει *adaptive instance normalization* (AdaIN) για τον έλεγχο οπτικών γνωρισμάτων. Αυτό επέτρεψε:

- Καλύτερο αποδιαχωρισμό τροποποιήσεων (π.χ. ανεξάρτητος έλεγχος πόζας και υφής),
- Δείγματα υψηλής ποιότητας σε πρωτοφανείς αναλύσεις,
- *Style mixing*, ώστε να αυξηθεί η ποικιλία των παραγόμενων εξόδων.

Βελτιώσεις στο StyleGAN2. Παρότι το StyleGAN βελτίωσε τον έλεγχο του στυλ, εμφάνιζε τεχνουργήματα(artifacts) και σύμπλεξη μη σχετικών χαρακτηριστικών. Το *StyleGAN2* αντιμετώπισε τα ζητήματα αυτά με:

- **Weight modulation/demodulation:** Αντικατέστησε τις ρητές στρώσεις AdaIN, περιορίζοντας επαναλαμβανόμενα τεχνουργήματα.
- **Εκπαίδευση σε σταθερή ανάλυση:** Αφαίρεσε το progressive growing για μια απλούστερη αλλά αποτελεσματική διαδικασία εκμάθησης.
- **Skip & residual συνδέσεις:** Διευκόλυνε τη ροή των gradients και σταθεροποίησε την εκπαίδευση.
- **Path length regularization:** Προέτρεψε συνεπή συμπεριφορά στις παρεμβολές λανθάνοντος κώδικα, μειώνοντας απότομες μεταβάσεις.

Πειράματα που συνέκριναν *StyleGAN* και *StyleGAN2* σε σύνολα όπως τα FFHQ ή LSUN cats/horses επιβεβαίωσαν μειωμένη Fréchet Inception Distance (FID) και χαμηλότερο Perceptual Path Length (PPL), υποδηλώνοντας υψηλότερη πιστότητα και πιο ομαλές τροποποιήσεις στον λανθάνοντα χώρο.

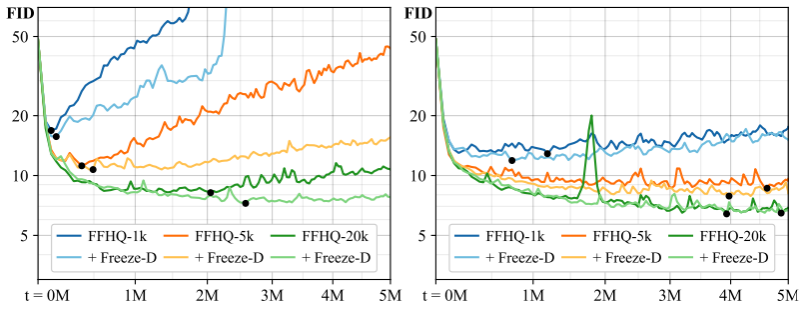
StyleGAN2-ADA: Προσαρμοστική Ενίσχυση Δεδομένων στον Discriminator

Το *StyleGAN2* απαιτεί μεγάλα σύνολα δεδομένων για σταθερή απόδοση. Ωστόσο, σε πολλούς τομείς—ιδίως στην *ιατρική απεικόνιση*—τα δεδομένα είναι περιορισμένα. Το **StyleGAN2-ADA** [53] υπερνικά αυτό το μειονέκτημα εισάγοντας ένα *προσαρμοστικό σχήμα επαύξησης* (adaptive augmentation) που:

- Εφαρμόζει «μη διαρρεύσιμες» επαυξήσεις (περιστροφές, αναστροφές, αλλοιώσεις χρώματος) *μόνο* στην είσοδο του discriminator,
- Ρυθμίζει δυναμικά την ένταση των επαυξήσεων ανάλογα με το επίπεδο υπερπροσαρμογής του D,
- Διατηρεί την κατανομή που μαθαίνει ο generator, χωρίς να «μολύνεται» από τις επαυξήσεις.

Οφέλη στην Ιατρική Απεικόνιση.

1. **Ποιοτική σύνθεση με λίγα δεδομένα:** Πολύτιμο για εξειδικευμένες ιατρικές περιοχές με λίγες εκατοντάδες δειγματοληπτικά δεδομένα.
2. **Δημιουργία αντιπαραδειγματικών εικόνων:** Ένας επεξεργάσιμος λανθάνων χώρος ωφελεί στη δημιουργία αντιπαραδειγματικών εικόνων που μπορεί να διευκολύνουν την διάγνωση ή τον σχεδιασμό θεραπείας.
3. **Διαμοιρασμός δεδομένων με διασφάλιση ιδιωτικότητας:** Φορείς μπορούν να μοιράζονται ρεαλιστικές συνθετικές εικόνες με ελάχιστο κίνδυνο αποκάλυψης πραγματικών πληροφοριών ασθενών.



(a) Without ADA (b) With ADA

Figure 1.2.8: Μεταφορά μάθησης στο StyleGAN2-ADA: σύγκριση σύγκλισης ενός προεκπαιδευμένου μοντέλου (π.χ. FFHQ) και προσαρμογής σε άλλη περιοχή (CELEBA-HQ).

Αναστροφή ΠΑΔ (GAN Inversion): Σύνδεση Πραγματικών Εικόνων με Λανθάνοντες Κώδικες

Πολλές σύνθετες εφαρμογές απαιτούν **αναστροφή ΠΑΔ (GAN inversion)**—δηλ. την εύρεση λανθάνοντος κώδικα z^* ώστε η παραγόμενη εικόνα $G(z^*)$ να αναπαράγει πιστά μια πραγματική εικόνα x. Η επιτυχής αναστροφή δίνει δυνατότητα για:

- **Σημασιολογική επεξεργασία (semantic editing)**: Ανεξάρτητος έλεγχος συγκεκριμένων χαρακτηριστικών σε πραγματικές εικόνες,
- **Παραγωγή αντιπαραδειγματικών**: Ελάχιστες στοχευμένες αλλαγές που αποκαλύπτουν πώς μικρές τροποποιήσεις επηρεάζουν την ταξινόμηση,
- **Ανάλυση λανθάνοντος χώρου**: Κατανόηση τυχόν μεροληψιών του μοντέλου.

Μέθοδοι Αναστροφής. Συνήθεις προσεγγίσεις:

1. **Βασισμένες σε βελτιστοποίηση (optimization-based)**: Επαναληπτική ελαχιστοποίηση της $\mathcal{L}(G(z), x)$ ως προς z, επιτυγχάνοντας υψηλή πιστότητα, αλλά πιο αργή σύγκλιση.
2. **Με χρήση αποκωδικοποιητή (encoder-based)**: Εκπαιδεύεται ένας encoder E που αντιστοιχίζει άμεσα τις εικόνες σε λανθάνοντα κώδικα, εξασφαλίζοντας γρήγορη πρόβλεψη, όμως συχνά λιγότερο ακριβή.
3. **Υβριδικές Στρατηγικές**: Αρχικοποίηση από έναν encoder και στη συνέχεια λεπτομερή βελτιστοποίηση, επιτυγχάνοντας ισορροπία ταχύτητας-ακρίβειας.

Βασικές προκλήσεις παραμένουν η *διατήρηση ταυτότητας*, η κάλυψη δειγμάτων *εκτός κατανομής* και η *διατήρηση επεξεργασιμότητας* (editability) του λανθάνοντος κώδικα.

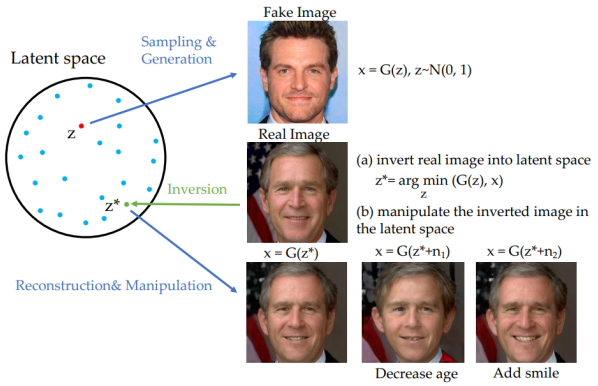


Figure 1.2.9: Ενδεικτική απεικόνιση της αναστροφής ΠΑΔ: αντιστοίχιση πραγματικής εικόνας x σε λανθάνοντα κώδικα z^*.

1.2.4 Ανταγωνιστική Εκπαίδευση και Ευρωστία

Η ενότητα αυτή αναδεικνύει μια κρισιμότατη πρόκληση στη βαθιά μάθηση: την ευπάθεια των νευρωνικών δικτύων σε **ανταγωνιστικά παραδείγματα**, δηλαδή εισόδους που έχουν υποστεί ελάχιστες αλλά «καλοσχεδιασμένες» διαταραχές, με σκοπό να ξεγελάσουν μοντέλα που υπό κανονικές συνθήκες εμφανίζουν υψηλή ακρίβεια. Τέτοιες αδυναμίες εγείρουν σοβαρούς προβληματισμούς σε πεδία υψηλής ασφαλείας, όπως η αυτόματη οδήγηση οχημάτων,η υγειονομική περίθαλψη και η κυβερνοασφάλεια, καθιστώντας την **ανταγωνιστική ανθεκτικότητα** θεμελιώδη προτεραιότητα έρευνας. Ακολουθεί μια πιο λεπτομερής επισκόπηση των βασικών εννοιών, των κυριότερων στρατηγικών άμυνας και των ανοιχτών προκλήσεων γύρω από την ανταγωνιστική ευρωστία.

Ανταγωνιστικά Παραδείγματα και Επιθέσεις

Ένα ανταγωνιστικό παράδειγμα προκύπτει μέσω προσθήκης μιας μικρής, συχνά αόρατης διαταραχής δ σε ένα έγκυρο δείγμα x, παράγοντας $x' = x + \delta$. Παρότι οι οπτικές αλλαγές είναι σχεδόν αόρατες, το νευρωνικό δίκτυο ταξινομεί λανθασμένα την είσοδο x'. Τυπικά, οι ανταγωνιστικές επιθέσεις στοχεύουν στη μεγιστοποίηση του σφάλματος ταξινόμησης:

$$\max_{\delta} \mathcal{L}\big(f(x+\delta), y\big) \quad \text{με την προϋπόθεση} \quad \|\delta\|_p \leq \epsilon. \tag{1.2.2}$$

όπου το δ οριοθετείται από την μεταβλητή ϵ (σύμφωνα με κάποια νόρμα ℓ_p). Ανάλογα με τη γνώση του σχεδιαστή της επίθεσης για το μοντέλο, οι επιθέσεις κατηγοριοποιούνται ως:

- **White-box επιθέσεις:** Ο επιτιθέμενος έχει πλήρη πρόσβαση στην αρχιτεκτονική και στα βάρη του μοντέλου, επιτρέποντας μεθόδους όπως *Fast Gradient Sign Method (FGSM)* [33] ή *Projected Gradient Descent (PGD)* [73].

- **Black-box επιθέσεις:** Ο επιτιθέμενος διαθέτει μόνο δυνατότητα κλήσης (queries) στο μοντέλο, παρατηρώντας τις τελικές πιθανότητες ή τις ετικέτες εξόδου του. Με αξιοποίηση μεθόδων μετάδοσης (transfer-based) ή μεθόδων βάσει ερωτημάτων (query-based) [89], κατασκευάζει ανταγωνιστικές εισόδους.

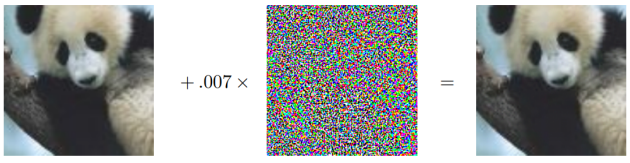


Figure 1.2.10: Δημιουργία ανταγωνιστικών παραδειγμάτων με ελάχιστα ορατές διαταραχές, από [33].

Ανταγωνιστική Ανθεκτικότητα

Η **ανταγωνιστική ανθεκτικότητα** (adversarial robustness) περιγράφει την ικανότητα ενός μοντέλου να διατηρεί υψηλή επίδοση κάτω από ανταγωνιστικές διαταραχές. Η έρευνα περιλαμβάνει τόσο θεωρητικά όρια (π.χ. για συγκεκριμένες νόρμες ή αρχιτεκτονικές) όσο και εμπειρικές μεθόδους (π.χ. ανταγωνιστική εκπαίδευση). Η αξιολόγηση συνήθως αφορά τη μέτρηση της ακρίβειας του μοντέλου για μια σειρά τιμών ϵ. Ένας ανθεκτικός ταξινομητής διατηρεί υψηλή ακρίβεια ακόμη και σε σχετικά μεγάλες τιμές ϵ.

Ανταγωνιστική Εκπαίδευση

Μεταξύ των πιο γνωστών στρατηγικών άμυνας συγκαταλέγεται η **ανταγωνιστική εκπαίδευση**, όπου παράγονται ανταγωνιστικά παραδείγματα και προστίθενται στο σύνολο εκπαίδευσης, ουσιαστικά «εκπαιδεύοντας» το μοντέλο να τα αντιμετωπίζει. Στοχεύουμε να επιλύσουμε:

$$\min_{\theta} \max_{\|\delta\| \leq \epsilon} \mathcal{L}\big(f(x+\delta;\theta), y\big), \tag{1.2.3}$$

με θ τις παραμέτρους του μοντέλου. Σε κάθε βήμα, δημιουργούνται ανταγωνιστικά παραδείγματα και ενημερώνονται τα βάρη. Παραλλαγές της μεθόδου είναι:

- **FGSM-based training**: Εφαρμογή ενός βήματος gradient για τη δημιουργία διαταραχών [33].
- **PGD-based training**: Διαδικασία πολλών βημάτων (*Projected Gradient Descent*) για πιο ισχυρές επιθέσεις, χρησιμοποιείται συχνά στην εκπαίδευση με χρήση της ℓ_∞- νόρμας[73].
- **Εκπαίδευση με αναβαθμιζόμενη αυστηρότητα (curriculum)**: Ξεκινά με μικρό ϵ και το αυξάνει σταδιακά, ώστε το μοντέλο να προσαρμόζεται ομαλότερα.

Άλλες Στρατηγικές Άμυνας

Πέρα από την ανταγωνιστική εκπαίδευση, έχουν προταθεί διάφορες ακόμα άμυνες:

- **Defensive distillation:** Εκπαίδευση δευτερεύοντος μοντέλου πάνω σε «εξομαλυνμένες» (softened) εξόδους του αρχικού δικτύου, με στόχο την απόκρυψη των πραγματικών gradients [88].
- **Μετασχηματισμοί εισόδου:** Μετριάζουν τις ανταγωνιστικές διαταραχές εφαρμόζοντας λειτουργίες όπως συμπίεση JPEG ή μείωση βάθους bit [38].
- **Προσεγγίσεις τυχαιότητας:** Προσθέτουν τυχαιότητα είτε στα δεδομένα εισόδου είτε στα στρώματα του δικτύου, δυσκολεύοντας τον επιτιθέμενο να υπολογίσει σταθερή κατεύθυνση gradients.
- **Certified defenses:** Παρέχουν εγγυήσεις ανθεκτικότητας υπό συγκεκριμένες νόρμες, π.χ. μέσω randomized smoothing [17].

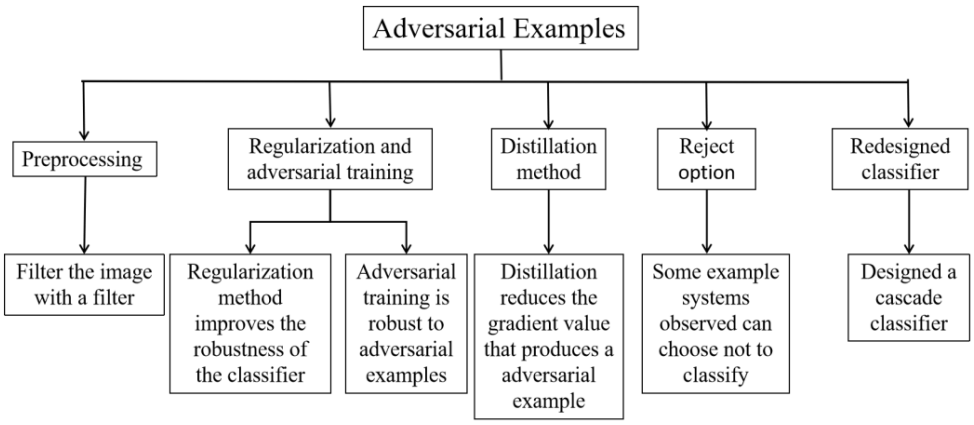


Figure 1.2.11: Κατηγοριοποίηση δημοφιλών μεθόδων άμυνας έναντι ανταγωνιστικών παραδειγμάτων, βασισμένη στο [15].

1.2.5 Ερμηνεύσιμη Τεχνητή Νοημοσύνη και Αντιπαραδειγματικές Εξηγήσεις

Η ενότητα αυτή διερευνά τα θεμέλια της **Ερμηνεύσιμης Τεχνητής Νοημοσύνης (XAI)** και την ολοένα αυξανόμενη σημασία των **αντιπαραδειγματικών εξηγήσεων (counterfactual explanations)**, ειδικά σε πεδία υψηλού ρίσκου όπως η ιατρική απεικόνιση. Η ανάλυση ξεκινά ορίζοντας τις έννοιες της ερμηνευσιμότητας και της επεξηγηματικότητας στη μηχανική μάθηση και στη συνέχεια δείχνουμε πώς οι αντιπαραδειγματικές προσεγγίσεις υπερβαίνουν τις παραδοσιακές τεχνικές (π.χ. χάρτες σημαντικότητας), παρέχοντας σαφείς, διαισθητικές υποδείξεις σχετικά με το *πώς* πρέπει να τροποποιηθεί μια είσοδος για να αλλάξει η απόφαση ενός μοντέλου.

Κίνητρα για Ερμηνεύσιμη Τεχνητή Νοημοσύνη

Τα σύγχρονα βαθιά νευρωνικά δίκτυα λειτουργούν συχνά σαν «μαύρα κουτιά», μοντελοποιώντας περίπλοκα πρότυπα που δεν είναι εύκολα κατανοητά από τον άνθρωπο [94, 98]. Ωστόσο, σε ιατρικές εφαρμογές είναι απαραίτητο τα αποτελέσματα ενός μοντέλου να είναι αξιόπιστα και ερμηνεύσιμα για τους κλινικούς ιατρούς. Γι' αυτόν τον λόγο, η έρευνα έχει εστιάσει σε στρατηγικές που αποκαλύπτουν τη διαδικασία λήψης αποφάσεων του μοντέλου, συνδυάζοντας την ακρίβεια πρόβλεψης με τη διαφάνεια.

Ερμηνευσιμότητα έναντι Επεξηγηματικότητας. Γενικά, η **ερμηνευσιμότητα** και η **επεξηγηματικότητα** αφορούν το πόσο εύκολα μπορεί κάποιος άνθρωπος να κατανοήσει τον λόγο για τον οποίο ένα μοντέλο κατέληξε σε μια συγκεκριμένη έξοδο [64]. Περιλαμβάνουν τόσο «εκ των υστέρων» (post-hoc) μεθόδους (π.χ. χάρτες σημαντικότητας) που αιτιολογούν τις αποφάσεις μετά την εκπαίδευση του μοντέλου, όσο και «εγγενώς ερμηνεύσιμα» (intrinsically interpretable) μοντέλα (π.χ. δέντρα απόφασης) των οποίων η δομή είναι εκ φύσεως διαφανής.

Συνηθισμένες Τεχνικές στη XAI

Στην όραση υπολογιστών, τεχνικές όπως η **Grad-CAM** και οι **χάρτες σημαντικότητας (saliency maps)** χρησιμοποιούνται ευρέως για τον εντοπισμό περιοχών μιας εικόνας που επηρεάζουν την έξοδο του δικτύου. Οι LIME [94] και SHAP [69] παρέχουν τοπικές προσεγγίσεις της συμπεριφοράς του μοντέλου. Παρότι όμως αυτές οι μέθοδοι αναδεικνύουν τα κρίσιμα εικονοστοιχεία, *δεν* προτείνουν τρόπο **αλλαγής** της εισόδου για διαφορετική απόφαση. Αυτό το κενό καλύπτουν οι **αντιπαραδειγματικές εξηγήσεις**.

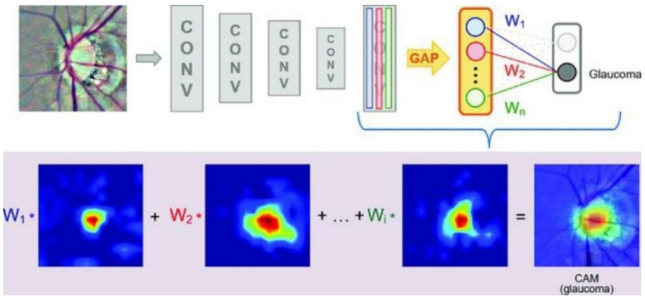


Figure 1.2.12: Εφαρμογή Grad-CAM σε διαγνωστική εικόνα, από [102].

Αντιπαραδειγματικές Εξηγήσεις

Οι **αντιπαραδειγματικές** εξηγήσεις στοχεύουν στο ερώτημα: *«Ποια ελάχιστη αλλαγή στην είσοδο θα άλλαζε την απόφαση του μοντέλου;»* [122]. Αντί να επισημαίνουν απλώς σημαντικές περιοχές, όπως οι χάρτες σημαντικότητας, οι αντιπαραδειγματικές προσεγγίσεις προτείνουν *εναλλακτικές εκδοχές* της εισόδου που αντιστρέφουν την έξοδο.Ο τρόπος λειτουργίας τους ταιριάζει με την ανθρώπινη λογική—αν μια μικρή μεταβολή αλλάζει τη διάγνωση από θετική σε αρνητική, τότε αυτό αποκαλύπτει τα καθοριστικά χαρακτηριστικά για την πρόβλεψη.

Βασικά Κριτήρια στην Ιατρική Απεικόνιση

Για να θεωρηθεί μια αντιπαραδειγματική εξήγηση κλινικά ουσιαστική [122, 98], πρέπει να διέπεται από τα παρακάτω χαρακτηριστικά:

- **Εγκυρότητα (Validity):** Να προκαλεί σαφή αλλαγή ετικέτας (π.χ. από καρκινικό σε μη καρκινικό).
- **Ρεαλισμός (Realism):** Να μοιάζει με αληθοφανή, ανατομικώς συνεπή εικόνα.
- **Εφικτότητα (Actionability):** Να αντιστοιχεί σε ρεαλιστικές ιατρικές επεμβάσεις (π.χ. μικρή μείωση μεγέθους μιας αλλοίωσης αντί για εξαφάνιση ενός ολόκληρου οργάνου).
- **Αραιότητα (Sparsity):** Να διατηρεί τις αλλαγές στο ελάχιστο, υπογραμμίζοντας ποια χαρακτηριστικά έχουν πραγματικά σημασία.
- **Αιτιατότητα (Causality):** Να τροποποιεί χαρακτηριστικά άμεσα σχετικά με την παθολογία, ώστε η αλλαγή πρόβλεψης να είναι ιατρικά ερμηνεύσιμη και όχι να προέρχεται από τυχαίες συσχετίσεις.

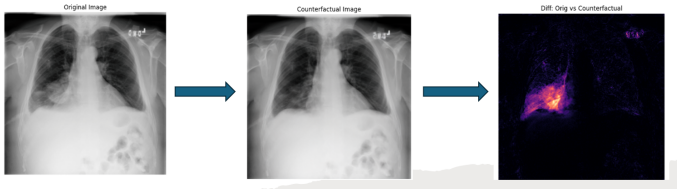


Figure 1.2.13: Παράδειγμα «υγιούς» αντιπαραδειγματικής ακτινογραφίας θώρακα, με τον αντίστοιχο χάρτη διαφορών.

Γιατί οι Αντιπαραδειγματικές Εξηγήσεις είναι Διαισθητικές

Οι άνθρωποι συχνά σκέφτονται με σενάρια «what if». Στον ιατρικό κλάδο, τέτοιοι υποθετικοί συλλογισμοί βοηθούν στην επιβεβαίωση ή αμφισβήτηση μιας διάγνωσης: «Αν αυτός ο όγκος ήταν στο μισό μέγεθος, θα τον θεωρούσε το μοντέλο μη καρκινικό;» Μια τέτοια συλλογιστική ενισχύει την επικοινωνία ανάμεσα στα εργαλεία ΤΝ και τους κλινικούς ειδικούς, αυξάνοντας την εμπιστοσύνη στη χρήση τους.

Προκλήσεις Υιοθέτησης

Παρά την αξία τους, οι αντιπαραδειγματικές εξηγήσεις είναι δύσκολες στην υλοποίηση για δεδομένα ιατρικής απεικόνισης για τους εξής λόγους:

- **Θέματα Ιδιωτικότητας:** Η δημιουργία ή επεξεργασία δεδομένων ασθενών πρέπει να συμμορφώνεται με αυστηρούς κανονισμούς (π.χ. HIPAA, GDPR).
- **Ρεαλιστική Επεξεργασία:** Οι επεμβάσεις υψηλής ανάλυσης, που διατηρούν την ανατομική συνέπεια, απαιτούν προηγμένα παραγωγικά μοντέλα και μεγάλη υπολογιστική ισχύ.
- **Κλινική Επικύρωση:** Οι ειδικοί οφείλουν να επιβεβαιώσουν ότι οι αλλαγές ανταποκρίνονται σε πραγματικές δυνατότητες και δεν δημιουργούν τεχνητά artifacts που θα μπορούσαν να παραπλανήσουν τη διάγνωση ή τον σχεδιασμό θεραπείας.

1.3 Μεθοδολογία

1.3.1 Εισαγωγή στο πλαίσιο SPRUCE

Σε αυτήν την ενότητα, εισάγουμε το **SPRUCE** (Sparse Realistic Uncoupled Counterfactual Explanations), ένα πλαίσιο για την παραγωγή ρεαλιστικών και ελάχιστα τροποποιημένων **αντιπαραδειγματικών εξηγήσεων** (counterfactual explanations) στο τομέα της ιατρικής απεικόνισης. Το SPRUCE στηρίζεται στις ιδέες του **PIECE**, μιας μεθόδου που εντοπίζει «αποκλίνοντα χαρακτηριστικά» (exceptional features) στο τελευταίο επίπεδο αναπαράστασης ενός **Νευρωνικού Δικτύου** προτού γίνει η τελική ταξινόμηση. Παρακάτω παρουσιάζεται μια λεπτομερής σύνοψη των κεντρικών ιδεών του PIECE και του τρόπου με τον οποίο το SPRUCE τις επεκτείνει.

Συνοπτική Παρουσίαση του PIECE

Το PIECE αρχικά δοκιμάστηκε σε σύνολα δεδομένων όπως το MNIST [21] και το CIFAR-10, επιδεικνύοντας την ικανότητά του να παράγει αληθοφανείς αντιπαραδειγματικές εξηγήσεις. Βασική αρχή του είναι ο εντοπισμός *πιθανώς σπάνιων* (δηλαδή «αποκλινόντων») χαρακτηριστικών σε μια δοσμένη εικόνα, και η αντικατάστασή τους με *αναμενόμενες* τιμές, όπως αυτές εμφανίζονται στην αντιπαραδειγματική κλάση c'.

Δύο Υποσυστήματα. Το PIECE προϋποθέτει:

- Έναν **ταξινομητή εικόνων** (π.χ. ένα **Συνελικτικό Νευρωνικό Δίκτυο**), που εξάγει τις πιθανότητες όλων των κλάσεων.
- Ένα **μοντέλο ΠΑΔ** (GAN) το οποίο ανακατασκευάζει ή συνθέτει εικόνες από λανθάνοντες κώδικες, επιτρέποντας την οπτικοποίηση των αλλαγών σε επίπεδο εικονοστοιχείων (pixel) για την παραγωγή αντιπαραδειγματικών εξηγήσεων.

Και τα δύο μοντέλα μπορούν να εκπαιδευτούν ανεξάρτητα, αρκεί τα σύνολα εκπαίδευσής τους να μοιράζονται την ίδια κατανομή.

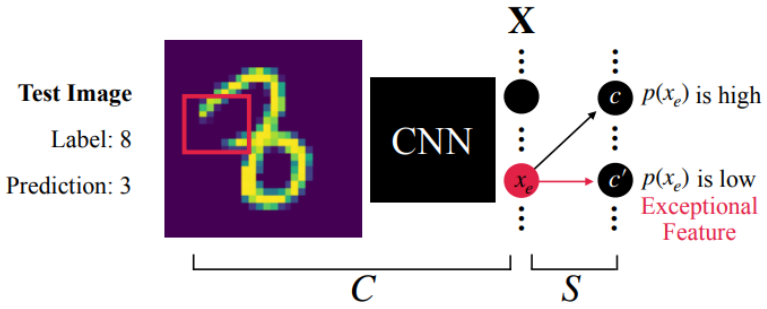


Figure 1.3.1: Η προσέγγιση του PIECE για τον εντοπισμό αποκλινόντων χαρακτηριστικών σε μια εικόνα(query image).

Βασικός Συμβολισμός

Θεωρούμε ένα **CNN** με δύο κύρια τμήματα:

- C: Όλα τα επίπεδα του δικτύου μέχρι το προτελευταίο επίπεδο χαρακτηριστικών, που εξάγει το **διάνυσμα χαρακτηριστικών** x.

- S: Το τελευταίο επίπεδο *softmax*, το οποίο ταξινομεί το x σε μια κατανομή πιθανοτήτων Y.

Έτσι, για μια εικόνα εισόδου I, η αναπαράσταση στο προτελευταίο επίπεδο είναι $x = C(I)$ και η έξοδος είναι $Y = S(C(I))$. Παράλληλα, ένας εκπαιδευμένος **generator** G (από ένα **Παραγωγικό Ανταγωνιστικό Δίκτυο**) αντιστοιχίζει έναν λανθάνοντα κώδικα z σε μια εικόνα $I = G(z)$.

Βασικά Βήματα του Αλγορίθμου PIECE.

1. **Αναστροφή ΠΑΔ (GAN inversion) και Επιλογή Αντιπαραδειγματικής Κλάσης:** Το PIECE πρώτα *αναστρέφει* την εικόνα I στον λανθάνοντα χώρο του GAN. Επιπλέον, προσδιορίζει την αντιπαραδειγματική κλάση c' — συχνά την πραγματική ετικέτα εάν ο ταξινομητής σφάλει ή κάποια άλλη «κοντινή» κλάση εάν η αρχική πρόβλεψη είναι σωστή.

2. **Εντοπισμός Αποκλινόντων Χαρακτηριστικών:** Χρησιμοποιώντας εκτιμήσεις των κατανομών ενεργοποίησης για την κλάση c', το PIECE εντοπίζει χαρακτηριστικά του x που αποκλίνουν σημαντικά (κάτω από ένα κατώφλι α) από την «τυπική» κατανομή. Μόνο αυτά θεωρούνται «αποκλίνοντα» και υπόκεινται σε αλλαγή.

3. **Αλλαγή των Αποκλινόντων Χαρακτηριστικών στις Αναμενόμενες Τιμές:** Ο αλγόριθμος τροποποιεί τις «παράταιρες» τιμές χαρακτηριστικών (π.χ. υπερβολικά υψηλές ή χαμηλές) ώστε να πλησιάζουν τις «φυσιολογικές» τιμές για την κλάση c'.Δηλαδή, η μέθοδος μεταβάλει την τιμή εκείνων των αποκλίνοντων χαρακτηριστικών που επηρεάζουν αρνητικά την ταξινόμηση της αρχικής εικόνας στην αντιπαραδειγματική κλάση c'

4. **Οπτικοποίηση στον Χώρο των Εικονοστοιχείων:** Τέλος, το τροποποιημένο διάνυσμα χαρακτηριστικών x' απεικονίζεται πάλι σε εικόνα I', μέσω μιας ακόμα αναστροφής ΠΑΔ. Αυτό παράγει μια «ελάχιστα τροποποιημένη» εικόνα με ουσιώδεις αλλαγές για την κλάση-στόχο.

Το Πλαίσιο SPRUCE

Το **SPRUCE** αξιοποιεί την ιδέα του *εντοπισμού αποκλινόντων χαρακτηριστικών* από το PIECE, αλλά την προσαρμόζει στις απαιτήσεις της ιατρικής απεικόνισης, δίνοντας έμφαση στην *αραιότητα* (sparsity) και στον *ρεαλισμό* του τελικού αποτελέσματος. Το Σχήμα 1.3.2 δείχνει το διάγραμμα ροής του πλαισίου, χωρισμένο σε τρία βασικά τμήματα.

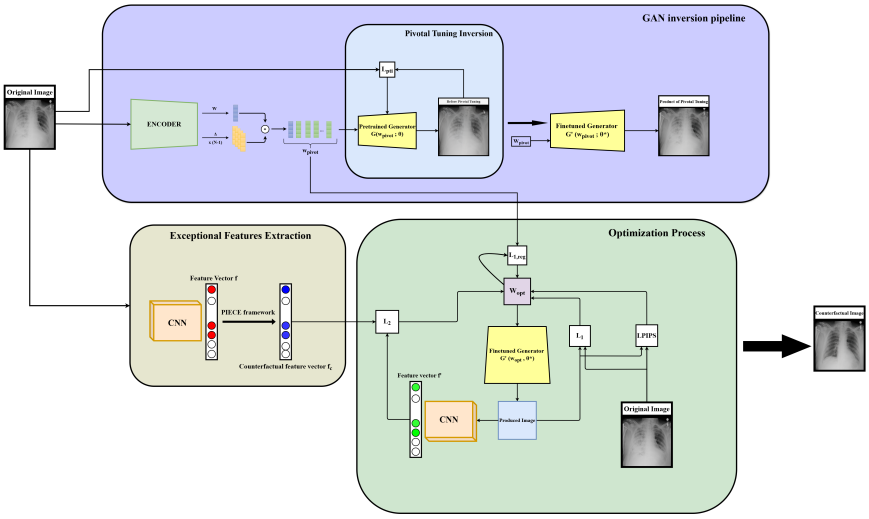


Figure 1.3.2: Ολοκληρωμένο διάγραμμα λειτουργίας του SPRUCE: από τη μοντελοποίηση χαρακτηριστικών, στην αναστροφή ΠΑΔ, μέχρι τη βελτιστοποίηση στον λανθάνοντα χώρο.

Βήμα 1: Εξαγωγή των Αποκλινόντων Χαρακτηριστικών

Για μια σωστά ταξινομημένη εικόνα I, το πλαίσιο δημιουργεί το «αντιπαραδειγματικό» διάνυσμα χαρακτηριστικών στον χώρο χαρακτηριστικών του ταξινομητή.

Βήμα 2: Αναστροφή ΠΑΔ

Στη συνέχεια, η αρχική εικόνα προβάλλεται στον λανθάνοντα χώρο ενός **StyleGAN2-ADA** μέσω ενός υβριδικού πλαισίου αναστροφής. Έτσι αποκτάμε μια επεξεργάσιμη λανθάνουσα αναπαράσταση που όταν δοθεί σαν είσοδο στον fine-tuned generator του ΠΑΔ ανακατασκευάζει πιστά την αρχική εικόνα.

Βήμα 3: Βελτιστοποίηση στον Λανθάνοντα Χώρο

Τέλος, χρησιμοποιείται μια εξειδικευμένη συνάρτηση βελτιστοποίησης (optimization objective) που επιβάλλει τον ρεαλισμό και την αραιότητα στην τελική επεξηγηματική εικόνα. Η βελτιστοποίηση του λανθάνοντος κώδικα οδηγεί σε μια εικόνα-αντιπαράδειγμα (I') που:

- Διαφέρει από την I κυρίως στα κρίσιμα σημεία που καθορίζουν την απόφαση του ταξινομητή,
- Διατηρεί υψηλή πιστότητα και συνέπεια με την κλάση c',
- Δεν αλλοιώνει αχρείαστα τις μη παθολογικές δομές.

1.3.2 Εξαγωγή Αποκλινόντων Χαρακτηριστικών

Για να εφαρμόσουμε την ιδέα του *exceptional feature extraction* από το PIECE [54] σε σύγχρονα μοντέλα ConvNeXt ακολουθούμε τα παρακάτω βήματα:

1. **Μοντελοποίηση των λανθανόντων χαρακτηριστικών** (Fitting the Latent Features' activations) ανά κλάση,
2. **Εντοπισμός αποκλινόντων χαρακτηριστικών** (identifying exceptional features) της αρχικής εικόνας I ως προς μια αντιπαραδειγματική κλάση c',

3. **Τροποποίηση μόνο των κρίσιμων outliers** που αναμένεται να ανατρέψουν την απόφαση του μοντέλου.

Συμβολισμός

Ας υποθέσουμε ότι F είναι ένας εκπαιδευμένος και «παγωμένος» (frozen) ταξινομητής τύπου ConvNeXt, ο οποίος δέχεται μια εικόνα I και επιστρέφει ένα διάνυσμα πιθανοτήτων $Y = F(I)$. Συμβολίζουμε:

- C: Το δίκτυο μέχρι το προτελευταίο επίπεδο,
- S: Το τελικό πλήρως συνδεδεμένο στρώμα (*Fully Connected Layer*) και το *softmax*,
- $x = C(I)$: Το λανθάνον διάνυσμα χαρακτηριστικών,
- $Y = S(x)$: Η πιθανότητα για κάθε κλάση, με $c = \arg\max(Y)$ την προβλεφθείσα κλάση.

Εφόσον μια μελλοντική αντιπαραδειγματική εικόνα I' αλλάζει την πρόβλεψη από c σε c', μας ενδιαφέρει να εντοπίσουμε ποια χαρακτηριστικά x'_i παρεκκλίνουν από τα τυπικά πρότυπα της κλάσης c'.

Μοντελοποίηση των Τιμών Ενεργοποίησης Λανθανόντων Χαρακτηριστικών

Για να περιγράψουμε στατιστικά κάθε χαρακτηριστικό x_i του λανθάνοντος χώρου του ταξινομητή, συγκεντρώνουμε το **σύνολο δεδομένων εκπαίδευσης** D και ορίζουμε:

$$L_c = \big\{\, x \mid x = C(I),\ S(x) = c,\ I \in D \big\}$$

για κάθε κλάση c. Υποθέτουμε ότι κάθε χαρακτηριστικό x_i ακολουθεί ξεχωριστή κατανομή X_{ci}. Καθώς το ConvNeXt χρησιμοποιεί συνάρτηση **GELU** (επιτρέποντας αρνητικές τιμές), δεν μπορούμε να εφαρμόσουμε απευθείας το hurdle model του αρχικού PIECE (που λειτουργούσε με ReLU). Συνεπώς:

- Μοντελοποιούμε την κατανομή των τιμών ενεργοποίησης του κάθε νευρώνα-χαρακτηριστικού με την χρήση των Gaussian Mixture Models (GMMs).Στα πειράματα μας χρησιμοποιήσαμε 5 Γκαουσιανές συνιστώσες για την μοντελοποίηση όλων των χαρακτηριστικών.
- Κάθε GMM ορίζεται από βάρη π_k, μέση τιμή μ_k και διακυμάνσεις σ_k^2.
- Η διαδικασία γίνεται μία φορά ανά κλάση, αποτυπώνοντας την «τυπική» κατανομή ενεργοποιήσεων των χαρακτηριστικών.

Έτσι, για την κλάση c και το χαρακτηριστικό i ισχύει:

$$p(x_i) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\big(x_i \mid \mu_k,\, \sigma_k^2\big).$$

Εντοπίζοντας τα Αποκλίνοντα Χαρακτηριστικά

Για μια εικόνα I και μια αντιπαραδειγματική κλάση c', εξετάζουμε κάθε χαρακτηριστικό x_i στο $x = C(I)$:

$$F(x_i) = \sum_{k=1}^{K} \pi_k \, \Phi\Big(\tfrac{x_i - \mu_k}{\sigma_k}\Big),$$

όπου $\Phi(\cdot)$ είναι η συνάρτηση κατανομής της κανονικής $\mathcal{N}(0,1)$. Για τον εντοπισμό των outliers (υψηλών ή χαμηλών), ορίζουμε:

$$P(x_i) = \min\big(F(x_i),\, 1 - F(x_i)\big).$$

Εάν $P(x_i) < t$ (μικρότερο από ένα κατώφλι t), τότε το x_i χαρακτηρίζεται «αποκλίνον».

Αλλάζοντας τα Αποκλίνοντα στην Αναμενόμενη Τιμή

Αφού εντοπίσουμε τα outliers, εξετάζουμε επιπρόσθετα το πρόσημο του βάρους w_i στο τελικό στρώμα S. Για παράδειγμα:

- Εάν το x_i είναι υπερβολικά υψηλό αλλά $w_i > 0$ (ίσως ωφελεί την κλάση c') και άρα δεν το αλλάζουμε.
- Εάν αντίθετα το x_i είναι υψηλό και $w_i < 0$, το θεωρούμε επιβλαβές και το αντικαθιστούμε με τη $\mathbb{E}[X_{c'i}]$, δηλαδή την αναμενόμενη τιμή της GMM για το χαρακτηριστικό i στην κλάση c'.

Με αυτή τη στρατηγική, μόνο οι τιμές που εμποδίζουν την ταξινόμηση στην c' τροποποιούνται.

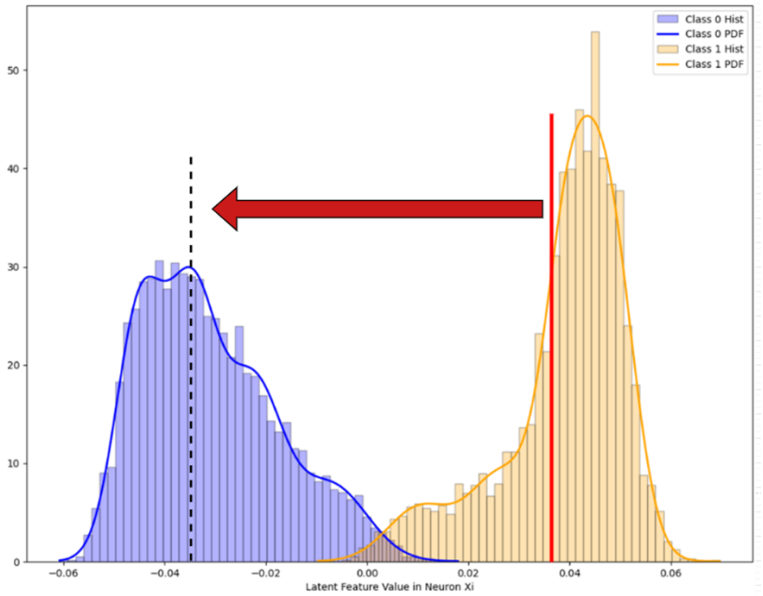


Figure 1.3.3: Διαγραμματική απεικόνιση του τρόπου που ένα «αποκλίνον» χαρακτηριστικό (κόκκινο) αλλάζει στην «αναμενόμενη» τιμή (διακεκομμένη γραμμή) για την αντιπαραδειγματική κλάση 0.

Τελικό Αποτέλεσμα. Με την αντικατάσταση των αποκλινόντων στοιχείων σε τιμές «φυσιολογικές» για την c' (διατηρώντας παράλληλα τις τυχόν ευεργετικές αποκλίσεις), λαμβάνουμε το *αντιπαραδειγματικό διάνυσμα χαρακτηριστικών* x'. Αυτό το διάνυσμα , όπως θα δούμε στην συνέχεια, θα λειτουργήσει ως οδηγός σε μία διαδικασία βελτιστοποίησης για την εύρεση της αντιπαραδειγματικής εικόνας.

1.3.3 Αναστροφή ΠΑΔ: E4E και PTI

Κίνητρο: Η Ανάγκη για Αναστροφή ΠΑΔ Υψηλής Πιστότητας και Επεξεργασιμότητας

Στο πλαίσιο του *PIECE*, η παραγωγή αντιπαραδειγματικών εικόνων βασίζεται σε μια διαδικασία βελτιστοποίησης που εφαρμόζεται απευθείας στο λανθάνον διάνυσμα της αρχικής εικόνας. Στον τομέα της ιατρικής απεικόνισης, αυτό απαιτεί μια λανθάνουσα αναπαράσταση που είναι ταυτόχρονα **ακριβής** (δηλαδή, πρέπει να ανακατασκευάζει την εικόνα εισόδου με ελάχιστη παραμόρφωση) και **επεξεργάσιμη** (δηλαδή, πρέπει να επιτρέπει ελεγχόμενες τροποποιήσεις για αντιπαραδειγματική σύνθεση). Οι παραδοσιακές μέθοδοι αναστροφής ΠΑΔ συχνά αντιμετωπίζουν δυσκολίες στην εξισορρόπηση αυτών των δύο ιδιοτήτων:

- **Προσεγγίσεις βασισμένες στη βελτιστοποίηση**, οι οποίες βελτιστοποιούν απευθείας τον λανθάνοντα κώδικα χρησιμοποιώντας ένα μόνο δείγμα, επιτυγχάνουν υψηλή πιστότητα ανακατασκευής αλλά παράγουν λανθάνοντες κώδικες που είναι δύσκολο να τροποποιηθούν.
- **Προσεγγίσεις βασισμένες σε αποκωδικοποιητές**, οι οποίες εκπαιδεύουν έναν αποκωδικοποιητή με μεγάλο αριθμό δειγμάτων, παρέχουν επεξεργάσιμες λανθάνουσες αναπαραστάσεις αλλά αποτυγχάνουν να ανακατασκευάσουν λεπτομέρειες.

- **Υβριδικές μέθοδοι**, οι οποίες πρώτα χρησιμοποιούν έναν αποκωδικοποιητή για να λάβουν το αρχικό λανθάνον διάνυσμα και στη συνέχεια εκτελούν απευθείας βελτιστοποίηση σε αυτό. Ακόμη και αυτή η μέθοδος δυσκολεύεται να βρει ένα ιδανικό σημείο ισορροπίας.

Στο προτεινόμενο πλαίσιο SPRUCE, χρησιμοποιούμε μια διαδικασία αναστροφής δύο σταδίων:

1. **Αποκωδικοποιητής E4E**: Παρέχει μια αρχική επεξεργάσιμη λανθάνουσα αναπαράσταση.
2. **Pivotal Tuning Inversion (PTI)**: Στοχευμένη εκπαίδευση του παραγωγικού μοντέλου για βελτιωμένη ακρίβεια ανακατασκευής.

Αναπαραστάσεις Λανθάνοντος Χώρου στο StyleGAN

Το StyleGAN ορίζει πολλαπλούς λανθάνοντες χώρους, ο καθένας με διαφορετική επίδραση στην ποιότητα, την επεξεργασιμότητα και τον ρεαλισμό των παραγόμενων εικόνων:

- **Χώρος Z**: Ο αρχικός λανθάνων χώρος, όπου οι λανθάνοντες κώδικες $z \sim \mathcal{N}(0, I)$ δειγματοληπτούνται από μια τυπική κανονική κατανομή. Αυτός ο χώρος είναι εξαιρετικά συσχετισμένος, καθιστώντας δύσκολες τις άμεσες τροποποιήσεις.
- **Χώρος W**: Ένας πιο δομημένος και απομπλεγμένος χώρος που λαμβάνεται μέσω ενός εκπαιδευμένου δικτύου απεικόνισης M, όπου $w = M(z)$. Αυτός ο χώρος επιτρέπει ελεγχόμενες τροποποιήσεις χαρακτηριστικών διατηρώντας τον ρεαλισμό.

Για τη βελτίωση των δυνατοτήτων ανακατασκευής και τον λεπτομερή έλεγχο των χαρακτηριστικών, το StyleGAN εισάγει επαυξημένους λανθάνοντες χώρους:

- **Χώρος $\mathbf{W}^+$**: Σε αντίθεση με τον χώρο **W**, όπου ένας μόνο λανθάνων κώδικας εφαρμόζεται σε όλα τα στρώματα του παραγωγικού μοντέλου, ο χώρος $\mathbf{W}^+$ αναθέτει ανεξάρτητους λανθάνοντες κώδικες σε κάθε στρώμα, επιτρέποντας λεπτομερέστερο έλεγχο των τοπικών χαρακτηριστικών της εικόνας.
- **Χώρος $\mathbf{W^k}$**: Μια περαιτέρω επέκταση όπου μόνο ένα υποσύνολο k στρωμάτων λαμβάνει διακριτούς λανθάνοντες κώδικες, παρέχοντας ισορροπία μεταξύ εκφραστικότητας και ελεγξιμότητας.

Αποκωδικοποιητής E4E για Αναστροφή ΠΑΔ

Ο **E4E (Encoder for Editing)** σχεδιάστηκε για την αντιμετώπιση των προκλήσεων αναστροφής, τοποθετώντας στρατηγικά λανθάνοντες κώδικες στον επαυξημένο λανθάνοντα χώρο, διασφαλίζοντας παράλληλα τη συμβατότητα με σημασιολογικές τροποποιήσεις.

Βασικές Αρχές Σχεδιασμού του E4E

1. **Ελαχιστοποίηση της Διακύμανσης στους Λανθάνοντες Κώδικες** - Σε αντίθεση με τους παραδοσιακούς αποκωδικοποιητές, ο E4E εκπαιδεύεται για προοδευτική βελτίωση των λανθανουσών αναπαραστάσεων, επιτρέποντας τη βελτιστοποίηση τόσο της πιστότητας ανακατασκευής όσο και της επεξεργασιμότητας.
2. **Ελαχιστοποίηση της Απόκλισης από τον χώρο W** - Μια ειδική συνάρτηση απώλειας ενθαρρύνει τους λανθάνοντες κώδικες να παραμένουν κοντά στον χώρο **W** ενώ αξιοποιούν την ευελιξία του χώρου $\mathbf{W}^+$, βελτιώνοντας την επεξεργασιμότητα.
3. **Αρχιτεκτονική Αποκωδικοποιητή Βασισμένη σε ResNet** - Ο E4E βασίζεται σε ένα ResNet-like σκελετό, παράγοντας έναν αρχικό λανθάνοντα κώδικα και πρόσθετες μετατοπίσεις που βελτιώνουν συγκεκριμένα στρώματα.

Συνολική Συνάρτηση Απώλειας στον E4E

Ο αποκωδικοποιητής E4E εκπαιδεύεται χρησιμοποιώντας μια συνάρτηση απώλειας σχεδιασμένη να εξισορροπεί δύο κρίσιμους στόχους:

- **Ελαχιστοποίηση Παραμόρφωσης**: Διασφαλίζοντας ότι η ανακατασκευασμένη εικόνα μοιάζει πολύ με την αρχική.

- **Διατήρηση Επεξεργασιμότητας**: Διατηρώντας έναν δομημένο λανθάνοντα κώδικα που επιτρέπει σημαντικές τροποποιήσεις.

Η συνολική συνάρτηση απώλειας ορίζεται ως σταθμισμένος συνδυασμός αυτών των δύο όρων:

$$L(x) = L_{\text{dist}}(x) + \lambda_{\text{edit}} L_{\text{edit}}(x), \tag{1.3.1}$$

όπου λ_{edit} ελέγχει την ισορροπία μεταξύ πιστότητας ανακατασκευής και ευελιξίας του λανθάνοντος κώδικα.

Απώλεια Παραμόρφωσης

Για την ελαχιστοποίηση του σφάλματος ανακατασκευής, η απώλεια παραμόρφωσης αποτελείται από τρία συστατικά:

$$L_{\text{dist}}(x) = \lambda_2 L_2(x) + \lambda_{lpips} L_{LPIPS}(x) + \lambda_{sim} L_{\text{sim}}(x). \tag{1.3.2}$$

- $L_2(x)$ - Τυπική απώλεια ανακατασκευής εικονοστοιχείων. - $L_{LPIPS}(x)$ - Αντιληπτική απώλεια για τη διασφάλιση δομικής ομοιότητας μεταξύ εικόνων. - $L_{\text{sim}}(x)$ - Απώλεια ταυτότητας, διασφαλίζοντας τη συνέπεια σε επίπεδο χαρακτηριστικών χρησιμοποιώντας ένα προ-εκπαιδευμένο δίκτυο.

Απώλεια Επεξεργασιμότητας

Για να διασφαλιστεί ότι οι λανθάνοντες κώδικες παραμένουν δομημένοι και επεξεργάσιμοι, η απώλεια επεξεργασιμότητας αποτελείται από:

$$L_{\text{edit}}(x) = \lambda_{\text{d-reg}} L_{\text{d-reg}}(x) + \lambda_{\text{adv}} L_{\text{adv}}(x). \tag{1.3.3}$$

- $L_{\text{d-reg}}(x)$ - Απώλεια Delta-κανονικοποίησης που περιορίζει τις μετατοπίσεις Δ_i, διασφαλίζοντας εγγύτητα στον χώρο W. - $L_{\text{adv}}(x)$ - Ανταγωνιστική απώλεια χρησιμοποιώντας έναν διακριτοποιητή λανθάνοντος χώρου για να διατηρήσει τους μαθημένους κώδικες στυλ εντός της εγγενούς κατανομής του StyleGAN.

Στοχευμένη Εκπαίδευση για Αναστροφή ΠΑΔ: Pivotal Tuning Inversion (PTI)

Κίνητρο για το PTI στην Αναστροφή Ιατρικών Εικόνων Ενώ ο E4E παρέχει μια επεξεργάσιμη λανθάνουσα αναπαράσταση, δεν εγγυάται πάντα ανακατασκευές υψηλής πιστότητας, ειδικά για ιατρικές εικόνες που περιέχουν λεπτομερείς, υψηλής ανάλυσης ανατομικές δομές. Στις εφαρμογές ιατρικής απεικόνισης, ακόμη και μικρά τεχνουργήματα ανακατασκευής μπορούν να αποκρύψουν κρίσιμες διαγνωστικές λεπτομέρειες. Αυτό απαιτεί μια μέθοδο αναστροφής που ενισχύει την ακρίβεια ανακατασκευής διατηρώντας παράλληλα την επεξεργασιμότητα.

Το Pivotal Tuning Inversion (PTI) αντιμετωπίζει αυτήν την πρόκληση με στοχευμένη εκπαίδευση του ίδιου του παραγωγικού μοντέλου, διασφαλίζοντας ότι ο ανεστραμμένος λανθάνων κώδικας παράγει μια εικόνα που είναι αδιάκριτη από την είσοδο ενώ εξακολουθεί να επιτρέπει ουσιαστικές αντιπαραδειγματικές τροποποιήσεις.

Μεθοδολογία

Το PTI αποτελείται από δύο κύρια βήματα:

1. **Αρχική Αναστροφή ΠΑΔ:** Η εικόνα αντιστρέφεται πρώτα χρησιμοποιώντας μια έτοιμη μέθοδο βασισμένη σε αποκωδικοποιητή (π.χ., E4E) για να ληφθεί ένας αρχικός λανθάνων κώδικας w στον λανθάνοντα χώρο W^+.
2. **Στοχευμένη Εκπαίδευση Παραγωγικού Μοντέλου:** Το παραγωγικό μοντέλο G εκπαιδεύεται στοχευμένα για καλύτερη ανακατασκευή της εικόνας-στόχου διατηρώντας τον λανθάνοντα κώδικα σταθερό, βελτιώνοντας έτσι την ισορροπία μεταξύ ελαχιστοποίησης παραμόρφωσης και επεξεργασιμότητας.

Κανονικοποίηση Τοπικότητας

Για την αποφυγή της υπερπροσαρμογής σε μία μόνο εικόνα, εισάγεται η κανονικοποίηση τοπικότητας. Αυτό διασφαλίζει ότι το παραγωγικό μοντέλο δεν χάνει την ικανότητά του να παράγει διαφορετικά δείγματα από το αρχικό σύνολο δεδομένων. Ο όρος κανονικοποίησης τοπικότητας ορίζεται ως:

$$\mathcal{L}_{reg} = \mathcal{L}_{\text{LPIPS}}(x_r, x_r^*) + \lambda_{L2}^R \mathcal{L}_{L2}(x_r, x_r^*). \tag{1.3.4}$$

όπου:

- $x_r = G(w_r, \theta)$, είναι η εικόνα που παράγεται από το αρχικό παραγωγικό μοντέλο με το λανθάνον διάνυσμα w_r
- $x_r^* = G(w_r, \theta^*)$, είναι η εικόνα που παράγεται από το εκπαιδευμένο παραγωγικό μοντέλο με το λανθάνον διάνυσμα w_r

Τελικά, η βελτιστοποίηση του παραγωγικού μοντέλου διατυπώνεται ως:

$$\theta^* = \arg\min_{\theta^*} L_{\text{pti}} + \lambda_{\text{reg}} L_{\text{reg}}. \tag{1.3.5}$$

1.3.4 Βελτιστοποίηση Λανθάνοντος Διανύσματος

Το τελικό βήμα της μεθόδου SPRUCE είναι η επαναληπτική βελτιστοποίηση, στο λανθάνοντα χώρο, του λανθάνοντος διανύσματος που ανακατασκευάζει την αρχική εικόνα με σκοπό την απόκτηση της λανθάνουσας αναπαράστασης της αντιπαραδειγματικής εικόνας. Για την παραγωγή αντιπαραδειγματικών εικόνων που εξηγούν ουσιαστικά την απόφαση του ταξινομητή, βελτιστοποιούμε το λανθάνον διάνυσμα του StyleGAN2-ADA, που έχουμε αποκτήσει χρησιμοποιώντας τη διαδικασία αναστροφής ΠΑΔ, με τέτοιο τρόπο ώστε:

- Τα χαρακτηριστικά της εικόνας που παράγεται από το βελτιστοποιημένο λανθάνον διάνυσμα να ευθυγραμμίζονται με τα αποκλίνοντα χαρακτηριστικά που εξάγονται μέσω της μεθόδου PIECE.
- Η τροποποιημένη εικόνα να παραμένει αντιληπτικά(perceptually) παρόμοια με την αρχική, αποφεύγοντας υπερβολικές παραμορφώσεις.
- Οι τροποποιήσεις να είναι αραιές, δηλαδή να αλλάζουν μόνο οι ουσιώδεις πτυχές που συμβάλλουν στην απόφαση του ταξινομητή.

Συνάρτηση Βελτιστοποίησης

Η διαδικασία βελτιστοποίησης πραγματοποιείται στο λανθάνον διάνυσμα που αρχικοποιείται ως w_p. Η συνάρτηση βελτιστοποίησης αποτελείται από τέσσερις βασικούς όρους απώλειας:

$$L = L_{\text{piece}} + \lambda_{\text{perc}} L_{\text{LPIPS}} + \lambda_{\text{latent}} L_{\text{latent}} + \lambda_{\text{image}} L_{\text{image}}. \tag{1.3.6}$$

Όρος για Ευθυγράμμιση Χαρακτηριστικών

Έχοντας κατασκευάσει το αντιπαραδειγματικό διάνυσμα χαρακτηριστικών x', χρησιμοποιούμε την ίδια απώλεια σε επίπεδο χαρακτηριστικών όπως στον αρχικό αλγόριθμο PIECE:

$$L_{\text{piece}} = \|C(G(w_e)) - x'\|_2^2. \tag{1.3.7}$$

Όπου:

- C αντιπροσωπεύει όλα τα στρώματα του παγωμένου ταξινομητή μέχρι το προτελευταίο στρώμα χαρακτηριστικών X.
- x' είναι το αντιπαραδειγματικό διάνυσμα χαρακτηριστικών.
- G είναι το προσαρμοσμένο παραγωγικό μοντέλο.

- w_e είναι το βελτιστοποιημένο λανθάνον διάνυσμα.

Αυτός ο όρος διασφαλίζει ότι τα χαρακτηριστικά της αντιπαραδειγματικής εικόνας ταιριάζουν με εκείνα που αναμένονται για την αντιπαραδειγματική κλάση, εξασφαλίζοντας έτσι ότι η πρόβλεψη του ταξινομητή θα αλλάξει.

Όρος Αντιληπτικής Ομοιότητας (LPIPS)

Για τη διατήρηση της δομικής ακεραιότητας και την αποφυγή σημαντικής απόκλισης της αντιπαραδειγματικής εικόνας από την αρχική, ενσωματώνουμε έναν όρο αντιληπτικής ομοιότητας χρησιμοποιώντας την μετρική LPIPS (Learned Perceptual Image Patch Similarity):

$$L_{\text{LPIPS}} = \text{LPIPS}(G(w_e), I). \tag{1.3.8}$$

Όπου:

- $G(w_e)$ είναι η παραγόμενη αντιπαραδειγματική εικόνα.
- I είναι η αρχική εικόνα από το σύνολο δεδομένων.

Η LPIPS διασφαλίζει ότι τα χαρακτηριστικά υψηλού επιπέδου διατηρούνται ενώ επιτρέπονται ουσιαστικές τροποποιήσεις. Για το δίκτυο βάσης του LPIPS, χρησιμοποιούμε έναν εξαγωγέα χαρακτηριστικών βασισμένο στο VGG αντί για το AlexNet, λόγω της ευρείας υιοθέτησης του VGG σε σχήματα βελτιστοποίησης που περιλαμβάνουν παραλλαγές του StyleGAN.

Όρος Κανονικοποίησης στο Λανθάνοντα Χώρο για Αραιότητα

Εφαρμόζουμε επίσης κανονικοποίηση L1 για τον περιορισμό ασήμαντων αλλαγών στο λανθάνοντα χώρο:

$$L_{\text{latent}} = \|w_e - w_p\|_1. \tag{1.3.9}$$

Αυτός ο όρος αποτρέπει το βελτιστοποιημένο λανθάνον διάνυσμα να απομακρυνθεί από την επεξεργάσιμη και σημασιολογικά πλούσια γειτονιά του αρχικού λανθάνοντος διανύσματος w_p, ενισχύοντας έτσι τον ρεαλισμό και την αληθοφάνεια των παραγόμενων αντιπαραδειγμάτων.

Όρος Κανονικοποίησης στον Χώρο των Εικονοστοιχείων για Αραιότητα

Για να διασφαλίσουμε ότι οι αντιπαραδειγματικές εικόνες παραμένουν πιστές στην αρχική εικόνα ενώ αντικατοπτρίζουν μόνο τις απαραίτητες τροποποιήσεις, εισάγουμε μια ποινή L1 στις διαφορές σε επίπεδο εικονοστοιχείων μεταξύ της παραγόμενης αντιπαραδειγματικής εικόνας και της αρχικής εικόνας:

$$L_{\text{image}} = \|G(w_e) - I\|_1. \tag{1.3.10}$$

Αυτή η κανονικοποίηση ενθαρρύνει την αραιότητα στις αλλαγές των εικονοστοιχείων, διασφαλίζοντας ότι τροποποιούνται μόνο οι πιο σχετικές περιοχές της εικόνας, εκείνες που ευθύνονται για την απόφαση του ταξινομητή.

Συνεπώς, το τελικό πρόβλημα βελτιστοποίησης διατυπώνεται ως:

$$w_e^* = \arg\min_{w_p} \Big[L_{\text{piece}} + \lambda_{\text{perc}} L_{\text{LPIPS}} + \lambda_{\text{latent}} L_{\text{latent}} + \lambda_{\text{image}} L_{\text{image}} \Big]. \tag{1.3.11}$$

Όπου:

- w_e^* είναι το λανθάνον διάνυσμα που αντιστοιχεί στην τελική αντιπαραδειγματική εικόνα.
- $\lambda_{\text{perc}}, \lambda_{\text{latent}}, \lambda_{\text{image}}$ είναι υπερπαράμετροι που ελέγχουν την ισορροπία μεταξύ ευθυγράμμισης χαρακτηριστικών, αντιληπτικής ομοιότητας και αραιότητας.

Τροφοδοτώντας το w_e^* στο προσαρμοσμένο παραγωγικό μοντέλο G, μπορούμε να οπτικοποιήσουμε την αντιπαραδειγματική εικόνα.

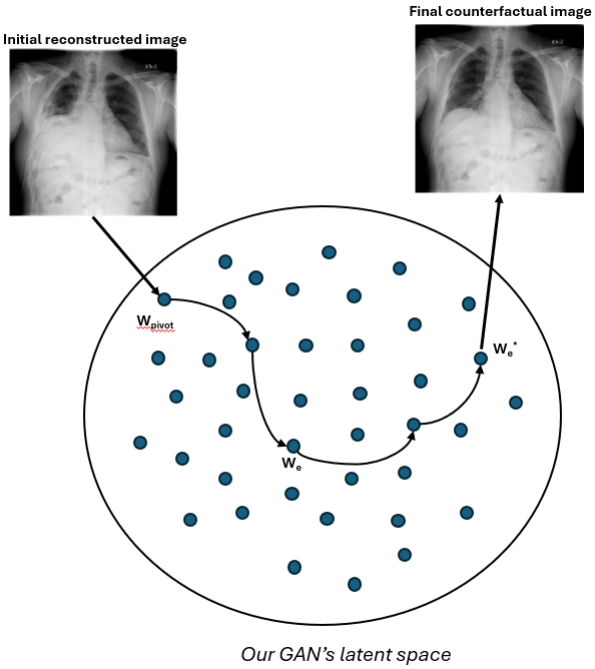


Figure 1.3.4: Οπτικοποίηση της διαδικασίας βελτιστοποίησης

1.3.5 Ανταγωνιστική Ανθεκτικότητα για Ουσιαστικές Αντιπαραδειγματικές Εξηγήσεις

Κίνητρο: Ο Ρόλος της Ανταγωνιστικής Ανθεκτικότητας

Όπως είδαμε στις προηγούμενες ενότητες του πλαισίου μας, το διάνυσμα χαρακτηριστικών του ταξινομητή χρησιμοποιείται για να καθοδηγήσει το παραγωγικό μοντέλο του StyleGAN2-ada προς την παραγωγή μιας εικόνας που ανήκει στην επιθυμητή αντιπαραδειγματική κλάση και είναι όσο το δυνατόν πιο κοντά στην αρχική εικόνα. Με άλλα λόγια, μπορούμε να πούμε ότι τα gradients του ταξινομητή είναι αυτά που καθοδηγούν το παραγωγικό μοντέλο προς τη σωστή κατεύθυνση. Συνεπώς, εάν τα gradients δεν είναι σημασιολογικά ευθυγραμμισμένα με τα χαρακτηριστικά μιας συγκεκριμένης κλάσης, τότε θα μπορούσε να οδηγήσει, όπως θα δούμε στην ενότητα των πειραμάτων, σε αντιπαραδείγματα που μοιάζουν οπτικά με την αρχική εικόνα όταν οι αλλαγές περιορίζονται να είναι ελάχιστες. Αυτό σημαίνει ότι εάν ο ταξινομητής είναι ευάλωτος σε ανταγωνιστικές, μη αντιληπτές από τον άνθρωπο τροποποιήσεις (ανταγωνιστικές επιθέσεις) [33], τότε τα παραγόμενα αντιπαραδείγματα θα αντικατοπτρίζουν αυτές τις ευπάθειες παρά τις πραγματικές αιτιατές συσχετίσεις.

Έχει αποδειχθεί ότι τα gradients των ανταγωνιστικά ανθεκτικών ταξινομητών έχουν ισχυρές παραγωγικές ιδιότητες και "μαθαίνουν" χαρακτηριστικά που είναι αντιληπτικά ευθυγραμμισμένα με τα χαρακτηριστικά μιας συγκεκριμένης κλάσης [100, 7]. Για αυτόν τον λόγο, χρησιμοποιούμε ανταγωνιστική εκπαίδευση για να βελτιώσουμε την ανθεκτικότητα του ταξινομητή και να διασφαλίσουμε ότι οι παραγόμενες εικόνες περιέχουν ουσιαστικές και αληθοφανείς αλλαγές. Συγκεκριμένα, χρησιμοποιούμε την μέθοδο TRADES (TRade-off-inspired Adversarial DEfense via Surrogate-loss minimization) [136], μια προηγμένη μέθοδο ανταγωνιστικής εκπαίδευσης που εξισορροπεί την ακρίβεια σε καθαρά δεδομένα και την ανθεκτικότητα.

TRADES: Εξισορρόπηση Καθαρής Ακρίβειας και Ανθεκτικότητας

Το πλαίσιο TRADES διατυπώνει την ανταγωνιστική εκπαίδευση ως μια εξισορρόπηση μεταξύ:

- **Τυπικής Ακρίβειας Ταξινόμησης:** Διασφαλίζει ότι ο ταξινομητής διατηρεί υψηλή ακρίβεια σε καθαρές (μη διαταραγμένες) εικόνες, κάτι που είναι απαραίτητο ειδικά στον τομέα της ιατρικής απεικόνισης.

- **Ανταγωνιστικής Ανθεκτικότητας:** Εκπαίδευση του ταξινομητή ώστε να είναι αμετάβλητος σε ανταγωνιστικές διαταραχές.

Μαθηματική Διατύπωση του TRADES

Η κεντρική ιδέα του TRADES διατυπώνεται ως μια κανονικοποιημένη συνάρτηση απώλειας αντικατάστασης, που συνδυάζει μια τυπική συνάρτηση απώλειας ταξινόμησης και έναν όρο κανονικοποίησης ανθεκτικότητας. Συγκεκριμένα, το TRADES βελτιστοποιεί την ακόλουθη συνάρτηση:

$$\min_f \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(f(x), y) + \beta \cdot \max_{x' \in \mathcal{B}_\epsilon(x)} \mathcal{L}_{\text{KL}}(f(x), f(x')) \right], \tag{1.3.12}$$

όπου:

- $f(x)$ αντιπροσωπεύει την προβλεπόμενη κατανομή πιθανότητας του ταξινομητή για την είσοδο x,
- $\mathcal{L}$ υποδηλώνει την τυπική συνάρτηση απώλειας ταξινόμησης (π.χ., cross-entropy),
- $\mathcal{L}_{\text{KL}}$ είναι η απόκλιση Kullback-Leibler που ποσοτικοποιεί τη διαφορά μεταξύ των εξόδων του ταξινομητή στο καθαρό παράδειγμα x και το ανταγωνιστικό παράδειγμα x',
- $\mathcal{B}_\epsilon(x)$ είναι η σφαίρα διαταραχής ακτίνας ϵ γύρω από την είσοδο x, που ορίζεται ως $\mathcal{B}_\epsilon(x) = \{x' : \|x' - x\|_p \leq \epsilon\}$,
- β είναι μια υπερπαράμετρος που ελέγχει την εξισορρόπηση μεταξύ της καθαρής ακρίβειας(clean accuracy) και ανταγωνιστικής ανθεκτικότητας.

Κατά τη διάρκεια της εκπαίδευσης, τα ανταγωνιστικά παραδείγματα x' δημιουργούνται εντός της σφαίρας διαταραχής χρησιμοποιώντας την Προβαλλόμενη Καθοδήγηση Διαβάθμισης (Projected Gradient Descent - PGD) [73]. Η PGD διαταράσσει επαναληπτικά τα δεδομένα εισόδου στην κατεύθυνση της διαβάθμισης της απώλειας KL για να βρει μια διαταραχή χειρότερης περίπτωσης. Αυτή η επαναληπτική δημιουργία ανταγωνιστικού παραδείγματος μπορεί να εκφραστεί τυπικά ως:

$$x'_{t+1} = \Pi_{\mathcal{B}_\epsilon(x)} \left(x'_t + \alpha \cdot \text{sign}(\nabla_{x'_t} \mathcal{L}_{\text{KL}}(f(x), f(x'_t))) \right), \tag{1.3.13}$$

όπου $\Pi_{\mathcal{B}_\epsilon(x)}(\cdot)$ υποδηλώνει την προβολή στη σφαίρα ακτίνας ϵ, α είναι το μέγεθος του βήματος και t υποδεικνύει τον αριθμό της επανάληψης.

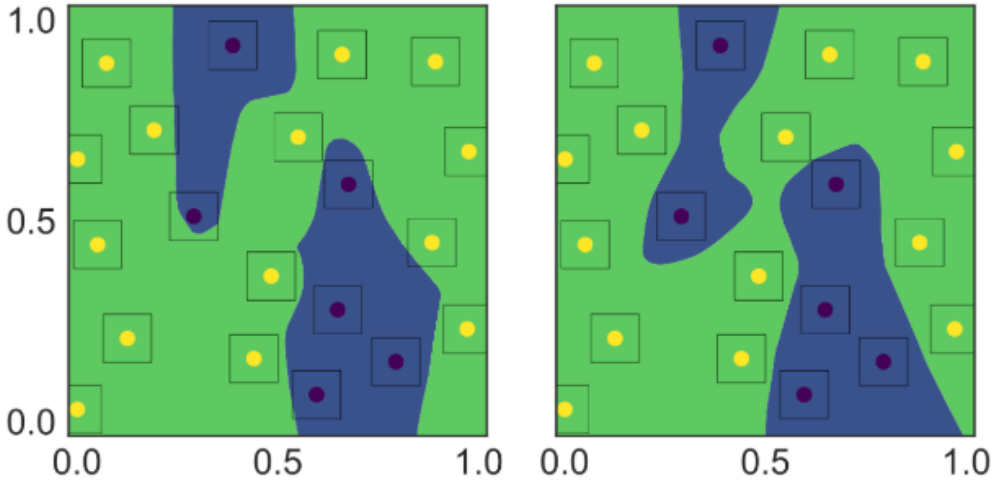


Figure 1.3.5: Αριστερή εικόνα: όριο απόφασης με απλή εκπαίδευση. Δεξιά εικόνα: όριο απόφασης με TRADES.

1.4 Πειράματα

1.4.1 Επισκόπηση Συνόλων Δεδομένων

Χρησιμοποιήσαμε τέσσερα δημόσια διαθέσιμα σύνολα δεδομένων ιατρικής απεικόνισης, καθένα από τα οποία περιλαμβάνει μοναδικές παθολογίες. Ακολουθεί μια συνοπτική περιγραφή των δεδομένων, μαζί με ενδεικτικές εικόνες που αναδεικνύουν την ποικιλομορφία τόσο των ανατομικών στοιχείων όσο και των παθολογιών.

Ακτινογραφίες Θώρακα

(α) Πνευμονία vs. Υγιής

- *Πηγή:* RSNA Pneumonia Detection Challenge [2].
- *Κλάσεις:* Πνευμονία, Υγιής.
- *Προεπεξεργασία:* Αλλαγή μεγέθους σε 256×256, μετατροπή σε RGB.
- *Βασικά Χαρακτηριστικά:*
  - **Πνευμονία:** Περιοχές αυξημένης πυκνότητας που προσομοιάζουν φλεγμονή των πνευμόνων.
  - **Υγιής:** Καθαροί πνεύμονες, φυσιολογικά αγγειακά πρότυπα.

(β) Μεγαλοκαρδία vs. Υγιής

- *Πηγή:* NIH Chest X-ray [126] και RSNA Challenge.
- *Κλάσεις:* Μεγαλοκαρδία, Υγιής.
- *Προεπεξεργασία:* Αλλαγή μεγέθους σε 256×256, μετατροπή σε RGB.
- *Βασικά Χαρακτηριστικά:*
  - **Μεγαλοκαρδία:** Μεγάλες διαστάσεις καρδιακής σιλουέτας, συχνά πέραν του 50% του πλάτους του θώρακα.
  - **Υγιής:** Διαυγείς πνεύμονες, καρδιά φυσιολογικού μεγέθους.

Figure 1.4.1: Ενδεικτικές ακτινογραφίες θώρακα: Υγιής (αριστερά), Πνευμονία (μεσαία), Μεγαλοκαρδία (δεξιά).

Οπτικές Τομογραφίες Συνοχής (ΟΤΣ) [55]

- *Πηγή:* Kaggle.
- *Κλάσεις:* Υγιής, Choroidal NeoVascularization (CNV),Diabetic Macular Edema (DME), Drusen.
- *Προεπεξεργασία:* Εστίαση στο κεντρικό σημείο (center-crop) της εικόνας , αλλαγή μεγέθους σε 256×256, μετατροπή σε RGB.

- *Βασικές Παθολογίες:*
  - **CNV (Choroidal Neovascularization):** Συσσώρευση υγρού που αλλοιώνει τις στιβάδες του αμφιβληστροειδούς.
  - **DME (Diabetic Macular Edema):** Πάχυνση του αμφιβληστροειδούς, κύστες γεμάτες υγρό.
  - **Drusen:** Εξωκυτταρικές εναποθέσεις κάτω από το μελάγχρου επιθήλιο του αμφιβληστροειδούς.
  - **Υγιής:** Ομοιόμορφες, μη παραμορφωμένες στιβάδες χωρίς υγρό ή εναποθέσεις.

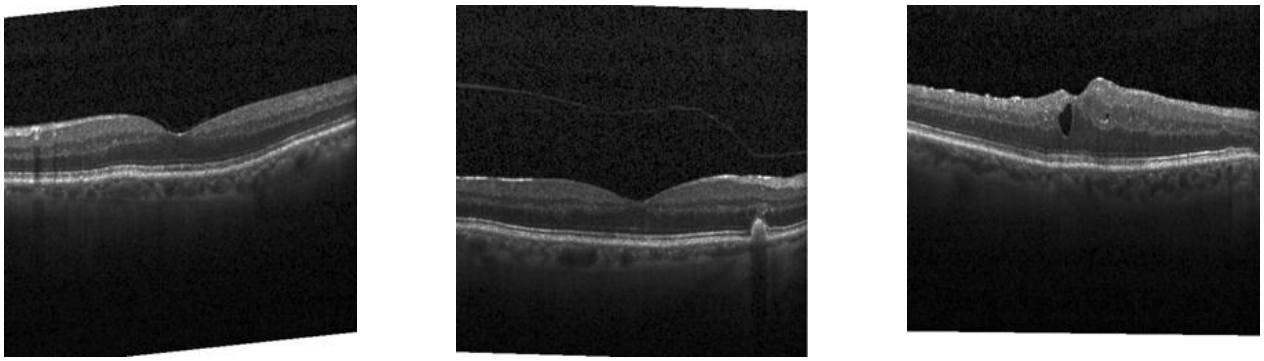


Figure 1.4.2: Παραδείγματα ΟΤΣ: Υγιές (αριστερά), Drusen (μεσαία), DME (δεξιά).

Μαγνητικές Τομογραφίες (MRI) Εγκεφάλου για Άνοια

- *Πηγή:* Kaggle.
- *Κλάσεις:* Χωρίς Βλάβη, Πολύ Ήπια, Ήπια, Μέτρια.
- *Προεπεξεργασία:* Αλλαγή μεγέθους σε 128×128, μετατροπή σε RGB.
- *Βασικά Χαρακτηριστικά:*
  - **Χωρίς Βλάβη:** Φυσιολογικός όγκος εγκεφάλου, ελάχιστη ή ανύπαρκτη ατροφία.
  - **Πολύ Ήπια / Ήπια**: Σταδιακή λέπτυνση του φλοιού, μέτρια διεύρυνση των κοιλιών.
  - **Μέτρια**: Έντονη ατροφία (ιδίως στις ιπποκαμπικές περιοχές), εμφανώς διευρυμένες κοιλίες.

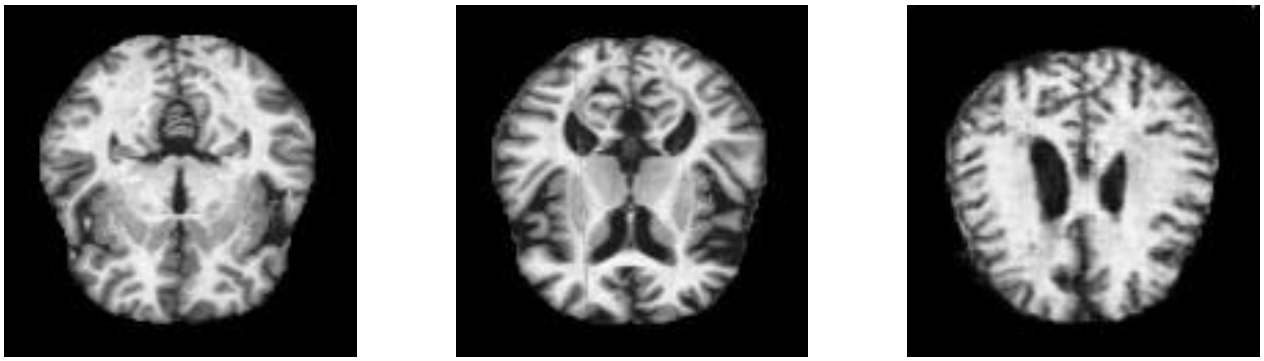


Figure 1.4.3: Παραδείγματα εγκεφαλικών MRI: Χωρίς Βλάβη (αριστερά), Πολύ Ήπια (μεσαία), Μέτρια (δεξιά).

1.4.2 Εκπαίδευση Ταξινομητή

Εκπαιδεύσαμε το μοντέλο **ConvNeXt-Base** τόσο σε απλή (μη ανθεκτική) ρύθμιση όσο και κάτω από ανταγωνιστικές επιθέσεις, προκειμένου να αξιολογήσουμε την ικανότητα του προτεινόμενου πλαισίου να παράγει αντιπαραδειγματικές εικόνες. Στην παρούσα ενότητα περιγράφονται αναλυτικά η διαδικασία εκπαίδευσης, τα

βασικά ευρήματα και ο ρόλος της ανταγωνιστικής εκπαίδευσης στην τελική ποιότητα των αντιπαραδειγματικών εικόνων.

Γενική Διαμόρφωση Εκπαίδευσης

Όλα τα σύνολα δεδομένων χωρίστηκαν σε υποσύνολα εκπαίδευσης (80%), επικύρωσης (10%) και ελέγχου (10%), όπως φαίνεται στον Πίνακα 1.2. Αρχικοποιήσαμε το ConvNeXt-Base με βάρη ενός αντίστοιχου προεκπαιδευμένου μοντέλου στο ImageNet και αντικαταστήσαμε το τελικό επίπεδο (fully connected) ώστε να ανταποκρίνεται στον αριθμό κλάσεων κάθε συνόλου δεδομένων. Η εκπαίδευση περιελάμβανε:

- **Χρονοδρομολόγηση** ρυθμού μάθησης και υπομονή 20 εποχών,
- **AdamW** Βελτιστοποιητή ,
- **Εξομάλυνση ετικέτας** (0.05) και **εκπαίδευση μικτής ακρίβειας** (mixed-precision) για βέλτιστη χρήση πόρων,
- **Επαύξηση δεδομένων**, π.χ. τυχαίες αποκοπές, αναστροφές, μετασχηματισμοί affine και μεταβολές χρώματος,
- **Κατανομή βαρών ανά κλάση** για την αντιμετώπιση ανισορροπίας σε ορισμένα σύνολα δεδομένων (π.χ. Υγιής vs. Πνευμονία).

Σύνολο Δεδομένων	Κλάση	Train	Val	Test	Ανάλυση
Ακτινογραφία Θώρακα (Πνευμονία vs. Υγιής)	Υγιής	7080	885	601	256×256
	Πνευμονία	4809	601	602	
Ακτινογραφία Θώρακα (Μεγαλοκαρδία vs. Υγιής)	Υγιής	2220	277	279	256×256
	Μεγαλοκαρδία	2220	277	279	
ΟΤΣ	Υγιής	21077	2634	2636	256×256
	CNV	29772	3721	3723	
	DME	9137	1142	1143	
	Drusen	6896	862	862	
MRI Εγκεφάλου για Άνοια	Χωρίς Βλάβη	2560	320	320	128×128
	Πολύ Ήπια	2406	300	302	
	Ήπια	2191	273	275	
	Μέτρια	2057	257	258	

Table 1.2: Διαχωρισμός σε σύνολα εκπαίδευσης,ελέγχου και επικύρωσης και λεπτομέρειες σχετικά με την ανάλυση εικόνων για κάθε σύνολο δεδομένων.

Απλή Εκπαίδευση Ταξινομητή

Αρχικά εκπαιδεύσαμε έναν "απλό" ταξινομητή (μη ανθεκτικό σε ανταγωνιστικές επιθέσεις) με χρήση συνάρτησης cross-entropy. Ο Πίνακας 1.3 παρουσιάζει τις βασικές υπερπαραμέτρους (π.χ. αρχικός ρυθμός μάθησης 5×10^{-4}, μέγεθος παρτίδας 64, 100 εποχές).

Μετρικές Απόδοσης. Τα αποτελέσματα ακριβείας (accuracy), precision, recall και F1-score στα σύνολα ελέγχου συνοψίζονται στον Πίνακα 1.4. Η διάκριση Πνευμονίας vs. Υγιούς έφτασε $\sim 96\%$ ακρίβεια, ενώ η ταξινόμηση Άνοιας σε MRI άγγιξε $\sim 99\%$.

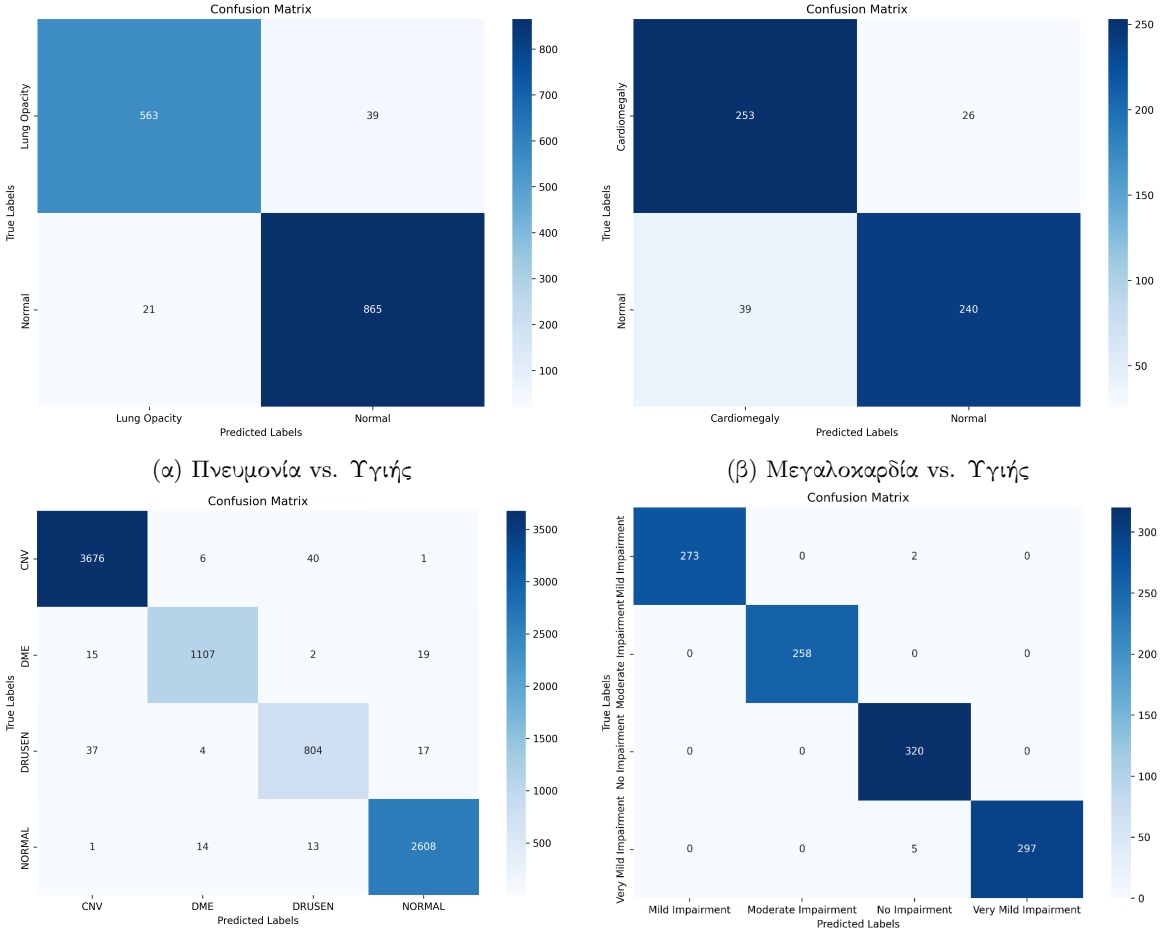
Στο Σχήμα 1.4.4 παρουσιάζονται οι πίνακες σύγχυσης (confusion matrices) για μία αναλυτικότερη εικόνα της κατανομής των προβλέψεων.

Υπερπαράμετρος	Τιμή
Συνάρτηση Απώλειας	Cross Entropy
Αρχικός Ρυθμός Μάθησης	5e-4
Εποχές	100
Warmup Εποχές	10
Βελτιστοποιητής (Optimizer)	AdamW
Μέγεθος Παρτίδας	64
Weight Decay	5e-2
Εξομάλυνση Ετικέτας	0.05

Table 1.3: Βασικές υπερπαράμετροι για την απλή εκπαίδευση του ταξινομητή.

	Ακρίβεια	Precision	Recall	F1-score
Πνευμονία	95.96	95.68	97.63	96.65
Μεγαλοκαρδία	88.35	90.22	86.02	88.07
ΟΤΣ	97.98	97.98	97.97	97.98
MRI Εγκεφάλου	99.39	99.40	99.39	99.40

Table 1.4: Επιδόσεις (%) στο σύνολο ελέγχου για τον απλά εκπαιδευμένο ταξινομητή.



(α) Πνευμονία vs. Υγιής

(β) Μεγαλοκαρδία vs. Υγιής

(γ) ΟΤΣ

(δ) MRI Εγκεφάλου

Figure 1.4.4: Πίνακες σύγχυσης για τον απλό ταξινομητή στα δεδομένα ελέγχου.

Ανταγωνιστική Εκπαίδευση με TRADES

Στη συνέχεια, εκπαιδεύσαμε το ίδιο μοντέλο ConvNeXt χρησιμοποιώντας την μέθοδο TRADES [136], εισάγοντας ανταγωνιστικά παραδείγματα κατά την διάρκεια της εκπαίδευσης με τη μέθοδο PGD. Ο Πίνακας 1.5 συνοψίζει τις σχετικές υπερπαραμέτρους (π.χ. μικρότερο μέγεθος παρτίδας 16, $\epsilon \in \{1/255,\ 2/255,\ 8/255\}$).

Υπερπαράμετρος	**Τιμή**
Συνάρτηση Απώλειας	TRADES
Αρχικός Ρυθμός Μάθησης	1e-4
Εποχές	100
Warmup Εποχές	10
Βελτιστοποιητής	AdamW
Μέγεθος Παρτίδας	16
Weight Decay	5e-2
Εξομάλυνση Ετικέτας	0.05
β	6.0
ϵ	$\{1/255, 2/255, 8/255\}$
Βήματα PGD	10

Table 1.5: Υπερπαράμετροι για την ανταγωνιστική (TRADES) εκπαίδευση.

Αποτελέσματα σε Καθαρές Εικόνες. Στον Πίνακα 1.6 παρουσιάζεται η ακρίβεια στο σύνολο ελέγχου για διάφορες τιμές ϵ. Παρότι ελαφρώς μειωμένη σε σχέση με την απλή εκπαίδευση, η μείωση είναι ήπια.

Σύνολο Δεδομένων	$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 8/255$
Υγιής vs Πνευμονία	94.48	94.02	94.09
Υγής vs Μεγαλοκαρδία	88.17	85.84	88.35
ΟΤΣ	97.44	97.73	96.95
MRI Εγκεφάλου	99.74	99.39	99.23

Table 1.6: Ακρίβεια (%) σε καθαρές εικόνες για ταξινομητές εκπαιδευμένους με την μέθοδο TRADES.

Παρατηρήσεις και Επόμενα Βήματα

Συνοψίζοντας, τόσο οι απλοί όσο και οι ανταγωνιστικά εκπαιδευμένοι ConvNeXt-Base ταξινομητές εμφανίζουν υψηλή ακρίβεια στα τέσσερα σύνολα δεδομένων. Ενώ η ανταγωνιστική εκπαίδευση μειώνει ελαφρώς την απόδοση σε καθαρές εικόνες, προσφέρει πιο "ανθεκτικά" (gradients)—ιδιαίτερα χρήσιμα για την παραγωγή *κλινικά ουσιαστικών* αντιπαραδειγματικών εικόνων, όπως θα φανεί στα επόμενα πειράματα.

1.4.3 Αποτελέσματα Αναστροφής ΠΑΔ

Αξιολογούμε το **pipeline αναστροφής ΠΑΔ (GAN inversion)**—μια *προσέγγιση βασισμένη σε encoder* ακολουθούμενη από *pivotal tuning* (PTI)—και στα τέσσερα σύνολα δεδομένων. Παρακάτω παρουσιάζονται τόσο ποσοτικά όσο και ποιοτικά αποτελέσματα της προτεινόμενης μεθόδου ανακατασκευής.

Ποσοτική Ανάλυση

Μετράμε την πιστότητα ανακατασκευής με τέσσερις βασικές μετρικές:

- **FID (Fréchet Inception Distance)**: Μικρότερες τιμές υποδηλώνουν καλύτερη ευθυγράμμιση των κατανομών των πραγματικών και των ανακατασκευασμένων εικόνων.
- **CMMD (Conditional MMD)**: Αξιολογεί την ομοιότητα στις κατανομές υπό συνθήκη κλάσης.
- **MSE (Μέσο Τετραγωνικό Λάθος)**: Αποτυπώνει το σφάλμα ανά εικονοστοιχείο μεταξύ πρωτότυπων και ανακατασκευασμένων.

- **LPIPS (Learned Perceptual Image Patch Similarity)**: Αντικατοπτρίζει την (perceptual)αντιληπτική ομοιότητα σε επίπεδο χαρακτηριστικών.

Στον Πίνακα 1.7 παρουσιάζονται τα αποτελέσματα για τέσσερις αντιπροσωπευτικές κλάσεις: *Μεγαλοκαρδία* και *Πνευμονία* (ακτινογραφίες θώρακα), *Drusen* (ΟΤΣ) και *Μέτρια Βλάβη* (MRI Εγκεφάλου). Παρατηρούμε τα εξής:

- **Ακτινογραφίες Θώρακα (Μεγαλοκαρδία, Πνευμονία)**: Γενικά επιτυγχάνεται χαμηλότερο FID, MSE και LPIPS, υποδηλώνοντας πιο απλά ανατομικά μοτίβα που είναι ευκολότερο να ανακατασκευαστούν.
- **ΟΤΣ (Drusen)**: Υψηλότερα MSE και FID, λόγω λεπτομερών υφών του αμφιβληστροειδούς.
- **MRI Εγκεφάλου (Μέτρια Άνοια)**: Πολύ χαμηλό FID (10.18), δείγμα αυξημένου ρεαλισμού.

Κλάση	**FID ↓**	**CMMD ↓**	**MSE ↓**	**LPIPS ↓**
Μεγαλοκαρδία (X-ray)	19.76	0.056	6.28e-4	0.0135
Πνευμονία (X-ray)	21.91	0.061	8.57e-4	0.0150
Drusen (ΟΤΣ)	26.27	1.031	5.02e-3	0.00712
Μέτρια Άνοια (MRI)	10.18	0.580	1.44e-3	0.00860

Table 1.7: Ποσοτικές μετρήσεις ανακατασκευής για αντιπροσωπευτικές κλάσεις στα τέσσερα σύνολα δεδομένων.

Ποιοτικά Παραδείγματα

Παρουσιάζουμε ένα δείγμα ανά σύνολο δεδομένων παρακάτω, όπου εμφανίζονται (1) η αρχική εικόνα, (2) η ανακατασκευή μόνο από τον encoder, (3) ο χάρτης διαφορών μετά το στάδιο του encoder, (4) η τελική ανακατασκευή έπειτα από την εφαρμογή του Pivotal Tuning, και (5) ο τελικός χάρτης διαφορών.

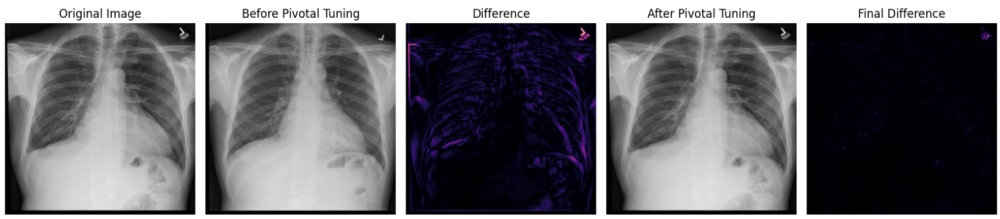


Figure 1.4.5: Αποτέλεσμα αναστροφής ΠΑΔ για ακτινογραφία με μεγαλοκαρδία. Προσέξτε τη βελτίωση μετά το Pivotal Tuning στην απόδοση των ορίων της καρδιάς.

Figure 1.4.6: Αναστροφή σε περίπτωση πνευμονίας. Το στάδιο PTI εξαλείφει μικρά τεχνητά σφάλματα και βελτιώνει τις λεπτομέρειες στο πεδίο των πνευμόνων.

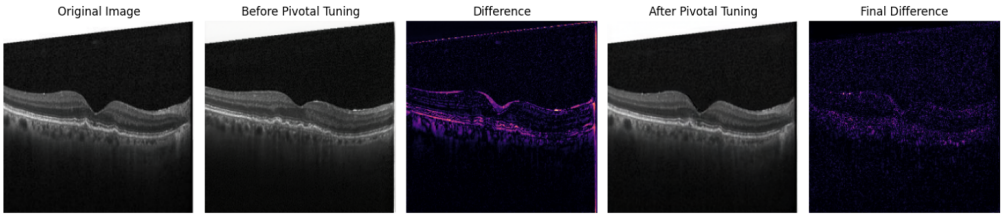


Figure 1.4.7: Ανακατασκευή εικόνας ΟΤΣ με Drusen. Η έκδοση μόνο από τον encoder παραλείπει λεπτομερείς υφές, ενώ το PTI αποκαθιστά καλύτερα την δομή του αμφιβληστροειδούς.

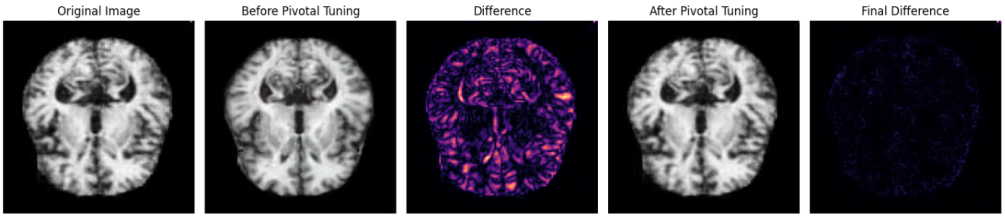


Figure 1.4.8: MRI με μέτρια βλάβη. Το PTI συμβάλλει στην πιο πιστή απόδοση του φλοιού και των κοιλιών του εγκεφάλου.

Κύριες Παρατηρήσεις

- **Επίδραση του Pivotal Tuning**: Σε όλα τα σύνολα δεδομένων, το PTI βελτιώνει αισθητά τις λεπτομέρειες και μειώνει τα τεχνητά σφάλματα (artifacts).
- **Πολυπλοκότητα Δεδομένων**: Στην περίπτωση των ακινογραφιών λάβαμε τα καλύτερα αποτελέσματα από άποψη ακρίβειας ανακατασκευής και ρεαλισμού. Οι μεγαλύτερες ανακρίβειες ανακατασκευής σε επίπεδο εικονοστοιχείων προέκυψαν στις εικόνες ΟΤΣ κυρίως λόγω της φύσης αυτών(λεπτή υφή ,αρκετός θόρυβος).
- **Κλινική Σημασία**:Και στα τέσσερα σύνολα δεδομένων καταφέραμε και ανακατασκευάσαμε επαρκώς τις βασικές ανατομικές δομές των εικόνων,γεγονός το οποίο κρίνεται αναγκαίο στην συνέχεια για την παραγωγή αραιών και ρεαλιστικών αντιπαραδειγματικών εικόνων.

Συνολικά, ο συνδυασμός encoder-based αναστροφής με PTI αποδίδει ακριβείς ανακατασκευές για πληθώρα ιατρικών εικόνων, εξασφαλίζοντας τόσο υψηλή πιστότητα όσο και δυνατότητα επεξεργασίας (editability) στο λανθάνων χώρο—βασικό στοιχείο για το αντιπαραδειγματικό μας πλαίσιο.

1.4.4 Παραγωγή Αντιπαραδειγματικών Εικόνων

Αξιολογούμε τις τελικές **αντιπαραδειγματικές εικόνες** του πλαισίου μας (SPRUCE) σε απλούς αλλά και ανταγωνιστικά εκπαιδευμένους ταξινομητές ConvNeXt-Base ($\epsilon \in \{0, 1/255, 2/255, 8/255\}$). Στόχος είναι να μετατρέψουμε εικόνες με παθολογία (μεγαλοκαρδία, πνευμονία, drusen, μέτρια άνοια) σε «υγιείς» εκδοχές και να αξιολογήσουμε τόσο ποσοτικά (FID, CMMD, Flip Ratio, L1 απόσταση, σιγουριά ταξινομητή) όσο και ποιοτικά την αποτελεσματικότητα τους.

Ποσοτικά Αποτελέσματα

Στους Πίνακες 1.8 και 1.9 παρουσιάζονται οι επιδόσεις και στα δύο σύνολα δεδομένων ακτινογραφιών θώρακα. Σημαντικά συμπεράσματα:

- **Επίδραση της Ανταγωνιστικής Ευρωστίας**: Μη μηδενικές τιμές ϵ οδηγούμαστε στην παραγωγή πιο ρεαλιστικών αντιπαραδειγματικών εικόνων (χαμηλότερο FID/CMMD) αλλά και σε ελαφρώς αυξημένη L1 απόσταση (λιγότερη «αραιότητα»).

- **Σταθερά Υψηλά Flip Ratios**: Σχεδόν σε ολές τις περιπτώσεις ξεπερνάται το 95% σε αντιστροφή απόφασης, υποδεικνύοντας αξιόπιστη αλλαγή ταξινόμησης.

- **Συσχέτιση Αραιότητας–Εμπιστοσύνης**: Μεγαλύτερα ϵ συχνά αυξάνουν την σιγουριά του ταξινομητή στην «υγιή» κλάση, με το αντάλλγμα για πιο εκτεταμένες επεμβάσεις στην αρχική εικόνα.

Epsilon	**FID ↓**	**CMMD ↓**	**Flip ratio ↑**	**L1 ↓**	**Σιγουριά ↑**
0	60.22	0.35	**100%**	**0.0118**	0.8601
1/255	56.34	0.315	**100%**	0.0168	0.9866
2/255	58.11	0.326	**100%**	0.0184	**0.9958**
8/255	**49.632**	**0.264**	98%	0.0269	0.9744

Table 1.8: Σύγκριση διαφορετικών τιμών epsilon και οι επιπτώσεις τους σε FID, CMMD, flip ratio, L1 και σιγουριά ταξινομητή (πνευμονία).

Epsilon	**FID ↓**	**CMMD ↓**	**Flip ratio ↑**	**L1 ↓**	**Σιγουριά ↑**
0	43.20	0.192	96.8%	**0.0103**	0.7513
1/255	40.069	0.185	**98.8%**	0.0112	0.7796
2/255	**39.410**	0.169	98.6%	0.0144	0.9221
8/255	40.256	**0.166**	98.6%	0.0188	**0.9569**

Table 1.9: Σύγκριση διαφορετικών τιμών epsilon και οι επιπτώσεις τους σε FID, CMMD, flip ratio, L1 και σιγουριά ταξινομητή (μεγαλοκαρδία).

Ποιοτικές Συγκρίσεις

Απλός vs. Ανταγωνιστικά Ανθεκτικός Ταξινομητής

Στα Σχήματα 1.4.9 και 1.4.10 συγκρίνονται αντιπαραδειγματικές εικόνες, όπως προκύπτουν από απλό ($\epsilon = 0$) και από ανταγωνιστικά ανθεκτικό ($\epsilon > 0$) ταξινομητή. Οι ανθεκτικοί ταξινομητές δίνουν πιο συνεκτικές και στοχευμένες στις παθολογικές περιοχές επεμβάσεις (π.χ. μέγεθος και σχήμα καρδιάς, περιοχή των πνευμόνων). Οι απλοί ταξινομητές εμφανίζουν ενίοτε διάχυτες, λιγότερο κλινικά σχετικές αλλαγές.

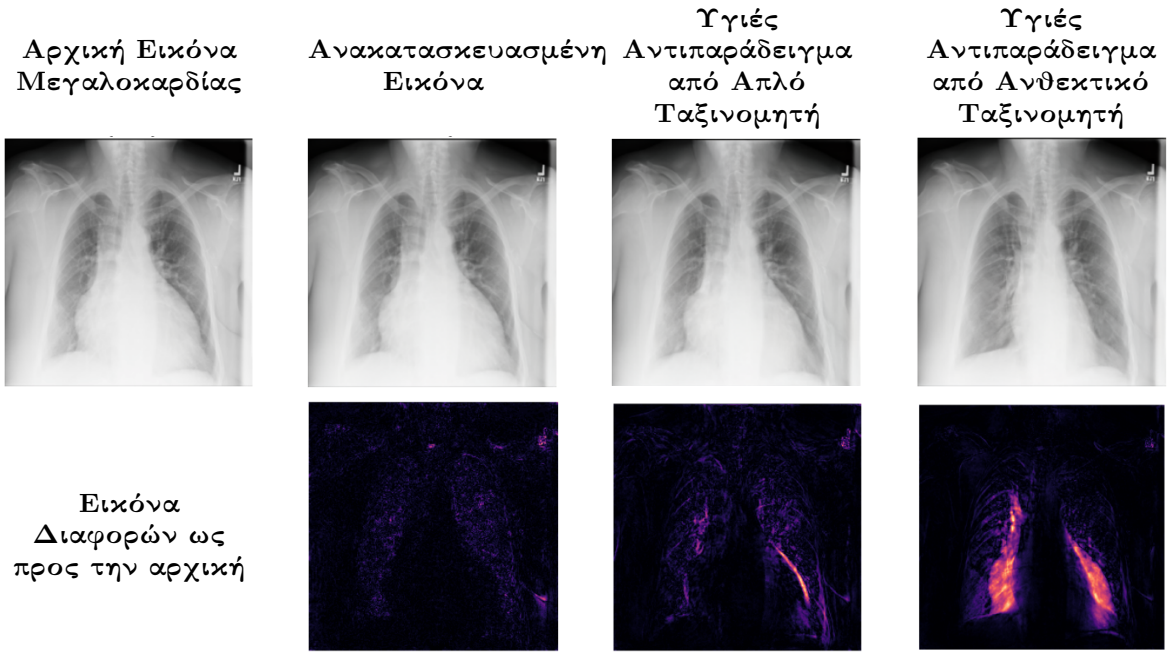


Figure 1.4.9: Σύγκριση υγιών αντιπαραδειγμάτων για μεγαλοκαρδία, μεταξύ απλού και ανταγωνιστικά ανθεκτικού ταξινομητή.

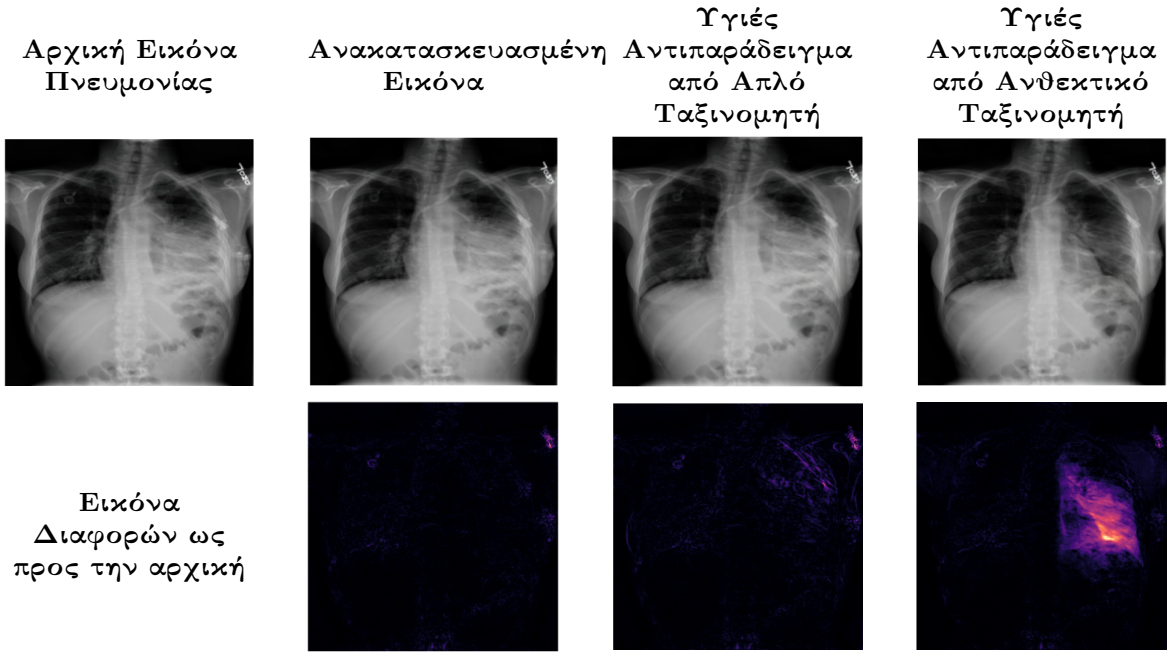


Figure 1.4.10: Σύγκριση υγιών αντιπαραδειγμάτων για πνευμονία, μεταξύ απλού και ανταγωνιστικά ανθεκτικού ταξινομητή.

Αντιπαραδείγματα για Ανθεκτικούς Ταξινομητές

Για καθένα από τα τέσσερα σύνολα δεδομένων, δημιουργήσαμε «υγιείς» αντιπαραδειγματικές εικόνες για ανταγωνιστικά εκπαιδευμένα μοντέλα ($\epsilon > 0$). Τα Σχήματα 1.4.11, 1.4.12, 1.4.13 και 1.4.14 δείχνουν ορισμένα αποτελέσματα:

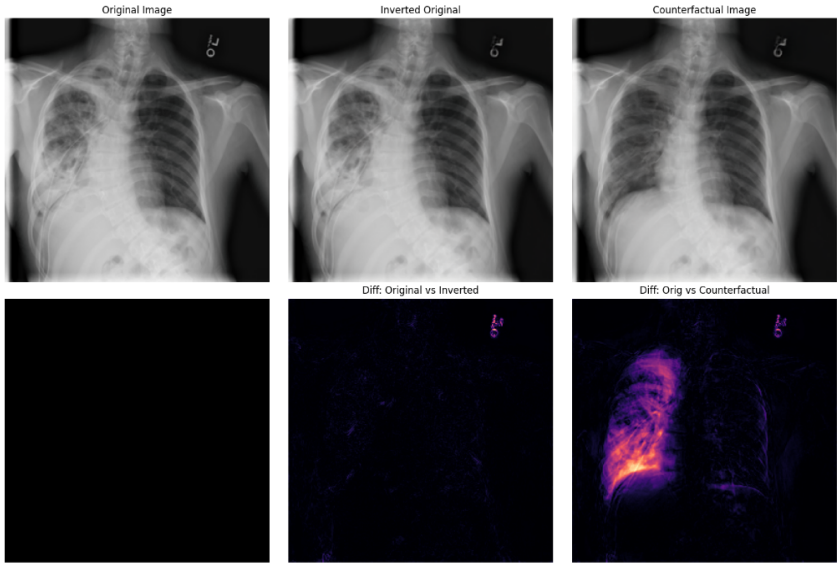


Figure 1.4.11: Υγιές αντιπαράδειγμα για περίπτωση πνευμονίας.

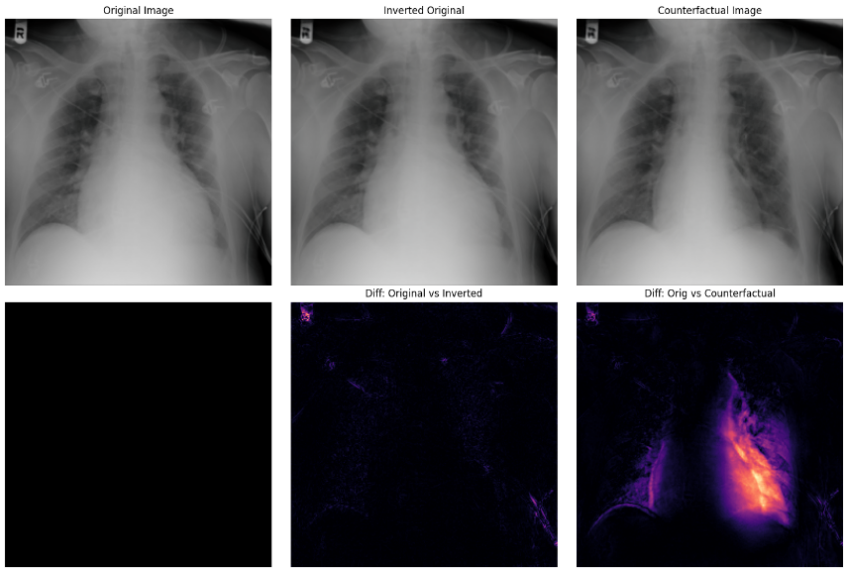


Figure 1.4.12: Υγιές αντιπαράδειγμα για εικόνα μεγαλοκαρδίας.

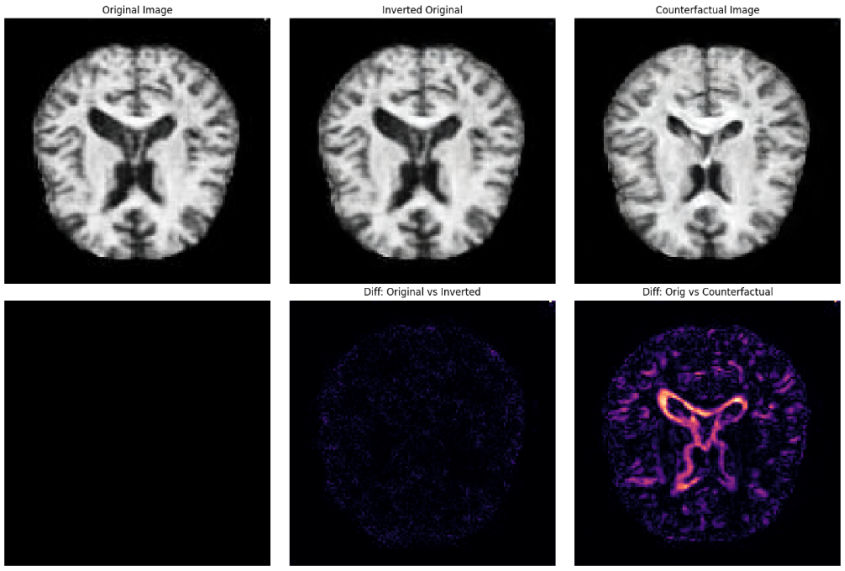


Figure 1.4.13: Υγιές αντιπαράδειγμα για μέτρια άνοια. Αναστρέφονται μερικώς οι ατροφικές περιοχές, ενώ το μέγεθος των κοιλιών επανέρχεται πιο κοντά στο φυσιολογικό.

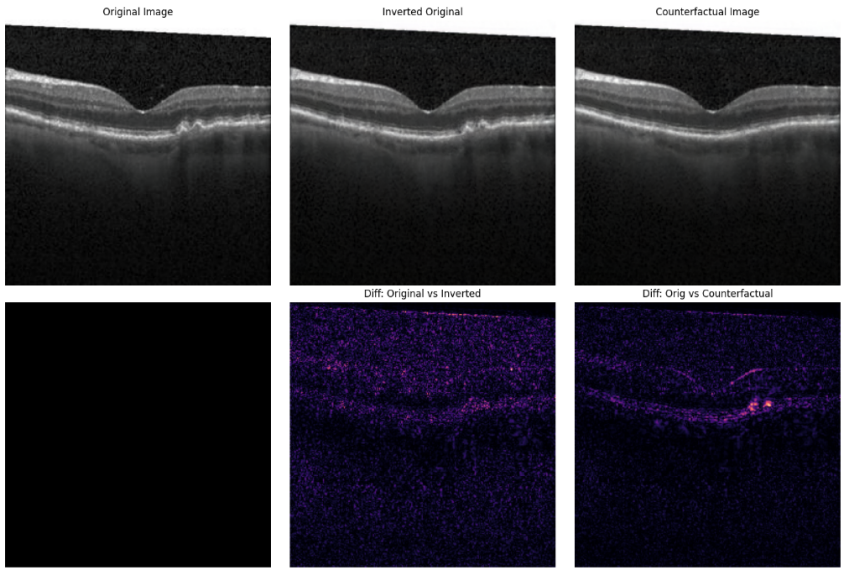


Figure 1.4.14: Υγιές αντιπαράδειγμα για Drusen. Οι εναποθέσεις μειώνονται, ενώ οι στιβάδες του αμφιβληστροειδούς γίνονται πιο ομοιόμορφες.

Παρατηρήσεις και Συμπεράσματα

- **Στοχευμένες Τροποποιήσεις με Ανθεκτικούς Ταξινομητές**: Οι χάρτες διαφορών εστιάζονται στις παθολογικές περιοχές, βελτιώνοντας την ερμηνευσιμότητα.
- **Συμβιβασμός για την τιμή** ϵ: Μεγαλύτερη τιμή ϵ ενδέχεται να ενισχύει τον ρεαλισμό, αλλά αυξάνει τις παραγόμενες αλλαγές (L1). Κάθε σύνολο δεδομένων μπορεί να απαιτεί άλλη επιλογή ϵ για σωστή ισορροπία.
- **Υψηλό Flip Ratio σε Όλες τις Κλάσεις**: Ακόμα και σε απαιτητικές περιπτώσεις (π.χ. Drusen ή άνοια), το πλαίσιο αντιστρέφει αξιόπιστα την απόφαση του ταξινομητή, παράγοντας έγκυρες αντιπαραδειγματικές εικόνες.

Συνολικά, τα πειράματα παραγωγής αντιπαραδειγματικών εικόνων δείχνουν *γιατί* η ανταγωνιστική ανθεκτικότητα είναι κρίσιμη: τα ανθεκτικά gradients ταξινομητών παράγουν κλινικά ρεαλιστικές τροποποιήσεις που μετατρέπουν παθολογικές εικόνες σε πειστικές «υγιείς» εκδοχές, διατηρώντας παράλληλα τα βασικά ανατομικά χαρακτηριστικά των ιατρικών απεικονίσεων.

1.4.5 Επίδραση των Όρων της Συνάρτησης Απώλειας

Σε αυτή την ενότητα μελετάμε πώς κάθε συνιστώσα της συνάρτησης απώλειας για τα αντιπαραδείγματα—δηλαδή ο όρος κανονικοποίησης στον χώρο των pixel (λ_{image}), ο όρος αντιληπτικής ομοιότητας (λ_{perc}) και ο όρος κανονικοποίησης στον λανθάνοντα χώρο (λ_{latent})—επηρεάζει τις παραγόμενες αντιπαραδειγματικές εικόνες. Διεξάγουμε τα πειράματα σε παραδείγματα μεγαλοκαρδίας και πνευμονίας, χρησιμοποιώντας τον καλύτερο ανθεκτικό ταξινομητή βάσει CMMD. Για κάθε πείραμα, μεταβάλουμε μία υπερπαράμετρο και κρατάμε σταθερές τις άλλες στις τιμές: $\lambda_{\text{perc}} = 0.1$, $\lambda_{\text{latent}} = 10^{-4}$, $\lambda_{\text{image}} = 10^{-6}$.

Ποσοτικά Αποτελέσματα

Η αξιολόγηση βασίζεται όπως και πριν στις μετρικές FID, CMMD, flip ratio, L1 απόσταση και σιγουριά ταξινομητή, ώστε να εξεταστούν ο ρεαλισμός, η αποτελεσματικότητα τροποποίησης και η αραιότητα. Ο Πίνακας 1.10 συνοψίζει τα αποτελέσματα για τις εικόνες πνευμονίας.

Μεταβαλλόμενη Υπερπαράμετρος	FID ↓	CMMD ↓	Flip (%) ↑	L1 ↓	Εμπιστοσύνη ↑
$\lambda_{\text{image}} = 10^{-8}$	**48.59**	0.240	98.8	0.0434	0.9807
$\lambda_{\text{image}} = 10^{-6}$	49.63	0.264	98.0	0.0269	0.9744
$\lambda_{\text{image}} = 10^{-5}$	56.64	0.333	95.0	**0.0125**	0.8913
$\lambda_{\text{perc}} = 0.01$	48.66	**0.213**	98.5	0.0294	0.9803
$\lambda_{\text{perc}} = 0.1$	49.63	0.264	98.0	0.0269	0.9744
$\lambda_{\text{perc}} = 1$	58.12	0.377	95.2	0.0229	0.8991
$\lambda_{\text{latent}} = 10^{-3}$	51.31	0.250	94.4	0.0318	0.8811
$\lambda_{\text{latent}} = 10^{-4}$	49.63	0.264	98.0	0.0269	0.9744
$\lambda_{\text{latent}} = 10^{-6}$	53.94	0.320	**99.2**	0.0215	**0.9841**

Table 1.10: Επίδραση κάθε συνιστώσας της συνάρτησης απώλειας στην ποιότητα των παραγόμενων αντιπαραδειγμάτων (πνευμονία).

Ποιοτικά Αποτελέσματα

Για επιπρόσθετη ανάλυση, οπτικοποιούμε τα αντιπαραδείγματα που δημιουργούνται μεταβάλλοντας το λ_{image}. Στο Σχήμα 1.4.15 απεικονίζεται η αρχική εικόνα μεγαλοκαρδίας, τα παραγόμενα αντιπαραδείγματα και οι χάρτες διαφορών για διαφορετικά επίπεδα κανονικοποίησης.

Συζήτηση

- **Όρος κανονικοποίησης σε pixel-level (λ_{image})**: Μικρότερες τιμές επιτρέπουν την παραγωγή πιο ρεαλιστικών ,σύμφωνα με τις αντίστοιχες μετρικές, αντιπαραδειγματικών εικόνων των οποίων όμως οι αλλαγές είναι και λιγότερο αραιές.Αυτό φαίνεται και από την σύγκριση των εικόνων διαφοράς του Σχήματος 1.4.15.Αντίθετα, αύξηση της συγκεκριμένης υπερπαραμέτρου οδηγεί σε μικρότερης έκτασης και έντασης τροποποιήσεις.

- **Αντιληπτικός Όρος (λ_{perc})**: Χαμηλότερες τιμές μπορούν να ενισχύσουν τον ρεαλισμό και το ποσοστό αλλαγής απόφασης, αλλά επεκτείνουν τις τροποποιήσεις. Υψηλότερες τιμές οδηγούν σε καλύτερη διατήρηση αντιληπτικών χαρακτηριστικών, αυξάνοντας την αραιότητα.

- **Όρος κανονικοποίησης στον λανθάνοντα χώρο (λ_{latent})**: Αυξημένη κανονικοποίηση στο λανθάνοντα χώρο μειώνει την σιγουριά του ταξινομητή και το flip rate και αυξάνει τον ρεαλισμό , ενώ πιο χαλαρή κανονικοποίηση επιτρέπει ισχυρότερες μετατοπίσεις στον λανθάνοντα χώρο υπονομεύοντας τον ρεαλισμό αλλά βελτιώνοντας την αραιότητα της παραγόμενης εικόνας.

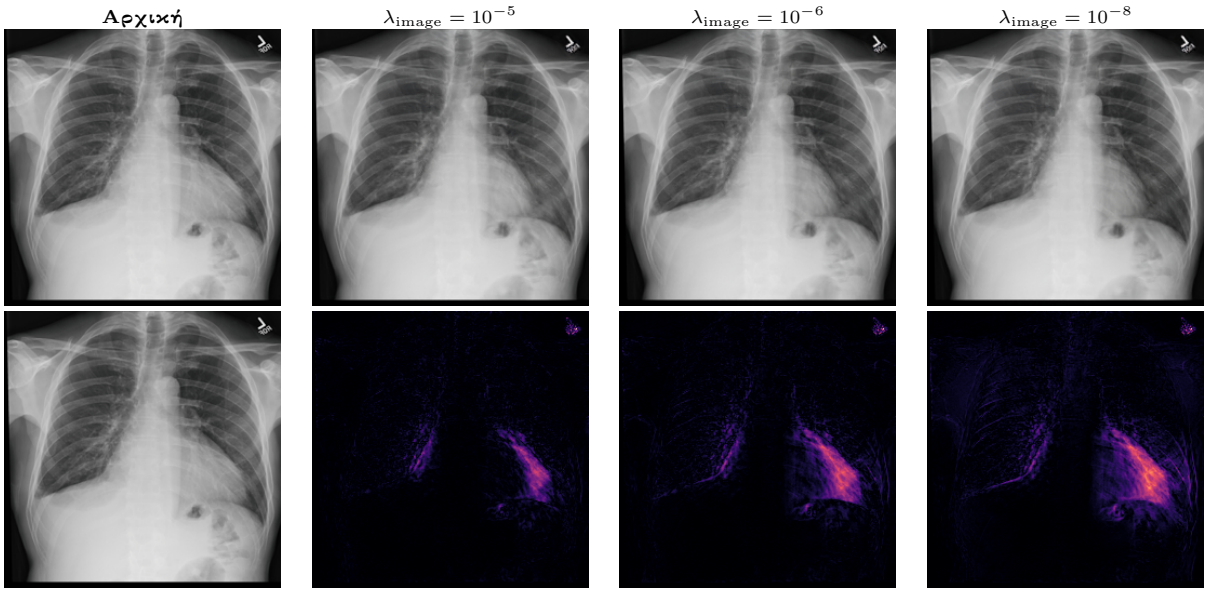


Figure 1.4.15: Επίδραση της λ_{image} στα παραγόμενα αντιπαραδείγματα μεγαλοκαρδίας και στις αντίστοιχες εικόνες διαφορών.

- **Βασικός Συμβιβασμός**: Οι παράμετροι αυτοί ρυθμίζουν την ισορροπία μεταξύ ρεαλισμού, έκτασης αλλαγών και αποτελεσματικότητας ταξινόμησης. Κάθε όρος συμβάλλει καθοριστικά στη δημιουργία οπτικά πειστικών και κλινικά χρήσιμων αντιπαραδειγμάτων.

Από την ανάλυση καθίσταται σαφές ότι ο σχεδιασμός της συνάρτησης απώλειας είναι ευέλικτος και αποτελεσματικός. Με κατάλληλη ρύθμιση των υπερπαραμέτρων, ο χρήστης μπορεί να δώσει προτεραιότητα είτε στην αραιότητα(sparsity) των αλλαγών,είτε στην σιγουριά της ταξινόμησης στην αντιπαραδειγματική κλάση είτε στο επίπεδο ρεαλισμού—ανάλογα με τις κλινικές ή ερμηνευτικές απαιτήσεις.

1.5 Συμπεράσματα

1.5.1 Συζήτηση

Στην παρούσα διπλωματική εργασία παρουσιάσαμε ένα καινοτόμο πλαίσιο για την παραγωγή ρεαλιστικών, αραιών και ερμηνεύσιμων ιατρικών αντιπαραδειγματικών εξηγήσεων. Η προσέγγισή μας συνδυάζει έναν σύγχρονο ταξινομητή τύπου ConvNeXt (τόσο σε απλή όσο και σε ανταγωνιστικά ανθεκτική εκδοχή) με ένα μοντέλο παραγωγής εικόνων StyleGAN2-ADA, ενισχυμένο από τεχνικές αναστροφής ΠΑΔ (E4E και Pivotal Tuning Inversion). Ο σχεδιασμός αυτός επιτρέπει την ανακατασκευή εικόνων υψηλής πιστότητας, διατηρώντας ταυτόχρονα κρίσιμες ανατομικές λεπτομέρειες που είναι απαραίτητες για την κλινική ερμηνεία.

Κύριες Συνεισφορές.

- Αναπτύξαμε μια διαδικασία για την **αναγνώριση και τροποποίηση των λεγόμενων αποκλινόντων χαρακτηριστικών** (μέσω της μεθόδου PIECE), με στόχο την παραγωγή αντιπαραδειγματικών εικόνων που αναδεικνύουν τις πλέον σχετικές ανατομικές περιοχές σε παθήσεις όπως η μεγαλοκαρδία, η πνευμονία και οι αμφιβληστροειδικές αλλοιώσεις.
- Μέσω της **ανταγωνιστικής εκπαίδευσης** (*TRADES*) του ταξινομητή, αποδείξαμε ότι τα πιο εύρωστα gradients ταξινομητών οδηγούν σε σημασιολογικές και κλινικά συσχετιζόμενες μεταβολές, σε αντίθεση με τις αντιπαραδειγματικές εικόνες απλών ταξινομητών που μπορεί να περιέχουν τυχαίες ή διάχυτες αλλαγές.
- **Επικυρώσαμε τη μέθοδο** τόσο ποσοτικά (με δείκτες όπως το FID, CMMD, Flip Ratio και L1 απόσταση) όσο και ποιοτικά. Τα αποτελέσματα έδειξαν ότι, υπό κατάλληλες τιμές υπερπαραμέτρων,οι

αλλάγες που περιέχουν οι αντιπαραδειγματικές εικόνες παραμένουν αραιές και κλινικά συναφείς διατηρώντας παράλληλα τα εξατομικευμένα χαρακτηριστικά του ασθενούς.

- Το πλαίσιο **αναστροφής ΠΑΔ** (E4E + PTI) επέτρεψε την ακριβή ανακατασκευή των αρχικών εικόνων,ενώ παράλληλα μάς έδωσε την δυνατότητα να ελέγξουμε και να επεξεργαστούμε με μεγάλη ευκολία τα διανύσματα στον λανθάνοντα χώρο του ΠΑΔ.

Σημασία και Αντίκτυπος. Το προτεινόμενο πλαίσιο συμβάλλει στο ευρύτερο πεδίο της *Ερμηνεύσιμης Τεχνητής Νοημοσύνης (XAI)* στην υγεία, ανταποκρινόμενο στην ανάγκη για αξιόπιστες και διαφανείς μεθόδους ερμηνείας σε ιατρικές εικόνες. Μέσα από την παραγωγή αραιών αλλά ανατομικά εύλογων επεμβάσεων, οι κλινικοί γιατροί μπορούν να κατανοούν καλύτερα το ελάχιστο σύνολο αλλαγών που επηρεάζει την απόφαση του μοντέλου, ενισχύοντας έτσι την εμπιστοσύνη σε μοντέλα που λειτουργούν ως "μαύρα κουτιά".

Περιορισμοί. Παρά τα ενθαρρυντικά αποτελέσματα, θα πρέπει να αναφερθούμε επίσης σε κάποιους περιορισμούς του πλαισίου. Οι μέθοδοι που βασίζονται σε ΠΑΔ δυσκολεύονται αρκετά σε περιπτώσεις σπάνιων παθολογιών ή σε ετερογενή σύνολα δεδομένων όπου τα δεδομένα εκπαίδευσης είναι αρκετά περιορισμένα. Επιπλέον, ενώ η ανταγωνιστική ευρωστία μπορεί να ενισχύει την ερμηνευσιμότητα, ενδέχεται να μειώνει ελαφρώς την ακρίβεια του ταξινομητή.

Τελικές Παρατηρήσεις. Συνολικά, τα πειράματά μας δείχνουν ότι ο συνδυασμός ανθεκτικών ταξινομητών με μια ισχυρή διαδικασία αναστροφής ΠΑΔ μπορεί να προσφέρει αξιόπιστες και κατατοπιστικές αντιπαραδειγματικές εξηγήσεις—ένα σημαντικό βήμα προς ασφαλέστερες εφαρμογές της τεχνητής νοημοσύνης στον τομέα της ιατρικής απεικόνισης. Μέσα από εκτενείς ποσοτικές και ποιοτικές αναλύσεις, αποδείξαμε τη βιωσιμότητα της προσέγγισής μας σε πολλαπλά σύνολα δεδομένων.

Εκτιμούμε ότι το προτεινόμενο πλαίσιο μπορεί να αποτελέσει τη βάση για διάφορες ιατρικές εφαρμογές, όπου ειδικοί θα μπορούν να εξετάζουν άμεσα πώς μικρές αλλαγές στις εικόνες επηρεάζουν την πρόβλεψη, ενισχύοντας την κατανόηση τόσο των δεδομένων του ασθενούς όσο και της συμπεριφοράς των μοντέλων.

1.5.2 Μελλοντικές Κατευθύνσεις

Βάσει των ευρημάτων και των περιορισμών αυτής της μελέτης, προτείνονται τα εξής πεδία μελλοντικής έρευνας:

- **Μελέτες Χρηστών με Επιστήμονες από τον τομέα της Υγείας:** Διεξαγωγή μελετών με ακτινολόγους ή άλλους ειδικούς για την αξιολόγηση του ρεαλισμού, της ερμηνευσιμότητας και της κλινικής εγκυρότητας των παραγόμενων αντιπαραδειγματικών εξηγήσεων από το πλαίσιο μας.
- **Προηγμένα Παραγωγικά Μοντέλα:** Διερεύνηση εναλλακτικών μοντέλων παραγωγής εικόνας (π.χ. diffusion models ή βελτιωμένες αρχιτεκτονικές ΠΑΔ) που ενδεχομένως αποδίδουν καλύτερα στις πολύπλοκες ιατρικές εικόνες.
- **Συγκριτική Αξιολόγηση Αρχιτεκτονικών:** Επέκταση του πλαισίου σε άλλες αρχιτεκτονικές ταξινομητών (π.χ. DenseNet, ResNet, ViT) ώστε να διασφαλιστεί η γενίκευση και η συνέπεια των αποτελεσμάτων.
- **Τρισδιάστατες Ιατρικές Απεικονίσεις:** Επέκταση του πλαισίου σε τρισδιάστατα δεδομένα (π.χ. αξονικές ή μαγνητικές τομογραφίες) για να διευρυνθεί η εφαρμογή σε πιο σύνθετα σενάρια ιατρικής απεικόνισης.
- **Συγκριτικές Μελέτες Ερμηνευσιμότητας:** Συγκριτική αξιολόγηση των αντιπαραδειγματικών εξηγήσεων έναντι άλλων δημοφιλών τεχνικών επεξηγηματικότητας (π.χ. Grad-CAM, Integrated Gradients), για πληρέστερη επικύρωση και ερμηνεία.

Chapter 2

Introduction

2.1 Motivation

Deep learning models are at the forefront of image applications and are particularly prevalent in medical image classification [74, 8, 92, 56, 133, 101, 87]. One major limitation of these models is that they function as black boxes i.e., their decision-making process is opaque to humans, and often rely heavily on statistical data dependencies inherent in the training data and do not capture domain-specific knowledge [124, 137]. Although this may be acceptable in other domains, in medical imaging poor decisions made by a classifier may negatively affect healthcare outcomes and pose risks to patients. This has prompted reactions by both the research community, with the development of techniques for eXplainable Artificial Intelligence (XAI) [37, 119], and the legal community, with the introduction of laws that establish the "right to explanation", the right to be given an explanation for a decision made by an AI system that significantly affects an individual.[1] This limitation poses a substantial barrier to the integration of these models into mainstream applications.

Many different approaches have been developed for explaining black box image classifiers. One such approach is counterfactual image generation [13, 47, 36, 27]. Counterfactual explanations provide a new input, similar to the original one, that produces a different output, essentially describing the necessary changes to the input to elicit changes to the output. This type of explanation is very similar to human decision processes and has been shown to be easily interpretable by humans [34, 84, 120]. Additionally, the pixel-wise differences between the original and the counterfactual image can serve as a saliency map [83], with the added benefit that the counterfactual image serves as a certificate of correctness for the saliency map.

Technically, any image that produces a different output from the original one can serve as a counterfactual, but in order to provide insights the counterfactual image ought to present only the absolute necessary changes to the input image to flip the classifier's output. As an example, for a classifier that detects pneumonia in chest CT scans, we would prefer the area of change to be completely within the lungs and not in the background or even the shape of the skeleton, as that could represent a completely different patient and not the same patient with a different outcome. A major limitation of many works in the relevant literature [5, 116, 35, 80, 14] is that they do not introduce any mechanisms in the counterfactual generation process which ensure that the counterfactual image presents sparse, perceptually similar, and visually coherent changes. In this work, we employ several regularization penalties in the counterfactual generation processes that induce the aforementioned properties, namely L1 regularization in both the image space and the latent space of the generator and Learned Perceptual Image Patch Similarity loss [138].

Another major limitation of many other works [86, 5, 61, 116, 105, 80, 14] is that they couple in some way the training of the counterfactual generator to the training of the model under inspection. Some works train the generator and the model jointly, thus being unable to apply their framework to any arbitrary pre-trained model, while others train the generator on the outputs of the model, having to retrain the generator for each new model under inspection.

[1]

2.2 Contribution

In this work, we employ and extend the method of exceptional feature extraction from PIECE [54], which allows a completely unsupervised pre-training of the counterfactual generator, and only requires the calculation of some compute-light statistics about the activations of the penultimate layer of the classifier. This technique also poses no restriction on the number of output classes of the classifier and is able to produce counterfactuals from any class to any other, which is not true for many other works.

We propose a novel framework, SPRUCE (SParse Realistic and Uncoupled Counterfactual Explanations), for sparse counterfactual medical image generation. Our contributions are summarized as follows:

- We introduce SPRUCE, a novel framework for generating sparse counterfactual medical images.
- We develop a specialized loss function that enforces both sparsity and image fidelity in the generated counterfactuals, minimizing irrelevant modifications.
- We employ Generative Adversarial Networks (GANs), along with Encoding for Editing (E4E) [117] and pivotal tuning [96], for superior performance in low-data scenarios, the ability to faithfully reconstruct and generate sharp realistic images, and for enhanced latent space controllability.
- We employ a technique from PIECE [54] that uncouples the training of the counterfactual explainer from the training of the classifier and allows us to use the same explainer for any binary or multi-class classifier trained on the same data distribution.
- We find that the ability to generate sparse, realistic counterfactuals is tied to classifier robustness, making our framework a valuable tool for vunlerability detection as well as model inspection.

2.3 Thesis Outline

This thesis is structured in 6 separate chapters. Below, follows a distinct summary of their content:

- Chapter 2 introduces the motivation behind this research, highlighting the significance of counterfactual explanations in medical imaging, outlines the contributions made, and summarizes the structure of the thesis.
- In the Theoretical Background 3 chapter the theoretical background required to understand this work is provided.It includes detailed analysis on Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), GAN inversion techniques, adversarial training methods, and Explainable Artificial Intelligence (XAI) with a particular emphasis on counterfactual explanations.
- In the Related Work 4 chapter we review relevant existing literature, categorizing previous works into visual counterfactual methods for general image domains and specific methodologies applied within medical imaging
- Chapter 5 details the methodology employed, introducing the proposed SPRUCE framework. This includes the architecture and training strategies for the ConvNeXt classifier, the StyleGAN2-ADA generator, GAN inversion pipeline (E4E and PTI), latent vector optimization process, and strategies to ensure sparsity and realism in generated counterfactuals.
- Chapter 6 includes all the experiments conducted to validate the proposed methods. Results are presented quantitatively and qualitatively, comparing the performance of plain and robust classifiers, evaluating GAN inversion quality, and thoroughly analyzing the effectiveness of the SPRUCE-generated counterfactuals.
- Finally, in chapter 7 we conclude the thesis, summarizing the main findings, discussing their implications, and suggesting potential directions for future research.

Chapter 3

Theoretical Background

Contents

3.1 Convolutional Neural Networks

3.1.1 The neuron

The fundamental processing unit of a Neural network, the neuron, is inspired from the observation of the brain's formation. So just like in the brain where the neurons form a complex, highly interconnected network and send electrical signals to each other to help humans process information, the neurons of the neural networks work together on a common problem, using computing systems to solve mathematical calculations. Specifically, a neural network's neuron processes a local patch of input data, uses learnable weights to compute a weighted sum also adding a learnable bias, applies an activation function, and contributes to the formation of feature maps. The activation function that is being applied interprets the result and presents it in a meaningful way, often introducing non-linearity into the network, allowing it to model more complex patterns.

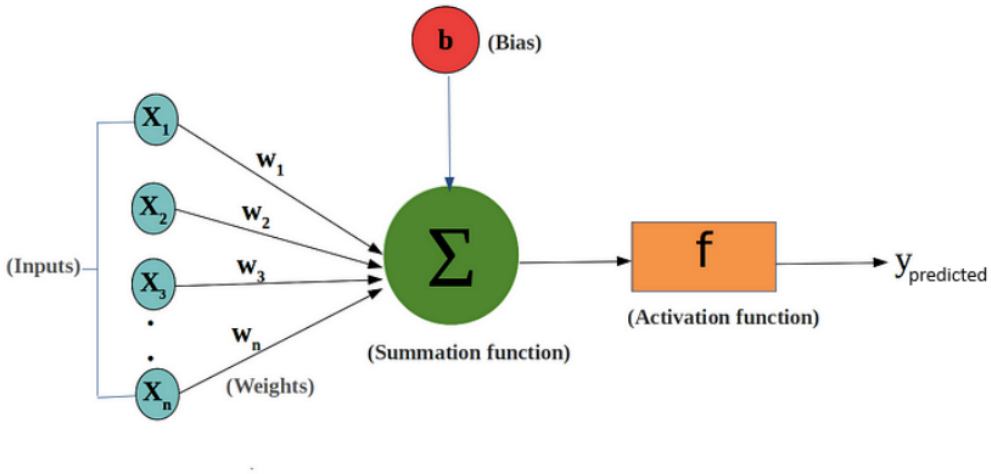


Figure 3.1.1: The neuron architecture, the neuron takes the weighted sum of the input data and the learnable weights, adds a learnable bias, applies an activation function and produces the final output.

Activation functions

Like mentioned above, activation functions are functions that are applied to the weighted sum of the input signals with the weights of the neuron and often they introduce a non-linearity to the computation of the result. Some activation functions that are usually used are:

ReLU Currently the most used in the world, ReLU thresholds at zero by outputting:

$$f(x) = \max(0, x) \tag{3.1.1}$$

This function is easily implemented and eliminates the problem of the saturating gradient. An issue with this function is that all the negative values given as input become zero immediately, which in turn affects the results by not mapping the negative values appropriately.

Leaky-ReLU Leaky-ReLU tries to encounter the issue with ReLU by multiplying the negative inputs with a small constant, instead of completely zeroing negative inputs.

$$f(x) = \mathbb{I}(x < 0)(\alpha x) + \mathbb{I}(x \geq 0)(x) \tag{3.1.2}$$

Sigmoid This activation function is mainly used, due to the fact, that it exists between 0 to 1. Therefore, it is especially used for models where we have to predict the probability as an output. Since probability of anything exists only between the range of 0 and 1, sigmoid is the right choice. Its formula is the following:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.1.3}$$

An issue with this function is that it can cause a neural network to get stuck at the training time.

Tanh The tanh function is an improved version of the sigmoid one, but being zero-centered as shown in 3.1.2 Tanh maps input to the range $[-1, 1]$, instead of $[0, 1]$, with the following formula:

$$\tanh(x) = 2\sigma(2x) - 1 \tag{3.1.4}$$

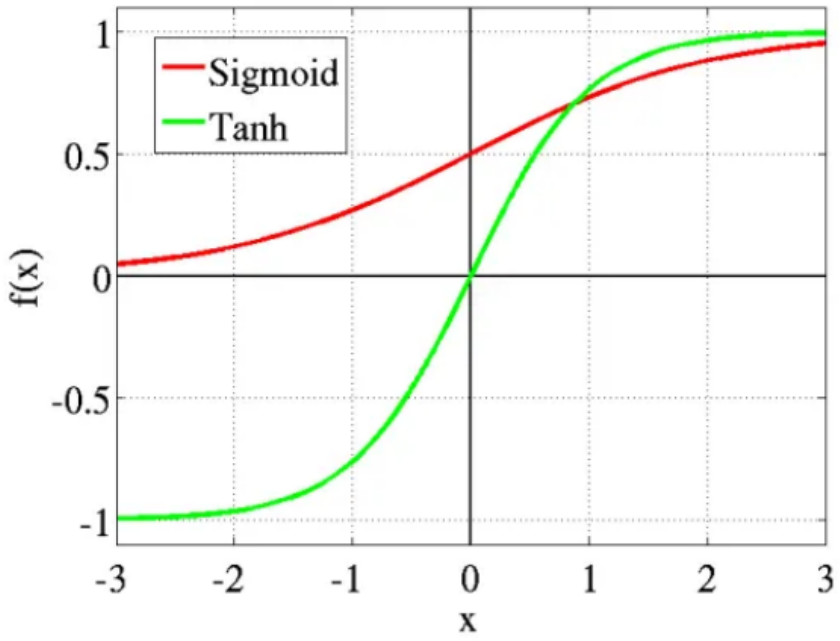


Figure 3.1.2: Sigmoid vs Tanh activation functions from [30].

However, it still may cause the gradient to saturate.

3.1.2 Neural Networks

Neural Networks [60] are consisted of layers of interconnected neurons, which work together to process input data, recognize patterns, and make decisions or predictions. The first of the layers, which receives the raw data is the input layer, followed by multiple intermediate layers, called Hidden Layers, where neurons process inputs independently and detect patterns. These layers perform transformations on the data and are not directly observable from the outside. Finally, the final layer, called the output layer, produces the network's predictions or decisions and its structure depends on the task (e.g., classification, regression). In order for the weights of the neural networks to be updated according to the backpropagation algorithm, the gradient of the loss is needed. The loss is calculated using a loss function, which compares the output of the last layer to the ground truth.

Loss functions

The generalized equation for the loss function is:

$$J(w^T, b) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}) \tag{3.1.5}$$

where w are the weights, b the bias, m is the total number of training set data points, $\hat{y}$ is the prediction and y is the ground truth.

Mean Squared Error The most popular loss function calculating the difference between the actual value and the predicted value, squaring it and in the end taking the mean of it:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y_{\text{actual}} - y_{\text{predicted}})^2 \tag{3.1.6}$$

Its disadvantages are that it is affected by outliers and that it can't be used to interpret or compare directly with actual value.

Cross Entropy Loss The Cross Entropy Loss is a loss function especially used for classification tasks, measuring the difference between two probability distributions: the predicted probability distribution and the actual distribution of the labels (often represented as one-hot encoded vectors). For each class, it calculates the negative log of the predicted probability corresponding to the actual class and sums these across all classes:

$$L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{3.1.7}$$

The fact that it is a smooth and differentiable function which provides useful gradients for learning and is more effective in handling class probabilities makes it preferable in classification tasks compared to other loss functions, like mean squared error (MSE).

Optimization - Gradient Descent

Optimization is the process of adjusting the parameters of a model (weights) in order to minimize or maximize some objective function, typically the loss function. This can be achieved using Gradient Descent, an optimization algorithm which tries to minimize a function by iteratively moving towards the steepest descent, as defined by the negative of the gradient.

$$\theta = \theta - \eta \nabla_\theta J(\theta) \tag{3.1.8}$$

where η is the learning rate. Gradient Descent is the procedure of repeatedly evaluating the gradient and then performing a parameter update.

Backpropagation Algorithm

The backpropagation algorithm is a widely used method for training artificial neural networks by minimizing the error between the predicted output and the actual target output. It involves computing the gradient of the loss function with respect to each weight by the chain rule, efficiently propagating errors backward through the network layers. This process allows the model to adjust its weights to reduce prediction error, ultimately improving its performance, as in large neural networks the relationship of some weights and the loss function is very hard to find. The Backpropagation algorithm works by two different passes, that are repeated for some epochs, the forward pass and the backward pass.

In the forward pass, we start by propagating the data inputs to the input layer, go through the hidden layer(s), measure the network's predictions from the output layer, and finally calculate the network error based on the predictions the network made. Once the network error is calculated, then the forward propagation phase has ended, and backward pass starts.

In the backward pass, the flow is reversed so that we start by computing the gradient of the loss function with respect to the output, then applying the chain rule is iteratively to compute the gradient of the loss function with respect to each layer's parameters and inputs and in the end the gradients are propagated backward through the network to update the weights.

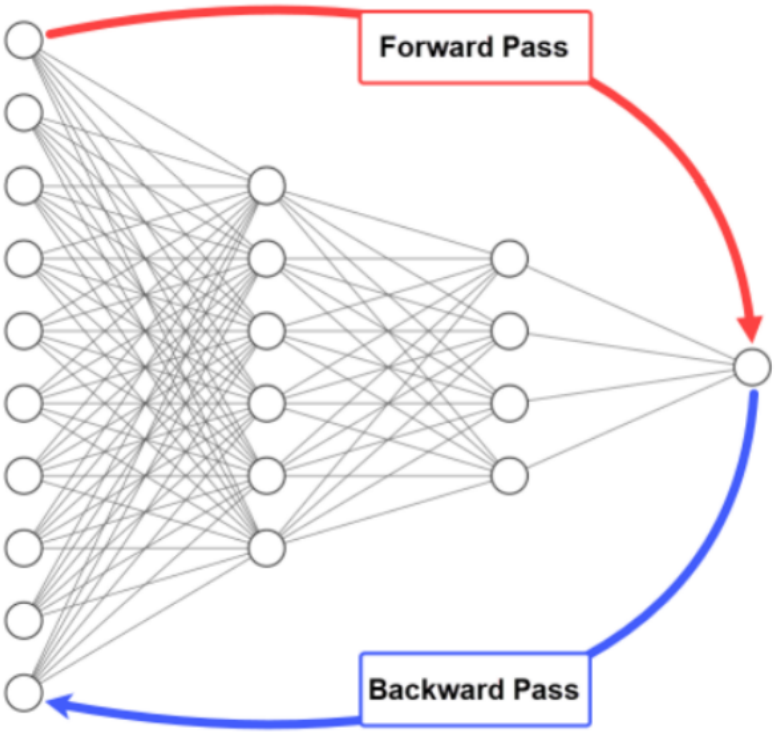


Figure 3.1.3: The forward and the backward pass.

So, firstly we compute the Gradient of the Loss with Respect to Output Layer Weights:

$$\frac{\partial L_{\mathbf{i}}}{\partial W_{\text{output}}} = \frac{\partial L_{\mathbf{i}}}{\partial O} \cdot \frac{\partial O}{\partial W_{\text{output}}} \tag{3.1.9}$$

where L the loss, W the weights and O the output.

And then we propagate the error backwards using the chain rule, so for a hidden layer h we have:

$$\frac{\partial L_{\mathbf{i}}}{\partial W_h} = \frac{\partial L_{\mathbf{i}}}{\partial O} \cdot \frac{\partial O}{\partial h} \cdot \frac{\partial h}{\partial W_h} \tag{3.1.10}$$

Regularization

The goal of a neural network is to learn a correspondence of input to output from training data and apply it on test data. Thus, it is important to be able to generalize its weights and not learn specifically the examples from the training data. When a model learns the noise and details of the training data to such an extent that it performs poorly on new, unseen data it is called overfitting. Regularization is a technique used in deep learning to prevent overfitting, by adding a penalty to the loss function during model training, discouraging

overly complex models and encouraging simpler, more generalizable models. Specifically, it adds an extra component to the loss function which prevents the weights from increasing excessively in their magnitude and thus update in a less flexible way.

L1 regularization In L1 regularization, for each weight w we add the term $\lambda_1|w|$ to the loss function. It has the intriguing property that it leads the weight vectors to become sparse during optimization (i.e. very close to exactly zero). In other words, neurons with L1 regularization end up using only a sparse subset of their most important inputs and become nearly invariant to the "noisy" inputs. In practice, if you are not concerned with explicit feature selection, L2 regularization can be expected to give superior performance over L1.

L2 regularization L2 regularization is perhaps the most common form of regularization. It can be implemented by penalizing the squared magnitude of all parameters directly in the loss function. That is, for every weight w in the network, the term $\frac{\lambda}{2}w^2$ is added to the loss function where λ is the regularization strength. The factor of $\frac{1}{2}$ is used so that the gradient of the term is simple λw. The L2 regularization has the intuitive interpretation of heavily penalizing peaky weight vectors and preferring diffuse weight vectors. This has the appealing property of encouraging the network to use all of its inputs a little rather than some of its inputs a lot. Additionally, during gradient descent parameter updates, using the L2 regularization ultimately means that every weight is decayed linearly: $w := w - \lambda x$ towards zero. It is possible to combine the L1 regularization with the L2 regularization: $\lambda_1|w| + \lambda_2 w^2$.

Dropout Dropout [111] is a popular technique for regularizing neural networks. During each training iteration, the network randomly "drops out" a fraction (dropout probability) of neurons (along with their connections) from the network. This forces the network to learn more robust features and reduces the likelihood of over-reliance on specific neurons or paths through the network. It can be thought as sampling a Neural Network within the full Neural Network, and only updating the parameters of the sampled network based on the input data.

Data Augmentation A technique where additional training examples are created by modifying the original training data (e.g., rotation, flipping, or cropping images). It helps the model generalize better by learning from a wider variety of examples, thereby reducing overfitting. Often it is performed in less populated categories of the training dataset in order to close the gap on the difference with the majority categories.

Batch size The batch size is a critical hyperparameter that refers to the number of training examples utilized in one forward/backward pass through the model. So, the model first makes its predictions to all the samples of each batch and after it is finished with that calculates the error by comparing the output (prediction) variables with the predictions. The value of the batch size can vary and choosing the appropriate one affects the model's performance, training time, and how well it generalizes to unseen data. A small batch size provides a more accurate estimate of the gradient and helps avoid local minima by introducing more noise in the computed descent process, but also results in slower training time due to less efficient computation and requires more frequent updates to model parameters. On the other hand, a large batch size makes full use of parallel computation hardware architectures, leading to faster computation, and also provides smoother and more stable gradient updates, but can lead to poorer generalization due to smoother loss landscapes and potentially getting stuck in sharp local minima.

3.1.3 Convolution

Convolution [31] is a mathematical operation applied on two functions (such as a filter or kernel and an input) to produce a third function that expresses how the shape of one is modified by the other. In simpler terms, it is a way of applying a filter (or kernel) to an input to create a feature map that highlights certain aspects or features of the input. Formally, in discrete 2-D convolutions the value of an unit at position (x, y) in the jth feature map in the ith layer, denoted as v_{ij}^{xy}, is given by the following formula:

$$v_{ij}^{xy} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ij}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \tag{3.1.11}$$

where tanh is the hyperbolic tangent function, b_{ij} is the bias for this feature map, m indexes over the set of feature maps in the $(i-1)$th layer connected to the current feature map, w_{ij}^{pq} is the value at the position (p, q) of the kernel connected to the kth feature map, and P_i and Q_i are the height and width of the kernel, respectively.

The kernel is spatially smaller than an image but is more in-depth. This means that, if the image is composed of three (RGB) channels, the kernel height and width will be spatially small, but the depth extends up to all three channels.

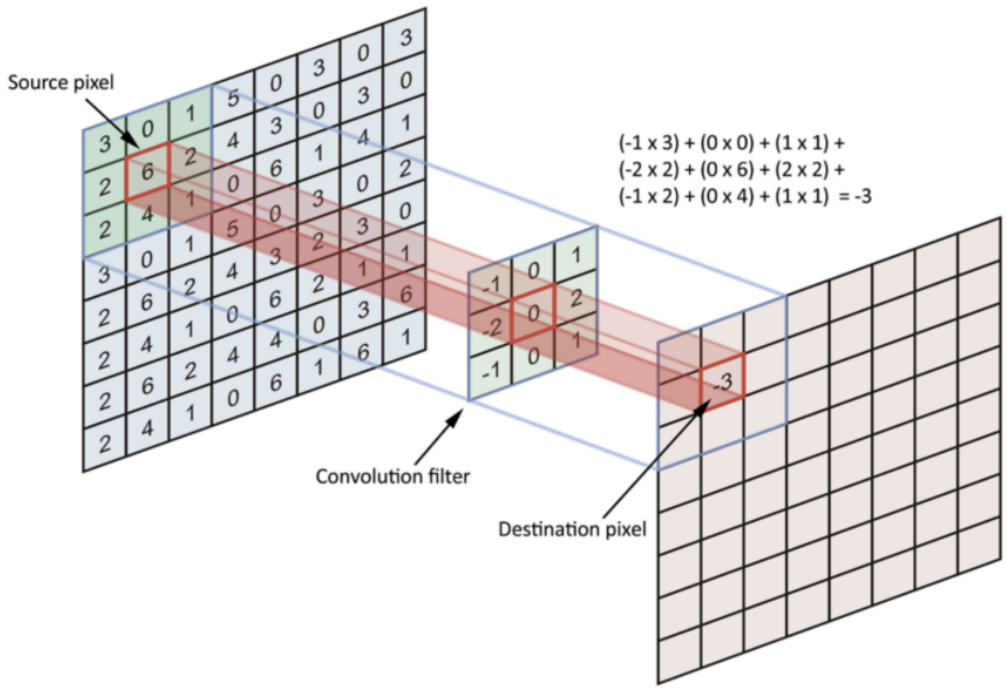


Figure 3.1.4: Convolution Visualization. The filter (kernel) is applied on an area of the input every time calculating a single point of the feature map from [10].

The convolution operation is used in CNNs to process an image and recognize on it specific characteristics. Unlike a regular Neural Network, the layers of a CNN have neurons arranged in 3 dimensions: width, height, depth. The neurons in a layer will only be connected to a small region of the layer before it, instead of all of the neurons in a fully-connected manner. A simple CNN is a sequence of layers, and every layer of a CNN transforms one volume of activations to another through a differentiable function. The types of layers used to build a CNN architecture include: Convolutional Layer, Pooling Layer, and Fully-Connected Layer.

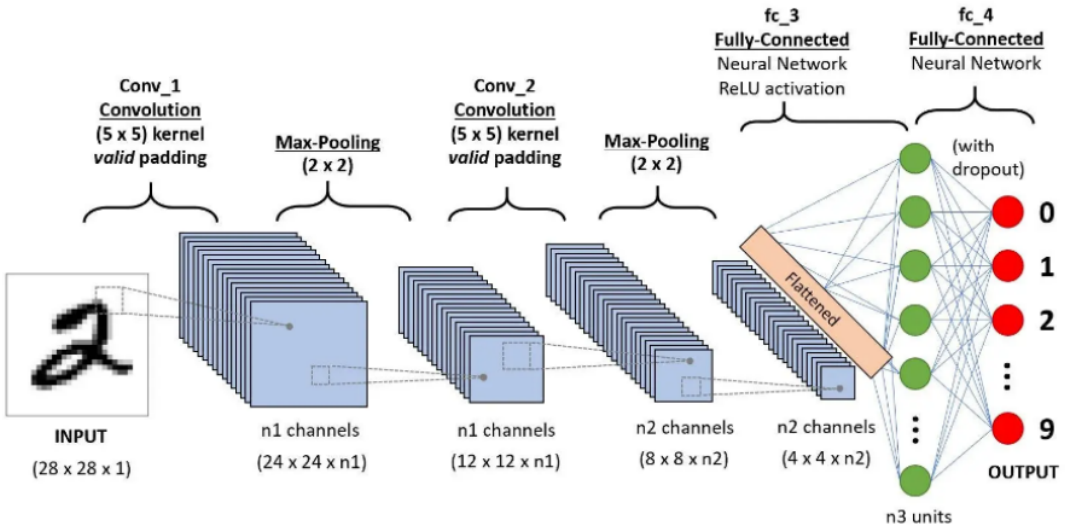


Figure 3.1.5: Convolution Network architecture

Due to the large number of pixels in an image, every neuron has a receptive field and is applied only to the corresponding layers. Therefore, it has $w \times h \times c$ number of weights, where $w \times h$ is the receptive field and c is the number of channels of the input image. That's why convolutional neural networks do not detect as sufficiently as attention mechanisms contextual information. The depth of the filter in each kernel, for example n1 and n2 in Figure 3.1.5, indicates that there are multiple neurons applied to the same patch of the input image. Therefore, the output image is also 3-dimensional with a depth equal to the depth of the filter. The purpose of multiple stacked filters is to detect multiple characteristics one output from each convolutional layer. Even though every neuron has a specific receptive field, by shifting the filter and computing the output of every patch there is no need for multiple neurons for every patch of the input image. The shifting is specified by the stride of the filter which indicates how many pixels the filter is slided across the image before it is reapplied. Additionally, another important hyperparameter is the padding, which can play an important role in controlling the output dimensions even more. It refers to adding extra pixels to the input image. Padding is usually zeros, which results in the increase of the final feature map compared to using no padding. Finally, assuming if we have an input of $w \times w \times D$ and number of kernels with a spatial size of F with stride S and amount of padding P, then the size of output volume can be determined by the following formula:

$$W = \frac{W - F + 2P}{S} + 1 \tag{3.1.12}$$

3.1.4 Pooling layer

The pooling layer is used to replace the output of the network at certain locations by deriving a summary statistic of the nearby outputs. This helps in reducing the spatial size of the representation, which decreases the required amount of computation and weights. The pooling operation is processed on every slice of the representation individually.

There are several pooling functions such as the average of the rectangular neighborhood, L2 norm of the rectangular neighborhood, and a weighted average based on the distance from the central pixel. However, the most popular process is max pooling, which reports the maximum output from the neighborhood.

3.1.5 Batch Normalization

Batch normalization is used to standardize the inputs to any particular layer. That entails the input to have zero mean and unit variance. BN layer transforms each input in the current mini-batch by subtracting the input mean in the current mini-batch and dividing it by the standard deviation. However, it is not proven

that the models perform better with zero mean and unit variance. It might perform better with some other mean and variance. Hence the BN layer also introduces two learnable parameters γ and β which adjust those parameters.

3.1.6 Fully connected layer

In the fully connected layer the neurons have full connectivity with all neurons in the preceding and succeeding layer as seen in regular FCNN. This is why it can be computed as usual by a matrix multiplication followed by a bias effect. The FC layer helps to map the representation between the input and the output.

3.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) were first introduced by Ian Goodfellow et al. in 2014 as a groundbreaking framework for generating realistic synthetic data [32]. GANs consist of two neural networks trained simultaneously through a process known as adversarial learning: a generator network (G) and a discriminator network (D).

3.2.1 Architecture and Training Procedure

The primary objective of the generator network is to produce synthetic data samples indistinguishable from real data, whereas the discriminator aims to correctly classify input data as either real or synthetic. Training is conducted through a two-player min-max game where the generator seeks to minimize the discriminator's accuracy, while the discriminator aims to maximize it. Formally, the objective is expressed as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{3.2.1}$$

Here, x represents real data samples, z is a latent vector sampled from a prior distribution (usually Gaussian), $D(x)$ is the discriminator's probability estimation of sample x being real, and $G(z)$ is the synthetic data generated from latent vector z.

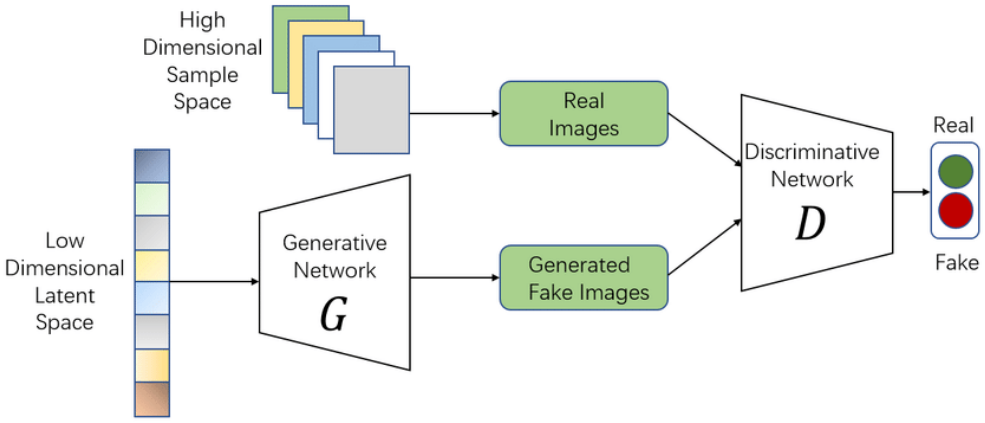


Figure 3.2.1: The architecture of vanilla GANs.

3.2.2 Variants

Since the original introduction, numerous GAN variants have been developed to address inherent training instabilities and to improve generated data quality. Notable variants include:

- **DCGAN (Deep Convolutional GAN)**: Introduces convolutional neural networks (CNNs) to GAN architectures, significantly improving visual quality and stability [93].
- **WGAN (Wasserstein GAN)**: Addresses training instability by utilizing the Earth Mover's distance (Wasserstein metric) as the loss function, which offers better convergence properties [3].
- **CycleGAN**: Facilitates image-to-image translation tasks without paired examples by enforcing cycle consistency [143].
- **PGGAN (Progressive Growing GAN)**: Improves high-resolution image synthesis by gradually increasing image size during training, leading to more stable training and finer details [51].
- **StyleGAN and Variants**: StyleGAN introduced a style-based architecture that enables fine-grained control over the synthesis process [49]. It was further refined in StyleGAN2 [52] and StyleGAN2-ADA [53], which incorporates adaptive augmentation strategies for improved performance on limited datasets. These variants are discussed in more detail in the following sections.

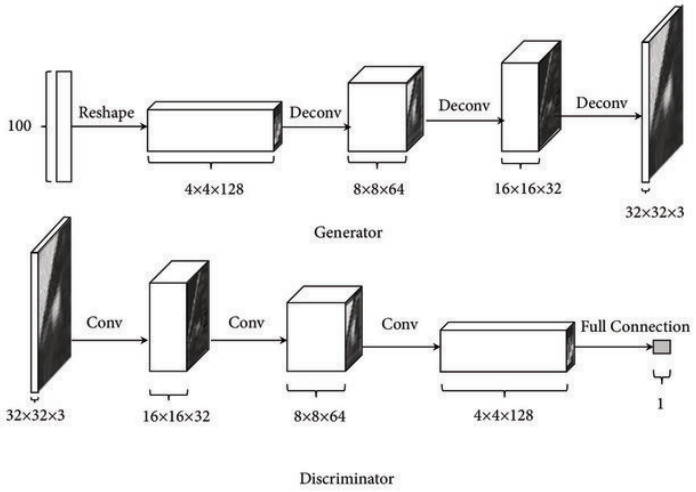


Figure 3.2.2: The architecture of DCGANs from [140].

3.2.3 Applications of GANs

GANs have shown versatility across numerous fields, including but not limited to:

- **Medical Imaging**: Synthesis of medical images for data augmentation, anomaly detection, and privacy-preserving data sharing [134].
- **Computer Vision**: Super-resolution, image inpainting, and domain adaptation [19].
- **Data Augmentation**: Generation of synthetic datasets to mitigate data scarcity and imbalance, particularly beneficial in scenarios with limited or sensitive datasets.

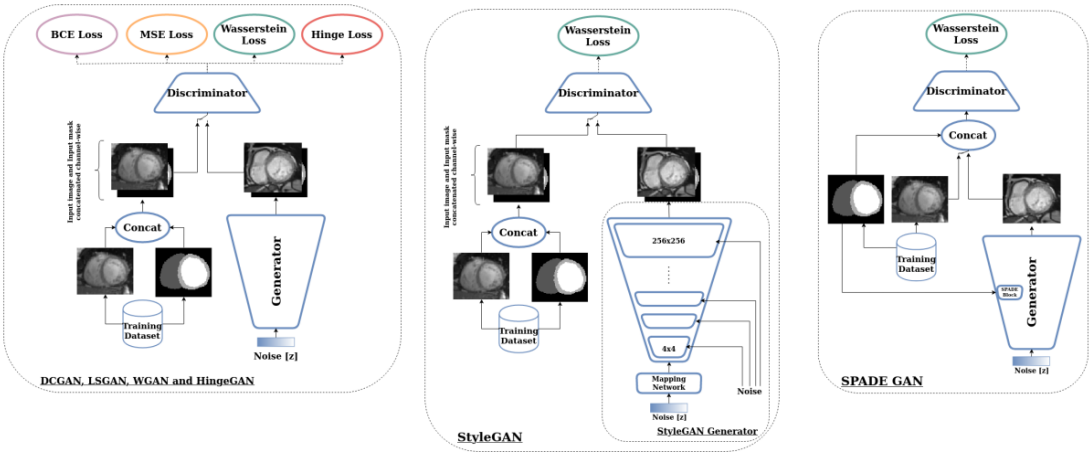


Figure 3.2.3: Various architectures for medical image synthesis from [108].

3.2.4 Challenges and Limitations

Despite their success, GANs encounter several challenges:

- **Mode Collapse**: The generator fails to capture the diversity of the dataset, producing repetitive outputs.

- **Training Instability**: Sensitivity to hyperparameter selection and difficulty achieving convergence.
- **Evaluation Difficulty**: Lack of definitive, universally accepted metrics to quantitatively assess generated data quality, often relying on human judgment or surrogate metrics like Inception Score (IS) and Fréchet Inception Distance (FID) [12].

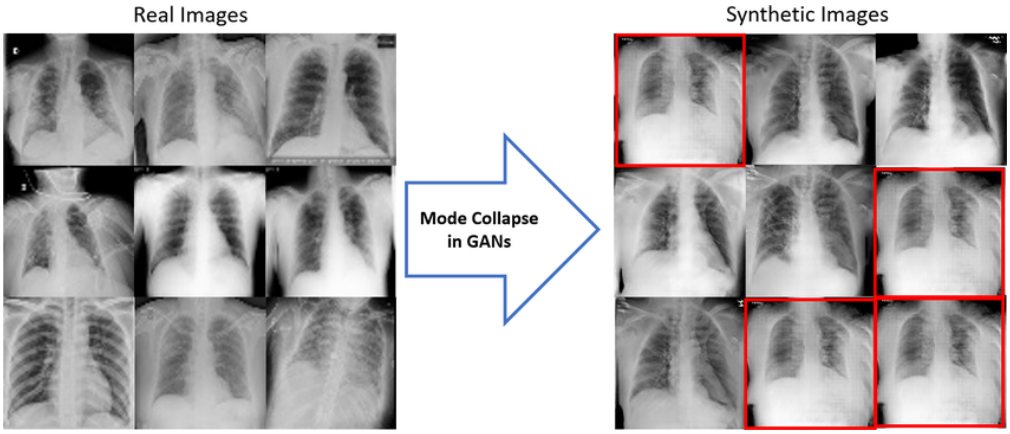


Figure 3.2.4: Mode collapse in GANs for generating X-ray images from [97].

3.3 GAN Inversion and Latent Representations

While the forward process of generating images from latent vectors using GANs has been extensively studied and applied, the inverse problem—known as **GAN inversion**—is equally crucial for many practical applications. GAN inversion refers to the task of finding a latent code z such that the output of the generator $G(z)$ closely matches a given real image x.

3.3.1 Latent Space Representations

GANs operate by mapping latent vectors sampled from a simple distribution (usually Gaussian or uniform) to the image space through a generator. This latent space, often denoted as $\mathcal{Z}$, encapsulates high-level semantic information in a compressed and often disentangled form. More advanced GAN architectures like StyleGAN introduce intermediate latent spaces (e.g., $\mathcal{W}$ and $\mathcal{W}^+$) that offer improved disentanglement and manipulation capabilities [49].

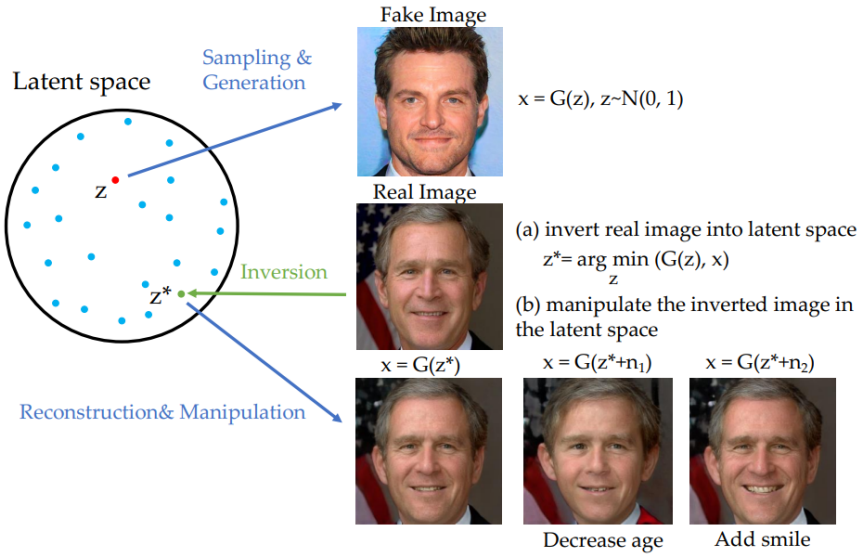


Figure 3.3.1: Illustration of GAN inversion from [132].

3.3.2 Fundamentals of GAN Inversion

The central goal of GAN inversion is to identify a latent code z^* such that the generated image $G(z^*)$ is as close as possible to a target image x. This process enables real image reconstruction, semantic editing, and analysis of the latent space. Formally, the problem is posed as an optimization:

$$z^* = \arg\min_z \mathcal{L}(G(z), x) \tag{3.3.1}$$

where $\mathcal{L}$ can be a pixel-wise loss, perceptual loss [48], or a combination of them. This problem is non-convex and can be approached in several ways:

- **Optimization-based inversion**: Solving the objective directly using gradient descent. It offers high reconstruction accuracy but can be computationally intensive.
- **Encoder-based inversion**: Training an encoder network to approximate the inverse mapping $E(x) \approx z^*$ for fast inference [142].
- **Hybrid approaches**: Combining encoder initialization with optimization refinement for better trade-offs.

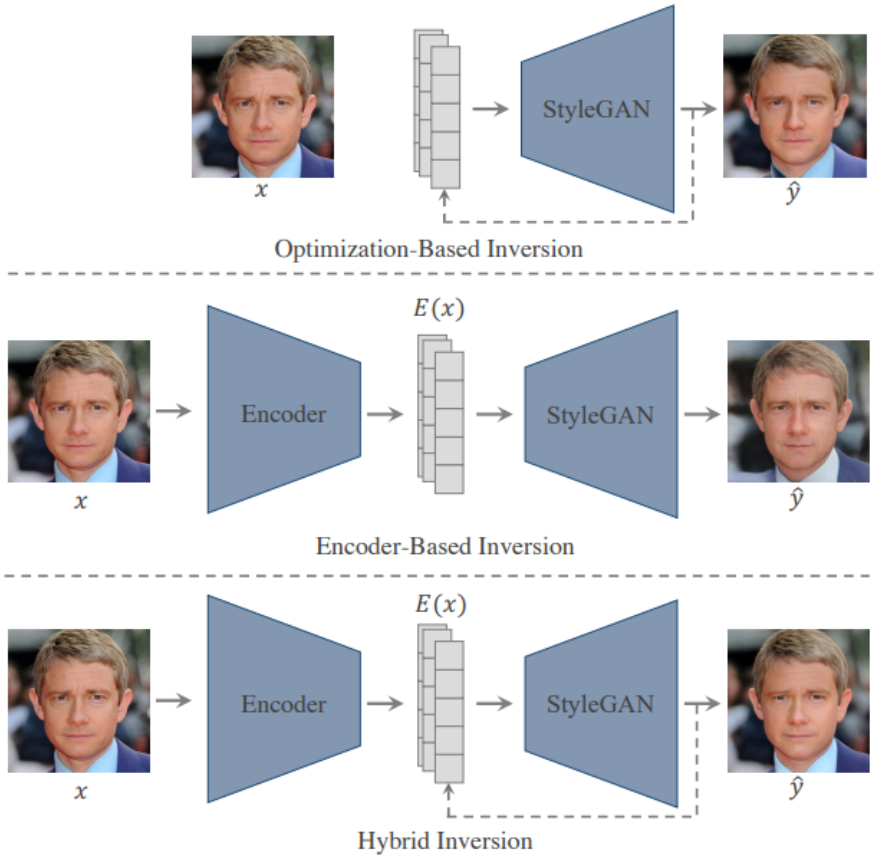


Figure 3.3.2: Various approaches of GAN inversion from [9].

3.3.3 Importance of Inversion

GAN inversion has become a cornerstone for many tasks such as image editing, attribute manipulation, and counterfactual generation. By embedding real images into the latent space, one gains access to the powerful semantic controls offered by the generator.

Despite its utility, GAN inversion presents challenges, including reconstruction fidelity, identity preservation, and handling of out-of-distribution inputs. These issues are areas of active research and refinement.

3.4 Adversarial Training and Robustness in Deep Learning

Deep learning models have achieved remarkable success across various tasks, but they are notably vulnerable to adversarial examples—intentionally crafted inputs that cause misclassifications while appearing benign to human observers [113, 33]. This susceptibility to adversarial perturbations poses risks in safety-critical applications such as autonomous driving, medical imaging, and cybersecurity.

3.4.1 Adversarial Examples and Attacks

An **adversarial example** is generated by adding a small, often imperceptible perturbation δ to a benign input x. The perturbed input $x' = x + \delta$ is designed to cause the neural network model f to produce an incorrect or targeted output:

$$\max_{\delta} \mathcal{L}(f(x+\delta), y) \quad \text{subject to } ||\delta||_p \leq \epsilon, \tag{3.4.1}$$

where $\mathcal{L}$ is typically a classification loss, y is the true label (or a targeted label for targeted attacks), $||\cdot||_p$ denotes the ℓ_p-norm, and ϵ defines the allowed perturbation budget.

Depending on the attacker's knowledge, adversarial attacks can be:

- **White-box attacks**: The attacker has full access to the model parameters and architecture (e.g., Fast Gradient Sign Method [33], Projected Gradient Descent [73]).
- **Black-box attacks**: The attacker can only query the model and observe outputs (e.g., transfer-based or query-based approaches [89]).

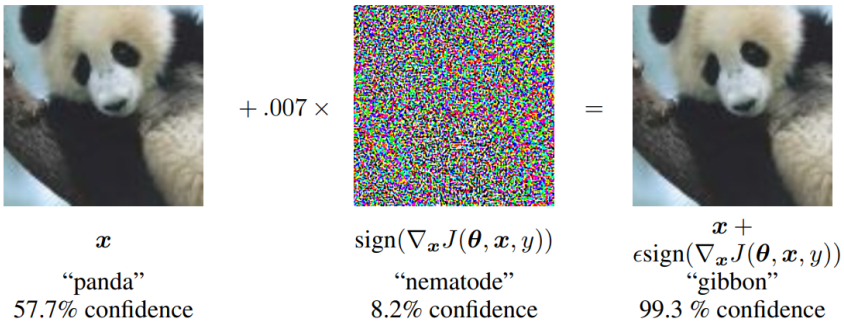


Figure 3.4.1: A demonstration of fast adversarial example generation from [33].

3.4.2 Adversarial Robustness

Adversarial robustness refers to a model's ability to maintain high performance in the presence of adversarially perturbed inputs. Robustness is typically measured by evaluating a model's accuracy against a suite of adversarial attacks with varying perturbation levels ϵ. The discovery of adversarial vulnerabilities has spurred research into both theoretical and empirical robustness criteria [29], including provable bounds for specific norms and classes of models.

3.4.3 Adversarial Training

Adversarial training is one of the most prominent defenses against adversarial examples. Introduced by Goodfellow et al. [33] and later refined by Madry et al. [73], adversarial training incorporates adversarial examples into the training set, effectively teaching the model to resist perturbations:

$$\min_{\theta} \max_{||\delta|| \leq \epsilon} \mathcal{L}(f(x+\delta;\theta), y), \tag{3.4.2}$$

where θ are the parameters of the model. By repeatedly generating adversarial examples and updating θ to minimize the loss on those adversarial inputs, models can become more robust.

- **FGSM-based training**: Uses the Fast Gradient Sign Method (FGSM) to generate adversarial examples during training [33].

- **PGD-based training**: Employs a multi-step Projected Gradient Descent (PGD) attack for stronger adversarial examples, often regarded as a gold standard for evaluating ℓ_∞-robustness [73].

- **Curriculum adversarial training**: Starts with smaller perturbations and gradually increases ϵ as training progresses.

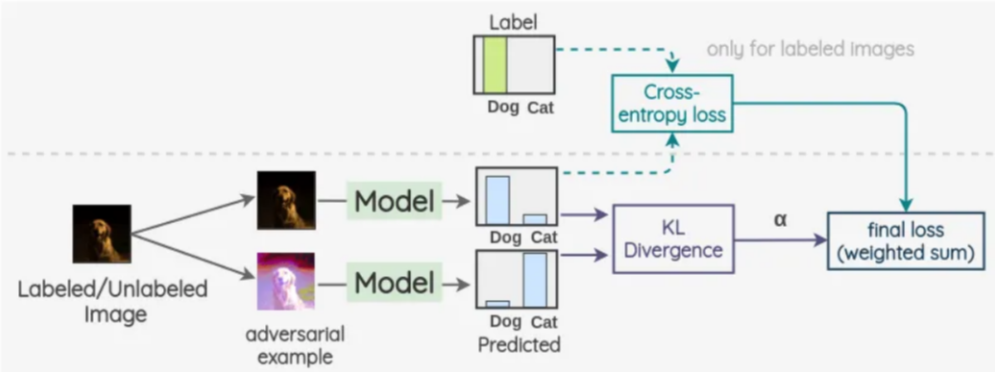


Figure 3.4.2: Adversarial training scheme

3.4.4 Other Defense Strategies

Beyond adversarial training, a variety of other defense mechanisms have been explored:

- **Defensive distillation**: Training a secondary model on softened outputs of the original model [88].

- **Input transformation**: Applying transformations (e.g., bit-depth reduction, JPEG compression) to inputs to remove adversarial perturbations [38].

- **Randomization-based methods**: Randomly altering inputs or model layers to make gradient-based attacks less effective.

- **Certified defenses**: Methods providing provable robustness guarantees for restricted perturbation classes [17, 110].

3.4.5 Challenges and Ongoing Research

Despite substantial progress, achieving robust models remains challenging. Many defenses have been circumvented by new attacks [6]. Ongoing research seeks to understand the fundamental trade-offs between accuracy, robustness, and computational cost, as well as how to extend robustness beyond ℓ_p-threat models (e.g., to realistic conditions such as rotations, translations, and other deformations).

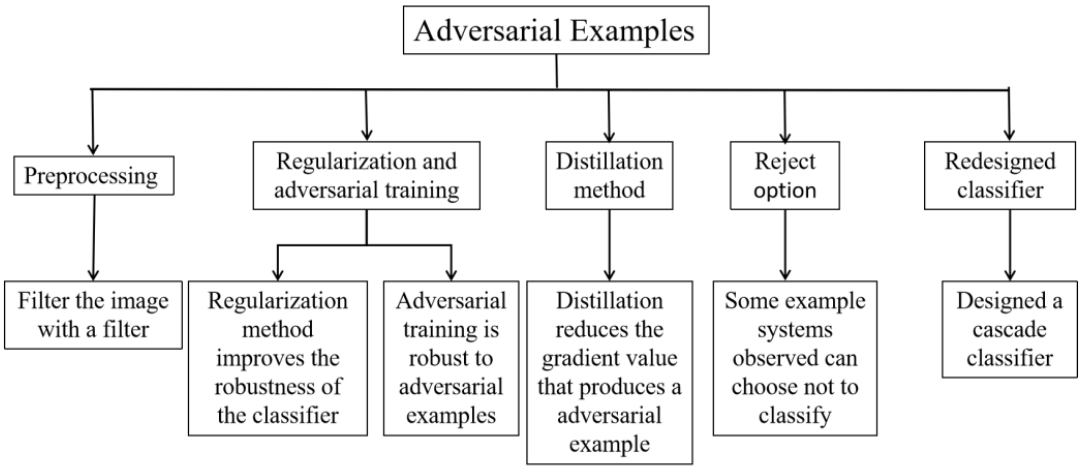


Figure 3.4.3: Defense methods on adversarial examples from [15]

3.5 Explainable AI (XAI) and Counterfactual Explanations

Explainable AI (XAI) aims to shed light on the often opaque decision-making processes of modern deep learning models. In high-stakes domains like medical imaging, interpretability is especially critical, as clinicians and stakeholders need to trust and understand model outputs [94, 98].

3.5.1 Definition of Interpretability and Explainability

In general, **interpretability** and **explainability** refer to the extent to which a human can comprehend the underlying reasons for a model's prediction [64]. A substantial body of research has emerged to make black-box models more transparent and understandable [62, 63, 78, 27, 77]. These efforts span a wide range of application domains, including medical imaging, audio analysis [70], multimodal learning [109], and natural language processing [58, 76, 71].

Broadly, there are two main strategies for achieving model explainability:

- **Post-hoc interpretations**: These involve generating visual or textual explanations after model training, without altering the model's internal architecture (e.g., saliency maps or attribution methods).
- **Intrinsically interpretable models**: These are models whose structures are inherently transparent and understandable (e.g., decision trees, rule-based systems), though they often trade off predictive power for interpretability.

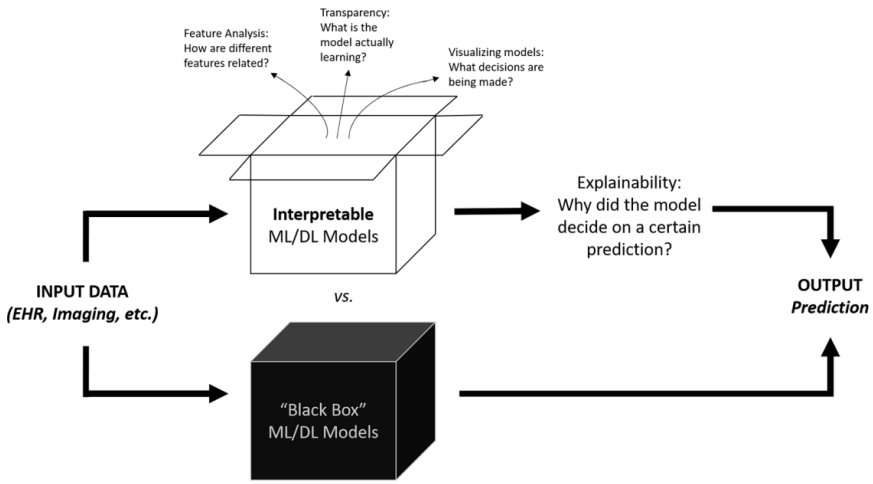


Figure 3.5.1: Black box AI models versus interpretable and explainable AI models from [44]

3.5.2 Overview of Common XAI Techniques

XAI methods in computer vision often focus on localizing influential regions in input images. Techniques like **Grad-CAM** (Gradient-weighted Class Activation Mapping)[104] and **saliency maps**[106] highlight important pixels or regions for a given prediction.

- **Grad-CAM**: Uses the gradients of target outputs flowing into the final convolutional layer to produce a coarse localization map.
- **Saliency Maps**: Compute the gradient of the output class score with respect to input pixels, visualizing where small input changes most affect the prediction.
- **Other Approaches**: LIME [94] and SHAP [69] approximate complex model behaviors around specific instances.

While these methods reveal which parts of an image contribute to a model's decision, they do not directly provide suggestions for how to *change* an input to achieve a different outcome. This is where counterfactual explanations stand out.

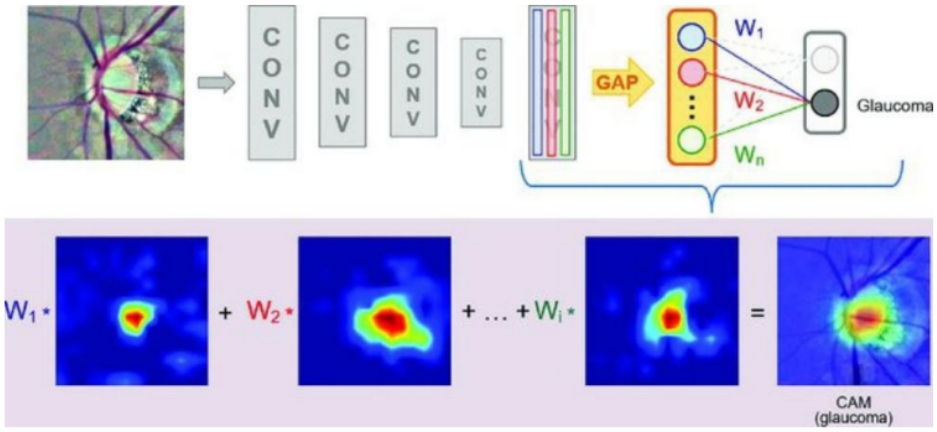


Figure 3.5.2: Use of the Grad-CAM method for a medical diagnosis problem from [102]

3.5.3 Counterfactual Explanations

Counterfactual explanations seek to answer the question: *"What minimal change to the input would alter the model's decision?"* [122]. Unlike saliency or activation maps, which highlight important regions, counterfactuals propose new inputs that lead to a desired (or undesired) outcome, often reflecting human-like reasoning [75, 28, 22, 24, 72].

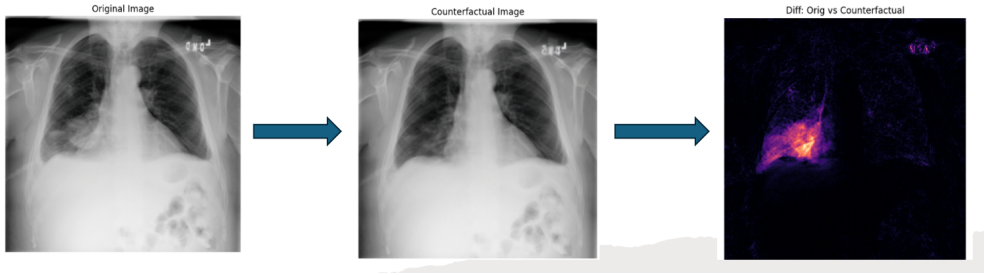


Figure 3.5.3: A healthy counterfactual chest X-ray along with the respective difference map.

Essential Criteria for Effective Counterfactuals

Counterfactual explanations in medical imaging should meet several key requirements to ensure they are both useful and trustworthy in clinical settings [122, 98]:

- **Validity**: The counterfactual image must change the model's classification outcome. If the original prediction was *positive* (e.g., malignant), the counterfactual should induce a *negative* (benign) outcome, and vice versa.
- **Realism**: Generated images must resemble authentic medical scans. In practice, this means the modification should preserve anatomical consistency and remain indistinguishable from real data, subject to expert scrutiny.
- **Actionability**: Counterfactual edits should align with clinically feasible scenarios. For example, shrinking a lesion by a slight margin may be actionable as a hypothetical surgery or treatment effect, but removing a critical anatomical structure entirely is not.

- **Sparsity**: The change from the original image should be minimal, making it clear which factors influenced the decision. This also aligns with interpretability: small, localized changes are more easily understood than broad, sweeping modifications.
- **Causality**: Ideally, counterfactual modifications reflect meaningful disease pathology rather than superficial, spurious changes. If removing a suspicious nodule decreases the classifier's likelihood of malignancy, it aligns with clinical logic.

Why Counterfactuals Are Intuitive

Counterfactuals resonate well with human reasoning, as we routinely answer "what if" questions. In clinical settings, such hypothetical scenarios are valuable for justifying or questioning a diagnosis: *"If this lesion were slightly smaller or differently shaped, the model might judge it as non-cancerous."*

Adoption Challenges in Medical Imaging

Despite their promise, counterfactual explanations for medical imaging face several hurdles:

- **Data Privacy and Ethics**: Generating or manipulating patient data must comply with strict privacy regulations.
- **Clinical Validation**: Any modifications to medical images require expert evaluation to ensure realistic anatomical changes.
- **Computational Complexity**: Generating high-resolution, anatomically plausible counterfactuals can be computationally expensive.

Chapter 4

Related Work

Visual counterfactual explanation (CE) methods have gained significant attention within the XAI community due to their intuitive interpretability and potential applicability across diverse domains. This chapter presents a comprehensive review of existing visual CE approaches, categorizing them into methods tested primarily on non-medical datasets and those explicitly designed for medical image analysis. The objective of this structured review is twofold: firstly, to highlight the breadth and evolution of techniques developed outside healthcare, emphasizing foundational concepts and innovations in algorithmic design; secondly, to explore methods tailored specifically for medical datasets, underscoring their unique considerations such as clinical validity, anatomical realism, and ethical implications. Through this review, we aim to identify common trends, methodological strengths, limitations, and open challenges, thus clearly positioning our proposed framework within the broader landscape of visual explainability research.

Contents

4.1 Visual Counterfactual Methods in General Image Domains

Initially, we will focus on counterfactual generation algorithms tested on non medical image datasets, describing their method and mentioning some key points referenced in the respective papers.Our goal in this section is to get a broader view of the visual CE approaches in the XAI community.The methods below are presented in chronological order of publication:

- **CEM**.Dhurandhar et al. [23] introduce one of the first algorithms for visual CE generation. This explanation process involves finding a minimal set of features in the input that are sufficient for the same classification (pertinent positives) and a minimal set of features that should be absent to prevent classification from changing (pertinent negatives).The paper argues that understanding what is absent (pertinent negatives) is just as critical for accurate explanations. The authors validate CEM on three diverse real-world datasets: handwritten digits **MNIST**, procurement fraud(which is a tabular dataset), and brain functional MRI imaging, demonstrating the method's effectiveness in generating precise and understandable explanations.The paper also discusses the broader implications and potential applications of CEM, suggesting it could be useful for tasks beyond just generating explanations, such as model comparison, debugging, and improvement.
- **CVE** [34].The Counterfactual Visual Explanation approach aims at finding counterfactual explanations for image classifiers by solving with greedy search the minimum-edit counterfactual problem. The idea of CVE is simple but effective. Given a classifier b and a randomly selected image x" with b(x")≠b(x), CVE searches (with two possible strategies) for the minimum changes to x replacing with pixels selected from x',i.e., x' = T (x, x"), that leads to x' such that b(x') = b(x") through a transformation T . This new approach is applied in four different datasets: **MNIST** , **SHAPES** which is a dataset of synthetic images which are depicting objects characterized by shape,color and size , **Omniglot** which contains 1623 different handwritten characters from 50 different alphabets and **CUB-200-2011**[123].

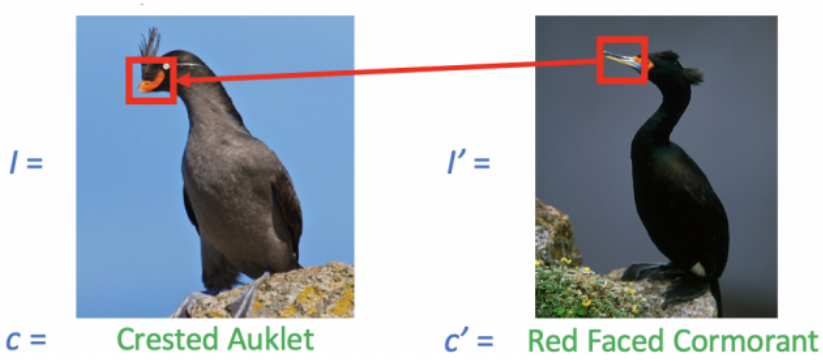


Figure 4.1.1: CVE approach generates counterfactual visual explanations for a query image I (left) – explaining why the example image was classified as class c (*Crested Auklet*) rather than class c' (*Red Faced Cormorant*) by finding a region in a distractor image I' (right) and a region in the query I (highlighted in red boxes) such that if the highlighted region in the left image looked like the highlighted region in the right image, the resulting image I^* would be classified more confidently as c'.

- **FACE**.Feasible and Actionable Counterfactual Explanations (face) [91] focuses on returning "actionable" counterfactuals by uncovering "feasible paths" for generating counterfactual. These feasible paths are the shortest path distances defined via density-weighted metrics. In this way, face extracts plausible counterfactuals that are coherent with the input data distribution. More it details faceworks as follows. First, it generates a graph over the data points by using KDE,k-NN or an ε -graph. The user can also select the prediction threshold of b, the density threshold, the weights of the features, and custom condition functions to specify actionability. Then, it updates the graph according to these constraints. Finally, face applies a shortest path algorithm to find all the data points that satisfy the requirements.face is an endogenous and data-agnostic counterfactual explanation method that can the-

oretically be used also to work on datasets with categorical features. As far as images are concerned, FACE is applied in **MNIST** dataset.

- **PCATTGAN**. In [4] the authors present a plausible counterfactual explainer relying on adversarial examples to retrieve counterfactuals. In particular, the pcattgan system comprehend an AttGAN model [40] and a multiobjective optimization model that infers the attribute modifications needed to produce plausible counterfactuals for the black-box b. The loss function accounts for validity, minimality and plausibility that is intended here as the implementation of credible changes not performed by a computer. This approach is tested on the **CelebA** dataset.
- **SCOUT**. The Self-aware disCriminant cOUnterfactual explanaTion method in [125], aims at returning discriminant counterfactual explanations for image classifiers. An explanation is produced by the computation of two discriminant explanations with the role of the input image x and an image with the desired class x', inverted. A discriminant explanation for images x and x' consists of a saliency map highlighting pixels highly informative for b(x)(b is the classifier of the problem) but uninformative for b(x'). Discriminant explanations are obtained through an optimization process performed with an explanation architecture combining features activation layers of the CNN explained.Experiments are performed on two datasets: The **CUB200-2011** datset [123] and the **ADE20K** dataset [141] with more than 1000 fine-grained scene categories.

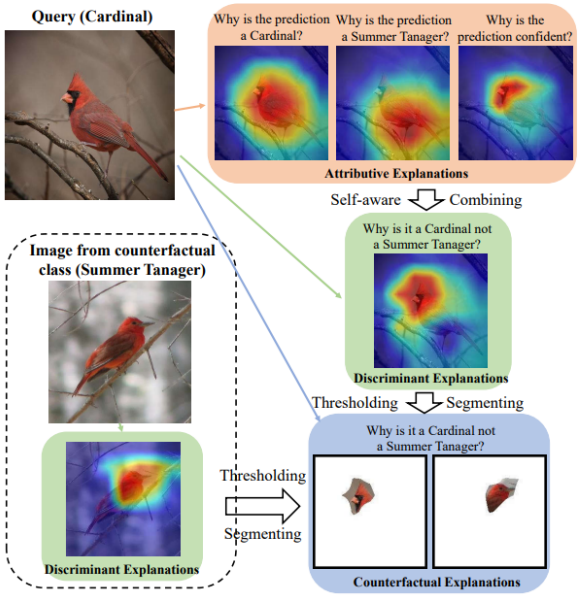


Figure 4.1.2: Counterfactual explanations using SCOUT method.

- **FRACE**.Fast ReAl-time Counterfactual Explanation which is introduced in [139] is basically an explainer for neural networks classifiers for images. The architecture of frace is a neural network itself, and it is aimed at minimizing a loss function accounting for validity and a minimal perturbation. FRACE searches for the perturbation through a starGAN used as residual generator to generate the perturbation that causes the change of class. FRACE also accounts for plausibility because of the adversarial training. Experiments show that it is markedly faster than SCOUT. Experiments were carried out on three different datasets : **MNIST** which is a large collection of handwritten digits, **EMNIST**[16] which is an extension of MNIST and another dataset which consists of various Chinese Characters.
- **PIECE**. Kenny and Keane [54] illustrate the PlausIble Exceptionality based Contrastive Explanation (piece) method for generating contrastive explanations for CNN working on image data. PIECE identifies feature-values with low probability in the latent features of the CNN representing the instance under analysis x, i.e., exceptional features, and attempts to modify them to be their expected values in the desired counterfactual class, i.e., normal features. We underline that the "features" treated by

piece are not directly parts of the input image, but their latent features activating the neurons of the CNN. Finally, piece exploits a GAN to generate the counterfactual images.PIECE is mainly tested on **MNIST** and **CIFAR10** dataset.We will explain this method in detail in the Methodology section, as it inspired us to create our SPRUCE framework.

- **Diverse Valuable Explanations (DiVE)**. In [66] a novel approach is presented aimed at improving the quality of counterfactual explanations in computer vision applications.DiVE focuses on generating diverse and valuable explanations that are valid, proximal (close to the original input), sparse (minimal changes), and non-trivial. This means it seeks to uncover less obvious but more informative factors affecting the model's predictions, such as revealing biases or spurious correlations. DiVE employs a β-Total Correlation Variational Autoencoder (β-TCVAE) to achieve a disentangled latent representation of the data. This helps in generating explanations that are more proximal and sparse.Additionally, it learns a latent perturbation constrained by three main losses: a counterfactual loss to fool the machine learning model, a proximity loss to maintain similarity with the original input, and a diversity loss to ensure diverse explanations.To avoid trivial explanations (like exaggerating an already existing attribute), DiVE masks out the most influential latent factors, thus encouraging modifications in other, potentially more insightful attributes.T The method was tested on **CelebA** and **Synbols**, demonstrating its ability to produce high-quality, valuable explanations superior to previous methods.

- **StylEX**.Lang et al. [59] introduce StylEx, a method that discovers and visualizes the semantic characteristics of an image , by training a generative model(StyleGan) to specifically explain multiple attributes that underlie classifier decisions.It is showed how an image can be modified in different ways to change its classifier output. The results show that the method finds attributes that align well with semantic ones, generate meaningful image-specific explanations, and are human-interpretable as measured in user-studies. Multiple datasets are used in this specific paper: **CelebA** , **CUB-200-2001** , **FFHQ** which contains multiple human faces with different characteristics , age, accessories,ethnicities and backgrounds , **AFHQ** which contains multiple animal faces and **Diabetic Retinopathy Dataset** which contains a large amount of retina images.

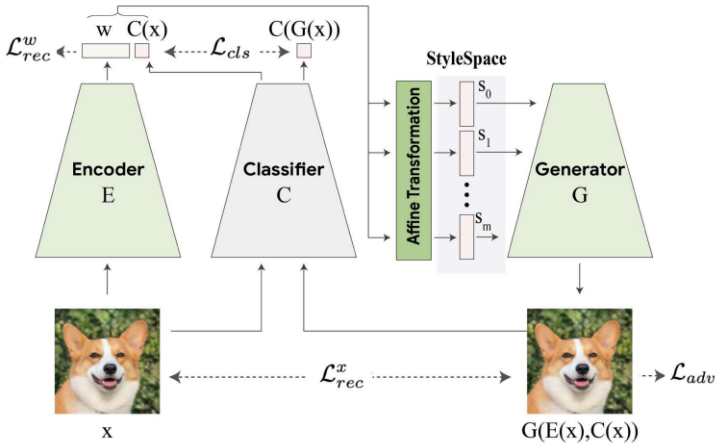


Figure 4.1.3: StylEx architecture

- **SEDC-T**.Vermeire and Martens [121] presents SEDCT, an extension of SEDC working on images and also usable for multiclass classifiers.The algorithm identifies a small set of image segments whose removal changes the classification result, thereby creating counterfactuals .It is mainly applied on various images of ImageNet dataset , showing impressive results. The algorithm is tested on images that were initially misclassified , finding the part of the image that should change in order for that to be put in the correct class.

- **Diff-SCM**.In [99] a new deep structural causal model (Diff-SCM) that builds on the latest advances of generative energy-based models is proposed. Counterfactual estimation is achieved by firstly inferring

latent variables with deterministic forward diffusion, then intervening on a reverse diffusion process using the gradients of an anti-causal predictor w.r.t the input. Furthermore, a metric is proposed for evaluating the generated counterfactuals.It is shown that Diff-SCM produces more realistic and minimal counterfactuals than baselines on **MNIST** data and can also be applied to **ImageNet** data.

- **SCVC**.A new framework for Semantically Consistent Visual Counterfactuals is introduced in [118]. The new approach,being an improvement of the SCOUT method discussed above, is based on two key ideas. First, the replaced and replacer regions are enforced to contain the same semantic part, resulting in more semantically consistent explanations. Second, multiple distractor images are used in a computationally efficient way and thus more discriminative explanations with fewer region replacements are obtained.Their approach is 27% more semantically consistent and an order of magnitude faster than SCOUT method on three fine-grained image recognition datasets: **CUB-200-2001** , **Stanford-Dogs** [57] which includes 22,000 annotated images of 120 species of dogs and **iNaturalist-2021** which contains over 2.7M images from 10k different species.
- **Diffusion Visual Counterfactual Explanations(DVCE)**. In [7], Boreiko et al. unveil a groundbreaking approach that capitalizes on the significant benefits of diffusion models, thereby addressing some of the limitations inherent in existing methods for producing Visual Counterfactual Explanations (VCEs).The authors introduce Diffusion Visual Counterfactual Explanations (DVCEs), applicable to any ImageNet classifier, utilizing a diffusion process. This method incorporates two critical modifications: firstly, a flexible parameterization strategy, enhanced by distance regulation and a delayed commencement of the diffusion process, which ensures the generation of images with minimal yet impactful semantic alterations. Secondly, they implement cone regularization through a model resistant to adversarial attacks to guide the diffusion towards significant, semantically relevant changes rather than inconsequential ones, thereby ensuring the produced images are not only realistic representations of the intended class but also elicit high confidence from the classifier.

4.2 Visual Counterfactual Methods for Medical Imaging

In this section we will refer to the methods that mainly focus on medical datasets.

- **TRaCE**. A great approach is made in [116]. The TraCE method introduces an innovative approach to generating counterfactual explanations specifically for clinical image predictors. Central to TraCE is its focus on integrating prediction uncertainties to ensure the reliability and meaningfulness of the explanations. This method is underpinned by three main components: an auto-encoding Convolutional Neural Network that creates a continuous latent space for image data, a predictive model equipped with uncertainty estimation using the Learn-by-Calibrating technique [115], and a counterfactual optimization strategy guided by an uncertainty-based calibration objective.Key to its approach is the construction of counterfactuals in a latent space, minimizing semantic discrepancies while altering the classification outcome.The empirical studies highlight TraCE's superiority in producing physically plausible and clinically valuable explanations compared to existing methods. The dataset that is being leveraged here is the the publicly available RSNA pneumonia detection challenge database which contains several chest X-ray images.

- **CLEAR**. A novel VCE method called CLEAR is introduced in [129] . CLEAR Image explains an image's classification probability by contrasting the image with a corresponding image generated automatically via adversarial learning. This enables both salient segmentation and perturbations that faithfully determine each segment's importance. CLEAR Image uses regression to determine a causal equation describing a classifier's local input-output behaviour. Counterfactuals ,supported by the causal equation, are also identified. CLEAR is applied on **CheXpert** dataset which contains 224,316 chest radiographs of 65,240 patients with both frontal and lateral views available.

- **Visual Explanations for diabetic retinopathy**.Another great approach is being made in [11]. Boreko et al. propose an ensemble method of plain and simultaneously adversarially robust models, exploiting the advantages of both. Through their technique they manage to maintain high accuracy producing meaningful visual counterfactuals. The paper is mainly focused on diabetic retinopathy and thus uses retinal fundus images. Three datasets are used: : the Kaggle DR detection challenge data with retinal images , the **Messidor** dataset and the **IDRiD** (Indian Diabetic Retinopathy Image Dataset) [90].

- **Gifsplanation**. In [18] an autoencoder approach to counterfactual generation for Chest X-rays is introduced.Given an arbitrary classifier, they propose a simple autoencoder and gradient update (Latent Shift) that can transform the latent representation of a specific input image to exaggerate or curtail the features used for prediction. They use this method to study chest X-ray classifiers and evaluate their performance. They conduct a reader study with two radiologists assessing 240 chest X-ray predictions to identify which ones are false positives (half are) using traditional attribution maps and their proposed method.

- **GANterfactual**. [80]The "GANterfactual" paper presents a model-agnostic approach to generating counterfactual explanations using adversarial image-to-image translation techniques, particularly for non-experts in medical fields.The method utilizes Generative Adversarial Networks (GANs) to transform the input image into a counterfactual version that the classifier would predict differently. This is achieved by including the classifier's decision into the objective function of the generative networks, allowing for automated and realistic counterfactual creation.The paper validates the effectiveness of GANterfactual through a computational evaluation and a user study, demonstrating its ability to generate more meaningful and satisfactory counterfactual explanations compared to state-of-the-art methods like LIME and LRP.The dataset used in the paper for evaluating the GANterfactual approach is the one published for the RSNA Pneumonia Detection Challenge.

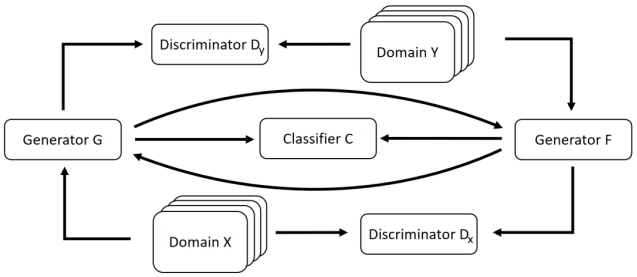


Figure 4.2.1: Schematic overview of the GANterfactual method.

- **Counterfactuals for Retinal Fundus and OCT Images using Diffusion Models** Ilanchezian et al. [45] explore the application of diffusion models for creating realistic counterfactual images in ophthalmic imaging, focusing on two modalities: color fundus photography (CFP) and Optical Coherence Tomography (OCT). The datasets employed are from EyePacs Inc., comprising over 180,000 retinal fundus images from a Diabetic Retinopathy (DR) screening program, and a dataset of 108,309 OCT images categorized into normal, choroidal neovascularization (CNV), drusen, and Diabetic Macular Edema (DME). The method leverages diffusion models, known for their realistic image generation, combined with classifiers trained to detect various eye diseases. This approach produces highly realistic counterfactuals that not only change the classifier's decision but do so by depicting or removing disease signs in a believable manner. The study's significance is underscored by a user study where domain experts confirmed the realism and clinical utility of the generated counterfactuals.
- **DiffExplainer.**In [26], Fang et al. introduce DiffExplainer, a novel counterfactual generation method designed to explain black-box classifier decisions, particularly in medical imaging. This approach combines diffusion-based generative modeling with teacher-student learning to produce high-quality and realistic counterfactual explanations. DiffExplainer first employs a Diffusion Autoencoder to learn a latent representation of the input images, ensuring stable training and effective reconstruction. A shallow student model, trained to mimic the predictions of the black-box (teacher) classifier through knowledge distillation, then manipulates these latent features to generate counterfactual images. These generated images highlight influential features by minimally altering the original inputs, thus changing the black-box model's predictions. DiffExplainer is particularly significant as it addresses common challenges associated with GAN-based methods, such as unstable training and poor reconstruction capabilities, and demonstrates effectiveness on complex medical tasks like CT imaging for lung disease prognosis.

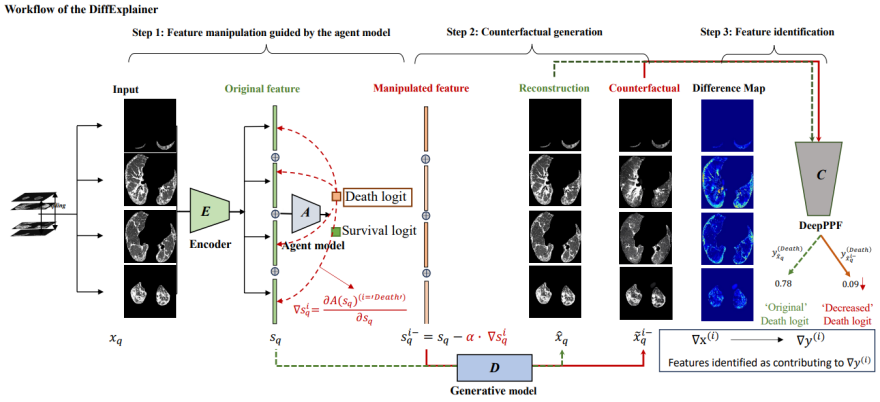


Figure 4.2.2: DiffExplainer framework.

- **Diffusion Autoencoder-based Counterfactuals (DAE).**Atad et al. [5] propose using a Diffusion Autoencoder (DAE) to generate counterfactual explanations directly within the latent space, without

relying on external models. Their approach takes advantage of the unsupervised learning capabilities of DAEs, which learn semantically meaningful latent representations from medical images. These latent spaces inherently support interpolation and manipulation, facilitating the generation of both binary and ordinal counterfactuals. The method involves training a linear classifier within this latent space to establish a decision boundary, which is subsequently used to produce realistic counterfactual examples by adjusting latent representations across or along decision boundaries. The authors demonstrate the versatility and interpretability of their approach on several medical imaging tasks, including classification and ordinal regression of conditions such as vertebral compression fractures and diabetic retinopathy. Importantly, this method simplifies the generation of counterfactual explanations, enhances interpretability, and supports continuous visualization of pathology progression without the need for additional supervised feature extraction.

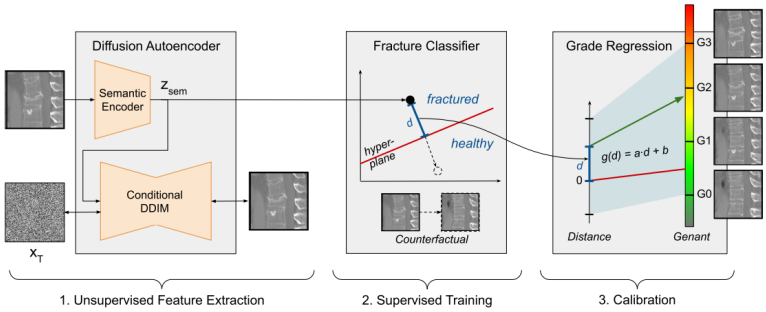


Figure 4.2.3: DAE framework.

4.3 Key Insights

As mentioned in the introduction, only a few of these works enforce some form of sparsity in the counterfactual edit, but these rely on supervised training of the explainer and residual maps, which makes these methods unable to generate counterfactuals for arbitrary classifiers.Some works rely on Autoencoders instead of GANs , but GANs are known to produce sharper images with better details, which we consider imperative in the medical domain.Recent works also employ Denoising Diffusion Implicit Models (DDIMs).Even though the image quality of DDIMs rivals that of GANs, latent space manipulation is harder, while image reconstruction is an issue due to their inherent probabilistic nature and it is still being researched.Some of the methods employing DDIMs report better metrics than GANs, but they do not employ any techniques for improving GAN inversion, and fail to account for sparsity.

Chapter 5

Methodology

The core of the workflow for the generation of the medical image counterfactuals is based on the work of Kenny and Keane who introduced a novel approach called PIECE (PlausIble Exceptionality-based Contrastive Explanations) that not only produced counterfactual explanations but also focused on the creation of semifactual images that offer a better understanding of the classifier's decision boundaries. We will first describe the steps of the original PIECE framework, and then make a thorough analysis of the components of our proposed image counterfactual generation scheme called SPRUCE (Sparse Realistic Uncoupled Counterfactual Explanations).

Contents

5.1 Introducing SPRUCE: Sparse Realistic Uncoupled Counterfactual Explanations

5.1.1 Overview of PIECE framework

As for most counterfactual explanation methods in the image space, PIECE was originally tested on the MNIST [21] and CIFAR-10 datasets,with the results indicating the capability of the method to produce plausible counterfactuals with meaningful and humanly interpretable changes along with the respective semi-factuals. The main idea of PIECE is to identify probabilistically low feature values in the test image (i.e. exceptional features) and modify them to be their expected values in the counterfactual class (i.e. normal features).

PIECE involves two distinct systems, an image classifier with predictions that need to be explained and a generative adversararial network (GAN) that helps visualize counterfactual or semifactual explanatory images in the pixel space. For the algorithm to work , the GAN needs to be trained on the same dataset (or larger but with the same distribution) as the one that the CNN has been trained on. The models can be trained separately giving us the freedom to exploit the most advanced classifiers and generative adversarial networks that will tackle the obstacle of medical image complexity.

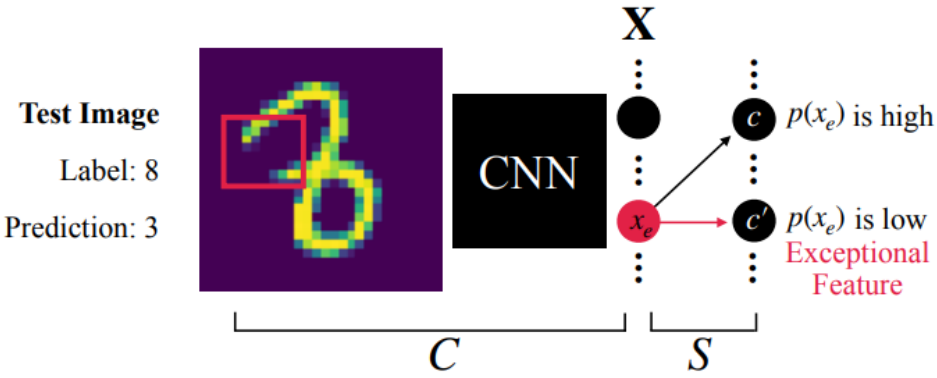


Figure 5.1.1: Identifying Exceptional Features in a query image

Setup and Notation. Let a **convolutional neural network (CNN)** consist of:

- A set of layers, up to the penultimate extracted feature layer **X** ,will be denoted as C as shown in Figure 5.1.1.

-A final output softmax classifier denoted as S.

For a given test image I, we extract the deep feature representation at layer X:

$$x = C(I) \tag{5.1.1}$$

where x is the feature vector at the last CNN layer before classification.

The classifier output is given by:

$$Y = S(x) = S(C(I)) \tag{5.1.2}$$

where Y is the probability distribution over all classes, and the predicted class is:

$$Y_c = \arg\max Y \tag{5.1.3}$$

Let G represent the generator of a trained GAN, which maps a latent vector z to an image:

$$I = G(z) \tag{5.1.4}$$

where z is the latent representation of I.The counterfactuals to a test image I, in class c, with latent features x,are denoted as I', c' and x' respectively.

Here are the four main steps of the original algorithm :

1. **Locating the query image in the GAN's latent space and identifying the counterfactual class.** Initially, it is necessary to find the latent vector z , such that $G(z) \approx I$. This procedure is called GAN inversion and in the original paper of PIECE a simple optimization approach is used :

$$z = \arg\min_{z_0} \|C(G(z_0)) - C(I)\|_2^2 + \|G(z_0) - I\|_2^2 \tag{5.1.5}$$

   where z_0 is a sample from the normal standard distribution.The method also requires a counterfactual class to be selected. If the classifier's prediciton is incorrect then the counterfactual class c' is trivially selected to be the true label of the query image. However, when the CNN's classification is correct for I, identifying c' becomes non-trivial.A method involving *gradient ascent* is used to solve this problem :

$$\arg\max_{z} \|S(C(G(z))) - Y_c\|_2^2 \tag{5.1.6}$$

   where Y_c is binary encoded as all 0s, and a 1 for the class c. During this optimization process, the first time a decision boundary is crossed, the new class is selected as c'.

2. **Identifying Exceptional Features.** The PIECE algorithm identifies *exceptional features* in a test image I by analyzing the probability of feature activations in the **counterfactual class** c'. This step ensures that counterfactual modifications focus only on **statistically uncommon features**, making the generated explanations **more plausible**.

   Given the fact that the CNNs used in the original method were ResNet50 backbones with ReLU activation functions, PIECE models feature activations in the CNN's last feature layer using a **hurdle model**, which consists of:

   (a) A **Bernoulli activation probability** θ_i that represents whether a neuron activates in class c'.

   (b) A **probability density function (PDF)** $f_i(x_i)$ modeling the distribution of activation values when the neuron is active.

   Each neuron's activation probability for class c' is expressed as:

$$p(x_i) = (1 - \theta_i)\delta(x_i)(0) + \theta_i f_i(x_i), \quad x_i \geq 0 \tag{5.1.7}$$

   where:

   - $\theta_i = \frac{q}{m}$ is the probability of activation, estimated from training data (q is the number of times neuron i activates, and m is the total number of samples in c').
   - $f_i(x_i)$ is the learned distribution of activations when $x_i > 0$.
   - $\delta(x_i)(0)$ is the **Kronecker delta function**, ensuring that inactive neurons are properly modeled.

   The distributions are estimated using maximum likelihood estimation (MLE) and validated via the Kolmogorov-Smirnov test to choose the best fit (Gaussian, Gamma, or Exponential).

   A feature x_i in the test image is exceptional if its probability under the counterfactual distribution is below a predefined threshold α (which usually takes values from 0.01 to 0.05). The following conditions determine whether a feature is exceptional:

   (a) **Neuron remains inactive when it typically activates in c':**

$$p(1 - \theta_i) < \alpha, \quad x_i = 0 \tag{5.1.8}$$

   (b) **Neuron activates when it typically remains inactive in c':**

$$p(\theta_i) < \alpha, \quad x_i > 0 \tag{5.1.9}$$

(c) **Neuron has an unusually low activation value for c':**

$$\theta_i F_i(x_i) < \alpha, \quad x_i > 0 \tag{5.1.10}$$

(d) **Neuron has an unusually high activation value for c':**

$$(1 - \theta_i) + \theta_i F_i(x_i) > 1 - \alpha, \quad x_i > 0 \tag{5.1.11}$$

where $F_i(x_i)$ is the cumulative distribution function (CDF) of the PDF $f_i(x_i)$.

By defining exceptional features in this manner, PIECE ensures that counterfactual modifications only target features that significantly deviate from expected values in c', leading to more realistic and interpretable counterfactual images.

3. **Changing the Exceptional to the Expected.** The algorithm adjusts exceptional features to their expected values in c' , but only if this change brings the CNN closer to reclassifying the image as c'. Features are prioritized based on probability, modifying those with the lowest probability first. This ordering is crucial for semi-factual explanations, where modifications stop before crossing the decision boundary.

4. **Visualizing the Explanation.** Finally, having constructed x', the explanation is visualized by solving the following optimization problem with gradient descent:

$$z' = \arg\min_z \|C(G(z)) - x'\|_2^2 \tag{5.1.12}$$

and inputting z' into G to visualize the explanation I'.

5.1.2 Overview of SPRUCE framework

We employ and extend the method of exceptional feature extraction from PIECE and thus create a novel framework named SPRUCE (Sparse Realistic Uncoupled Counterfactual Explanations) that will help us generate medical image counterfactuals with an emphasis on sparsity and realism. In Figure 5.1.2 we present the pipeline of our proposed methodology.

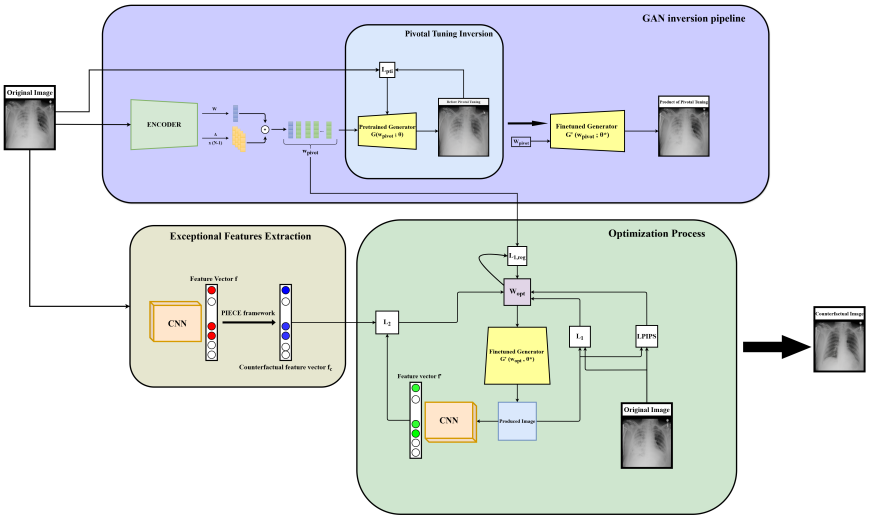


Figure 5.1.2: The pipeline of SPRUCE

Observing the pipeline of SPRUCE one can clearly understand that it is composed of three main steps :

- **Step 1: Extraction of Exceptional Features.**For a correctly classified image, we compute the feature vector of the counterfactual image in the classifier's latent space using the methodology introduced from PIECE.
- **Step 2: GAN inversion.**We project the original query image into our GAN's latent space , obtaining a latent representation that can reconstruct the image accurately while still allowing meaningful, targeted edits.We use a hybrid GAN inversion pipeline that is made up of two modules:
  1. An **encoder-based** inversion for an initial editable latent representation.
  2. **Pivotal tuning** to significantly enhance reconstruction realism without severely reducing editability.
- **Step 3: Latent Space Optimization.**We optimize the latent code of the query image using a specialized loss function that enforces both realism and sparsity in the counterfactual.Using the optimized latent code, we generate a realistic counterfactual that minimally alters the input image while changing the classifier's decision.

In Sections 5.3,5.5,5.6 we will explain in detail each step of SPRUCE and in Sections 5.2,5.4 we are going to analyze the architectural design of the classifier and GAN model respectively that we employed in our framework.Finally, in Section 5.7 we will refer to the importance of adversarial robustness for the generation of meaningful and realistic counterfactuals.

5.2 ConvNeXt classifier

The first part of the PIECE framework that we discussed in the previous section has to do with the selection of a classifier network whose decisions we are going to explain through our counterfactuals. Our intention was to use the most recent and state-of-the-art model that has been evaluated and tested on image classification tasks.For SPRUCE, we chose the ConvNeXt model as our primary classification model over other traditional backbones such as ResNet50 [39], VGG [107], and DenseNet121 [43] mainly due to its superior performance in modern vision tasks and its architectural improvements inspired by Vision Transformers (ViTs)[25] and more specifically form hierarchical vision transformers like Swin [65].

5.2.1 Architectural Design of ConvNeXt

Introduction

ConvNeXt is a modernized convolutional neural network (CNN) that integrates design principles from Vision Transformers (ViTs) while retaining the efficiency and simplicity of traditional ConvNets. Unlike pure Transformers, ConvNeXt does not rely on self-attention mechanisms but instead enhances CNNs through a series of structural modifications that improve scalability, expressiveness, and computational efficiency.

We will first present the key architectural innovations of ConvNeXt, highlighting how it refines CNNs to match the performance of hierarchical Transformers while maintaining interpretability and ease of implementation.

Modernizing ConvNets: Key Architectural Changes

The ConvNeXt architecture is built upon a ResNet foundation but incorporates refinements inspired by ViTs and Swin Transformers. These enhancements can be grouped into two main categories:

- **Macro-level changes:** Improvements to overall network structure.
- **Micro-level changes:** Modifications at the block level.

Macro-Level Design Changes

(a) Hierarchical Multi-Stage Design

- ConvNeXt follows a hierarchical architecture, similar to standard CNNs, but adjusts the stage compute ratio.
- The number of blocks per stage is modified to **(3, 3, 9, 3)** to align with hierarchical Transformers like Swin.

(b) Patchified Convolutional Stem

- Traditional ResNets begin with a 7×7 convolution followed by max pooling.
- ConvNeXt replaces this with a **"patchify" layer**—a 4×4 convolution with stride 4, similar to ViTs' patch tokenization.

(c) Larger Convolutional Kernels

- Instead of using 3×3 kernels, ConvNeXt **increases the kernel size to 7×7** for depthwise convolutions.
- This mimics the global receptive field of self-attention in Transformers.

(d) ResNeXt in ConvNeXt

- **Depthwise Convolutions:** Replaces ResNeXt's grouped convolutions with depthwise convolutions, where each channel operates independently.
- **Separation of Spatial and Channel Processing:** Uses depthwise and 1×1 pointwise convolutions, mimicking self-attention in Transformers while reducing FLOPs.
- **Increased Network Width:** Expands model width to match Swin Transformers, improving accuracy.

(e) Inverted Bottleneck Structure

- Inspired by MobileNetV2, ConvNeXt expands channels before applying convolutions.
- Uses a 4× expansion ratio, similar to ViTs' MLP blocks.

Figure 5.2.1: Inverted bottleneck structure

Micro-Level Design Changes

(a) GELU Activation Instead of ReLU

- Traditional CNNs use ReLU, but ConvNeXt adopts GELU, a smoother activation function used in ViTs.
- GELU allows better gradient flow and improves feature representations.

The Gaussian Error Linear Unit (GELU) activation function is defined as:

$$\text{GELU}(x) = x \cdot \Phi(x) = x \cdot \frac{1}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \tag{5.2.1}$$

where $\Phi(x)$ is the cumulative distribution function of the Gaussian distribution, and $\text{erf}(x)$ is the error function.One can also approximate the function with the following formula :

$$\text{GELU}(x) \approx \frac{x}{2}\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right) \tag{5.2.2}$$

As we will see later, the use of GeLU activation will play an important role in the way that we are going to model the feature activation values , needed for the PIECE algorithm implementation.

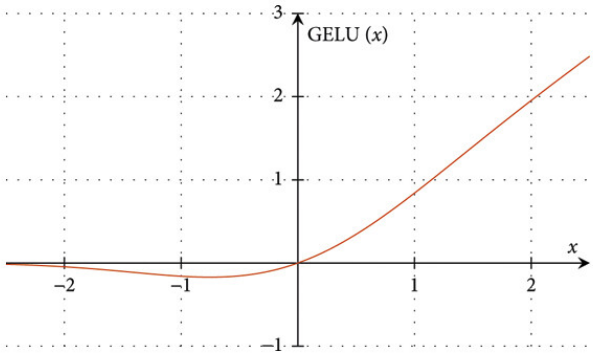


Figure 5.2.2: GeLU activation function

(b) Reduction of Activation Functions and Normalization Layers

- Traditional CNNs apply activations after every convolutional layer. ConvNeXt removes redundant activation functions and uses a single GeLU activation in each block.
- ConvNeXt only uses one Batch Normalization layer (instead of two) before the conv 1×1 layers.

(c) Layer Normalization instead of Batch Normalization

- Following the Transformer's structure, ConvNeXt replaces the traditional Batch Normalization component with Layer Normalization in each residual block.

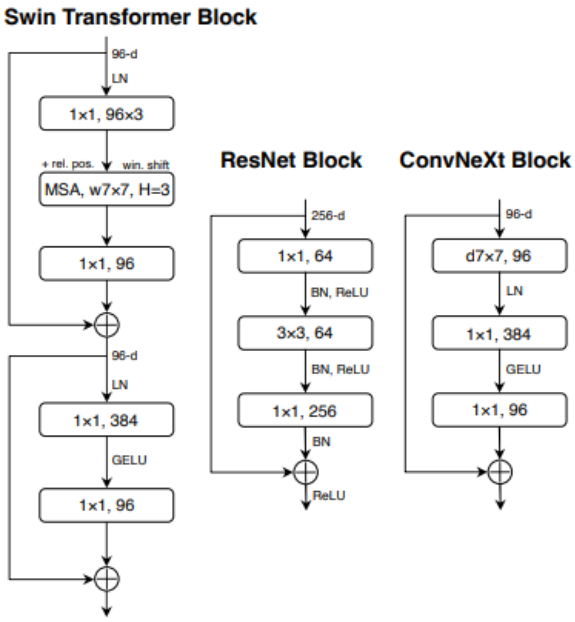


Figure 5.2.3: Block designs for a ResNet, a Swin Transformer and a ConvNeXt

5.2.2 Comparison with Traditional CNNs

In Figure 5.2.3 we can see the final convnext block in comparison with the ResNet and Swin Transformer block. The table below summarizes key differences between traditional CNNs and ConvNeXt:

Feature	**Traditional CNNs**	**ConvNeXt**
Convolution Type	Standard Convs	Depthwise Separable Convs
Normalization	BatchNorm	LayerNorm
Activation	ReLU	GELU
Downsampling	Max Pooling / Strided Conv	Patchified Convolution
Residual Blocks	Bottleneck Residuals	Inverted Bottlenecks

Table 5.1: Key Architectural Differences between Traditional CNNs and ConvNeXt.

5.2.3 ConvNeXt variant selection

For our experiments, we selected the **ConvNeXt-Base** model, a mid-sized variant of the ConvNeXt architecture that provides a strong balance between accuracy and computational efficiency. This model retains the core architectural improvements of ConvNeXt while maintaining a practical parameter size for training on medical datasets.

Model Specifications

ConvNeXt-Base follows a hierarchical design with four feature extraction stages, each increasing in complexity. The number of channels at each stage follows the configuration:

$$(128, 256, 512, 1024) \tag{5.2.3}$$

where each stage consists of multiple ConvNeXt blocks. The network depth is distributed as (3, 3, 27, 3), ensuring efficient representation learning across scales. ConvNeXt Base-model has 89M parameters.

Performance on ImageNet

The ConvNeXt-Base model was evaluated on ImageNet-1K and ImageNet-22K datasets, demonstrating competitive performance with state-of-the-art models.The top-1 accuracy and computational cost achieved at each dataset are the following.

Model	Dataset	Image Size	Top-1 Accuracy (%)	FLOPs (G)
Swin-Base	ImageNet-1K	224^2	83.5	15.4
ConvNeXt-Base	ImageNet-1K	224^2	83.8	15.4
Swin-Base	ImageNet-22K	224^2	85.2	15.4
ConvNeXt-Base	ImageNet-22K	224^2	85.8	15.4

Table 5.2: Performance comparison of ConvNeXt-Base and Swin Transformer on ImageNet-1K and ImageNet-22K.

These results highlight the effectiveness of ConvNeXt-Base in large-scale image classification and its capability to outperform hierarchical transformers , as well as traditional ConvNets, in benchmark classification tasks.

Performance on robust benchmarks

Another key factor in selecting this variant is its inherent robustness to image corruption. As we will discuss later, ensuring robustness to small perturbations is crucial for generating sparse yet meaningful counterfactuals. The original ConvNeXt paper emphasizes that the larger model variants generally outperform both traditional ConvNets and Transformers on various robustness classification benchmarks. Therefore, to achieve a balance between model size and robustness, we selected the Base variant as our primary classification model.

Model	ImageNet-A (%)	ImageNet-R (%)	ImageNet-Sketch (%)
ResNet-50	0.0	36.1	24.1
Swin Transformer	35.8	46.6	32.4
ConvNeXt-Base	36.7	51.3	38.2

Table 5.3: Robustness performance of ConvNeXt-Base on ImageNet-A, ImageNet-R, and ImageNet-Sketch.

5.3 Extraction of Exceptional Features for ConvNeXt

Modifying slightly the notation used in [54], let F be a pre-trained, frozen classifier that accepts as input an image I and outputs a probability vector $Y = F(I)$, with the predicted class being $c = \textbf{argmax}(Y)$. We assume that F is a Neural Network whose final layer is a fully connected linear layer followed by a SoftMax activation function. Let C be the part of the network up to the penultimate layer, let S be the final linear layer, and let x be the input to the final layer i.e., $x = C(I)$, $Y = S(x)$. We will also call x the latent features of image I. A counterfactual image I' is an image such that $c' = \textbf{argmax}\, Y' \neq c$, with $Y' = F(I')$.

In order to guide the counterfactual generation process to the desired counterfactual class, we identify and modify the "exceptional" features of the latent features x of image I, a concept originally introduced by [54]. This process consists of three key steps:

- For each class of the classifier, fit a probabilistic model on the latent features of images it classifies to that class.
- For any given input image with latent features x and any desired counterfactual class c', identify the "exceptional" features of x with respect to c'. These are the elements of x that significantly deviate from the values that an image classified as c' would have.
- Among these exceptional features, selectively modify only those that influence the classifier's decision, ensuring the model is guided toward the counterfactual class.

5.3.1 Fitting the Latent Features' activations

Let D be a dataset consisting of images that belong to the same distribution that the classifier was trained on. Let L_c be the set of the latent features of all images in D classified as class c i.e., $L_c = \{x |\ x = C(I),\ S(x) = c,\ I \in D\}$. We assume that x follows a different distribution X_c for each class c and we fit each X_c using the statistics of L_c. Since x is a vector, X_c is a multivariate distribution, but we assume that the distributions X_{ci} of each element x_i of x are independent. For each X_{ci} we fit a Gaussian Mixture Model (GMM) with K components:

$$p(x_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2),$$

where:

- π_k are the mixture weights such that $\sum_k \pi_k = 1$.
- $\mathcal{N}(x_i \mid \mu_k, \sigma_k^2)$ is a Gaussian component with mean μ_k and variance σ_k^2.
- K is the number of Gaussian components.

In the original paper of PIECE, the authors assumed that the activation function of the penultimate layer was a Rectified Linear Unit (ReLU), which only takes positive values, and could be fit using a hurdle model. Since ConvNeXt models use a smooth variant of ReLU that takes negative values as well, a Gaussian Error Linear Unit (GELU) [41], we needed to employ a different and more flexible distribution, with the GMM giving us the most consistent results. It is important to note that fitting the probabilistic models only needs to happen once for each classifier and not for each input image.In the experiments we found empirically that 5 gaussian components are enough to effectively model the activation values of the convnext's features.

5.3.2 Identifying Exceptional Features.

For a given input image I and a desired counterfactual class c', we identify the exceptional features of its latent features x by evaluating the statistical likelihood of each element x_i of x with respect to the fitted model for $X_{c'i}$. To determine whether an observed latent feature x_i is exceptional, we compute its two-tailed probability based on the CDF of the fitted GMM in order to detect both low and high exceptional values:

$$F(x_i) = \sum_{k=1}^{K} \pi_k \Phi\left(\frac{x_i - \mu_k}{\sigma_k}\right),$$

where $\Phi(\cdot)$ is the standard Gaussian CDF.

The two-tailed probability is then computed as:

$$P(x_i) = \min(F(x_i), 1 - F(x_i)).$$

- If x_i is much smaller than the expected values then $F(x_i)$ will be close to 0.
- If x_i is much larger than the expected values then 1-$F(x_i)$ will be close to 0.

A latent feature is considered exceptional if:

$$P(x_i) < t,$$

where t is the significance threshold.

5.3.3 Changing the Exceptional to the Expected.

Once we identify all features whose two-tailed probability $P(x_i)$ is below the threshold t, we further filter them by considering the sign of weights w_i of the final layer S. Specifically, if an outlier is "high" but also has a positive weight (or is "low" with a negative weight), it may be beneficial for class c' and is therefore left unchanged. Conversely, if an outlier contradicts its weight's sign (e.g., a high outlier with a negative weight), it is deemed detrimental and replaced with its expected value under the fitted GMM:

$$x'_i \leftarrow \mathbb{E}[X_{c'i}].$$

This two-stage process ensures that the modified latent feature vector x' shifts toward the typical distribution of class c', while preserving meaningful outliers that support classification into c'. Upon completion of this process, we obtain the counterfactual feature vector in the feature space of the trained classifier. This feature vector will help us guide the explainer—in our case, the generator—toward producing the final explanatory image in pixel space via an optimization scheme that will be applied on the the projection of the original image into the GAN's latent space.

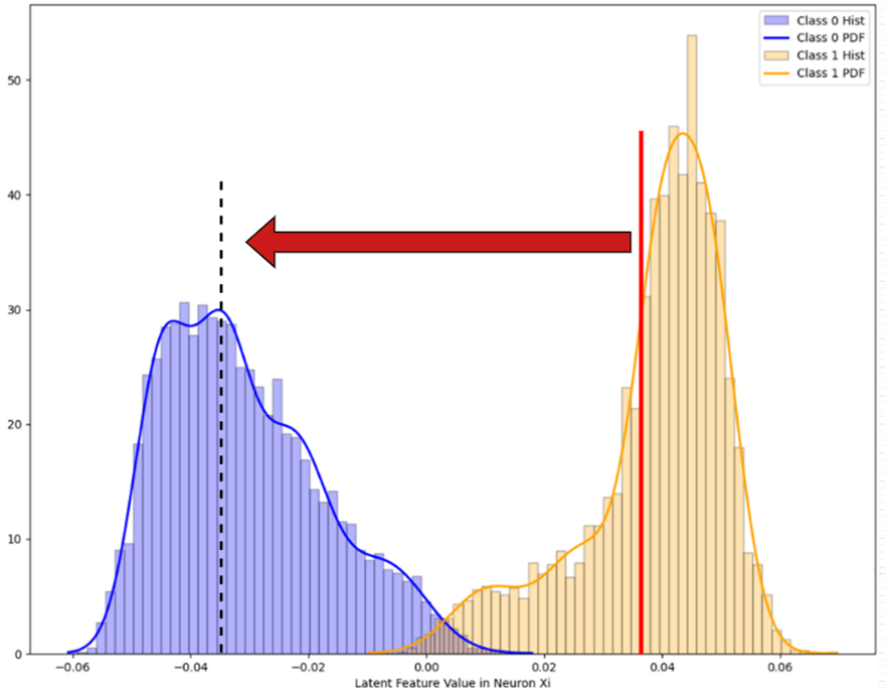


Figure 5.3.1: Changing the exceptional to the expected.With the blue and orange lines we have modelled the activation value of a neuron x_i for both classes of the dataset.The activation of the neuron for the query image I is indicated with red and the expected value of the counterfactual class 0 is indicated with the dotted line.Neuron x_i is considered exceptional and thus its value is changed to the expected.

5.4 StyleGAN2-ADA model

After selecting the classifier model whose decisions we aim to explain, it is essential to choose a Generative Adversarial Network (GAN) capable of generating realistic counterfactual explanations in image space. The selected GAN must effectively capture the complexity and variability of the medical imaging domain while ensuring high-fidelity image synthesis. Additionally, it must address a significant challenge inherent to medical datasets: limited sample availability. Considering these requirements, we selected StyleGAN2-ADA [53] as our primary GAN model, as it combines high-quality image generation with adaptive discriminator augmentation (ADA), enabling stable training even on small datasets.

5.4.1 Introduction to StyleGAN

Generative Adversarial Networks (GANs) traditionally consist of two components: a *generator* that synthesizes candidate images from random noise, and a *discriminator* that classifies images as real or fake. *StyleGAN* [49] introduced a **style-based** generator architecture that enables more explicit control over the generated images at various scales.

5.4.2 Style-Based Generator Concept

Mapping Network. Instead of feeding the latent vector $\mathbf{z}$ (sampled from some prior, e.g., $\mathcal{N}(0, \mathbf{I})$) directly to the generator, StyleGAN first transforms $\mathbf{z}$ into an intermediate latent vector $\mathbf{w}$ via a learned *mapping network*. This intermediate space helps disentangle factors of variation and makes the generator's outputs more controllable.

Adaptive Instance Normalization (AdaIN). Once the mapping network outputs $\mathbf{w}$, an *affine transformation* produces scale and bias parameters (γ, β) for each layer in the generator. These parameters modulate activations through

$$\text{AdaIN}(\mathbf{x}) = \gamma \left(\frac{\mathbf{x} - \mu(\mathbf{x})}{\sigma(\mathbf{x})} \right) + \beta, \tag{5.4.1}$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the per-channel mean and standard deviation of the feature map $\mathbf{x}$. This mechanism allows each layer to independently control different visual features, resulting in semantically meaningful style manipulation.

Progressive Growing & Style Mixing. To stabilize high-resolution image synthesis, the model can be trained in a *progressive* manner, starting from lower resolutions and incrementally adding layers in the same manner that a Progressive GAN [50] is trained . Additionally, *style mixing* uses different $\mathbf{w}$ vectors for different layers, encouraging the network to avoid encoding all information in a single stage, leading to more diverse image synthesis.

5.4.3 Advantages of the Style-Based Approach

- **Disentangled Control:** Mapping $\mathbf{z}$ to $\mathbf{w}$ simplifies the latent representation, making it easier to interpret and edit visual features.
- **High-Quality Images:** StyleGAN demonstrated high fidelity and detail at resolutions previously not achievable by earlier GANs.
- **Rich Semantic Edits:** The $\mathbf{w}$ space often correlates with distinct semantic factors (e.g., pose, hair style, lighting), supporting intuitive manipulations.

5.4.4 Refinements in StyleGAN2

StyleGAN introduced a novel *style-based generator* architecture that enabled unprecedented control over image synthesis. Despite its success, researchers noted several limitations:

- **Repeating "Droplet" Artifacts:** Generated images sometimes exhibited faint circular or blob-like artifacts, especially noticeable in backgrounds or regions of smooth color.

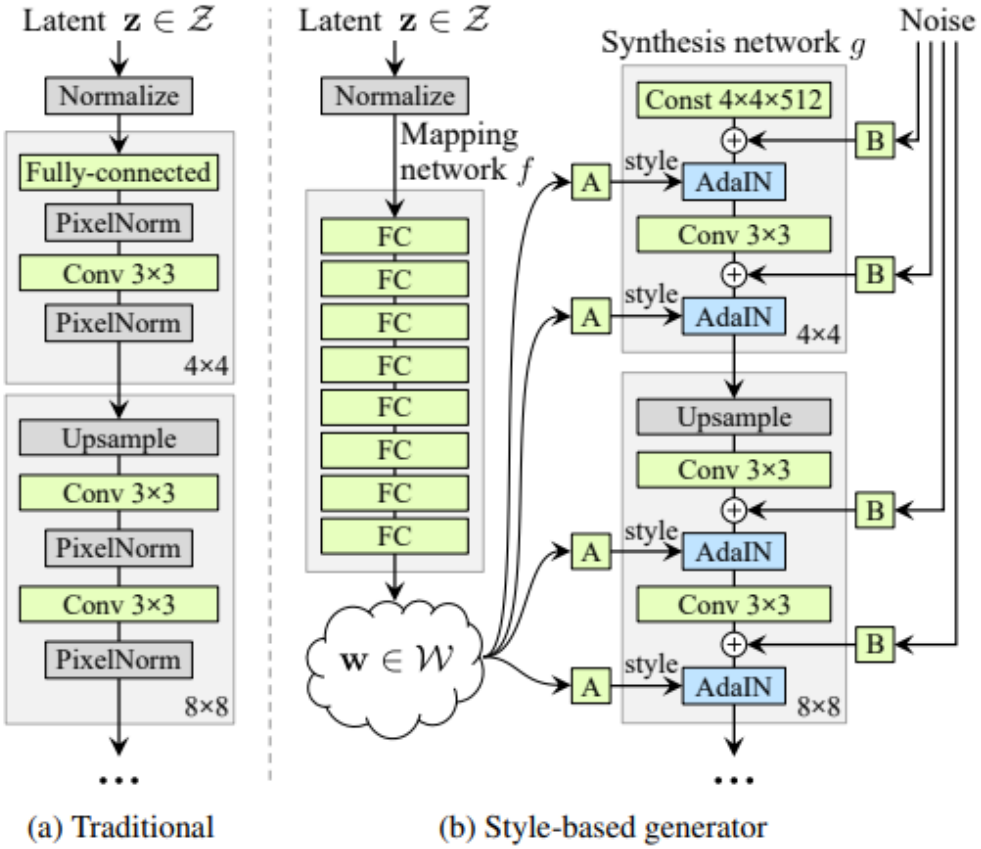


Figure 5.4.1: High-level schematic of the StyleGAN generator. The mapping network transforms the latent code $\mathbf{z}$ into intermediate latents $\mathbf{w}$, which then modulate convolutional layers via AdaIN.

- **Progressive Growing Complexity:** The original training strategy used *progressive growing* (starting with low resolution and incrementally adding layers), which adds training complexity and can introduce transient artifacts at scale transitions.
- **Residual Entanglement:** Although AdaIN improved style control, small entanglements in the latent representations could still cause correlated changes in unrelated features.

StyleGAN2 [52] addressed these artifacts with several key innovations.

- **Weight Modulation and Demodulation** In StyleGAN, each layer applied *Adaptive Instance Normalization* (AdaIN) using an affine transformation of the intermediate latent vector $\mathbf{w}$ to produce scale and bias parameters (γ, β). While effective, this design could introduce normalization artifacts and repetitive textural patterns (often called "droplets").

  Replacement of AdaIN. StyleGAN2 replaces explicit AdaIN layers with **weight modulation and demodulation**:

$$\mathbf{W}'_{ijk} = s_i \cdot \mathbf{W}_{ijk}, \quad \mathbf{W}''_{ijk} = \frac{\mathbf{W}'_{ijk}}{\sqrt{\sum_{i,k} (\mathbf{W}'_{ijk})^2 + \epsilon}}, \tag{5.4.2}$$

  where:

  - $\mathbf{W}_{ijk}$ are the original convolution kernel weights for input channel i, output channel j, and kernel element k.
  - s_i is a learned *style coefficient* derived from the latent $\mathbf{w}$.

– ϵ is a small constant to avoid division by zero.

This scheme effectively scales the weights based on the style and then *demodulates* them, preventing any channel from dominating or introducing periodic artifacts.

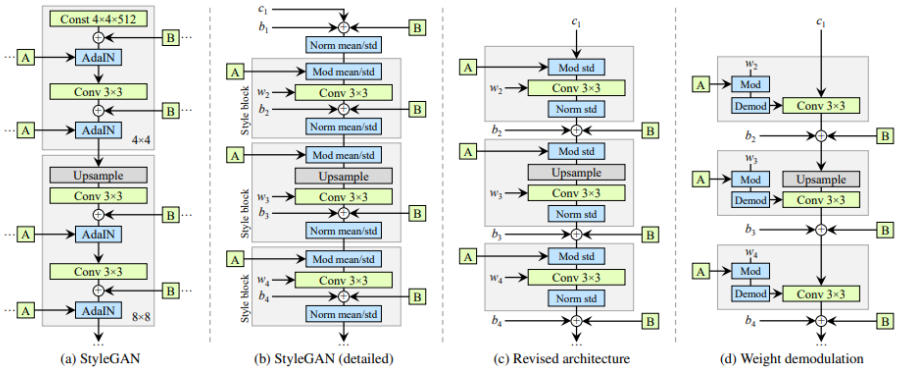


Figure 5.4.2: Redesigned architecture of the StyleGAN's synthesis network

Benefits.

– **Reduced Artifacts:** Eliminates many of the visible water-drop patterns previously caused by AdaIN's per-channel normalization.

– **Consistent Feature Representation:** Modulated convolutions preserve feature statistics more naturally, providing smoother outputs across the network.

- **Revised Architecture with Skip and Residual Connections** Rather than relying on *progressive growing*, StyleGAN2 employs a **fixed resolution** throughout training, combined with improved skip and residual connections:

  – **Skip Connections:** Early layers can pass low-level feature information directly to later layers, aiding gradient flow and retaining fine details.

  – **Residual Connections:** Certain blocks are restructured as shown in Figure 5.4.3c to have short pathways for features, improving training stability and convergence.

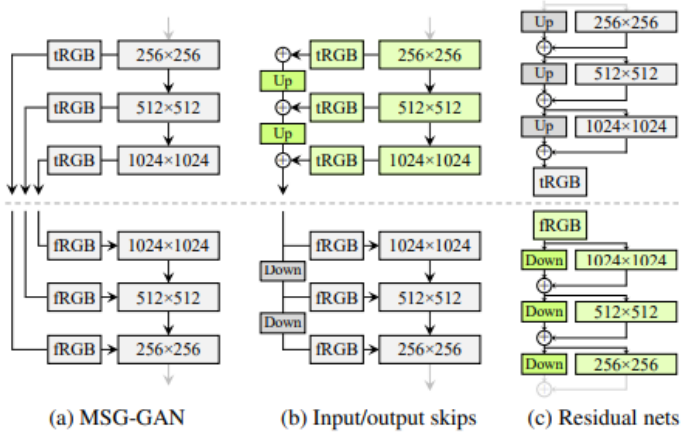


Figure 5.4.3: Three generator (above the dashed line) and discriminator architectures. Up and Down denote bilinear up and downsampling, respectively. In residual networks these also include 1×1 convolutions to adjust the number of feature maps. tRGB and fRGB convert between RGB and high-dimensional per-pixel data.

By adopting these connections, StyleGAN2 avoids some of the transitional artifacts that can appear when resolutions are scaled up progressively, and it streamlines the overall training procedure.

- **Path Length Regularization** One of the notable new regularization techniques in StyleGAN2 is *path length regularization*. This involves computing how changes in the intermediate latent vector **w** affect the generated images and penalizing overly large or abrupt changes. Concretely, it encourages a consistent "step size" in image space when interpolating between latent codes, yielding:
  - **Smooth Latent Interpolations:** Interpolations between two **w** vectors produce coherent morphs, without abrupt jumps in color or geometry.
  - **Better Latent Geometry:** The generator learns a smoother mapping function, making the latent space more intuitive for editing or semantic manipulations.

5.4.5 Comparison of StyleGAN and StyleGAN2 Performance

To quantitatively demonstrate the improvements of StyleGAN2 over StyleGAN, we refer to two commonly used metrics in GAN evaluation:

- **Fréchet Inception Distance (FID)[42]**: Measures the distance between the feature distributions of real and generated images. Lower FID indicates better image quality and diversity.
- **Perceptual Path Length (PPL)[49]**: Assesses the smoothness and disentanglement of the generator's latent space. Lower PPL typically indicates more predictable and semantically consistent interpolations.

Both models were evaluated on diverse datasets at high resolutions (e.g., FFHQ [49] and LSUN [135] categories). Table 5.4 summarizes FID and PPL scores for StyleGAN (original) and StyleGAN2 (best configuration) across several datasets.

Table 5.4: FID and PPL scores comparing StyleGAN vs. StyleGAN2 across common datasets.

Dataset	**Resolution**	**FID ↓**		**PPL ↓**	
		StyleGAN	**StyleGAN2**	**StyleGAN**	**StyleGAN2**
FFHQ	1024×1024	4.40	2.84	212.1	145
LSUN Cat	256×256	8.53	6.93	924	439
LSUN Horse	256×256	3.83	3.43	1405	338
LSUN Car	512×384	3.27	2.32	1484.5	415.5

As shown in Table 5.4, StyleGAN2 consistently achieves lower FID (implying higher image quality and diversity) and lower PPL (indicating smoother, more disentangled latent interpolations) across all tested datasets. These results highlight the significant performance gains delivered by StyleGAN2's revised architecture.

5.4.6 StyleGAN2-ADA: Adaptive Discriminator Augmentation

Motivation for StyleGAN2-ADA

While *StyleGAN2* improved upon its predecessor by reducing structural artifacts and improving overall image quality, it still heavily relied on large-scale datasets to train effectively. GANs, including StyleGAN2, require tens or hundreds of thousands of high-quality images to generalize well. However, many real-world applications, particularly in medical imaging, suffer from data scarcity due to the following constraints:

- **Limited Availability:** High-quality labeled medical datasets are difficult to collect and often contain only a few hundred or thousand samples.
- **Privacy and Ethical Concerns:** Sharing medical images is restricted due to patient privacy regulations (e.g., HIPAA, GDPR).
- **Data Imbalance:** Many medical conditions are rare, making it difficult to gather balanced datasets for training deep learning models.

The key challenge in low-data regimes is that the **discriminator** tends to *overfit* to the limited training samples. When this happens, the discriminator memorizes the dataset instead of learning meaningful feature representations. Consequently, the generator fails to learn properly, leading to training instability and reduced image diversity.

To address this issue, StyleGAN2-ADA (Adaptive Discriminator Augmentation) [53] introduces an adaptive augmentation strategy that stabilizes training and enables high-quality image generation from significantly smaller datasets.

Adaptive Discriminator Augmentation (ADA)

The core idea of ADA is to apply image augmentations to the discriminator's input images while ensuring that these augmentations do not propagate to the generator. Standard data augmentation methods, such as rotations, flips, and color distortions, can leak into the generated images if applied improperly. ADA prevents this issue by:

- Applying augmentations **only to the discriminator's training images**, ensuring the generator never directly observes augmented data.
- Introducing an adaptive probability p that controls the strength of augmentations. p starts at zero and increases only if the discriminator shows signs of overfitting.
- Using invertible and stochastic augmentations that do not bias the generator's learned distribution.

Overfitting Heuristic and Adaptive Control

The augmentation probability p is not fixed but instead *dynamically* adjusted based on the discriminator's behavior. Two heuristics guide the adjustment:

$$r_v = \frac{\mathbb{E}[D_{\text{train}}] - \mathbb{E}[D_{\text{validation}}]}{\mathbb{E}[D_{\text{train}}] - \mathbb{E}[D_{\text{generated}}]} \qquad r_t = \mathbb{E}[\text{sign}(D_{\text{train}})] \tag{5.4.3}$$

where D_{train},$D_{\text{validation}}$,$D_{\text{generated}}$ represents the discriminator's output for the training set,validation set and generated images respectively. If the discriminator becomes too confident (i.e., consistently assigns high confidence to real images and low confidence to generated ones), p is increased, thereby intensifying augmentations to regularize the discriminator. Conversely, if training stabilizes, p is reduced to prevent excessive augmentations that could slow convergence.

Types of Non-Leaking Augmentations

ADA uses a set of carefully selected augmentations that do not leak into the generator's learned distribution. These augmentations fall into the following categories:

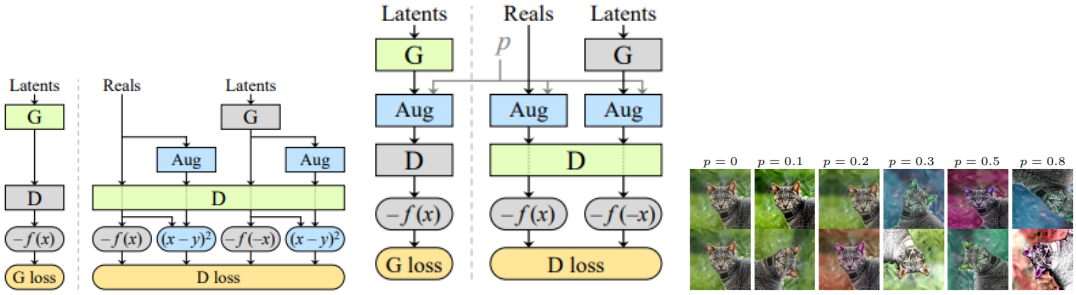
- **Pixel Blitting**: Horizontal flips, integer translations, 90° rotations.
- **Geometric Transformations**: Isotropic scaling, affine transformations.
- **Color Transformations**: Brightness, contrast, gamma correction, color jittering.
- **Filtering**: Gaussian blur, image sharpening.

5.4.7 Advantages of StyleGAN2-ADA for Medical Imaging

The ability to train high-quality GANs on small datasets makes StyleGAN2-ADA particularly well-suited for medical imaging applications, where dataset sizes are inherently limited. Some key benefits include:

1. Data Augmentation for Medical AI Models

Medical imaging datasets often contain only a few hundred labeled examples, making it difficult to train robust AI models. StyleGAN2-ADA can generate realistic synthetic medical images, expanding dataset size and improving model generalization.



(a) balanced Consistency regularization(bCR) (b) Stochastic discriminator augmentations (c) Effect of augmentation probability p

Figure 5.4.4: (a) Previous augmentation methods applied during GAN training . (b) Adaptive Discriminator Augmentation (ADA). (c) Diverse set of augmentations controlled by probability p

2. Counterfactual Image Generation

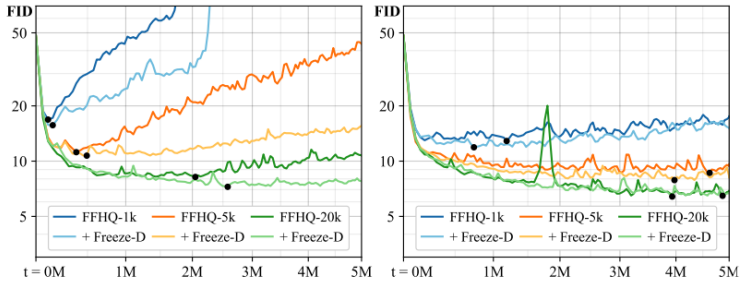
In medical research, generating counterfactual images—modifying specific patient attributes while preserving others—is crucial for studying disease progression, anomaly detection, and treatment planning. The disentangled latent space of StyleGAN2-ADA facilitates such edits.

3. Privacy-Preserving Data Synthesis

Since real medical images contain sensitive patient data, sharing datasets across institutions is often restricted. GAN-generated synthetic medical images can preserve privacy while retaining the statistical properties of real patient data, enabling safer collaborative research.

5.4.8 StyleGAN2-ada transfer-learning

Transfer learning significantly reduces the amount of training data required by initializing a model with weights pretrained on a different dataset, rather than using random initialization. Prior to the introduction of StyleGAN2-ADA, several studies explored the effectiveness of transfer learning in the context of GANs [128, 82, 85, 127]. The original StyleGAN2-ADA paper demonstrated that the most effective strategy involved freezing the highest-resolution layers of the discriminator (Freeze-D), leading to significantly improved performance compared to training from scratch. Notably, the success of transfer learning in GANs appears to depend more on the diversity of the source dataset rather than the similarity between the source and target domains.



(a) Without ADA (b) With ADA

Figure 5.4.5: Transfer learning FFHQ starting from a pre-trained model on CELEBA-HQ dataset.(a) Training convergence for StyleGAN2. (b) Training convergence for StyleGAN2-ada model.

StyleGAN2-ada on medical images. Several studies have investigated the use of StyleGAN2-ADA in medical imaging, taking advantage of its ability to generate high-fidelity images even when training data

is scarce. These works have primarily focused on data augmentation, improving classification models, and synthesizing diverse medical images for various downstream tasks. For example, [130] evaluated the model's performance on several publicly available medical imaging datasets, including SILVER07 (liver CT scans), ChestX-ray-14 (chest X-ray images), and the Medical Image Segmentation Decathlon (brain tumors). Similarly, [114] explored the effectiveness of StyleGAN2-ADA for brain MRI images with tumors, aiming to enhance limited brain MRI datasets through augmentation. Additionally, [131] utilized StyleGAN2-ADA for out-of-distribution detection in CT scans from various anatomical regions.

All these studies have demonstrated the remarkable effectiveness of adaptive discriminator augmentations (ADA), especially when combined with transfer learning from models trained on non-medical datasets such as FFHQ and CELEBA-HQ. Inspired by these findings, we adopt a similar training strategy for our StyleGAN model, leveraging both ADA augmentations and transfer learning to improve the quality and generalization of our generated medical images.

5.5 GAN Inversion: E4E and PTI

5.5.1 Motivation: The Need for High-Fidelity and Editable GAN Inversion

In the *PIECE* framework, counterfactual image generation relies on an optimization process applied directly to the latent vector of the original image.In the context of medical imaging, this necessitates a latent representation that is both **accurate** (i.e., it should reconstruct the input image with minimal distortion) and **editable** (i.e., it should allow controlled modifications for counterfactual synthesis). Traditional GAN inversion methods often struggle with the trade-off between these two properties:

- **Optimization-based approaches**, which directly optimize the latent code using a single sample, yield high reconstruction fidelity but result in latent codes that are difficult to edit.
- **Encoder-based approaches**, which train an encoder over a large number of samples, provide editable latent representations but fail to reconstruct fine details.
- **Hybrid approaches**, which first use an encoder to receive the initial latent vector and then perform a direct optimization on it.Even this method struggles to find a "sweet-spot" in this trade-off.

Related Work Many works have considered specifically the task of StyleGAN inversion with the aim of harnessing the high visual quality and editability of the model. In [1] it is demonstrated that inverting images in the native latent space of StyleGAN W introduces significant artifacts in the inverted image. It has been shown that inverting images into the extended embedding space $W+$ leads to more accurate reconstruction and better image preservation. [79] were the first to implement a direct optimization into the extended latent space $W+$. Also, [95] were the first to train an encoder for $W+$ inversion which was demonstrated to solve a variety of image-to-image translation tasks. Although the $W+$ inversion achieves minimal distortion , it has been shown that the results of latent manipulation on $W+$ inversions are inferior compared to the same manipulations over latent codes from the StyleGAN native space W.

To overcome this,in our proposed SPRUCE framework, we employ a two-step inversion process:

1. **E4E Encoder [117]**: Provides an initial editable latent representation.
2. **Pivotal Tuning Inversion (PTI) [96]**: Fine-tunes the generator for improved reconstruction accuracy.

5.5.2 Latent Space Representations in StyleGAN

StyleGAN defines multiple latent spaces, each impacting the quality, editability, and realism of the generated images. Understanding these spaces is crucial for selecting an appropriate GAN inversion method.

Native Latent Spaces

StyleGAN operates with two primary latent spaces:

- **Z-space**: The original latent space, where latent codes $z \sim \mathcal{N}(0, I)$ are sampled from a standard normal distribution. This space is highly entangled, making direct modifications challenging.
- **W-space**: A more structured and disentangled space obtained via a learned mapping network M, where $w = M(z)$. This space enables controlled attribute modifications while preserving realism.

Mathematically, the mapping from **Z**-space to **W**-space is defined as:

$$w = M(z), \quad z \sim \mathcal{N}(0, I), \tag{5.5.1}$$

where M represents the learned mapping function.

Extended Latent Spaces

To improve reconstruction capabilities and allow fine control over attributes, StyleGAN introduces extended latent spaces:

- $\mathbf{W}^+$**-space**: Unlike $\mathbf{W}$-space, where a single latent code is applied to all generator layers, $\mathbf{W}^+$ assigns independent latent codes to each layer of the generator, allowing finer control over local image features.
- $\mathbf{W^k}$**-space**: A further extension where only a subset k of layers receive distinct latent codes, providing a balance between expressiveness and controllability.

The $\mathbf{W}^+$-space is mathematically defined as:

$$w^+ = (w_1, w_2, ..., w_n), \quad w_i \in \mathbf{W}, \tag{5.5.2}$$

where each w_i represents the style latent code for a specific generator layer.

The $\mathbf{W}^+$-space enables high-quality image reconstructions by giving per-layer control over style and structure. However, it introduces a key trade-off between distortion minimization and editability, which we discuss next.

5.5.3 The GAN Inversion Trade-offs

The process of GAN inversion involves mapping a real image I to a latent code w such that the generated image $G(w)$ closely matches I. However, this process is constrained by two competing objectives:

- **Distortion minimization**: Ensuring that the reconstructed image is visually identical to the input.
- **Editability**: Maintaining a latent representation that allows for meaningful counterfactual modifications.

The latent space used for inversion significantly affects the trade-off between these objectives:

- $\mathbf{W}$-space offers structured and interpretable latent codes, making it ideal for semantic edits, but it lacks the expressiveness required for fine-grained reconstructions.
- $\mathbf{W}^+$-space allows near-exact reconstructions but often leads to inferior latent manipulations, as the structure of the latent space is less constrained.

Thus, while $\mathbf{W}^+$ inversion achieves minimal distortion, it has been shown that the results of latent manipulations are often less effective compared to latent codes from the $\mathbf{W}$-space [1, 95].

5.5.4 E4E Encoder for GAN Inversion

E4E (Encoder for Editing) was introduced to tackle these inversion challenges by strategically placing latent codes within the extended latent space while ensuring compatibility with semantic modifications.

Key Design Principles of E4E

1. **Minimizing Variation in Latent Codes** - Unlike traditional encoders, E4E is trained to progressively refine latent representations, allowing it to optimize both reconstruction fidelity and editability.

2. **Minimizing Deviation from W** - A dedicated loss function encourages latent codes to remain closer to $\mathbf{W}$ while leveraging the flexibility of $\mathbf{W}^+$, improving editability.

3. **ResNet-Based Encoder Architecture** - E4E builds on a ResNet-like backbone, producing an initial latent code and additional offsets that fine-tune specific layers.

5.5.5 Total Loss Function in E4E

The E4E encoder is trained using a loss function designed to balance two critical objectives:

- **Distortion Minimization**: Ensuring that the reconstructed image closely resembles the original.

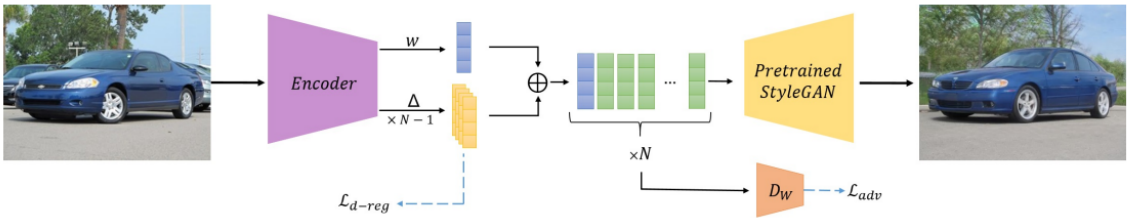


Figure 5.5.1: The E4E network architecture. The encoder receives an input image and outputs a single style code w together with a set of offsets $\Delta_1, \ldots, \Delta_{N-1}$, where N denotes the number of StyleGAN's style modulation layers. The final latent representation is obtained by replicating the w vector N times and adding each Δ_i to its corresponding entry.

- **Editability Preservation**: Maintaining a structured latent code to enable meaningful modifications.

The total loss function is defined as a weighted combination of these two terms:

$$L(x) = L_{\text{dist}}(x) + \lambda_{\text{edit}} L_{\text{edit}}(x), \tag{5.5.3}$$

where λ_{edit} controls the trade-off between reconstruction fidelity and latent code flexibility.

Distortion Loss

To minimize reconstruction error, the distortion loss consists of three components:

$$L_{\text{dist}}(x) = \lambda_2 L_2(x) + \lambda_{lpips} L_{LPIPS}(x) + \lambda_{sim} L_{\text{sim}}(x). \tag{5.5.4}$$

- $L_2(x)$ - Standard pixel-wise reconstruction loss. - $L_{LPIPS}(x)$ - Perceptual loss to enforce structural similarity between images. - $L_{\text{sim}}(x)$ - Identity loss, ensuring feature-level consistency using a pre-trained network.

The identity loss is defined as:

$$L_{\text{sim}}(x) = 1 - \left(C(G(E4E(x))), C(G(x))\right), \tag{5.5.5}$$

where:

- C is a pre-trained ResNet-50 network that extracts feature embeddings.
- G is the pre-trained StyleGAN2 generator.
- This term ensures that the generated image maintains identity consistency with the original.

Editability Loss

To ensure that latent codes remain structured and editable, the editability loss consists of:

$$L_{\text{edit}}(x) = \lambda_{\text{d-reg}} L_{\text{d-reg}}(x) + \lambda_{\text{adv}} L_{\text{adv}}(x). \tag{5.5.6}$$

- $L_{\text{d-reg}}(x)$ - Delta-regularization loss that constrains offsets Δ_i, ensuring proximity to W-space. - $L_{\text{adv}}(x)$ - Adversarial loss using a latent discriminator to keep the learned style codes within the native StyleGAN distribution.

5.5.6 Fine-Tuned GAN Inversion: Pivotal Tuning Inversion (PTI)

Motivation for PTI in Medical Image Inversion While E4E provides an editable latent representation, it does not always guarantee high-fidelity reconstructions, especially for medical images that contain fine-grained, high-resolution anatomical structures. In medical imaging applications, even minor reconstruction artifacts can obscure critical diagnostic details. This necessitates an inversion method that enhances reconstruction accuracy while maintaining editability.

Pivotal Tuning Inversion (PTI) [96] addresses this challenge by fine-tuning the generator itself, ensuring that the inverted latent code produces an image that is indistinguishable from the input while still allowing meaningful counterfactual modifications. Unlike conventional GAN inversion methods that focus solely on mapping the image to a pre-trained latent space, PTI locally adjusts the generator's weights, enabling precise reconstructions even when the StyleGAN model struggles with the dataset.

Methodology

PTI consists of two primary steps:

1. **Initial GAN Inversion:** The image is first inverted using an off-the-shelf encoder-based method (e.g., E4E) to obtain an initial latent code w in the latent space W^+.
2. **Generator Fine-Tuning:** The generator G is fine-tuned to better reconstruct the target image while keeping the latent code stable, thus improving the trade-off between distortion minimization and editability.

In the original PTI paper the authors also suggest the solution of the following optimization problem as an alternative way to get the initial latent vector w_p from the editable latent space W :

$$w_p = \arg\min_w L_{\text{LPIPS}}(I, G(w, \theta)) + \lambda_n L_n(n), \tag{5.5.7}$$

where:

- $G(w, \theta)$ is the image generated using a pre-trained generator G with weights θ.
- L_{LPIPS} represents the perceptual loss used to optimize reconstruction fidelity.
- $L_n(n)$ is a noise regularization term that prevents unwanted artifacts in the generated image.
- λ_n is a hyperparameter that controls the contribution of noise regularization.

Pivotal Tuning: Generator Fine-Tuning

After obtaining an initial latent code w, a direct inversion may still produce images with minor distortion. Therefore, in the second step, PTI freezes the latent code w_p and fine-tunes the generator weights θ to improve the reconstruction fidelity while keeping the inverted code unchanged:

$$L_{\text{pti}} = L_{\text{LPIPS}}(x, x^p) + \lambda_{L2} L_{L2}(x, x^p), \tag{5.5.8}$$

where:

- $x^p = G(w_p, \theta^*)$ is the reconstructed image using the fine-tuned generator.
- L_{L2} enforces pixel-wise reconstruction accuracy.

Since the generator is initialized with pre-trained weights θ, PTI ensures that the fine-tuned model still retains its original generative capacity while adapting locally to better reconstruct the input image.

Locality Regularization

To prevent overfitting to a single image, locality regularization is introduced. This ensures that the generator does not lose its ability to produce diverse samples from the original dataset. In each iteration,a normally distributed random vector z is sampled and StyleGAN's mapping network f is used to produce a corresponding latent code $w_z = f(z)$. Then, interpolation between w_z and the pivotal latent code w_p is applied using the interpolation parameter α, to obtain the interpolated code w_r:

$$w_r = w_p + \alpha \frac{w_z - w_p}{\|w_z - w_p\|_2}. \tag{5.5.9}$$

Then, the locality regularization loss is defined as:

$$\mathcal{L}_{reg} = \mathcal{L}_{\text{LPIPS}}(x_r, x_r^*) + \lambda_{L2}^R \mathcal{L}_{L2}(x_r, x_r^*). \tag{5.5.10}$$

where:

- $x_r = G(w_r, \theta)$, is the image generated from the original generator given the latent vector w_r
- $x_r^* = G(w_r, \theta^*)$, is the image generated from the finetuned generator given the latent vector w_r

Finally, the generator optimization is formulated as:

$$\theta^* = \arg\min_{\theta^*} L_{\text{pti}} + \lambda_{\text{reg}} L_{\text{reg}}. \tag{5.5.11}$$

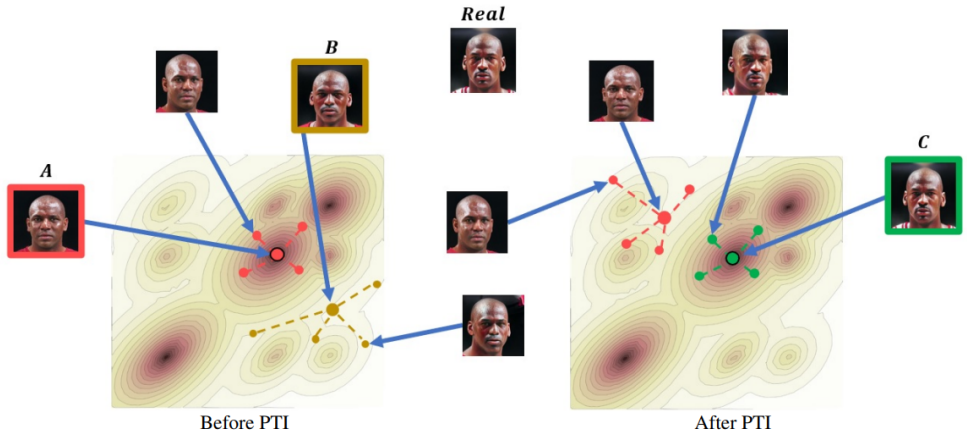


Figure 5.5.2: An illustration of the PTI method. StyleGAN's latent space is portrayed in two dimensions, where the warmer colors indicate higher densities of W, i.e. regions of higher editability. On the left, we illustrate the generated samples before pivotal tuning. We can see the Editability-Distortion trade-off. A choice must be made between Identity "A" and Identity "B". "A" resides in a more editable region but does not resemble the "Real" image. "B" resides in a less editable region, which causes artifacts, but induces less distortion. On the right - After the pivotal tuning procedure. "C" maintains the same high editing capabilities of "A", while achieving even better similarity to "Real" compared to "B".

5.6 Latent Vector Optimization

The final step of SPRUCE is to iteratively optimize ,in the latent space , the latent vector that reconstructs the initial image in order to get the latent representation of the counterfactual image.To generate counterfactual images that meaningfully explain the classifier's decision, we optimize the latent vector of the StyleGAN2-ADA generator, that we have obtained using the steps described in 5.5 , in such a way that:

- The features of the image that is generated from the optimized latent vector align with the exceptional features extracted through PIECE.
- The modified image remains perceptually similar to the original image, avoiding excessive distortions.
- The modifications are sparse, meaning that only the essential aspects contributing to the classifier's decision are altered.

Optimization Objective

The optimization process is performed on the latent vector which is initialized as w_p . The optimization objective consists of four key loss terms:

$$L = L_{\text{piece}} + \lambda_{\text{perc}} L_{\text{LPIPS}} + \lambda_{\text{latent}} L_{\text{latent}} + \lambda_{\text{image}} L_{\text{image}}. \tag{5.6.1}$$

We explain the significance of each term in the following paragraphs.

5.6.1 Feature Alignment Loss

Having constructed the counterfactual feature vector x_p , we use the same feature-level loss as in the original PIECE algorithm :

$$L_{\text{piece}} = \|C(G(w_e)) - x'\|_2^2. \tag{5.6.2}$$

Here:

- C represents all the layers of the frozen classifier up to the penultimate feature layer X.
- x' is the counterfactual feature vector
- G is the finetuned generator
- w_e is the optimized latent vector.

This term ensures that the features of the counterfactual image match those expected for the target class, thus ensuring that the classifier's prediction will change.

5.6.2 Perceptual Similarity Loss (LPIPS)

To preserve structural integrity and prevent the counterfactual image from deviating too far from the original, we incorporate a perceptual similarity loss using the Learned Perceptual Image Patch Similarity (LPIPS) metric [138]:

$$L_{\text{LPIPS}} = \text{LPIPS}(G(w_e), I). \tag{5.6.3}$$

where:

- $G(w_e)$ is the generated counterfactual image.
- I is the original image taken from the dataset.

LPIPS ensures that high-level perceptual features are preserved while enabling meaningful modifications. For the LPIPS backbone network, we utilize a VGG-based feature extractor instead of AlexNet. This choice is motivated by VGG's widespread adoption in optimization schemes involving StyleGAN variants, as well as its deeper architecture, which captures more expressive hierarchical features, making it better suited for complex image domains such as medical imaging.

5.6.3 Latent Space Sparsity Regularization

We also apply an L1 regularization to limit any insignificant changes in the latent space:

$$L_{\text{latent}} = \|w_e - w_p\|_1. \tag{5.6.4}$$

This term prevents the optimized latent vector to drift away from the editable and semantically rich latent neighborhood of the initial latent vector w_p, thus boosting the realism and plausibility of the generated counterfactuals.

5.6.4 Image Space Sparsity Regularization

To ensure that counterfactual images remain faithful to the original while reflecting only the necessary modifications, we introduce an L1 penalty on pixel-wise differences between the generated counterfactual image and the original image:

$$L_{\text{image}} = \|G(w_e) - I\|_1. \tag{5.6.5}$$

This regularization encourages sparsity in pixel changes, ensuring that only the most relevant regions of the image, those responsible for the classifier's decision, are altered.

Thus, the final counterfactual optimization problem is formulated as:

$$w_e^* = \arg\min_{w_p} \Big[L_{\text{piece}} + \lambda_{\text{perc}} L_{\text{LPIPS}} + \lambda_{\text{latent}} L_{\text{latent}} + \lambda_{\text{image}} L_{\text{image}} \Big]. \tag{5.6.6}$$

where:

- w_e^* is the latent vector corresponding to the final counterfactual image.
- $\lambda_{\text{perc}}, \lambda_{\text{latent}}, \lambda_{\text{image}}$ are hyperparameters that control the trade-offs between feature alignment, perceptual similarity, and sparsity.

By inputting w_e^* into the finetuned generator G we can visualize the counterfactual image.

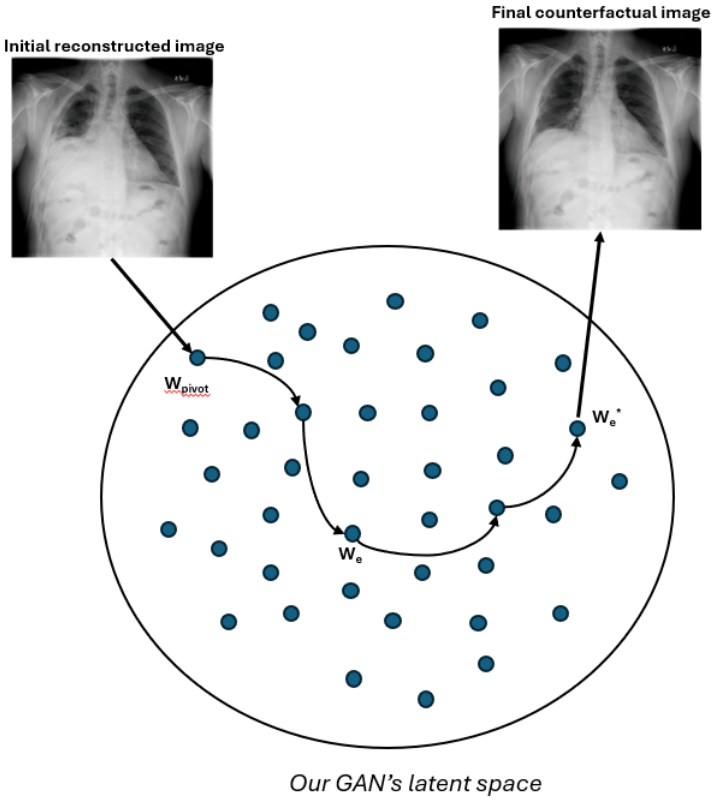


Figure 5.6.1: Visualization of the optimization process

Optimization Strategy

To optimize the latent vector, we employ the AdamW optimizer [68] with a cosine annealing learning rate schedule [67] for smooth convergence and gradient clipping to prevent exploding gradients .

The final optimization is performed iteratively for T steps, yielding the optimized latent vector w_e^*, which is then used to synthesize the counterfactual image.

Semifactual generation

During each optimization step, the generated image is evaluated by the classifier, and its corresponding probability score is recorded. This iterative process allows us to systematically track how the classifier's decision transitions from the original class to the counterfactual class, providing valuable insights into the model's decision boundaries and the impact of the applied modifications.

5.7 Adversarial Robustness for Meaningful Counterfactual Explanations

5.7.1 Motivation: The Role of Adversarial Robustness

As we have seen in the previous sections of our framework, the classifier's feature vector is used to drive the generator of the StyleGAN2-ada model towards producing an image that belongs to the desired counterfactual class and that is as close as possible to the original image. In other words, we can say that the gradients of the classifier are those that guide the generator to the right direction. Hence if the gradients are not perceptually aligned with the features of a particular class then it could result , as we will see in the experiments section , in counterfactuals that look visually similar to the original image when the changes are constrained to be minimal. This means that if the classifier is vulnerable to adversarial human-imperceptible modifications (adversarial attacks) [33] then the generated counterfactuals will reflect those vulnerabilities rather than true causal attributions.

It has been proven that the gradients of adversarially robust classifiers have strong generative properties and capture features that are perceptually aligned with the features of a certain class [100, 7]. For that reason,we employ adversarial training to improve the classifier's robustness and ensure that the generated images contain meaningful and plausible changes. Specifically, we use TRADES (TRade-off-inspired Adversarial DEfense via Surrogate-loss minimization) [136], a state-of-the-art adversarial training method that balances clean accuracy and robustness.

5.7.2 TRADES: Balancing Clean Accuracy and Robustness

The TRADES framework formulates adversarial training as a trade-off between:

- **Standard Classification Accuracy:** Ensure that the classifier maintains high accuracy on clean (unperturbed) images, which is indispensable especially in the domain of medical imaging.
- **Adversarial Robustness:** Training the classifier to be invariant to adversarial perturbations.

This perturbation forces the model to minimize the discrepancy between its response to clean and adversarial examples.

5.7.3 Mathematical Formulation of TRADES

The core idea of TRADES is formulated as a regularized surrogate loss, combining a standard classification loss and a robustness regularization term. Specifically, TRADES optimizes the following objective:

$$\min_f \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(f(x), y) + \beta \cdot \max_{x' \in \mathcal{B}_\epsilon(x)} \mathcal{L}_{\mathrm{KL}}(f(x), f(x')) \right], \tag{5.7.1}$$

where:

- $f(x)$ represents the classifier's predicted probability distribution for input x,
- $\mathcal{L}$ denotes the standard classification loss (e.g., cross-entropy),
- $\mathcal{L}_{\mathrm{KL}}$ is the Kullback-Leibler divergence that quantifies the divergence between the classifier outputs on the clean example x and the adversarial example x',
- $\mathcal{B}_\epsilon(x)$ is the ϵ-bounded perturbation ball around input x, defined as $\mathcal{B}_\epsilon(x) = \{x' : \|x' - x\|_p \leq \epsilon\}$,
- β is a hyperparameter controlling the trade-off between natural accuracy and adversarial robustness.

During training, adversarial examples x' are generated within the perturbation ball using Projected Gradient Descent (PGD) [73]. PGD iteratively perturbs the input data in the direction of the gradient of the KL divergence loss to find a worst-case adversarial perturbation. This iterative adversarial example generation can be formally expressed as:

$$x'_{t+1} = \Pi_{\mathcal{B}_\epsilon(x)} \left(x'_t + \alpha \cdot \text{sign}(\nabla_{x'_t} \mathcal{L}_{\text{KL}}(f(x), f(x'_t))) \right), \tag{5.7.2}$$

where $\Pi_{\mathcal{B}_\epsilon(x)}(\cdot)$ denotes projection onto the ϵ-ball, α is the step size, and t indicates the iteration number.

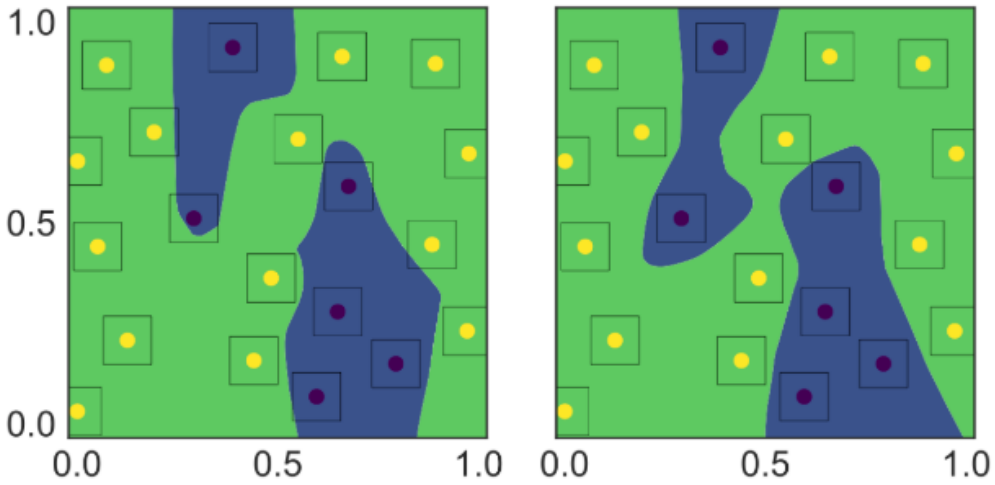


Figure 5.7.1: Left figure: decision boundary by natural training. Right figure: decision boundary by TRADES.

5.7.4 Comparison to Standard Adversarial Training

Unlike conventional adversarial training methods that simply minimize loss on adversarial examples, TRADES explicitly balances robustness and accuracy, making it particularly suitable for our application. Standard adversarial training methods suffer from:

- Excessive accuracy drop on clean images when increasing robustness.
- Gradient obfuscation, which can limit counterfactual interpretability.
- Overfitting to adversarial attacks, reducing generalization.

TRADES mitigates these issues by controlling the KL divergence loss with the hyperparameter β, allowing us to tune robustness without severely degrading classification accuracy.

Chapter 6

Experiments and Results

Contents

6.1 Datasets

6.1.1 Introduction

In this section, we describe the datasets used for evaluating the proposed framework for counterfactual generation. Four publicly available medical image datasets were utilized, covering different medical imaging modalities. The datasets include two binary classification datasets based on chest X-ray images, a four-class Optical Coherence Tomography (OCT) dataset, and a four-class Brain MRI dataset for dementia classification. Apart from referencing numerical details it is also essential to describe the visual characteristics that define each class within the datasets , as these will help us evaluate the plausibilty and meaningfulness of the produced counterfactuals.

6.1.2 Dataset Descriptions

Chest X-ray Datasets

Lung Opacity vs. Normal Dataset

- **Source:** RSNA Pneumonia Detection Challenge [2]. The original RSNA Pneumonia Detection Challenge dataset contains 29,700 frontal-view x-ray images of 26,600 patients. The training daa is split into three classes : Normal , Lung Opacity and No Lung Opacity / Not Normal .We used only the classes Normal and Lung Opacity since we only wanted to train the classifier to distinct between lungs suffering from pneumonia and healthy lungs. Other anomalies that do not result in opacities in the lungs are excluded from the training task to keep it a binary classification problem. All duplicates from the same patients were removed as well.
- **Classes:** Lung Opacity, Normal
- **Number of Samples:** 6012 images of Lung Opacity, 8851 images of Normal
- **Preprocessing**: We resized the original images from 1024×1024 to 256×256 resolution and made sure that all images are RGB.
- **Description:**
  - **Lung Opacity:** Features include regions of increased opacity in lung fields, reduced contrast, and irregular texture patterns indicative of pneumonia.
  - **Normal:** Clear lung fields with well-defined vascular structures, normal lung aeration, and no abnormal opacities.

Cardiomegaly vs. Normal Dataset

- **Source:** NIH Chest X-ray Dataset[126] and RSNA Pneumonia Detection Challenge[2].The cardiomegaly images were sourced from the NIH Chest X-ray Benchmark Dataset. Due to the limited number of images labeled exclusively as "Cardiomegaly," additional images were included that contained this label alongside other lung-related conditions. The normal images were obtained from the RSNA Pneumonia Detection Challenge to maintain consistency and balance between classes.
- **Classes:** Cardiomegaly, Normal
- **Number of Samples:** 2776 images of Cardiomegaly, 2776 images of Normal
- **Preprocessing**: We resized the original images from 1024×1024 to 256×256 resolution and made sure that all images are RGB.
- **Description:**
  - **Cardiomegaly:** Enlarged cardiac silhouette relative to thoracic cavity, often exceeding 50% of thoracic width.
  - **Normal:** Proportionate heart size within normal limits, clear lung fields, and absence of abnormal cardiac enlargement.

Optical Coherence Tomography (OCT) Dataset

[55]

- **Source:** Kaggle
- **Classes:** Normal, Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), Drusen
- **Number of Samples:** 26347 images of Normal, 37216 images of CNV, 11422 images of DME, 8620 images of Drusen
- **Preprocessing**: All images were center cropped to square dimensions 256×256 and converted to RGB. Original images resolution was either 496×512 or 496×768.
- **Description:**
  - **Choroidal Neovascularization (CNV):** Presence of hyperreflective fluid accumulations, distorted retinal layers, and irregular subretinal deposits.
  - **Diabetic Macular Edema (DME):** Thickened retina with cystoid spaces, hyporeflective lesions indicating fluid buildup.
  - **Drusen:**Extracellular deposits seen between the retinal pigment epithelium and Bruch's membrane.
  - **Normal:** Uniformly structured retinal layers with no signs of pathology.

Brain MRI Dementia Dataset

- **Source:** Kaggle
- **Classes:** No Impairment, Very Mild, Mild, Moderate
- **Number of Samples:** 3200 images of No Impairment, 3008 images of Very Mild, 2739 images of Mild, 2572 images of Moderate
- **Preprocessing**: We resized the original images to 128×128 resolution and converted them to RGB.
- **Description:**
  - **No Impairment:** Normal brain volume with well-preserved cortical structures and no significant atrophy.
  - **Very Mild Impairment:** Subtle cortical thinning, slight hippocampal volume reduction.
  - **Mild Impairment:** Noticeable atrophy in temporal lobes, enlarged ventricles, and white matter hyperintensities.
  - **Moderate Impairment:** Severe hippocampal atrophy, significant cortical thinning, and prominent ventricular enlargement.

6.1.3 Dataset Sample Images

To provide a visual reference of the datasets used, Figures 6.1.1 through 6.1.11 present sample images for each class within the respective datasets.

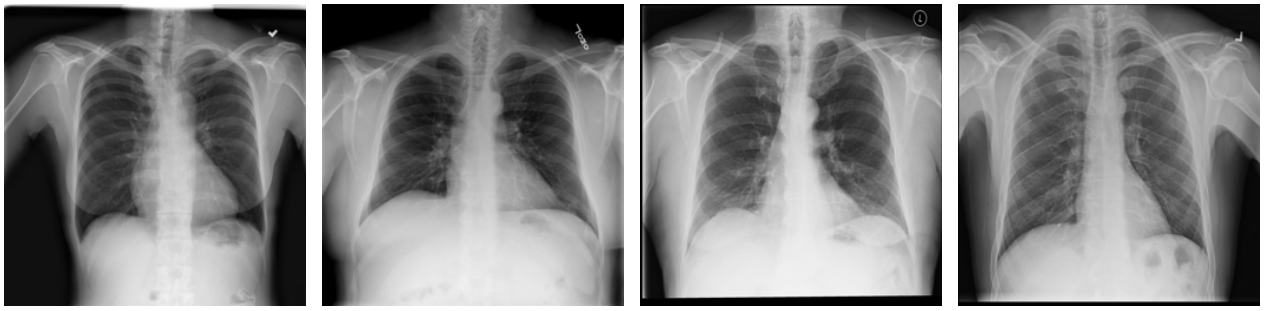


Figure 6.1.1: Healthy chest x-ray images.

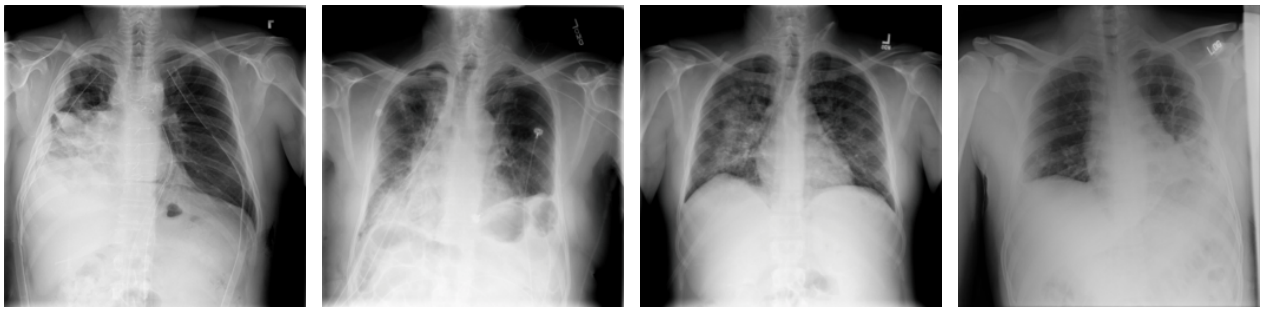


Figure 6.1.2: Chest x-ray images with lung opacity.

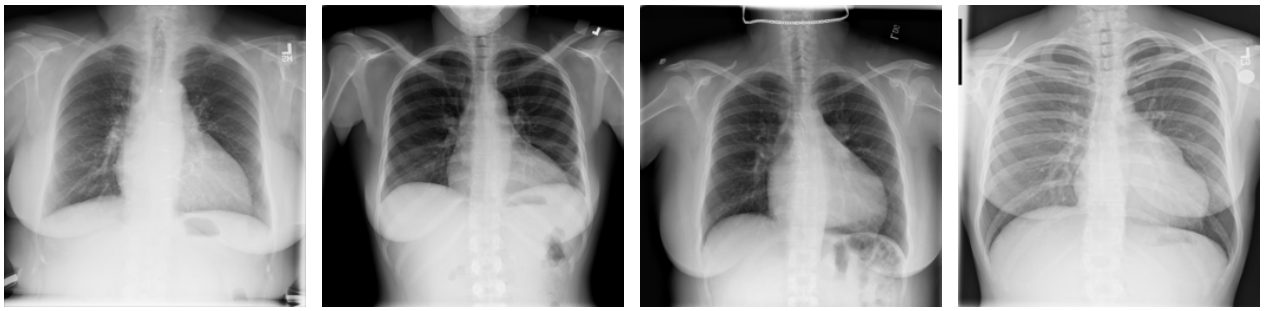


Figure 6.1.3: Chest x-ray images with cardiomegaly.

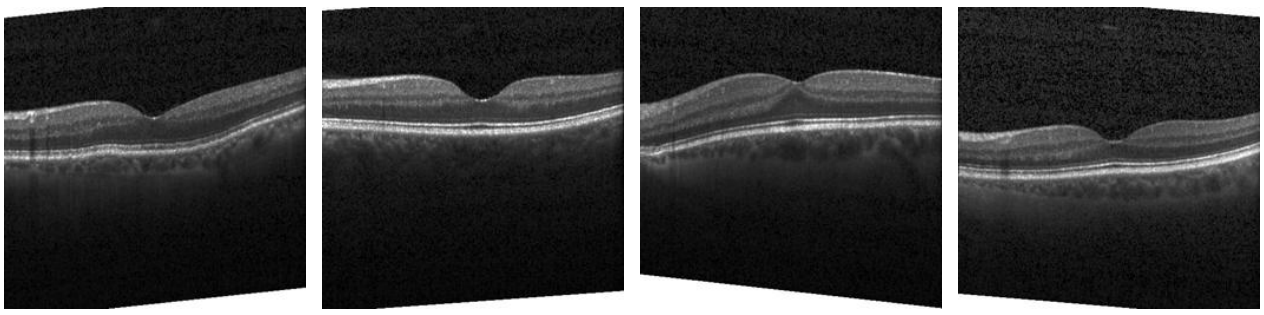


Figure 6.1.4: Healthy OCT samples.

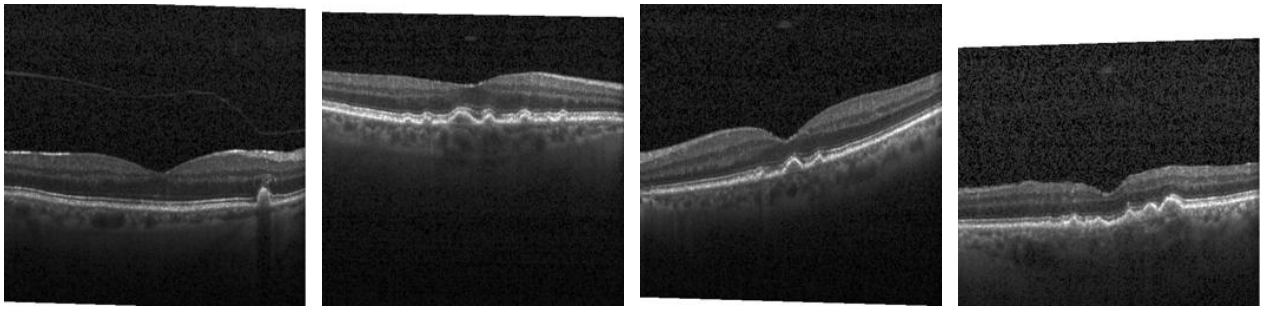


Figure 6.1.5: OCT samples with drusen.

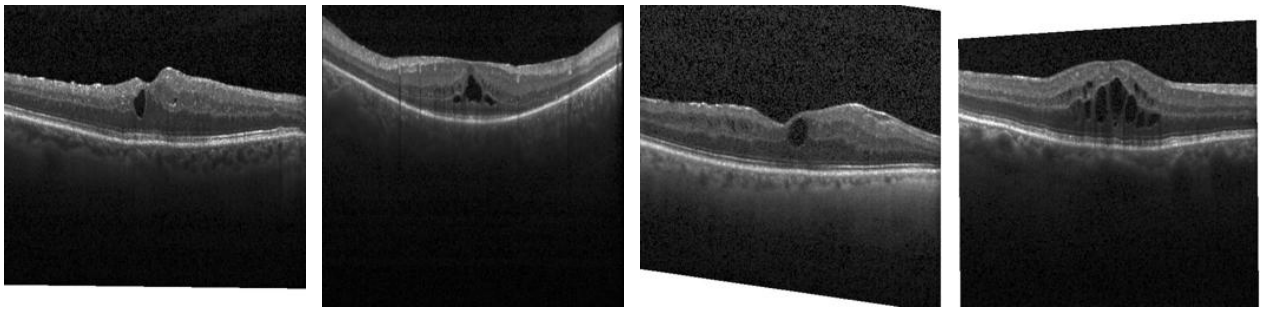


Figure 6.1.6: OCT samples with signs of Diabetic Macular Edema(DME).

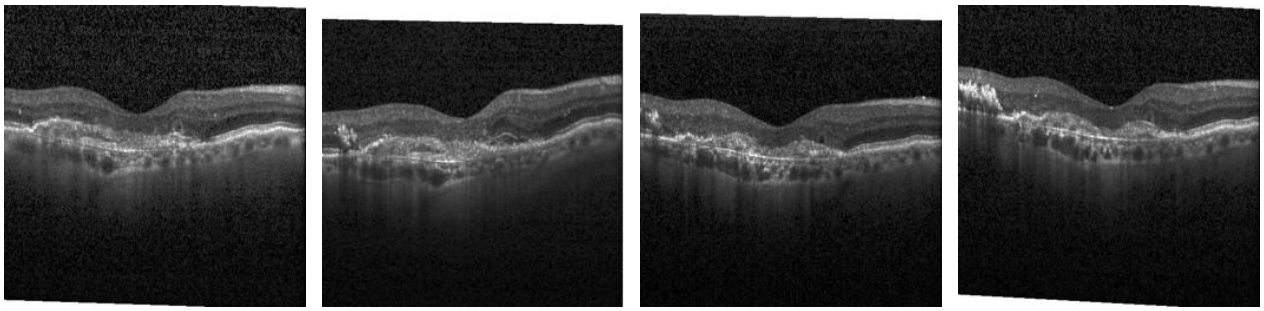


Figure 6.1.7: OCT samples with signs of Choroidal Neovascularization (CNV).

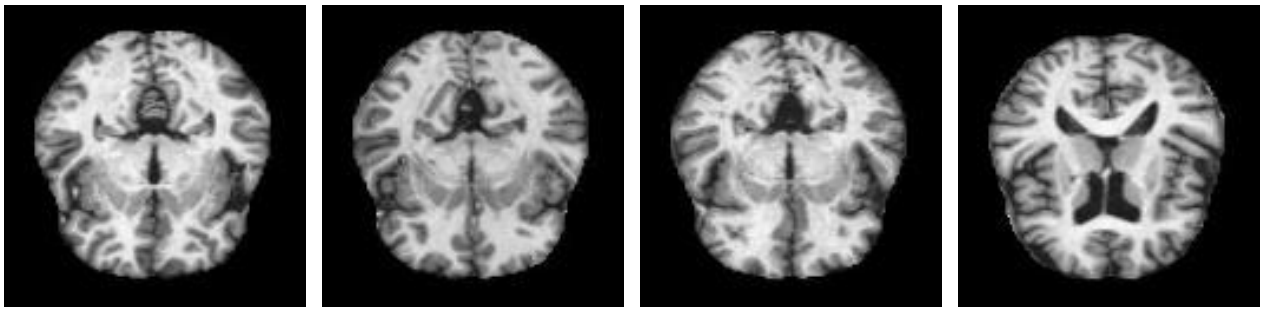


Figure 6.1.8: Brain MRIs with no signs of dementia.

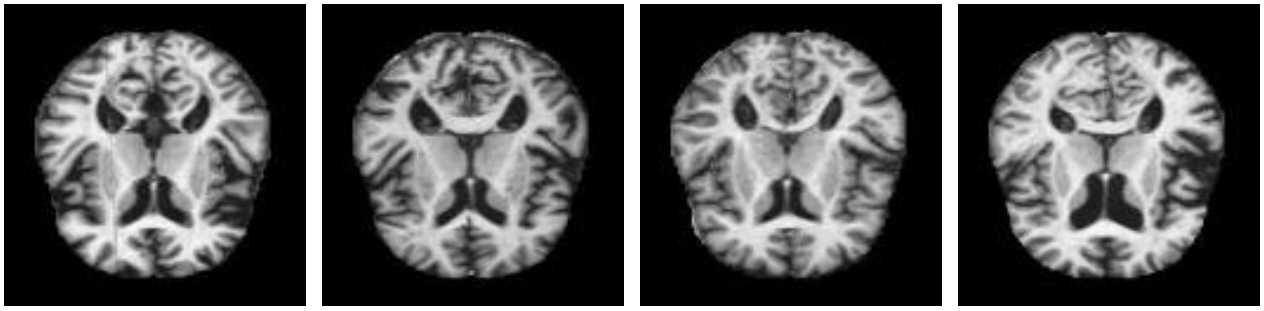


Figure 6.1.9: Brain MRIs with very mild impairment.

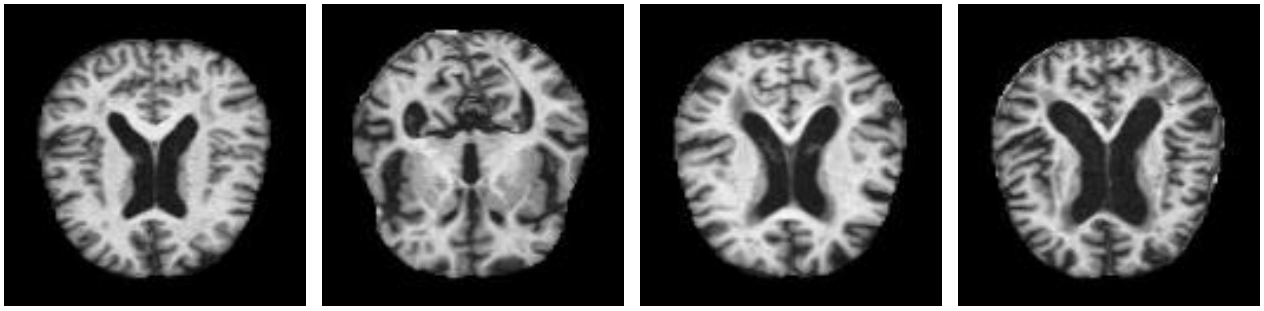


Figure 6.1.10: Brain MRIs with Mild Impairment.

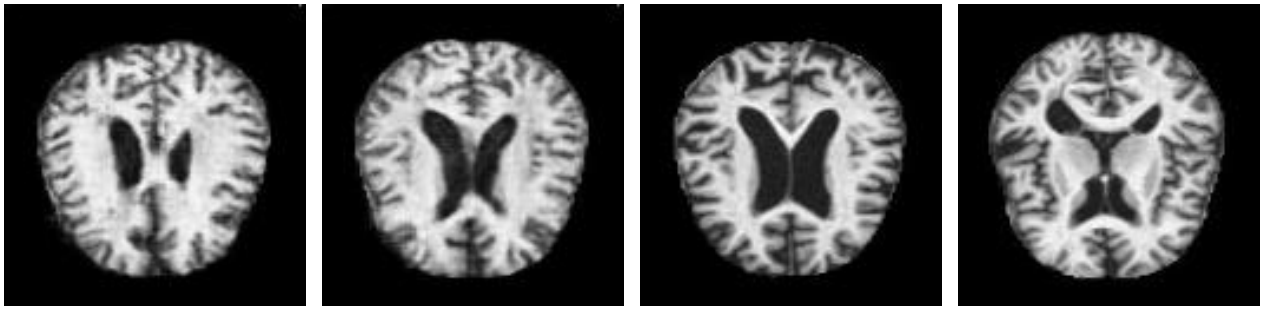


Figure 6.1.11: Brain MRIs with moderate impairment.

6.2 Classifier training

As we discussed in Section 5.2,the model that we chose to use in our counterfactual framework is the ConvNeXt-Base model. In this Section we are going to provide all the techniques and parameters used during both the training of a plain and an adversarially robust model. We will present the results in each case using various evaluation metrics and plots .Finally we will showcase several saliency maps[106] that will help us understand the image regions on which the model gives more attention to , but will also prove the necessity of using adversarially robust classifiers for plausible and meaningful counterfactuals.

6.2.1 General Training Configuration

The training of the classifier followed a consistent setup across both standard and adversarial training schemes. The following parameters and techniques remained constant throughout:

Dataset Splitting All datasets were divided into three subsets:

- **Training Set:** 80% of the dataset, used for model training.
- **Validation Set:** 10%, used for hyperparameter tuning and early stopping.
- **Test Set:** 10%, used for evaluating final model performance.

Dataset	Class	Training	Validation	Test	Resolution
Chest X-ray (Lung Opacity vs. Healthy)	Healthy	7080	885	601	256×256
	Lung Opacity	4809	601	602	
Chest X-ray (Cardiomegaly vs. Healthy)	Healthy	2220	277	279	256×256
	Cardiomegaly	2220	277	279	
OCT Dataset	Normal	21077	2634	2636	256×256
	CNV	29772	3721	3723	
	DME	9137	1142	1143	
	Drusen	6896	862	862	
Brain MRI Dementia Dataset	No Impairment	2560	320	320	128×128
	Very Mild Impairment	2406	300	302	
	Mild Impairment	2191	273	275	
	Moderate Impairment	2057	257	258	

Table 6.1: Number of images per dataset split (Train, Validation, Test) for each class with resolution details.

Model Fine-tuning To leverage pretrained knowledge and accelerate convergence, we performed transfer learning by initializing our model with pretrained weights from the ImageNet-1k V1 dataset [20].The classifier head was replaced with a fully connected layer corresponding to the number of classes in each dataset.

Optimization and Regularization Techniques To enhance model generalization and robustness, the following optimization strategies were applied:

- **Learning Rate Scheduling:** Cosine annealing learning rate scheduler .
- **Weight Decay:** Applied to prevent overfitting and improve generalization.
- **Early Stopping:** Terminated training when validation loss did not improve after 20 epochs
- **Warmup Epochs:** Gradual increase in learning rate at the start of training.
- **Mixed Precision Training [81]:** Reduced memory footprint and improved computational efficiency.
- **Label Smoothing:** Applied to prevent overconfidence in model predictions by assigning a small probability to incorrect labels, thereby improving model calibration and robustness.

- **Class Weighting:** In the cases of the Normal vs Lung Opacity and the OCT dataset a weighted cross-entropy loss was used to confront the imbalancies across different classes. The weight for each class was computed based on its frequency in the training set.

Data Augmentation To enhance model generalization, data augmentation techniques were applied to the training set:

- **Random Resized Cropping:** Scaling image patches between 75% and 100%.
- **Random Horizontal Flipping:** Probability of 30%.
- **Random Affine Transformations:** Rotation (10 degrees), translation (2%), scaling (0.95–1.05), and shear (5 degrees).
- **Color Jitter:** Adjustments in brightness and contrast within range 0.2.
- **Normalization:** ImageNet normalization with mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225].

6.2.2 Plain Classifier Training

This section details the hyperparameters and results obtained from training a classifier only on "clean" images using the strategy described in the previous section .

The hyperparameters used during training are presented in Table 6.2.

Hyperparameter	Value
Loss Function	Cross Entropy Loss
Initial Learning Rate	5e-4
Epochs	100
Warmup Epochs	10
Optimizer	AdamW
Batch Size	64
Weight Decay	5e-2
Label Smoothing	0.05

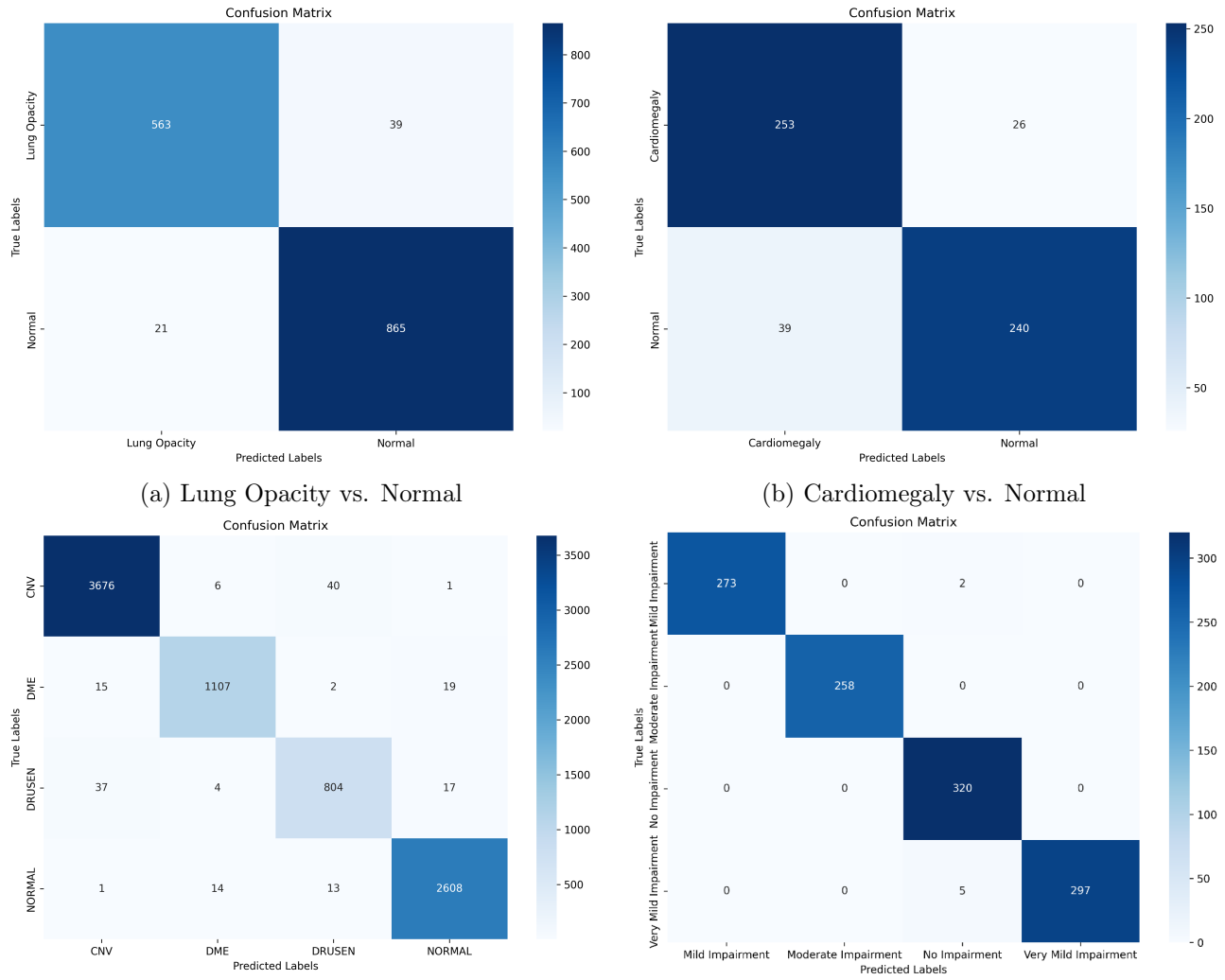
Table 6.2: Hyperparameters that yielded the best model accuracy across all datasets.

Results of Plain Training The trained classifier was evaluated on the test set,using the accuracy, precision ,f1 and recall metrics. Table 6.3 summarizes the results on each dataset.

Dataset	Accuracy	Precision	Recall	F1-score
Lung Opacity vs. Normal	95.96	95.68	97.63	96.65
Cardiomegaly vs. Normal	88.35	90.22	86.02	88.07
OCT Dataset	97.98	97.98	97.97	97.98
Brain MRI Dementia Dataset	99.39%	99.40%	99.39%	99.40%

Table 6.3: Performance metrics (accuracy, precision, recall, and F1-score) for each dataset.

We also plotted the confusion matrices in order to get a better view on the model's performance. Results are shown in 6.2.1



(a) Lung Opacity vs. Normal

(b) Cardiomegaly vs. Normal

(c) OCT Dataset

(d) Brain MRI Dementia Dataset

Figure 6.2.1: Confusion matrices for the plain classifier.

6.2.3 Adversarial Training with TRADES

This section presents the hyperparameters and results obtained from training a classifier using adversarial training with the TRADES method.For all experiments,the trade-off parameter β was set to 6, following recommendations from [112] and [45], as this value provides a well-balanced compromise between adversarial robustness and clean accuracy. Our experimental results further validate this balance.We decided to train our classifiers on three different ϵ values : $\{1/255, 2/255, 8/255\}$. For each ϵ value and for each dataset we present the respective clean accuracy achieved from our model.

Hyperparameters for Adversarial Training The hyperparameters used during TRADES training are presented in Table 6.4.

Hyperparameter	Value
Loss Function	TRADES Adversarial Loss
Initial Learning Rate	1e-4
Epochs	100
Warmup Epochs	10
Optimizer	AdamW
Batch Size	16
Weight Decay	5e-2
Label Smoothing	0.05
β (Trade-off parameter)	6.0
ϵ (Perturbation bound)	1/255, 2/255, 8/255
α (PGD Step Size)	$\epsilon/4$
PGD Steps	10
Norm Type	L_∞

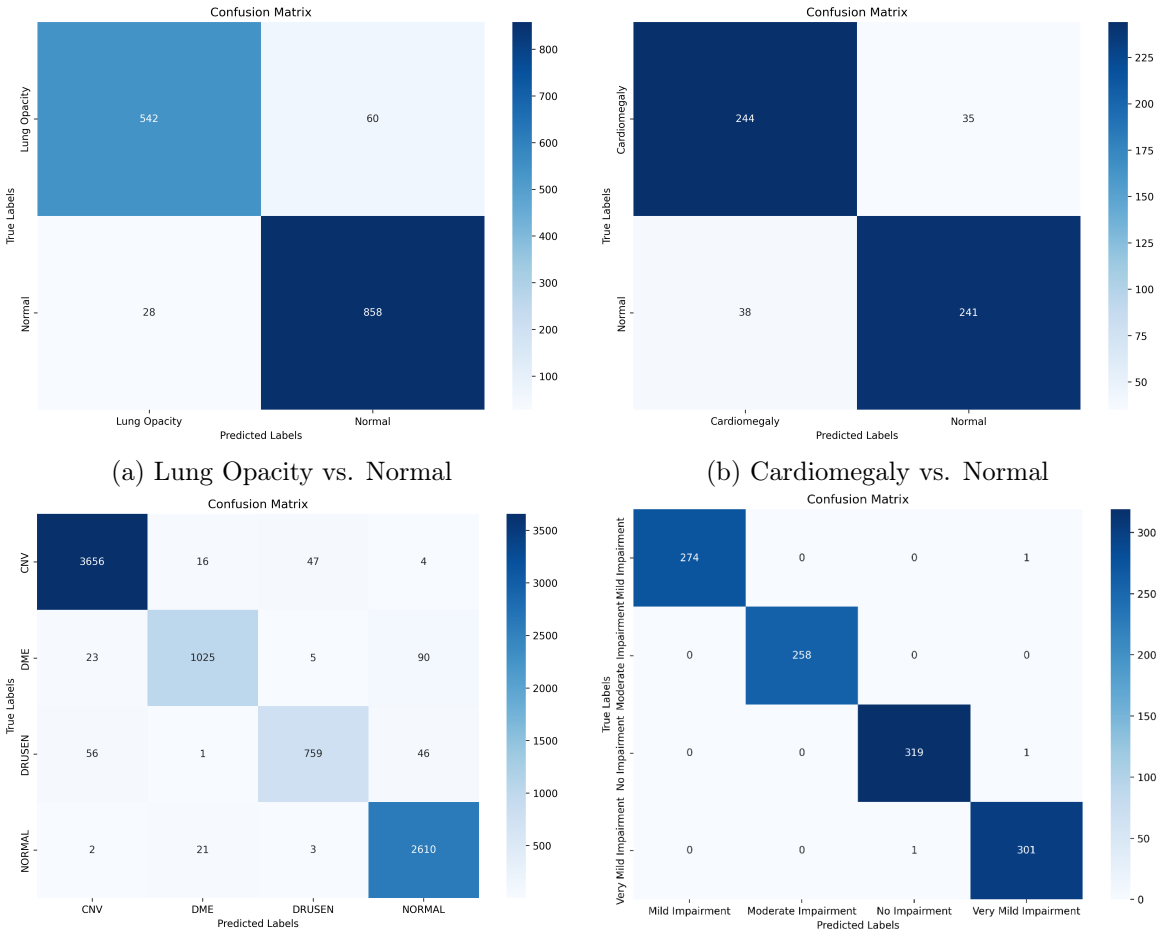
Table 6.4: TRADES hyperparameters used for adversarial training.

Results of Adversarial Training The classifier trained with TRADES was evaluated on both clean and adversarially perturbed test images. Table 6.5 presents the clean accuracy results for all ϵ values .

Dataset	$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 8/255$
Lung Opacity vs. Normal	94.48	94.02	94.09
Cardiomegaly vs. Normal	88.17	85.84	88.35
OCT Dataset	97.44	97.73	96.95
Brain MRI Dementia Dataset	99.74	99.39	99.23

Table 6.5: Accuracy of the adversarially trained classifier on clean test images for $\epsilon = 1/255, 2/255, 8/255$.

Confusion Matrices on Clean Images To analyze the model's performance on clean images, we provide confusion matrices for each dataset, shown in Figure 6.2.2.



(a) Lung Opacity vs. Normal

(b) Cardiomegaly vs. Normal

(c) OCT Dataset

(d) Brain MRI Dementia Dataset

Figure 6.2.2: Confusion matrices for adversarially trained classifier ($\epsilon = 8/255$) evaluated on clean images.

As expected, the model's accuracy on clean images has shown a slight decline compared to the plain classifier. However, this reduction is not substantial and remains within an acceptable range. Furthermore, as demonstrated in the subsequent section, the gradients of the robust model capture meaningful feature representations that align conceptually with the characteristics of each class.

6.3 StyleGAN2-ada training

In this section, we present the methodology and results of our GAN model's training. As discussed in previous sections, we selected StyleGAN2-ADA as the generative model for producing counterfactual medical images.

6.3.1 Training configuration

Datasets used

A significant advantage of our proposed SPRUCE framework is that it does not require coupled training between the classifier whose decisions we aim to interpret and the generative model used to produce explanatory images. This decoupling enables us to train the generative model, specifically StyleGAN2-ADA, on datasets that share the same data distribution as those used for classifier training but are substantially larger and more diverse. Consequently, a single trained generator can be utilized to explain the decisions of multiple classifiers, each trained on different subsets of the original extensive dataset.

We leveraged this advantage specifically for chest X-ray image datasets. Given that the generative quality of GAN models largely depends on the size of the training dataset, we trained our StyleGAN2-ADA model using the entire NIH Chest X-ray dataset, which contains $115,875$ frontal-view X-ray images collected from $32,717$ patients. Each image is annotated with one or multiple pathologies. Another important feature of our approach is that it enables unconditional training of the GAN model, requiring only raw images without associated labels. This unconditional training strategy allowed our model to generate more realistic images, as indicated by improved Fréchet Inception Distance (FID) scores. The resulting StyleGAN2-ADA model, trained on the NIH dataset, was subsequently applied to both classifiers evaluated on the binary chest X-ray datasets presented in Section 6.1.

For the OCT and Brain MRI modalities, we used the same datasets that were employed for classifier training, as they represent the largest publicly available datasets within their respective domains. In total, we trained three separate StyleGAN2-ADA models—one for each medical imaging modality explored in our experiments.

Image Modality	Number of images	Resolution
Chest X-rays	115,875	256×256
Optical Coherence Tomography images	83,605	256×256
Brain MRIs	11,519	128×128

Table 6.6: Datasets used for StyleGAN2-ada training.

Transfer Learning

As emphasized in [130], transfer learning during the training of StyleGAN2-ADA on medical image datasets is a powerful technique that accelerates convergence of the training loss and enhances both the perceptual quality and anatomical fidelity of the generated images. Notably, it has been shown that transfer learning can be effective even when the source domain is unrelated to medical imaging. Based on this insight,during all our experiments, we initialized our training with pretrained weights from a StyleGAN2 model previously trained on the FFHQ dataset.

Augmentations

In Section 5.4.6, we discussed various augmentation strategies employed by the StyleGAN2-ADA model to enhance the realism of generated images, particularly in settings with limited training data. In our experiments, we adopted the Adaptive Discriminator Augmentation (ADA) framework, integrating a broad set of stochastic image augmentations to promote training stability and improve generalization. These augmentations encompassed geometric transformations— horizontal flips, 90-degree rotations, arbitrary angle rotations, translations, anisotropic scaling, and fractional shifts—as well as color-related adjustments, including brightness, contrast, hue, saturation, and luminance inversion. The augmentation intensity was dynamically modulated during training based on discriminator feedback, aiming to maintain a target signal-to-noise ratio by adjusting the augmentation probability toward a predefined threshold (ADA target = 0.6).

Other details

All experiments were conducted in a V100 GPU for approximately 25000 kimg which stands for "thousands of real images shown to the discriminator". As mentioned in [53] in typical cases, 25000 kimg or more is needed to reach convergence, but the results are already quite reasonable around 5000 kimg, especially when using transfer learning.Training on the chest x-rays and oct images took approximately 10 days each , while on the brain mris 6 days were enough for convergence.

6.3.2 Results

During training, we monitored the Fréchet Inception Distance (FID) [103], a widely adopted metric for evaluating the performance of GANs in natural image synthesis. Specifically, we computed the FID between 50,000 images generated by the model under training and real images from the training dataset. The final model selected for use was the one that achieved the lowest FID score, indicating the highest fidelity and realism in the generated outputs.

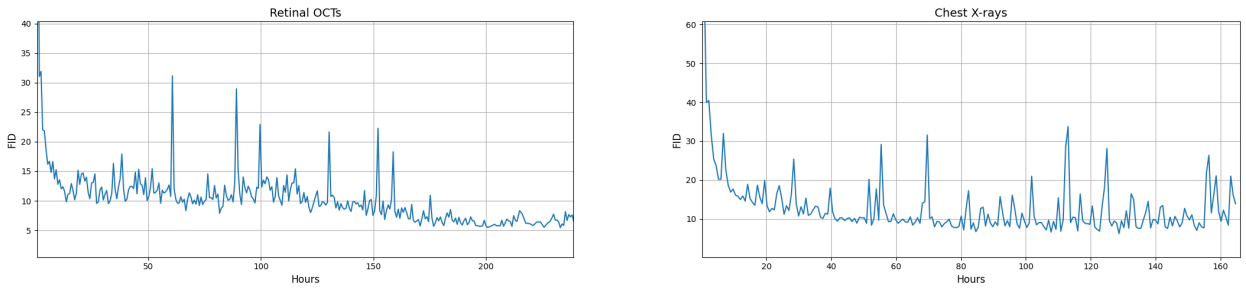


Figure 6.3.1: FID metric during StyleGAN2-ada training.

The models that we later used for the counterfactual generation and image reconstruction achieved an FID of 6.24 for the chest X-rays, 5.46 for the OCT images and 7.63 for the brain MRIs.In the following figure we have generatd some sample images from the final models after training.

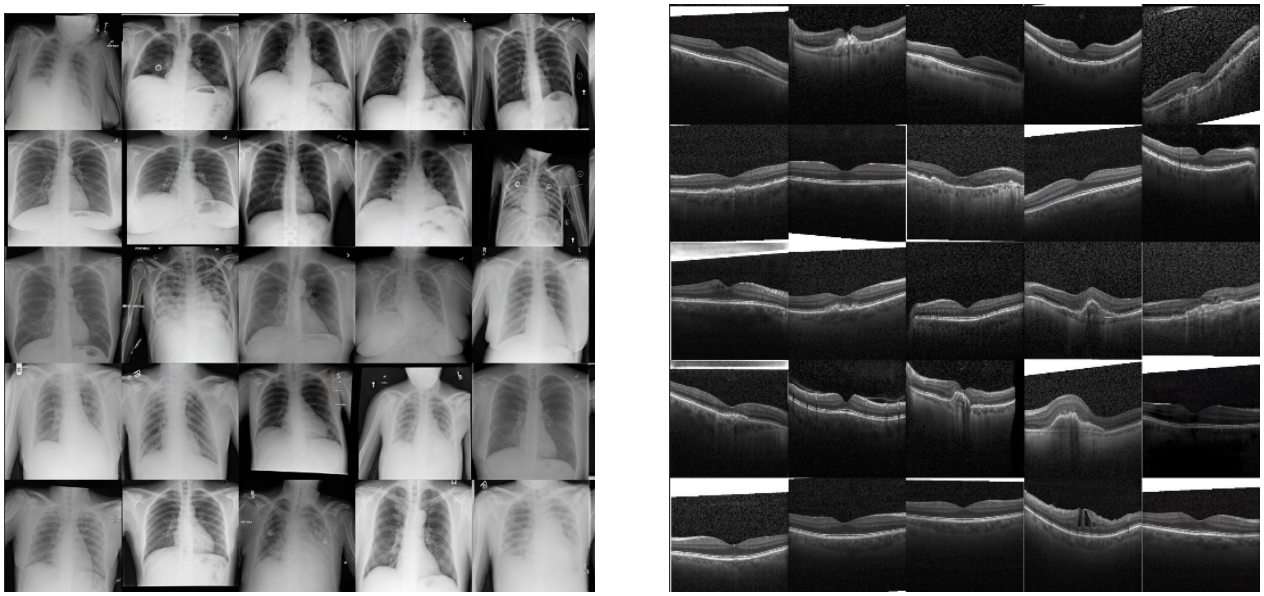


Figure 6.3.2: Chest X-rays and OCT images unconditionally generated from the trained StyleGAN2-ada models.

6.4 GAN inversion

In this section we are going to evaluate the reconstruction results of our gan inversion pipeline on all four datasets using both qualitative and quantitative analysis.

6.4.1 Quantitative Results

For the quantitative evaluation of our results we utilized the following four metrics :

- **Frechet Inception Distance (FID)** The Frechet Inception Distance (FID) [42] quantifies the similarity between two image sets, evaluating realism and diversity. It compares distributions of features extracted from an intermediate layer of the pretrained Inception-v3 network:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \tag{6.4.1}$$

  where μ_r, Σ_r and μ_g, Σ_g denote means and covariances of real and generated images, respectively. Lower FID values indicate better image quality. For the computation of the FID values in our experiments we used the repository [103].

- **Conditional Maximum Mean Discrepancy (CMMD)** Conditional Maximum Mean Discrepancy (CMMD) [46] assesses differences between conditional probability distributions of generated and real images using a Gaussian kernel k:

$$\text{CMMD}^2(P, Q \mid Y) = \|\mathbb{E}[k(X, X') \mid Y] - 2\mathbb{E}[k(X, Z) \mid Y] + \mathbb{E}[k(Z, Z') \mid Y]\| \tag{6.4.2}$$

  where X, X' are from distribution P, and Z, Z' from distribution Q, conditioned on class labels Y. Smaller CMMD indicates greater conditional similarity.

- **Mean Squared Error (MSE)** Mean Squared Error (MSE) measures pixel-wise accuracy between original and reconstructed images:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_i - \hat{I}_i)^2 \tag{6.4.3}$$

  where I_i and $\hat{I}_i$ are pixel intensities of original and inverted images, respectively, and N is the total number of pixels. Lower MSE denotes higher reconstruction fidelity.

- **Learned Perceptual Image Patch Similarity (LPIPS)** Learned Perceptual Image Patch Similarity (LPIPS) evaluates perceptual similarity based on deep network activations:

$$\text{LPIPS}(I, \hat{I}) = \sum_l \frac{1}{H_l W_l}\sum_{h,w}\|w_l \odot (\phi_l(I)_{h,w} - \phi_l(\hat{I})_{h,w})\|_2^2 \tag{6.4.4}$$

  where $\phi_l(I)$ and $\phi_l(\hat{I})$ are activations from layer l of a pretrained network, w_l are learned weights, and H_l, W_l denote spatial dimensions. Lower LPIPS indicates higher perceptual similarity.

In Table 6.7 we have summarized the quantitative results of the reconstructed images from each selected class. From the quantitative results presented, we observe that the GAN inversion performance varies significantly across datasets, influenced primarily by the nature and complexity of each dataset. Specifically, chest X-ray datasets (Cardiomegaly, Lung Opacity) achieve the lowest values for FID, CMMD, MSE, and LPIPS metrics, indicating higher realism, better distributional alignment, and superior pixel-wise and perceptual reconstruction. This superior performance can be attributed to the relatively simpler structural features and lower-frequency content of chest X-rays.

In contrast, the OCT datasets (Drusen, DME, CNV) exhibit higher FID, CMMD, and MSE values, suggesting a greater challenge in accurately reconstructing the intricate textures and detailed retinal structures inherent in OCT images. Similarly, brain MRI images show intermediate performance, highlighting the moderate complexity of structural features in these images.

Class of inverted images	FID ↓	CMMD ↓	MSE ↓	LPIPS ↓
Cardiomegaly (chest X-rays)	19.76	**0.056**	6.28e-4	0.0135
Lung Opacity (chest X-rays)	21.91	0.061	8.57e-4	0.015
OCT - Drusen	26.27	1.031	5.02e-3	**0.00712**
OCT - DME	34.52	0.836	5.56e-3	0.00713
OCT - CNV	33.77	0.855	6.04e-3	0.007562
Moderate Dementia (brain MRIs)	**10.18**	0.58	1.44e-3	0.0086

Table 6.7: Comparison of different inverted-image classes showing FID, CMMD, MSE, and LPIPS.

6.4.2 Qualitative Results

In this section, we visually assess the effectiveness of our proposed GAN inversion pipeline by examining representative examples from our dataset. Each qualitative example is presented across five columns: (1) the original input image; (2) the initial inverted image obtained directly from the encoder, prior to pivotal tuning; (3) a difference heatmap highlighting discrepancies between the original and encoder-only inverted image; (4) the final inversion image obtained after pivotal tuning; and (5) a final difference heatmap illustrating remaining differences post-tuning. In Figures 6.4.1 - 6.4.5 we have collected some examples from all datasets.

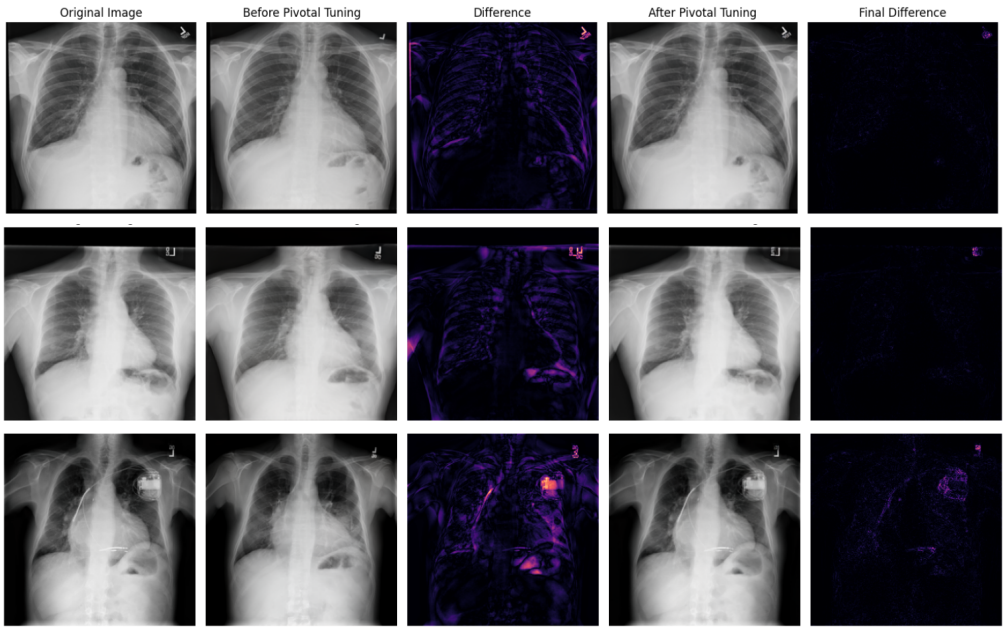


Figure 6.4.1: Qualitative GAN inversion results for cardiomegaly chest X-Ray images.

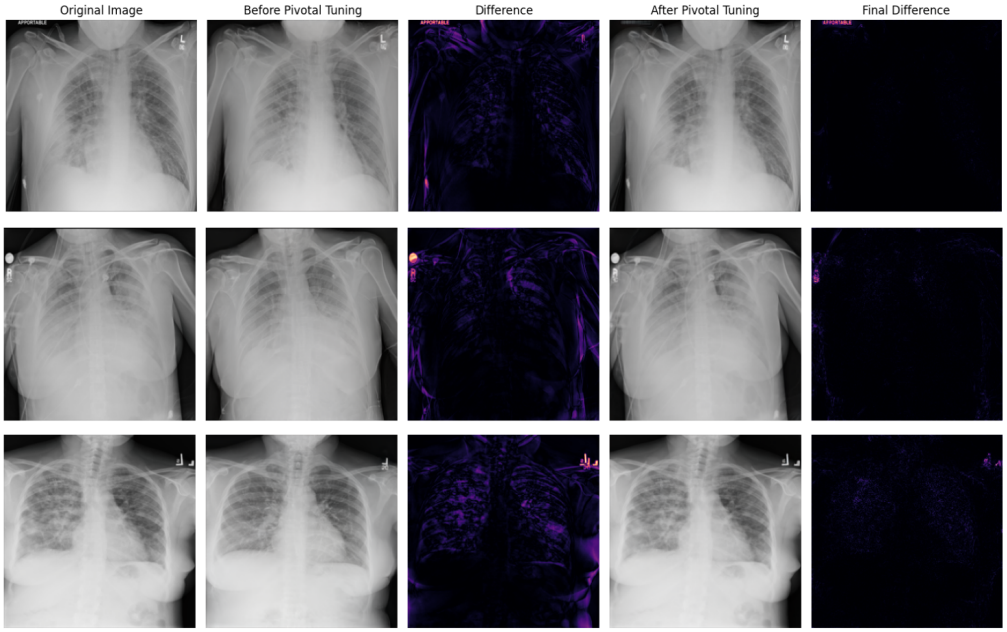


Figure 6.4.2: Qualitative GAN inversion results for lung opacity chest X-Ray images.

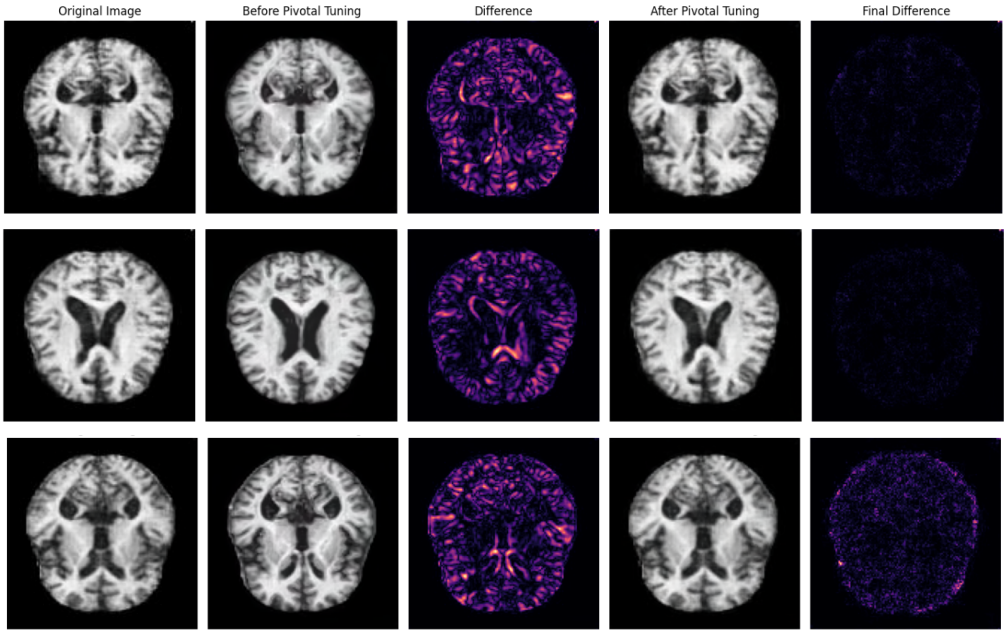


Figure 6.4.3: Qualitative GAN inversion results for brain MRI moderate dementia images.

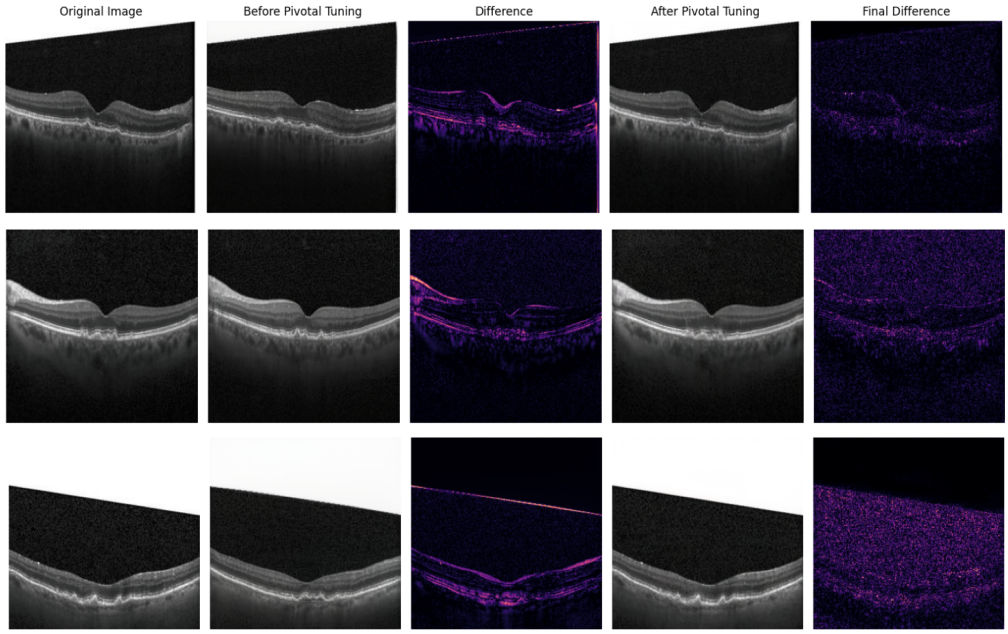


Figure 6.4.4: Qualitative GAN inversion results for OCT images belonging in drusen class.

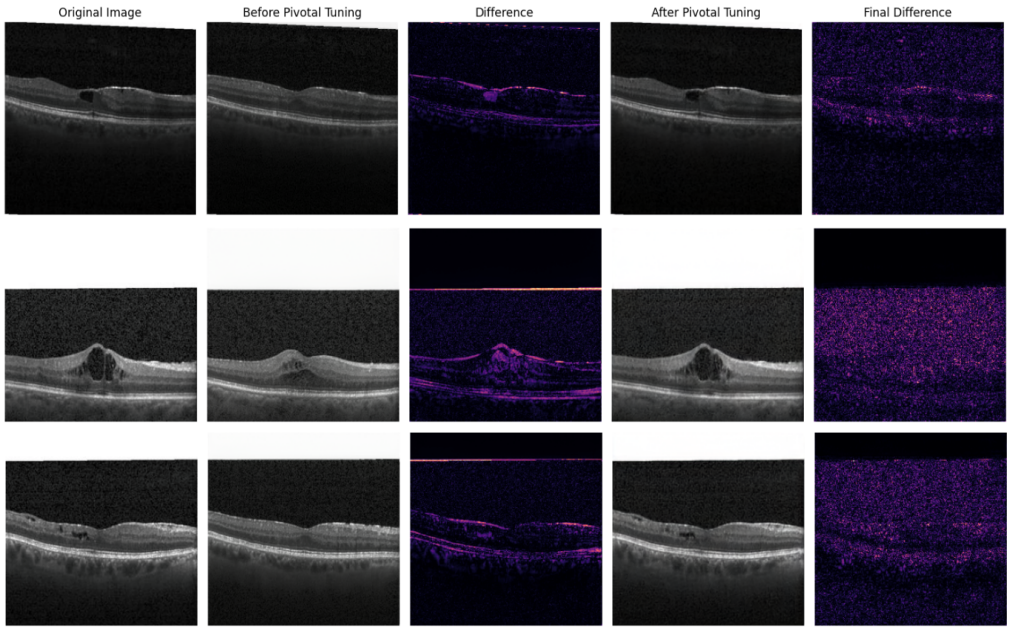


Figure 6.4.5: Qualitative GAN inversion results for OCT images belonging in DME class.

Comparing the difference map of the encoder-based reconstructed images with the originals and the difference map of the final reconstructed images with the originals, we can clearly observe the significant improvement in reconstruction accuracy and perceptual similarity that the Pivotal Tuning process offers to us. This is true accross all four datasets. Especially in chest X-ray images, the inverted images are almost identical to the original ones with the only differences being connected to artifacts that are rarely present in the data distribution like some words and letters outside the body of the X-ray. In brain MRIs the reconstructed images are also identical to the initial ones with minor differences. For OCT images, the qualitative results reveal slightly larger discrepancies in reconstruction compared to chest X-rays and brain MRIs. This is

primarily due to the intricate textures and fine-grained structural details inherent in OCT data, such as distinct retinal layers.Although minor residual differences remain, especially in high-frequency areas, pivotal tuning effectively enhances the realism and clinical relevance of the reconstructed OCT images.

6.5 Counterfactual Generation

In this section, we evaluate the generated counterfactual images both quantitatively and qualitatively. Specifically, we generate counterfactual explanations using two variants of the ConvNeXt-Base model: a standard (plainly trained) classifier and an adversarially trained classifier evaluated at three distinct epsilon (ε) levels: $\{1/255,\ 2/255,\ 8/255\}$. For our evaluation, we generate healthy counterfactual images based on 500 randomly selected and correctly classified examples from four distinct classes: cardiomegaly, lung opacity, drusen, and moderate dementia. For all the experiments we used the following hyperparameter values : $\lambda_{\text{perc}} = 0.1,\ \lambda_{\text{latent}} = 10^{-4},\ \lambda_{\text{image}} = 10^{-6}$.

6.5.1 Quantitative Results

For our quantitative analysis , we used the FID and CMMD metric to asses the realism of the produced counterfactual images. We also used the following three evaluation metrics :

- **Flip Ratio (%):** Measures the percentage of generated counterfactual images that successfully cause a classifier to change its prediction to the desired counterfactual class.Higher flip ratios indicate more effective counterfactual generation, demonstrating that the generated modifications are strongly influential in changing classifier decisions.

- **L1 Distance:** Quantifies sparsity by measuring the pixel-wise differences between the original and counterfactual images.Lower L1 distances represent sparser counterfactual edits, meaning the generated images remain close to the original, altering only the minimal set of pixels necessary to change the classifier's prediction.

- **Classifier Confidence:** Indicates the classifier's predicted probability/confidence for the counterfactual class.Higher classifier confidence suggests that the generated counterfactuals convincingly represent the target class, highlighting meaningful semantic alignment with the classifier's learned decision boundaries.

The results are shown in the following Tables. For $\epsilon = 0$ we refer to the plainly trained classifier.

Epsilon	FID ↓	CMMD ↓	Flip ratio ↑	L1 ↓	Confidence ↑
0	60.22	0.35	**100%**	**0.0118**	0.8601
1/255	56.34	0.315	**100%**	0.0168	0.9866
2/255	58.11	0.326	**100%**	0.0184	**0.9958**
8/255	**49.632**	**0.264**	98%	0.0269	0.9744

Table 6.8: Comparison of different epsilon values and their effects on FID, CMMD, flip ratio, L1, and classifier confidence for the counterfactual class on the lung opacity dataset.

Epsilon	FID ↓	CMMD ↓	Flip ratio ↑	L1 ↓	Confidence ↑
0	43.20	0.192	96.8%	**0.0103**	0.7513
1/255	40.069	0.185	**98.8%**	0.0112	0.7796
2/255	**39.410**	0.169	98.6%	0.0144	0.9221
8/255	40.256	**0.166**	98.6%	0.0188	**0.9569**

Table 6.9: Comparison of different epsilon values and their effects on FID, CMMD, flip ratio, L1, and classifier confidence for the counterfactual class on the cardiomegaly dataset.

Epsilon	FID ↓	CMMD ↓	Flip ratio ↑	L1 ↓	Confidence ↑
0	52.34	1.321	**99.5%**	**0.0251**	0.9810
1/255	**48.54**	1.138	99.2%	0.0295	0.9751
2/255	50.28	**1.125**	98.7%	0.0306	**0.9817**
8/255	49.83	1.132	97.96%	0.0291	0.9523

Table 6.10: Comparison of different epsilon values and their effects on FID, CMMD, flip ratio, L1, and classifier confidence for the counterfactual class on the oct dataset.

Epsilon	FID ↓	CMMD ↓	Flip ratio ↑	L1 ↓	Confidence ↑
0	54.34	0.856	**100%**	**0.0321**	0.9910
1/255	51.79	0.753	**100%**	0.0362	**0.9985**
2/255	**51.59**	**0.744**	**100%**	0.0370	0.9562
8/255	53	0.767	**100%**	0.0368	0.9737

Table 6.11: Comparison of different epsilon values and their effects on FID, CMMD, flip ratio, L1, and classifier confidence for the counterfactual class on the brain mri dataset.

Based on the results from all four datasets (Lung Opacity, Cardiomegaly, OCT, and Brain MRI), we draw the following general observations regarding the influence of the adversarial training parameter ϵ on the generated counterfactuals:

- **Trade-off Between Realism and Classifier Influence:** In several datasets (e.g., Lung Opacity), larger ϵ values (such as 8/255) yield improved realism, evidenced by lower FID and CMMD scores, but at the cost of a slightly reduced flip ratio. This suggests a trade-off, where stronger adversarial robustness introduces more classifier-aligned changes but can reduce the sparsity or subtlety of edits.
- **Sparsity vs. Confidence:** As ϵ increases, the L1 distance tends to rise, indicating less sparse counterfactuals (i.e., more extensive pixel changes). At the same time, classifier confidence in the counterfactual class increases, meaning the generated images align more strongly with the classifier's learned features. This highlights the tension between minimal modification and decision impact.
- **No Universal Best ϵ:** The optimal ϵ value varies by dataset. For instance, in the Cardiomegaly dataset, an intermediate value (2/255) provides a good balance between low FID and high flip ratio. Meanwhile, in the OCT dataset, the best FID is achieved at lower ϵ, while higher confidence appears at 2/255. This illustrates that the ideal ϵ is task- and dataset-dependent, based on the desired balance of realism, sparsity, and classifier influence.

- **Improvements Over $\epsilon = 0$:** In most cases, non-zero ϵ values outperform the $\epsilon = 0$ baseline in terms of either realism or classifier confidence. This supports the notion that adversarial training introduces more informative gradients, which benefit the generation of meaningful counterfactual images.

- **Consistently High Flip Ratios:** Across all datasets and ϵ values, flip ratios remain consistently high (typically above 98%), indicating the effectiveness of the counterfactual generation process in altering the classifier's decision. This further emphasizes the utility of adversarially robust classifiers in guiding semantically meaningful image modifications.

These findings reinforce the importance of adversarially robust classifiers in generating semantically coherent and clinically plausible counterfactuals. However, fine-tuning of ϵ is crucial to achieving a desirable balance between image fidelity, classifier influence, and edit sparsity.

6.5.2 Qualitative Results

Plain vs Robust Classifiers

Initially, we will present some visual results of counterfactuals generated for the same image but for different classifier variants: one trained plainly and one trained on adversarial attacks.

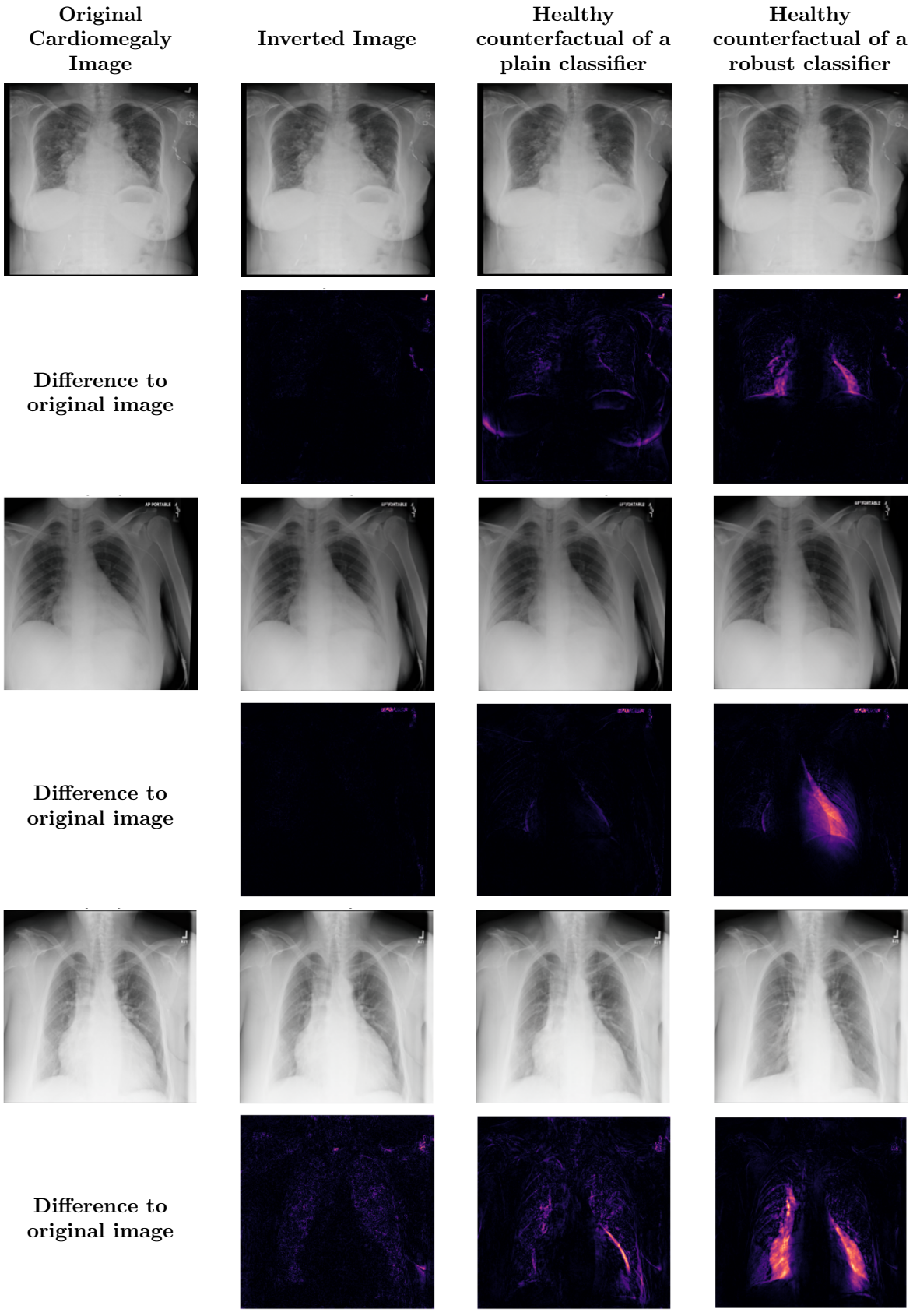


Figure 6.5.1: Comparison of healthy counterfactuals from cardiomegaly cases between plain and robust classifiers.

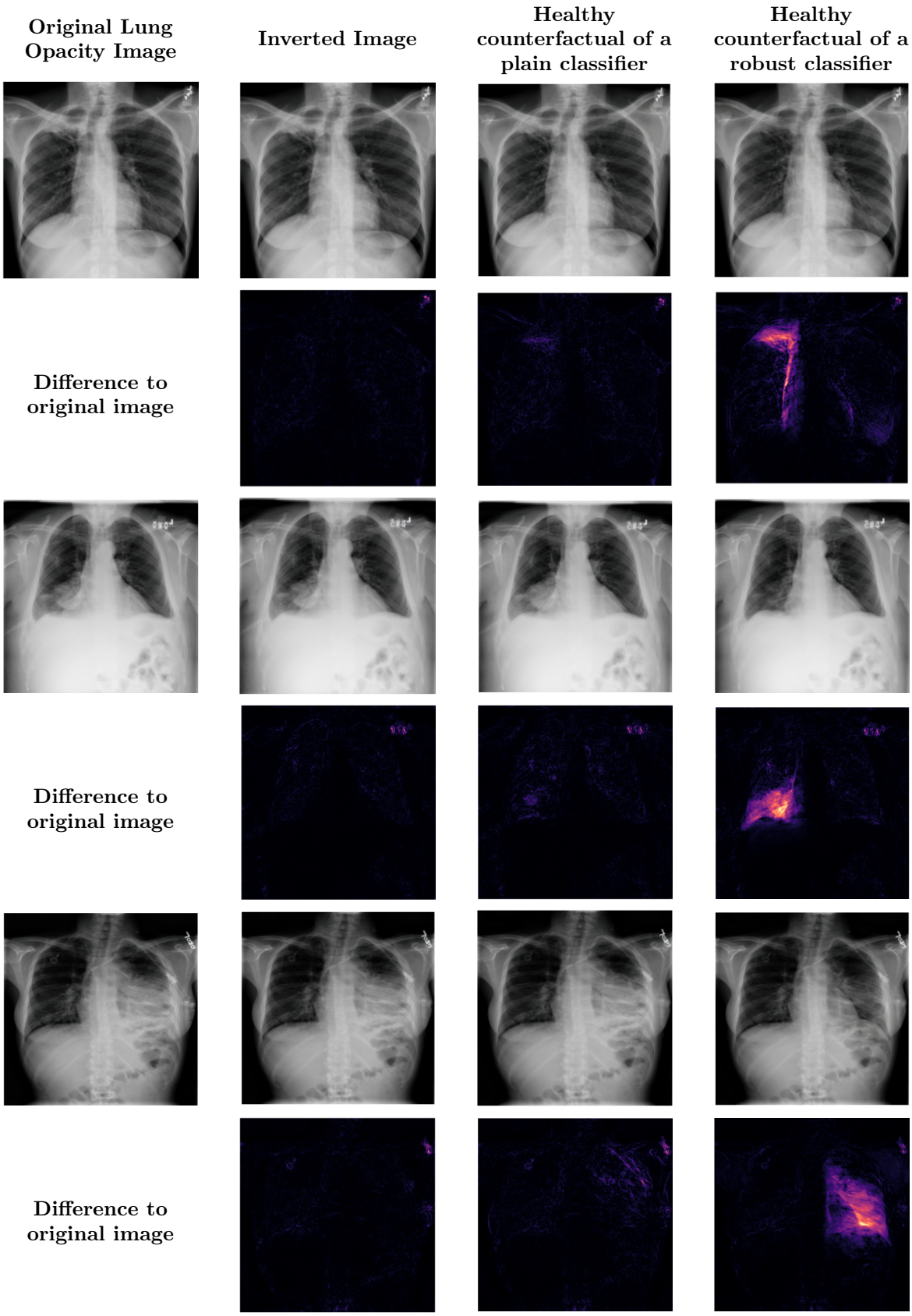


Figure 6.5.2: Comparison of healthy counterfactuals from lung opacity cases between plain and robust classifiers.

Comparing the counterfactual images generated for a plain classifier and those generated for an adversarially robust classifier , we can make the following key observations :

- **Counterfactual Realism:** For both cardiomegaly and lung opacity, the robust classifier generally produces more realistic "healthy" images compared to the plain classifier. The latter occasionally introduces subtle artifacts or over-smoothing.

- **Localization of Changes:** In the difference maps, the robust classifier focuses adjustments around key regions (heart area for cardiomegaly, lung fields for opacity). By contrast, the plain classifier spreads its alterations more diffusely, indicating potentially less targeted corrections.

- **Spurious vs. Meaningful Modifications:** The plain classifier can exhibit spurious changes, sometimes affecting regions not clinically tied to the pathology. The robust classifier, however, typically shows more pathologically relevant edits (e.g., adjusting the cardiac silhouette in cardiomegaly, or clarifying opacities in lung images).

- **Interpretability for Clinicians:** Because the robust classifier better localizes edits to disease-relevant areas, its difference maps tend to be more interpretable. For cardiomegaly, changes concentrate on the enlarged heart region. For lung opacity, shifts occur predominantly within the lung fields.

- **Confidence in Model Explanations:** The more localized and anatomically consistent modifications in robust counterfactuals may enhance clinician trust. This suggests that robust training methods can provide explanations aligning more closely with medical understanding of disease features.

More counterfactuals for robust classifiers.

We will now present a number of healthy counterfactual images produced for robust models from all the datasets that we experimented on.

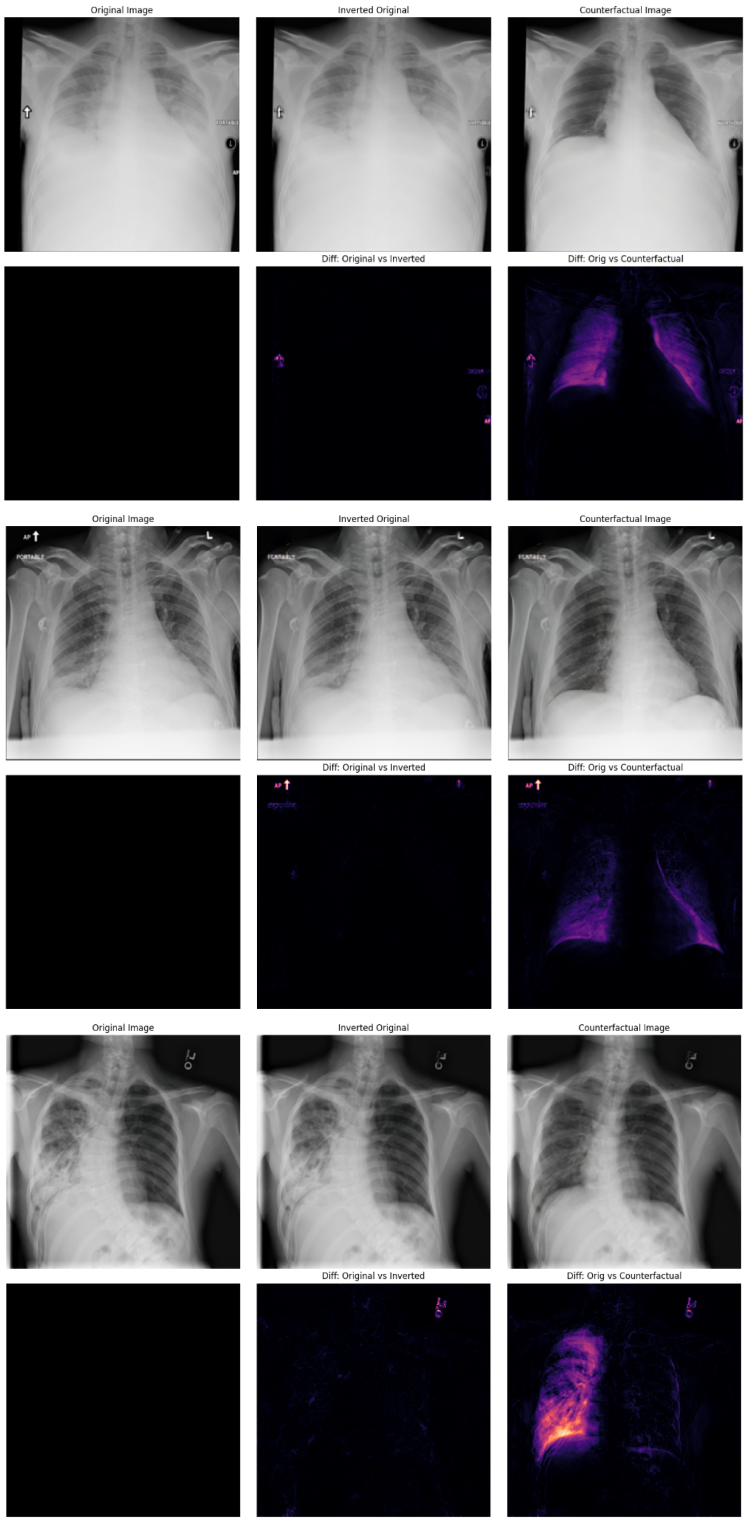


Figure 6.5.3: Healthy counterfactual images from lung opacity cases for robust classifier

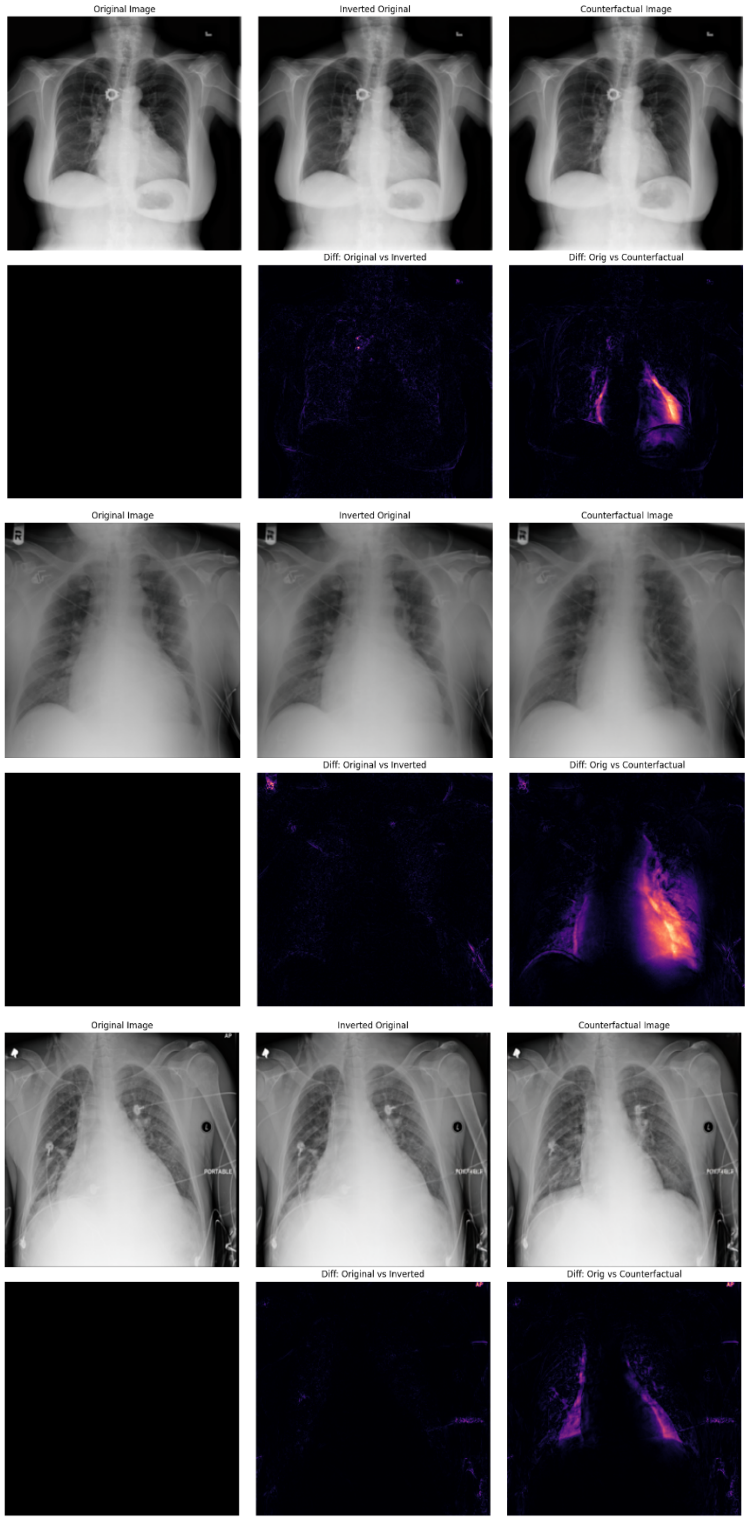


Figure 6.5.4: Healthy counterfactual images from cardiomegaly cases for robust classifier

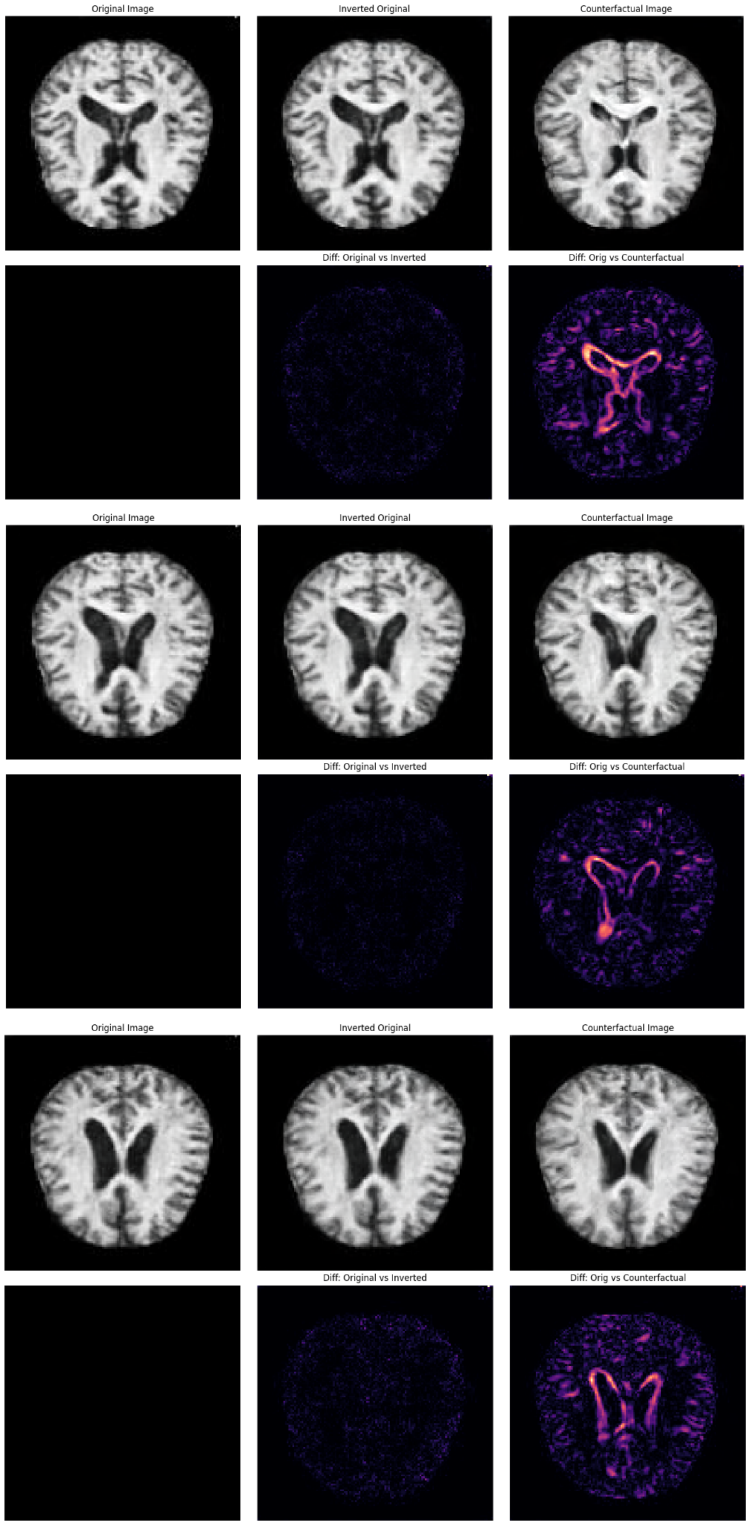


Figure 6.5.5: Healthy counterfactual images from moderate dementia cases for robust classifier

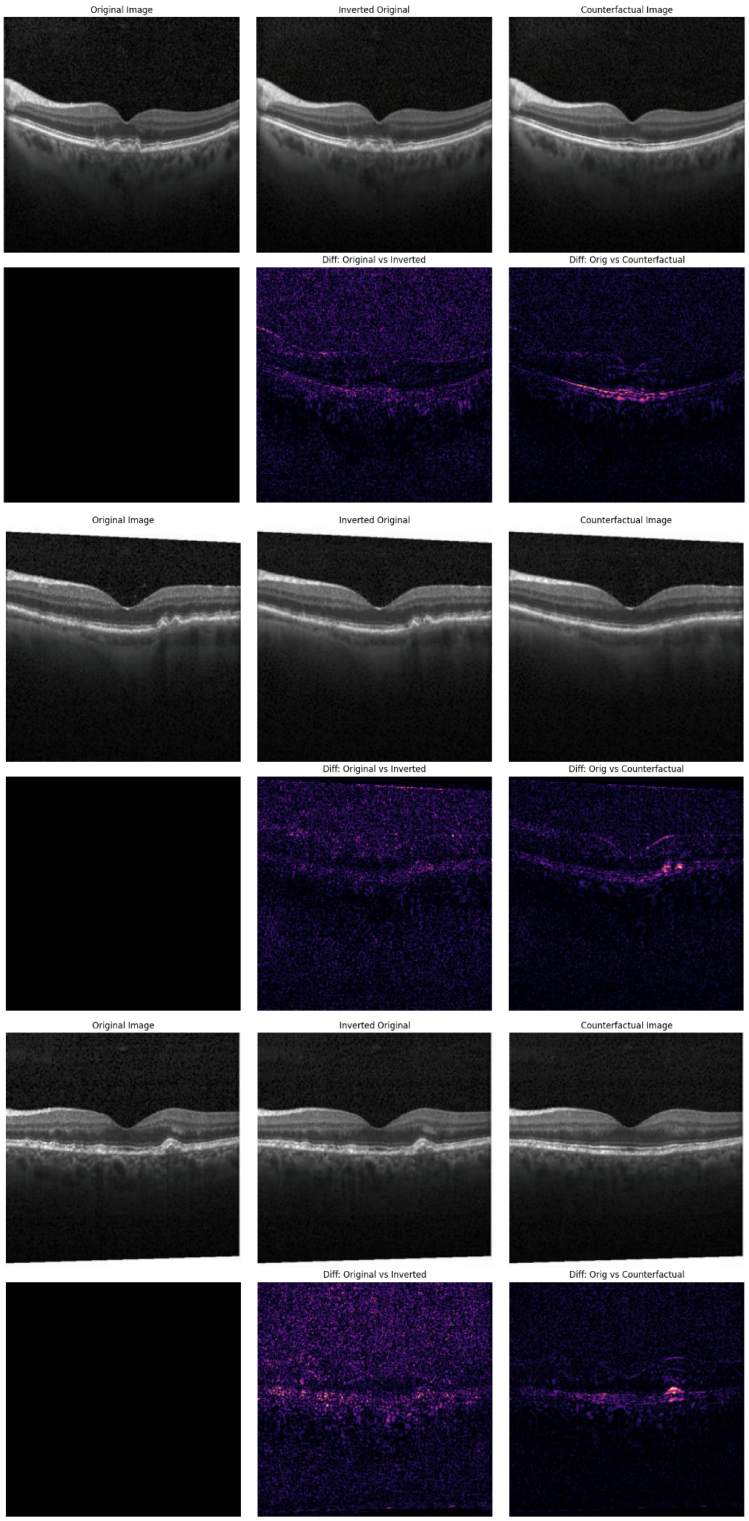


Figure 6.5.6: Healthy counterfactual images from drusen cases for robust classifier

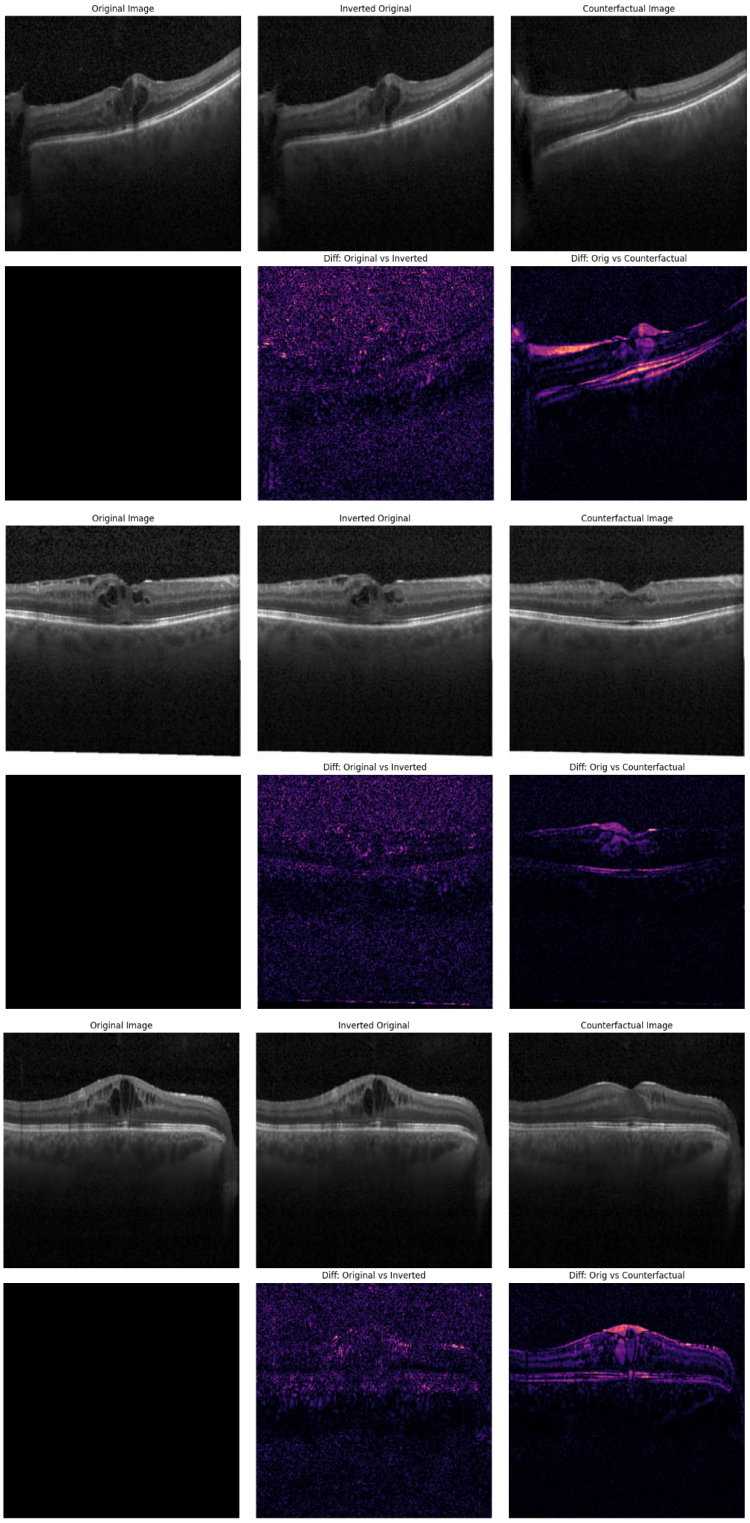


Figure 6.5.7: Healthy counterfactual images from DME cases for robust classifier

By examining the generated counterfactual images alongside their corresponding difference maps, it is evident that the proposed framework produces sparse and visually realistic counterfactuals. Moreover, the highlighted modifications are clinically meaningful and consistently relevant across all four evaluated datasets.

6.6 Impact of loss components

Our intention in this section is to evaluate how each loss term of the optimization function that we use to produce the counterfactual explanations, affects the quality of the generated explanatory image. For that, we chose the adversarially robust model that yielded the most realistic healthy counterfactuals ,for the cardiomegaly and lung opacity images ,based on the CMMD metric and experimented with different hyperparameter values.We start our evaluation from the following values : $\lambda_{\text{perc}} = 0.1$, $\lambda_{\text{latent}} = 10^{-4}$, $\lambda_{\text{image}} = 10^{-6}$. In each experiment we change one of the three hyperparameters and keep the others fixed.

6.6.1 Quantitative results.

For the quantitative analysis in this step we used the same evaluation metrics as in the previous experiments in Section 6.5. The results are summarized in the following tables.

Hyperparameter changed	FID ↓	CMMD ↓	Flip (%) ↑	L1 ↓	Probability ↑
$\lambda_{\text{image}} = 10^{-8}$	**48.59**	0.240	98.8	0.0434	0.9807
$\lambda_{\text{image}} = 10^{-6}$	49.632	0.264	98	0.0269	0.9744
$\lambda_{\text{image}} = 10^{-5}$	56.64	0.333	95	**0.0125**	0.8913
$\lambda_{\text{perc}} = 0.01$	48.66	**0.213**	98.5	0.0294	0.9803
$\lambda_{\text{perc}} = 0.1$	49.632	0.264	98	0.0269	0.9744
$\lambda_{\text{perc}} = 1$	58.12	0.377	95.2	0.0229	0.8991
$\lambda_{\text{latent}} = 10^{-3}$	51.31	0.250	94.4	0.0318	0.8811
$\lambda_{\text{latent}} = 10^{-4}$	49.632	0.264	98	0.0269	0.9744
$\lambda_{\text{latent}} = 10^{-6}$	53.94	0.320	**99.2**	0.0215	**0.9841**

Table 6.12: Comparison of different hyperparameter settings and their effects on FID, CMMD, Flip (%), L1, and classifier probability for the counterfactual class on the lung opacity dataset.

Hyperparameter changed	FID ↓	CMMD ↓	Flip (%) ↑	L1 ↓	Probability (%) ↑
$\lambda_{\text{image}} = 10^{-8}$	**38.37**	0.164	99.6	0.0246	92.60
$\lambda_{\text{image}} = 10^{-6}$	39.41	0.169	98.6	0.0144	92.21
$\lambda_{\text{image}} = 10^{-5}$	42.05	0.186	98.6	**0.0091**	89.48
$\lambda_{\text{perc}} = 0.01$	39.88	**0.151**	**99.8**	0.0162	92.71
$\lambda_{\text{perc}} = 0.1$	39.41	0.169	98.6	0.0144	92.21
$\lambda_{\text{perc}} = 1$	40.78	0.203	98.2	0.0126	88.39
$\lambda_{\text{latent}} = 10^{-3}$	38.61	0.155	98.0	0.0210	86.53
$\lambda_{\text{latent}} = 10^{-4}$	39.41	0.169	98.6	0.0144	92.21
$\lambda_{\text{latent}} = 10^{-6}$	41.76	0.192	**99.8**	0.0121	**93.38**

Table 6.13: Comparison of different hyperparameter settings and their effects on FID, CMMD, Flip (%), L1, and classifier probability for the counterfactual class on the cardiomegaly dataset.

Effects of λ_{image} values

Decreasing λ_{image} (from 10^{-5} to 10^{-8}) consistently results in:

- **Improved Realism:** Significant reductions in FID and CMMD values indicate improved realism and closer alignment with the target class distribution.
- **Higher Flip Ratios and Classifier Confidence:** Lower values of λ_{image} yield higher flip rates and greater classifier probabilities.
- **Increased Image Deviation (Higher L_1):** Looser regularization results in larger pixel-wise deviations, reducing the sparsity of counterfactual modifications.

Effects of λ_{perc} values

Reducing λ_{perc} (from 1 to 0.01) generally leads to:

- **Enhanced Image Realism:** Notable improvements in realism metrics (FID and CMMD) as the perceptual similarity constraint is relaxed.
- **Increased Effectiveness:** Higher flip ratios and classifier probabilities, indicating more convincingly classified counterfactuals.
- **Moderately Increased Image Deviation (Higher L_1):** Lower perceptual constraints allow greater divergence from the original images, reducing sparsity.

Effects of λ_{latent} values

Increasing λ_{latent} (from 10^{-6} to 10^{-3}) typically results in:

- **Improved Realism:** Enhanced realism metrics (lower FID and CMMD), suggesting better latent-space alignment with the target distribution.
- **Reduced Classifier Confidence and Flip Ratio:** Higher latent-space constraints limit image modifications, reducing flip success and lowering classifier confidence.
- **Greater Pixel-Level Deviation (Higher L_1):** Counterintuitively, stricter latent-space regularization sometimes necessitates larger pixel-wise changes to maintain latent consistency.

The analysis of the previous sections confirms the flexibility and effectiveness of our custom loss function during the optimization process. With careful hyperparameter tuning, users can decide whether the resulting counterfactuals are minimally altered or significantly changed, depending on their desired balance of realism, classifier confidence, and similarity to the original image.

6.6.2 Qualitative results

In order to visually assess the effect of each hyperparameter value on the generated counterfactuals we chose a cardiomegaly image and produced the respective counterfactual changing each time a different hyperparameter term.Here are the results :

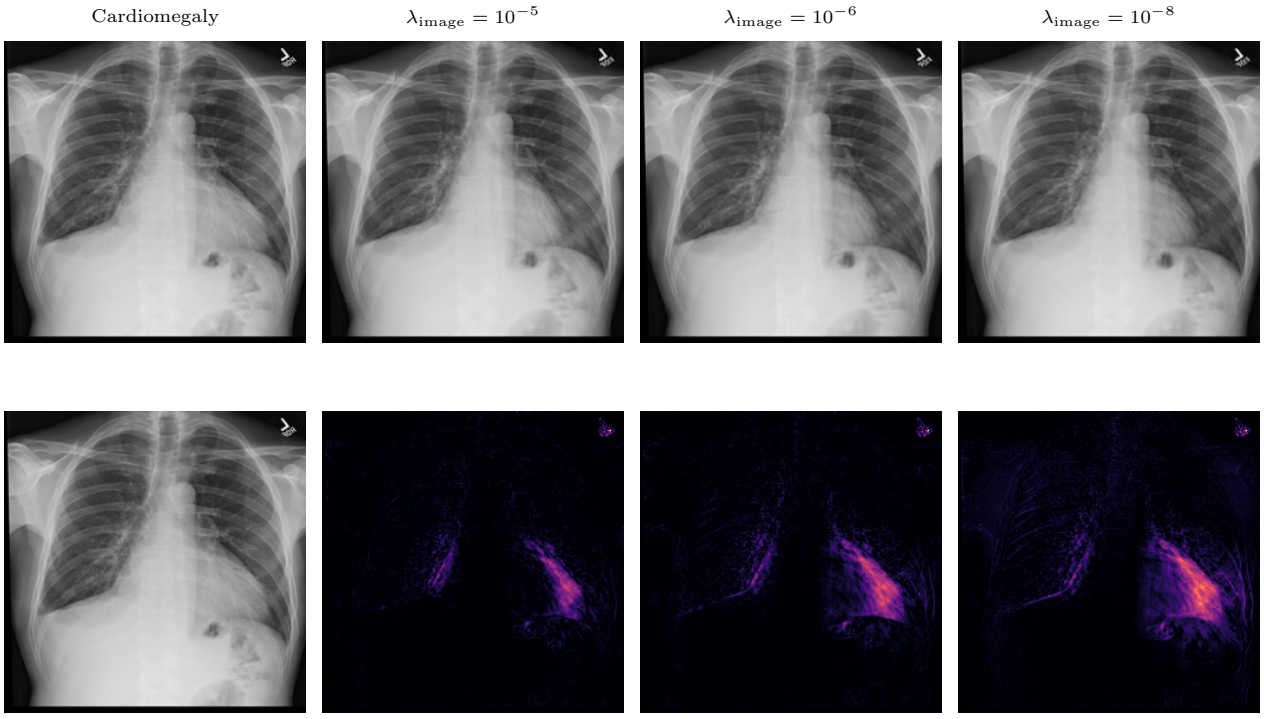


Figure 6.6.1: Counterfactuals for different λ_{image} values

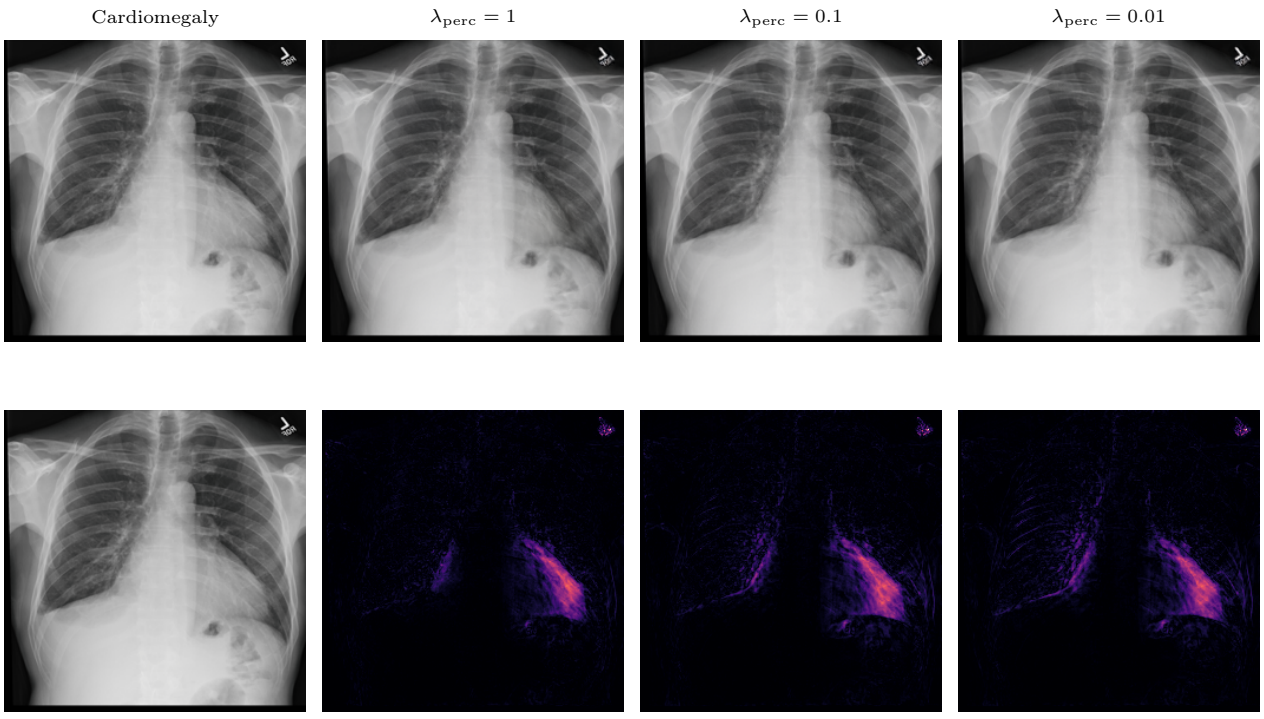


Figure 6.6.2: Counterfactuals for different λ_{perc} values

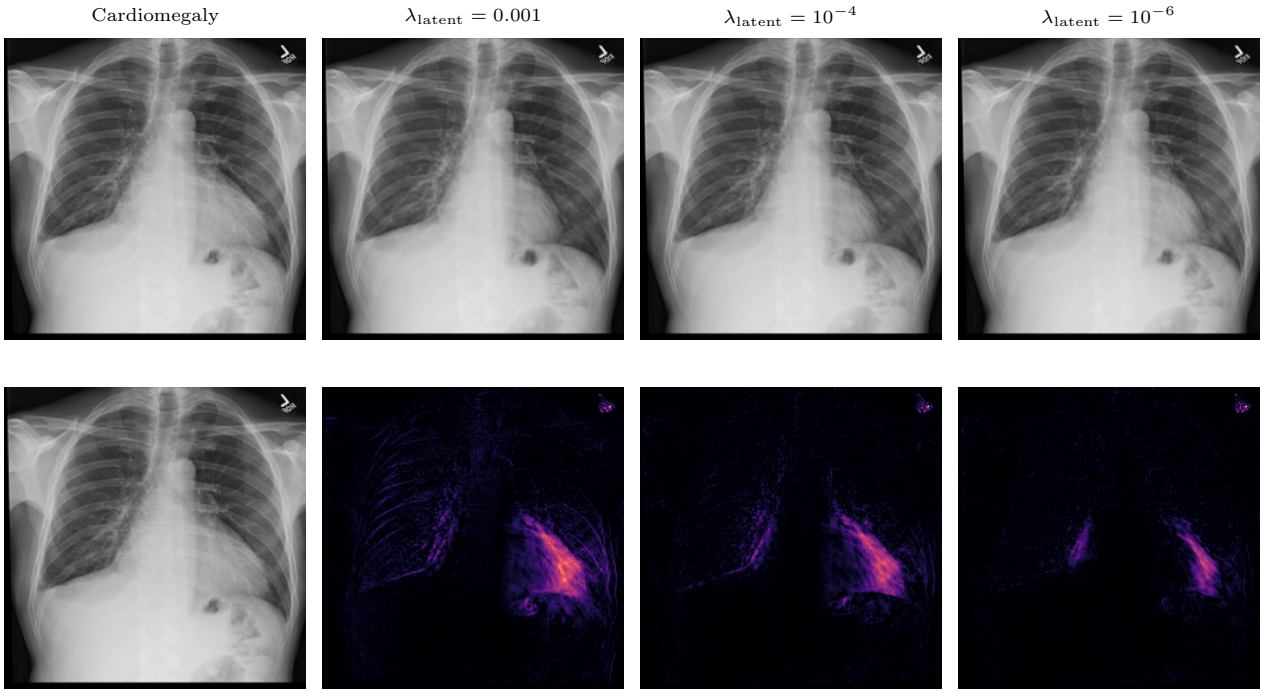


Figure 6.6.3: Counterfactuals for different λ_{latent} values

By looking at the images above we can make the following observations :

- **Influence of $\boldsymbol{\lambda_{\text{image}}}$:** As λ_{image} increases from 10^{-8} to 10^{-5}, the counterfactuals appear more closely aligned with the original image, suggesting that the framework imposes a stronger penalty for pixel-level changes. In the difference maps, the heart region still shows noticeable edits, but the overall alterations become subtler. This indicates that higher λ_{image} values help preserve global structure while still targeting the pathological region.
- **Influence of $\boldsymbol{\lambda_{\text{perc}}}$:** When $\lambda_{\text{perc}} = 0.01$, the counterfactuals tend to alter a broader area around the heart, reflected by more intense regions in the difference map. Increasing λ_{perc} to 0.1 or 1 results in more localized and refined changes, suggesting that emphasizing perceptual features too heavily (λ_{perc} large) can lead to slightly more diffuse modifications, whereas a moderate or smaller λ_{perc} encourages sharper, pathology-focused edits.
- **Influence of $\boldsymbol{\lambda_{\text{latent}}}$:** Moving from $\lambda_{\text{latent}} = 0.001$ to 10^{-6} shows how the strength of latent-space regularization shapes the counterfactual. Smaller λ_{latent} values permit more freedom to change latent representations, often yielding a pronounced difference map around the heart. Larger λ_{latent} (e.g., 10^{-3}) can yield more conservative modifications but may slightly broaden the edited area.
- **Visual Realism vs. Targeted Edits:** Across all hyperparameters, a trade-off is evident: settings that allow more extensive modification (smaller λ values) produce larger—but sometimes noisier—edits, while settings that strongly penalize changes (larger λ values) preserve the original structure but may miss subtle pathological corrections. Finding a balance is crucial for generating both realistic and clinically meaningful counterfactuals.
- **Localizing Cardiomegaly:** In every set of parameters, the most pronounced modifications consistently appear around the enlarged heart region, implying that the model effectively learns to focus on the pathological area. The variance in color intensity and shape of this region reflects how different weights shift the emphasis between realism and effective disentangling of disease-relevant features.

Chapter 7

Conclusion

7.1 Discussion

In this thesis, we presented a novel framework for generating realistic, sparse, and interpretable medical image counterfactuals. Our approach combines a state-of-the-art ConvNeXt classifier (both plainly trained and adversarially robust variants) with a StyleGAN2-ADA generator, enhanced by advanced inversion techniques (E4E and Pivotal Tuning Inversion). This design enables high-fidelity reconstructions of complex medical images while preserving critical anatomical details essential for clinical interpretability.

Key Contributions.

- We established a pipeline for **identifying and modifying class-specific features** (via PIECE) to generate counterfactual images that can highlight the most relevant anatomical regions in pathologies such as cardiomegaly, lung opacity, and retinal abnormalities.
- By employing **adversarial training** (*TRADES*) on the classifier, we demonstrated how robust gradient signals drive more localized and clinically meaningful edits. This contrasts with counterfactuals from plain classifiers, which can introduce spurious or diffuse modifications.
- We **validated the method** quantitatively (using FID, CMMD, Flip ratio, L1 distance, and classifier probability) and qualitatively (via visual inspection and difference maps). Our results highlight that, under the correct hyperparameter settings, counterfactual edits remain sparse, preserving patient-specific traits while addressing the pathology.
- The **GAN inversion scheme** (E4E + PTI) ensured high-fidelity reconstruction for each original image, mitigating artifacts typically observed in medical contexts where data scarcity often hampers generative performance.

Significance and Impact. The proposed framework contributes to the broader field of *Explainable AI (XAI)* in healthcare, addressing the critical need for trustworthy and interpretable solutions in medical imaging. By generating sparse yet anatomically plausible edits, clinicians can better visualize the minimal set of changes that shift a model's prediction, thus increasing the transparency and reliability of black-box classifiers.

Limitations. Despite encouraging results, several challenges remain. GAN-based methods can struggle with rare pathologies or highly heterogeneous datasets where training data is limited. Additionally, while adversarial robustness often improves interpretability, it must be balanced against potential reductions in overall classifier accuracy.

Concluding Remarks. Overall, our experiments underscore that integrating robust classification models with an effective GAN inversion procedure can yield reliable and informative counterfactuals—an important

step toward safer AI deployments in clinical practice. Through comprehensive quantitative and qualitative analyses, we have demonstrated the viability of our approach across multiple datasets and shown how careful hyperparameter tuning can emphasize either sparsity, realism, or classifier alignment.

We envision this framework paving the way for interactive medical workflows where domain experts can rapidly explore how subtle changes in imaging features influence diagnostic predictions, ultimately fostering a deeper understanding of both patient data and model behavior.

7.2 Future Work

Based on the findings and limitations of this study, several directions for future exploration are proposed:

- **User Studies with Medical Specialists:** Conduct user studies with radiologists or domain experts to assess the realism, interpretability, and clinical plausibility of the generated counterfactuals.
- **Advanced Generative Models:** Investigate alternative generative approaches (e.g., diffusion models or improved GAN architectures) that could outperform StyleGAN in capturing intricate medical details.
- **Cross-Architecture Benchmarking:** Evaluate how the proposed counterfactual method generalizes across various classifier architectures (e.g., DenseNet, ResNet, ViT), ensuring robustness and consistency.
- **3D Medical Modalities:** Extend the framework to three-dimensional data (e.g., CT or MRI scans) to handle volumetric representations, increasing its applicability in a broader range of medical imaging scenarios.
- **Comparative Interpretability Studies:** Compare the proposed counterfactual-based saliency with alternative interpretability methods such as Grad-CAM, Integrated Gradients, or other popular techniques for enhanced validation and context.

Chapter 8

Bibliography

[1] Abdal, R., Qin, Y., and Wonka, P. *Image2StyleGAN++: How to Edit the Embedded Images?* 2020. arXiv: 1911.11544 [cs.CV]. URL:

[2] Anouk Stein, M. et al. *RSNA Pneumonia Detection Challenge*. Kaggle. 2018.

[3] Arjovsky, M., Chintala, S., and Bottou, L. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[4] Arrieta, A. B. and Ser, J. D. "Plausible Counterfactuals: Auditing Deep Learning Classifiers with Realistic Adversarial Examples". In: *2020 International Joint Conference on Neural Networks (IJCNN)* (2020), pp. 1–7. URL:

[5] Atad, M. et al. "Counterfactual Explanations for Medical Image Classification and Regression using Diffusion Autoencoder". In: *ArXiv* abs/2408.01571 (2024). URL:

[6] Athalye, A., Carlini, N., and Wagner, D. *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples*. 2018. arXiv: 1802.00420 [cs.LG]. URL:

[7] Augustin, M. et al. *Diffusion Visual Counterfactual Explanations*. 2022. arXiv: 2210.11841 [cs.CV]. URL:

[8] Azizi, S. et al. *Big Self-Supervised Models Advance Medical Image Classification*. 2021. arXiv: 2101.05224 [eess.IV]. URL:

[9] Bermano, A. H. et al. *State-of-the-Art in the Architecture, Methods and Applications of StyleGAN*. 2022. arXiv: 2202.14020 [cs.CV]. URL:

[10] Biswas, A. and Islam, M. S. "An Efficient CNN Model for Automated Digital Handwritten Digit Classification". In: *Journal of Information Systems Engineering and Business Intelligence* (2021). URL:

[11] Boreĭko, V. B. et al. "Visual explanations for the detection of diabetic retinopathy from retinal fundus images". In: *medRxiv*. 2022. URL:

[12] Borji, A. "Pros and cons of GAN evaluation measures". In: *Computer Vision and Image Understanding* 179 (2019), pp. 41–65.

[13] Chen, J. et al. "Meta-Causal Learning for Single Domain Generalization". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 7683–7692. URL:

[14] Chen, J. et al. "Explaining the Black-box Smoothly- A Counterfactual Approach". In: *Medical image analysis* 84 (2021), p. 102721. URL:

[15] Chen, K. et al. "A Survey on Adversarial Examples in Deep Learning". In: *Journal on Big Data* (2020). URL:

[16] Cohen, G. et al. "EMNIST: Extending MNIST to handwritten letters". In: *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), pp. 2921–2926. URL:

[17] Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. *Certified Adversarial Robustness via Randomized Smoothing*. 2019. arXiv: 1902.02918 [cs.LG]. URL:

[18] Cohen, J. P. et al. "Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays". In: *International Conference on Medical Imaging with Deep Learning*. 2021. URL:

[19] Creswell, A. et al. "Generative adversarial networks: An overview". In: *IEEE signal processing magazine* 35.1 (2018), pp. 53–65.

[20] Deng, J. et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 248–255.

[21] Deng, L. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[22] Dervakos, E. et al. *Choose your Data Wisely: A Framework for Semantic Counterfactuals*. 2023. arXiv: 2305.17667 [cs.AI]. URL:

[23] Dhurandhar, A. et al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: *Neural Information Processing Systems*. 2018. URL:

[24] Dimitriou, A. et al. *Structure Your Data: Towards Semantic Graph Counterfactuals*. 2024. arXiv: 2403.06514 [cs.CV]. URL:

[25] Dosovitskiy, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL:

[26] Fang, Y. et al. "DiffExplainer: Unveiling Black Box Models Via Counterfactual Generation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 208–218.

[27] Filandrianos, G. et al. "Conceptual Edits as Counterfactual Explanations." In: 2022.

[28] Filandrianos, G. et al. *Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors*. 2023. arXiv: 2305.17055 [cs.CL]. URL:

[29] Gilmer, J. et al. *Motivating the Rules of the Game for Adversarial Example Research*. 2018. arXiv: 1807.06732 [cs.LG]. URL:

[30] Glorot, X., Bordes, A., and Bengio, Y. "Deep Sparse Rectifier Neural Networks". In: *International Conference on Artificial Intelligence and Statistics*. 2011. URL:

[31] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

[32] Goodfellow, I. et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[33] Goodfellow, I. J., Shlens, J., and Szegedy, C. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML]. URL:

[34] Goyal, Y. et al. "Counterfactual Visual Explanations". In: *ArXiv* abs/1904.07451 (2019). URL:

[35] Gu, Y. et al. "BiomedJourney: Counterfactual Biomedical Image Generation by Instruction-Learning from Multimodal Patient Journeys". In: *ArXiv* abs/2310.10765 (2023). URL:

[36] Guidotti, R. "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Min. Knowl. Discov.* 38 (2024), pp. 2770–2824. URL:

[37] Guidotti, R. et al. "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys (CSUR)* 51 (2018), pp. 1–42. URL:

[38] Guo, C. et al. *Countering Adversarial Images using Input Transformations*. 2018. arXiv: 1711.00117 [cs.CV]. URL:

[39] He, K. et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[40] He, Z. et al. "AttGAN: Facial Attribute Editing by Only Changing What You Want". In: *IEEE Transactions on Image Processing* 28 (2017), pp. 5464–5478. URL:

[41] Hendrycks, D. and Gimpel, K. *Gaussian Error Linear Units (GELUs)*. 2023. arXiv: 1606.08415 [cs.LG]. URL:

[42] Heusel, M. et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. 2018. arXiv: 1706.08500 [cs.LG]. URL:

[43] Huang, G. et al. *Densely Connected Convolutional Networks*. 2016. arXiv: 1608.06993 [cs.CV].

[44] Hui, A. et al. "Ethical Challenges of Artificial Intelligence in Health Care: A Narrative Review". In: *Ethics in Biology, Engineering and Medicine: An International Journal* (2022). URL:

[45] Ilanchezian, I. et al. *Generating Realistic Counterfactuals for Retinal Fundus and OCT Images using Diffusion Models*. 2023. arXiv: 2311.11629 [cs.CV]. URL:

[46] Jayasumana, S. et al. "Rethinking FID: Towards a Better Evaluation Metric for Image Generation". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 9307–9315. URL:

[47] Jia, Y. et al. "Counterfactual Causal Adversarial Networks for Domain Adaptation". In: *International Conference on Neural Information Processing*. 2022. URL:

[48] Johnson, J., Alahi, A., and Fei-Fei, L. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*. 2016. arXiv: 1603.08155 [cs.CV]. URL:
[49] Karras, T., Laine, S., and Aila, T. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2018. arXiv: 1812.04948 [cs.NE].
[50] Karras, T. et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. 2017. arXiv: 1710.10196 [cs.NE].
[51] Karras, T. et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. 2018. arXiv: 1710.10196 [cs.NE]. URL:
[52] Karras, T. et al. *Analyzing and Improving the Image Quality of StyleGAN*. 2019. arXiv: 1912.04958 [cs.CV].
[53] Karras, T. et al. *Training Generative Adversarial Networks with Limited Data*. 2020. arXiv: 2006.06676 [cs.CV]. URL:
[54] Kenny, E. M. and Keane, M. T. *On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning*. 2020. arXiv: 2009.06399 [cs.LG]. URL:
[55] Kermany, D. S., Zhang, K., and Goldbaum, M. H. "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification". In: 2018. URL:
[56] Khanal, B. et al. *Improving Medical Image Classification in Noisy Labels Using Only Self-supervised Pretraining*. 2023. arXiv: 2308.04551 [eess.IV]. URL:
[57] Khosla, A. et al. "Novel Dataset for Fine-Grained Image Categorization : Stanford Dogs". In: 2012. URL:
[58] Koulakos, A. et al. "Enhancing adversarial robustness in Natural Language Inference using explanations". In: *arXiv preprint arXiv:2409.07423* (2024).
[59] Lang, O. et al. "Explaining in Style: Training a GAN to explain a classifier in StyleSpace". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 673–682. URL:
[60] LeCun, Y. et al. "Gradient-based learning applied to document recognition". In: *Proc. IEEE* 86 (1998), pp. 2278–2324. URL:
[61] Lenis, D. et al. "Domain aware medical image classifier interpretation by counterfactual impact analysis". In: *ArXiv* abs/2007.06312 (2020). URL:
[62] Liartis, J. et al. "Semantic Queries Explaining Opaque Machine Learning Classifiers." In: *DAO-XAI*. 2021.
[63] Liartis, J. et al. "Searching for explanations of black-box classifiers in the space of semantic queries". In: *Semantic Web* 15.4 (2024), pp. 1085–1126.
[64] Lipton, Z. C. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG]. URL:
[65] Liu, Z. et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV].
[66] López, P. R. et al. "Beyond Trivial Counterfactual Explanations with Diverse Valuable Explanations". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 1036–1045. URL:
[67] Loshchilov, I. and Hutter, F. *SGDR: Stochastic Gradient Descent with Warm Restarts*. 2017. arXiv: 1608.03983 [cs.LG]. URL:
[68] Loshchilov, I. and Hutter, F. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG]. URL:
[69] Lundberg, S. and Lee, S.-I. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI]. URL:
[70] Lyberatos, V. et al. "Perceptual musical features for interpretable audio tagging". In: *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE. 2024, pp. 878–882.
[71] Lymperaiou, M. et al. *Towards explainable evaluation of language models on the semantic similarity of visual concepts*. 2022. arXiv: 2209.03723 [cs.CL]. URL:
[72] Lymperaiou, M. et al. *Counterfactual Edits for Generative Evaluation*. 2023. arXiv: 2303.01555 [cs.CV]. URL:
[73] Madry, A. et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706.06083 [stat.ML]. URL:

[74] Manzari, O. N. et al. "MedViT: A robust vision transformer for generalized medical image classification". In: *Computers in Biology and Medicine* 157 (May 2023), p. 106791. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2023.106791. URL:

[75] Mastromichalakis, O. M., Liartis, J., and Stamou, G. "Beyond One-Size-Fits-All: Adapting Counterfactual Explanations to User Objectives". In: *arXiv preprint arXiv:2404.08721* (2024).

[76] Mastromichalakis, O. M. et al. "GOSt-MT: A Knowledge Graph for Occupation-related Gender Biases in Machine Translation". In: *arXiv preprint arXiv:2409.10989* (2024).

[77] Mastromichalakis, O. M. et al. "Rule-Based Explanations of Machine Learning Classifiers Using Knowledge Graphs". In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1. 2024, pp. 193–202.

[78] Menis Mastromichalakis, O. et al. "Semantic Prototypes: Enhancing Transparency Without Black Boxes". In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024, pp. 1680–1688.

[79] Menon, S. et al. *PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models*. 2020. arXiv: 2003.03808 [cs.CV]. URL:

[80] Mertes, S. et al. "GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning". In: *Frontiers in Artificial Intelligence* 5 (2020). URL:

[81] Micikevicius, P. et al. *Mixed Precision Training*. 2018. arXiv: 1710.03740 [cs.AI]. URL:

[82] Mo, S., Cho, M., and Shin, J. *Freeze the Discriminator: a Simple Baseline for Fine-Tuning GANs*. 2020. arXiv: 2002.10964 [cs.CV]. URL:

[83] Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL:

[84] Mothilal, R. K., Sharma, A., and Tan, C. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2019). URL:

[85] Noguchi, A. and Harada, T. *Image Generation From Small Datasets via Batch Statistics Adaptation*. 2019. arXiv: 1904.01774 [cs.CV]. URL:

[86] Oh, K., Yoon, J. S., and Suk, H.-I. "Learn-Explain-Reinforce: Counterfactual Reasoning and its Guidance to Reinforce an Alzheimer's Disease Diagnosis Model". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2023), pp. 4843–4857. DOI: 10.1109/TPAMI.2022.3197845.

[87] Pantelaios, D. et al. "Hybrid CNN-ViT models for medical image classification". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–4.

[88] Papernot, N. and McDaniel, P. *On the Effectiveness of Defensive Distillation*. 2016. arXiv: 1607.05113 [cs.CR]. URL:

[89] Papernot, N. et al. *Practical Black-Box Attacks against Machine Learning*. 2017. arXiv: 1602.02697 [cs.CR]. URL:

[90] Porwal, P. et al. "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research". In: *Data* 3 (2018), p. 25. URL:

[91] Poyiadzi, R. et al. "FACE: Feasible and Actionable Counterfactual Explanations". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2019). URL:

[92] Puttagunta, M. K. and Subban, R. "Medical image analysis based on deep learning approach". In: *Multimedia Tools and Applications* 80 (2021), pp. 24365–24398. URL:

[93] Radford, A., Metz, L., and Chintala, S. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[94] Ribeiro, M. T., Singh, S., and Guestrin, C. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG]. URL:

[95] Richardson, E. et al. *Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation*. 2021. arXiv: 2008.00951 [cs.CV]. URL:

[96] Roich, D. et al. *Pivotal Tuning for Latent-based Editing of Real Images*. 2021. arXiv: 2106.05744 [cs.CV]. URL:

[97] Saad, M. M., O'Reilly, R., and Rehmani, M. H. "A survey on training challenges in generative adversarial networks for biomedical image analysis". In: *Artificial Intelligence Review* 57 (Jan. 2024). DOI: 10.1007/s10462-023-10624-y.

[98] Samek, W. and Müller, K.-R. "Towards Explainable Artificial Intelligence". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 2019, pp. 5–22. ISBN: 9783030289546. DOI: 10.1007/978-3-030-28954-6_1. URL:

[99] Sanchez, P. and Tsaftaris, S. A. "Diffusion Causal Models for Counterfactual Estimation". In: *CLEaR*. 2022. URL:
[100] Santurkar, S. et al. *Image Synthesis with a Single (Robust) Classifier*. 2019. arXiv: 1906.09453 [cs.CV]. URL:
[101] Sapountzakis, G., Theofilou, P.-A., and Tzouveli, P. "Covid-19 Detection From X-Rays Images Using Deep Learning Methods". In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE. 2023, pp. 1–5.
[102] Saucedo, J. A. M. and Kose, U. "Numerical Grad-Cam Based Explainable Convolutional Neural Network for Brain Tumor Diagnosis". In: *Mob. Networks Appl.* 29 (2022), pp. 109–118. URL:
[103] Seitzer, M. *pytorch-fid: FID Score for PyTorch*. Version 0.3.0. Aug. 2020.
[104] Selvaraju, R. R. et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL:
[105] Shvetsov, D. et al. "COIN: Counterfactual inpainting for weakly supervised semantic segmentation for medical images". In: *World Conference on Explainable Artificial Intelligence*. Springer. 2024, pp. 39–59.
[106] Simonyan, K., Vedaldi, A., and Zisserman, A. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. arXiv: 1312.6034 [cs.CV]. URL:
[107] Simonyan, K. and Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV]. URL:
[108] Skandarani, Y., Jodoin, P.-M., and Lalande, A. *GANs for Medical Image Synthesis: An Empirical Study*. 2021. arXiv: 2105.05318 [eess.IV]. URL:
[109] Sotirou, T. et al. "Musiclime: Explainable multimodal music understanding". In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025, pp. 1–5.
[110] Spanos, N. et al. "Complex Style Image Transformations for Domain Generalization in Medical Images". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2024), pp. 5036–5045. URL:
[111] Srivastava, N. et al. "Dropout: A simple way to prevent neural networks from overfitting". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.
[112] Stutz, D., Hein, M., and Schiele, B. *Relating Adversarially Robust Generalization to Flat Minima*. 2021. arXiv: 2104.04448 [cs.LG]. URL:
[113] Szegedy, C. et al. *Intriguing properties of neural networks*. 2014. arXiv: 1312.6199 [cs.CV]. URL:
[114] Tariq, U. et al. "Brain Tumor Synthetic Data Generation with Adaptive StyleGANs". In: *Artificial Intelligence and Cognitive Science*. Springer Nature Switzerland, 2023, pp. 147–159. ISBN: 9783031264382. DOI: 10.1007/978-3-031-26438-2_12. URL:
[115] Thiagarajan, J. J., Venkatesh, B., and Rajan, D. "Learn-By-Calibrating: Using Calibration As A Training Objective". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 3632–3636. URL:
[116] Thiagarajan, J. J. et al. "Training calibration-based counterfactual explainers for deep learning models in medical image analysis". In: *Scientific Reports* 12 (2021). URL:
[117] Tov, O. et al. *Designing an Encoder for StyleGAN Image Manipulation*. 2021. arXiv: 2102.02766 [cs.CV]. URL:
[118] Vandenhende, S. et al. "Making Heads or Tails: Towards Semantically Consistent Visual Counterfactuals". In: *ArXiv* abs/2203.12892 (2022). URL:
[119] Velden, B. H. M. van der et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical image analysis* 79 (2021), p. 102470. URL:
[120] Verma, S., Dickerson, J. P., and Hines, K. E. "Counterfactual Explanations for Machine Learning: A Review". In: *ArXiv* abs/2010.10596 (2020). URL:
[121] Vermeire, T. and Martens, D. "Explainable Image Classification with Evidence Counterfactual". In: *ArXiv* abs/2004.07511 (2020). URL:
[122] Wachter, S., Mittelstadt, B., and Russell, C. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018. arXiv: 1711.00399 [cs.AI]. URL:
[123] Wah, C. et al. "The Caltech-UCSD Birds-200-2011 Dataset". In: 2011. URL:

[124] Wang, J. et al. "Generalizing to Unseen Domains: A Survey on Domain Generalization". In: *IEEE Transactions on Knowledge and Data Engineering* 35 (2021), pp. 8052–8072. URL:

[125] Wang, P. and Vasconcelos, N. "SCOUT: Self-Aware Discriminant Counterfactual Explanations". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 8978–8987. URL:

[126] Wang, X. et al. "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017, pp. 3462–3471. DOI: 10.1109/cvpr.2017.369. URL:

[127] Wang, Y. et al. *Transferring GANs: generating images from limited data*. 2018. arXiv: 1805.01677 [cs.CV]. URL:

[128] Wang, Y. et al. *MineGAN: effective knowledge transfer from GANs to target domains with few images*. 2020. arXiv: 1912.05270 [cs.CV]. URL:

[129] White, A. et al. "Contrastive counterfactual visual explanations with overdetermination". In: *Machine Learning* 112 (2021), pp. 3497–3525. URL:

[130] Woodland, M. et al. "Evaluating the Performance of StyleGAN2-ADA on Medical Images". In: *Simulation and Synthesis in Medical Imaging*. Springer International Publishing, 2022, pp. 142–153. ISBN: 9783031169809. DOI: 10.1007/978-3-031-16980-9_14. URL:

[131] Woodland, M. et al. *StyleGAN2-based Out-of-Distribution Detection for Medical Imaging*. 2023. arXiv: 2307.10193 [eess.IV]. URL:

[132] Xia, W. et al. *GAN Inversion: A Survey*. 2022. arXiv: 2101.05278 [cs.CV]. URL:

[133] Yang, Y. et al. "DiffMIC: Dual-Guidance Diffusion Network for Medical Image Classification". In: *ArXiv* abs/2303.10610 (2023). URL:

[134] Yi, X., Walia, E., and Babyn, P. "Generative adversarial network in medical imaging: A review". In: *Medical image analysis* 58 (2019), p. 101552.

[135] Yu, F. et al. *LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop*. 2016. arXiv: 1506.03365 [cs.CV]. URL:

[136] Zhang, H. et al. *Theoretically Principled Trade-off between Robustness and Accuracy*. 2019. arXiv: 1901.08573 [cs.LG]. URL:

[137] Zhang, M. et al. "Adaptive Risk Minimization: Learning to Adapt to Domain Shift". In: *Neural Information Processing Systems*. 2020. URL:

[138] Zhang, R. et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *CVPR*. 2018.

[139] Zhao, Y. "Fast Real-time Counterfactual Explanations". In: *ArXiv* abs/2007.05684 (2020). URL:

[140] Zheng, G. et al. "A Novel Computer-Aided Diagnosis Scheme on Small Annotated Set: G2C-CAD". In: *BioMed Research International* 2019 (2019). URL:

[141] Zhou, B. et al. "Scene Parsing through ADE20K Dataset". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 5122–5130. URL:

[142] Zhu, J. et al. "In-domain GAN Inversion for Real Image Editing". In: *Proceedings of European Conference on Computer Vision (ECCV)*. 2020.

[143] Zhu, J.-Y. et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.