Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Τεχνικές κανονικοποίησης συνέπειας για την ημιεπιβλεπόμενη σημασιολογική κατάτμηση εικόνας

Διπλωματικη Εργασια

του

ΝΙΚΟΛΑΟΥ ΣΠΥΡΟΥ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής ΕΜΠ

Αθήνα, Μάρτιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Τεχνικές κανονικοποίησης συνέπειας για την ημιεπιβλεπόμενη σημασιολογική κατάτμηση εικόνας

Διπλωματικη Εργασια

του

ΝΙΚΟΛΑΟΥ ΣΠΥΡΟΥ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2η Μαρτίου του 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Κόλλιας	Αθανάσιος Βουλόδημος	Γεώργιος Στάμου
Καθηγητής ΕΜΠ	Επίκουρος Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ

Αθήνα, Μάρτιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Νικόλαος Σπύρου

1η Μαρτίου του 2023

Περίληψη

Η ευρεία ανάπτυξη των βαθιών συνελικτικών δικτύων και ειδικότερα των αρχιτεκτονικών κωδικοποιητή-αποκωδικοποιητή έχει οδηγήσει στην εκτεταμένη χρήση τους για την επίλυση του προβλήματος της σημασιολογικής κατάτμησης εικόνας. Παρόλα αυτά, προκειμένου να εκπαιδευτούν μοντέλα με ικανοποιητική ικανότητα γενίκευσης για το συγκεκριμένο πρόβλημα απαιτούνται μεγάλου πλήθους σύνολα επισημασμένων δεδομένων. Στην περίπτωση της σημασιολογικής κατάτμησης η επισήμανση δεδομένων είναι μια αρκετά χρονοβόρα διαδικασία, καθώς σε κάθε εικονοστοιχείο πρέπει να ανατεθεί μία σημασιολογική κλάση. Για αυτόν τον λόγο, ιδιαίτερη έμφαση δίνεται σε ανάπτυξη μεθόδων ημιεπιβλεπόμενης μάθησης, όπου το δίκτυο εκπαιδεύεται πάνω σε λιγότερα επισημασμένα και σε περισσότερα μη επισημασμένα δεδομένα, με στόχο να μπορεί να παρουσιάζει εξίσου καλή ικανότητα γενίκευσης. Στην παρούσα διπλωματική εξετάζεται η μέθοδος της κανονικοποίησης συνέπειας (consistency regularization) για την αξιοποίηση των μη επισημασμένων δεδομένων, σύμφωνα με την οποία το δίκτυο καλείται να παράγει όμοιες προβλέψεις για διαταραγμένες (perturbed) εκδοχές της εικόνας εισόδου (διαταραχές εισόδου). Συγκεκριμένα, χρησιμοποιούμε το παράδειγμα εκπαίδευσης της ασθενούς-ισχυρής συνέπειας (weak-to-strong consistency), κατά το οποίο το δίκτυο δέχεται στην είσοδο μία ασθενώς διαταραγμένη εκδοχή (λιγότερο έντονη επαύξηση, ασθενής επαύξηση) και η πρόβλεψη του για αυτή χρησιμοποιείται ως ψευδοετικέτα για την επίβλεψη της αντίστοιχης εξόδου της ισχυρά διαταραγμένης εκδοχής (πιο έντονη επαύξηση, ισχυρή επαύξηση). Για τη δημιουργία της ισχυρά επαυξημένης εκδοχής μιας εικόνας, πειραματιζόμαστε με διάφορες μεθόδους επαύξησης, όπως με μετασχηματισμούς χρώματος (color distortions, color augmentation), καθώς και με τεχνικές μίξης ζευγών εικόνων, όπως οι ClassMix, CutMix, οι οποίες συνδυάζουν εικόνες με σκοπό την παραγωγή νέων πιο πολύπλοκων τεχνητών εικόνων με διευρυμένο σημασιολογικό περιεχόμενο. Παράλληλα, πειραματιζόμαστε και με την περίπτωση όπου το δίκτυο τροφοδοτείται με δύο ισχυρά επαυξημένες εκδοχές που έχουν παραχθεί είτε με την ίδια, είτε και με διαφορετική μέθοδο επαύξησης. Αναφέρουμε και συγκρίνουμε τα πειραματικά αποτελέσματα στα σύνολα δεδομένων Pascal VOC 2012, CelebAMask-HQ και QaTa-COV19-v2.

Λέξεις Κλειδιά

σημασιολογική κατάτμηση εικόνας, βαθιά συνελικτικά δίκτυα, δίκτυα κωδικοποιητή-αποκωδικοποιητή, ημιεπιβλεπόμενη μάθηση, κανονικοποίηση συνέπειας, διαταραχές εισόδου, ασθενής-ισχυρή συνέπεια, ψευδοετικέτα, ισχυρή επαύξηση, ασθενής επαύξηση, μετασχηματισμοί χρώματος, ClassMix, CutMix.

Abstract

The widespread development of deep convolutional networks, especially encoder-decoder architectures, has led to their extensive use in solving the problem of image semantic segmentation. However, large labeled datasets are required to train models with satisfactory generalization ability for this problem. In the case of semantic segmentation, data labeling is a time-consuming process, as each image pixel must be assigned to a semantic class. Therefore, particular emphasis is placed on the development of semi-supervised learning methods, where the network is trained on fewer labeled and more unlabeled data, with the aim of exhibiting equally good generalization ability. In this thesis, the consistency regularization method is examined to exploit unlabeled data, according to which the network is called to produce similar predictions for perturbed input images. Specifically, we use the example of weak-to-strong consistency training, where the network accepts a weakly perturbed version (less intense augmentation, weak augmentation) in the input and its prediction for it is used as a pseudo-label to supervise the corresponding output of the strongly perturbed version (more intense augmentation, strong augmentation). To create the strongly augmented version of an image, we experiment with various augmentation methods, such as color distortions(color augmentation), as well as image mixing techniques, such as ClassMix, CutMix, which combine images in order to produce new and more complex artificial images with an extended semantic content. At the same time, we are experimenting also with the case where the network is fed with two strongly augmented versions that have been generated either with the same or different augmentation method. We report and compare the experimental results on the Pascal VOC 2012, CelebAMask-HQ, and QaTa-COV19-v2 datasets.

Keywords

image semantic segmentation, deep convolutional networks, encoder-decoder architectures, semi-supervised learning, consistency regularization, input perturbations, weak-strong consistency, pseudo-label, strong augmentation, weak augmentation, color distortions, color augmentation, ClassMix, CutMix

στην οικογένεια μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Στέφανο Κόλλια για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να πραγματοποιηθεί η εκπόνηση της στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης (AILS) του ΕΜΠ. Επίσης θα ήθελα να ευχαριστήσω ιδιαίτερα την κ. Παρασκευή Τζούβελη για την καθοδήγησή της, τις χρήσιμες συμβουλές της και γενικότερα για την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου, τον αδερφό μου και τους φίλους μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Μάρτιος 2023

Νικόλαος Σπύρου

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της διπλωματικής

Η σημασιολογική κατάτμηση εικόνας αποτελεί έναν πολύ σημαντικό τομέα της όρασης υπολογιστών με πολλές εφαρμογές σε προβλήματα του πραγματικού κόσμου. Η διαμέριση της εικόνας σε υποπεριοχές και η αναγνώριση αντικειμένων και περιοχών ενδιαφέροντος σε αυτή, μέσω αλγορίθμων μηχανικής μάθησης, συνεισφέρει στην καλύτερη κατανόηση του περιεχομένου της που είναι απαραίτητη για τη δημιουργία αξιόπιστων συστημάτων τεχνητής νοημοσύνης, τα οποία θα μπορούν να επιλύουν προβλήματα της καθημερινότητας. Μερικοί από τους τομείς που η σημασιολογική κατάτμηση βρίσκει ευρεία εφαρμογή είναι τα αυτο-οδηγούμενα οχήματα (Self-driving vehicles) [28, 29, 30, 31], η τμηματοποίηση ιατρικών εικόνων (medical image segmentation) [32, 33], η τμηματοποίηση δορυφορικών εικόνων(land cover segmentation) [34, 35], καθώς και η κατανόηση σκηνών (scene understanding) [36]. Η ραγδαία ανάπτυξη των βαθιών νευρωνικών δικτύων έχει οδηγήσει τα τελευταία χρόνια στη ευρεία χρήση τους σε προβλήματα σημασιολογικής κατάτμησης με τα αποτελέσματα να είναι πολύ ικανοποιητικά. Παρόλα αυτά, προκειμένου να έχουμε αξιόπιστα μοντέλα που θα παρουσιάζουν καλή ικανότητα γενίκευσης απαιτούνται μεγάλες ποσότητες επισημασμένων δεδομένων (labeled data) για την εκπαίδευση των βαθιών νευρωνικών δικτύων. Σε ένα πρόβλημα όπως η σημασιολογική κατάτμηση, η δημιουργία ετικετών (data labeling) θεωρείται αρκετά χρονοβόρα διαδικασία, καθώς χρειάζεται σε κάθε εικονοστοιχείο pixel να αποδοθεί μία ετικέτα. Χαρακτηριστικά, για το σύνολο δεδομένων Cityscapes [37], αναφέρεται ότι για να γίνει η επισήμανση όλων των pixel μίας εικόνας απαιτούνται κατα μέσο όρο 90 λεπτά. Για αυτόν τον λόγο τα τελευταία χρόνια έχει πραγματοποιηθεί μεγάλη έρευνα γύρω από αλγορίθμους ημιεπιβλεπόμενης μάθησης που αξιοποιούν ένα σχετικά μικρό πλήθος επισημασμένων(labeled) δεδομένων σε συνδυασμό με ένα αρκετά μεγαλύτερο πλήθος μη επισημασμένων (unlabeled) δεδομένων με σκόπο την εκπαίδευση μοντέλων που θα μπορούν να παρουσιάζουν παρόμοια ικανότητα γενίκευσης με τα πλήρως επιβλεπόμενα μοντέλα, μετριάζοντας έτσι την ανάγκη για μεγάλους όγκους επισημασμένων δεδομένων.

Στην παρούσα διπλωματική εργασία εξετάζουμε τη μέθοδο της κανονικοποίησης συνέπειας (consistency regularization) [38], μίας πολύ διαδεδομένης τεχνικής για ημιεπιβλεπόμενη μάθηση, όπου το δίκτυο καλείται να μάθει να παράγει όμοιες εξόδους για διαφορετικές επαυξημένες (augmented) ή διαταραγμένες (perturbed) εκδοχές ενός δείγματος εισόδου. Πιο συγκεκριμένα, βασιζόμαστε στο παράδειγμα εκπαίδευσης ασθενούς-ισχυρής συνέπειας

του FixMatch[13], όπου η πρόβλεψη για μία ασθενώς επαυξημένη εικόνα χρησιμοποιείται ως ψευδοετικέτα για την επίβλεψη της εξόδου της αντίστοιχης ισχυρά επαυξημένης εκδοχής. Για τη δημιουργία της ισχυρά επαυξημένης εκδοχής μιάς εικόνας πειραματιζόμαστε με τις τεχνικές επαύξησης ClassMix [16], CutMix[15], καθώς και με την επαύξηση χρώματος (color augmentation) [39]. Επιπροσθέτως, αξιοποιώντας την ιδέα που προτείνεται στο [14] για τη δημιουργία δύο ισχυρά επαυξημένων εκδοχών και την ταυτόχρονη τροφοδότηση τους στο δίκτυο, πειραματιζόμαστε με τις διαφορετικές μεθόδους επαύξησης που μπορούν να παραχθούν αυτές οι δύο εκδοχές (π.χ ClassMix + CutMix). Αναφέρουμε τα πειραματικά μας αποτελέσματα στα σύνολα δεδομένων Pascal VOC 2012 [19], CelebAMask-HQ [20] και QaTa-COV19 [21, 22] και παρουσιάζουμε επιπλέον ποιοτικά αποτελέσματα που παράγει το δίκτυο κατάτμησης.

Τέλος να αναφέρουμε ότι αρχιτεκτονικές βαθιών νευρωνικών δικτύων έχουν υλοποιηθεί και χρησιμοποιηθεί σε διάφορες εφαρμογές από μέλη του Εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης (AILS) του ΕΜΠ. Ειδικότερα επιβλεπόμενες τεχνικές CNN και CNN-RNN έχουν εφαρμοστεί για κατηγοριοποίηση αντικειμένων, στην ιατρική διάγνωση νευροεκφυλιστικών ασθενειών, όπως της νόσου του Πάρκινσον [40, 41, 42, 43, 44] ή της Covid-19 [45, 46, 47, 48], περιλαμβάνοντας κατάτμηση 2D ή 3D εικόνων. Έμφαση έχει δοθεί στην διαφάνεια και στην προσαρμογή των μοντέλων [49, 50, 51] αλλά και στην ανάπτυξη πλέον σύνθετων αρχιτεκτονικών, μπαϋεσιανών, με κάψουλες και αβεβαιότητα [52, 53, 54, 55]. Βαθιές ημι- και αυτο- επιβλεπόμενες 3D νευρωνικές αρχιτεκτονικές, αλλά και αρχιτεκτονικές κωδικοποιητή- αποκωδικοποιητή έχουν εφαρμοστεί στην ανίχνευση βλαβών σε πυρηνικούς αντιδραστήρες [56, 57], στην πρόβλεψη της παραγωγής στον αγροτικό τομέα [58, 59] και στην αναγνώριση και σύνθεση συναισθήματος [60, 61], ενώ άλλες εφαρμόζονται σε προβλήματα ανάλυσης εικόνων και αλληλεπίδρασης ανθρώπου-υπολογιστή [62, 63].

1.2 Οργάνωση του τόμου

Η εργασία περιλαμβάνει τα εξής κεφάλαια:

- **Κεφάλαιο** 2: Μια συνοπτική παρουσίαση θεωρητικών εννοιών που είναι απαραίτητες για τη καλύτερη κατανόηση της εργασίας.
- **Κεφάλαιο** 3: Παρουσιάση των βασικών μεθόδων κανονικοποίησης συνέπειας που συναντόνται στη βιβλιογραφία.
- **Κεφάλαιο** 4: Παρουσίαση των συνόλων δεδομένων που χρησιμοποιούνται στην εργασία.
- **Κεφάλαιο** 5: Παρουσίαση της βασικής μεθοδολογίας που ακολουθούμε.
- **Κεφάλαιο** 6: Παρουσίαση πειραματικών αποτελεσμάτων.
- **Κεφάλαιο** 7: Τελικά συμπεράσματα και μελλοντικές επεκτάσεις.

Μέρος Ι

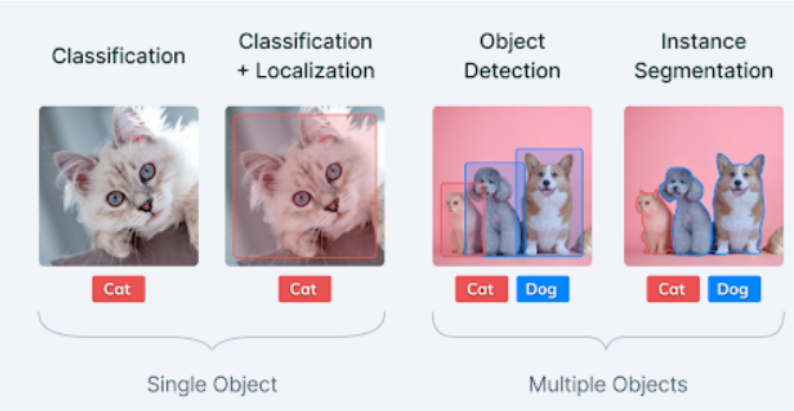
Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

2.1 Κατάτμηση εικόνας

Πρόκειται για το πρόβλημα της όρασης υπολογιστών και της ψηφιακής ανάλυσης εικόνας, το οποίο στοχεύει στην ομαδοποίηση όμοιων περιοχών ή τμημάτων μιάς εικόνας. Μπορούμε να θεωρήσουμε ότι μία ψηφιακή εικόνα αναπαρίσταται από ένα σύνολο από pixels, τα οποία ανήκουν σε κάποια κλάση. Συνεπώς το πρόβλημα της δόμησης όμοιων τμημάτων εικόνας είναι ισοδύναμο με την κατηγοριοποίηση σε επίπεδο pixel. Η κατάτμηση εικόνας αποτελεί μία επέκταση του προβλήματος της κατηγοριοποίησης εικόνας, καθώς εκτός από την κατηγοριοποίηση πραγματοποιείται και ο εντοπισμός της ακριβής θέσης και των ορίων των αντικειμένων που απεικονίζονται [1].



Εικόνα 2.1: *Παράδειγμα της κατηγοριοποίησης (classification) και της κατάτμησης (segmentation)[1]*

2.1.1 Είδη κατάτμησης εικόνας

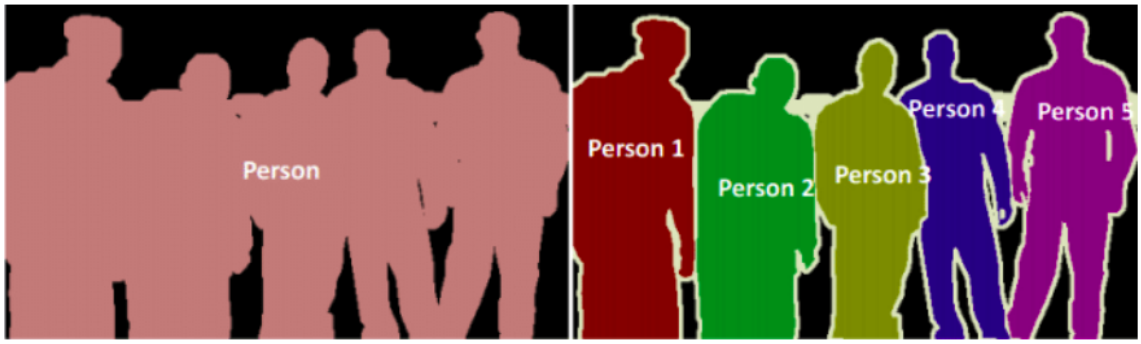
Οι 2 βασικοί τρόποι για να τμηματοποιηθεί μία εικόνα είναι οι εξής:

- **Σημασιολογική κατάτμηση** (Semantic Segmentation).
- **Κατάτμηση παραδειγμάτων** (Instance Segmentation).

Σημασιολογική κατάτμηση εικόνας: Κατηγοριοποίηση των pixels μιας εικόνας σε σημασιολογικές κλάσεις. Τα pixels που ανήκουν σε κάποια συγκεκριμένη κλάση απλώς

ταξινομούνται σε αυτή χωρίς να λαμβάνεται υπόψην περαιτέρω πληροφορία [1].

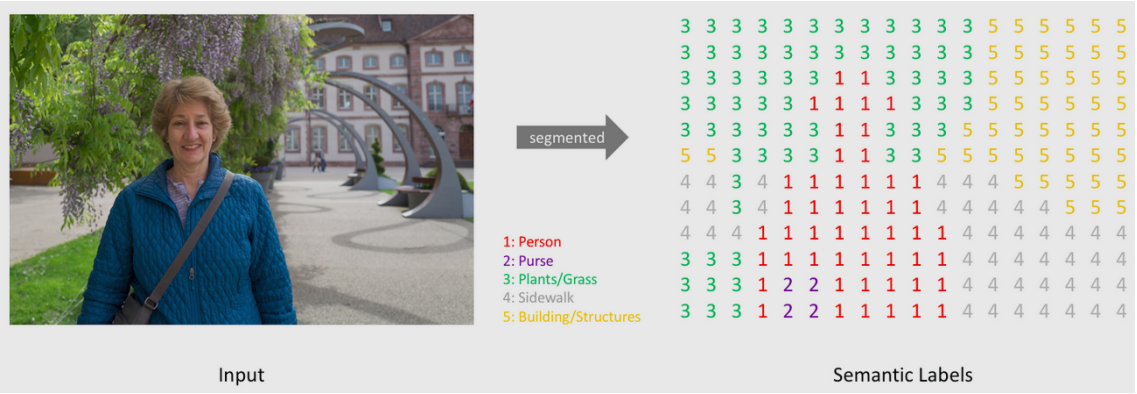
Κατάτμηση παραδειγμάτων εικόνας: Η κατηγοριοποίηση των pixels γίνεται με βάση το πλήθος των αντικειμένων (παρουσιών) που συναντώνται σε μία εικόνα χωρίς να λαμβάνονται υπόψην κλάσεις. Ένας αλγόριθμος τμηματοποίησης παραδειγμάτων δεν ξέρει την κλάση στην οποία ανήκει μια περιοχή, αλλά μπορεί να διαχωρίσει επικαλυπτόμενες ή πολύ παρόμοιες περιοχές αντικειμένων με βάση τα όριά τους [1].



Εικόνα 2.2: *Αριστερά :Semantic Segmentation Δεξιά :Instance Segmentation [2]*

2.2 Συνελικτικά Νευρωνικά Δίκτυα για την επίλυση του προβλήματος της σημασιολογικής κατάτμησης εικόνας

Ένα συνελικτικό δίκτυο παίρνει ως είσοδο μία RGB εικόνα και προσπαθεί να παράξει ως έξοδο έναν χάρτη τμηματοποίησης(segmentation map), χωρικών διαστάσεων ίδιων με της εικόνας εισόδου, ο οποίος περιέχει σε κάθε του θέση έναν ακέραιο δείκτη (class index) που υποδηλώνει την κλάση που έχει ταξινομηθεί το αντίστοιχο pixel της αρχικής εικόνας. Μπορούμε να πούμε ότι το πρόβλημα της σημασιολογικής κατάτμησης αποτελεί ουσιαστικά ένα πρόβλημα κατηγοριοποίησης (classification) σε επίπεδο pixel [3].



Εικόνα 2.3: *Παράδειγμα τμηματοποίησης με 5 σημασιολογικές κλάσεις [3]*

Υποθέτουμε τα ότι έχουμε τα εξής:

- Εικόνα εισόδου *RGB* με χωρικές διαστάσεις (H,W), έστω $I \in R^{HxWx3}$
- Σύνολο C με πλήθος K σημασιολογικές κλάσεις, έστω $C = \{0, 1, 2, .., K-1\}$

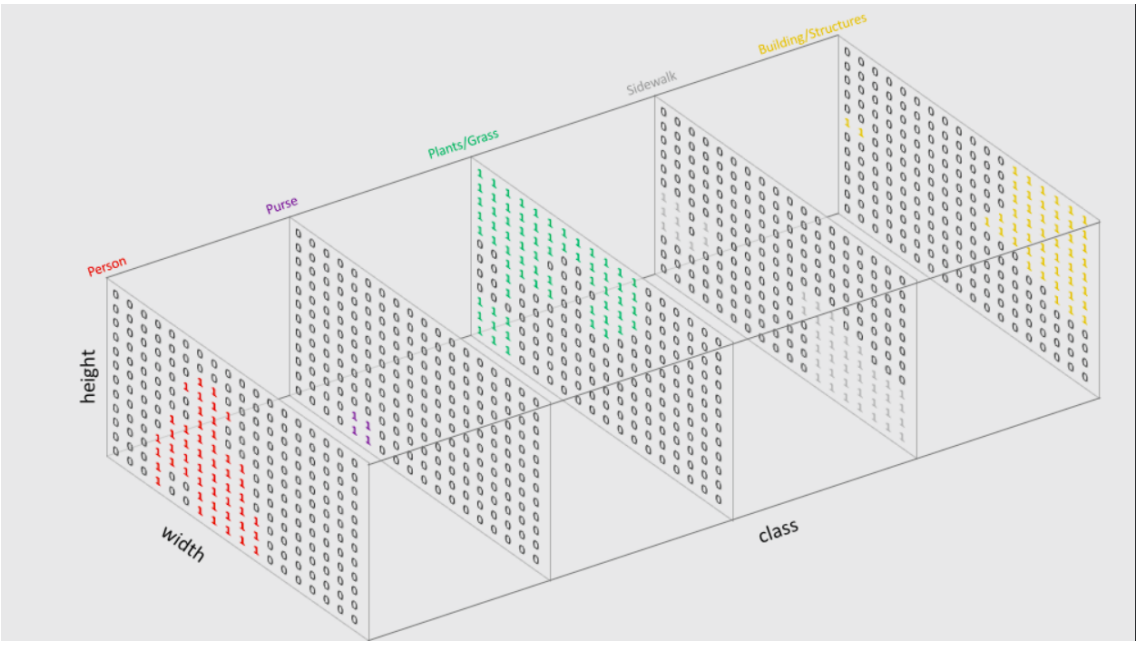
- Συνελικτικό δίκτυο με παραμέτρους θ, έστω $f(\partial)$

Εφαρμόζοντας το δίκτυο f στην εικόνα εισόδου I, τότε έχουμε:

$$y = f(I) \in R^{HxWx|C|} \tag{2.1}$$

Η έξοδος του δικτύου αποτελείται από χάρτες χαρακτηριστικών (feature maps) K-καναλιών (πλήθος σημασιολογικών κλάσεων) και χωρικών διαστάσεων ίδιων με της εικόνας εισόδου. Αν εφαρμόσουμε την συνάρτηση *argmax* για κάθε pixel της εξόδου y του δικτύου θα παραχθεί ο χάρτης τμηματοποίησης y^*(segmented map) που έχουμε αναφέρει.

$$y^* = argmax(f(I)) = argmax(y) \in R^{HxWx1} \tag{2.2}$$

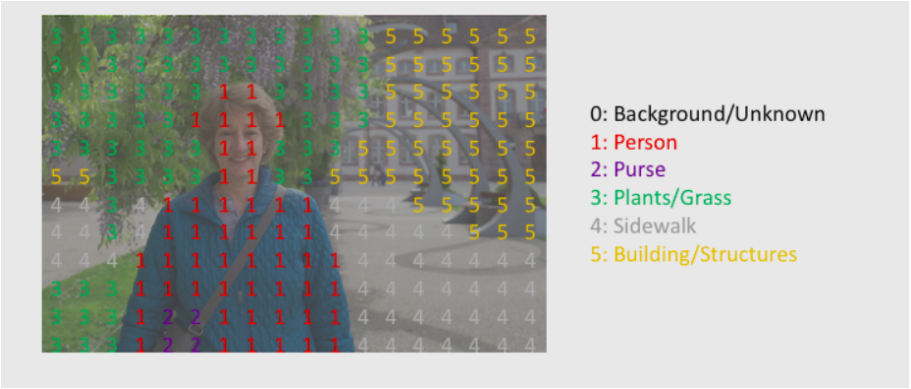


Εικόνα 2.4: *One-Hot encoded πρόβλεψη του δικτύου πριν την εφαρμογή της argmax [3]*

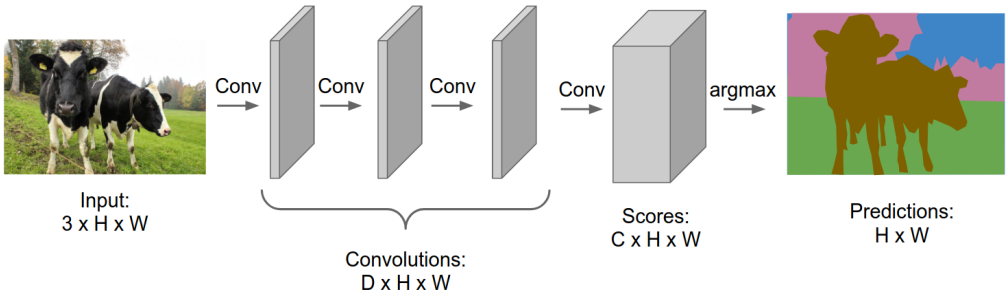
Επομένως, η one-hot encoded πρόβλεψη του δικτύου που διάθετει κανάλια όσα και οι κλάσεις του προβλήματος μπορεί να αναχθεί σε μία εικόνα ενός μόνο καναλιού (single-channel) μετά την εφαρμογή της *argmax* σε κάθε pixel. Το αποτέλεσμα εκτός από χάρτης τμηματοποίησης ονομάζεται και μάσκα τμηματοποίησης (segmentation mask) 2.5 [3].

2.2.1 Αρχιτεκτονικές Encoder - Decoder

Στόχος είναι να σχεδιάσουμε αρχιτεκτονικές συνελικτικών δικτύων που θα επιλύουν το πρόβλημα. Μία πρώτη προσέγγιση είναι να κατασευάσουμε ένα δίκτυο, το οποίο θα αποτελείται μόνο από διαδοχικά συνελικτικά επίπεδα που θα έχουν το ίδιο γέμισμα(padding), έτσι ώστε οι χωρικές διαστάσεις να διατηρούνται σταθερές κατά το πέρασμα της εικόνας από το δίκτυο. Με αυτό τον τρόπο το δίκτυο μπορεί να μάθει μία αντιστοίχιση απο τον χώρο της εικόνας εισόδου R^{HxWx3} στον χώρο του τελικού χάρτη τμηματοποίησης R^{HxWx1}. Μία τέτοια αρχιτεκτονική φαίνεται στην εικόνα 2.6 [3].



Εικόνα 2.5: *Η τελική μάσκα μετά την εφαρμογή της argmax [3]*



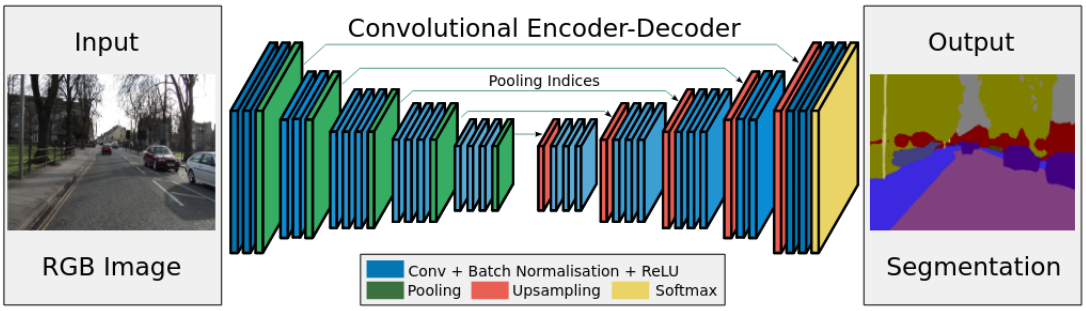
Εικόνα 2.6: *[4]*

Βλέπουμε, λοιπόν, ότι η αρχική χωρική ανάλυση της εικόνας διατηρείται κατά το πέρασμα της απο το δίκτυο, κάτι που στην πράξη αποδεικνύεται αρκέτα ακριβό υπολογιστικά.

Γενικά στις σύγχρονες αρκτεκτονικές βαθιών συνελικτικών δικτύων γίνεται χρήση τεχνικών υποδειγματοληψίας (downsampling) έτσι ώστε να μειώνεται η διαστατικότητα των ενδιάμεσων χαρτών χαρακτηριστικών (feature maps) που παράγονται και να ελαφρύνεται το υπολογιστικό κόστος του forward περάσματος της εικόνας. Τέτοιες τεχνικές είναι είτε η χρήση συνελικτικών επιπέδων με stride > 1, είτε πράξεις pooling. Σε περιπτώσεις απλής κατηγοριοποίησης (classification) η μείωση της διαστατικότητας δεν αποτελεί πρόβλημα, καθώς στην έξοδο του το δίκτυο απλά πρέπει να αποφανθεί για το περιεχόμενο της εικόνας εισόδου [3]. Σε ένα πρόβλημα σημασιολογικής τμηματοποίησης όμως, όπου πρέπει να γίνει κατηγοριοποίηση του κάθε pixel ξεχωριστά, όπως έχουμε αναφέρει σε προηγούμενη παράγραφο η έξοδος του δικτύου θα πρέπει να είναι ένας χάρτης τμηματοποίησης που θα έχει χωρική ανάλυση ίδια με αυτή της αρχικής εικόνας[3]. Συνεπώς, με κάποιο τρόπο τα ενδιάμεσα χαμηλής χωρικής ανάλυσης χαρακτηριστικά που έχουν παραχθεί χρειάζεται να αναδειγματοληπτηθούν (upsampling) στην αρχική χωρική ανάλυση της εικόνας εισόδου [3].

Η πιο διαδεδομένη προσέγγιση για προβλήματα σημασιολογικής τμηματοποίησης είναι αρχιτεκτονικές, οι οποίες διαθέτουν δομή κωδικοποιητή - αποκωδικοποιητή (encoder-decoder). Μία από τις πρώτες δουλειές που πρότειναν τέτοιου είδους δίκτυα είναι το SegNet[5], όπου η εικόνα εισόδου περνάει αρχικά απο διαδοχικά συνελικτικά (convolutional layers with stride) και υποδειγματοληπτικα επίπεδα (pooling layers), έτσι ώστε να γίνεται συμπίεση της πληροφορίας σε χάρτες χαρακτηριστικών (feature maps) χαμηλής χωρικής διαστατικότητας. Το τμήμα του δικτύου που επιτελεί τα παραπάνω είναι ουσιαστικά ο en-

coder. Εν συνεχεία, το δεύτερο τμήμα του δικτύου ο decoder αναλαμβάνει να αποσυμπιέσει και να ανακατασκευάσει την έξοδο που παίρνει απο τον encoder. Με διαδοχικές πράξεις αναδειγματοληψίας (upsampling) ο decoder παράγει στην έξοδο του το τελικό χάρτη τμηματοποίησης (segmentation map) που έχει τη χωρική ανάλυση της εικόνας εισόδου.



Εικόνα 2.7: *Η αρχιτεκτονική encoder-decoder, όπως αυτή παρουσιάζεται στο SegNet [5]*

Τα δίκτυα της παραπάνω μορφής ονομάζονται πλήρως συνελικτικά δίκτυα (Fully Convolutional Networks), καθώς αποτελούνται αποκλειστικά απο συνελικτικά επίπεδα και δεν έχουν κανένα πλήρως συνδεδεμένο επίπεδο (fully connected layer).

Ας υποθέσουμε τα εξής:

- Εικόνα εισόδου *RGB* $I \in R^{HxWx3}$
- C σημασιολογικές κλάσεις
- Συνελικτικο δίκτυο με παραμέτρους ∂, έστω $f(\partial)$, το οποίο αποτελείται από 2 επιμέρους τμήματα:
  - δίκτυο Encoder, με παραμέτρους $\partial 1$, έστω $e(\partial 1)$
  - δίκτυο Decoder, με παραμέτρους $\partial 2$, έστω $d(\partial 1)$
  - $\partial = \partial 1 \cup \partial 2$

Εφαρμόζουμε τον encoder e στην εικόνα εισόδου και παίρνουμε τα ενδιάμεσα χαρακτηριστικά χαμηλής χωρικής διαστατικότητας.

$$y_f = e(I) \in R^{H/4xW/4xD} \tag{2.3}$$

Υποθέτουμε, ότι ο encoder υποτετραπλασιάζει την αρχική χωρική ανάλυση της εικόνας και ότι τα παραγόμενα ενδιάμεσα χαρακτηριστικά βρίσκονται σε έναν χώρο υψηλής διαστατικότητας, έστω D.

Έν συνέχεια τα ενδιαμέσα χαρακτηριστικά τροφοδοντούνται στον decoder. Ο decoder θα παράξει χάρτες χαρακτηριστικών χωρικής ανάλυσης ίδιας με της εικόνας εισόδου, ενώ τα κανάλια θα είναι ίσα με τον αριθμό των κλάσεων. Ουσιαστικά, κάθε κανάλι αντιπροσωπεύει μία κλάση. Επίσης, στην έξοδο του decoder εφαρμόζονται διαδοχικά σε επίπεδο pixel οι συναρτήσεις *softmax* και *argmax* για να προκύψει ο τελικός χάρτης τμηματοποίησης

(segmentation map).

$$y = softmax(d(y_f)) \in R^{HxWxC} \tag{2.4}$$

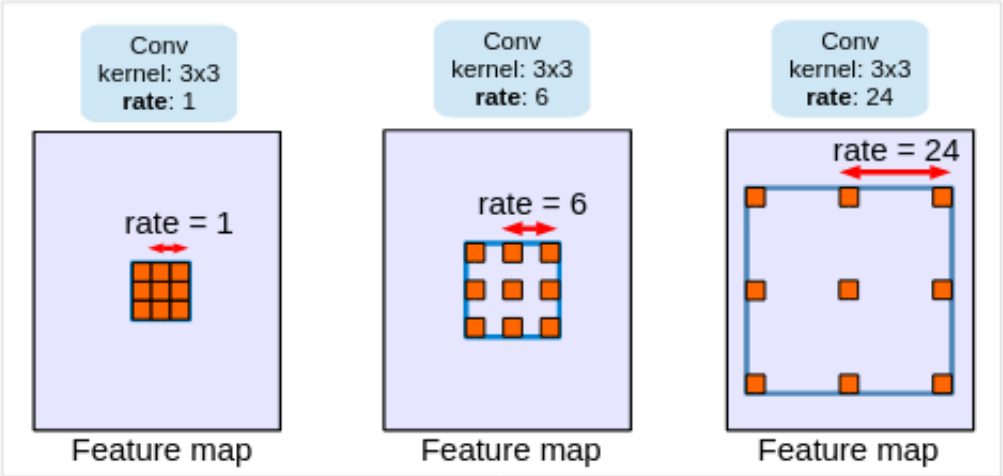
$$y_o = argmax(y) \in R^{HxW} \tag{2.5}$$

2.2.2 Τα δίκτυα DeepLabV3 και DeepLabV3plus

Όπως, αναφέραμε στην προηγούμενη παράγραφο, τα πλήρως συνελικτικά δίκτυα(FCN), τα οποία ακολουθούν την αρχιτεκτονική encoder-decoder είναι αυτά που χρησιμοποιούνται κατ' εξοχήν στα περισσότερα προβλήματα σημασιολογικής κατάτμησης. Ένα πρόβλημα που αντιμετώπιζουν αυτές οι αρχιτεκτονικές είναι ότι οι χάρτες χαρακτηριστικών (features maps) που παράγονται, καθώς προχωράμε σε βαθύτερα επίπεδα του δικτύου, υπόκεινται σε συνελίξεις και πράξεις pooling που έχουν ως αποτέλεσμα τη μείωση της χωρικής ανάλυσης τους και συνεπώς, την απώλεια πληροφορίας που είναι πολύ σημαντική για ένα πρόβλημα τμηματοποίησης, όπου θέλουμε ο τελικός χάρτης τμηματοποίησης να είναι καλής ποιότητας. Οι αρχιτεκτονικές τύπου Deeplab[7], που παρουσιάζουμε εδώ, προσπαθούν να μετριάσουν το παραπάνω πρόβλημα με την εισαγωγή των εξής δύο ιδεών:

- Διεσταλμένες συνελίξεις (Atrous Convolutions ή Dilated Convolutions)[7]
- Spatial Pyramid Pooling[7]

Dilated Convolutions: Πρόκειται για συνελίξεις που πραγματοποιούνται μεταξύ των χαρτών χαρακτηριστικών και φίλτρων, τα οποία περιέχουν μηδενικά μεταξύ δύο διαδοχικών μη μηδενικών τιμών τους. Το πλήθος των μηδενικών που εισάγονται, καθορίζεται από έναν αριθμό που ονομάζεται dilation rate και συμβολίζεται με r. Συγκεκριμένα, ανάμεσα σε κάθε ζεύγος διαδοχικών τιμών ενός αρχικού φίλτρου, εισάγονται r-1 μηδενικά σε κάθε διάσταση, έτσι ώστε ένα προκύψει το dilated φίλτρο. Προφανώς, αν επιλεγεί $r = 1$, τότε αυτό σημαίνει ότι χρησιμοποιούμε ακριβώς το αρχικό φίλτρο, χωρίς dilation.

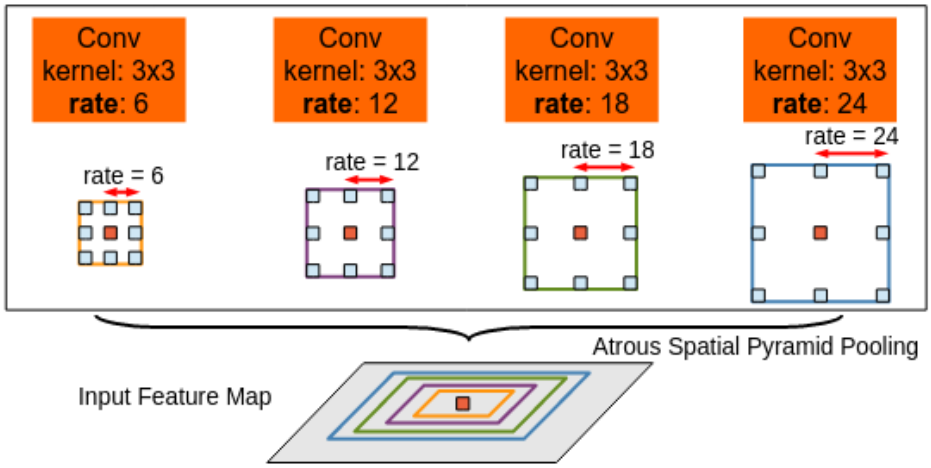


Εικόνα 2.8: *Φίλτρο αρχικών χωρικών διαστάσεων 3x3, με προσθήκη διαφόρων ρυθμών διαστολής (dilation rates) [6]*

Αυτό που επιτυγχάνεται με την εισαγωγή του ρυθμού διαστολής είναι ότι μεγαλώνει το πεδίο θέασης(field of view) του φίλτρου με αποτέλεσμα να διευρύνεται η πληροφορία που

εξάγεται μετά από κάθε συνέλιξη. Επομένως, κάνοντας χρήση τέτοιων συνελίξεων μπορούμε χωρίς πράξεις pooling να διατηρούμε την ίδια χωρική ανάλυση και ταυτόχρονα να εξάγουμε πληροφορία σε πολλαπλές κλίμακες, αυξάνοντας τον ρυθμό διαστολής. Επιπλέον, η διεύρυνση των πυρήνων με αυτό τον τρόπο δεν επιφέρει υπολογιστική επιβάρυνση, καθώς δεν αυξάνονται σε πλήθος οι τιμές των φίλτρων (βάρη του δικτύου).

Dilated Spatial Pyramid Pooling: Αξιοποιώντας το παραπάνω είδος συνέλιξης μπορεί να δομηθεί το Dilated Spatial Pyramid Pooling(DSPP). Ουσιαστικά, πρόκειται για μία συνιστώσα ή επίπεδο(layer), η οποία αποτελείται από παράλληλες διεσταλμένες συνελίξεις με διαφορετικό ρυθμό διαστολής και εφαρμόζεται σε κάποια χαρακτηριστικά εισόδου. Ο στόχος είναι αξιοποιώντας τους διαφορετικούς ρυθμούς r να μπορεί να εξαχθεί πληροφορία σε πολλαπλές κλίμακες.



Εικόνα 2.9: *Η δομή του Atrous Spatial Pyramid Pooling[7]*

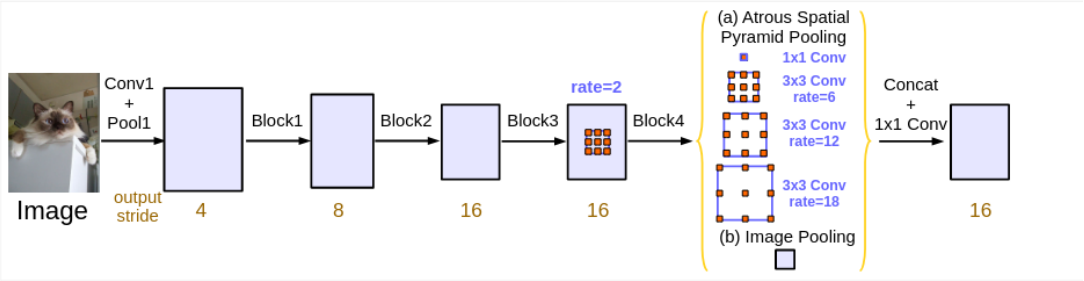
DeeplabV3 [6]

Το δίκτυο DeepLabV3[6] ουσιαστικά είναι η εξέλιξη της οικογένειας δικτύων DeepLab [7] που αποτελούν state of the art αρχιτεκτονικές για τη σημασιολογική κατάτμηση εικόνας. Η συγκεκριμένη αρχιτεκτονική αποτελείται από 2 βασικές συνιστώσες. Ένα βαθύ συνελικτικό δίκτυο (π.χ ResNet[23]), το οποίο ονομάζεται και ραχοκοκαλιά(backbone), καθώς και τη δομή Dilated Spatial Pyramid Pooling. Ουσιαστικά, στα χαρακτηριστικά εξόδου που παράγει το ResNet, εφαρμόζεται το Dilated Spatial Pyramid Pooling, για την εξαγωγή πληροφοριάς σε πολλαπλές κλίμακες.

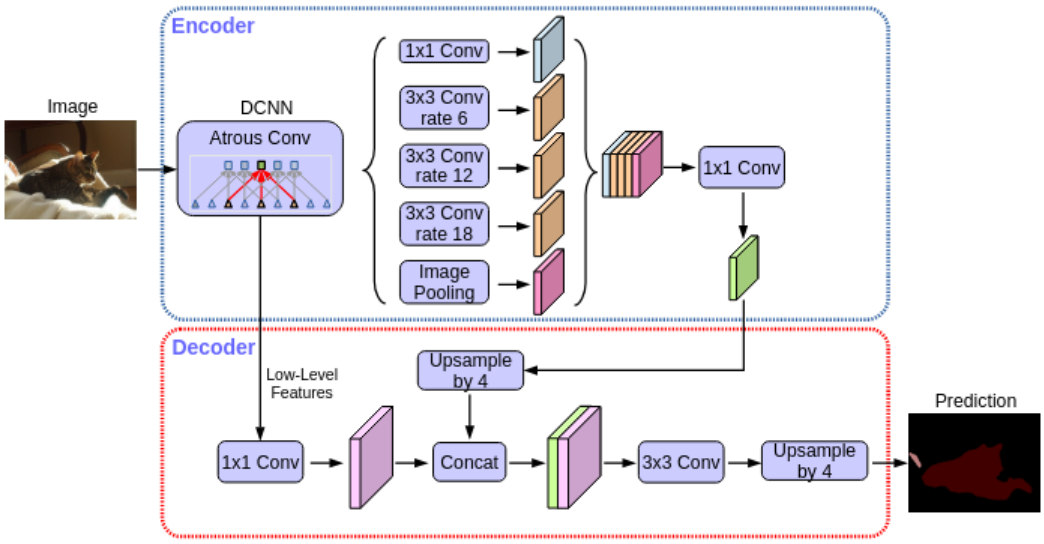
DeepLabV3plus [8]

Το DeepLabV3plus[8] αποτελεί την εξέλιξη της αρχιτεκτονικής DeepLabV3, καθώς την ενσωματώνει σε μία ενιαία αρχιτεκτονική τύπου encoder-decoder. Το δίκτυο DeepLabV3, λοιπόν, αποτελεί τον encoder, ο οποίος με χρήση της δομής Dilated Spatial Pyramid Pooling μπορεί να απομονώνει χαρακτηριστικά σε πολλαπλές κλίμακες. Ο decoder αυτό που κάνει είναι να παίρνει την έξοδο του Dilated Spatial Pyramid Pooling να κάνει μία υπερδειγματοληψία και να ενώνει αυτά τα χαρακτηριστικά με κάποια χαμηλότερου επιπέδου (π.χ

ακμές, καμπύλες κ.λ.π)(low - level features) που έχουν εξαχθεί από τα αρχικά επίπεδα του encoder. Συγκεκριμένα, αυτά είναι χαρακτηριστικά που παράγονται στα αρχικά επίπεδα του backbone ResNet και συνδυάζονται με αυτά που έχει εξάγει ο encoder για να μετριαστεί η απώλεια πληροφορίας σε ότι έχει να κάνει με τα όρια των αντικειμένων προς κατάτμηση. Τα χαμηλού επιπέδου χαρακτηριστικά, επειδή είναι υψηλής διαστατικότητας, έχουν δηλαδή πολλά κανάλια, φιλτράρονται από μία $1x1$ συνέλιξη, ώστε να μειωθεί η διαστατικότητα τους. Αυτό γίνεται, έτσι ώστε να μην υπερκαλύψουν τα υψηλού επιπέδου χαρακτηριστικά(high-level features) που έχουν εξαχθεί από το DSPP. Τέλος, μετά από κάποιες ακόμα συνελίξεις και μία τελική υπερδειγματοληψία (upsampling) παράγεται ο τελικός χάρτης τμηματοποίησης.



Εικόνα 2.10: *Η αρχιτεκτονική DeepLabV3 [6]*



Εικόνα 2.11: *Η αρχιτεκτονική DeepLabV3plus [8]*

2.3 Ημιεπιβλεπόμενη Μάθηση (Semi-Supervised Learning)

Πρόκειται για ένα σύνολο αλγορίθμων μηχανικής μάθησης που συνδυάζουν τα χαρακτηριστικά της επιβλεπομένης (supervised learning) και της μη-επιβλεπόμενης μάθησης (unsupervised learning). Ουσιαστικά, αποτελεί μία υβριδική μέθοδο μάθησης, η οποία αξιοποιεί ταυτόχρονα τόσο δεδομένα με ετικέτα (labeled data), όσο και δεδομένα χωρίς ετικέτα (unlabeled data) κατά την εκπαίδευση του μοντέλου. Στόχος είναι η χρήση των επιπλέον μη επισημασμένων δεδομένων να επιφέρουν βελτίωση στην απόδοση του μοντέλου. Σε ένα πρόβλημα ημιεπιβλεπόμενης μάθησης έχουμε ένα συνολο δεδομένων D, όπου $D = D_l \cup D_u$. Το D_l πρόκειται για το σύνολο των επισημασμένων (labeled) δεδομένων μεγέθους N, έστω $D_l = \{x_{1l}, x_{2l}, ..., x_{lN}\}$ με αντίστοιχες ετικέτες $\Upsilon = \{y_{l1}, y_{l2}, ..., y_{lN}\}$. Επιπλέον, έχουμε το σύνολο των μη επισημασμένων (unlabeled) δεδομένων D_u μεγέθους M, έστω $D_u = \{x_{1u}, x_{2u}, ..., x_{uM}\}$ για τα οποία δεν διαθέτουμε ετικέτες. Γενικά, θέλουμε το πλήθος των μη επισημασμένων δεδομένων να είναι αρκετά μεγαλύτερο από αυτό των επισημασμένων, $N << M$.

Στόχος είναι να αξιοποιηθούν τα μη επισημασμένα δεδομένα, ώστε να παραχθεί ένα μοντέλο f_∂ ($f_\partial : X \rightarrow Y$, X χώρος χαρακτηριστικών, Y χώρος ετικετών), το οποίο θα έχει καλύτερη προβλεπτική ικανότητα από ότι θα είχε εάν χρησιμοποιούσαμε μόνο τα επισημασμένα δεδομένα κατά την εκπαίδευση και θα προσεγγίζει σε απόδοση το μοντέλο που εκπαιδεύεται με όλες τις ετικέτες διαθέσιμες[64]. Για να έχουμε βελτίωση στην απόδοση του μοντέλου σε σχέση με την απλή επιβλεπόμενη εκπαίδευση, χρειάζεται η γνώση πάνω στην $P(x)$(κατανομή δεδομένων) που αποκτάται από την αξιοποίηση των μη επισημασμένων δεδομένων να είναι χρήσιμη κατά την αποτίμηση της κατανομής $P(y|x)$[65, 66]. Με απλά λόγια χρειάζεται τα μη επισημασμένα δεδομένα να προέρχονται από κατανομή σχετική με το πρόβλημα που θέλουμε να λύσουμε.

Η γενική ιδέα πίσω από τις μεθόδους που εξετάζουμε και χρησιμοποιούμε στην παρούσα διπλωματική είναι ότι το μοντέλο εκπαιδεύεται με σκοπό την βελτιστοποίηση μίας συνδυαστικής συνάρτησης κόστους (loss function) που αποτελείται απο δύο όρους [64].

$$L_{total} = L_{labeled} + \eta(t) \cdot L_{unlabeled} \tag{2.6}$$

Ο πρώτος όρος είναι η κλασική συνάρτηση κόστους που χρησιμοποιείται στην επιβλεπόμενη μάθηση (π.χ cross entropy, mean square error) και υπολογίζεται πάνω στα επισημασμένα δεδομένα. Ο δεύτερος όρος αποτελεί τη συνάρτηση κόστους που αξιοποιεί τα μη επισημασμένα δεδομένα και στόχος των αλγορίθμων ημιεπιβλεπόμενης μάθησης είναι να σχεδιάσουν κατάλληλα το συγκεκριμένο κόστος. Η συνεισφορά της $L_{unlabeled}$ στη συνολική ημιεπιβλεπόμενη συνάρτηση κόστους συνήθως καθορίζεται από ένα βάρος $\eta(t)$. Το συγκεκριμένο βάρος μπορεί να είναι είτε μία σταθερά, είτε μία συνάρτηση ράμπας εξαρτώμενη από το βήμα εκπαίδευσης t που προσαρμόζει ανάλογα την επίδραση του όρου $L_{unlabeled}$ κατά τη διάρκεια της εκπαίδευσης.

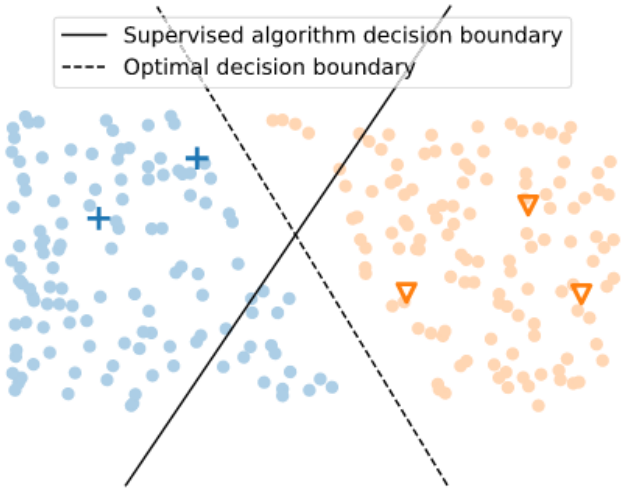
2.3.1 Βασικές υποθέσεις στην ημιεπιβλεπόμενη μάθηση

Για να έχει νόημα η εφαρμογή των αλγορίθμων ημιεπιβλεπόμενης μάθησης χρειάζεται να ληφθούν υπόψην ορισμένες υποθέσεις που αφορούν τη δομή των δεδομένων εκπαίδευσης.

- **Υπόθεση Ομαλότητας** (Smoothness Assumption) [65, 66]: Εάν δύο δείγματα $x1$, $x2$ που βρίσκονται σε μία περιοχή του χώρου, όπου υπάρχει πυκνή συγκέντρωση δεδομένων (high density cluster , high density region) είναι κοντά, τότε κοντά θα πρέπει να είναι και οι αντίστοιχες προβλέψεις $y1$, $y2$ [65, 66]. Πρακτικά αυτό μας λέει ότι αν τα $x1, x2$ είναι κοντά στον χώρο πρέπει να έχουν την ίδια ετικέτα σε ένα πρόβλημα κατηγοριοποίησης, δηλαδή τα $y1, y2$ να είναι ίδια.

- **Υπόθεση Συστάδας** (Cluster Assumption) [65, 66]: Αν δύο δείγματα βρίσκονται στην ίδια συστάδα (cluster) σε έναν χώρο, τότε πιθανώς να ανήκουν στην ίδια κλάση [65, 66]. Ουσιαστικά αυτό αποτελεί μία ειδική περίπτωση της υπόθεσης ομαλότητας. Θεωρούμε ότι τα δείγματα είσοδου σχηματίζουν συστάδες στον χώρο και κάθε συστάδα αντιπροσωπεύει μία κλάση[64].

- **Υπόθεση διαχωρισμού χαμηλής πυκνότητας** (Low - Density Separation Assumption): Εδώ υποθέτουμε ότι το όριο απόφασης (decision boundary) που παράγεται από κάποιον ταξινομητή πρέπει να βρίσκεται σε μία περιοχή χαμηλής πυκνότητας του χώρου, δηλαδή σε περιοχές που δεν έχουν σχηματιστεί συστάδες (clusters) δεδομένων. Έδω έχουμε άμεση συσχέτιση με την υπόθεση συστάδας, καθώς εάν ένα όριο απόφασης διερχόταν από μία περιοχή υψηλής πυκνότητας (high -density region), δηλαδή από μία συστάδα, θα τη διαμέριζε σε δύο ξεχωριστές κλάσεις με αποτέλεσμα να υπήρχαν δείγματα της ίδιας συστάδας τα οποία θα είχαν διαφορετική κλάση και θα είχαμε έτσι παραβίαση της υπόθεσης συστάδας [64].

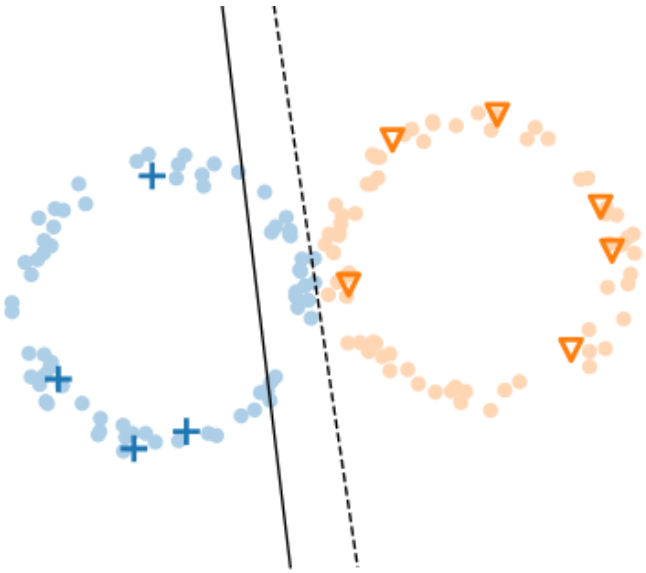


Εικόνα 2.12: *Υποθέσεις ομαλότητας (smoothness) και χαμηλής πυκνότητας (low-density)[9]*

Στο παραπάνω σχήμα οι τελείες με το αχνό χρώμα αντιπροσωπεύουν τα μη επισημασμένα και οι σταυροί με τα τρίγωνα τα επισημασμένα δεδομένα για ένα πρόβλημα δυαδικής κατηγοριοποίησης. Η μη διακεκομμένη γραμμή απεικονίζει ένα βέλτιστο όριο

απόφασης που μπορεί να έβρισκε ένας επιβλεπόμενος αλγόριθμος αξιοποιώντας μόνο τα επισημασμένα δεδομένα. Από την άλλη η διακεκομμένη γραμμή αντιπροσωπεύει ένα όριο αποφάσης κάποιου ημιεπιβλεπόμενου αλγορίθμου, ο οποίος κάνει χρήση και των μη επισημασμένων δεδομένων, ωθώντας το όριο απόφασης σε περιοχές χαμηλής πυκνότητας, βασιζόμενος στις υποθέσεις ομαλότητας (smoothness assumption), συστάδας (cluster assumption) και διαχωρισμού χαμηλής πυκνότητας (low-density separation).

- **Υπόθεση Manifold** (Manifold Assumption): Γενικά στα πρόβληματα μηχανικής μάθησης του πραγματικού κόσμου τα δεδομένα εκπαίδευσης που καλούμαστε να διαχειριστούμε είναι υψηλής διαστατικότητας. Για παράδειγμα μία εικόνα *RGB* χωρικών διαστάσεων $(256, 256)$ ανήκει στον χώρο $R^{256x256x3}$, ο οποίος έχει διάσταση $256x256x3 = 196608$. Σύμφωνα με αυτή την υπόθεση τα δείγματα που βρίσκονται σε έναν χώρο υψηλής διαστατικότητας (high-dimensional) συνήθως συγκεντρώνονται σε υποδομές (substructures), οι οποίες είναι μικρότερης διαστατικότητας (lower-dimension). Αυτές, οι υποδομές είναι τοπολογικοί χώροι που ονομάζονται manifolds[9]. Στην περίπτωση της ημιεπιβλεπόμενης μάθησης, η υπόθεση αναφέρει ότι ο χώρος εισόδου μπορεί να αποτελείται από πολλαπλά χαμηλότερης διαστατικότητας manifolds στα οποία βρίσκονται τα δεδομένα και ότι τα δεδομένα που βρίσκονται στο ίδιο manifold ανήκουν στην ίδια κλάση (έχουν ίδιο label) [9]. Επομένως, εάν καθορίσουμε το πλήθος των manifolds που απεικονίζονται τα δεδομένα, καθώς και ποια δείγματα ανήκουν στο κάθε manifold θα μπορέσουμε να συμπεράνουμε τις κλάσεις που ανήκουν τα μη επισημασμένα δεδομένα με βάση τα επισημασμένα με τα οποία συνυπάρχουν στο ίδιο manifold[9].



Εικόνα 2.13: *Υπόθεση Manifold[9]*

Στην παραπάνω εικόνα βλέπουμε με τη διακεκομμένη γραμμή το βέλτιστο όριο απόφασης που βρίσκει ο ημιεπιβλεπόμενος αλγόριθμος. Τα δεδομένα απεικονίζονται σε 2 δισδιάστατα manifolds και η κατηγοριοποίηση των μη επισημασμένων δειγμάτων

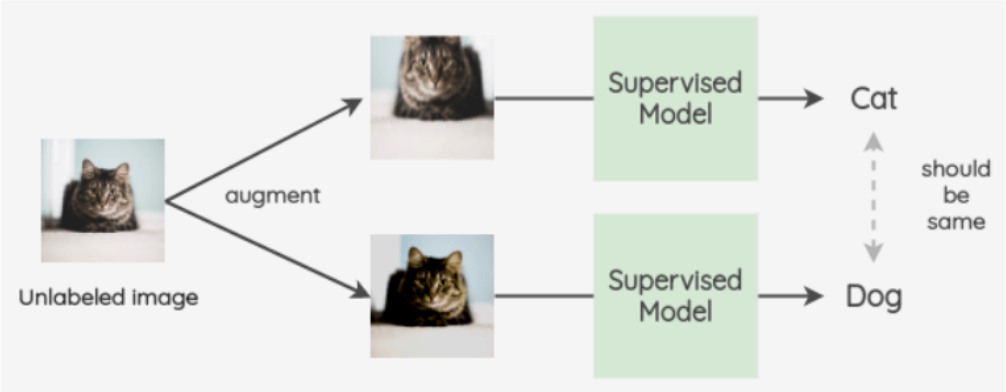
γίνεται με βάση τα επισημασμένα δείγματα που υπάρχουν στο αντίστοιχο manifold.

2.3.2 Κανονικοποίηση Συνέπειας (Consistency Regularization)

Μία πολύ διαδεδομένη προσέγγιση στην ημιεπιβλεπόμενη μάθηση είναι η κανονικοποίηση συνέπειας (consistency regularization). Η γενική ιδέα στην οποία βασίζονται όλοι οι αλγόριθμοι αυτής της κατηγορίας είναι ότι οι μικρές μεταβολές ή διαταραχές (perturbations) που μπορούν να εφαρμοστούν πάνω σε κάποιο μη επισημασμένο δείγμα εισόδου (εφαρμογή κάποιου μετασχηματισμού επαύξησης δεδομένων, προσθήκη θορύβου γκαουσιανού σε μια εικόνα κ.λ.π) δεν θα επηρεάσουν σε μεγάλο βαθμό την τελική πρόβλεψη του μοντέλου[64]. Όπως έχουμε αναφέρει και στις υποθέσεις συστάδας και διαχωρισμού χαμηλής πυκνότητας, τα όρια απόφασης βρίσκονται σε χαμηλής πυκνότητας περιοχές και διαχωρίζουν μεταξύ τους τις συστάδες που αποτελούν τις υψηλής πυκνότητας περιοχές και αντιπροσωπεύουν τις διαφορετικές κλάσεις. Συνεπώς, κάτω από την εφαρμογή κάποιας μικρής διαταραχής (perturbation) θεωρούμε ότι δεν θα πρέπει να παραβιάζονται οι παραπάνω υποθέσεις και η πιθανότητα το διαταραγμένο (perturbed) δείγμα να ανήκει σε διαφορετική κλάση θα πρέπει να είναι μικρή[64]. Αξιοποιώντας, λοιπόν, τα μη επισημασμένα δεδομένα κατά την εκπαίδευση, στόχος είναι το μοντέλο να έχει τη δυνατότητα να παράγει κοντινές προβλέψεις για κάποιο δείγμα και τη διαταραγμένη (perturbed) εκδοχή αυτού, δηλαδή συνεπείς προβλέψεις για παρόμοια δείγματα εισόδου.

Έν γένει, όπως θα δούμε και στη συνέχεια διαταραχές (perturbations) εκτός από τα δείγματα εισόδου μπορούμε να εφαρμόσουμε τόσο σε επίπεδο δικτύου (network perturbation)[18], όταν αναφερόμαστε σε νευρωνικά δίκτυα, καθώς και σε επίπεδο ενδιάμεσων χαρακτηριστικών που μπορεί να παράγει κάποιο δίκτυο (feature perturbation)[17].

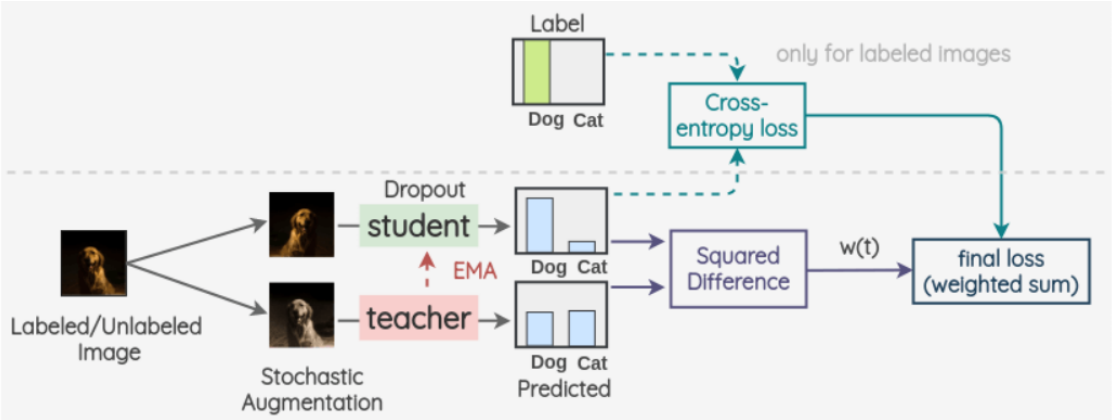
Εάν υποθέσουμε ότι έχουμε κάποιο μη επισημασμένο δείγμα u_1 και τη διαταραγμένη εκδοχή αυτού u_{1per} και τα τροφοδοτήσουμε σε ένα δίκτυο $f(\partial)$, τότε ο στόχος θα είναι το δίκτυο να παράξει συνεπείς προβλέψεις ελαχιστοποιώντας την απόσταση αυτών, δηλαδή την ποσότητα $D(f_\partial(u_1), f_\partial(u_{1per}))$. Οπότε ο δεύτερος όρος της 2.6 που αξιοποιεί τα μη επισημασμένα δεδομένα συνήθως είναι μία συνάρτηση που αξιολογεί απόσταση δύο κατανομών (π.χ Mean Square Error, Kullback-Leiber divergence, Jensen-Shannon divergence, cross-entropy) [64] .



Εικόνα 2.14: *Η γενική ιδέα της κανονικοποίησης συνέπειας (consistency regularization) σε ένα πρόβλημα κατηγοριοποίησης [10]*

Το μοντέλο Mean Teacher [67] Στις μεθόδους κανονικοποίησης συνέπειας, όπως είδαμε εφαρμόζουμε διαταραχές στα μη επισημασμένα δεδομένα και εν συνεχεία τα τροφοδοτούμε στο δίκτυο, απαιτώντας ίδιες εξόδους μεταξύ του δείγματος και της διαταραγμένης εκδοχής αυτού. Το μοντέλο που χρησιμοποιείται σε αυτές τις περιπτώσεις είναι αυτό του Student - Teacher. Ουσιαστικά, πρόκειται για δύο δίκτυα που στις περισσότερες φορές έχουν κοινές παραμέτρους θ και στόχος είναι να υπάρχει συνέπεια στις εξόδους του student με τον teacher για τις διαταραγμένες (perturbed) εισόδους. Ομοίως, το μοντέλο Mean Teacher [67] αποτελείται από τα δίκτυα Student και Teacher τα οποία έχουν την ίδια αρχιτεκτονική με τη διαφορά ότι τα βάρη του teacher είναι ο εκθετικός κινητός μέσος όρος (exponential moving average a.k.a EMA)[68] των βαρών του student. Ουσιαστικά τα βάρη του teacher αποτελούν τον μέσο όρο των βαρών του student πάνω σε όλα τα βήματα εκπαίδευσης (training steps)[67], για αυτό και έχει την ονομασία mean teacher. Δηλαδή, τα βάρη του teacher στο βήμα εκπαίδευσης t δίνονται από τη σχέση $\theta_t^{teacher} = \lambda * \theta_{t-1}^{teacher} + (1-\lambda) * \theta_t^{student}$ [67]. Το λ αποτελεί συντελεστή εξομάλυνσης και είναι υπερπαράμετρος. Σύμφωνα με την αρχική δημοσίευση[67], ο μέσος όρος των βαρών πάνω σε όλα τα βήματα εκπαίδευσης μπορεί να παράξει πιό ακριβή μοντέλα teachers σε σχέση με το αν χρησιμοποιούνταν απλά τα τελικά βάρη του student[67].

Η μεθοδολογία έδω είναι ότι για κάθε δείγμα επισημασμένο και μη επισημασμένο παράγονται τυχαία 2 επαυξημένες εκδοχές (input perturbation). Η πρώτη εκδοχή τροφοδοτείται στο δίκτυο student και η δεύτερη στον teacher. Το κόστος συνέπειας (consistency loss) που χρησιμοποιείται ειναι το *MSE* για τις προβλέψεις των δικτύων student και teacher, καθώς θέλουμε να ωθήσουμε τα δίκτυα να παράγουν ίδιες εξόδους. Για την περίπτωση των επισημασμένων δεδομένων γίνεται υπολογισμός και του κόστους cross entropy με βάση τις διαθέσιμες ετικέτες. Το τελικό κόστος αποτελείται απο τη συνάρτηση απωλειών cross entropy και τον όρο κανονικοποίησης consistency loss σύμφωνα με την γενική σχέση 2.6.[10]



Εικόνα 2.15: *Η μέθοδος Mean Teacher [10]*

Η γενική συνάρτηση κόστους έδω μπορεί να εκραστεί ως εξής [64]:

$$L = \frac{1}{|D_l|} \sum_{(x,y)\in D_l} CE(y, f_{\theta^{student}}(x)) + w(t) * \frac{1}{|D_u|} \sum_{x\in D_u} d_{MSE}(f_{\theta^{student}}(x)), f_{\theta^{teacher}}(x_{perturbed})) \quad (2.7)$$

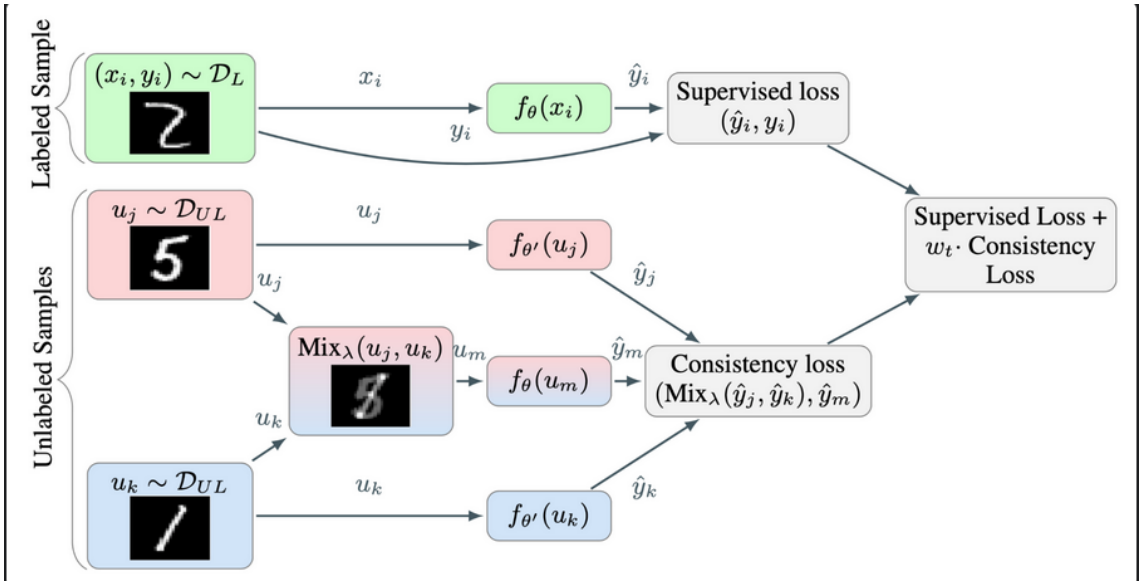
Το παράδειγμα εκπαίδευσης συνέπειας Interpolation Consistency Training (ICT)[11] Ας υποθέσουμε ότι έχουμε μία πράξη παρεμβολής (interpolation) δύο δειγμάτων όπως η MixUp [69], η οποία εκφράζεται ως εξής: $Mix(a,b) = \lambda * a + (1-\lambda) * b$, όπου βάρος $\lambda \sim Beta(a,a)$ για $a \in [0,\infty]$ [64]. Τότε, στόχος του ICT είναι να εκπαιδεύσει το μοντέλο να παραμένει συνεπές κάτω από πράξεις παρεμβολής μεταξύ των μη επισημασμένων δειγμάτων. Δηλαδή, για ένα δίκτυο student f_θ και για τον αντίστοιχο *EMA* teacher $f_{\theta'}$ απαιτείται συνέπεια μεταξύ των ποσοτήτων $f_\theta(Mix_\lambda(u_1,u_2))$ και $Mix_\lambda(f_{\theta'}(u_1),f_{\theta'}(u2))$, όπου u_1,u_2 δύο τυχαία μη επισημασμένα δείγματα [11].

Σύμφωνα με το [11] τέτοιου είδους πράξεις μίξης δειγμάτων μπορούν να αξιοποιηθούν ως διαταραχές για ζεύγη μη επισημασμένων δειγμάτων. Γενικά, τα δείγματα που βρίσκονται σε χαμηλής πυκνότητας περιοχές, δηλαδή κοντά στα όρια απόφασης είναι κατάλληλα για να εφαρμοστεί σε αυτά κανονικοποίηση συνέπειας. Αυτό ισχύει, καθώς εάν θεωρήσουμε ένα δείγμα u_1, το οποίο υποθέτουμε ότι βρίσκεται κοντά στο όριο απόφασης τότε εάν δημιουργήσουμε μία διαταραγμένη εκδοχη αυτού, έστω $u_1+\delta$, τότε είναι πιθανό το $u_1+\delta$ να περάσει από την άλλη πλευρά του ορίου απόφασης παραβιάζοντας τη συνθήκη διαχωρισμού χαμηλής πυκνότητας. Από την άλλη σε υψηλής πυκνότητας περιοχές μία τέτοια παραβίαση δεν θα συνέβαινε, καθώς τα δείγματα βρίσκονται μακρία απο το όριο απόφασης και μικρές διαταραχές δεν θα οδηγούσαν πέρα από αυτό[11]. Γενικά, θέλουμε μια διαταραχή $u_1+\delta$ κάποιου δείγματος u_1 που βρίσκεται κοντά στο όριο απόφασης, ώστε το $u_1+\delta$ να βρίσκεται από την αντίθετη πλευρά του ορίου απόφασης[11]. Η ιδέα απλά να χρησιμοποιούμε τυχαίες διαταραχές δεν είναι αποδοτική, καθώς το υποσύνολο των κατευθύνσεων που οδηγούν κοντά στο όριο απόφασης είναι ένα εξαιρετικά μικρό ποσοστό του συνολικού χώρου[11]. Αυτό που προτείνεται είναι να θεωρησουμε μία διαταραχή(perturbation) ενός δείγματος u_1, την πράξη Mix με κάποιο άλλο δείγμα u_2. Δηλαδή, $u_1+\delta = Mix_\lambda(u_1,u_2)$. Για προβλήματα που έχουν μεγάλο αριθμό κλάσεων και παρόμοια κατανομή μεταξύ των δειγμάτων της κάθε κλάσης είναι πολύ πιθανό τα δείγματα $(u1,u2)$ να βρίσκονται σε διαφορετική συστάδα και να ανήκουν σε διαφορετική κλάση[64]. Αν υποθέσουμε, επίσης, και ότι ένα εκ των 2 δειγμάτων(π.χ το u_1) βρίσκεται κόντα και σε μία χαμηλής πυκνότητας περιοχή (κατάλληλο δείγμα για εφαρμοστεί κανονικοποίηση συνέπειας), τότε το να πραγματοποιήσουμε μία πράξη μίξης π.χ με το u_2 θα οδηγήσει το όριο απόφασης σε μια χαμηλής πυκνότητας περιοχή. Συνεπώς, οι πράξεις παρεμβολής (interpolation) αποτελούν καλές διαταραχές (perturbations) για την κανονικοποίηση συνέπειας (consistency regularization)[11].

Το γενικό παράδειγμα εκπαίδευσης του ICT αποτελείται από δύο κλάδους (branches), ενώ και εδώ χρησιμοποιείται το μοντέλο Mean Teacher. Ο πρώτος κλάδος αξιοποιεί τα επισημασμένα δεδομένα. Συγκεκριμένα, αυτά τροφοδοτούνται στο δίκτυο student $f_{\theta^{student}}$ και υπολογίζεται το cross-entropy κόστος με βάση τις διαθέσιμες ετικέτες. Στον δεύτερο κλάδο που αξιοποιούνται τα μη επισημασμένα δεδομένα, αρχικά επιλέγονται τυχαία 2 δείγματα τα οποία γίνονται Mix και το παραγόμενο δείγμα τροφοδοτείται στο δίκτυο student. Παράλληλα, τα 2 αρχικά δείγματα δίνονται ως είσοδο στο δίκτυο teacher $f_{\theta^{ema-teacher}}$, το οποίο παράγει τις αντίστοιχές εξόδους, οι οποίες και αυτές γίνονται Mix. Τέλος, απαιτείται συνέπεια μεταξύ της εξόδου του student και της εξόδου που προέκυψε από την ανάμειξη των δύο εξόδων του teacher. Και σε αυτή την περίπτωση το τελικό κόστος αποτελείται από την cross-entropy και το κόστος συνέπειας μεταξύ student-teacher.

Μαθηματικά το συνολικό κόστος για το ICT μπορεί να εκφραστεί ως εξής [11] [64]:

$$L = \frac{1}{|D_l|} \sum_{(x,y)\in D_l} CE(y, f_{\partial^{student}}(x))$$
$$+ w(t) * \frac{1}{|D_u|} \sum_{(u1,u2)\in D_u} d_{MSE}(f_{\partial^{student}}(Mix_{\lambda}(u_1, u_2)), Mix_{\lambda}(f_{\partial^{ema-teacher}}(u_1), f_{\partial^{ema-teacher}}(u_2)) \quad (2.8)$$



Εικόνα 2.16: *Η μέθοδος Interpolation Consistency Training [11]*

2.3.3 Ψευδοετικέτες (Pseudolabels)

Η χρησιμοποίηση των προβλέψεων του μοντέλου πάνω στα μη επισημασμένα δεδομένα κατά τη διάρκεια της εκπαίδευσης ονομάζεται pseudo-labeling[70, 71]. Ουσιαστικά, αυτό που γίνεται είναι η κατανομή που παράγει η συνάρτηση softmax στην έξοδο του δικτύου να περνάει από μία πράξη argmax και εν συνεχεία να παράγεται ένας δείκτης (index) που υποδηλώνει την κλάση. Αυτό αποτελεί την ψευδοετικέτα (pseudolabel). Ακόμη μπορεί να εκφραστεί και ως one-hot διάνυσμα.

Για παράδειγμα για ένα πρόβλημα κατηγοριοποίησης C κλάσεων, έστω ότι έχουμε ένα μη επισημασμένο δείγμα u και το τροφοδοτούμε στο δίκτυο, το οποίο παράγει μία κατανομή softmax για αυτό το δείγμα.

$$y_{modelprediction} = f_{\partial}(u) \in R^C$$
$$\sum_C y_{modelprediction_C} = 1$$
$$y_{pseudo} = y_{argmax} = argmax(y_{modelprediction}) \in \{1, 2, 3, ..., C\}$$

Το y_{argmax} ουσιαστικά αποτελεί την ψευδοετικέτα, καθώς δηλώνει την κλάση με τη μεγαλύτε-

ρη βεβαιότητα (confidence), σύμφωνα με την πρόβλεψη του μοντέλου. Όπως, αναφέραμε η ψευδοετικέτα μπορεί να εκφραστεί και ώς one-hot διάνυσμα, δηλαδή ώς ένα διάνυσμα με μηδενικά σε όλες τις θέσεις εκτός από τη θέση y_{argmax}.

Παρόμοια για ένα πρόβλημα σημασιολογικής κατάτμησης C κλάσεων θα έχουμε:

$$y_{modelprediction} = f_{\partial}(u) \in R^{HxWxC}$$
$$y_{pseudo} = y_{argmax} = argmax(y_{modelprediction}) \in R^{HxW}$$

Όπως έχουμε αναφέρει παράγεται ένας χάρτης τμηματοποιήσης (segmentation map), όπου σε κάθε pixel της αρχικής εικόνας έχει ανατεθεί μία κλάση. Η χρήση ψευδοετικετών κατά την εκπαίδευση του μοντέλου μπορεί να θεωρηθεί και μία μορφή ελαχιστοποίησης εντροπίας (Entropy Minimization) [72], καθώς το μοντέλο ενθαρρύνεται να παράγει προβλέψεις υψηλής βεβαιότητας (high-confidence) για τα μη επισημασμένα δεδομένα, ελαχιστοποιώντας την εντροπία των πιθανοτήτων των κλάσεων, ωθώντας το όριο απόφασης σε χαμηλής πυκνότητας (low-density) περιοχές. Αυτο συμβαίνει, καθώς οι προβέψεις του μοντέλου για τη βεβαιότητα της κάθε κλάσης αποτελούν μέτρο που δηλώνει την επικάλυψη των κλάσεων (class overlap). Το να μειωθεί η εντροπία αυτών των προβλέψεων(δηλαδή να παράγονται υψηλής βεβαιότητας προβλέψεις) είναι ισοδύναμο με τη μείωση της επικάλυψης των κλάσεων και συνεπώς της ικανοποίησης των υποθέσεων συστάδας (cluster) και χαμηλής πυκνότητας διαχωρισμού (low-denisty separation)[73, 70].

Μία συνήθης και απλή τεχνική που αξιοποιεί ψευδοετικέτες είναι το μοντέλο να εκπαιδεύεται πρώτα στα επισημασμένα δεδομένα και εν συνεχεία να χρησημοποιείται για να παράξει ψευδοετικέτες για τα μη επισημασμένα δεδομένα. Έπειτα, τα αρχικά επισημασμένα δεδομένα εμπλουτίζονται με τις παραγόμενες ψευδοετικέτες και όλα μαζί χρησιμοποιούνται εκ νέου για την περαιτέρω εκπαίδευση του μοντέλου με τον συνήθη επιβλεπόμενο τρόπο.

Μία διαφορετική και πιό σύγχρονη προσέγγιση που έχει επιφέρει αρκετά state-of-the art αποτελέσματα είναι ο συνδυασμός της μεθόδου των ψευδοετικετών με την κανονικοποίηση συνέπειας που έχουμε αναφέρει. Τέτοιου είδους αλγόριθμοι σε κάθε βήμα εκπαίδευσης αξιοποιούν παράλληλα τόσο τα επισημασμένα, όσο και τα μη επισημασμένα δεδομένα, κάνοντας χρήση της γενικής συνάρτησης κόστους 2.6.

Ας πάρουμε ως παράδειγμα το μοντέλο Mean Teacher. Προκείμένου να αξιοποιηθούν οι ψευδοετικέτες ο όρος του κόστους συνέπειας αποτελείται από το κόστος cross-entropy μεταξύ των προβλέψεων του student για τα διαταραγμένα (perturbed) μη επισημασμένα δείγματα και των ψευδοετικετών που παράγονται από τον teacher για την αρχική εκδοχή των μη επισημασμένων δειγμάτων. Γενικά, η λογική είναι ότι το δίκτυο teacher παράγει ψευδοετικέτες και ο student μαθαίνει αξιοποιώντας αυτές κατά την εκπαίδευση[73]. Κάνουμε χρήση ορισμένων συμβολισμών από το [64]. Η συνάρτηση κόστους μπορεί να τώρα να γραφτεί ως:

$$L = \frac{1}{|D_l|} \sum_{(x,y)\in D_l} CE(y, f_{\partial^{student}}(x)) + w(t) * \frac{1}{|D_u|} \sum_{x\in D_u} CE(y_{pseudo}, f_{\partial^{student}}(x_{perturbed})) \quad (2.9)$$

Αντίστοιχα για το ICT θα έχουμε:

$$L = \frac{1}{|D_l|} \sum_{(x,y)\in D_l} CE(y, f_{\partial^{student}}(x)) + w(t) * \frac{1}{|D_u|} \sum_{(u1,u2)\in D_u} CE(Mix_{\lambda}(y_{pseudo1}, y_{pseudo2}), f_{\partial^{student}}(Mix_{\lambda}(u_1, u_2)) \quad (2.10)$$

Κεφάλαιο 3

Σχετική βιβλιογραφία μεθόδων κανονικοποίησης συνέπειας για την ημιεπιβλεπόμενη κατάτμηση εικόνας

Η ημιεπιβλεπόμενη σημασιολογική κατάτμηση εικόνας (semi-supervised semantic segmentation) αποτελεί ένα πολύ σύγχρονο ερευνητικό πεδίο, το οποίο τα τελευταία χρόνια έχει γνωρίσει σημαντική πρόοδο, ειδικά με την εισαγωγή και την ευρεία χρησιμοποίηση τεχνικών βαθιάς μάθησης για την επίλυση του προβλήματος. Σε αυτό το κεφάλαιο θα παρουσιάσουμε κάποιες βασικές σχετικές εργασίες που στηρίζονται στην ιδέα της κανονικοποίησης συνέπειας (consistency regularization), τις οποίες θα ομαδοποιήσουμε ανάλογα με τη γενική μεθοδολογία την οποία ακολουθούν.

3.1 Σχετικές εργασίες

Οι μέθοδοι που βασίζονται στην κανονικοποίηση συνέπειας (consistency regularization) έχουν ως κοινό στοιχείο, όπως έχουμε αναφέρει, την αξιοποίηση διαταραχών (perturbations), οι οποίες μπορούν να εφαρμοστούν στα παρακάτω επίπεδα.

- **Διαταραχές σε επίπεδο εισόδου (input-level)**: Προσθήκη γκαουσιανού θορύβου σε μία εικόνα, εφαρμογή άλλων μετασχηματισμών επαύξησης δεδομένων σε εικόνες εισόδου(data augmentation).
- **Διαταραχές σε επίπεδο ενδιάμεσων χαρακτηριστικών (feature-level)**:Προσθήκη θορύβου σε κάποια ενδιάμεσα παραγόμενα χαρακτηριστικά από το δίκτυο, εφαρμογή dropout σε κάποια χαρακτηριστικά.
- **Διαταραχές σε επίπεδο δικτύου (network-level)**: Αρχικοποίηση του ίδιου δικτύου με διαφορετικές παραμέτρους.

3.1.1 Προσεγγίσεις βασισμένες σε διαταραχές επιπέδου εισόδου (input-level perturbations)

Η βασική ιδεά για όλες τις μεθόδους που αξιοποιούν διαταραχές σε επίπεδο εισόδου προέρχεται απο την ημιεπιβλεπόμενη κατηγοριοποιήση (semi-supervised classification) και είναι η δουλειά των Kihyuk Sohn et al. [13]. Η συγκεκριμένη μέθοδος που ονομάζεται

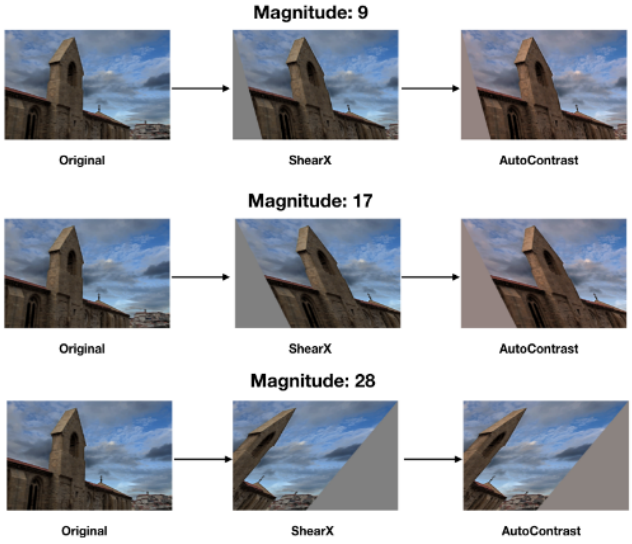
FixMatch συνδυάζει τη μέθοδο της κανονικοποίησης συνέπειας με την τεχνική των ψευδοετικετών.

Οι συγγραφείς εισάγουν τις έννοιες δύο ειδών επαύξησης δεδομένων (data augmentation).

- **Ασθενής επαύξηση δεδομένων (Weak data augmentation)**
- **Ισχυρή επαύξηση δεδομένων (Strong data augmentation)**

Η ασθενής επαύξηση αποτελείται από μετασχηματισμούς οριζόντιας ή κάθετης αναστροφής (horizontal flipping, vertical flipping) μίας εικόνας, καθώς και μετασχηματισμούς μετατόπισης (shifting transforms) κατά τον οριζόντιο ή κατακόρυφο άξονα. Από την άλλη η ισχυρή επαύξηση αναφέρεται ότι περιλαμβάνει μετασχηματισμούς αντίθεσης (contrast), φωτεινότητας(brightness), χρώματος (color) κ.λ.π μίας εικόνας εισόδου. Συγκεκριμένα, για την παραγωγή ισχυρά επαυξημένων εικόνων οι τεχνικές που χρησιμοποιούνται είναι η RandAugment[12] και η CTAugment [74], οι οποίες επιλέγουν τυχαία ένα πλήθος μετασχηματισμών που θα εφαρμοστούν διαδοχικά.

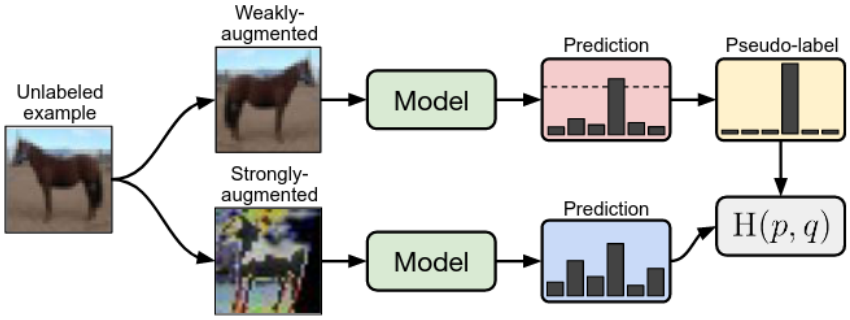
Η ισχύς των μετασχηματισμών αυτών καθορίζεται από μία παράμετρο M, η οποία στο RandAugment [12] ονομάζεται magnitude. Όπως είναι εμφανές και στο παρακάτω παράδειγμα από το RandAugment, όσο αυξάνεται η τιμή του magnitude, τόσο πιο έντονος γίνεται ο μετασχηματισμός.



Εικόνα 3.1: *Παράδειγμα της επίδρασης του magnitude στη περίπτωση της ισχυρής επαύξησης[12]*

Η ιδέα, λοιπόν, του FixMatch είναι ότι για κάποιο μη επισημασμένο δείγμα παράγονται 2 επαυξημένες εκδοχές αυτού, η ασθενώς επαυξημένη και η ισχυρά επαυξημένη. Η ασθενώς επαυξημένη εκδοχή τροφοδοτείται στο δίκτυο, το οποίο παράγει μια πρόβλεψη για αυτή. Η συγκεκριμένη πρόβλεψη φιλτράρεται με ένα κατώφλι (threshold), και μετατρέπεται σε ψευδοετικέτα. Με αυτόν τον τρόπο μόνο οι υψηλής βεβαιότητας (high-confidence) προβλέψεις του δικτύου για την ασθενώς επαυξημένη εικόνα μετατρέπονται σε ψευδοετικέτες. Έν συνεχεία, στο δίκτυο τροφοδοτείται η ισχυρά επαυξημένη εκδοχή για την οποία

παράγεται η αντίστοιχη πρόβλεψη. Η ψευδοετικέτα που παράχθηκε για την ασθενώς επαυξημένη εκδοχή χρησιμοποιείται για να επιβλέψει, μέσω του κόστους cross-entropy την έξοδο του δικτύου για την ισχυρά επαυξημένη εκδοχή. Διαισθητικά, μπορούμε να πούμε ότι το δίκτυο είναι σε θέση να παράγει ικανοποιητικές ψευδοετικέτες για τις ασθενώς επαυξημένες εικόνες, αλλά δυσκολεύεται να παράξει ψευδοετικέτες για τις ισχυρά επαυξημένες εκδοχές αυτών. Συνεπώς, καλείται να μάθει να παράγει ικανοποιητικές προβλέψεις και για τις ισχυρά επαυξημένες εισόδους με επίβλεψη μέσω των καλής ποιότητας ψευδοετικετών που έχουν παραχθεί για την ασθενώς επαυξημένη είσοδο. Η συγκεκριμένη μέθοδος αποτελεί ένα παράδειγμα εκπαίδευσης που ονομάζεται ασθενής-ισχυρή κανονικοποίηση συνέπειας (weak-to-strong consistency regularization).



Εικόνα 3.2: *Το γενικό pipeline του FixMatch [13]*

Η ιδέα του FixMatch μπορεί να επεκταθεί και στο πρόβλημα της κατάτμησης εικόνας (image segmentation). Μία από τις δουλειές που ασχολείται με αυτό, είναι η δημοσίευση των Lihe Yang et al.[14], η οποία εμβαθύνει στο FixMatch για την περίπτωση της σημασιολογικής κατάτμησης, καθώς και σε επεκτάσεις αυτού, όπως θα δούμε και στη συνέχεια.

Όπως έχουμε αναφέρει, χρησιμοποιείται η ιδέα της ασθενούς-ισχυρής συνέπειας για την αξιοποίηση των μη επισημασμένων δεδομένων. Η γενική συνάρτηση κόστους ουσιαστικά ακολουθεί το παράδειγμα της 2.6. Ο επιβλεπόμενος όρος αποτελείται από το κόστος cross-entropy για τα επισημασμένα δεδομένα , ενώ ο μη επιβλεπόμενος όρος μπορεί να αποτελείται εν γένει από μία συνάρτηση κόστους που μετράει τη μεταξύ απόσταση των κατανομών εξόδου του ασθενούς(weak) και ισχυρού(strong) κλάδου αντίστοιχα. Σύμφωνα, με το [14] έχουμε τα εξής:

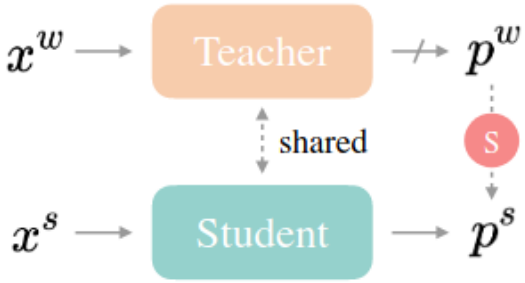
- A^{weak}, A^{strong}: Συναρτήσεις ασθενούς και ισχυρής επαύξησης δεδομένων για μία εικόνα εισόδου.
- Δίκτυα $F^{'}$ και F: teacher και student αντίστοιχα. Στην περίπτωση του FixMatch τα δίκτυα teacher και student είναι ακριβώς ίδια (αρχιτεκτονική και παράμετροι).
- $p^{weak} = F^{'}(A^{weak}(x_u))$: έξοδος του δικτύου teacher για την ασθενώς επαυξημένη εικόνα.
- $p^{strong} = F(A^{strong}(A^{weak}(x_u)))$: έξοδος του δικτύου student για την ισχυρά επαυξημένη εικόνα. Αυτό που βλέπουμε εδώ είναι η ισχυρή επαύξηση εφαρμόζεται πάνω από την ασθενή (on the top of).

Ο ημιεπιβλεπόμενος όρος της συνάρτησης κόστους για μία δέσμη (batch) B_u μη επισημασμένων δεδομένων μπορεί να εκφραστεί ως [14]:

$$L_u = \frac{1}{|B_u|} \sum_{u \in B_u} \mathbf{1}[(max(p^{weak} \geq \tau_{thres})] H(p^{weak}, p^{strong}) \tag{3.1}$$

, όπου H συνάρτηση εντροπίας. Συνήθως, πρόκειται για το κόστος cross-entropy.

Βλέπουμε ότι, γίνεται χρήση ενός κατωφλίου (threshold), το οποίο φιλτράρει τις ψευδοετικέτες (τους χάρτες τμηματοποίησης) που παράγει ο teacher για τις ασθενώς επαυξημένες εικόνες. Με αυτό τον τρόπο εξασφαλίζεται ότι στον υπολογισμό του κόστους λαμβάνονται υπόψην μόνο τα pixels για τα οποία ο teacher έχει παράξει πρόβλεψη με υψηλή βεβαιότητα. Επομένως, τυχόν θορυβώδεις προβλέψεις δεν επηρεάζουν την εκπαίδευση του μοντέλου.



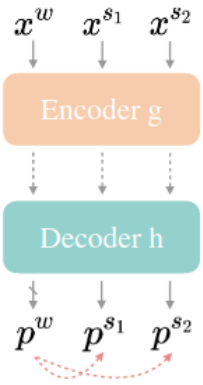
Εικόνα 3.3: *FixMatch όπως παρουσιάζεται στο [14]*

Βλέπουμε, λοιπόν, ότι στον teacher δίνεται η ασθενώς επαυξημένη εικόνα και παράγεται η αντίστοιχη ψευδοετικέτα, η οποία χρησιμοποιείται για την επίβλεψη της εξόδου του student για την ισχυρά επαυξημένη εικόνα. Να τονίσουμε επίσης ότι, ο αλγόριθμος οπισθοδιάδοσης σφάλματος (backpropagation)[75] που χρησιμοποείται για την εκπαίδευση των νευρωνικών δικτύων εφαρμόζεται μόνο κατά μήκος του δικτύου student (strong augmented κλάδος) που είναι άλλωστε και το δίκτυο που καλείται να μάθει να παράγει σωστές προβλέψεις για τις ισχυρά επαυξημένες εικόνες μέσω της επίβλεψης από τις ψευδοετικέτες που παρέχονται από τον teacher. Έν συνεχεία οι παράμετροι του student ανατίθονται στο δίκτυο teacher κατα μήκος του οποίου, όπως είπαμε δεν εκτελείται οπισθοδιάδοση σφάλματος.

Επίπλέον, μία από τις επεκτάσεις που περιγράφονται στη δημοσίευση των Lihe Yang et al.[14] είναι η δημιουργία δύο ισχυρά επαυξημένων εκδοχών μίας εικόνας εισόδου και τροφοδότησης αυτών στο δίκτυο student. Παράλληλα, οι έξοδοι του student για τις δύο ισχυρά επαυξημένες εκδοχές επιβλέπονται από την ψευδοετικέτα που παράγει ο teacher όταν του δίνεται είσοδος η ασθενώς επαυξημένη εικόνα. Ουσιαστικά, οι δύο ισχυρά επαυξημένες εκδοχές παράγονται από την αρχική ασθενώς επαυξημένη εικόνα, εφαρμόζοντας πάνω σε αυτή 2 φορές την μη-ντετερμινιστική (οι μετασχηματισμοι που θα εφαρμοστούν όπως είδαμε επιλέγονται τυχαία) συνάρτηση A^{strong}. Η συγκεκριμένη ιδέα διαπιστώνεται πειραματικά ότι βελτιώνει το baseline FixMatch, καθώς διερευνάται καλύτερα ο χώρος των διαταραχών σε επίπεδο εισόδου του δικτύου, αξιοποιούνται σε μεγαλύτερο βαθμό οι διάφορες ισχυρές διαταραχές, δίνεται μεγαλύτερη πληροφορία στο μοντέλο, ενώ οι συγγραφείς εικάζουν ότι η απαίτηση συνέπειας μεταξύ της ασθενώς επαυξημένη εκδοχής με των δύο ισχυρά επαυξημένων πιθανόν να επιφέρει και συνέπεια ανάμεσα στις δύο ισχυρά επαυξημένες εκδοχές,

λειτουργώντας ως μία επιπλέον μορφή κανονικοποίησης συνέπειας. Με την προσθήκη και της δεύτερης ισχυρά επαυξημένης εικόνας ο όρος του κόστους που αξιοποιεί τα μη επισημασμένα δεδομένα γίνεται ως εξής με βάση το [14]:

$$L_u = \frac{1}{|B_u|} \sum_{u \in B_u} \mathbf{1}[(max(p^{weak} \geq \tau_{thres})]((H(p^{weak}, p^{strong1}) + H(p^{weak}, p^{strong2})) \quad (3.2)$$



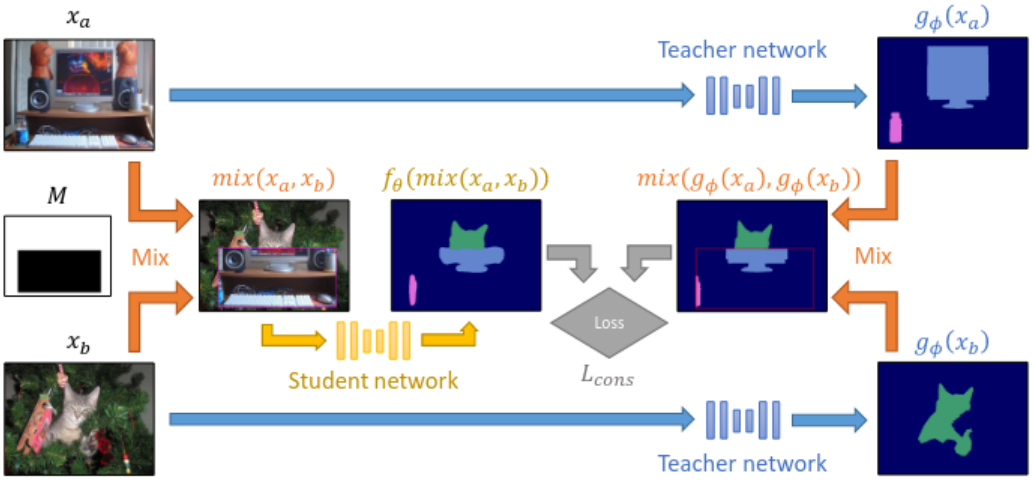
Εικόνα 3.4: *FixMatch με δύο ισχυρά επαυξημένες εκδοχές της εικόνας εισόδου [14]*

Γενικά, στο FixMatch και σε συναφείς μεθόδους που αξιοποιούν διαταραχές σε επίπεδο εισόδου και ειδικότερα το παράδειγμα της ασθενούς-ισχυρής συνέπειας, καθοριστικό ρόλο έχει η επιλογή των μετασχηματισμών για την παραγωγή των ισχυρά επαυξημένων εισόδων. Όπως αναφέρεται και στο [14] και έχει διαπιστωθεί και πειραματικά ότι η χρήση του κλάδου που δέχεται ως είσοδο ισχυρές διαταραχές είναι απαραίτητη, καθώς αυτές προσθέτουν επιπλέον πληροφοριά και βοηθούν στην αποτελεσματική αξιοποίηση των πλεονεκτημάτων της κανονικοποίησης συνέπειας(consistency regularization). Η απουσία ισχυρών διαταραχών μετατρέπει την ιδέα του FixMatch σε μία απλοϊκή μέθοδο self-training[76], η οποία ενδεχομένως να έχει πολύ χειρότερα αποτελέσματα.

Οι Geoffrey French et al.[15] προτείνουν μεθόδους μίξης δειγμάτων, παρόμοιες με τη MixUp [69] που έχουμε αναφέρει σε προηγούμενο κεφάλαιο, ως μία καλή επιλογή για δημιουργία ισχυρά επαυξημένων εικόνων σε προβλήματα σημασιολογικής κατάτμησης(semantic segmentation). Συγκεκριμένα, προτείνεται η επέκταση της χρήσης της μεθόδου CutMix[77] στο πρόβλημα της κατάτμησης εικόνας. Η CutMix αναμειγνύει δύο μη επισημασμένες εικόνες εξάγοντας μία ορθογώνια περιοχή (rectangular patch) από τη μία και προσαρτώντας την στην άλλη. Η επιλογή του μεγέθους της ορθογώνιας περιοχής και της θέσης του πάνω στην εικόνα γίνεται με τυχαίο τρόπο.

Άν υποθέσουμε, ότι έχουμε δύο μη επισημασμένες εικόνες u_1, u_2, τότε η cutmixed εικόνα θα παράγεται ως εξής σύμφωνα με το [15]: $CutMix(u_1, u2, M) = M * u1 + (1 - M) * u_2$, όπου M μία δυαδική μάσκα που περιλαμβάνει άσους στις θέσεις που καθορίζουν την ορθογώνια περιοχή και μηδενικά οπουδήποτε αλλού. Ένας, τέτοιου είδους μετασχηματισμός παράγει μία ισχυρά επαυξημένη εικόνα, έστω $u^{strong_{perturbed}} = CutMix(u_1, u2, M)$. Έν συνεχεία ακολουθώντας το παράδειγμα του ICT [11], οι συγγραφείς προτείνουν την ενσωμάτωση του CutMix σε αυτό. Μετά τη δημιουργία του cutmixed δείγματος, οι επιμέρους μη επισημασμένες ει-

κόνες δίνονται στο δίκτυο teacher, έστω f'. Ο teacher θα παράξει τις εξόδους $f'(u_1), f'(u_2)$, οι οποίες θα γίνουν και αυτές cutmixed $CutMix(f'(u_1), f'(u_2))$, ώστε να προκύψει η αντίστοιχη ψευδοετικέτα που θα επιβλέψει την έξοδο του student.



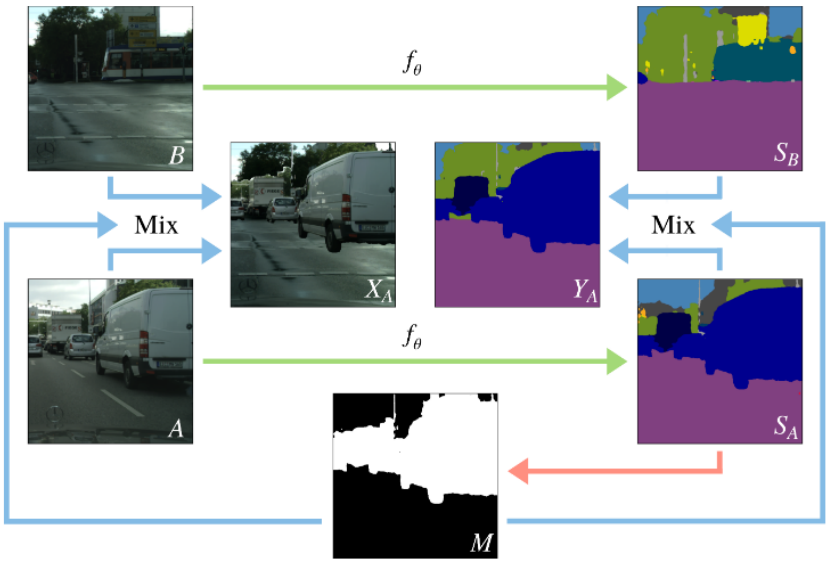
Εικόνα 3.5: *Ενσωμάτωση του CutMix στο παράδειγμα του Interpolation Consistency Training [15]*

Μία άλλη πράξη μίξης που χρησιμοποιείται για την παραγωγή ισχυρά επαυξημένων δειγμάτων είναι η δουλειά των Viktor Olsson et al.[16], οι οποίοι προτείνουν τη μέθοδο επαύξησης δεδομένων ClassMix για το πρόβλημα της σημασιολογικής κατάτμησης εικόνας. Η ClassMix όμοια με την CutMix, χρησιμοποιεί ζεύγη μη επισημασμένων εικόνων, προκειμένου να παραχθεί μία καινούρια ισχυρά επαυξημένη εικόνα. Η διαφορά των δύο μεθόδων οφείλεται στον τρόπο με το οποίο παράγεται η μάσκα μίξης M. Στην περίπτωση του ClassMix, αντί για μία ορθογώνια περιοχή, εξάγονται σημασιολογικές κλάσεις από τη μία εικόνα και προσαρτούνται στην άλλη, δημιουργώντας ένα εντελώς καινούριο δείγμα. Συγκεκριμένα, επιλέγονται οι μισές από τις σημασιολογικές κλάσεις που υπάρχουν στην πρώτη εικόνα και μεταφέρονται στη δεύτερη.

Σύμφωνα με το [16], λοιπόν, δύο μη επισημασμένες εικόνες επιλέγονται τυχαία, έστω A, B και τροφοδοτούνται στο δίκτυο teacher $f'(\partial)$ (και σε αυτή την περίπτωση ακολουθείται το μοντέλο του Mean Teacher [67] και του ICT [11]), ώστε να παραχθούν οι ποσότητες $S_A = f'_\partial(A), S_B = f'_\partial(B)$ που ουσιαστικά πρόκειται για τους χάρτες τμηματοποίησης. Εν συνεχεία, στο S_A εφαρμόζεται μία *argmax* σε επίπεδο pixel, ώστε να παραχθεί η μάσκα $S'_A = argmax(S_A)$, η οποία σε κάθε θέση (i, j) έχει τον δείκτη της κλάσης (class index) που ταξινομήθηκε το αντίστοιχο pixel της αρχικής εικόνας A με βάση την argmax. Η δυαδική μάσκα μίξης M παράγεται με την τυχαία επιλογή των μισών κλάσεων από αυτές που υπάρχουν στην S'_A, θέτοντας τα pixels αυτών των κλάσεων την τιμή 1 στην M και 0 οπουδήποτε αλλού, απομονώνοντας έτσι pixels της εικόνας A που έχουν ταξινομηθεί στην επιλεγμένη κλάση. Η μάσκα M χρησιμοποιείται για την μίξη των αρχικών εικόνων και την παραγωγή της ισχυρά επαυξημένης εικόνας $X_A = M * A + (1 - M) * B$, καθώς και για τη μίξη των S_A, S_B για τη δημιουργία της αντίστοιχης ψευδοετικέτας $Y_A = M * S_A + (1 - M) * S_B$ της X_A.

Όπως, είπαμε επειδή ακολουθούνται τα μοντέλα Mean Teacher και ICT, η εικόνα μίξης

X_A θα τροφοδοτηθεί στο δίκτυο $f(\theta)$ student, η έξοδος του οποίου θα επιβλέπεται από την ψευδοετικέτα Y_A, η οποία παράγεται όπως είδαμε από τη μίξη των εξόδων του teacher για τις δύο αρχικές εικόνες A, B.



Εικόνα 3.6: *Η μέθοδος μίξης ClassMix [16]*

Η συνολική συνάρτηση κόστους όπως περιγράφεται στη δημοσίευση είναι η παρακάτω[16]:

$$L(\theta) = E[l(f_\theta(X_L), Y_L) + \lambda \cdot l(f_\theta(X_A), Y_A)] \tag{3.3}$$

, όπου (X_L, Y_L) οι επισημασμένες εικόνες μαζί με την αντίστοιχη ετικέτα και οι (X_A, Y_A) οι classmixed εικόνες μαζί με την αντίστοιχη ψευδοετικέτα. Το λ πρόκειται για την υπερπαράμετρο που καθορίζει τη συνεισφορά του μη επιβλεπόμενου όρου και το l το κόστος σε επίπεδο pixel cross-entropy [16].

$$l(Pred, Y) = -\frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{C} Y(i,j,c) \cdot \log Pred(i,j,c) \tag{3.4}$$

3.1.2 Προσεγγίσεις βασισμένες σε διαταραχές ενδιάμεσων χαρακτηριστικών (feature-level perturbations)

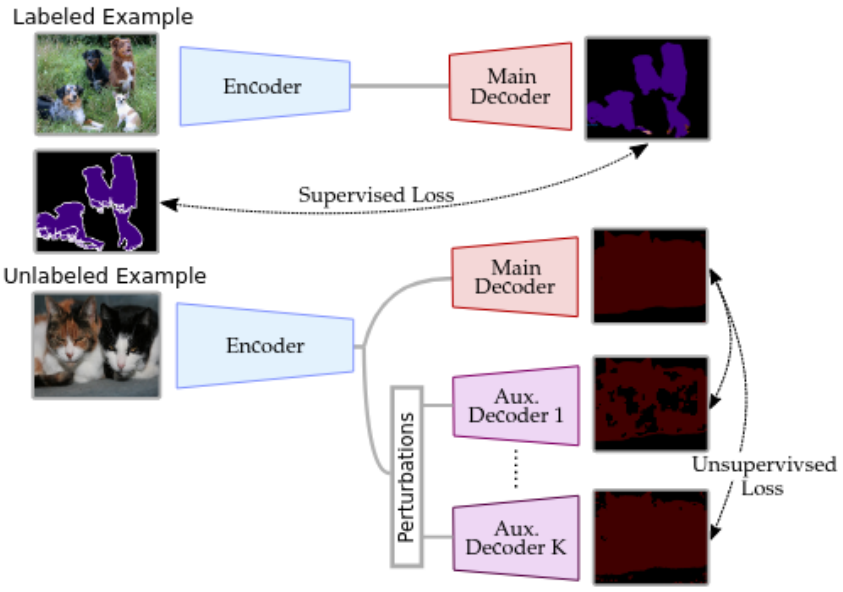
Η δημοσίευση των Yassine Ouali et al.[17] προτείνει μία μεθοδολογία που στηρίζεται σε εφαρμογή διαταραχών σε ενδιάμεσες αναπαραστάσεις (intermediate representations, intermediate features) που παράγονται από κάποιο επίπεδο του δικτύου τμηματοποίησης. Όπως γνωρίζουμε, ένα δίκτυο τμηματοποίησης αποτελείται από δύο μέρη: Τον encoder και τον decoder. Η ιδέα που προτείνεται από τους συγγραφείς στο [17] είναι η εφαρμογή διαφορετικών διαταραχών στα ενδιάμεσα χαρακτηριστικά που παράγονται στην έξοδο του encoder.

Πιο συγκεκριμένα, το δίκτυο τμηματοποίησης αποτελείται από έναν κοινό encoder (shared encoder) και από τον βασικό decoder (main decoder). Τα επισημασμένα δεδομένα αξιοποιούνται με τον κλασικό τρόπο, δηλαδή τροφοδοτούνται στο δίκτυο shared encoder - main

Αλγοριθμος 3.1: *Ο αλγόριθμος του ClassMix, όπως διατυπώνεται στο [16]*

Είσοδος: $A, B, f_{\partial'}$ (*δύο unlabeled εικόνες προς μίξη και δίκτυο* $f_{\partial'}$)
Έξοδος: X_A, Y_A (*classmixed εικόνα με το αντίστοιχο pseudolabel*)
$S_A = f_{\partial'}(A)$, $S_B = f_{\partial'}(B)$
$S'_A = argmax(S_A)$ (*pixel-wise argmax, πάνω στις κλάσεις*)
$C = set(S'_A)$ (*σύνολο C των διαφορετικών κλάσεων που υπαρχούν στη πρόβλεψη του δικτύο* S_A)
$c = randomSelect(C)$ (*Τυχαιά επιλογή υποσυνόλου κλάσεων με πλήθος* $|c| = |C|/2$)
for $(i,j) \in S'_A$ **do**
  if $S'_A(i,j) \in c$ **then**
    $M(i,j) = 1$
  else
    $M(i,j) = 0$
  end if
end for
$X_A = M * A + (1 - M) * B$, $Y_A = M * S_A + (1 - M) * S_B$
return X_A, Y_A

decoder και υπολογίζεται το κόστος cross-entropy με βάση τις αντίστοιχες ψευδοετικέτες. Για την αξιοποίηση των μη επισημασμένων δεδομένων γίνεται χρήση επιπλέον βοηθητικών decoders (auxiliary decoders). Αυτό που γίνεται είναι η εφαρμογή διάφορετικών διαταραχών σε πλήθος ίσο με τον αριθμό των βοηθητικών decoders που υπάρχουν, στην κοινή έξοδο του encoder με αποτέλεσμα να παράγονται διαφορετικές διαταραγμένες εκδοχές των χαρακτηριστικών εξόδου του encoder. Εν συνεχεία, οι διαταραγμένες εκδοχές δίνονται ως είσοδο στο βασικό και στους βοηθητικούς decoders και επιβάλλεται συνέπεια μεταξύ της εξόδου του βασικού decoder και των εξόδων των βοηθητικών decoders. Το παραπάνω παράδειγμα εκπαίδευσης ονομάζεται απο τους συγγραφείς Cross-Consistency Training (CCT)[17].



Εικόνα 3.7: *Η βασική ιδεά του Cross Consistency Training (CCT)[17]*

Όπως, βλέπουμε το δίκτυο f αποτελείται από έναν κοινό encoder, έναν βασικό decoder και K σε πλήθος βοηθητικούς decoders που δέχονται στην είσοδο τους τα διαταραγμένα

χαρακτηριστικά από την έξοδο του encoder. Ο κλάδος που αξιοποιεί τα επισημασμένα δεδομένα είναι ουσιαστικά ο shared encoder - main decoder και μπορεί να συμβολιστεί ως $f = goh$. Ώς σύνθεση, δηλαδή του encoder h και του βασικού decoder g. Επιπλέον, υπάρχουν K βοηθητικοι decoders, όπως είπαμε που μπορούν να συμβολιστούν ως $g^k_{aux}, k \in [1, ..., K]$. Έστω μη επισημασμένο δείγμα x^u_i που το τροφοδοτούμε στον κοινό encoder. Τότε ο encoder θα παραξει μία κρυφή αναπαράσταση (hidden representation), έστω $z_i = h(x^u_i)$. Εφαρμόζοντας, σε αυτή την αναπαράσταση K σε πλήθος διαταραχές θα προκύψουν K διαφορετικές διαταραγμένες εκδοχές της εξόδου του encoder, έστω z^k_i. Κάθε ένα από τα z^k_i τροφοδότειται στο αντίστοιχο k-βοηθητικό decoder.

Η συνολική συνάρτηση κόστους σύμφωνα με το [17] μπορεί να εκφραστεί ως εξής:

$$L = L_s + \omega \cdot L_u$$

$$L_s = \frac{1}{|D_l|} \sum_{x^l_i, y_i \in D_l} CE(y_i, f(x^l_i)) \tag{3.5}$$

$$L_u = \frac{1}{|D_u|} \frac{1}{K} \sum_{x^u_i \in D_u} \sum_{k=1}^{K} d(g(z_i), d(g^k_{aux}(p^r(z_i))) \tag{3.6}$$

Ο μη επιβλεπόμενος όρος χρησημοποιεί μία συνάρτηση που αξιολογεί την απόσταση δύο κατανομών (π.χ Mean Square Error). Συγκεκριμένα, το $g(z_i)$ αποτελεί την έξοδο του βασικού decoder, ενώ το $g^k_{aux}(p^r(z_i))$ την έξοδο του k-οστού βοηθητικού decoder για τη k-διαταραγμένη εκδοχή της εξόδου z_i, η οποία προκύπτει από την εφαρμογή μίας στοχαστικής συνάρτησης διαταραχής p^r στο z_i[17]. Όπως, έχει αναφερθεί στόχος είναι η ελαχιστοποίηση της απόστασης, δηλαδη η συνέπεια (consistency), μεταξύ της εξόδου του βασικού decoder και των επιμέρους εξόδων των βοηθητικών decoders. Το ω αποτελεί το βάρος που καθορίζει τη συνεισφορά του μη επιβλεπόμενου όρου στο συνολικό κόστος.

Ενδεικτικά αναφέρουμε κάποιες από τις συναρτήσεις διαταραχών(perutbation) που χρησιμοποιούνται στο [17]:

- F-Noise: Δειγματοληπτείται ομοιόμορφος θόρυβος $N\ U(-0.3, 0.3)$ ίδιων διαστάσεων με τα χαρακτηριστικά εξόδου z του encoder. Η διαταραγμένη εκδοχή προκύπτει ως εξής: $z^{perturbed} = (z * N) + z$
- Εφαρμογή τυχαίου Spatial Dropout[78] στα χαρακτηριστικά z.

3.1.3 Προσεγγίσεις βασισμένες σε διαταραχές σε επίπεδο δικτύου (network-level perturbations)

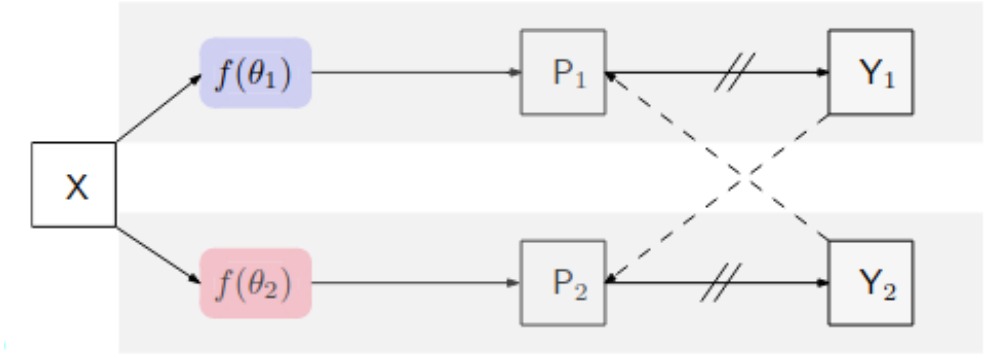
Μία σημαντική δουλειά που ανήκει σε αυτή την κατηγορία είναι αυτή των Xiaokang Chen et al.[18]. Οι συγγραφείς εδώ, προτείνουν τη χρήση δύο δικτύων κατάτμησης ίδιας αρχιτεκτονικής, τα οποία έχουν διαφορετική αρχικοποίηση παραμέτρων, έστω $f(\partial 1), f(\partial 2)$. Στα δίκτυα δίνεται ως είσοδος η ίδια εικόνα και απαιτείται συνέπεια στις εξόδους τους. Αυτό, όπως έχουμε πει θεωρείται μία μορφή διαταραχής (perturbation) σε επίπεδο δικτύου. Επιπλέον, γίνεται και χρήση ψευδοετικετών. Συγκεκριμένα, οι ψευδοετικέτες που παράγονται από το ένα δίκτυο χρησιμοποιούνται για την επίβλεψη του δεύτερου δικτύου και το αντίστρο-

φο, μέσω του κόστους cross entropy. Για αυτό και η ιδέα ονομάζεται διασταυρούμενη ψευδοεπίβλεψη (Cross Pseudo Supervision).

Αν υποθέσουμε, ότι έχουμε μία μη επισημασμένη (unlabeled) εικόνα εισόδου X, η οποία πιθανόν να έχει υποστεί κάποιο είδος επαύξησης, τότε αν τροφοδοτηθεί στα δύο δίκτυα θα ισχύει:

$$P1 = f_{\partial 1}(X)$$
$$P2 = f_{\partial 2}(X)$$

Τα $P1, P2$ αποτελούν τους χάρτες βεβαιότητας ή εξόδους softmax (confidence maps) που πάραξαν τα δίκτυα για την κοινή είσοδο. Από τα $P1, P2$ μπορούν να προκύψουν οι one-hot ψευδοετικέτες $\Upsilon 1, \Upsilon 2$, οι οποίες χρησιμοποιούνται για την επίβλεψη των δικτύων $f_{\partial 2}$ και $f_{\partial 1}$ αντίστοιχα. Να τονίσουμε ότι ο αλγόριθμος οπισθοδιάδοσης σφάλματος (backpropagation) εκτελείται κατά μήκος και των δύο δικτύων.



Εικόνα 3.8: *Η ιδεά του Cross Pseudo Supervision[18]*

Η συνάρτηση κόστους, όπως συνήθως αποτελείται από έναν επιβλεπόμενο και έναν μη επιβλεπόμενο όρο. Για τα επισημασμένα δεδομένα υπολογίζεται το κόστος cross-entropy με βάση τις διαθέσιμες ετικέτες. Σύμφωνα με το [18] έχουμε:

$$L_s = \frac{1}{|D_l|} \sum_{X \in D_l} \frac{1}{W \cdot H} \sum_{i=0}^{W \cdot H} l_{CE}(p_{1i}, y^*_{1i}) + l_{CE}(p_{2i}, y^*_{2i}) \tag{3.7}$$

Τα p_{1i}, p_{2i} αποτελούν διανύσματα βεβαιότητας (confidence) για κάθε pixel i που ανήκει στα $P1, P2$. Αντίστοιχα, τα y^*_{1i}, y^*_{2i} αποτελούν τις one-hot ψευδοετικέτες για κάθε pixel i στις διαθέσιμες ετικέτες $\Upsilon^*_1, \Upsilon^*_2$.

Για την περίπτωση των μη επισημασμένων δεδομένων το κόστος διασταυρούμενης ψευδοεπίβλεψης μπορεί να γραφτεί σύμφωνα με το [18] ως εξής:

$$L_{cps} = \frac{1}{|D_u|} \sum_{X \in D_u} \frac{1}{W \cdot H} \sum_{i=0}^{W \cdot H} l_{CE}(p_{1i}, y_{2i}) + l_{CE}(p_{2i}, y_{1i}) \tag{3.8}$$

Όπως βλέπουμε, οι one-hot ψευδοετικέτες y_{1i} για κάθε pixel στο $\Upsilon 1$, επιβλέπουν τις εξόδους p_{2i}, ενώ οι y_{2i} τις p_{1i}. Στη δημοσίευση αναφέρεται, επίσης, ότι το κόστος L_{cps} υπολογίζεται

και για τα επισημασμένα δεδομένα. Το συνολικό κόστος γράφεται ως $L_{total} = L_s + \lambda \cdot L_{cps}$, όπου λ το βάρος συνεισφοράς του ημιεπιβλεπόμενου όρου.

3.1.4 Συνδυαστικές προσεγγίσεις

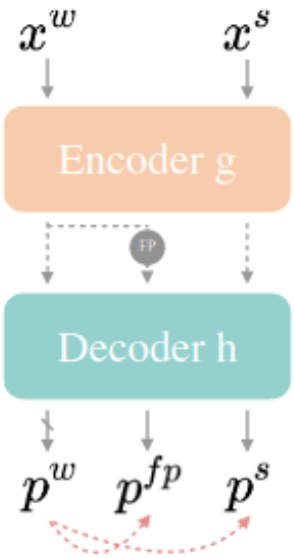
Πολλές φορές μπορεί να αποδειχτεί αποδοτικός ένας συνδυασμός διαφορετικών ειδών διαταραχών. Μια τέτοια περίπτωση αποτελεί η επέκταση που προτείνεται ξανά από το [14], όπου πραγματοποιείται συνδυασμός διαταραχών επιπέδου εισόδου (input-level) και επιπέδου χαρακτηριστικών (feature-level). Σύμφωνα με τους συγγραφείς, με τον παραπάνω συνδυασμό μπορεί να εξερευνηθεί ένας μεγαλύτερος χώρος διαταραχών (perturbation space) και να διασφαλιστεί η συνέπεια (consistency) των προβλέψεων σε διαφορετικά επίπεδα. Για την διαταραχή των χαρακτηριστικών χρησιμοποιείται ένα απλό επίπεδο dropout στην έξοδο του encoder του δικτύου κατάτμησης, προκειμένου να αποκόπτονται κάθε φορά διαφορετικά χαρακτηριστικά στον κλάδο της ασθενούς επαυξημένης εκδοχής (weak augmented).

Αν θεωρήσουμε τον encoder g και P την συνάρτηση διαταραχής χαρακτηριστικών τότε τα διατεραγμένα χαρακτηριστικά της εξόδου του encoder θα είναι τα εξής σύμφωνα με το [14].

$$e^{fp} = P(g(x^{weak}))$$

Η αντίστοιχη έξοδος του decoder h με είσοδο τα παραπάνω διαταραγμένα χαρακτηριστικά θα είναι [14]:

$$p^{fp} = h(e^{fp})$$



Εικόνα 3.9: *Συνδυασμός διαταραχών επιπέδου εισόδου και επιπέδου χαρακτηριστικών [14]*

Η συνάρτηση κόστους σε αυτή την περίπτωση μπορεί να γραφτεί ως εξής [14]:

$$L_u = \frac{1}{|B_u|} \sum_{u \in B_u} \mathbf{1}[(max(p^{weak} \geq \tau_{thres})]((H(p^{weak}, p^{strong}) + H(p^{weak}, p^{fp})) \quad (3.9)$$

Βλέπουμε, λοιπόν ότι η ασθενώς επαυξημένη εκδοχή επιβλέπει τόσο την έξοδο της ισχυρής εκδοχής, όσο και την έξοδο που προκύπτει από εφαρμογή διαταραχών στα χαρακτηριστικά.

Μια επίσης πολύ διαδεδομένη τεχνική που χρησιμοποιείται τα τελευταία χρόνια στο τομέα της ημιεπιβλεπόμενης κατάτμησης εικόνας είναι ο συνδυασμός όλων των παραπάνω τεχνικών της κανονικοποίησης συνέπειας που έχουμε παρουσιάσει με την ιδέα της συγκριτικής μάθησης (Contrastive Learning) [79]. Διαισθητικά, αυτό που προσπαθεί να πετύχει το contrastive learning είναι να εκπαιδεύσει το δίκτυο να παράγει χαρακτηριστικά που βρίσκονται κοντά το ένα με το άλλο σε έναν χώρο από embeddings, για τα δείγματα που ανήκουν στην ίδια κλάση, και αντίστοιχα να απομακρύνει στον χώρο τα χαρακτηριστικά που παράγονται για δείγματα που ανήκουν σε διαφορετικές κλάσεις. Για παράδειγμα, στην περίπτωση της κατάτμησης εικόνας σκοπός είναι τα χαρακτηριστικά των pixels της ίδιας κλάσης να ωθούνται κόντα στον χώρο, ενώ τα χαρακτηριστικά των pixels διαφορετικών κλάσεων να απομακρύνονται μεταξύ τους. Το παραπάνω επιτυγχάνεται με την προσθήκη ενός ακόμη όρου στο συνολικό κόστος που ονομάζεται InfoNCE [80]. Κάποιες, από τις δουλειές που αξιοποιούν αυτόν τον όρο στο συνολικό κόστος είναι οι [81, 82, 83]. Με τη χρήση αυτού του όρου είναι δυνατό το δίκτυο κατάτμησης να παράγει καλύτερα δομημένους χώρους ενδιάμεσων χαρακτηρικών με πιο εμφανείς συστάδες για τα pixels ίδιας κλάσης, έτσι ώστε να είναι πιο εύκολος ο διαχωρισμός τους και συνεπώς το αποτέλεσμα της κατάτμησης να είναι καλύτερης ποιότητας.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

Επισκόπηση συνόλων δεδομένων

Στο παρόν κεφάλαιο αναφερόμαστε στα σύνολα δεδομένων σημασιολογικής κατάτμησης που χρησιμοποιήσαμε για τη διεξαγωγή των πειραμάτων. Πιο συγκεκριμένα, δίνουμε γενικές πληροφορίες για κάθε σύνολο, παρουσιάζουμε τις σημασιολγογικές κλάσεις που περιλαμβάνει κάθε ένα από αύτα, καθώς και ορισμένα παραδείγματα εικόνων με τις αντίστοιχες ετικέτες κατάτμησης τους. Επιπλέον, αναφερόμαστε σε επιλογές που κάναμε όσον αφορά την προετοιμασία των δεδομένων για την εκπαίδευση.

4.1 Pascal VOC 2012

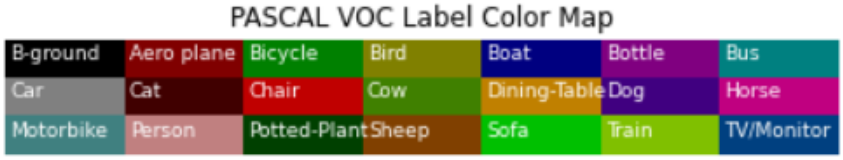
Το Pascal VOC 2012 [19] αποτελεί ένα βασικό σύνολο δεδομένων για το πρόβλημα της σημασιολογικής κατάτμησης εικόνων αντικειμένων από τον φυσικό κόσμο. Περιλαμβάνει 21 σημασιολογικές κλάσεις συμπεριλαμβανομένης της κλάσης background. Το πρωτότυπο

background:0	dining table:11
airplane:1	dog:12
bicycle:2	horse:13
bird:3	motorbike:14
boat:4	person:15
bottle:5	potted plant:16
bus:6	sheep:17
car:7	sofa:18
cat:8	train:19
chair:9	tv/monitor:20
cow:10	-

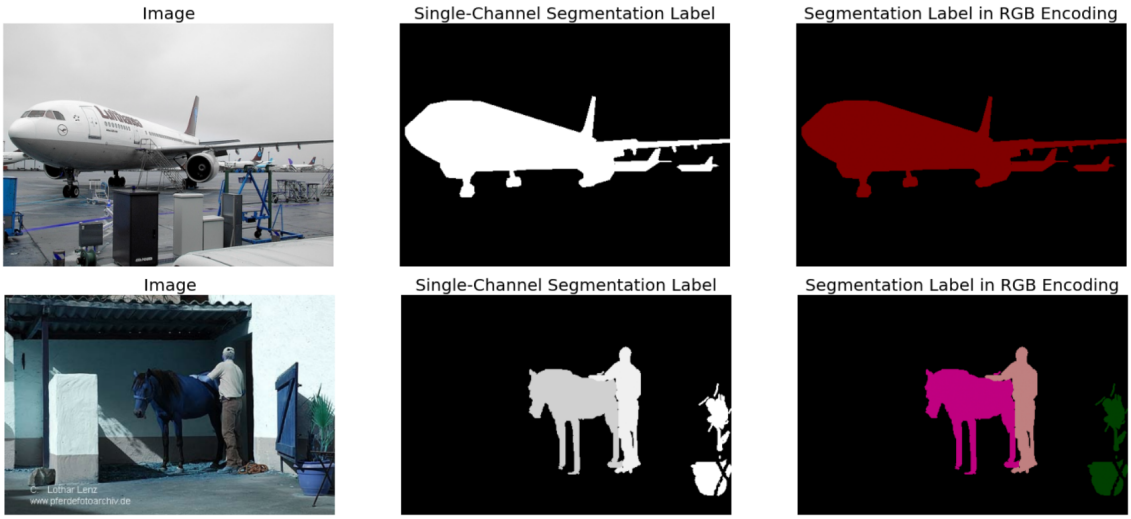
Πίνακας 4.1: *Σημασιολογικές κλάσεις στο σύνολο Pascal*

σύνολο περιέχει 1464 εικόνες για εκπαίδευση (train) και 1449 για επικύρωση (validation). Ακολουθώντας τις πρακτικές της σχετικής βιβλιογραφίας [18, 15, 84] κάνουμε χρήση των επιπλέον ετικετών κατάτμησης που προσφέρονται από το σύνολο SBD [85], μίας επαυξημένης εκδοχής του Pascal, με αποτέλεσμα το τελικό σύνολο εκπαίδευσης που θα πειραματιστούμε

να αποτελείται από 10582 εικόνες. Όπως βλέπουμε και στον πίνακα 4.1 κάθε κλάση προσδιορίζεται από κάποιον δείκτη (class index) στο διάστημα $[0-20]$. Επιπλέον, σε κάθε κλάση αντιστοιχίζεται κάποιο χρώμα *rgb*, ώστε να μπορεί να οπτικοποιηθεί καλύτερα στις ετικέτες κατάτμησης. Παρακάτω παρουσιάζεται η επίσημη αντιστοίχιση.



Εικόνα 4.1: *Η αντιστοίχιση κλάσεων και χρωμάτων για το σύνολο Pascal [19]*



Εικόνα 4.2: *Παρδείγματα εικόνων και ετικετών από το σύνολο Pascal [19]*

Στις παραπάνω εικόνες βλέπουμε δύο μορφές κωδικοποίησης για τις ετικέτες κατάτμησης. Η πρώτη κωδικοποίηση πρόκειται για μία εικόνα ενός μόνο καναλιού, όπου η τιμή του κάθε pixel (i,j) αντιστοιχεί σε έναν δείκτη κλάσης (class index) στο εύρος $[0-20]$. Για να είναι εμφανή τα αντικείμενα που περιέχονται έχει γίνει αναπαράσταση της εικόνας σε αποχρώσεις του γκρι (gray scale). Επιπλέον , η συγκεκριμένη μάσκα κατάτμησης χρησιμοποιείται κατα τη διάρκεια της εκπαίδευσης ως στόχος (target), καθώς μπορεί εύκολα να μετατραπεί σε one-hot . Η δεύτερη μορφή της ετικέτας κατάτμησης βρίσκεται σε *RGB* κωδικοποίηση και για την οπτικοποίηση των κλάσεων γίνεται χρήση των χρωμάτων που έχουν αντιστοιχηθεί σε κάθε μία απο αυτές. Συνήθως, η συγκεκριμένη μορφή χρησιμοποιείται για την καλύτερη παρουσίαση των αποτελεσμάτων κατάτμησης.

4.2 CelebAMask-HQ

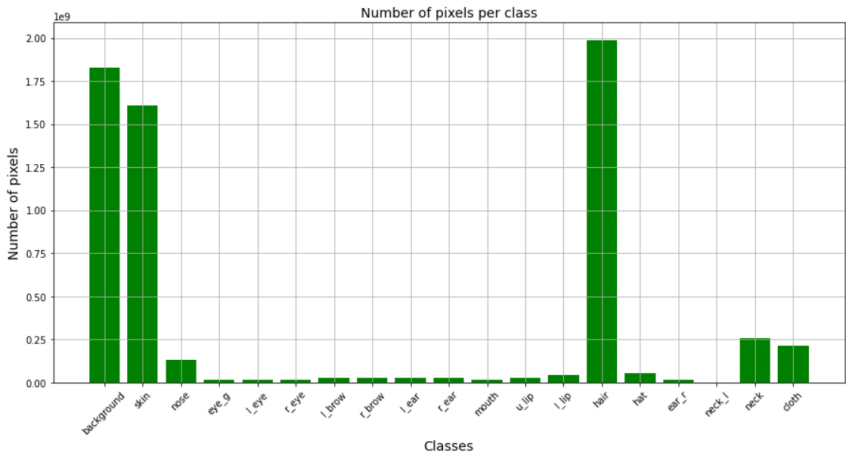
Το συγκεκριμένο σύνολο δεδομένων [20] περιέχει εικόνες υψηλής ανάλυσης (high-resolution HQ) προσώπων διασήμων ατόμων, οι οποίες έχουν επιλεχθεί από το μεγάλης κλίμακας σύνολο δεδομένων CelebA [86]. Το CelebAMask-HQ θεωρείται επίσης ένα μεγάλης κλίμακας

σύνολο, καθώς αποτελείται από 30000 εικόνες προσώπων διαστάσεων $(512, 512)$ για τα οποία προσφέρονται οι αντίστοιχες μάσκες κατάτμησης μερών του ανθρώπινου προσώπου και διάφορων αντικειμένων που μπορεί να υπάρχουν σε αυτό. Πιο συγκεκριμένα περιλαμβάνονται οι παρακάτω 19 σημασιολογικές κλάσεις.

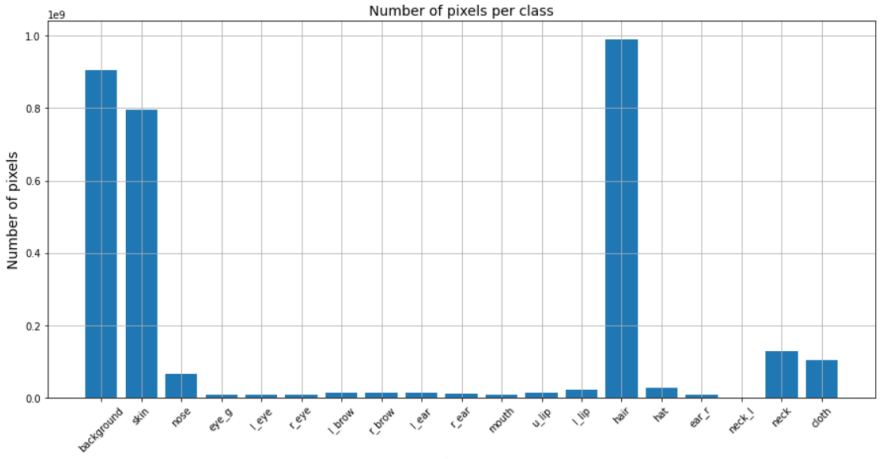
background:0	upper lip:11
skin:1	lower lip:12
nose:2	hair:13
eye glasses:3	hat:14
left eye:4	earring:15
right-eye:5	necklace:16
left brow:6	neck:17
right brow:7	cloth:18
left ear:8	-
right ear:9	-
mouth:10	-

Πίνακας 4.2: *Σημασιολογικές κλάσεις στο σύνολο CelebAMask-HQ [20]*

Τα δεδομένα ειναι διαχωρισμένα σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου με μεγέθη 24183, 2993 και 2824 αντίστοιχα. Εξαιτίας του μεγάλου πλήθους δεδομένων καθίσταται πρακτικά δύσκολος ο πειραματισμός πάνω σε όλο το σύνολο, λόγω και της μη ύπαρξης αρκετών πόρων για τη διαχείρηση δεδομένων αυτής της κλίμακας. Συνεπώς, επιλέγουμε να εφαρμόσουμε τους αλγορίθμους πάνω σε ένα υποσύνολο του αρχικού CelebAMask-HQ, το οποίο δημιουργούμε δειγματοληπτώντας τυχαία 12000 εικόνες από το σύνολο εκπαίδευσης. Παρακάτω παρουσιάζουμε την κατανομή των pixels ανά κλάση για το αρχικό σύνολο εκπαίδευσης, καθώς και για το υποσύνολο που δημιουργήσαμε.



Εικόνα 4.3: *Κατανομή πλήθους pixel για το ολικό σύνολο εκπαίδευσής*



Εικόνα 4.4: *Κατανομή πλήθους pixel για το υποσύνολο εκπαίδευσής*

Παρατηρούμε ότι η μορφή της κατανομής του πλήθους των pixels ανα κλάση είναι πανομοιότυπη. Η κλάση με τον μικρότερο αριθμό από pixel είναι η necklace με 1088278 pixels στο αρχικό σύνολο εκπαίδευσης και 526676 για το υποσύνολο που δημιουργήσαμε. Επειδή, διαπιστώσαμε ότι κατά την εκπαίδευση, το μοντέλο δεν ήταν ικανό να αναγνωρίσει τη συγκεκριμένη κλάση από τα υπάρχοντα δεδομένα και για αυτό επιλέξαμε να αναφέρουμε τα πειραματικά αποτελέσματα για το πρόβλημα 18 σημασιολογικών κλάσεων. Στη συνέχεια, παρουσιάζουμε την αντιστοίχιση χρώματος με κλάση, καθώς και παραδείγματα εικόνων και ετικετών κατάτμησης από το σύνολο δεδομένων.

Εικόνα 4.5: *Η αντιστοίχιση κλάσεων και χρωμάτων για το σύνολο CelebAMask-HQ [20]*

Εικόνα 4.6: *Παρδείγματα εικόνων και ετικετών από το σύνολο CelebAMask-HQ [20]*

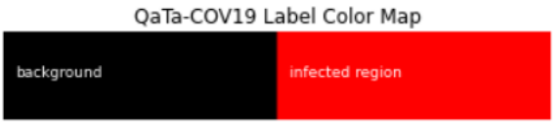
4.3 QaTa-COV19 Dataset

Το QaTa-COV19 [21, 22] πρόκειται για ένα σύνολο δεδομένων από το πεδίο της ιατρικής, το οποίο έχει δημιουργηθεί από ερευνητές των πανεπιστημίων του Κατάρ, της πόλης Τάμπερε στη Φιλανδία και της εταιρείας Hamad Medical Corporation. Το συγκεκριμένο σύνολο αποτελείται από ακτινογραφίες θώρακος ασθενών για τις οποίες παρέχονται μάσκες κατάτμησης της περιοχής των πνευμόνων που έχει μολυνθεί εξαιτίας του Covid-19. Η αρχική έκδοση του συνόλου αποτελούνταν από 4603 ακτινογραφίες, ενώ η δεύτερη έκδοση του που χρησιμοποιούμε επεκτάθηκε στις 9258 ακτινογραφίες. Οι μάσκες κατάτμησης περιλαμβάνουν δύο μόνο σημασιολογικές κλάσεις (binary segmentation), την κλάση background και την κλάση που αντιπροσωπεύει την μολυσμένη περιοχή του πνεύμονα (infected region).

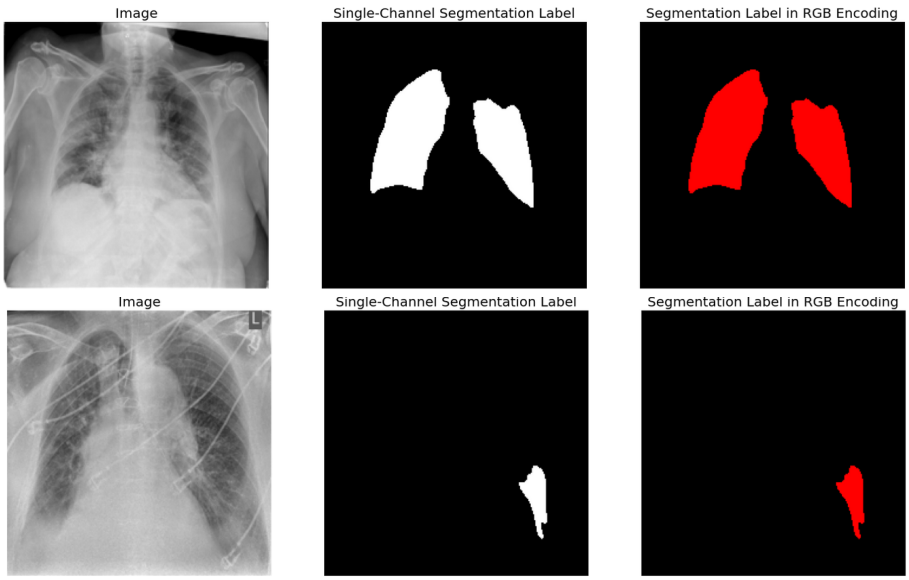
background:0	infected region:1

Πίνακας 4.3: *Σημασιολογικές κλάσεις στο σύνολο QaTa-COV19 [21, 22]*

Τα δεδομένα χωρίζονται σε ένα σύνολο εκπαίδευσης που περιέχει 7145 εικόνες και σε ένα σύνολο ελέγχου που περιλαμβάνει 2113. Επειδή, σε αυτό το σύνολο δεδομένων έχουμε μόνο δύο κλάσεις επιλέγουμε να αντιστοιχίσουμε το background το μαύρο χρώμα και στην περιοχή μόλυνσης το κόκκινο.



Εικόνα 4.7: *Η αντιστοίχιση κλάσεων και χρωμάτων για το σύνολο QaTa-COV19[21, 22]*



Εικόνα 4.8: *Παρδείγματα εικόνων και ετικετών από το σύνολο QaTa-COV19 [19]*

Κεφάλαιο 5

Γενική μεθοδολογία και υλοποίηση

Σε αυτό το κεφάλαιο παρουσιάζουμε τη γενική μεθοδολογία, καθώς και τις λεπτομέρειες της υλοποίησης μας, τόσο για την επιβλεπόμενη, όσο και για την ημιεπιβλεπόμενη προσέγγιση που επιλέξαμε να ακολουθήσουμε κατά τον πειραματισμό μας με τα προαναφερθέντα σύνολα δεδομένων.

5.1 Διαχωρισμός δεδομένων σε επισημασμένα (labeled) και μη επισημασμένα (unlabeled)

Προκειμένου να εφαρμόσουμε τεχνικές ημιεπιβλεπόμενης μάθησης χρειάζεται να έχουμε ένα επαρκές σε μέγεθος σύνολο από μη επισημασμένα δεδομένα τα οποία θα αξιοποιούνται μαζί με τα επισημασμένα κατα τη διάρκεια της εκπαίδευσης. Γι' αυτό τον λόγο επιλέγουμε να πραγματοποιήσουμε κάποιες διαμερίσεις στα σύνολα εκπαίδευσης (training sets) των συνόλων δεδομένων που χρησιμοποιούμε (datasets), έτσι ώστε κάθε σύνολο εκπαίδευσης να αποτελείται από τα επιμέρους σύνολα των επισημασμένων (labeled) και μη επισημασμένων (unlabeled) δειγμάτων.

Παρακάτω αναφέρονται για κάθε σύνολο δεδομένων οι διαμερίσεις που πραγματοποιήσαμε. Όπως, φαίνεται έχουμε κάνει περισσότερες από μία διαμερίσεις για κάθε σύνολο, προκειμένου να μπορεί η ημιεπιβλεπόμενη μέθοδος να αξιολογηθεί υπό διάφορες ποσότητες επισημασμένων και μη επισημασμένων δεδομένων.

Πίνακας 5.1: *Διαμερίσεις επισημασμένων/μη επισημασμένων δεδομένων για το σύνολο Pascal VOC 2012*

Dataset	1/32	1/16	1/8	1/4
Pascal	330 labels	662 labels	1300 labels	2645 labels

Πίνακας 5.2: *Διαμερίσεις επισημασμένων/μη επισημασμένων δεδομένων για το σύνολο CelebAMask-HQ*

Dataset	1/64	1/32	1/16
CelebAMask-HQ	187 labels	375 labels	750 labels

Πίνακας 5.3: *Διαμερίσεις επισημασμένων/μη επισημασμένων δεδομένων για το σύνολο QaTa-COV19*

Dataset	1/32	1/16	1/8	1/4
QaTa-COV19	223 labels	446 labels	893 labels	1786 labels

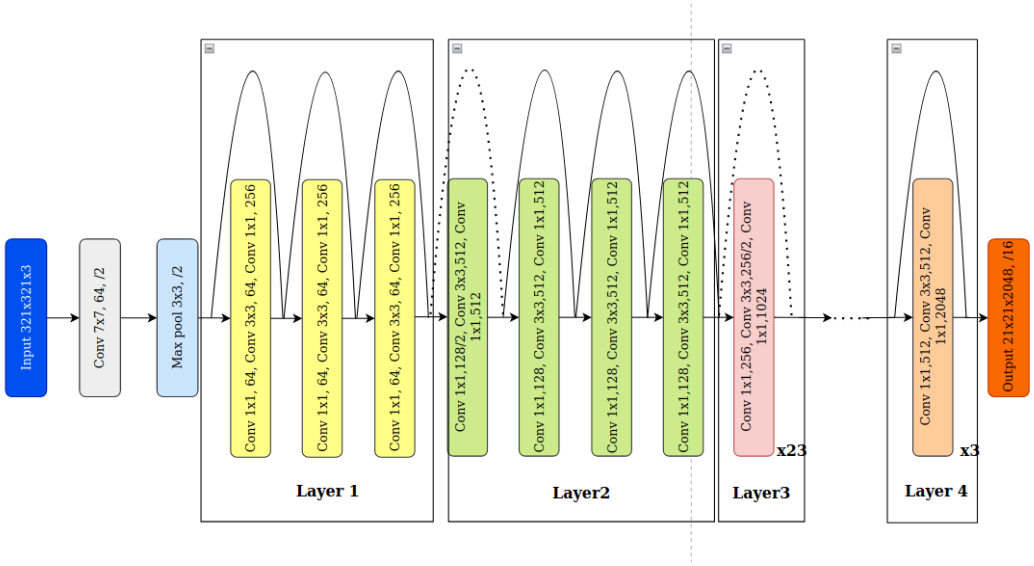
Οι συγκεκριμένες διαμερίσεις εκφράζουν το ποσοστό του αρχικού συνόλου εκπαίδευσης (training set) για το οποίο χρησιμοποιούνται οι ετικέτες των δειγμάτων για κάθε σύνολο δεδομένων. Τα δείγματα τα οποία θεωρούνται επισημασμένα (labeled) δειγματοληπτούνται με τυχαίο τρόπο από το αρχικό σύνολο εκπαίδευσης. Για την περίπτωση της αμειγώς επιβλεπόμενης μάθησης, κατά την εκπαίδευση γίνεται αποκλειστικά χρήση των επισημασμένων δεδομένων, το πλήθος των οποίων αναφέρεται στους παραπάνω πίνακες. Από την άλλη, κατά την εφαρμογή της ημιεπιβλέπομενης μάθησης αξιοποιούνται επιπλέον και τα εναπομέιναντα μη επισημασμένα δεδομένα. Για παράδειγμα, το σύνολο δεδομένων pascal που αποτελείται απο ένα σύνολο εκπαίδευσης με 10582 εικόνες, αν θεωρήσουμε μία διαμέριση 1/32 για αυτό, θα σημαίνει ότι οι $10582 div 32 = 330$ από αυτές τις εικόνες θα θεωρηθούν επισημασμένες και θα μπορούν να αξιοποιηθούν οι ετικέτες τους, ενώ οι εναπομείναντες $10582 - 330 = 10252$ θα χρησιμοποιηθούν ως μη επισημασμένες.

5.2 Αρχιτεκτονική δικτύου κατάτμησης (Segmentation network architecture)

Το δίκτυο που επιλέγουμε να χρησιμοποιήσουμε για την εκπαίδευση στα παραπάνω σύνολα δεδομένων είναι το DeepLabv3plus 2.11, το οποίο έχουμε αναφέρει και στο κεφάλαιο του θεωρητικού υποβάθρου. Σε αυτή την παράγραφο θα εμβαθύνουμε σε περισσότερες δομικές λεπτομέρειες αυτής της αρχιτεκτονικής. Το DeepLabv3plus, όπως έχουμε πει αποτελείται από τρείς επιμέρους συνιστώσες:

- Το δίκτυο ¨ραχοκοκαλιάς¨ backbone
- Τη δομή Atrous Spatial Pyramid Pooling (ASPP) που μαζί με το backbone αποτελούν τον κωδικοποιητή (encoder)
- Τον αποκωδικοποιήτη (decoder)

Δίκτυο backbone Για το δίκτυο backbone που αποτελεί τον βασικό εξαγωγέα χαρακτηριστικών (feature extractor) κάνουμε χρήση της αρχιτεκτονικής ResNet101[23]. Επιλέγουμε να αρχικοποιήσουμε το δίκτυο με τα προεκπαιδεύμενα βάρη από το σύνολο δεδομένων ImageNet[87], εφαρμόζοντας έτσι μεταφορά μάθησης (transfer learning)[88] από το σύνολο ImageNet στα σύνολα Pascal, CelebAMask-HQ και QaTa-COV19. Να τονίσουμε ότι κατα την εκπαίδευση κάνουμε unfreeze τα βάρη ώστε οι τιμές τους να ανανεώνονται κανονικά, εκτελώντας έτσι fine-tuning πάνω στα σύνολα δεδομένων που εξετάζουμε. Παρακάτω παρουσιάζουμε ένα γενικό διάγραμμα για την αρχιτεκτονική ResNet101 που χρησιμοποιούμε.

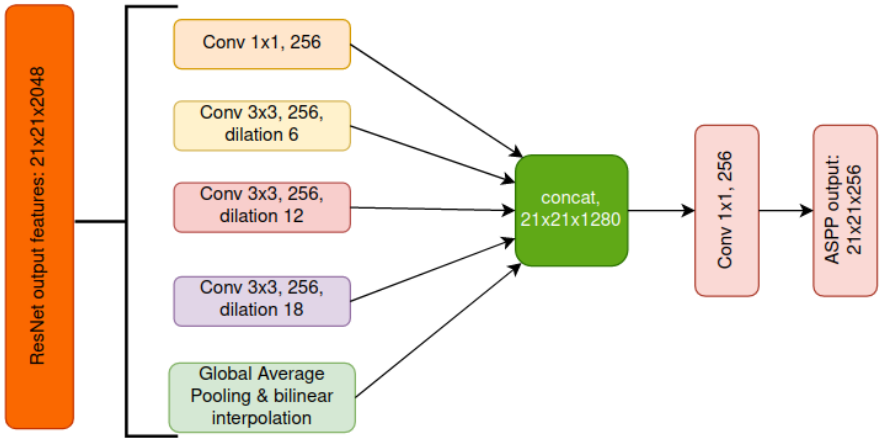


Εικόνα 5.1: *Η αρχιτεκτονική ResNet101[23]*

Όπως φαίνεται και στην εικόνα 5.1, έχουμε αφαιρέσει την κεφαλή κατηγοριοποίησης (classification head) του resnet και στην έξοδο παίρνουμε τα χαρακτηριστικά που παράγει η τελευταία ομάδα συνελίξεων της αρχιτεκτονικής όπως ορίζεται στο [23] την οποία έχουμε ονομάσει layer4. Επίσης, βλέπουμε ότι για μία εικόνα διαστάσεων $(321, 321)$, τα χαρακτηριστικά της εξόδου είναι χωρικής διαστατικότητας $(21, 21)$, κάτι που σημαίνει ότι το δίκτυο κάνει μία συμπίεση της τάξης 1 προς 16 (stride εξόδου).

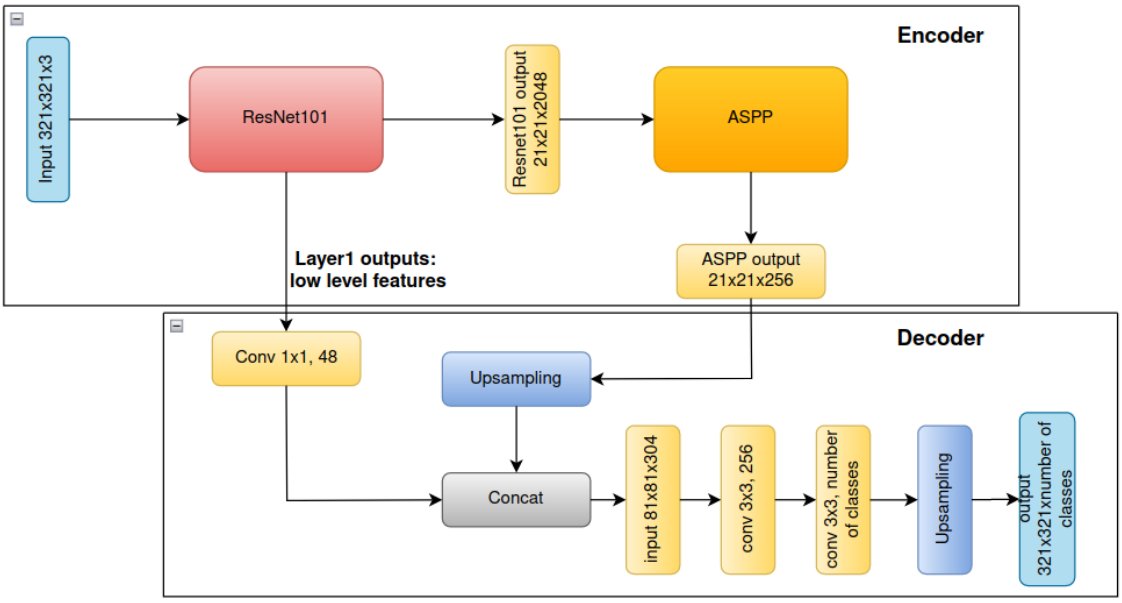
Η δομή ASPP Για τη δομή ASPP κάνουμε χρήση διευρημένων συνελίξεων (dilated convolutions) με ρυθμούς $r = 6, 12, 18$, με μέγεθος πυρήνα (kernel) $3x3$, οι οποίες εφαρμόζονται στα χαρακτηριστικά εξόδου του ResNet101, έτσι ώστε να διευκολυνθεί ο εντοπισμός περιεχομένου πολλαπλών κλιμάκων (multiscale context). Επιπλέον εκτελούμε μία συνέλιξη με μέγεθος πυρήνα $1x1$ (pointwise[8]) και ένα Global Average Pooling, το οποίο ουσιαστικά υπολογίζει το μέσο όρο όλων των χαρακτηριστικών. Οι έξοδοι που παράγονται από κάθε συνέλιξη του ASPP είναι διαστατικότητας $21x21x256$. Τα παραγόμενα χαρακτηριστικά ενώνονται με αποτέλεσμα να φτάνουν τη διαστατικότητα τους σε $5 * 256 = 1280$. Εν συνεχεία σε αυτά εφαρμόζεται μία συνέλιξη $1x1$, η οποία επαναφέρει τη διαστατικότητα σε $21x21x256$.

Το δίκτυο decoder Ο αποκωδικοποιητής, πρώτα εφαρμόζει μία $1x1$ συνέλιξη στα χαρακτηριστικά χαμηλού επιπέδου (low level features) που παράγονται στην έξοδο layer1 του δικτύου ResNet101, προκειμένου να μειωθεί ο αριθμός των καναλιών από 256 σε 48. Αυτό, όπως έχουμε πει γίνεται προκειμένου να φιλτραριστούν σε έναν βαθμό τα υψηλης διαστατικότητας χαρακτηριστικά που εξάγονται από τα αρχικά επίπεδα, για τα οποία αν διατηρηθεί η αρχική διαστατικότητα, πιθανόν να δυσκολεύουν την εκπαίδευση. Η προσθήκη των χαμηλού επιπέδου χαρακτηριστικών βοηθά το μοντέλο να μπορεί να εντοπίσει καλύτερα τα όρια των αντικειμένων, όπως ακμές, γραμμές κ.λ.π. Στη συνέχεια, η έξοδος του ASPP υπερδειγματοληπτείται, προκειμένου να ταιριάξει σε χωρικές διαστάσεις με τα χαμηλού επιπέδου



Εικόνα 5.2: *Η δομή ASPP [6]*

χαρακτηριστικά $(81, 81)$ με τα οποία συνενώνεται. Έπειτα, μετά από κάποιες ακόμα συνελίξεις και μία τελική υπερδειγματοληψία παράγεται το αποτέλεσμα κατάτμησης. Παρακάτω, βλέπουμε ένα γενικό σχήμα της αρχιτεκτονικής. Η χωρικές διαστάσεις των χαρακτηριστικών προκύπτουν για παράδειγμα εισόδου χωρικών διαστάσεων $(321, 321)$.



Εικόνα 5.3: *Η δομή του Deeplabv3plus για είσοδο διαστατικότητας* $(321, 321)$ [8]

5.3 Επιβλεπόμενη προσέγγιση (Supervised approach)

Όπως έχουμε αναφέρει κατα την εφαρμογή επιβλεπόμενης μάθησης για την επίλυση του προβλήματος της κατάτμησης εικόνας για τα σύνολα δεδομένων που έχουμε παρουσιάσει, γίνεται χρήση αποκλειστικά των επισημασμένων δειγμάτων (labeled samples) για κάθε διαμέριση που δημιουργήσαμε.

Πιο συγκεκριμένα, ακολουθώντας το συνηθισμενο πρότυπο εκπαίδευσης για την αμειγώς επιβλεπόμενη μάθηση, τα επισημασμένα δεδομένα τροφοδούνται στο δίκτυο κατάτμησης DeepLabv3plus, το οποίο παράγει τις αντίστοιχες μάσκες κατάτμησης (segmentation maps), οι οποίες επιβλέπονται από τις διαθέσιμες ετικέτες μέσω του κόστους cross-entropy, το οποίο υπολογίζεται σε επίπεδο pixel (pixel-wise cross entropy).

Όπως είναι γνωστό, τα δεδομένα δίνονται στο δίκτυο σε δέσμες (batches). Σε κάθε δέσμη δεδομένων, πριν δοθεί ως είσοδος στο δίκτυο, πραγματοποιούνται ορισμένοι μετασχηματισμοί επαύξησης (data augmentation transforms) πάνω στα δείγματα που περιέχει. Οι συγκεκριμένοι μετασχμηματισμοί είναι στοχαστικοί. Συνεπώς, σε κάθε δέσμη που δέχεται το δίκτυο ως είσοδο κατά τη διάρκεια μίας εποχής εκπαίδευσης, περιέχονται κάθε φορά διαφορετικές επαυξημένες εκδοχές των εικόνων (τυχαιότητα στη εφαρμογή των μετασχηματισμών), γεγονός που αυξάνει τη ποικιλότητα των δειγμάτων που ¨βλέπει¨ το δίκτυο σε κάθε επανάληψη εκπαίδευσης (training iteration), βοηθώντας το να αποφύγει την υπερεκπαίδευση (overfitting) και να ενισχύσει την ικανότητα γενίκευσης του. Παρακάτω παρουσιάζουμε τους μετασχηματισμούς που εφαρμόζονται στις εικόνες. Να σημειωθεί εδώ ότι ακολουθούμε τις συνήθεις τεχνικές επαύξησης που χρησιμοποιούνται και στα [16, 82, 39].

- Κανονικοποίηση των τιμών των pixels στο εύρος $[0, 1]$ και στη συνέχεια κανονικόποιηση κάθε καναλιού της εικόνας με βάση τη μέση τιμή $\mu = [0.485, 0.456, 0.406]$ και τη τυπικη απόκλιση $\sigma = [0.229, 0.224, 0.225]$ του σύνολου δεδομένων ImageNet. Αυτό αποτελεί μία συνήθης πρακτική όταν το δίκτυο αρχικοποιείται με βάρη προεκπαιδευμένα στο ImageNet.

- Τυχαία αλλαγή κλίμακας (μεγέθους) της εικόνας (Random Scaling)[89]. Ο μετασχηματισμός καθορίζεται από έναν παράγοντα, έστω a, ο οποίος εκφράζει το πόσες φορές θα μειωθούν ή θα αυξήθουν οι διαστάσεις μιας εικόνας. Συνήθως, η παράμετρος a επιλέγεται τυχαία από ένα εύρος πραγματικών αριθμών. Για παράδειγμα αν υποθέσουμε ότι επιλέγεται τυχαία κάποια τιμή του a από το διάστημα $[minscale, maxscale2]$, τότε η νέα εικόνα θα έχει διαστάσεις (aH, aW).

- Τυχαιά περικοπή της εικόνας (Random Cropping)[89]. Εφαρμόζεται, αμέσως μετά την τυχαία αλλαγή κλίμακας και αποκόπτει τυχαία μια περιοχή της εικόνας (image patch) μεγέθους $(CropSizeHeight, CropSizeWidth)$. Στην περίπτωση που οι διαστάσεις αποκοπής είναι μεγαλύτερες από τις διαστάσεις της εικόνας τότε ο εναπομείναντας χώρος γεμίζεται με μηδενικά (image padding).

- Οριζόντια ανατροπή της εικόνας (Horizontal flipping)[89]. Η εικόνα αναστρέφεται γύρω από τον y άξονα.

- Μετασχηματισμοί χρώματος με προσαρμογές στη φωτεινότητα (brightness), την αντίθεση (contrast), τον κορεσμό (saturation), και τη χροιά (hue) της εικόνας. Οι συγκεκριμένοι μετασχηματισμοί χρώματος γίνονται με χρήση της συνάρτησης *ColorJitter*[24] της Pytorch

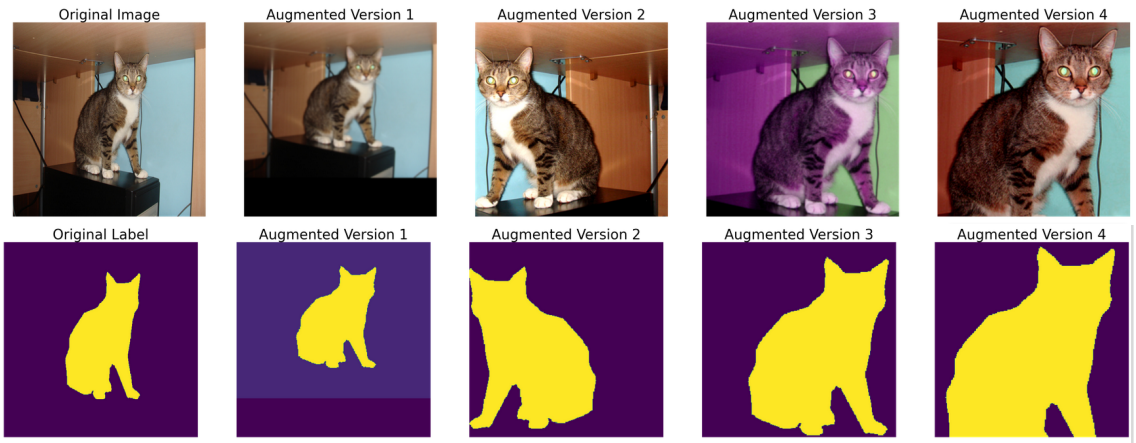
- Γκαουσιανό θόλωμα της εικόνας (gaussian blurring[90])

Αλγοριθμος 5.1: *Γενική συνάρτηση βασικών μετασχηματισμών*

Είσοδος: I (*εικόνα εισόδου*)
Έξοδος: I^{tr} (*μετασχηματισμένη εικόνα*)
  $I^{tr} = I$
  $I^{tr} = RandScaling(I^{tr})$, $I^{tr} = RandCropping(I^{tr})$
  if $rand.uniform() \leq 0.5$ **then**
    $I^{tr} = HorizontalFlip(I^{tr})$
  end if
  if $rand.uniform() \leq 0.5$ **then**
    $I^{tr} = ColorJitter(I^{tr})$
  end if
  if $rand.uniform() \leq 0.5$ **then**
    $I^{tr} = GaussianBlurr(I^{tr})$
  end if
  return I^{tr}

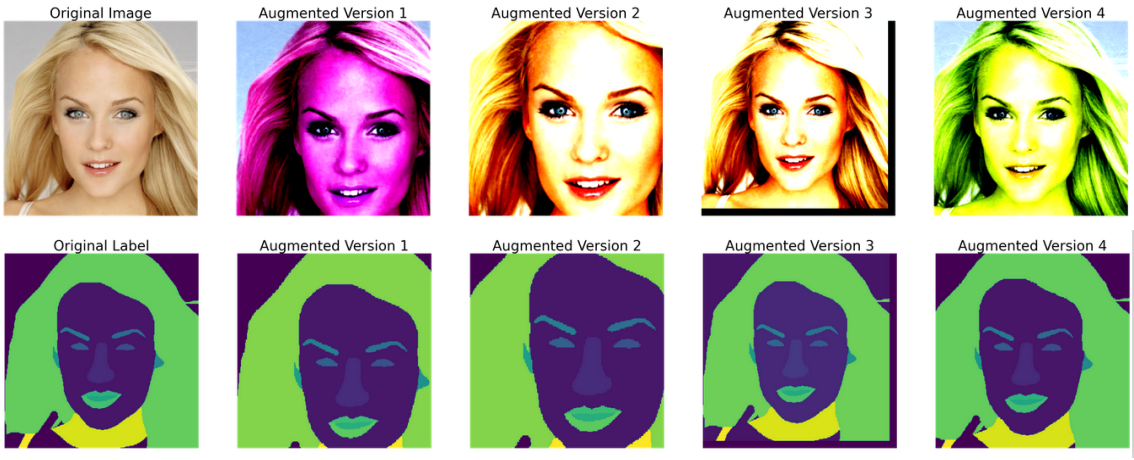
Παρακάτω παραθέτουμε για κάθε σύνολο δεδομένων, τους βασικούς μετασχηματισμούς που εφαρμόσαμε σε κάθε περίπτωση, καθώς και παραδείγματα πιθανών επαυξημένων εκδοχών που μπορεί να προκύψουν. Να τονίσουμε ότι στην περίπτωση των χωρικών μετασχηματισμών (random scaling, random cropping, horizontal flipping) χρειάζεται να πραγματοποιηθούν οι ίδιοι μετασχηματισμοί και στις αντίστοιχες ετικέτες των εικόνων εισόδου.

- **Pascal**: Για το σύνολο δεδομένων Pascal επιλέγουμε να εφαρμόσουμε τυχαία αλλαγή κλίμακας (random scaling) με παράμετρο $a \in [0.5, 1.8]$, ενώ για τη τυχαία αποκοπή (random cropping) ορίζουμε το μέγεθος της περιοχής αποκοπής (patch) σε $(321, 321)$. Επίσης, εφαρμόζουμε οριζόντια αναστροφή (horizontal flipping). Οι παράμετροι για τους μετασχηματισμους χρώματος (color jittering) και το γκαουσιανό θόλωμα (gaussian blurring) τους δανειζόμαστε από τις δουλειές [16][82].



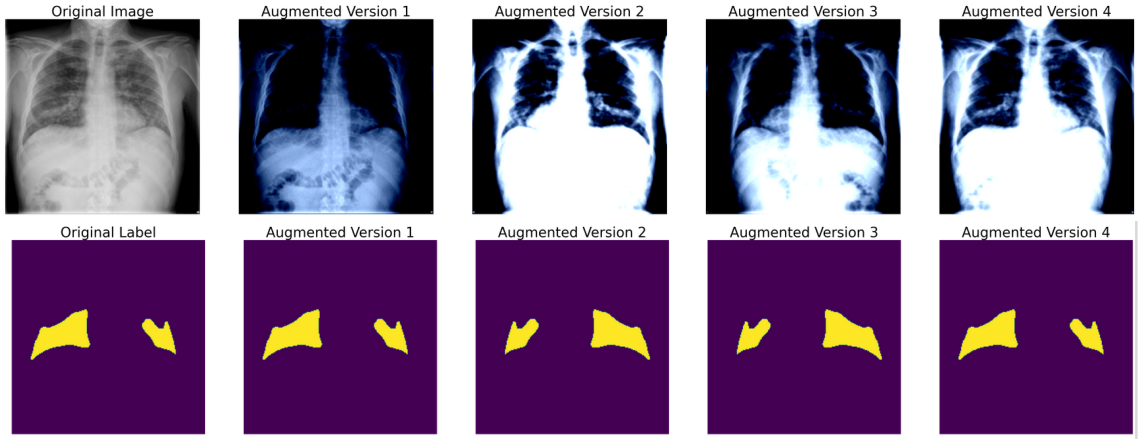
Εικόνα 5.4: *Πιθανές επαυξημένες εκδοχές και αντίστοιχες ετικέτες κάποιου τυχαίου δείγματος από το σύνολο Pascal*

- **CelebAMask-HQ**: Τυχαία αλλαγή κλίμακας με $a \in [0.8, 1.5]$ και τυχαία αποκοπή μεγέθους $(321, 321)$. Όμοίως, μετασχηματισμοί χρώματος και θολώματος από [16][82].



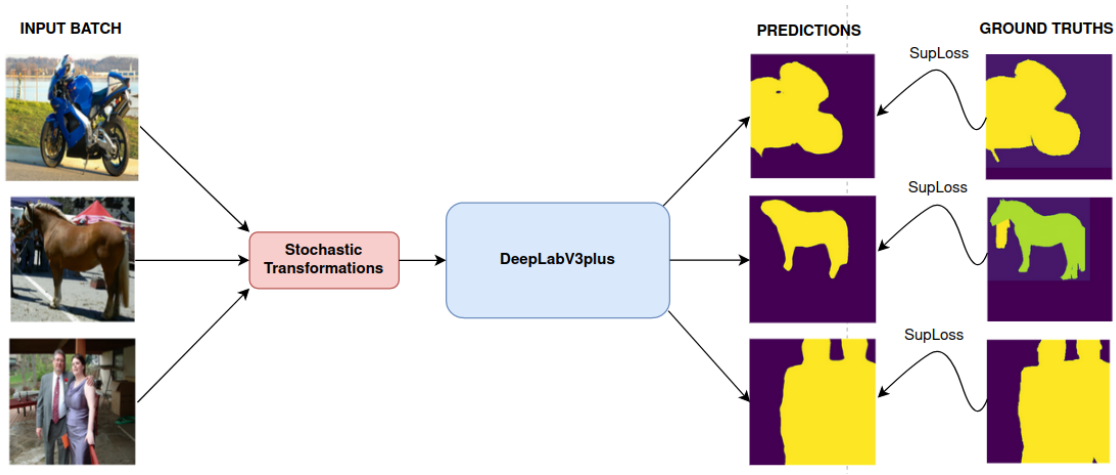
Εικόνα 5.5: *Πιθανές επαυξημένες εκδοχές και αντίστοιχες ετικέτες κάποιου τυχαίου δείγματος από το σύνολο CelebAMask-HQ*

- **QaTa-COV19 Dataset**: Εδώ εφαρμόζουμε τυχαιά αναστροφή εικόνας (horizontal flipping) και μετασχηματισμούς χρώματος και θολώματος από [16][82]



Εικόνα 5.6: *Πιθανές επαυξημένες εκδοχές και αντίστοιχες ετικέτες κάποιου τυχαίου δείγματος από το σύνολο QaTa-COV19 Dataset*

Στο παρακάτω σχήμα απεικονίζεται αφαιρετικά το γενικό παράδειγμα της επιβλεπόμενης προσέγγισης, όπου αξιοποιούμε αποκλειστικά τα επισημασμένα δεδομένα κατα την εκπαίδευση για την κάθε διαμέριση.



Εικόνα 5.7: *Γενικό παράδειγμα της αμειγώς επιβλεπόμενης προσέγγισης*

Υποθέτουμε ότι έχουμε μία δέσμη από εικόνες *RGB* (batch) μεγέθους B, έστω $X = [x_1, x_2, ..., x_B]$, $X \in R^{BxHxWx3}$ με τις αντίστοιχες ετικέτες $Y = [y_1, y_2, ..., y_B], Y \in R^{BxHxWxC}$ (C ο αριθμός των σημασιολογικών κλάσεων) σε one-hot μορφή και το δίκτυο κατάτμησης f_∂. Επιπλέον, θεωρούμε τις προβλέψεις του δίκτυου για την δέσμη εισόδου X, έστω $P = [f_\partial(x_1), f_\partial(x_2), ..., f_\partial(x_B)]$ με $P \in R^{BxHxWxC}$. Το κόστος cross-entropy σε επίπεδο pixel για τη τρέχουσα δέσμη εικόνων μπορεί να γραφτεί ως εξής:

$$L_{sup}(\Upsilon, P) = \frac{1}{B}\sum_{b=1}^{B} -\frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} Y_{b,i,j,c} \cdot \log(P_{b,i,j,c}) \tag{5.1}$$

Άν θεωρήσουμε ότι το σύνολο δεδομένων έχει χωριστεί και N δέσμες μεγέθους B, τότε το συνολικό κόστος θα είναι απλά ο μέσος όρος των ποσοτήτων $L_{sup}(\Upsilon, P)$ για τις N δέσμες.

$$L_S = \frac{1}{N}\sum_{n=1}^{N} L_{sup}(Y_n, P_n) \tag{5.2}$$

Για την επιβλεπόμενη εκπαίδευση του δικτύου ορίζουμε και για τα τρία σύνολα δεδομένων το μέγεθος δέσμης (batch size) στις 20 εικόνες, ενώ κάνουμε χρήση του βελτιστοποιητή (optimizer) SGD. Επιπλέον για το σύνολο Pascal εκπαιδεύουμε το δίκτυο για 40.000 ενημερώσεις της κλίσης (gradient updates, training iterations), ενώ για τα σύνολα CelebAMask-HQ, QaTa-COV19 20.000 ενημερώσεις. Όσον αφορά το ρυθμό εκμάθησης (learning rate) επιλέγουμε να χρησιμοποιήσουμε την πολυωνυμική πολιτική (polynomial learning rate policy)[91] με αρχικούς ρυθμούς 0.002 για το Pascal και 0.01 για τα CelebAMask-HQ, QaTa-COV19.

5.4 Ημιεπιβλεπόμενη προσέγγιση (Semi-Supervised approach)

Προκειμένου να αξιοποιήσουμε τα επιπλέον μη επισημασμένα δεδομένα με βάση τις διαμερίσεις που παρουσιάσαμε στην παράγραφο 5.1 και να δούμε σε τι βαθμό μπορούν να συνεισφέρουν στην απόδοση του μοντέλου για τα τρία σύνολα δεδομένων, επιλέγουμε να εξετάσουμε μεθόδους που βασίζονται σε διαταραχές επιπέδου εισόδου (input-level perturbations). Πιο συγκεκριμένα, ακολουθώντας το παράδειγμα της ασθενούς προς ισχυρής συνέπειας (weak-to-strong consistency)[14, 13] και του FixMatch [13], πειραματιζόμαστε με διάφορες τεχνικές ισχυρής επαύξησης στα μη επισημασμένα δεδομένα και πως κάθε μία απο αυτές επιδρά στην εκπαίδευση του δικτύου. Ιδιαίτερη έμφαση δίνουμε σε τεχνικές επαύξησης που περιλαμβάνουν μίξη δεδομένων, όπως οι ClassMix[16] και CutMix[15].

5.4.1 Οι υποθέσεις συστάδας (cluster) και χαμηλής πυκνότητας διαχωρισμού (low density separation) σε προβλήματα σημασιολογικής κατάτμησης

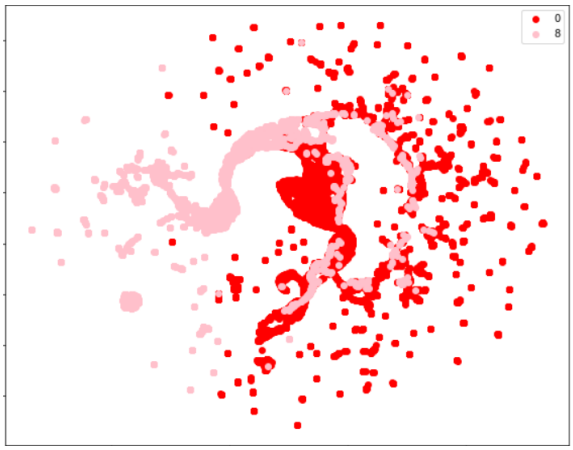
Στην παράγραφο 2.3.1, διατυπώνουμε τις υποθέσεις συστάδας και διαχωρισμού χαμηλής πυκνότητας, αναφέροντας θεωρητικά ότι πρέπει να ικανοποιούνται για την κατανομή εισόδου των δεδομένων, έτσι ώστε η εφαρμογή ημιεπιβλεπόμενης μάθησης να βοηθάει στην καλύτερη επίλυση του προβλήματος. Παρόλα αυτά η συντριπτική πλειοψηφία προβλημάτων μηχανικής μάθησης που συναντόνται στην πράξη διαθέτουν υψηλής διαστατικότητας (high-dimensional) δεδομένα τα οποία δεν σχηματίζουν συστάδες, οι οποίες χωρίζονται από περιοχές χαμηλής πυκνότητας στο χώρο εισόδου. Όσον αφορά τα προβλήματα σημασιολογικής κατάτμησης στη δουλειά των Geoff French et al. [15] αναφέρεται ότι οι συγκεκριμένες συνθήκες δεν ισχύουν για την κατανομή εισόδου των pixels της εικόνας προς κατάτμηση. Αυτό δεν σημαίνει όμως, ότι δεν μπορεί να εφαρμοστεί η κανονικοποίηση συνέπειας με επιτυχία σε τέτοιου είδους προβλήματα κατάτμησης. Οι μέθοδοι κανονικοποίησης συνέπειας εκπαιδεύουν το δίκτυο να παράγει όμοιες εξόδους για μία εικόνα και τη διαταραγμένη (perturbed) εκδοχή αυτής, αυξάνοντας έτσι την ευστάθεια του και την ικανότητα γενίκευσης. Επιπλέον, τα βαθιά δίκτυα κατάτμησης αρχιτεκτονικής encoder-decoder έχουν τη δυνατότητα να παράγουν ενδιάμεσες αναπαραστάσεις χαρακτηριστικών για την εικόνα εισόδου και να κατηγοριοποιούν το κάθε pixel σωστά ανεξάρτητα τη μορφή της κατανομής εισόδου.

Παρακάτω, παραθέτουμε δύο παραδείγματα εικόνων από τα σύνολα Pascal και CelebAMask-HQ μαζί με τον χάρτη βεβαιότητας (confidence map) που παράγει το δίκτυο κατάτμησης. Θέτοντας ένα κατώφλι (threshold π.χ 0.80) φιλτράρουμε τους χάρτες εντοπίζοντας έτσι τις περιοχές με pixels χαμηλής βεβαιότητας (low confidence). Βλέπουμε, ότι οι περιοχές χαμηλής βεβαιότητας ταυτίζονται σε ικανοποιητικό βαθμό με τα όρια των κλάσεων (class boundaries) που περιέχονται στην εικόνα. Γενικά μπορούμε να συμπεράνουμε ότι τα pixels που βρίσκονται στα όρια των κλάσεων είναι αυτά που παρουσιάζουν τη μεγαλύτερη αβεβαιότητα και να υποθέσουμε ότι αυτά τα pixels βρίσκονται σε περιοχές χαμηλής πυκνότητας (low density regions) στο χώρο των χαρακτηριστικών εξόδου, οι οποίες χωρίζουν τις περιοχές υψηλής πυκνότητας (high density regions), όπου πιθανώς να βρίσκονται τα pixels με υψηλή βεβαιότητα (high confidence).

Εικόνα 5.8: *Παράδειγμα κατωφλιομένου χάρτη βεβαιότητας για εικόνα από το σύνολο Pascal*

Εικόνα 5.9: *Παράδειγμα κατωφλιομένου χάρτη βεβαιότητας για εικόνα από το σύνολο CelebAMask-HQ*

Το παραπάνω συμπέρασμα μπορεί να διαπιστωθεί σε ένα βαθμό απο την οπτικοποίηση που πραγματοποιούμε για τις δύο παραπάνω εικόνες σε επίπεδο κατανομής pixel εισόδου και σε επίπεδο των χαρακτηριστικών που παράγονται μετά το συνελικτικό επίπεδο $3x3, 256$ του decoder, τα οποία έχουν 256 κανάλια. Η οπτικοποίηση στις δύο διαστάσεις γίνεται με τη μέθοδο μείωση διαστατικότητας UMAP [92].

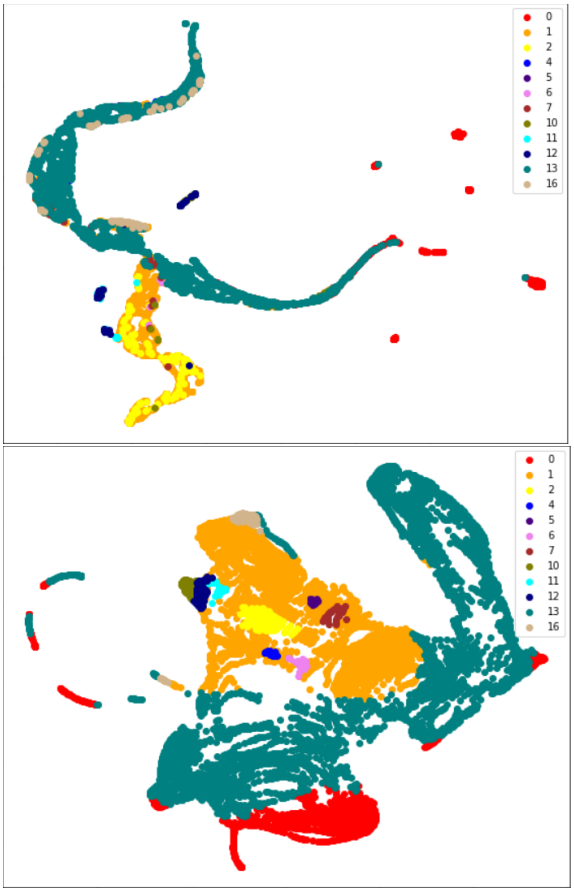


Εικόνα 5.10: *Οπτικοποίηση της κατανομής εισόδου των pixels για το παράδειγμα από το Pascal με χρήση UMAP*

Βλέπουμε ότι τα pixels εισόδου δεν σχηματίζουν εμφανείς συστάδες σε σχέση με τα χαρακτηριστικά των pixels που παράγονται από το προτελευταίο συνελικτικό επίπεδο του decoder, όπου διακρίνουμε πιο εμφανή συσταδοποίηση.

Εικόνα 5.11: *Οπτικοποίηση της κατανομής εισόδου των χαρακτηριστικών των pixels του προτελευταίου συνελικτικού επιπέδου του decoder για το παράδειγμα από το Pascal με χρήση UMAP*

Παρόμοια συμπεριφορά διακρίνουμε και για την εικόνα από το CelebAMask-HQ.



Εικόνα 5.12: *Οπτικοποίηση της κατανομής εισόδου των pixels και των χαρακτηριστικών που παράγει ο decoder για το παράδειγμα από το CelebAMask-HQ με χρήση της τεχνικής μείωσης διαστατικότητας UMAP*

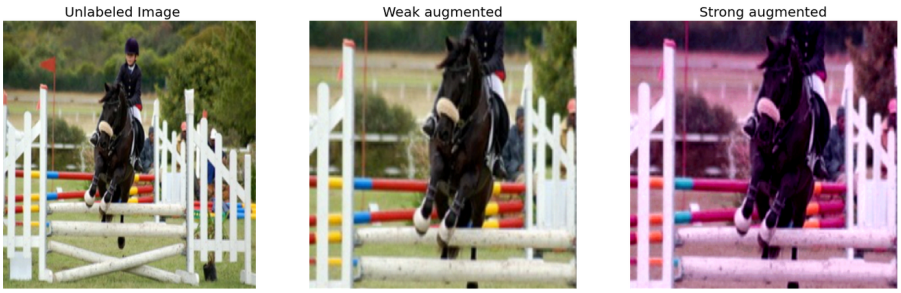
5.4.2 Εφαρμογή ισχυρών διαταραχών επιπέδου εισόδου (strong input-level perturbations)

Η γενική ιδέα μας είναι να αξιοποιήσουμε το παράδειγμα εκπαίδευσης του FixMatch[13, 14] και της ασθενούς-ισχυρής συνέπειας (weak-to-strong consistency) για τη σημασιολογική κατάτμηση και να πειραματιστούμε με διάφορες διαταραχές που μπορούμε να εφαρμόσουμε για την ισχυρή επαύξηση των μη επισημασμένων εικόνων εισόδου. Βασική συνιστώσα εδώ αποτελεί το μοντέλο student-teacher. Όπως έχουμε αναφέρει οι προβλέψεις που παράγει ο teacher για τις ασθενώς επαυξημένες μη επισημασμένες εικόνες, μετατρέπονται σε ψευδοετικέτες, οι οποίες χρησιμοποιούνται για την επίβλεψη των εξόδων του student, ο οποίος δέχεται ως είσοδο ισχυρά επαυξημένες εικόνες. Η οπισθοδιάδοση σφάλματος πραγματοποιείται αποκλειστικά κατά μήκος του δικτύου student. Η λογική πίσω από αυτό, είναι ότι για τις ασθενώς επαυξημένες εικόνες, οι οποίες περιλαμβάνουν λιγότερο έντονη επαύξηση, το δίκτυο teacher έχει την ικανότητα να παράγει πιο εύκολα προβλέψεις υψηλής βεβαιότητας (high confidence predictions) για αυτές. Αντιθέτως, για τις εντονότερα επαυξημένες εικόνες, δηλαδή τις ισχυρά επαυξημένες είναι πιο δύσκολο να παραχθούν αποτελέσματα υψηλής ποιότητας. Συνεπώς, στόχος είναι να εκπαιδεύσουμε το δίκτυο student να μάθει να παράγει υψηλής ποιότητας προβλέψεις για αυτές, αξιοποιώντας τις ψευδοετικέτες καλής ποιότητας που παράγει ο teacher για τις ασθενώς επαυξημένες εικόνες.

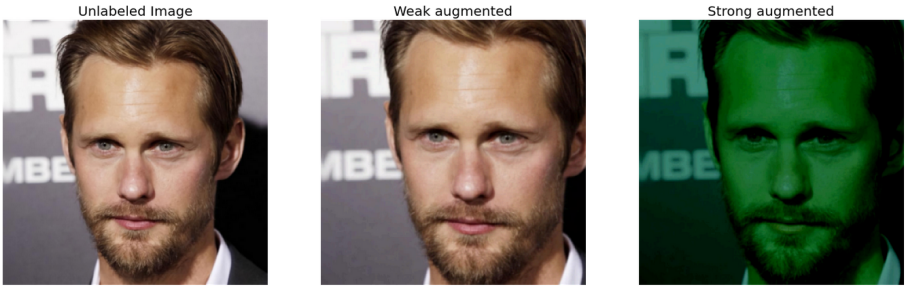
Για να έχει το παραπάνω σχήμα εκπαίδευσης καλά αποτελέσματα, αξιοποιώντας τα μη επισημασμένα δεδομένα προς βελτίωση της απόδοσης, χρειάζεται να επιλεχθούν με προσοχή οι μέθοδοι ισχυρής επαύξησης, οι οποίες έχουν κύριο ρόλο για την εξέλιξη της εκπαίδευσης. Είναι σημαντικό το μοντέλο να εκτεθεί σε διαφορετικές παραλλαγές των δεδομένων, ώστε να γίνει ανθεκτικό σε μεταβολές που μπορεί να συναντήσει και να μάθει να γενικεύει σε νέα δείγματα, βελτιώνοντας την απόδοση του. Έν συνεχεία παραθέτουμε τα είδη των διαταραχών που χρησιμοποιούμε για τον κλάδο του student (ισχυρός κλάδος).

Ισχυρή επαύξηση με μετασχηματισμούς χρώματος (Color Augmentation) και θολώματος (Blurring)

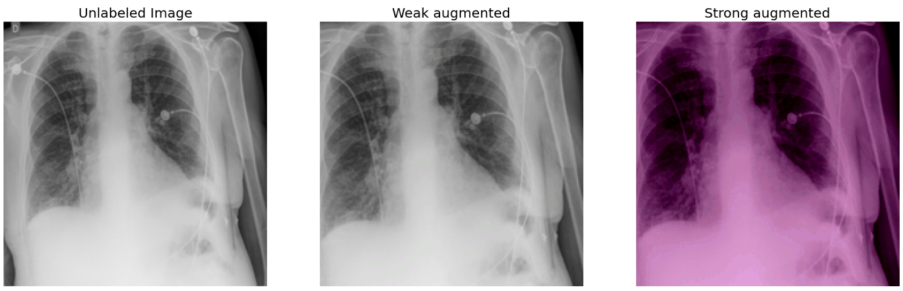
Εδώ κάνουμε χρήση της συνάρτησης ColorJittering[24], η οποία προσαρμόζει τυχαιά κάθε φορά τις παραμέτρους της φωτεινότητας (brightness), της αντίθεσης (contrast), του κορεσμού (saturation), και της χροιάς (hue) της εικόνας. Επιπλέον, χρησιμοποιούμε τη συνάρτηση Gaussian Blurring της Pytorch [25] που εφαρμόζει γκαουσιανό θόλωμα [90]. Οι μετασχηματισμοί χρώματος που εφαρμόζει η ColorJittering, μεταβάλλουν ικανοποιητικά την εμφάνιση της εικόνας, επιτρέποντας στο δίκτυο να μάθει να αναγνωρίζει τα αντικείμενα όχι μόνο βασιζόμενο στο σύνηθες χρώμα που μπορεί να έχουν, απλά απομνημονεύοντας το [39]. Από την άλλη, η εφαρμογή γκαουσιανού θολώματος κάνει πιο απαιτητική την αναγνώριση, καθώς οι ακμές των αντικειμένων γίνονται λιγότερο ξεκάθαρες. Οι παραπάνω μετασχηματισμοί αυξάνουν την αβεβαιότητα του δικτύου, ωθώντας το να μαθαίνει πιο γενικές αναπαραστάσεις από τα μη επισημασμένα δεδομένα. Όσον αφορά την ασθενή επαύξηση εφαρμόζουμε τυχαία αλλαγή κλίμακας (RandomScaling) [25] και τυχαία εξαγωγή περιοχής (Random Cropping) [89]. Παρακάτω παραθέτουμε παραδείγματα ασθενούς και ισχυρής επαύξησης (Color jitter, Gaussian Blur) για τα τρία σύνολα δεδομένων.



Εικόνα 5.13: *Παραδείγματα ασθενούς και ισχυρής επαύξησης για το σύνολο Pascal*



Εικόνα 5.14: *Παραδείγματα ασθενούς και ισχυρής επαύξησης για το σύνολο CelebAMask-HQ*



Εικόνα 5.15: *Παραδείγματα ασθενούς και ισχυρής επαύξησης για το σύνολο QaTa-COV19*

Παρακάτω παρουσιάζουμε τον τρόπο που αξιοποιούνται τα μη επισημασμένα δεδομένα κατά την εκπαίδευση. Όπως, βλέπουμε η ισχυρή επαύξηση εφαρμόζεται πάντα πάνω στην ασθενώς επαυξημένη εικόνα. Το δίκτυο teacher, όπως έχουμε πεί παράγει προβλέψεις για τις ασθενώς επαυξημένες εικόνες, οι οποίες μετατρέπονται σε ψευδοετικέτες που επιβλέπουν την έξοδο του student. Όπως φαίνεται τα βάρη του teacher είναι ο εκθετικός κινητός μέσος όρος (EMA [68]) των βαρών του student. Για αυτό και δεν εκτελούμε οπισθοδιάδοση σφάλματος κατά μήκος του κλάδου του teacher. Όσον αφορά τα επισημασμένα δεδομένα, ακολουθείται ακριβώς η μεθοδολογία που περιγράψαμε στην παράγραφο 5.3 για την αξιοποίηση τους. Κάνουμε χρήση κάποιων συμβολισμών από [64]. Ο μη επιβλεπόμενος όρος της συνάρτησης κόστους μπορεί να γραφτεί ως εξής, όπου L_{CE} το κόστος cross-entropy:

$$L_U = \frac{1}{|D_u|} \sum_{x \in D_u} L_{CE}(y^{pseudo}, f_{\partial}^{student}(x^{strong})) \tag{5.3}$$

Το συνολικό κόστος μπορεί να γραφτεί ώς:

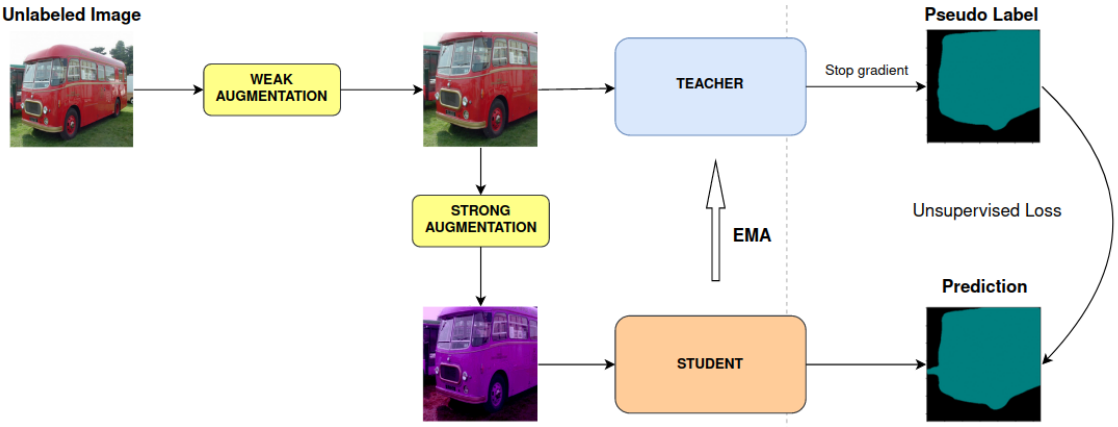
$$L_{total} = L_S + \lambda \cdot L_U$$

Ο όρος L_S αποτελεί το κόστος cross-entropy, όπως το έχουμε διατύπωσει στην εξίσωση 5.2. Ως βάρος συνεισφοράς λ του μη επιβλεπόμενου όρου έχουμε επιλέξει ακολουθώντας τα [16, 15, 82, 93], όπου ορίζεται ως το ποσοστό των pixels των ψευδοετικετών, τα οποία έχουν βεβαιότητα (confidence) πάνω από ένα υψηλό κατώφλι (π.χ 0.96). Δηλαδή, αν θεωρήσουμε μία ψευδοετικέτα p, τότε το λ για την τρέχουσα ψευδοετικέτα θα είναι:

$$\lambda = \frac{1}{H \cdot W} \sum_i \sum_j \mathbf{1}[max(p_{ij} \geq \tau_{thres})] \tag{5.4}$$

, όπου p η ψευδοετικέτα που παράγει ο teacher για κάποια εικόνα και $H \cdot W$, το συνολικό πλήθος pixels στην τρέχουσα εικόνα.

Με αυτό τον τρόπο καταφέρνουμε να μην επηρεάζεται το μοντέλο από θορυβώδεις ψευδοετικέτες που παράγονται στην αρχή της εκπαίδευσης.



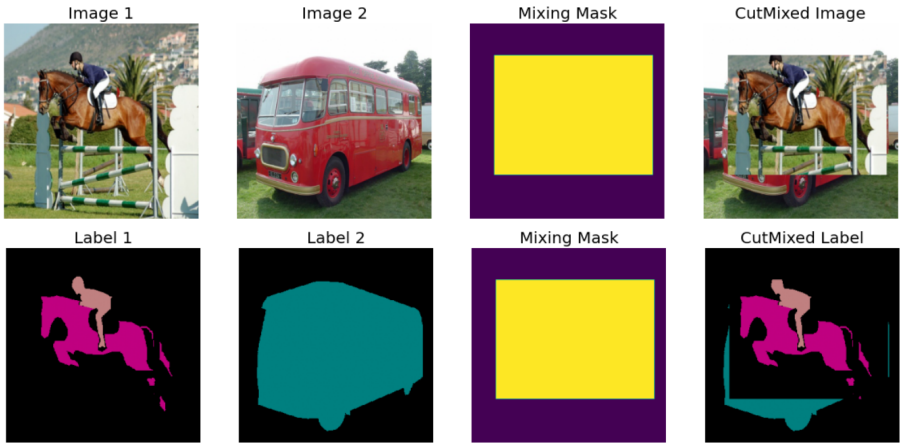
Εικόνα 5.16: *Ο μη επιβλεπόμενος κλάδος για τη περίπτωση εφαρμογής μετασχηματών χρώματος και θολώματος ως τεχνικές ισχυρής επαύξησης*

Ισχυρή επαύξηση με CutMix και ClassMix

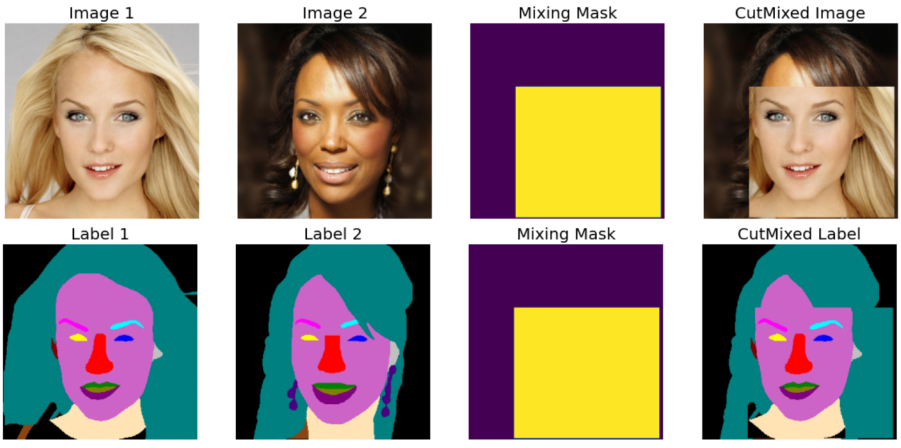
Οι μέθοδοι CutMix [15] και ClassMix [16], όπως έχουμε αναφέρει και στην παράγραφο 3.1.1, αποτελούν δύο αρκετά διαδεδομένες τεχνικές επαύξησης δεδομένων που αξιοποιούν συνήθως δύο δείγματα και τα αναμειγνύουν με αποτέλεσμα την δημιουργία ενός καινούριου δείγματος. Τα δύο επιμέρους δείγματα αναμειγνύονται με τη βοήθεια μίας δυαδικής μάσκας M, η οποία επιλέγει την περιοχή της μίας εικόνας που θα μεταφερθεί στην άλλη, ώστε να προκύψει μία εντελώς νεά εικόνα. Η διαφορά των δύο μεθόδων έγκειται στο τρόπο παραγωγής της δυαδικής μάσκας M. Για δύο εικόνες x_1, x_2 θα έχουμε:

$$x_{new} = M \cdot x_1 + (1 - M) \cdot x_2$$
$$y_{new} = M \cdot y_1 + (1 - M) \cdot y_2$$

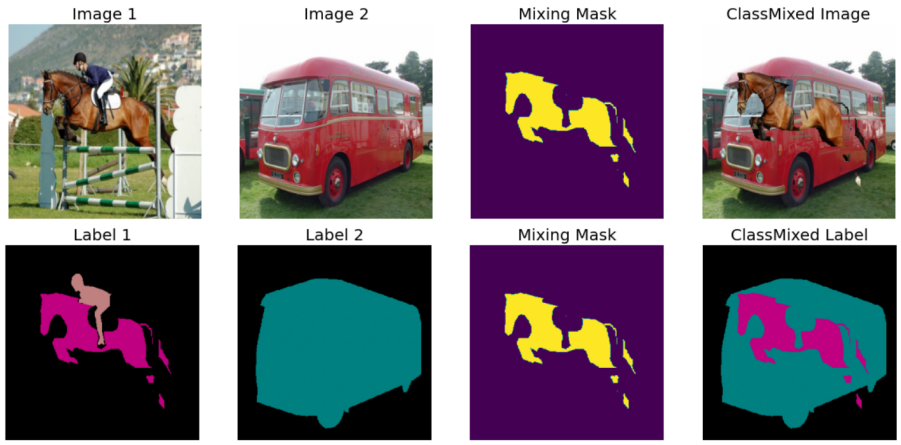
Ο πολλαπλασιασμός που εκτελείται ανάμεσα στη μάσκα M και στις εικόνες είναι element-wise και στόχος είναι να απομονωθεί η περιοχή της πρώτης εικόνας που θα προσαρτηθεί στη δεύτερη. Όπως, έχουμε αναφέρει και στην 3.1.1, στην περίπτωση του Cutmix, εξάγεται τυχαία μία ορθογώνια περιοχή από την πρώτη εικόνα και προστίθεται στη δεύτερη, ενώ στο ClassMix μπορεί να εξάγεται μία ολόκληρη περιοχή που καταλαμβάνει μία σημασιολογική κλάση και να μεταφέρεται στη δεύτερη. Η μορφή των παραγόμενων εικόνων, καθώς και της μάσκας μίξης μπορούν να κατανοηθούν καλύτερα από τα παρακάτω παραδείγματα.



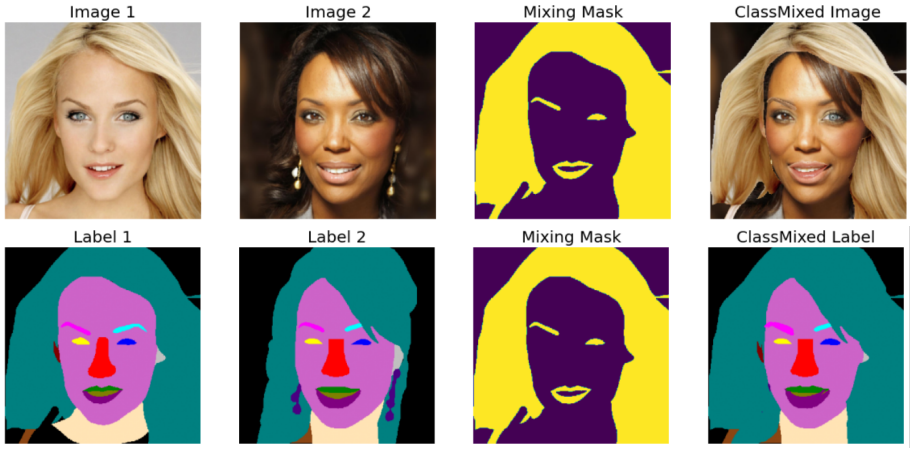
Εικόνα 5.17: *Παράδειγμα CutMix για δύο τυχαίες εικόνες από το σύνολο Pascal*



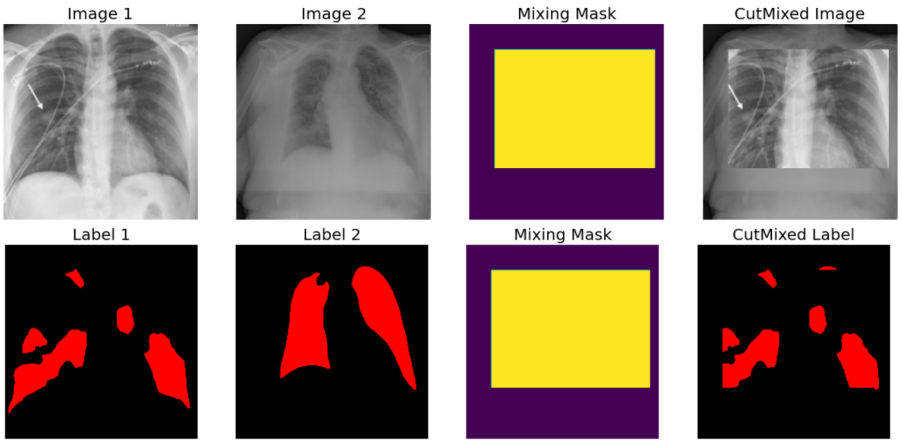
Εικόνα 5.18: *Παράδειγμα CutMix για δύο τυχαίες εικόνες από το σύνολο CelebAMask-HQ*



Εικόνα 5.19: *Παράδειγμα ClassMix για δύο τυχαίες εικόνες από το σύνολο Pascal*

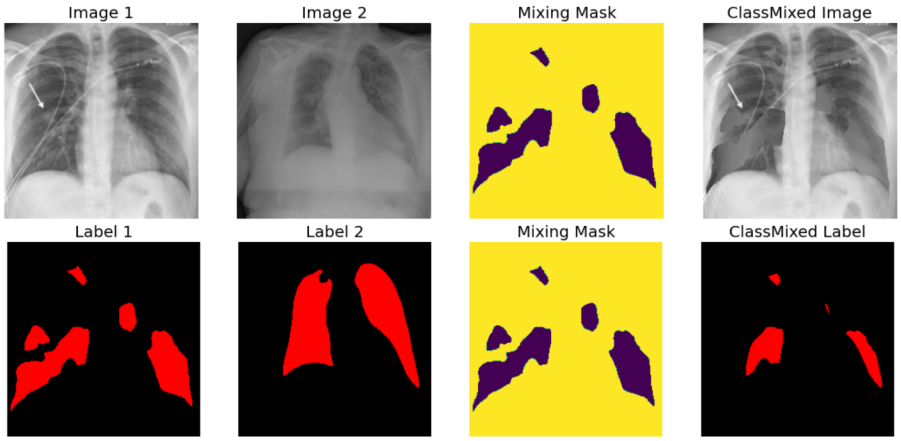


Εικόνα 5.20: *Παράδειγμα ClassMix για δύο τυχαίες εικόνες από το σύνολο CelebAMask-HQ*



Εικόνα 5.21: *Παράδειγμα CutMix για δύο τυχαίες εικόνες από το σύνολο QaTa-COV19*

Όπως βλέπουμε για την περίπτωση του Cutmix παρατηρούμε στις 5.17, 5.18 ότι η δυαδική μάσκα μίξης περιέχει άσους αποκλειστικά στα σημεία που ορίζουν την επιφάνεια της ορθογώνιας περιοχής που αποκόπτεται από την πρώτη εικόνα και χρησιμοποιείται τόσο για την ανάμειξη των εικόνων, όσο και των αντίστοιχων ετικετών. Από την άλλη, όσον αφορά το ClassMix βλέπουμε ότι η δυαδική μάσκα περιέχει άσους στις περιοχές που ορίζονται



Εικόνα 5.22: *Παράδειγμα ClassMix για δύο τυχαίες εικόνες από το σύνολο QaTa-COV19*

από σημασιολογικές κλάσεις στην αρχική εικόνα. Σύμφωνα με τον αλγόριθμο 3.1 κάθε φορά μεταφέρονται οι μισές κλάσεις που υπάρχουν από την ετικέτα της πρώτης εικόνας, στη δεύτερη εικόνα. Για παράδειγμα στην 5.19, η εικόνα από το σύνολο Pascal περιέχει τρείς κλάσεις (background, person, horse). Το ClassMix μεταφέρει $3div2$ κλάσεις στη δεύτερη εικόνα και στη συγκεκριμένη περίπτωση επιλέγεται τυχαία η κλάση horse.

Για την περίπτωση της ασθενούς-ισχυρής συνέπειας επιλέγουμε να χρησιμοποιήσουμε τις παραπάνω τεχνικές, προκειμένου να παράξουμε τις ισχυρά επαυξημένες εκδοχές των μη επισημασμένων εικόνων. Όπως, είδαμε τόσο το CutMix, όσο και το ClassMix μπορούν να παράξουν πρωτότυπες εικόνες που φέρουν μεγάλη ποικιλομορφία και έχουν διευρυμένο σημασιολογικό περιεχόμενο, ωθώντας το δίκτυο κατάτμησης στην εκμάθηση αρκετά πιο δύσκολων παραδειγμάτων και στη δυνατότητα αναγνώρισης αντικειμένων σε διαφορετικές συνθήκες και περιβάλλοντα. Ακολουθούμε το παράδειγμα εκπαίδευσης σύμφωνα με τα [16, 15, 82] που στηρίζεται ουσιαστικά στην ιδέα του ICT [11]. Σε δύο τυχαία επιλεγμένες μη επισημασμένες εικόνες εφαρμόζουμε μία από τις παραπάνω τεχνικές, παράγοντας μία νέα εικόνα μίξης που θεωρείται ισχυρά επαυξημένη (εφαρμόζουμε επιπλέον μετασχηματισμούς χρώματος και θολώματος), η οποία δίνεται ως είσοδος στο δίκτυο student. Παράλληλα, οι δύο επιμέρους εικόνες τροφοδοτούνται στο δίκτυο teacher που παράγει τις αντίστοιχες ψευδοετικέτες, οι οποίες αναμειγνύονται εκ νέου, ώστε να προκύψει η ψευδοετικέτα που θα επιβλέψει την έξοδο του student. Με αυτόν τον τρόπο, το δίκτυο student μαθαίνει να παράγει ικανοποιητικές προβλέψεις για τις αναμειγμένες εικόνες με την βοήθεια των ψευδοετικέτων που του παρέχονται από τον teacher, οι οποίες έχουν παραχθεί από τις επιμέρους ασθενώς επαυξημένες εικόνες. Για την ανάμειξη των εικόνων επιλέγουμε να ακολουθήσουμε τη λογική του [16], όπου στην εκάστοτε δέσμη δεδομένων (batch) κάθε εικόνα αναμειγνύεται με την αμέσως επόμενη της, εξασφαλίζοντας έτσι ότι πάντα θα πραγματοποιείται μίξη διαφορετικών εικόνων.

Παρακάτω, παρουσιάζουμε πιο αναλυτικά την διαδικασία εκπαίδευσης. Ας υποθέσουμε ότι εφαρμόζουμε τη μέθοδο ClassMix και έχουμε μία δέσμη μη επισημασμένων εικόνων μεγέθους 4 από το σύνολο CelebAMask-HQ.

Όπως, έχουμε αναφέρει κάθε εικόνα γίνεται classmixed με την αμέσως επόμενη της, συνεπώς, θα έχουμε τις εξής μίξεις: $(u1, u2), (u2, u3), (u3, u4)$ και $(u4, u1)$.

Εικόνα 5.23: *Δέσμη μη επισημασμένων δεδομένων από το σύνολο CelebAMask-HQ*

Εικόνα 5.24: *Δέσμη δεδομένων μετά την εφαρμογή του ClassMix*

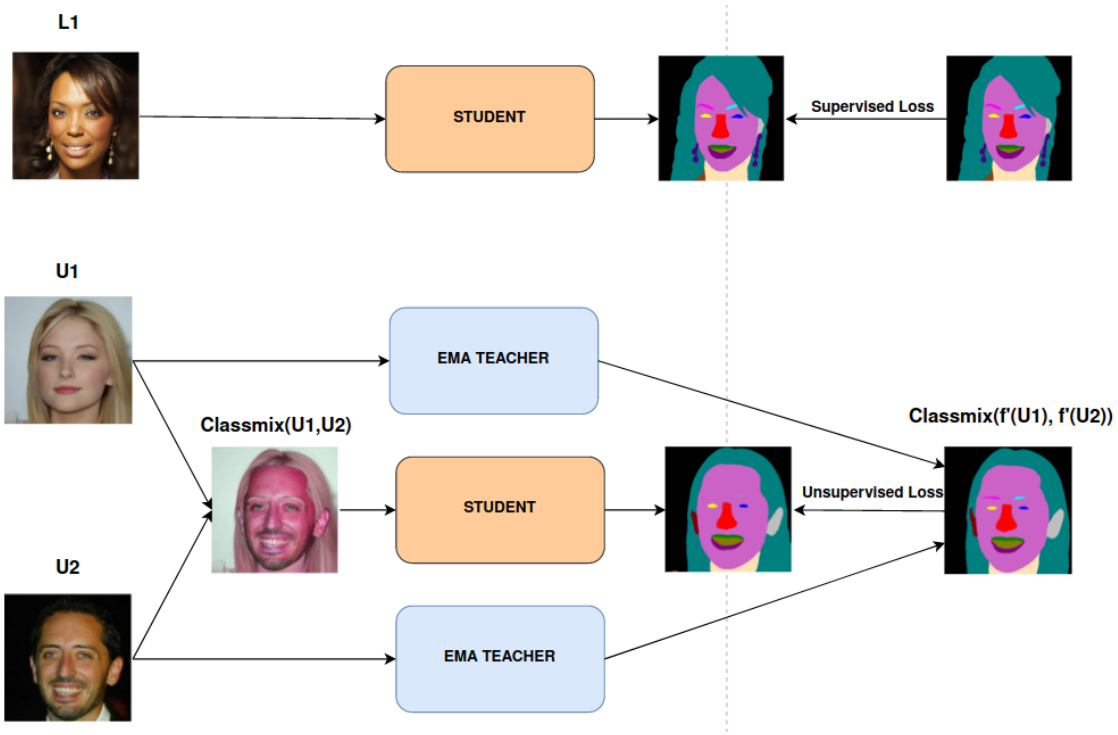
Στη συνέχεια εφαρμόζουμε και επιπλέον μετασχηματισμούς χρώματος και θολώματος, έτσι ώστε να προκύψει η τελική μορφή της ισχυρά επαυξημένης δέσμης μη επισημασμένων δεδομένων που θα δοθεί ως είσοδος στον student.

Εικόνα 5.25: *Δέσμη δεδομένων μετά τους μετασχηματισμούς Colorjitter[24] και Gaussian Blur[25]*

Κάνουμε χρήση ορισμένων συμβολισμών από το [64]. Το συνολικό κόστος μπορεί να γραφτεί ως εξής, όπου l_{ce} το κόστος cross-entropy:

$$L_{total} = \frac{1}{|D_L|} \sum_{(x,y)\in D_L} l_{CE}(y, f_{\partial}^{student}(x)) + \lambda \cdot \frac{1}{|D_U|} \sum_{(u1,u2)\in D_U} l_{CE}(ClassMix(f_{\partial'}^{teacher}(u1), f_{\partial'}^{teacher}(u2)), f_{\partial}^{student}(ClassMix(u1,u2))) \tag{5.5}$$

Όπου, όπως έχουμε αναφέρει ο μη επιβλεπόμενος όρος πολλαπλασιάζεται με τον συντελεστή λ 5.4.



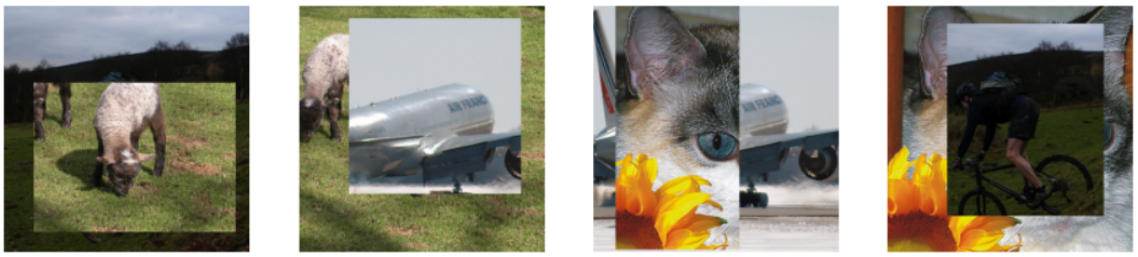
Εικόνα 5.26: *Παράδειγμα εκπαίδευσης με εφαρμογή Classmix ως ισχυρή επαύξηση*

Ισχυρή επαύξηση με συνδυασμούς των τεχνικών ClassMix και CutMix

Στη δουλειά των Lihe Yang et al. [14] προτείνεται ως επέκταση του κλασικού FixMatch, η τροφοδότηση του δικτύου με δύο ή περισσότερες ισχυρά επαυξημένες εκδοχές μίας εικόνας, προκειμένου να γίνει μεγαλύτερη αξιοποίηση των πλεονεκτήματων που προσφέρει η ισχυρή επαύξηση στην ημιεπιβλεπόμενη κατάτμηση, καθώς το μοντέλο μπορεί να γενικεύει καλύτερα, μαθαίνοντας ένα μεγαλύτερο εύρος παραλλαγών των δεδομένων. Ουσιαστικά, η ιδεά προέρχεται από την ημιεπιβλεπόμενη κατηγοριοποίηση και τις δημοσίευσεις MixMatch[94], ReMixMatch[74], όπου το δίκτυο εκτίθεται σε πολλαπλές επαυξημένες εκδοχές μίας αρχικής εικόνας. Στο [14], οι συγγραφείς, προκειμένου να κατασκευάσουν τις δύο ισχυρά επαυξημένες εκδοχές κάνουν χρήση του CutMix και για τις δύο. Εμείς, ακολουθώντας τη συγκεκριμένη μεθοδολογία των δύο επαυξημένων εκδοχών επιλέγουμε να πειραματιστούμε με διάφορους συνδυασμούς για την παραγωγή κάθε μίας εκ των δύο αυτών εκδοχών. Συγκεκριμένα, δοκιμάζουμε να παράξουμε τη μία ισχυρά επαυξημένη εκδοχή με ClassMix και την άλλη με CutMix. Διαισθητικά, εκμεταλλευόμενοι την ποικιλομορφία των εικόνων που παράγονται από τις παραπάνω μεθόδους, θέλουμε να εξετάσουμε κατά πόσο το δίκτυο μπορεί να ωφεληθεί, μαθαίνοντας παράλληλα να τμηματοποιεί τόσο τις ClassMixed εικόνες, όσο και τις CutMixed. Επιπλέον, πειραματιζόμαστε και με την περίπτωση που οι δύο ισχυρά επαυξημένες εκδοχές προκύπτουν με εφαρμογή του ClassMix (θα προκύπτουν διαφορετικές εικόνες, καθώς οι κλάσεις που θα αναμειχθούν επιλέγονται τυχαία κάθε φορά). Παρακάτω, παρουσιάζουμε τη διαδικασία εκπαίδευσης για το συνδυασμό ClassMix-CutMix, φέρνοντας ως παράδειγμα εικόνες από το σύνολο Pascal. Θεωρούμε, πάλι, ότι έχουμε μία δέσμη ασθενώς επαυξημένων μη επισημασμένων εικόνων μεγέθους 4.

Εικόνα 5.27: *Δέσμη μη επισημασμένων δεδομένων από το σύνολο Pascal*

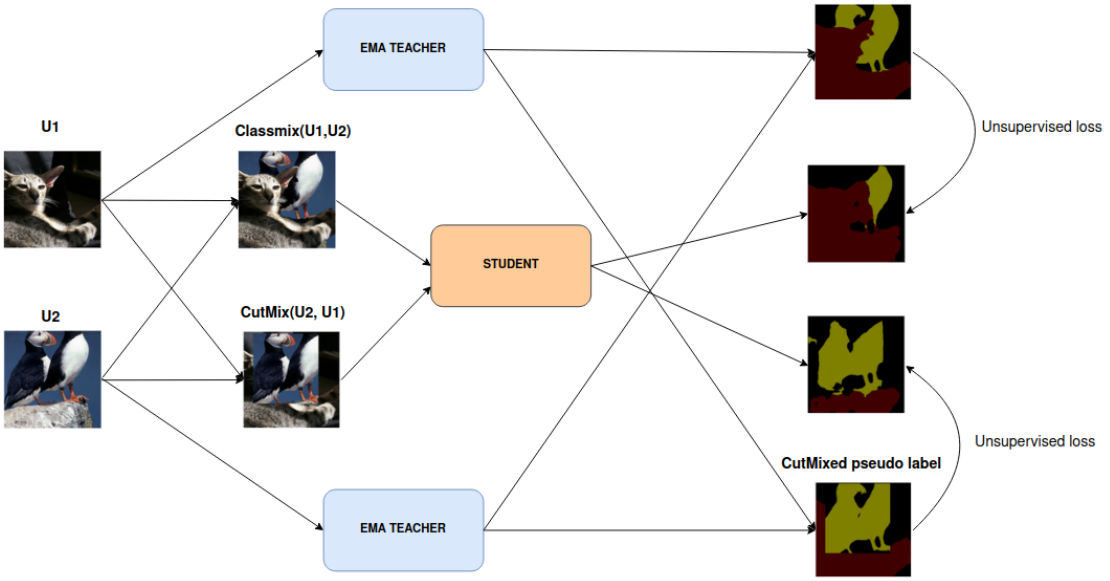
Εικόνα 5.28: *Δέσμη των παραγόμενων Classmixed δεδομένων*



Εικόνα 5.29: *Δέσμη των παραγόμενων CutMixed δεδομένων*

Το δίκτυο student, λοιπόν εκτίθεται ταυτόχρονα στις δύο παραπάνω εκδοχές των εικόνων με σκοπό να μάθει να τις τμηματοποιεί κάτω υπο την επίβλεψη των ψευδοετικετών που παράγει ο teacher, τόσο για τις ClassMixed, όσο και για τις CutMixed εικόνες. Σε αυτό το σημείο μπορούμε να κάνουμε την εξής παρατήρηση. Προκειμένου να παράξουμε εικόνες με επιπλέον ποικιλομορφία μπορούμε να επιλέξουμε οι εικόνες Cutmixed να παράγονται με τη μεταφορά κάποιου patch από τη δεύτερη στην πρώτη εικόνα, έτσι ώστε ο student να μπορεί να μάθει να τμηματοποιεί τα ίδια αντικείμενα υπο διαφορετικά περιβάλλοντα στο background. Γι΄ αυτό και για τη δημιουργία της δέσμης των ClassMixed εικόνων αναμειγνύουμε τις εικόνες (u_i, u_{i+1}), ενώ για την CutMixed δέσμη τις (u_{i+1}, u_i). Αντίστοιχα, το ίδιο μπορεί να γίνει και με τη χρήση άλλων συνδυασμών (ClassMixed-ClassMixed, CutMixed-CutMixed).

Το συνολικό κόστος για το παραπάνω πρότυπο εκπαίδευσης μπορεί να γραφτεί(κάνουμε χρήση κάποιων συμβολισμών από το [64]) σύμφωνα με το [14] ως εξής:



Εικόνα 5.30: *Κλάδος αξιοποίησης των μη επισημασμένων δεδομένων με την τεχνική των δύο ισχυρά επαυξημένων εκδοχών [14]. Στη συγκεκριμένη περίπτωση για τη δημιουργία των δύο ισχυρών διαταραχών αξιοποιείται το ClassMix και το CutMix αντίστοιχα.*

$$
\begin{aligned}
L_{total} &= \frac{1}{|D_L|} \sum_{(x,y)\in D_L} l_{CE}(y, f_{\partial}^{student}(x)) \\
&+ \lambda(\frac{1}{2} \cdot \frac{1}{|D_U|} \sum_{(u1,u2)\in D_U} l_{CE}(ClassMix(f_{\partial'}^{teacher}(u1), f_{\partial'}^{teacher}(u2)), f_{\partial}^{student}(ClassMix(u1,u2))) \\
&+ \frac{1}{2} \cdot \frac{1}{|D_U|} \sum_{(u1,u2)\in D_U} l_{CE}(CutMix(f_{\partial'}^{teacher}(u1), f_{\partial'}^{teacher}(u2)), f_{\partial}^{student}(CutMix(u1,u2)))) \qquad (5.6)
\end{aligned}
$$

, όπου l_{CE} το κόστος cross-entropy και λ 5.4 το βάρος συνεισφοράς του μη επιβλεπόμενου όρου. Επιπλέον, πολλαπλασιάζουμε τους δύο όρους του μη επιβλεπόμενου κόστους με $\frac{1}{2}$, ώστε να έχουν ίση συνεισφορά. Αντίστοιχα, μπορεί να εκφραστεί το σύνολικό κόστος για οποιαδήποτε τεχνική μίξης χρησιμοποιηθεί για τη παραγωγή των δύο ισχυρά επαυξημένων εικόνων.

Επιμέρους λεπτομέρειες υλοποίησης Όσον αφορά επιπλέον λεπτομέρειες υλοποίησης για την ημιεπιβλεπόμενη προσέγγιση που περιγράψαμε σε αυτή την παράγραφο κάνουμε χρήση του βελτιστοποιητή (optimizer) SGD και στα τρία σύνολα δεδομένων. Οι αρχικοί ρυθμοί εκμάθησης είναι 0.002 για το Pascal και 0.01 για τα CelebAMask-HQ και Qata-COV19 αντίστοιχα, ενώ κάνουμε χρήση της πολυωνυμικής πολιτικής για τη σταδιακή μείωση του ρυθμού[91] κατά την εκπαίδευση.

Επίσης, όπως έχουμε πει στην περίπτωση της ημιεπιβλεπόμενης μάθησης έχουμε άνισες ποσότητες επισημασμένων και μη επισημασμένων δεδομένων (τα μη επισημασμένα είναι αρκετά περισσότερα από τα επισημασμένα). Στη δική μας περίπτωση επιλέγουμε να ορίσουμε ως εποχή το πέρασμα όλων των επισημασμένων δεδομένων από το δίκτυο, όπως γίνεται και

στην υλοποίηση της δημοσίευσης [82]. Στην αρχή κάθε εποχής δειγματοληπτούμε τυχαία με τη βοήθεια ενός RandomSampler [95] το μεγαλύτερο σύνολο μη επισημασμένων δεδομένων, επιλέγοντας τυχαία ίσο πλήθος με αυτό των επισημασμένων δεδομένων. Συνεπώς, σε κάθε εποχή το δίκτυο βλέπει όλα τα επισημασμένα δεδομένα και ίσο αριθμό διαφορετικών μη επισημασμένων δεδομένων κάθε φορά. Με αυτό τον τρόπο μπορούμε να αξιοποιήσουμε ικανοποιητικά την πληροφορία που μας παρέχεται από το συνολικό πλήθος των μη επισημασμένων δεδομένων. Επιλέγουμε δέσμη εκπαίδευσης με συνολικό μέγεθος 20, η οποία αποτελείται από 10 επισημασμένα και 10 μη επισημασμένα δείγματα. Τέλος για το σύνολο Pascal εκπαιδεύουμε το δίκτυο για 40.000 ενημερώσεις της κλίσης (gradient updates, training iterations), ενώ για τα CelebAMask-HQ και QaTa-COV19 για 20.000 ενημερώσεις. Όλα τα πειράματα εκτελέστηκαν στον υπερυπολογιστή ARIS[96] σε κόμβους τύπου ml node, μέσω SLURM JOBS, κάνοντας χρήση μίας GPU τύπου NVIDIA Volta V100[97]. Επιπλέον λεπτομέρειες υλοποίησης μπορούν να βρεθούν στο σύνδεσμο :https://github.com/nysp78/semi-supervised-semantic-segmentation.

Κεφάλαιο 6

Παρουσίαση και σχολιασμός πειραματικών αποτελεσμάτων

6.1 Μετρικές αξιολόγησης

Η βασική μετρική που χρησιμοποιούμε για να αξιολογήσουμε την απόδοση των μοντέλων, όσον αφορά την ποιότητα των μασκών κατάτμησης που παράγουν, είναι η **Intersection-Over-Union (IOU)** ή αλλιώς Jaccard Index [98]. Η μετρική IOU ουσιαστικά εκφράζει το ποσοστό επικάλυψης μεταξύ της μάσκας κατάτμησης (ground truth mask) που χρησιμοποιείται ως ετικέτα (label) και της πρόβλεψης που παράγει το δίκτυο κατάτμησης (prediction mask). Επομένως, μέσω του IOU μπορούμε σε ικανοποιητικό βαθμό να αντιληφθούμε πόσο καλά μπορεί να τμηματοποιήσει το δίκτυο τα αντικείμενα της εικόνας.

$$IOU = \frac{GroundTruth \cap Prediction}{GroundTruth \cup Prediction} \tag{6.1}$$

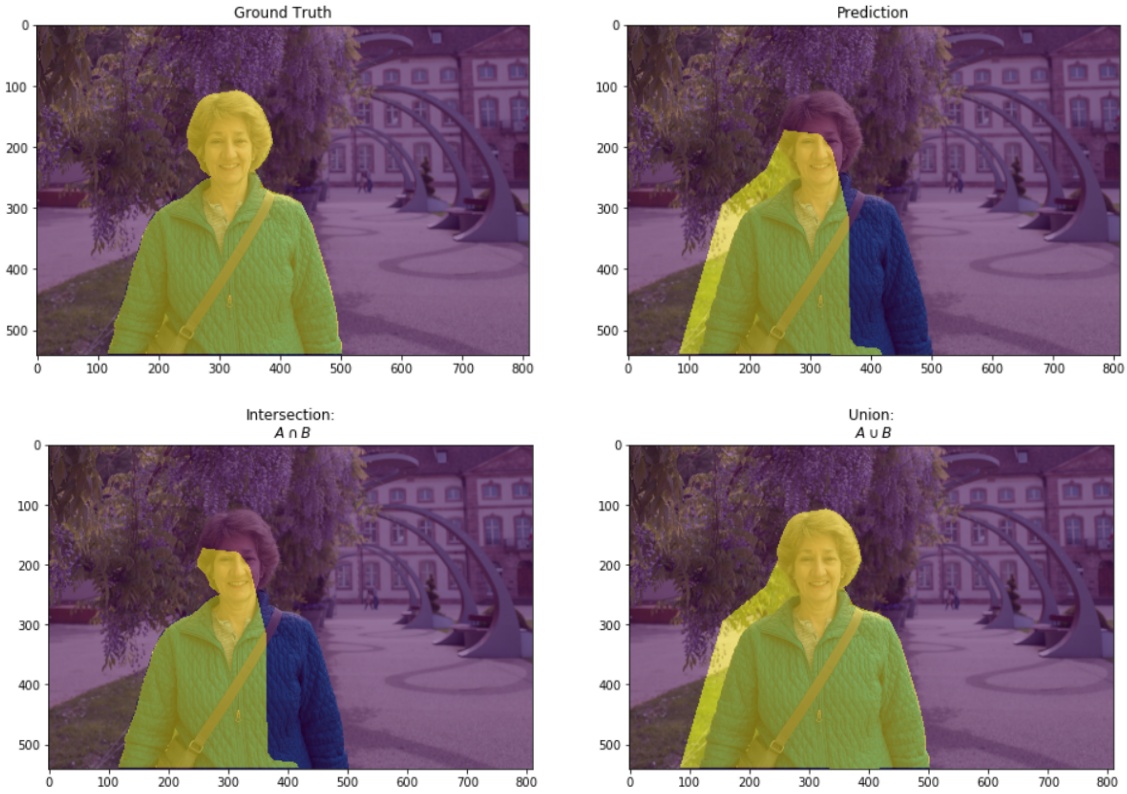
Το IOU υπολογίζεται για κάθε σημασιολογική κλάση ξεχωριστά, ως ο λόγος του πλήθους των κοινών και στις δύο μάσκες pixels, προς το πλήθος των pixels που βρίσκονται είτε στη μια είτε στην άλλη. Στη συνέχεια λαμβάνεται ο αριθμητικός μέσος όρος (meanIOU) μεταξύ των επιμέρους IOU της κάθε κλάσης.

Κάνοντας χρήση των παρακάτω εννοιών:

- **True Positives (TP)**: Πλήθος pixels που ταξινομήθηκαν στη σωστή κλασή, έστω C_i.
- **False Positives (FP)**: Πλήθος pixels που ταξινομήθηκαν λανθασμένα στη κλάση C_i, αλλά άνηκαν σε διαφορετική.
- **False Negatives (FN)**: Πλήθος pixels που άνηκαν στη κλάση C_i, αλλά ταξινομήθηκαν λανθασμένα σε διαφορετική κλάση.

Το score IOU για κάποια κλάση C_i μπορεί να εκφραστεί ως εξής:

$$IOU_{C_I} = \frac{TP_{C_i}}{TP_{C_i} + FP_{C_i} + FN_{C_i}} \tag{6.2}$$



Εικόνα 6.1: *Αναπαράσταση της μετρικής IOU, εικόνα από [26]. Η μετρική για τη συγκεκριμένη κλάση υπολογίζεται ως ο λόγος του πλήθους των pixels της μάσκας $A \cap B$, προς το αντίστοιχο πλήθος της μάσκας $A \cup B$.*

Μπορούν επίσης να χρησημοποιηθούν και άλλες μετρικές για την αξιολόγηση της απόδοσης ενός μοντέλου κατάτμησης, όπως το Dice coefficient [99] που εκφράζει πάλι ποσοστό επικάλυψης ως εξής:

$$DICE_{C_I} = \frac{2 \cdot TP_{C_i}}{2 \cdot TP_{C_i} + FP_{C_i} + FN_{C_i}} \tag{6.3}$$

Σε αυτή την περίπτωση η τομή TP υπολογίζεται δύο φορές, για αυτό και δεν αφαιρείται από τον παρονομαστή. Επιπλέον, χρησιμοποιείται και το $F1 - score$ [100], το οποίο εκφράζεται ως ο αρμονικός μέσος μεταξύ ακρίβειας (presicion) και ανάκλησης (recall). Η ακρίβεια πρόκειται για τον λόγο μεταξύ των ορθώς ταξινομημένων pixels σε μια κλάση προς τον ολικό αριθμό από pixels που ταξινομήθηκαν σε αυτή. Η ανάκληση εκφράζει τον λόγο μεταξύ των ορθώς ταξινομημένων pixel σε μια κλάση προς τον ολικό αριθμό pixel που ανήκουν σε αυτή την κλάση.

$$Precision_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + FP_{C_i}}$$

$$Recall_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + FN_{C_i}}$$

$$F1 - score_{C_i} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6.4}$$

6.2 Πειραματικά αποτελέσματα για το σύνολο Pascal VOC 2012

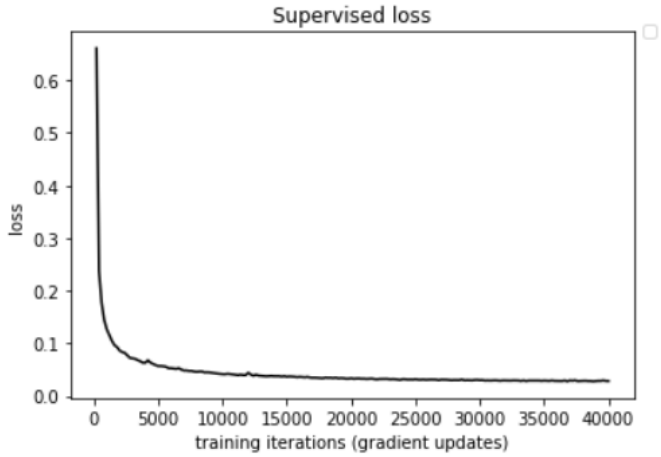
Παραθέτουμε συγκεντρωτικά τα πειραματικά αποτελέσματα για το σύνολο δεδομένων Pascal, αναφέροντας τη μετρική mIOU πάνω στο επίσημο σύνολο επικύρωσης (validation set) για τις τέσσερις διαμερίσεις. Για κάθε διαμέριση έχουμε εκπαιδεύσει ένα επιβλεπόμενο μοντέλο αποκλειστικά στα διαθέσιμα επισημασμένα δεδομένα και επιπλέον ημιεπιβλεπόμενα μοντέλα με διαταραχές επιπέδου εισόδου (color augmentation, ClassMix, CutMix) που αξιοποιούν τα μη επισημασμένα δεδομένα. Τα παρακάτω αποτελέσματα για τις ημιπεπιβλεπόμενες μεθόδους έχουν παραχθεί με την αξιολόγηση του καλύτερου μοντέλου teacher που αποθηκεύτηκε (checkpoint) κατα τη διάρκεια της εκπαίδευσης πάνω στο σύνολο επικύρωσης του Pascal.

Πίνακας 6.1: *Πειραματικά αποτελέσματα (mIOU) για το σύνολο επικύρωσης του PASCAL VOC 2012. Σε παρένθεση φαίνεται το πλήθος των επισημασμένων δεδομένων σε κάθε διαμέριση.*

Method	1/32(330)	1/16(662)	1/8(1323)	1/4(2645)	all labels(10582)
Supervised	60.03	66.20	69.58	72.92	77.85
Color augmentation	67.15	71.31	74.63	76.01	-
CutMix	70.92	73.21	74.67	76.11	-
ClassMix	71.94	73.82	74.65	76.49	-
ClassMix + CutMix	**72.56**	**74.62**	**75.84**	**77.30**	-
ClassMix + ClassMix	72.54	74.02	75.62	76.29	-
CutMix + CutMix	70.78	73.16	75.33	77.15	-

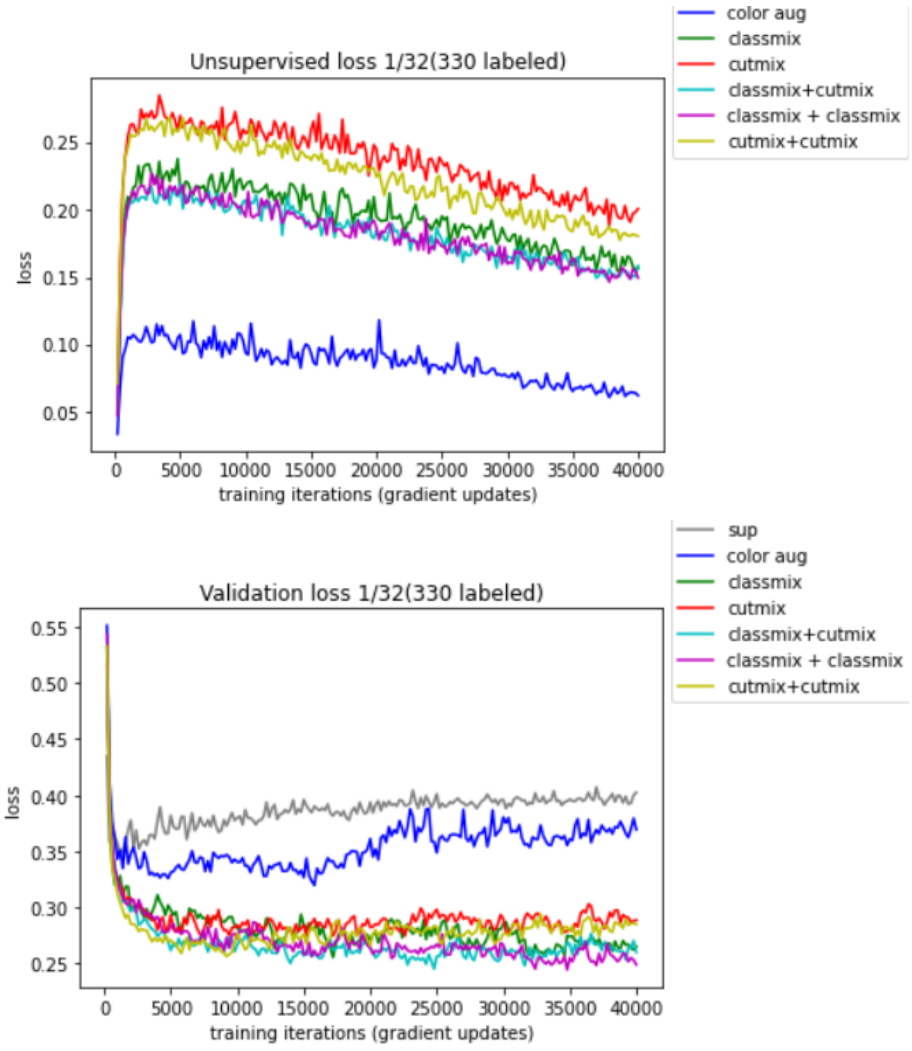
Παρατηρήσεις και σχολιασμός Με μία πρώτη ματιά μπορούμε να δούμε ότι η χρήση κανονικοποίησης συνέπειας με αξιοποίηση των επιπλέον μη επισημασμένων δεδομένων επιφέρει σημαντικές βελτιώσεις σε σχέση με την κλασική επιβλεπόμενη εκπαίδευση (supervised baseline). Οι διαταραχές που εφαρμόζονται είτε μέσω μετασχηματισμών χρώματος (color augmentation), είτε με χρήση των τεχνικών μίξης CutMix, ClassMix πάνω στα μη επισημασμένα δεδομένα, βοηθούν στη μείωση της υπερπροσαρμογής του μοντέλου στα λίγα διαθέσιμα επισημασμένα δεδομένα και ενισχύουν την ικανότητα γενίκευσης σε δεδομένα που καλείται να τμηματοποιήσει για πρώτη φορά. Επιπροσθέτως, η χρήση δύο ισχυρά επαυξημένων εκδοχών, φαίνεται ότι βοηθά στην περαιτέρω βελτίωση της απόδοσης.

Προκειμένου να κατανοηθούν καλύτερα τα παραπάνω αποτελέσματα παρουσιάζουμε τις καμπύλες μάθησης για το επιβλεπόμενο και μη επιβλεπόμενο κόστος, καθώς και την αντίστοιχη απώλεια στο σύνολο επικύρωσης (validation) για την κάθε μέθοδο. Αρχικά παραθέτουμε την καμπύλη του επιβλεπόμενου κόστους πάνω στα επισημασμένα δεδομένα που χρησιμοποιεί το δίκτυο και για κάθε διαμέριση έχει την παρακάτω μορφή. Βλέπουμε, ότι το μοντέλο προσαρμόζεται πλήρως στα λίγα επισημασμένα δεδομένα.

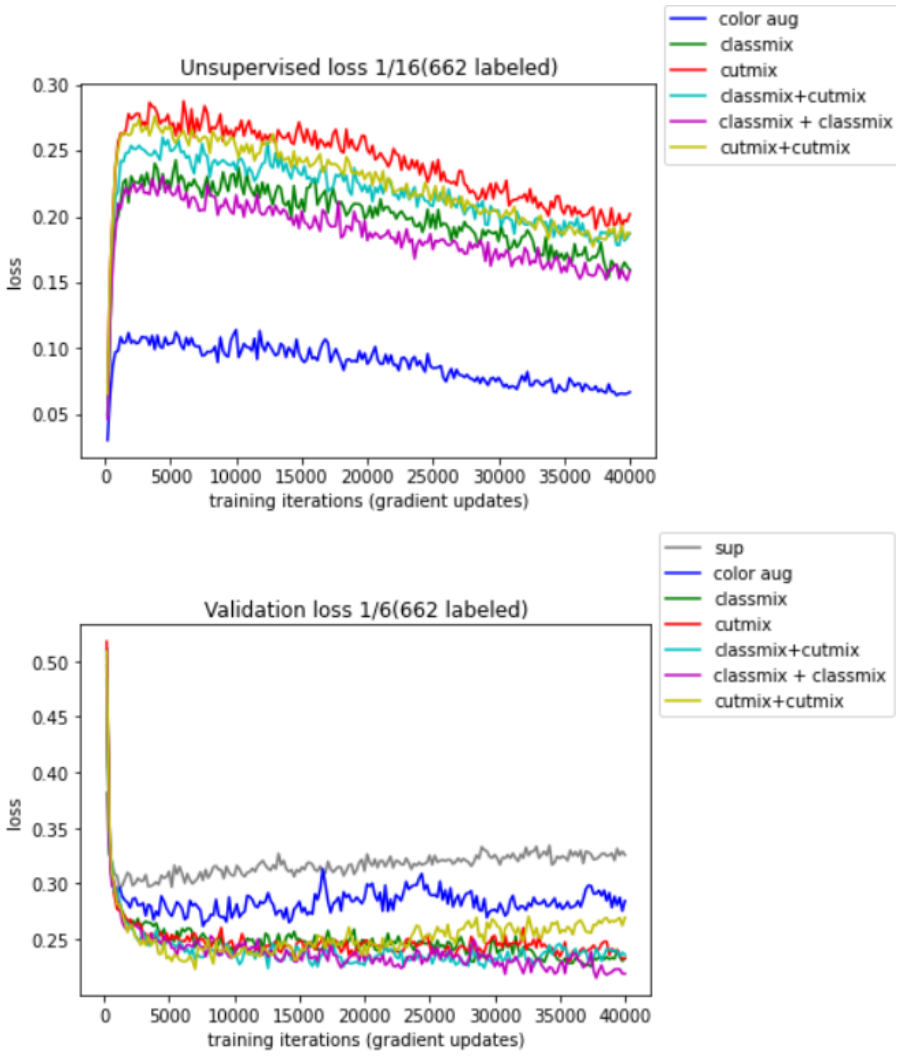


Εικόνα 6.2: *Καμπύλη μάθησης για τα επισημασμένα δεδομένα (supervised loss)*

Στη συνέχεια παραθέτουμε τις καμπύλες του μη επιβλεπόμενου κόστους και της απώλειας επικύρωσης (validation loss), αρχικά για τις διαμερίσεις 1/32, 1/16.



Εικόνα 6.3: *Καμπύλες μη επιβλεπόμενης απώλειας και απώλειας επικύρωσης για τη διαμέριση 1/32*



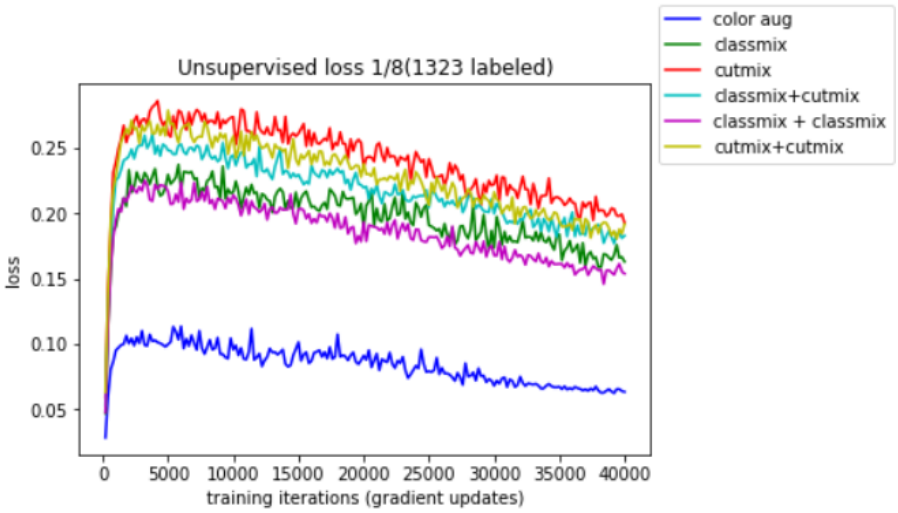
Εικόνα 6.4: *Καμπύλες μη επιβλεπόμενης απώλειας και απώλειας επικύρωσης για τη διαμέριση 1/16*

Με βάση τις παραπάνω καμπύλες μπορούμε να πούμε ότι όλες οι μέθοδοι κανονικοποίησης συνέπειας που εφαρμόζονται βοηθούν πολύ στην ελάττωση της υπερπροσαρμογής (overfitting). Πιο συγκεκριμένα, βλέπουμε και στις δύο διαμερίσεις, όπως είναι φυσικό, το supervised only μοντέλο (γκρί γραμμή) να υπερπροσαρμόζεται πάρα πολύ στα λίγα επισημασμένα δεδομένα με αποτέλεσμα να μην μπορεί να μειώσει την απώλεια στα δεδομένα επικύρωσης. Η χρήση μετασχηματισμών χρώματος (μπλέ γραμμή) ως ισχυρή διαταραχή έχει τη δυνανότητα σε κάποιο βαθμό να συνεισφέρει στη μείωση αυτής της υπερπροσαρμογής και στην αύξηση του mIOU (πίνακας 6.4). Περαιτέρω, αύξηση της ικανότητας γενίκευσης και του mUOI προσφέρουν οι μέθοδοι μίξης (CutMix, ClassMix), αλλά και η εκπαίδευση σε δύο ισχυρά επαυξημένες εκδοχές που προκύπτουν με συνδυασμούς αυτών (ClassMix-CutMix, ClassMix-ClassMix, CutMix-CutMix), καθώς παράγονται εικόνες με μεγαλύτερο σημασιολογικό περιεχόμενο και με περισσότερη ποικολομορφία. Όπως, φαίνεται η εκπαίδευση του student πάνω σε τέτοιες πολύπλοκες εικόνες βοηθά κατά πολύ στη βελτίωση της γενίκευσης.

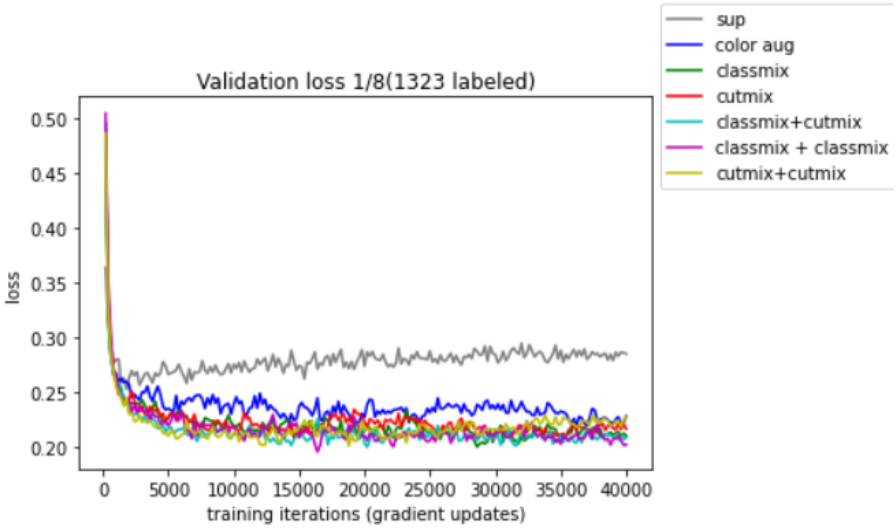
Όσον αφορά τη μορφή του μη επιβλεπόμενου κόστους, παρατήρουμε ότι στις αρχικές επαναλήψεις εκπαίδευσης (training iterations), η συνεισφορά του είναι μικρή και αυτό ο-

φείλεται στο συντελεστή λ (Εξ. 5.4 - ποσοστό pixels ψευδοετικέτας που έχουν βεβαιότητα πάνω από ένα κατώφλι). Στην αρχή της εκπαίδευσης ο teacher δεν παράγει ακόμα καλής ποιότητας ψευδοετικέτες και έχει μεγάλη αβεβαιότητα για τις κλάσεις των pixels, επομένως σε αυτό το διάστημα η βελτιστοποίηση οδηγείται κυρίως από τον επιβλεπόμενο όρο (supevised loss). Σταδιακά, ο teacher μαθαίνει να παράγει καλύτερης ποιότητας ψευδοετικέτες με μεγαλύτερη βεβαιότητα και το μη επιβλεπόμενο κόστος λαμβάνεται περισσότερο υπόψην κατά την εκπαίδευση, με τον student να προσπαθεί να προσαρμοστεί στις ψευδοετικέτες που του παρέχονται από τον teacher και να μειώσει την απόσταση από αυτές (μείωση απώλειας). Οι μικρότερες τιμές απώλειας παρουσιάζονται για την παραγωγή της ισχυρης εκδοχής με μετασχηματισμούς χρώματος, ενώ οι απώλειες για τις μεθόδους CutMix, ClassMix και τους συνδυασμούς τους, βλέπουμε ότι κυμαίνονται σε μεγαλύτερες τιμές. Οι μικρές τιμές απώλειας για το color augmentation μπορούμε να συμπεράνουμε ότι οφείλονται στη μικρή απόσταση που έχουν οι προβλέψεις του student από τις ψευδοετικέτες του teacher. Ο μετασχηματισμός χρώματος δεν εισάγει τόσο μεγάλη ποικολομορφία, συνεπώς είναι αρκετά εύκολο για τον student να μάθει να παράγει καλές προβλέψεις για τον συγκεκριμένο ισχυρό μετασχηματισμό. Επομένως, λαμβάνοντας υπόψην και την υψηλότερη απώλεια επικύρωσης, μπορούμε να πούμε ότι για τη συγκεκριμένη επαύξηση ο student υπερπροσαρμόζεται στις ψευδοετικέτες του teacher και παρόλο το μικρό μη επιβλεπόμενο κόστος δεν μπορεί να γενικεύσει καλύτερα. Απο την άλλη οι απώλειες για τις μεθόδους CutMix, ClassMix, ClassMix+CutMix κ.λ.π, βλέπουμε ότι λαμβάνουν μεγαλύτερες τιμές, καθώς είναι πιο δύσκολο για τον student να μάθει τμηματοποιήσει τις αναμειγμένες εικόνες και παρόλο το μεγαλύτερο κόστος παρατηρούμε ότι η ικανότητα γενίκευσης του δικτύου βελτιώνεται σε μεγάλο βαθμό σε σύγκριση με τον απλο μετασχηματισμό χρώματος. Οι πιο ισχυροί μετασχηματισμοί μίξης εισάγουν επιπλέον κανονικοποίηση και επιτρέπουν στο δίκτυο να μάθεί καλύτερες αναπαραστάσεις και να γενικεύσει καλύτερα.

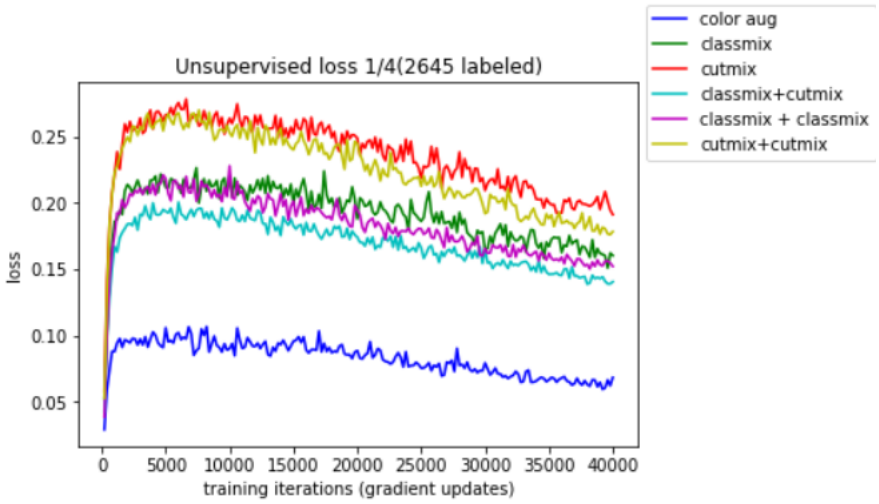
Στη συνέχεια παραθέτουμε τις αντίστοιχες καμπύλες για τις διαμερίσεις 1/8 και 1/4.



Εικόνα 6.5: *Καμπύλες μη επιβλεπόμενης απώλειας για τη διαμέριση 1/8*

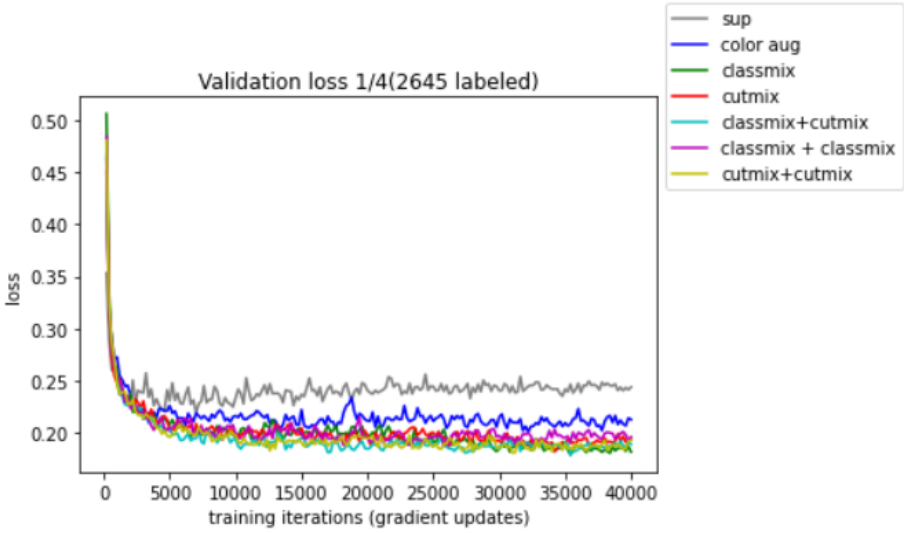


Εικόνα 6.6: *Καμπύλες απώλειας επικύρωσης για τη διαμέριση 1/8*



Εικόνα 6.7: *Καμπύλες μη επιβλεπόμενης απώλειας για τη διαμέριση 1/4*

Για τις διαμερίσεις με περισσότερα διαθέσιμα επισημασμένα δεδομένα για την εκπαίδευση του δικτύου το επιβλεπόμενο μοντέλο (supervised only) γενικεύει καλύτερα σε σχέση με τις διαμερίσεις λιγότερων επισημασμένων δεδομένων. Αυτό είναι κάτι αναμενόμενο, καθώς όπως βλέπουμε και στο πίνακα 6.4, καθώς αυξάνονται τα επισημασμένα δεδομένα, το δίκτυο μπορεί να μαθαίνει την επιπλέον πληροφορία που προσφέρουν και να αυξάνει την ικανότητα γενίκευσης. Για αυτό και παρατηρείται, ότι οι υπόλοιπες μέθοδοι που αξιοποιούν και τα μη επισημασμένα δεδομένα επιφέρουν λιγότερη βελτίωση στη γενίκευση σε σχέση με τις διαμερίσεις λιγότερων δεδομένων. Παρόλα αυτά, βλέπουμε ξανά ότι οι μέθοδοι κανονικοποίησης συνέπειας ακόμη και σε περιπτώσεις μεγαλύτερου πλήθους επισημασμένων δεδομένων μπορούν να βοηθήσουν. Η τάση που παρατηρείται από τις γραφικές της συνάρτησης απώλειας επικύρωσης, αλλά και απο το mIOU του πίνακα 6.4 είναι ότι ο μετασχηματισμός χρώματος επιφέρει παρόμοια γενίκευση και απόδοση mIOU σε σύγκριση με τις τεχνικές μίξης CutMix, ClassMix, οι οποίες όμως παραμένουν ακόμη και εδώ λίγο καλύτερες. Σε αυτές τις διαμε-

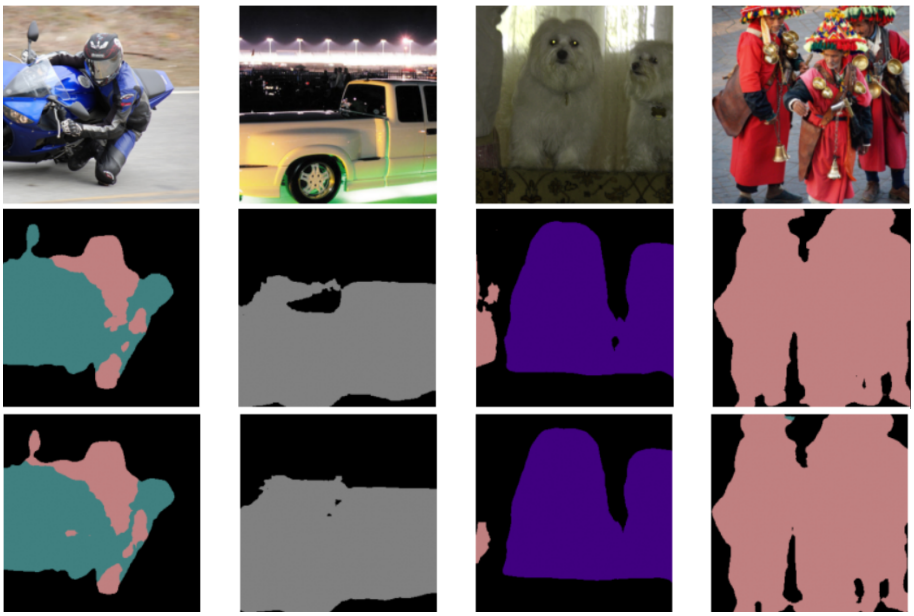


Εικόνα 6.8: *Καμπύλες επικύρωσης για τη διαμέριση 1/4*

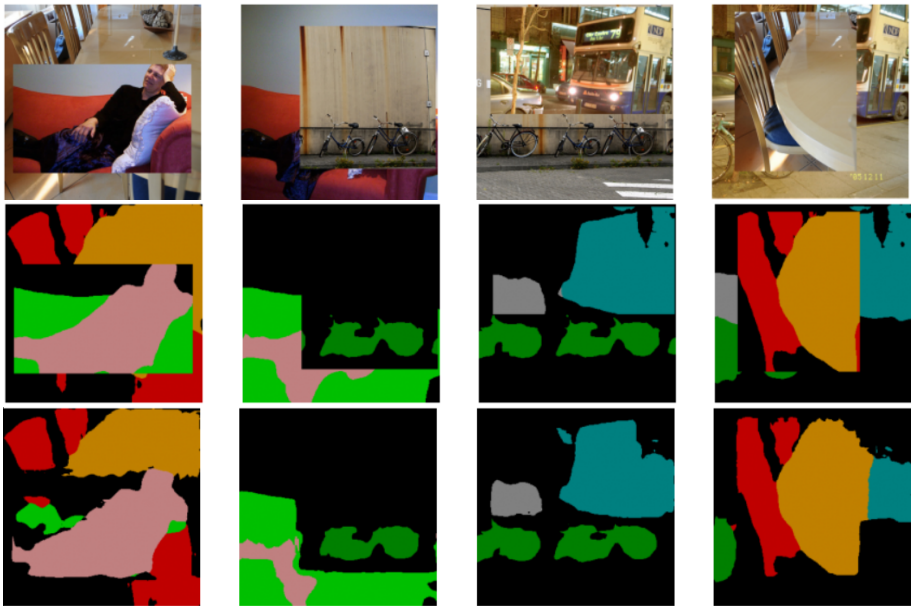
ρίσεις όπου έχουμε περισσότερα επισημασμένα δεδομένα διαθέσιμα, το δίκτυο είναι σε θέση μόνο από την εκπαίδευση πάνω σε αυτά να παρουσιάζει καλύτερη ικανότητα γενίκευσης, επομένως για να αυξήσουμε περαιτέρω αυτή την ικανότητα ενδεχομένως να χρειάζεται να εφαρμόσουμε ακόμη πιο ισχυρούς μετασχηματισμούς που θα εισάγουν περισσότερη ποικιλομορφία. Αυτό φαίνεται να επιτυγχάνεται σε ένα βαθμό με την εκπαίδευση του student πάνω σε δύο ισχυρα επαυξημένες εκδοχές. Βλέπουμε ότι οι μέθοδοι ClassMix-CutMix, ClassMix-ClassMix, CutMix-CutMix προσφέρουν ελαφρώς μεγαλύτερη ικανότητα γενίκευσης σε σχέση με τις CutMix, ClassMix για τις διαμερίσεις με περισσότερα επισήμασμένα δεδομένα. Εξαίρεση αποτέλει στη διαμέριση 1/4 η μέθοδος ClassMix-Classmix, η οποία δίνει λίγο χειρότερη απόδοση(76.29) σε σύγκριση με την ClassMix(76.49). Αυτο μπορεί να υποδηλώνει μια παραπάνω υπερπροσαρμογή στην τεχνική ClassMix, όταν χρησιμοποιείται δύο φορές στις ίδιες ασθενώς επαυξημένες εικόνες. Εδώ να αναφέρουμε ότι σε κάθε εικόνα στο σύνολο Pascal περιέχονται λίγες κλάσεις, το background και άλλες 2-3 το πολύ κλάσεις. Συνεπώς, η εφαρμογή του ClassMix στο ίδιο ζεύγος εικόνων δύο φορές είναι πιθανόν να συντελεί στη μεταφορά ίδιων κλάσεων κάθε φορά με αποτέλεσμα την μη επίτευξη περαιτέρω ποικιλομορφίας.

Γενικά από το πίνακα των συγκεντρωντικών αποτελεσμάτων 6.4 παρατηρούμε ότι για το συγκεκριμένο σύνολο η μέθοδος των δύο επαυξημένων εκδοχών ClassMix-Cutmix επιφέρει τη μεγαλύτερη ικανότητα γενίκευσης για όλες τις διαμερίσεις. Ο συνδυασμός δύο διαφορετικών μεθόδων μίξης, ενδεχομένως να προσφέρει εικόνες με μεγαλύτερη ποικολομορφία για την εκπαίδευση του student με αποτέλεσμα την αποφυγή ενδεχόμενης υπερπροσαρμογής στον ίδιο τύπο επαύξησης και στην ελαφρώς καλύτερη απόδοση mIOU στο σύνολο Pascal.

Παρακάτω παραθέτουμε ορισμένα παραδείγματα από τις ψευδοετικέτες που παράγει ο teacher και τις αντίστοιχες προβλέψεις που κάνει ο student για κάθε είδος ισχυρής επαύξησης που χρησιμοποιήσαμε, για ένα μοντέλο εκπαιδεύμενο με τη μέθοδο ClassMixed-Cutmixed με 1323 επισημασμένα δεδομένα (1/8).

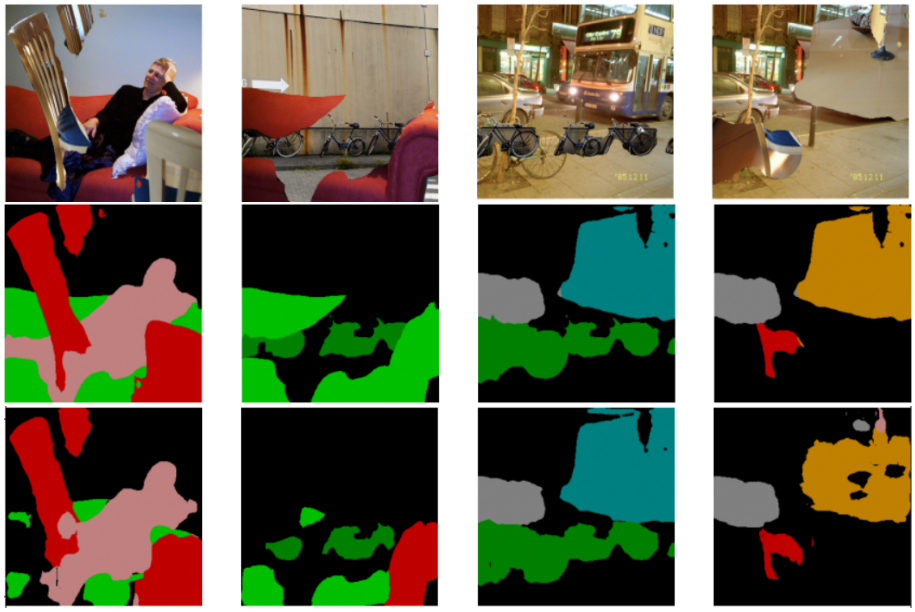


Εικόνα 6.9: *Οι ψευδοετικέτες που παράγει ο teacher και ακριβώς από κάτω οι αντίστοιχες προβλέψεις του student για εικόνες που έχει εφαρμοστεί μόνο μετασχηματισμός χρώματος*



Εικόνα 6.10: *Οι ψευδοετικέτες που παράγει ο teacher και ακριβώς από κάτω οι αντίστοιχες προβλέψεις του student για cutmixed εικόνες*

Στη διαμέριση 1/8 παρατηρείται η μεγαλύτερη βελτίωση στην απόδοση mIOU από τη μέθοδο των δύο ισχυρά επαυξημένων εκδοχών ClassMixed-CutMixed(75.84) σε σύγκριση με τα μεμονωμένα ClassMix(74.65) και Cutmix(74.67). Παίρνοντας ως παράδειγμα τη συγκεκριμένη διαμέριση παραθέτουμε κάποια στοιχεία σχετικά με τη βεβαιότητα (confidence) των ψευδοετικετών που παράγει ο teacher και των προβλέψεων που κάνει ο student για τις cutmixed, classmixed και color augmented εικόνες.



Εικόνα 6.11: *Οι ψευδοετικέτες που παράγει ο teacher και ακριβώς από κάτω οι αντίστοιχες προβλέψεις του student για classmixed εικόνες*

Πίνακας 6.2: *Μέση βεβαιότητα (mean confidence) των παραγόμενων ψευδοετικετών του teacher και των αντίστοιχων προβλέψεων του student για την περίπτωση της διαμέρισης 1/8*

Method	pseudolabels confidence	predictions confidence
Color augmentation	89.08	85.41
CutMix	85.94	75.84
ClassMix	84.65	71.83
ClassMix + CutMix	88.25, 88.03	80.06, 80.32

Από τον παραπάνω πίνακα βλέπουμε ότι κατά την εκπαίδευση του μοντέλου με μετασχηματισμούς χρώματος ο teacher παράγει ψευδοετικέτες με μέγαλη μέση βεβαιότητα(89.08), ενώ υψηλής βεβαιότητας(85.41) είναι και οι αντίστοιχες προβλέψεις του student για τις color augmented εικόνες. Από την άλλη παρατηρούμε ότι για τις μεθόδους Cutmix και ClassMix τόσο οι ψευδοετικέτες(85.94, 84,65) όσο και οι προβλέψεις(75.84, 71.83) έχουν χαμηλότερη μέση βεβαιότητα, κάτι αναμενόμενο, αφού οι συγκεκριμένες μέθοδοι παράγουν πιο δύσκολες εικόνες επιφέροντας μεγαλύτερη αβεβαιότητα στο student, ο οποίος καλείται να μάθει να τις τμηματοποιεί. Παράλληλα, ο student εκπαιδευμένος με τη μέθοδο ClassMix-CutMix βλέπουμε ότι παράγει προβλέψεις τόσο για τις ClassMixed εικόνες, όσο και για τις CutMixed με μεγαλύτερη βεβαιότητα(80.06, 80.32). Αυτό ενδεχομένως να συμβαίνει, καθώς η ταυτόχρονη έκθεση του student τόσο σε classmixed, όσο και σε cutmixed εικόνες, ενισχύει την ικανότητα του να παράγει προβλέψεις με υψηλότερη βεβαιότητα και για τις δύο εκδοχές εικόνων. Πρέπει να υπογραμμίσουμε εδώ, ότι η υψηλή βεβαιότητα στις προβλέψεις δεν σημαίνει απαραίτητα ότι το μοντέλο θα γενικεύει καλά και σε δεδομένα που βλέπει για πρώτη φορά. Για παράδειγμα, στην περίπτωση του color augmentation βλέπουμε ότι ο student παράγει προβλέψεις με τη μεγαλύτερη βεβαιότητα, παρόλα αυτά στο σύνολο επικύρωσης αδυνατεί να έχει τόσο καλή γενίκευση, όσο οι μέθοδοι ClassMix, Cutmix, στις οποίες οι προβλέψεις του student έχουν μικρότερη μέση βεβαιότητα. Επιπλέον, με τη μέθο-

δο ClassMix-CutMix ο student μπορεί να τμηματοποιεί με μεγαλύτερη βεβαιότητα και τις δύο εκδοχές εικόνων και παράλληλα να γενικεύει καλύτερα από τις άλλες μεθόδους. Γενικά, η υπερβολική βεβαιότητα(over-confidence) δεν συνεπάγεται απαραίτητα καλύτερη ικανότητα γενίκευσης. Πολλές φορές μάλιστα υποδηλώνει την υπερπροσαρμογή του δικτύου στα δεδομένα εκπαίδευσης. Αντίθετα, η μείωση της βεβαιότητας σε κάποιο βαθμό, μέσω πιο περίπλοκων τεχνικών επαύξησης μπορεί να επιτρέψει στο μοντέλο να γενικεύει καλύτερα, όπως άλλωστε στις περιπτώσεις των μεθόδων ClassMix, CutMix, ClassMix+CutMix.

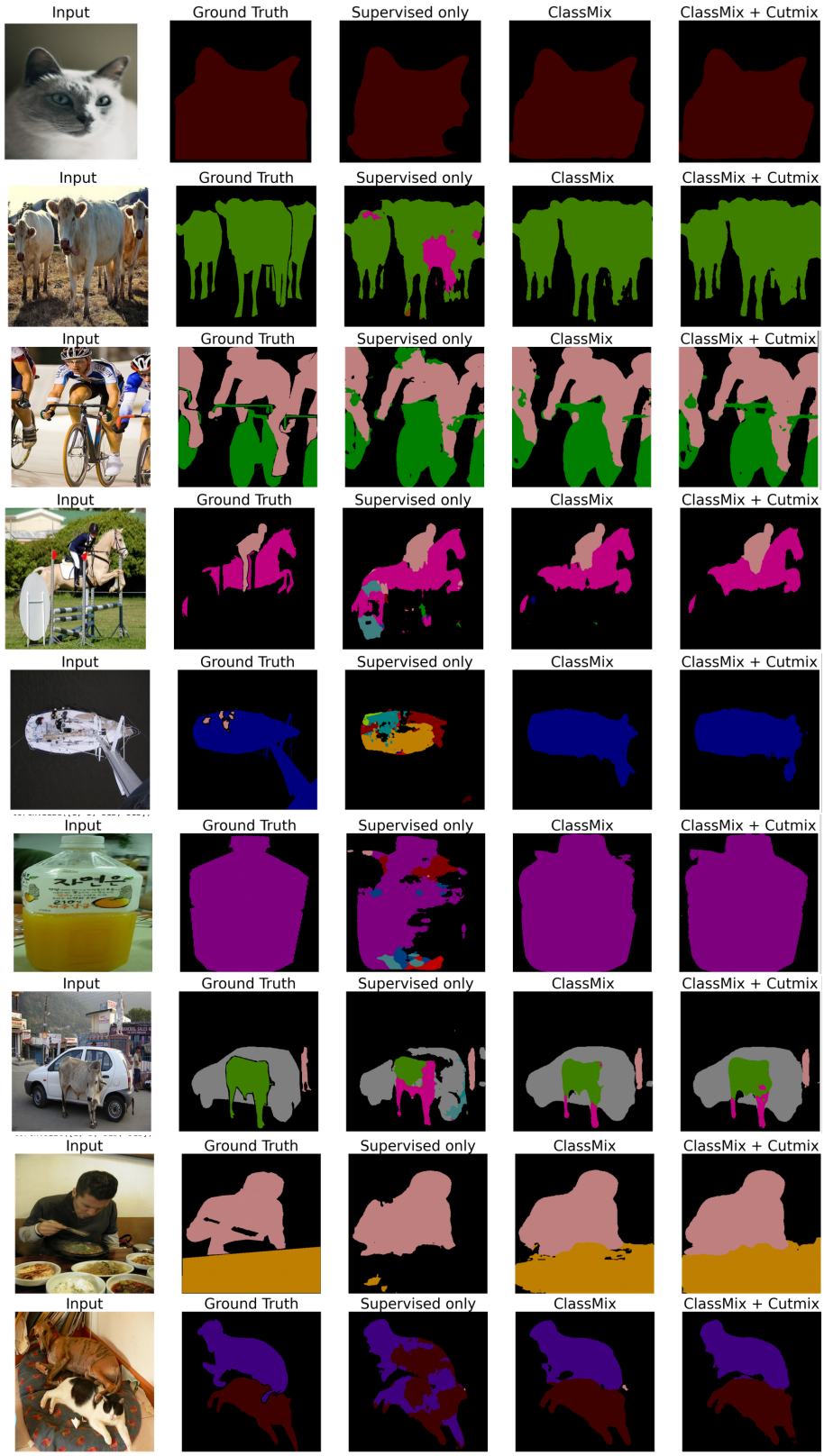
Έν συνεχεία για τις μεθόδους ClassMix, CutMix και ClassMix+CutMix για τη διαμέριση 1/8 παραθέτουμε τη μετρική IOU για τις επιμέρους 21 κλάσεις του προβήματος.

Πίνακας 6.3: *Η μετρική mIOU για κάθε σημασιολογική κλάση του Pascal για τις μεθόδους Supevised, CutMix, ClassMix, ClassMix + Cutmix στην περίπτωση της διαμέρισης 1/8*

	Sup only	CutMix	ClassMix	ClassMix+Cutmix
background:0	0.9271	0.9401	0.9392	**0.9421**
airplane:1	0.8823	0.8859	0.8712	**0.8941**
bicycle:2	0.4115	0.4079	0.4152	**0.4210**
bird:3	0.8623	0.8815	0.8659	**0.8916**
boat:4	0.6313	0.6674	0.6673	**0.7086**
bottle:5	0.7795	0.7464	0.7750	**0.7912**
bus:6	0.8695	**0.9393**	0.9269	**0.9393**
car:7	0.8243	0.8620	**0.8637**	0.8624
cat:8	0.8914	0.8991	0.8723	**0.9055**
chair:9	0.2915	**0.3565**	0.3310	0.3526
cow:10	0.7400	0.8276	0.8212	**0.8480**
dining table:11	0.3277	0.5403	**0.5451**	0.5407
dog:12	0.8297	0.8512	0.8282	**0.8667**
horse:13	0.7425	0.8465	0.8201	**0.8336**
motorbike:14	0.8160	0.8265	**0.8392**	0.8191
person:15	0.8286	0.8555	0.8560	**0.8587**
potted plant:16	0.3404	0.4511	**0.5822**	0.5486
sheep:17	0.7695	**0.8678**	0.8130	0.8510
sofa:18	0.4188	**0.5135**	0.4640	0.4843
train:19	0.7302	0.8340	**0.8588**	0.8330
tv/monitor:20	0.6991	0.6942	0.7104	**0.7359**

Παρατηρούμε ότι η εκπαίδευση του δικτύου με δύο ισχυρά επαυξημένες εκδοχές επιφέρει βελτίωση στις περισσότερες κλάσεις. Έν συνεχεία παραθέτουμε κάποια ποιοτικά αποτελέσματα που παράγει το δίκτυο για τυχαίες εικόνες από το σύνολο επικύρωσης.

Εικόνα 6.12: *Ποιοτικά αποτελέσματα που παράγει το δίκτυο για εικόνες από το σύνολο επικύρωσης του Pascal VOC 2012*



6.3 Πειραματικά αποτελέσματα για το σύνολο CelebAMask-HQ

Για το συγκεκριμένο σύνολο παρουσιάζουμε τη μετρική mIOU για τις 18 κλάσεις πάνω στο επίσημο σύνολο επικύρωσης του CelebAMask-HQ, όπως περιγράψαμε στην παράγραφο 4.2.

Πίνακας 6.4: *Πειραματικά αποτελέσματα (mIOU) για το σύνολο επικύρωσης του CelebAMask-HQ*

Method	1/64(187)	1/32(375)	1/16(750)	all labels(12000)
Supervised	72.20	74.29	75.91	80.10
Color perturbation	66.62	72.36	74.63	-
CutMix	75.68	76.90	77.83	-
ClassMix	**76.56**	**77.62**	**78.46**	-
ClassMix + CutMix	76.42	77.47	78.25	-

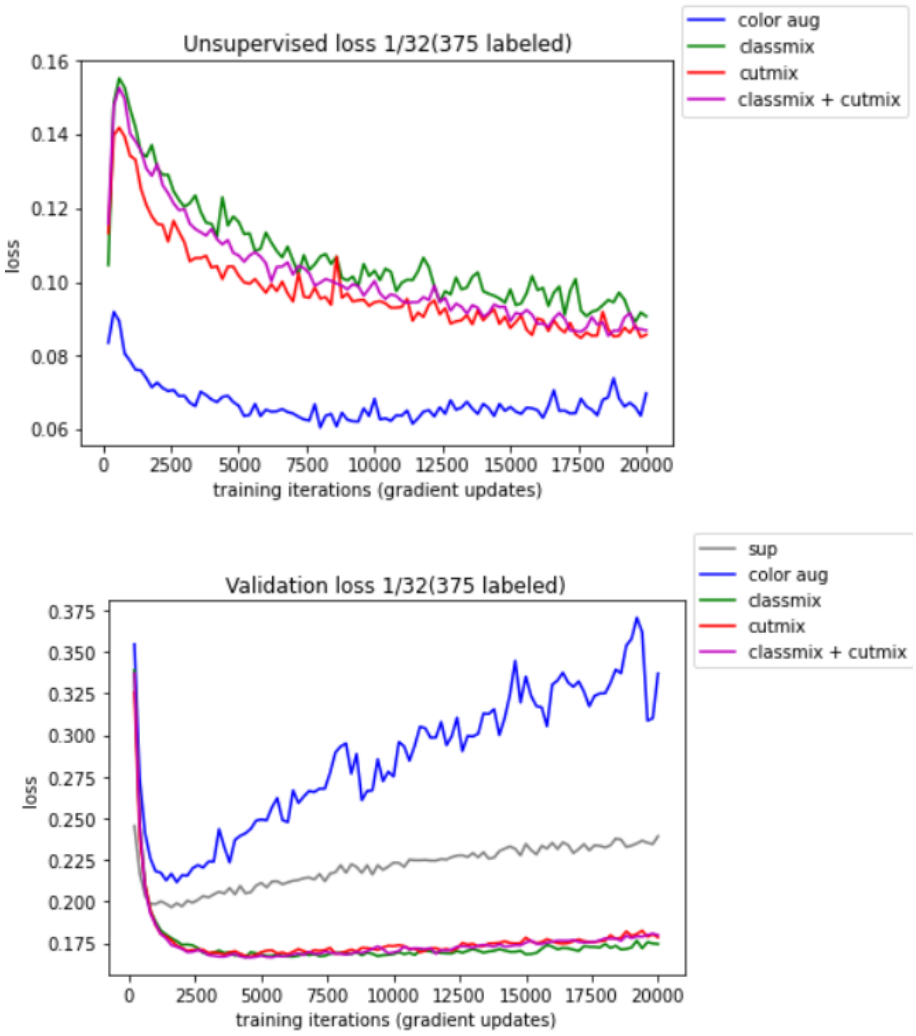
Παρατηρήσεις και σχολιασμός Όπως βλέπουμε από τα παραπάνω αποτελέσματα οι διαταραχές εισόδου ClassMix και CutMix, καθώς και ο συνδυασμός ClassMix + CutMix είναι οι μέθοδοι που επιφέρουν βελτιώσεις σε σύγκριση με το απλό επιβλεπόμενο μοντέλο για τις τρεις διαμερίσεις. Οι ClassMix και ClassMix+CutMix επιφέρουν λίγο καλύτερη γενίκευση από την CutMix, ενώ επίσης παρατηρούμε ότι σε αυτό το σύνολο ο συνδυασμός ClassMix+CutMix δεν βελτιώνει την απόδοση περαιτέρω σε σχέση με το απλό ClassMix. Όσον αφορά τώρα τη διαταράχη εισόδου με εφαρμογή μετασχηματισμών χρώματος (color augmentation/pertrubation) βλέπουμε ότι δεν αποδίδει καλά και μάλιστα το μοντέλο που εκπαιδεύεται με αυτή παρουσιάζει χειρότερη ικανότητα γενίκευσης, ακόμη και από τη απλή επιβλεπόμενη μέθοδο. Ο λόγος που συμβαίνει αυτό είναι γιατί η επαύξηση χρώματος φαίνεται να μην προσδίδει την απαραίτητη ποικιλομορφία στις εικόνες, ώστε το μοντέλο να παρουσιάζει ευστάθεια και να μπορεί να γενικεύει καλύτερα. Όπως, έχουμε αναφέρει στο παράδειγμα εκπαίδευσης ασθενούς-ισχυρής συνέπειας είναι απαραίτητη η ύπαρξη ισχυρών μεθόδων επαύξησης, έτσι ώστε ο teacher να μπορεί να παράγει υψηλής ποιότητας ψευδοετικέτες και ο student με την έκθεση του σε πιο πολύπλοκες εικόνες να γίνεται πιο ανθεκτικός σε μεταβολές και να μαθαίνει καλύτερες αναπαράστασεις που θα τον βοηθήσουν στην καλύτερη γενίκευση σε δεδομένα που βλέπει για πρώτη φορά. Σε αντίθετη περίπτωση, αν οι ισχυροί μετασχηματισμοί αντικατασταθούν από πιο ασθενείς μετασχηματισμούς που δεν προσφέρουν παραπάνω πληροφορία στο δίκτυο, τότε είναι πολύ πιθανό ο student να παράγει προβλέψεις με μεγάλη βεβαιότητα (over-confidence) για τις λιγότερο δύσκολες προς κατάτμηση ασθενώς επαυξημένες εικόνες. Επομένως, είναι πιθανό να παρουσιάσει υπερπροσαρμογή σε ψευδοετικέτες του teacher που θα είναι λιγότερο ακριβείς, λόγω της έλλειψης ισχυρών μετασχηματισμών.

Στην περίπτωση μας προφανώς συμβαίνει κάτι τέτοιο. Οι μεταβολές του χρώματος στις εικόνες που εφαρμόζουμε για να δημιουργήσουμε την ισχυρά επαυξημένη εκδοχή της εικόνας δεν προσφέρουν χρήσιμη πληροφορία που θα βοηθήσει τον student να τμηματοποιεί μέρη του ανθρώπινου προσώπου, επιφέροντας υπερπροσαρμογή στις ενδεχομένως όχι τόσο ποιοτικές ψευδοετικέτες του teacher και μείωση κατά πολύ της ικανότητας γενίκευσης. Αντίθετα,

οι τεχνικές ClassMix, CutMix παράγουν ισχυρά επαυξημένες εικόνες που παρουσιάζουν μεγάλη ποικιλία συνδυασμών από μέρη του ανθρώπινου προσώπου, δίνοντας τα κατάλληλα σήματα εισόδου στο δίκτυο που θα βοηθήσουν στη γενίκευση.

Οι παραπάνω παρατηρήσεις μπορούν να διαπιστωθούν και από τα πειραματικά αποτελέσματα μας. Παίρνοντας, ως παράδειγμα τη διαμέριση 1/32 με 375 ετικέτες παρουσιάζουμε τις καμπύλες μάθησης της μη επιβλεπόμενης απώλειας, καθώς και της απώλειας στο σύνολο επικύρωσης.

Εικόνα 6.13: *Καμπύλες μη επιβλεπόμενης απώλειας και απώλειας επικύρωσης για τη διαμέριση 1/32(375 ετικέτες)*



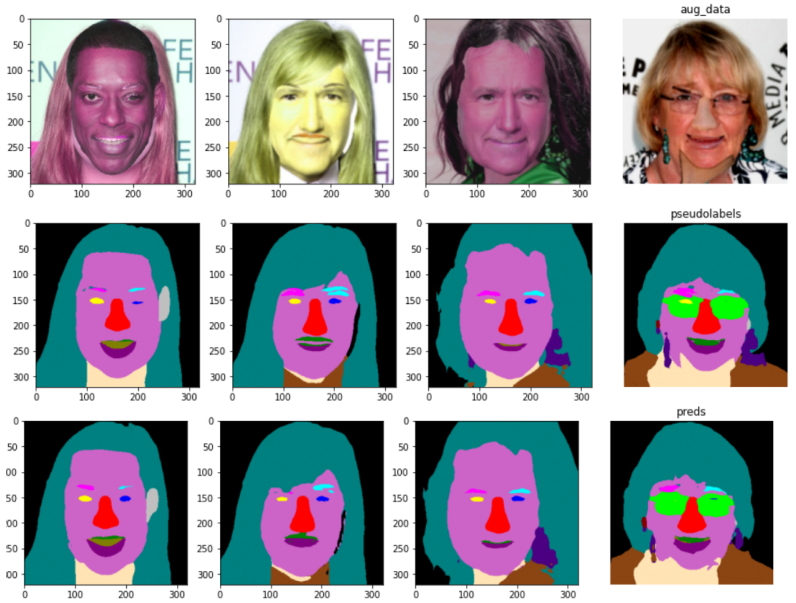
Παρατηρώντας τα δύο παραπάνω σχήματα, αντιλαμβανόμαστε ότι η μέθοδος που αξιοποιεί επαύξηση χρώματος (μπλέ γραμμή), παρουσιάζει μεγάλη υπερπροσαρμογή στις ψευδοετικέτες που παράγονται από τον teacher και αποτυγχάνει να γενικεύσει. Για αυτό και διακρίνουμε αρκετά μικρές τιμές μη επιβλεπόμενης απώλειας και μεγάλες τιμές απώλειας επικύρωσης. Η υπερπροσαρμογή στις ψευδοετικέτες του teacher οδηγεί το μοντέλο να παρουσιάσει χειρότερη γενίκευση ακόμη και από την επιβλεπόμενη μέθοδο. Από την άλλη, οι μέθοδοι ClassMix, CutMix, βλέπουμε ότι παρουσιάζουν μεγαλύτερες τιμές μη επιβλεπόμενης απώλειας, η οποία όμως σταδιακά μειώνεται, πράγμα που σημαίνει ότι ο student μα-

θαίνει σταδιακά την καλύτερη τμηματοποίηση των αναμειγμένων εικόνων υπο την επίβλεψή του teacher , ενώ παρατηρείται αισθητή μείωση της υπερπροσαρμογής και αρκετά καλύτερη ικανότητα γενίκευσης σε σχέση με το επιβλεπόμενο μοντέλο.

Εν συνεχεία παρουσιάζουμε τις ψευδοετικέτες που παράγει ο teacher και τις αντίστοιχες προβλέψεις του student για τις μεθόδους color perturbation και ClassMix κατά τη διάρκεια της εκπαίδευσης. Ενδεικτικά παραθέτουμε τα αποτελέσματα γις τις δύο μεθόδους μετά από 1600 επαναλήψεις εκπαίδευσης (training iterations, gradient updates).

Εικόνα 6.14: *Ψευδοετικέτες και αντίστοιχες προβλέψεις κατά τη διάρκεια της εκπαίδευσης με τη μέθοδο επαύξησης χρώματος*

Εικόνα 6.15: *Ψευδοετικέτες και αντίστοιχες προβλέψεις κατά τη διάρκεια της εκπαίδευσης με τη μέθοδο ClassMix*



Μπορούμε να παρατηρήσουμε ότι το μοντέλο που χρησιμοποιεί ως μέθοδο ισχυρής επα-

ύξησης μετασχηματισμούς χρώματος αδυνατεί να παράξει καλές ψευδοετικέτες όσον αφορά ,κυρίως την τμηματοποίηση των ματιών, των φρυδιών και σε ένα βαθμό του στόματος. Αντίθετα, βλέπουμε ότι οι teacher-student που εκπαιδεύονται με τη μέθοδο ClassMix για τις ίδιες επαναλήψεις εκπαίδευσης μπορούν να παράξουν καλύτερης ποιότητας ψευδοετικέτες(teacher) και καλύτερης ποιότητας προβλέψεις(student). Επομένως, οι αρχικές μας παρατηρήσεις σχετικά με την υπερπροσαρμογή του student σε μέτριας ποιότητας ψευδοετικέτες για τη μέθοδο color perturbation διαπιστώνονται και από τα παραπάνω ποιοτικά αποτελέσματα.

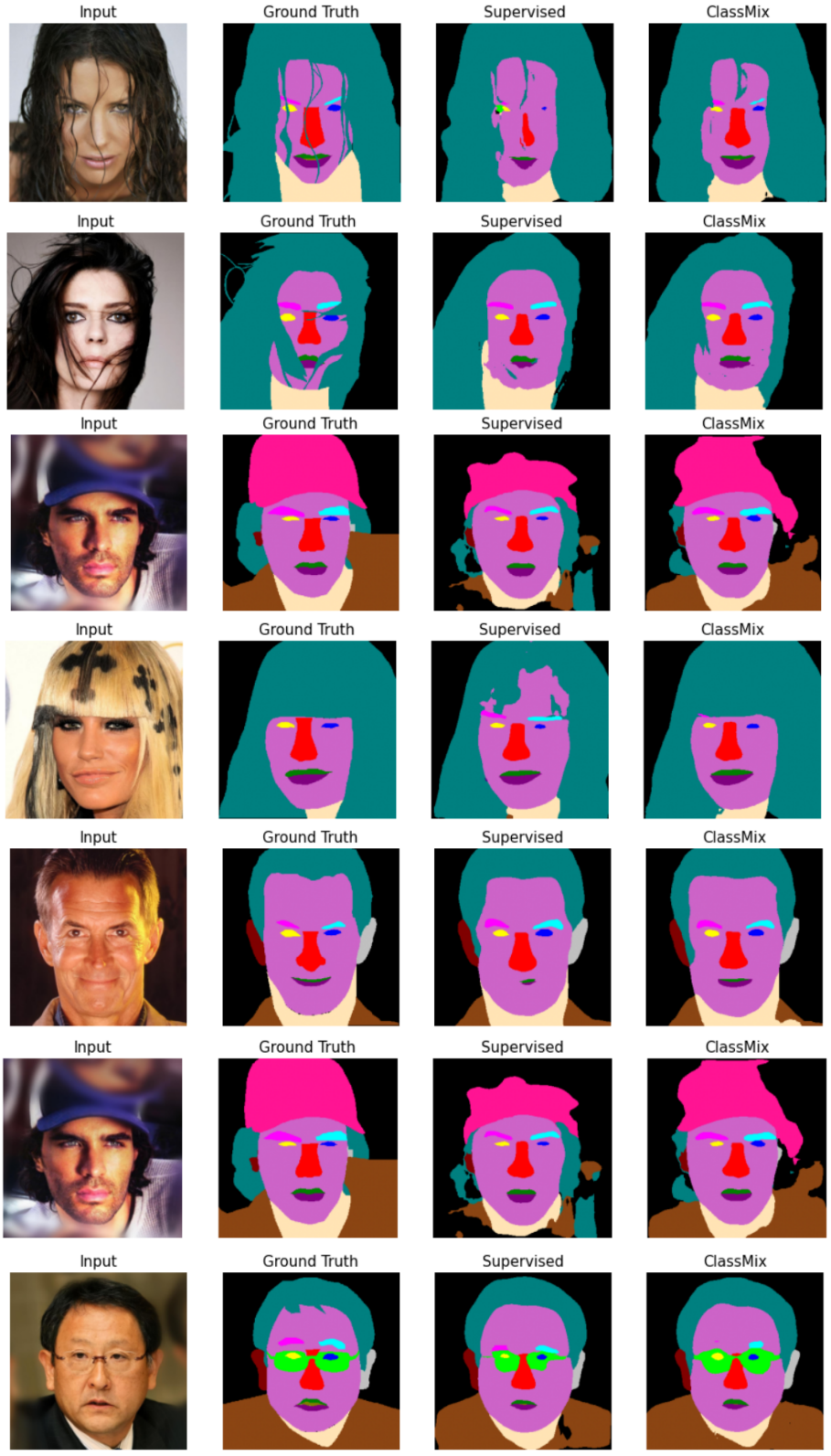
Για τη διαμέριση 1/64(187 ετικέτες) παρατηρείται η μεγαλύτερη βελτίωση σε σύγκριση με το επιβλεπόμενο μοντέλο. Για τη συγκεκριμένη διαμέριση παραθέτουμε αναλυτικά τη μετρική IOU για κάθε κλάση για τις μεθόδους supervised και ClassMix.

Πίνακας 6.5: *Μετρική IOU της κάθε κλάσης στο σύνολο CelebAMask-HQ για τις μεθόδους Supervised και ClassMix(187 ετικέτες)*

	Sup only	ClassMix
background:0	0.8980	**0.9183**
skin:1	0.9085	**0.9174**
nose:2	0.8652	**0.8725**
eye glasses:3	0.6915	**0.7633**
left eye:4	0.7649	**0.7672**
right-eye:5	0.7622	**0.7669**
left brow:6	0.6863	**0.7090**
right brow:7	0.6854	**0.7082**
left ear:8	0.7004	**0.7435**
right ear:9	0.6959	**0.7356**
mouth:10	0.7644	**0.7715**
upper lip:11	0.7207	**0.7550**
lower lip:12	0.7564	**0.7936**
hair:13	0.8796	**0.9022**
hat:14	0.5269	**0.7512**
earring:15	0.2898	**0.3679**
neck:16	0.7743	**0.8041**
cloth:17	0.6248	**0.7334**

Παρακάτω για τα ίδια μοντέλα (supervised, ClassMix) παραθέτουμε ορισμένα ποιοτικά αποτελέσματα που παράγει το δίκτυο τμηματοποίησης για τυχαίες εικόνες από το σύνολο επικύρωσης του CelebAMask-HQ

Εικόνα 6.16: *Ποιοτικά αποτελέσματα που παράγει το δίκτυο για εικόνες από το σύνολο επικύρωσης του CelebAMask-HQ*



6.4 Πειραματικά αποτελέσματα για το σύνολο QaTa-COV19-v2

Παραθέτουμε το μέσο IOU πάνω στο σύνολο επικύρωσης του QaTa-COV19-v2

Πίνακας 6.6: *Πειραματικά αποτελέσματα (mIOU) για το σύνολο επικύρωσης του QaTa-COV19*

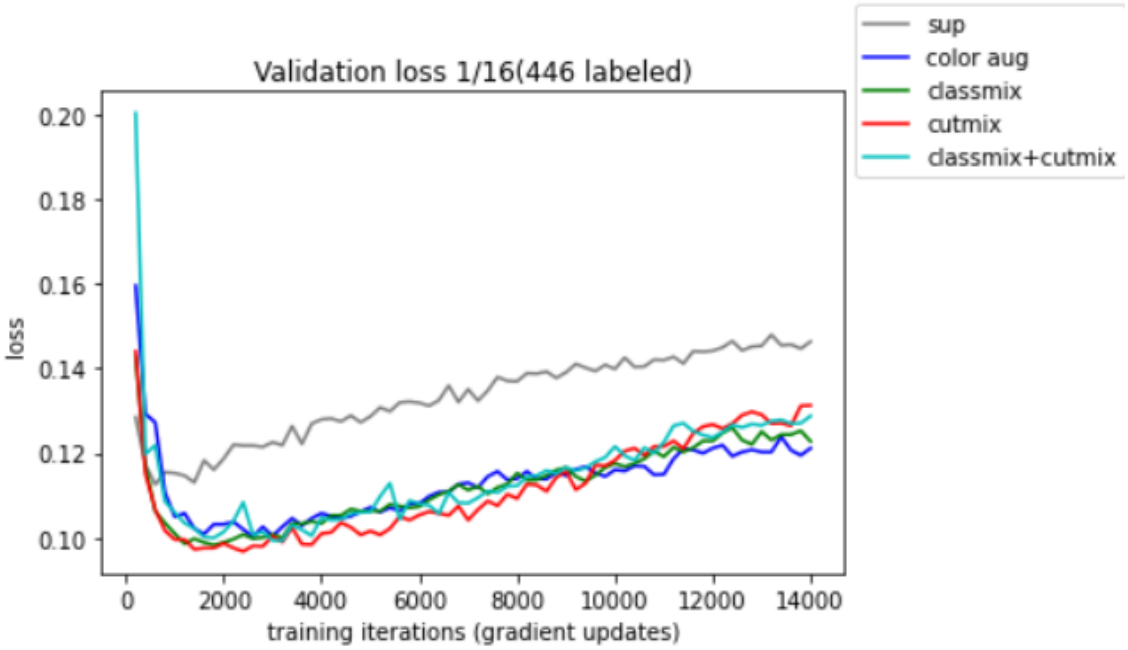
Method	1/32(223)	1/16(446)	1/8(893)	1/4(1786)	all labels(7145)
Supervised	81.16	83.42	83.85	85.11	86.76
Color perturbation	83.25	**84.65**	85.23	86.17	-
CutMix	82.82	84.48	85.19	**86.33**	-
ClassMix	**83.42**	84.47	85.16	86.14	-
ClassMix + CutMix	82.94	84.50	**85.27**	86.29	-

Παρατηρούμε ότι η αξιοποίηση των επιπλέον μη επισημασμένων δεδομένων με εφαρμογή κανονικοποίησης συνέπειας επιφέρει βελτίωση σε σχέση με την απλή επιβλεπόμενη εκπαίδευση. Όπως βλέπουμε, τόσο η εφαρμογή μετασχηματισμών χρώματος, όσο και η εφαρμογή των πιο σύνθετων μεθόδων μίξης ClassMix, CutMix, επιφέρουν παρόμοιες αυξήσεις στην απόδοση του μοντέλου. Αυτό σημαίνει ότι η επαύξηση χρώματος για το συγκεκριμένο σύνολο δεδομένων προσφέρει ικανοποιητική ποικιλομορφία, ώστε να αξιοποιηθεί η πληροφορία των μη επισημασμένων δεδομένων, ενώ η εφαρμογή των τεχνικών ClassMix και CutMix δεν προσδίδουν επιπρόσθετη πληροφορία κατά την εκπαίδευση. Όπως φαίνεται, ακόμη και η απλή επιβλεπόμενη μέθοδος μπορεί να γενικεύει σε ικανοποιητικό βαθμό στο σύνολο επικύρωσης, για αυτό και υπάρχει λιγότερη αύξηση στην απόδοση σε σχέση με τα προηγούμενα δύο σύνολα.

Ενδεικτικά παραθέτουμε τις καμπύλες της μη επιβλεπόμενης απώλειας και της απώλειας επικύρωσης για τη διαμέριση 1/16 (446 ετικέτες).

Εικόνα 6.17: *Καμπύλη μη επιβλεπόμενης απώλειας για τη διαμέριση για το σύνολο QaTa-COV19*

Εικόνα 6.18: *Καμπύλη απώλειας επικύρωσης στο σύνολο QaTa-COV19*



Παρατηρούμε ότι όλες οι μέθοδοι έχουν παρόμοιες τιμές μη επιβλεπόμενης απώλειας, η οποία μειώνεται φυσιολογικά, κάτι που σημαίνει ότι ο student μαθαίνει ομαλά από τις ψευδοετικέτες του teacher. Παράλληλα, βλέπουμε πως εξίσου όλες οι μέθοδοι επιφέρουν μείωση της υπερπροσαρμογής, σε σχέση με το επιβλεπόμενο μοντέλο και καλύτερη ικανότητα γενίκευσης.

Παρακάτω για την ίδια διαμέριση παραθέτουμε το IOU για τις δύο κλάσεις του προβλήματος.

Πίνακας 6.7: *Μετρική IOU της κάθε κλάσης στο σύνολο QaTa-COV19 για τη διαμέριση 1/16(446)*

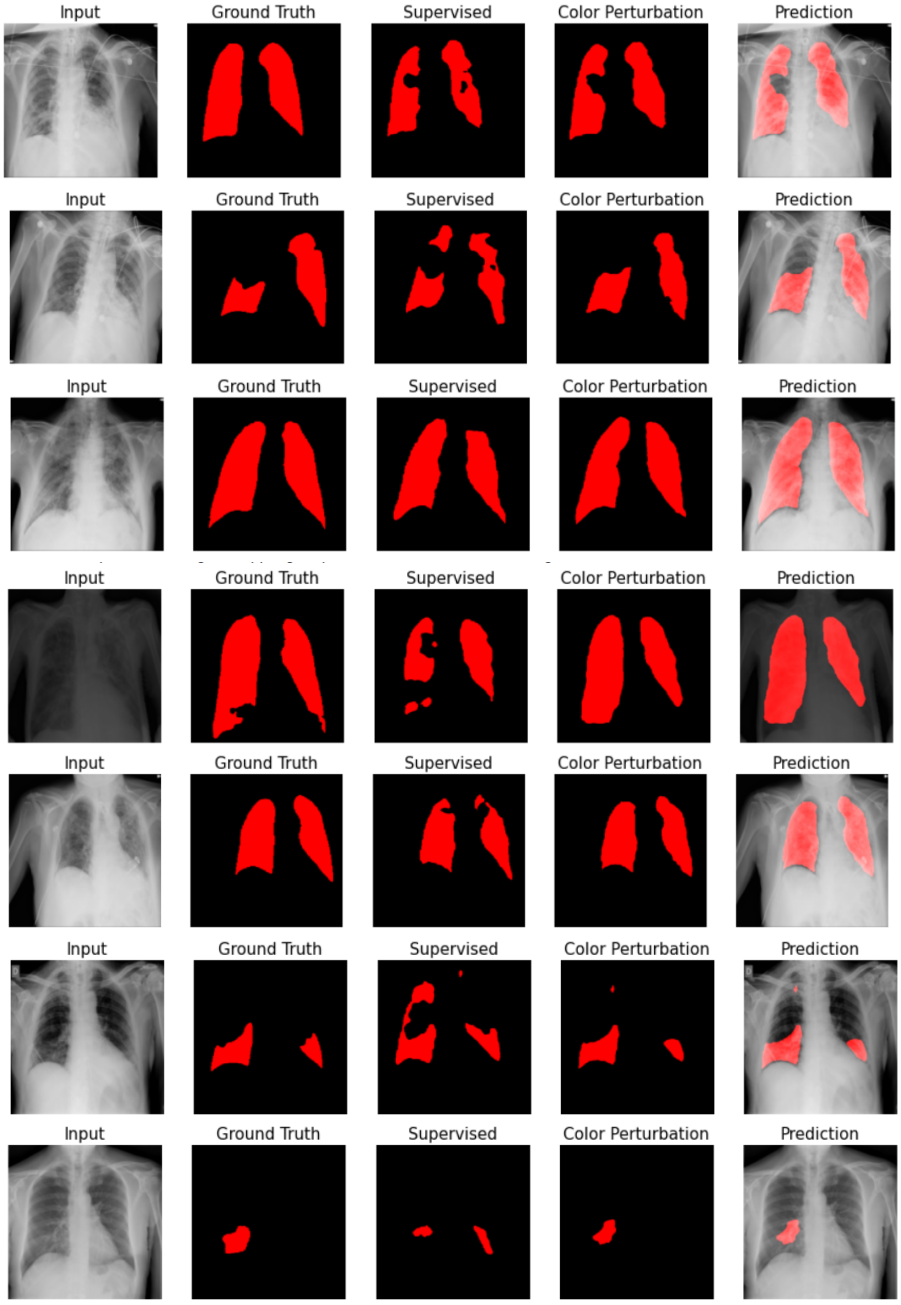
	Sup only	Color Aug	ClassMix	CutMix	ClassMix+CutMix
background:0	0.9551	**0.9595**	0.9588	0.9590	0.9590
infection:1	0.7134	**0.7337**	0.7307	0.7307	0.7312

Πίνακας 6.8: *Μετρική Recall της κάθε κλάσης στο σύνολο QaTa-COV19 για τη διαμέριση 1/16(446)*

	Sup only	Color Aug	ClassMix	CutMix	ClassMix+CutMix
background:0	0.9748	**0.9797**	0.9790	0.9791	0.9791
infection:1	0.8358	0.8442	0.8444	0.8444	**0.8447**

Παρατηρώντας και τη μετρική recall βλέπουμε ότι οι ημιεπιβλεπόμενες τεχνικές συνεισφέρουν στην αύξηση της συγκεκριμένης μετρικής και για τις δύο κλάσεις. Αυτό σημαίνει ότι έχουμε λιγότερο κακή ταξινόμηση στα pixel που ανήκουν στην κλάση infection και κατηγοριοποιούνται στην background (FN), σημαντικό για προβλήματα ιατρικής φύσεως.

Εικόνα 6.19: *Ποιοτικά αποτελέσματα που παράγει το δίκτυο για εικόνες από το σύνολο επικύρωσης του QaTa-COV19 για τις μεθόδους supervised και color perturbation στη διαμέριση 1/16(446 ετικέτες)*



Κεφάλαιο 7

Τελικά συμπεράσματα και μελλοντικές επεκτάσεις

7.1 Συμπεράσματα

Ανακεφαλαιώνοντας, στην παρούσα διπλωματική μελετήσαμε την επίδραση τεχνικών ημιεπιβλεπόμενης μάθησης, βασισμένες στην κανονικοποίηση συνέπειας, οι οποίες χρησιμοποιούν διαταραχές επιπέδου εισόδου (input-level perturbations) κατά το πρότυπο της ασθενούς-ισχυρής συνέπειας για την αξιοποίηση των μη επισημασμένων δεδομένων. Από τα πειράματα που εκτελέσαμε στα τρία σύνολα δεδομένων παρατηρήσαμε γενικά ότι η κανονικοποίηση συνέπειας βοήθησε στη βελτίωση της απόδοσης σε σχέση την απλή επιβλεπόμενη εκπαίδευση. Η μέθοδος που επιλέγεται για την παραγωγή της ισχυρά επαυξημένης εκδοχής της εικόνας εισόδου έχει διαφορετική επίδραση στην απόδοση του μοντέλου.

Πιο συγκεκριμένα, όσον αφορά το σύνολο δεδομένων Pascal παρατηρούμε ότι όλες οι μέθοδοι ισχυρής επαύξησης που εφαρμόζονται (μετασχηματισμός χρώματος, ClassMix, CutMix, ClassMix+CutMix αποδεικνύονται ιδιαίτερα αποτελεσμάτικές 6.4, καθώς βελτιώνουν σε μεγάλο βαθμό την απόδοση σε σχέση με το επιβλεπόμενο μοντέλο. Σε όλες τις διαμερίσεις επισημασμένων/μη επισημασμένων δεδομένων βλέπουμε ότι η μέθοδος που αξιοποιεί δύο ισχυρά επαυξημένες εκδοχές, σύμφωνα με το [14] ClassMix+CutMix αποδίδει καλύτερα.

Στα πειράματα του συνόλου CelebAMask-HQ παρατηρούμε ότι η εφαρμογή μετασχηματισμού χρώματος (color augmentation) δεν αποδίδει καθόλου καλά και μάλιστα οδηγεί σε χειρότερη ικανότητα γενίκευσης σε σχέση με την απλή επιβλεπόμενη μάθηση. Αυτό, όπως αναφέραμε συμβαίνει επειδή για το πρόβλημα της τμηματοποίησης μερών του ανθρώπινου προσώπου η απλή μεταβολή του χρώματος δεν παρέχει την απαραίτητη πληροφορία στο δίκτυο, ώστε να μπορεί να παράξει ποιοτικές ψευδοετικέτες και να μάθει να τμηματοποιεί με επιτυχία τις μη επισημασμένες εικόνες με αποτέλεσμα να παρατηρείται υπερπροσαρμογή σε μη ακριβείς ψευδοετικέτες και συνεπώς απώλεια της ικανότητας γενίκευσης. Αντιθέτως, η χρήση των τεχνικών μίξης ClassMix, CutMix για την παραγωγή των ισχυρά επαυξημένων εικόνων αποδεικνύεται αρκετά αποδοτική, καθώς με την ανάμειξη εικόνων διαφορετικών προσώπων μπορούν να παραχθούν δείγματα με μεγαλύτερη ποικιλομορφία που συνδυάζουν τμήματα του ανθρώπινου προσώπου που διαφέρουν από άτομο σε άτομο, σε σχήμα και σε μέγεθος(μάτια, αυτιά, στόμα κ.λ.π). Επομένως, το δίκτυο δέχεται πιο χρήσιμη πληροφορία για την εκμάθηση της τμηματοποίησης αυτών των μερών.

Όσον αφορά το σύνολο QaTa-COV19, παρατηρούμε ότι οι τεχνικές ισχυρής επαύξησης προσφέρουν την ίδια βελτίωση στην απόδοση, με τις μεθόδους μίξης ClassMix, CutMix να μην βοηθούν περαιτέρω όπως στα προηγούμενα σύνολα. Για το συγκεκριμένο σύνολο, λοιπόν, μπορούμε να πούμε ότι ένας σχετικά απλός μετασχηματισμός χρώματος είναι επαρκής για να προσδώσει την απαραίτητη ποικιλομορφία, ώστε το δίκτυο να επωφεληθεί από την επιπλέον πληροφορία των μη επισημασμένων δεδομένων.

Γενικά, καταλήγουμε στο συμπέρασμα, ότι το παράδειγμα εκπαίδευσης ασθενούς-ισχυρής συνέπειας για την αξιοποίηση των μη επισημασμένων δεδομένων μπορεί να επιφέρει σημαντικές βελτιώσεις στην απόδοση του μοντέλου σε σύγκριση με την απλή επιβλεπόμενη εκπαίδευση. Μείζονος σημασίας είναι η επιλογή της μεθόδου για τη δημιουργία της ισχυρά επαυξημένης εκδοχής ενός δείγματος. Οι μετασχηματισμοί που εφαρμόζονται χρειάζεται να προσφέρουν την απαραίτητη ποικιλομορφία, ώστε να αποφεύγεται η υπερπροσαρμογή σε κακής ποιότητας ψευδοετικέτες που μπορούν να οδηγήσουν σε κακή ικανότητα γενίκευσης και συνεπώς στην υποβάθμιση του συγκεκριμένου παραδείγματος εκπαίδευσης σε μία αφελή διαδικασία self-training[76]. Επιπλέον, όσον αφορά το είδος της ισχυρής επαύξησης που πρέπει να επιλεχθεί, αυτό εξαρτάται σε μεγάλο βαθμό από την κατανόμη των δεδομένων του κάθε συνόλου και τη φύση του εκάστοτε προβλήματος τμηματοποίησης και πρέπει να διαπιστώνεται μέσω πειραμάτων, με τις μεθόδους μίξης ClassMix, CutMix να αποτελούν συνήθως καλές επιλογές για προβλήματα εικόνων του φυσικού κόσμου (natural images).

7.2 Μελλοντικές επεκτάσεις

Κάποιες πιθανές μελλόντικες επεκτάσεις που μπορούν να γίνουν για περαιτέρω διερεύνηση και πειραματισμό είναι οι παρακάτω:

- Εφαρμογή του παραδείγματος της ασθενούς-ισχυρής συνέπειας (strong-to-weak consistency) και των τεχνικών ClassMix και CutMix για την παραγωγή των ισχυρά επαυξημένων δειγμάτων σε περισσότερα σύνολα δεδομένων από διαφορετικά πεδία. Για παράδειγμα σε σύνολα δεδομένων δορυφορικών εικόνων (landcover segmentation), σε επιπλέον σύνολα τμηματοποίησης ανθρώπινου προσώπου (human face segmentation), καθώς και σε άλλα σύνολα τμηματοποίησης ιατρικών εικόνων με σκοπό την αξιοποίηση μη επισημασμένων δεδομένων για τη βελτίωση της απόδοσης.

- Περαιτέρω διευρεύνηση για την αποδοτικότητα των πολλαπλών ισχυρών εκδοχών[14] που δίνονται ως είσοδο στο δίκτυο student και επιπλέον πειραματισμός με διάφορους συνδυασμούς τεχνικών επαύξησης για την παραγωγή αυτών των ισχυρά επαυξημένων εκδοχών (π.χ ClassMix-CutMix, ClassMix-ClassMix, Classmix-Color augmentation κ.λ.π).

- Διερεύνηση της αποδοτικότητας του συνδυασμού διαταραχών επιπέδου εισόδου(input-level perturbations) και διαταραχών επιπέδου ενδιάμεσων χαρακτηριστικών (feature-level perturbations) που περιγράφεται στο [14] στα σύνολα CelebAMask-HQ, QaTa-COV19, καθώς και σε άλλα σύνολα δεδομένων που δεν έχει δοκιμαστεί μέχρι τώρα.

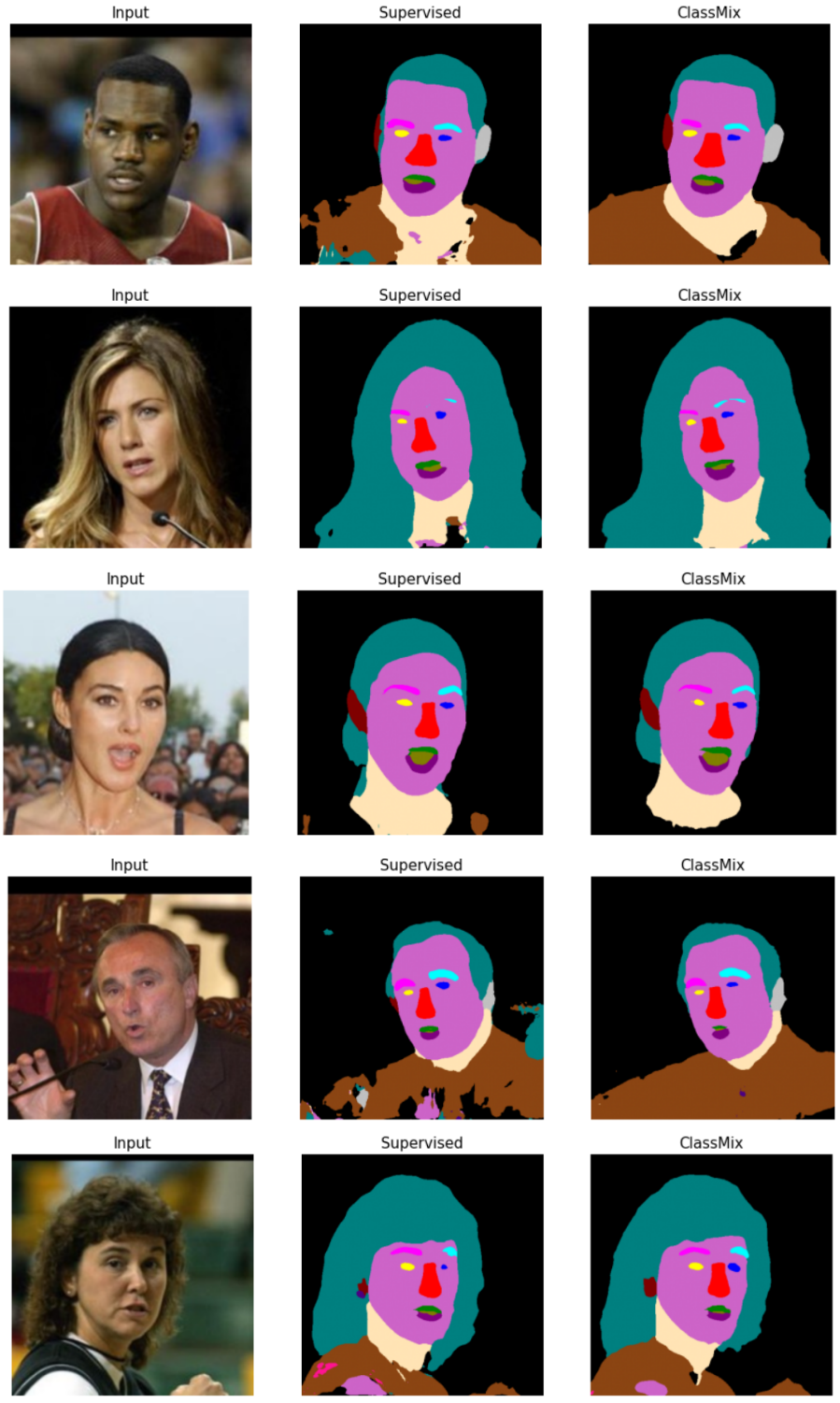
- Εφαρμογή των τεχνικών ClassMix, CutMix, ClassMix+CutMix με χρήση διαφορετικών αρχιτεκτονικών κατάτμησης εκτός από το DeepLabv3plus.
- Πειραματισμός με διαφορετικές συναρτήσεις απώλειας για τη μέτρηση του σφάλματος μεταξύ ψευδοετικετών του teacher και προβλέψεων του student. Για παράδειγμα αντί για το κόστος cross-entropy μπορεί να επιτευχθεί η συνέπεια μεταξύ student-teacher με χρήση του MSE ή της Focal loss [101] (μέσω αυτής της απώλειας μπορεί να δοθεί βαρύτητα στις κλάσεις που το δίκτυο αντιμετωπίζει δυσκολία στην τμηματοποίηση τους), καθώς και άλλες συναρτήσεις κόστους.
- Εφαρμογή κανονικοποίησης συνέπειας για την περίπτωση της ημιεπιβλεπόμενης κατηγοριοποίησης ή τμηματοποίησης βίντεο.

Παραρτήματα

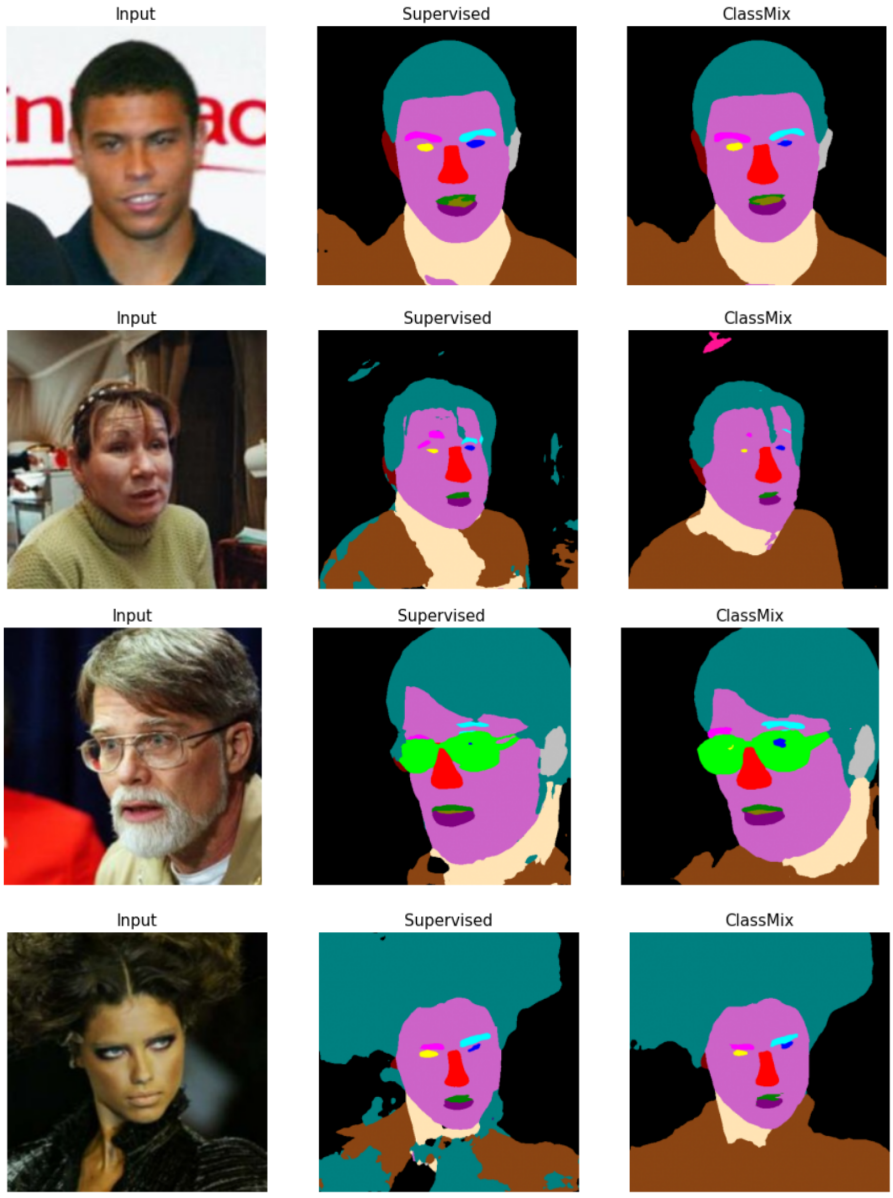
Παράρτημα Α΄

Ποιοτικά αποτελέσματα για εικόνες από το σύνολο δεδομένων Labeled Faces in the Wild

Εικόνα Α΄.1: *Ποιοτικά αποτελέσματα που παράγει το δίκτυο εκπαιδευμένο στο σύνολο CelebAMask-HQ με 187 επισημασμένα δείγματα για εικόνες από το σύνολο Labeled Faces in the Wild [27]*



Εικόνα Α΄.2: *Ποιοτικά αποτελέσματα που παράγει το δίκτυο εκπαιδευμένο στο σύνολο CelebAMask-HQ με 187 επισημασμένα δείγματα για εικόνες από το σύνολο Labeled Faces in the Wild [27]*



Βιβλιογραφία

[1] Hmrishav Bandyopadhyay. *An Introduction to Image Segmentation: Deep Learning vs. Traditional [+Examples]*. https://www.v7labs.com/blog/image-segmentation-guide#h5, 2023. Ημερομηνία πρόσβασης: 3-1-2023.

[2] Pulkit Sharma. *Computer Vision Tutorial: A Step-by-Step Introduction to Image Segmentation Techniques (Part 1)*. https://www.analyticsvidhya.com/blog/2019/04/introduction-image-segmentation-techniques-python, 2022. Ημερομηνία πρόσβασης: 3-1-2023.

[3] Jeremy Jordan. *An overview of semantic image segmentation*. https://www.jeremyjordan.me/semantic-segmentation/, 2018. Ημερομηνία πρόσβασης: 3-1-2023.

[4] Serena Yeung Fei-Fei Li, Justin Johnson. *Lecture 11: Detection and Segmentation*. http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf, 2017. Ημερομηνία πρόσβασης: 3-1-2023.

[5] Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla. *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[6] Liang Chieh Chen, George Papandreou, Florian Schroff and Hartwig Adam. *Rethinking atrous convolution for semantic image segmentation*. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L. Yuille. *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[8] Liang Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff and Hartwig Adam. *Encoder-decoder with atrous separable convolution for semantic image segmentation*. *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[9] Jesper Engelen and Holger Hoos. *A survey on semi-supervised learning*. *Machine Learning*, 109, 2020.

[10] Amit Chaudhary. *Semi-Supervised Learning in Computer Vision*. https://amitness.com/2020/07/semi-supervised-learning/, 2020. Ημερομηνία πρόσβασης: 3-1-2023.

[11] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio and David Lopez-Paz. *Interpolation consistency training for semi-supervised learning*. *Neural Networks*, 145:90–106, 2022.

[12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens and Quoc V Le. *Randaugment: Practical automated data augmentation with a reduced search space*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[13] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin and Chun Liang Li. *Fixmatch: Simplifying semi-supervised learning with consistency and confidence*. *Advances in neural information processing systems*, 33:596–608, 2020.

[14] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang and Yinghuan Shi. *Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation*. *CVPR*, 2023.

[15] Geoffrey French, Timo Aila, Samuli Laine, Michal Mackiewicz and Graham Finlayson. *Consistency regularization and cutmix for semi-supervised semantic segmentation*. *arXiv preprint arXiv:1906.01916*, 2(4):5, 2019.

[16] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto and Lennart Svensson. *ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning*. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1369–1378, 2021.

[17] Yassine Ouali, Celine Hudelot and Myriam Tami. *Semi-Supervised Semantic Segmentation With Cross-Consistency Training*. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Xiaokang Chen, Yuhui Yuan, Gang Zeng and Jingdong Wang. *Semi-Supervised Semantic Segmentation With Cross Pseudo Supervision*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, 2021.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[20] Cheng Han Lee, Ziwei Liu, Lingyun Wu and Ping Luo. *MaskGAN: Towards Diverse and Interactive Facial Image Manipulation*. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] Aysen Degerli, Mete Ahishali, Serkan Kiranyaz, Muhammad EH Chowdhury and Moncef Gabbouj. *Reliable covid-19 detection using chest x-ray images*. *2021 IEEE International Conference on Image Processing (ICIP)*, pages 185–189. IEEE, 2021.

[22] *QaTa-COV19 Dataset: Qatar University and Tampere University and Hamad Medical Corporation*. https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset. Ημερομηνία πρόσβασης: 24-1-2023.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Deep residual learning for image recognition*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] *Transforming and augmenting images:ColorJitter*. https://pytorch.org/vision/main/generated/torchvision.transforms.ColorJitter.html. Ημερομηνία πρόσβασης: 17-1-2023.

[25] *Transforming and augmenting images*. https://pytorch.org/vision/main/transforms.html. Ημερομηνία πρόσβασης: 17-1-2023.

[26] Jeremy Jordan. *Evaluating image segmentation models*. https://www.jeremyjordan.me/evaluating-image-segmentation-models/, 2018. Ημερομηνία πρόσβασης: 6-2-2023.

[27] Gary B. Huang, Manu Ramesh, Tamara Berg and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Τεχνική Αναφορά με αριθμό 07-49, University of Massachusetts, Amherst, 2007.

[28] Senay Cakir, Marcel Gauß, Kai Häppeler, Yassine Ounajjar, Fabian Heinle and Reiner Marchthaler. *Semantic Segmentation for Autonomous Driving: Model Evaluation, Dataset Generation, Perspective Comparison, and Real-Time Capability*. *arXiv preprint arXiv:2207.12939*, 2022.

[29] Çağrı Kaymak and Ayşegül Uçar. *Semantic Image Segmentation for Autonomous Driving Using Fully Convolutional Networks*. *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–8, 2019.

[30] Sanchit Gautam, Tarosh Mathuria and Shweta Meena. *Image Segmentation for Self-Driving Car*. *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6, 2022.

[31] Giulia Rizzoli, Francesco Barbato and Pietro Zanuttigh. *Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives*. *Technologies*, 10(4), 2022.

[32] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng and Asoke K. Nandi. *Medical image segmentation using deep learning: A survey*. *IET Image Processing*, 16(5):1243–1267, 2022.

[33] Natalia Salpea, Paraskevi Tzouveli and Dimitrios Kollias. *Medical Image Segmentation: A Review of Modern Architectures*. *Computer Vision - ECCV 2022 Workshops*Leonid Karlinsky, Tomer Michaeli and Ko Nishino, editors, pages 691–708, Cham, 2023. Springer Nature Switzerland.

[34] Wuttichai Boonpook, Yumin Tan, Attawut Nardkulpat, Kritanai Torsri, Peerapong Torteeka, Patcharin Kamsing, Utane Sawangwit, Jose Pena and Montri Jainaen. *Deep Learning Semantic Segmentation for Land Use and Land Cover Types Using Landsat 8 Imagery*. *ISPRS International Journal of Geo-Information*, 12(1), 2023.

[35] Vasilis Pollatos, Loukas Kouvaras and Eleni Charou. *Land cover semantic segmentation using ResUNet*. *arXiv preprint arXiv:2010.06285*, 2020.

[36] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen and Nong Sang. *Context Prior for Scene Segmentation*. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[37] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth and Bernt Schiele. *The Cityscapes Dataset for Semantic Urban Scene Understanding*. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[38] Yue Fan, Anna Kukleva, Dengxin Dai and Bernt Schiele. *Revisiting consistency regularization for semi-supervised learning*. *International Journal of Computer Vision*, pages 1–18, 2022.

[39] Geoffrey French and Michal Mackiewicz. *Colour augmentation for improved semi-supervised semantic segmentation.* *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - (Volume 4)*, pages 356 – 363, 2022.

[40] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos D. Kollias and Georgios L. Tagaris. *Deep neural architectures for prediction in healthcare*. *Complex & Intelligent Systems*, 4:119–131, 2018.

[41] Athanasios Tagaris, Dimitrios Kollias and Andreas Stafylopatis. *Assessment of Parkinson's Disease Based on Deep Neural Networks*. pages 391–403, 2017.

[42] Athanasios Tagaris, Dimitrios Kollias, Andreas Stafylopatis, Georgios Tagaris and Stefanos Kollias. *Machine Learning for Neurodegenerative Disorder Diagnosis — Survey of Practices and Launch of Benchmark Dataset*. *International Journal on Artificial Intelligence Tools*, 27, 2018.

[43] Ilianna Kollia, Andreas Georgios Stafylopatis and Stefanos Kollias. *Predicting Parkinson's Disease using Latent Information extracted from Deep Neural Networks*. *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[44] James Wingate, Ilianna Kollia, Luc Bidaut and Stefanos D. Kollias. *A Unified Deep Learning Approach for Prediction of Parkinson's Disease*. *CoRR*, αβσ/1911.10653, 2019.

[45] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian and Stefanos Kollias. *Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 537–544, 2021.

[46] Dimitrios Kollias, Anastasios Arsenos and Stefanos Kollias. *Ai-mia: Covid-19 detection & severity analysis through medical imaging*. *arXiv preprint arXiv:2206.04732*, 2022.

[47] Anastasios Arsenos, Dimitrios Kollias and Stefanos Kollias. *A Large Imaging Database and Novel Deep Neural Architecture for Covid-19 Diagnosis*. *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.

[48] Dimitrios Kollias, Anastasios Arsenos and Stefanos Kollias. *AI-MIA: Covid-19 detection and severity analysis through medical imaging*. *Computer Vision-ECCV 2022 Workshops: Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, pages 677–690. Springer, 2023.

[49] Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis and Stefanos Kollias. *Adaptation and contextualization of deep neural network models*. *2017 IEEE symposium series on computational intelligence (SSCI)*, pages 1–8. IEEE.

[50] D Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, I Kollia, L Sukissian, J Wingate and S Kollias. *Deep Transparent Prediction through Latent Representation Analysis*. *arXiv preprint arXiv:2009.07044*, 2020.

[51] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate and S Kollias. *Transparent adaptation in deep medical image diagnosis*. *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, pages 251–267. Springer, 2020.

[52] Fabio De Sousa Ribeiro, Francesco Caliv&39;a, Mark Swainson, Kjartan Gudmundsson, Georgios Leontidis and Stefanos Kollias. *Deep bayesian self-training*. *Neural Computing and Applications*, 32(9):4275–4291, 2020.

[53] Fabio De Sousa Ribeiro, Georgios Leontidis and Stefanos Kollias. *Capsule routing via variational bayes*. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3749–3756, 2020.

[54] Fabio De Sousa Ribeiro, Georgios Leontidis and Stefanos Kollias. *Introducing routing uncertainty in capsule networks*. *Advances in Neural Information Processing Systems*, 33:6490–6502, 2020.

[55] Nikolaos Simou and Stefanos Kollias. *Fire: A fuzzy reasoning engine for impecise knowledge*. Citeseer.

[56] Francesco Caliva, Fabio Sousa De Ribeiro, Antonios Mylonakis, Christophe Demazi'ere, Paolo Vinai, Georgios Leontidis and Stefanos Kollias. *A deep learning approach to anomaly detection in nuclear reactors*. *2018 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2018.

[57] Stefanos Kollias, Miao Yu, James Wingate, Aiden Durrant, Georgios Leontidis, Georgios Alexandridis, Andreas Stafylopatis, Antonios Mylonakis, Paolo Vinai and Christophe Demaziere. *Machine learning for analysis of real nuclear plant data in the frequency domain*. *Annals of Nuclear Energy*, 177:109293, 2022.

[58] Bashar Alhnaity, Stefanos Kollias, Georgios Leontidis, Shouyong Jiang, Bert Schamp and Simon Pearson. *An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth*. *Information Sciences*, 560:35–50, 2021.

[59] Bashar Alhnaity, Simon Pearson, Georgios Leontidis and Stefanos Kollias. *Using deep learning to predict plant growth and yield in greenhouse environments*. *International Symposium on Advanced Technologies and Management for Innovative Greenhouses: GreenSys2019 1296*, pages 425–432, 2019.

[60] Andreas Psaroudakis and Dimitrios Kollias. *MixAugment & Mixup: Augmentation Methods for Facial Expression Recognition*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2375, 2022.

[61] Dimitrios Kollias. *Abaw: Learning from synthetic data & multi-task learning challenges*. *European Conference on Computer Vision*, pages 157–172. Springer, 2023.

[62] G Caridakis, A Raouzaiou, K Karpouzis and S Kollias. *Synthesizing Gesture Expressivity Based on Real Sequences*. *Workshop Programme*, volume 10, page 19.

[63] *Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback*. *Proceedings IEEE International Conference on Multimedia Computing and Systems*, volume 2, pages 954–958. IEEE, 1999.

[64] Yassine Ouali, Céline Hudelot and Myriam Tami. *An overview of deep semi-supervised learning*. *arXiv preprint arXiv:2006.05278*, 2020.

[65] Philippe Thomas. *Semi-Supervised Learning by Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (Review)*. *IEEE Transactions on Neural Networks*, 20:542, 2009.

[66] Alexander Zien Olivier Chapelle, Bernhard Schölkopf. *Semi-Supervised Learning - MIT Press*. 2006.

[67] Antti Tarvainen and Harri Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*. *Advances in neural information processing systems*, 30, 2017.

[68] James Chen. *What is EMA? How to Use Exponential Moving Average With Formula*. https://www.investopedia.com/terms/e/ema.asp, 2022. Ημερομηνία πρόσβασης: 3-1-2023.

[69] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin and David Lopez-Paz. *mixup: Beyond empirical risk minimization*. *arXiv preprint arXiv:1710.09412*, 2017.

[70] Dong Hyun Lee. *Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks*. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013.

[71] Avi Bewtra. *The Ultimate Guide to Semi-Supervised Learning*. https://www.v7labs.com/blog/semi-supervised-learning-guide, 2022. Ημερομηνία πρόσβασης: 4-1-2023.

[72] Yves Grandvalet and Yoshua Bengio. *Semi-supervised Learning by Entropy Minimization*. *Advances in Neural Information Processing Systems*L. Saul, Y. Weiss and L. Bottou, editors, volume 17. MIT Press, 2004.

[73] Lilian Weng. *Learning with not Enough Data Part 1: Semi-Supervised Learning*. *lilianweng.github.io*, 2021.

[74] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang and Colin Raffel. *ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring*. *International Conference on Learning Representations*, 2020.

[75] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1η έκδοση, 2007.

[76] Massih Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver and Yury Maximov. *Self-training: A survey*. *arXiv preprint arXiv:2202.12040*, 2022.

[77] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe and Youngjoon Yoo. *Cutmix: Regularization strategy to train strong classifiers with localizable features*. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[78] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun and Christoph Bregler. *Efficient object localization using convolutional networks*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[79] Ting Chen, Simon Kornblith, Mohammad Norouzi and Geoffrey Hinton. *A simple framework for contrastive learning of visual representations*. *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[80] Aaron van den Oord, Yazhe Li and Oriol Vinyals. *Representation learning with contrastive predictive coding*. *arXiv preprint arXiv:1807.03748*, 2018.

[81] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng and Yu Xiong Wang. *Pixel Contrastive-Consistent Semi-Supervised Semantic Segmentation*. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7253–7262, 2021.

[82] Shikun Liu, Shuaifeng Zhi, Edward Johns and Andrew J Davison. *Bootstrapping Semantic Segmentation with Regional Contrast*. *International Conference on Learning Representations*, 2022.

[83] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao and Xinyi Le. *Semi-Supervised Semantic Segmentation Using Unreliable Pseudo Labels*. *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[84] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui and Liwei Wang. *Semi-supervised semantic segmentation via adaptive equalization learning*. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021.

[85] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji and Jitendra Malik. *Semantic Contours from Inverse Detectors*. *International Conference on Computer Vision (ICCV)*, 2011.

[86] Ziwei Liu, Ping Luo, Xiaogang Wang and Xiaoou Tang. *Deep Learning Face Attributes in the Wild*. *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[87] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li and Li Fei-Fei. *Imagenet: A large-scale hierarchical image database*. *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[88] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong and Qing He. *A comprehensive survey on transfer learning*. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[89] *Overview of augmentations in ML*. `https://hasty.ai/docs/mp-wiki/augmentations/overview-of-augmentations-in-ml`. Ημερομηνία πρόσβασης: 17-1-2023.

[90] Estevão S Gedraite and Murielle Hadad. *Investigation on the effect of a Gaussian Blur in image filtering and segmentation*. *Proceedings ELMAR-2011*, pages 393–396. IEEE, 2011.

[91] Purnendu Mishra and Kishor Sarawadekar. *Polynomial learning rate policy with warm restart for deep neural network*. *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 2087–2092. IEEE, 2019.

[92] Leland McInnes, John Healy and James Melville. *Umap: Uniform manifold approximation and projection for dimension reduction*. *arXiv preprint arXiv:1802.03426*, 2018.

[93] Sudhanshu Mittal, Maxim Tatarchenko and Thomas Brox. *Semi-Supervised Semantic Segmentation With High- and Low-Level Consistency*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1369–1379, 2021.

[94] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver and Colin A Raffel. *Mixmatch: A holistic approach to semi-supervised learning*. *Advances in neural information processing systems*, 32, 2019.

[95] *torch.utils.data*. https://pytorch.org/docs/stable/data.html#torch.utils.data.Sampler. Ημερομηνία πρόσβασης: 3-2-2023.

[96] *GRNET*. https://hpc.grnet.gr/. Ημερομηνία πρόσβασης: 13-2-2023.

[97] *GRNET Hardware Overview*. https://doc.aris.grnet.gr/system/hardware/. Ημερομηνία πρόσβασης: 13-2-2023.

[98] *Jaccard index*. https://en.wikipedia.org/wiki/Jaccard_index. Ημερομηνία πρόσβασης: 3-2-2023.

[99] *Sørensen-Dice coefficient*. https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient. Ημερομηνία πρόσβασης: 6-2-2023.

[100] *F-score*. https://en.wikipedia.org/wiki/F-score. Ημερομηνία πρόσβασης: 6-2-2023.

[101] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He and Piotr Dollár. *Focal loss for dense object detection*. *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.