ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Επανεκπαίδευση μοντέλων YOLO για ανίχνευση ατόμων σε περιβάλλον Edge AI

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΜΠΟΥΜΠΑΛΟΥ ΝΕΚΤΑΡΙΟΥ

Επιβλέπων : Παναγιώτης Τσανάκας

Καθηγητής ΗΜΜΥ ΕΜΠ

Αθήνα, Οκτώβριος 2025

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Επανεκπαίδευση μοντέλων YOLO για ανίχνευση ατόμων σε περιβάλλον Edge AI

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΜΠΟΥΜΠΑΛΟΥ ΝΕΚΤΑΡΙΟΥ

Επιβλέπων : Παναγιώτης Τσανάκας

Καθηγητής ΗΜΜΥ ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 31η Οκτωβρίου 2025.

.....	.....	.....
Παναγιώτης Τσανάκας ΔΕΠ	Γεώργιος Ματσόπουλος ΔΕΠ	Ανδρέας-Γεώργιος Σταφυλοπάτης ΔΕΠ
Καθηγητής, ΗΜΜΥ ΕΜΠ	Καθηγητής, ΗΜΜΥ ΕΜΠ	Ομότιμος Καθηγητής, ΗΜΜΥ ΕΜΠ

Αθήνα, Οκτώβριος 2025

(Υπογραφή)

.....

Νεκτάριος Μπούμπαλος

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

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στην επανεκπαίδευση μοντέλων YOLO βάσει δύο διακριτών πειραματικών σεναρίων, με στόχο την ανάπτυξη συστήματος εντοπισμού έκνομων επιβατών στην οροφή κινούμενων συρμών σε πραγματικό χρόνο, σε περιβάλλον Edge AI. Τα δύο σενάρια διαφοροποιούνται τόσο ως προς το μέγεθος και την ποικιλία των χρησιμοποιούμενων συνόλων δεδομένων, όσο και ως προς την παραμετροποίηση της διαδικασίας επανεκπαίδευσης, επιτρέποντας τη διερεύνηση της επίδρασής τους στην απόδοση και τη γενικευσιμότητα των μοντέλων σε πραγματικές συνθήκες λειτουργίας. Παράλληλα, πραγματοποιείται η δημιουργία εξειδικευμένου dataset από εναέρια δεδομένα και παρουσιάζεται μεθοδολογία ημιαυτόματης επισημείωσης, η οποία αξιοποιείται για την αποτελεσματική και ταχεία παραγωγή αξιόπιστων δεδομένων εκπαίδευσης.

Λέξεις Κλειδιά: YOLO, επανεκπαίδευση μοντέλων, Edge AI, εντοπισμός ατόμων, σύνολα δεδομένων, ημιαυτόματη επισημείωση, αξιολόγηση μοντέλου

Abstract

This thesis focuses on the retraining of YOLO models based on two distinct experimental scenarios, aiming at the development of a real-time system for detecting unauthorized passengers on the roof of moving trains within an Edge AI environment. The two scenarios differ both in the size and diversity of the datasets used, as well as in the parameterization of the retraining process, allowing for the investigation of their impact on model performance and generalization in real operational conditions. Additionally, a specialized dataset is constructed from aerial video data, and a semi-automatic annotation methodology is presented, which is employed to enable efficient and rapid production of reliable training data.

.

Keywords: YOLO, model retraining, Edge AI, person detection, datasets, semi-automatic annotation, model evaluation

Πίνακας περιεχομένων

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

1

Εισαγωγή

1.1 Παρουσίαση του προβλήματος και εφαρμογών στην καταμέτρηση πλήθους

Η καταμέτρηση πλήθους (crowd counting) αποτελεί μία από τις πιο κρίσιμες και μελετημένες εφαρμογές της υπολογιστικής όρασης και της τεχνητής νοημοσύνης, με ολοένα αυξανόμενη σημασία σε τομείς όπως η δημόσια ασφάλεια, η διαχείριση μαζικών συγκεντρώσεων, η επιτήρηση μεταφορικών μέσων και η εμπορική ανάλυση καταναλωτικής συμπεριφοράς [38]. Η δυνατότητα ακριβούς εκτίμησης του αριθμού ατόμων σε δημόσιους ή ελεγχόμενους χώρους είναι καθοριστικής σημασίας για την πρόληψη επικίνδυνων καταστάσεων, τη βελτιστοποίηση της διαχείρισης πόρων και τη διασφάλιση της ανθρώπινης ζωής.

Παραδοσιακές προσεγγίσεις, βασισμένες σε στατιστικά μοντέλα ή φυσικούς αισθητήρες όπως υπέρυθροι ή θερμικοί, παρουσιάζουν σημαντικούς περιορισμούς, ιδιαίτερα σε περιβάλλοντα με υψηλή οπτική συμφόρηση, επικαλύψεις αντικειμένων ή έντονες μεταβολές φωτισμού. Αντιθέτως, η αξιοποίηση συνελικτικών νευρωνικών δικτύων (Convolutional Neural Networks – CNNs) και αλγορίθμων ταχείας ανίχνευσης, όπως το YOLO, έχει επιφέρει νέα επίπεδα ακρίβειας και αποδοτικότητας, καθιστώντας δυνατή την επεξεργασία και ανάλυση εικόνων και βίντεο σε πραγματικό χρόνο.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι περιπτώσεις εναέριας παρακολούθησης, όπως η αναγνώριση ανθρώπων μέσω drone, όπου συνυπάρχουν στατικά και κινητά αντικείμενα, όπως συρμοί και επιβάτες. Σε τέτοια σενάρια απαιτείται συνδυασμός μοντέλων ανίχνευσης και τμηματοποίησης με υψηλή υπολογιστική ισχύ και δυνατότητα λειτουργίας στο άκρο του συστήματος (edge-computing). Η πλατφόρμα NVIDIA Jetson Orin Nano αποτελεί χαρακτηριστικό παράδειγμα τέτοιας τεχνολογίας, επιτρέποντας την ανάπτυξη συστημάτων βαθιάς μάθησης με υποστήριξη CUDA και εξειδικευμένες βιβλιοθήκες, προσφέροντας παράλληλα χαμηλή κατανάλωση ενέργειας και φορητότητα, ιδανική για κινητά ή απομακρυσμένα περιβάλλοντα.

1.2 Αντικείμενο και στόχοι της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία εξετάζει πτυχές της αξιολόγηση ενός συστήματος καταμέτρησης πλήθους και εντοπισμού έκνομων επιβατών σε εναέρια βιντεοσκόπηση, το οποίο λειτουργεί σε πραγματικό χρόνο και αξιοποιεί σύγχρονες τεχνικές βαθιάς μάθησης. Το σύστημα λαμβάνει ως είσοδο βίντεο από drone που περιλαμβάνει ανθρώπους και συρμούς (τρένα), με στόχο την ακριβή αναγνώριση επιβατών πάνω στα βαγόνια και την ανίχνευση πιθανών έκνομων ενεργειών.

Για τη βελτίωση της ακρίβειας του συστήματος, πραγματοποιείται επανεκπαίδευση (fine-tuning) προ-εκπαιδευμένων μοντέλων με ειδικά επισημασμένα δεδομένα, προσαρμοσμένα σε εναέρια σενάρια παρακολούθησης. Η μεθοδολογία βασίζεται σε ένα επαναληπτικό pipeline που συνδυάζει τμηματοποίηση της επιφάνειας των τρένων με το μοντέλο SAM-L, ανίχνευση ανθρώπων μέσω του μοντέλου YOLOv11x, ημιαυτόματη επισημείωση δεδομένων, μετατροπή σε μορφή COCO, χειροκίνητη επιβεβαίωση μέσω COCO Annotator και, τέλος, επαναληπτική εκπαίδευση του μοντέλου στα νέα δεδομένα. Το τελικό σύστημα είναι σε θέση να εκτελεί δυναμικό segmentation στο τρένο και να εντοπίζει ανθρώπους σε πραγματικό χρόνο, παρέχοντας λειτουργικότητα εντοπισμού παρατυπιών.

Οι κύριοι στόχοι της εργασίας επικεντρώνονται στην εκπαίδευση και αξιολόγηση του μοντέλου YOLOv11 για την αναγνώριση ανθρώπων και τρένων από αεροφωτογραφίες, και στη δημιουργία ενός ημιαυτόματου pipeline επισημείωσης που αξιοποιεί τη συνεργατική χρήση μοντέλων εντοπισμού (detection) και τμηματοποίησης (segmentation). Τέλος, πραγματοποιείται αξιολόγηση σε νέο dataset που περιλαμβάνει διαφορετικές στάσεις και γωνίες λήψης, ώστε να εξεταστεί η ικανότητα γενίκευσης του συστήματος.

1.2.1 Συνεισφορά της εργασίας

Η παρούσα διπλωματική εργασία συνεισφέρει στη δημιουργία και εφαρμογή μιας ολοκληρωμένης διαδικασίας εκπαίδευσης μοντέλων YOLO για την καταμέτρηση πλήθους σε εναέρια δεδομένα, με στόχο τον εντοπισμό έκνομων επιβατών σε οροφές συρμών σε περιβάλλον Edge AI. Επιπλέον, προτείνει και υλοποιεί μια μεθοδολογία ημιαυτόματης επισημείωσης δεδομένων. Παρέχει συγκριτική αξιολόγηση μοντέλων τύπου YOLO τόσο σε ανίχνευση όσο και σε τμηματοποίηση μέσω δύο διαφορετικών πειραματικών σεναρίων, με διαφορετικό όγκο και ποικιλία δεδομένων, καθώς και με διαφορετική παραμετροποίηση στις συναρτήσεις επανεκπαίδευσης, ώστε να αναδειχθούν καλές πρακτικές.

1.3 Οργάνωση κειμένου

Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο που στηρίζει τη μελέτη, περιλαμβάνοντας τις τεχνικές συνελικτικών νευρωνικών δικτύων (CNNs), την αρχιτεκτονική των μοντέλων τύπου YOLO, καθώς και τις θεμελιώδεις έννοιες της επιτάχυνσης μέσω CUDA και της διαδικασίας εκπαίδευσης.

Το Κεφάλαιο 3 αναφέρεται στις σχετικές εργασίες και τη σύγχρονη ερευνητική δραστηριότητα στον χώρο της καταμέτρησης πλήθους, της ανίχνευσης και τμηματοποίησης αντικειμένων, καθώς και στην ανάπτυξη τέτοιων συστημάτων σε πραγματικό χρόνο και σε πλατφόρμες edge computing. Η

ενότητα αυτή συνδέει τη θεωρητική βάση με τις τρέχουσες τάσεις της έρευνας και τις υπάρχουσες προσεγγίσεις στον τομέα.

Στο Κεφάλαιο 4 παρουσιάζεται αναλυτικά η μεθοδολογία και η διαδικασία υλοποίησης του προτεινόμενου συστήματος. Περιγράφεται το σύστημα που επιτρέπει την εκτέλεση σε πραγματικό χρόνο μέσω της πλατφόρμας Jetson Orin Nano, η διαδικασία επισημείωσης (annotation), η επανεκπαίδευση του μοντέλου YOLOv11, καθώς και η αξιολόγηση των μοντέλων που αναπτύχθηκαν, με ανάλυση των αποτελεσμάτων που προέκυψαν από κάθε στάδιο εκπαίδευσης.

Στο Κεφάλαιο 5 παρουσιάζονται τα συνολικά συμπεράσματα της έρευνας. Συνοψίζονται τα βασικά ευρήματα, αξιολογείται η επιτυχία των επιμέρους στόχων και συζητούνται οι δυνατότητες βελτίωσης και μελλοντικής επέκτασης του συστήματος, τόσο σε επίπεδο δεδομένων όσο και σε επίπεδο υπολογιστικής υλοποίησης.

Τέλος, στο Παράρτημα παρατίθενται συμπληρωματικά τεχνικά στοιχεία που υποστηρίζουν την υλοποίηση της εργασίας. Περιλαμβάνεται η περιγραφή της υλικοτεχνικής υποδομής, η διαδικασία εγκατάστασης και ρύθμισης της πλατφόρμας Jetson Orin Nano, καθώς και οι βιβλιοθήκες και τα εργαλεία βαθιάς μάθησης που χρησιμοποιήθηκαν. Επιπλέον, παρουσιάζονται ενδεικτικά τα διαγράμματα και τα confusion matrices που αναλύονται στο Κεφάλαιο 5, ώστε ο αναγνώστης να έχει πλήρη εποπτεία των αποτελεσμάτων και της διαδικασίας αξιολόγησης των μοντέλων.

2

Θεωρητικό υπόβαθρο

Η παρούσα εργασία βασίζεται σε πέντε βασικές θεματικές περιοχές της υπολογιστικής όρασης και της μηχανικής μάθησης. Πρώτον, εξετάζονται οι μέθοδοι ανίχνευσης και καταμέτρησης ανθρώπων σε εικόνες, οι οποίες αποτελούν το θεωρητικό υπόβαθρο για την κατανόηση του προβλήματος αναγνώρισης ανθρώπων σε μεταβαλλόμενα εξωτερικά περιβάλλοντα. Δεύτερον, παρουσιάζεται η αρχιτεκτονική της οικογένειας YOLO, η οποία αποτελεί την κύρια επιλογή για την υλοποίηση του συστήματος ανίχνευσης λόγω της ισορροπίας ανάμεσα στην ακρίβεια και την ταχύτητα. Τρίτον, μελετώνται οι τεχνικές ανάπτυξης και βελτιστοποίησης μοντέλων σε edge συσκευές, καθώς η υλοποίηση απαιτεί λειτουργία σε πραγματικό χρόνο με περιορισμένους υπολογιστικούς πόρους. Τέταρτον, αναλύονται οι στρατηγικές επανεκπαίδευσης (retraining), που επιτρέπουν στο σύστημα να βελτιώνεται και να παραμένει λειτουργικό σε διαφορετικές συνθήκες λήψης. Τέλος, παρουσιάζονται οι μέθοδοι αξιολόγησης μοντέλων ανίχνευσης, οι οποίες χρησιμοποιούνται για την ποσοτική εκτίμηση της απόδοσης και της γενίκευσης του συστήματος.

2.1 Θεωρητικές Προσεγγίσεις στην Καταμέτρηση Πλήθους

Η καταμέτρηση πλήθους αποτελεί σύνθετο πρόβλημα της υπολογιστικής όρασης, καθώς στοχεύει στην εκτίμηση του αριθμού ατόμων σε σκηνές με έντονη ποικιλία κλίμακας, προοπτικής και σύγκλεισης. Η ιστορική της εξέλιξη συνδέεται στενά με την πρόοδο των Συνελικτικών Νευρωνικών Δικτύων (CNNs). Η καθοριστική τομή ήρθε με το AlexNet [1], το οποίο σηματοδότησε την έναρξη της εποχής της βαθιάς μάθησης, μειώνοντας το σφάλμα top-5 στο ImageNet από 26% σε 15%. Ακολούθησαν τα VGGNet [2] και ResNet [3], που εισήγαγαν βαθύτερες και σταθερότερες αρχιτεκτονικές, επιτρέποντας στα CNNs να αποτυπώνουν πολύπλοκα χωρικά και εννοιολογικά μοτίβα. Η ραγδαία αυτή εξέλιξη οδήγησε

στη διαμόρφωση τριών κύριων οικογενειών μεθόδων καταμέτρησης πλήθους: των detection-based, regression-based και density-based προσεγγίσεων. Στο πλαίσιο των detection-based μεθόδων, η εμφάνιση των one-stage ανιχνευτών YOLO [5] και η ανάπτυξη πολυκλιμακικών αρχιτεκτονικών όπως οι Feature Pyramid Networks (FPN) [4] καθόρισαν τη μετάβαση σε real-time συστήματα εντοπισμού και καταμέτρησης με αυξημένη αποδοτικότητα και ακρίβεια.

Στις μεθόδους που βασίζονται στην ανίχνευση αντικειμένων, ο υπολογισμός του πλήθους γίνεται εντοπίζοντας κάθε άτομο ξεχωριστά και στη συνέχεια μετρώντας μόνο τις σωστές ανιχνεύσεις. Οι διπλές ή επικαλυπτόμενες προβλέψεις αφαιρούνται με διαδικασίες όπως το NMS ή το Soft-NMS. Το NMS κρατά μόνο την πιο σίγουρη ανίχνευση για κάθε άτομο και απορρίπτει όσες επικαλύπτονται ή έχουν χαμηλότερη εμπιστοσύνη, ώστε να μην εμφανίζονται πολλές προβλέψεις για το ίδιο αντικείμενο [5, 6]. Πρόσθετη βελτίωση στην ακρίβεια του εντοπισμού επιτυγχάνεται με τη χρήση πιο εξελιγμένων συναρτήσεων απώλειας για τα bounding boxes, όπως οι GIoU, DIoU και CIoU. Σε αντίθεση με τις βασικές συναρτήσεις που αξιολογούν μόνο το ποσοστό επικάλυψης μεταξύ πρόβλεψης και πραγματικού αντικειμένου, αυτές λαμβάνουν υπόψη και τη σχετική απόσταση των πλαισίων, καθώς και τη γεωμετρία και τη μορφή τους μέσα στην εικόνα. Έτσι, το μοντέλο μπορεί να τοποθετεί τα bounding boxes με μεγαλύτερη ακρίβεια και σταθερότητα [7,8]. Η επεξεργασία χαρακτηριστικών σε πολλαπλές κλίμακες μέσω των δομών FPN και PANet βοηθά το μοντέλο να αναγνωρίζει αντικείμενα ανεξάρτητα από το μέγεθός τους ή τη θέση τους μέσα στην εικόνα. Το FPN/PANet είναι δομές που συνδυάζουν χαρακτηριστικά από διαφορετικά επίπεδα ανάλυσης μέσα στο δίκτυο, ώστε το μοντέλο να εντοπίζει αποτελεσματικά τόσο μικρά όσο και μεγάλα αντικείμενα μέσα στην ίδια εικόνα [4,13]. Κύριος περιορισμός παραμένει η υποβάθμιση απόδοσης σε πολύ υψηλές πυκνότητες ή ακραίες παραμορφώσεις προοπτικής.

Οι regression-based προσεγγίσεις στοχεύουν στην απευθείας εκμάθηση της σχέσης μεταξύ των οπτικών χαρακτηριστικών μιας εικόνας και συνεχών μεταβλητών-στόχων, όπως το συνολικό πλήθος ή το κέντρο μάζας ενός πλήθους. Σε αντίθεση με τις detection-based ή segmentation-based μεθόδους, δεν παράγουν ενδιάμεσα bounding boxes ή μάσκες, αλλά υπολογίζουν άμεσα το επιθυμητό μέγεθος Παρότι έτσι χάνεται η ακριβής χωρική πληροφορία (δηλαδή πού βρίσκεται κάθε άτομο), αυτές οι μέθοδοι είναι πιο ανθεκτικές σε περιπτώσεις συνωστισμού ή απόκρυψης, γιατί βασίζονται στη συνολική κατανομή της εικόνας και όχι σε μεμονωμένες ανιχνεύσεις.

Οι density-based προσεγγίσεις επιχειρούν να εκτιμήσουν όχι άμεσα το πλήθος, αλλά έναν χάρτη πυκνότητας (density map), δηλαδή μια εικόνα στην οποία κάθε pixel περιέχει μια τιμή που αντιστοιχεί στην «πιθανότητα» ή την εκτιμώμενη πυκνότητα παρουσίας ατόμων γύρω από εκείνο το σημείο. Αν αθροίσουμε όλες τις τιμές του χάρτη, προκύπτει το συνολικό πλήθος ατόμων στην εικόνα. Για παράδειγμα, αν ένα άτομο αντιπροσωπεύεται από μια κατανομή

Gauss, τότε η κορυφή της κατανομής δηλώνει τη θέση του, ενώ η εξάπλωσή της δείχνει την περιοχή όπου υπάρχει πιθανότητα παρουσίας του. Η ιδέα θεμελιώθηκε στο Learning to Count [10], ενώ μοντέλα όπως MCNN και CSRNet έδειξαν ότι η διατήρηση μεγάλου αντιληπτικού πεδίου μέσω των dilated convolutions βελτιώνει την εκτίμηση σε περιβάλλοντα υψηλής πυκνότητας. Η διευρυμένη συνέλιξη (dilated convolutions) είναι μια παραλλαγή της κλασικής πράξης συνέλιξης που χρησιμοποιείται στα συνελικτικά νευρωνικά δίκτυα (CNNs). Ο στόχος της είναι να επιτρέπει στο δίκτυο να έχει μεγαλύτερο αντιληπτικό πεδίο – δηλαδή να «βλέπει» μεγαλύτερο μέρος της εικόνας – χωρίς να αυξάνει τον αριθμό των παραμέτρων ή το υπολογιστικό κόστος [9, 11]. Παρότι παρουσιάζουν αυξημένη ανθεκτικότητα σε occlusions και αποδίδουν με αξιοπιστία σε περιπτώσεις έντονου συνωστισμού, στερούνται ρητής εντόπισης των ατόμων και συνήθως απαιτούν αυξημένο υπολογιστικό φόρτο, γεγονός που δυσχεραίνει την υλοποίηση σε πραγματικό χρόνο, ειδικά σε edge πλατφόρμες.

Πέρα από τις διαφορετικές τεχνικές προσέγγισης, η καταμέτρηση πλήθους παρουσιάζει και σημαντικές προκλήσεις που σχετίζονται με τη δυσκολία των πραγματικών σκηνών. Οι πιο κρίσιμες αφορούν τη σύγκλειση (occlusion), τη μεγάλη μεταβολή στην κλίμακα των αντικειμένων, τις παραμορφώσεις λόγω προοπτικής και τις διαφορές μεταξύ διαφορετικών datasets (domain shift). Οι Feature Pyramid Networks (FPN) και Path Aggregation Networks (PANet) είναι δομές νευρωνικών δικτύων που συνδυάζουν χαρακτηριστικά από διαφορετικά επίπεδα ανάλυσης, επιτρέποντας στο μοντέλο να αναγνωρίζει με ακρίβεια αντικείμενα διαφόρων μεγεθών και να διαχωρίζει άτομα που επικαλύπτονται μεταξύ τους. Η σύγκλειση αντιμετωπίζεται πιο αποτελεσματικά όταν τα δίκτυα αξιοποιούν πολυκλιμακικές πυραμιδικές αναπαραστάσεις, όπως οι Feature Pyramid Networks (FPN) και οι Path Aggregation Networks (PANet), οι οποίες προσφέρουν πιο πλούσια χωρική πληροφορία και βοηθούν στη διάκριση ατόμων που επικαλύπτονται μεταξύ τους [26,27].

Η μεταβλητότητα στην κλίμακα των αντικειμένων αντιμετωπίζεται με τεχνικές που επιτρέπουν στο μοντέλο να αναλύει την εικόνα σε πολλαπλές κλίμακες και να συνδυάζει πληροφορίες από διαφορετικά επίπεδα ανάλυσης. Ένα παράδειγμα είναι το Spatial Pyramid Pooling (SPP), το οποίο βοηθά το δίκτυο να κρατά λεπτομέρειες για μικρά αντικείμενα χωρίς να αυξάνει σημαντικά το υπολογιστικό κόστος. Οι παραμορφώσεις προοπτικής μειώνονται επειδή το μοντέλο «βλέπει» τα αντικείμενα ταυτόχρονα σε διαφορετικά μεγέθη και θέσεις, ενώ η χρήση τεχνικών ενίσχυσης δεδομένων (augmentations), όπως αλλαγές στη γωνία λήψης ή στην απόσταση, το βοηθά να εξοικειώνεται με διαφορετικές συνθήκες και να προσαρμόζεται πιο αποτελεσματικά.

Τέλος, το domain shift, δηλαδή οι διαφορές στην κατανομή δεδομένων μεταξύ εκπαίδευσης και εφαρμογής, μετριάζονται με στρατηγικές transfer learning και domain adaptation, καθώς και με τεχνικές συνεχούς μάθησης (όπως regularization ή selective rehearsal), που συμβάλλουν

στη σταθεροποίηση των αναπαραστάσεων και στη διατήρηση της απόδοσης σε νέα περιβάλλοντα. Transfer learning σημαίνει ότι χρησιμοποιούμε ένα ήδη εκπαιδευμένο μοντέλο και το προσαρμόζουμε σε ένα νέο πρόβλημα, εξοικονομώντας χρόνο και δεδομένα. Domain adaptation σημαίνει ότι το μοντέλο μαθαίνει να αποδίδει σωστά όταν αλλάζει το περιβάλλον ή η μορφή των δεδομένων, χωρίς να χρειάζεται πλήρη επανεκπαίδευση [20,24].

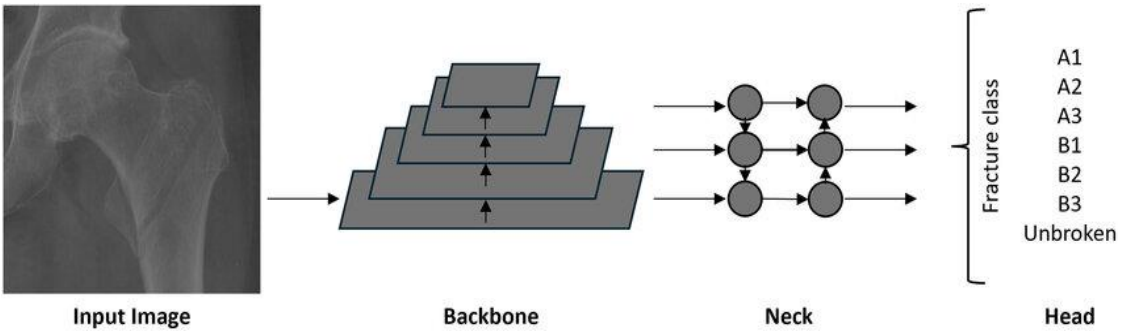
2.2 Ανάλυση της Αρχιτεκτονικής YOLOv11

Η επιλογή της οικογένειας YOLO στην παρούσα εργασία βασίζεται στις πρακτικές απαιτήσεις του συστήματος. Το τελικό μοντέλο καλείται να λειτουργεί σε πραγματικό χρόνο και μάλιστα σε συσκευές με περιορισμένους πόρους, όπως είναι οι edge πλατφόρμες. Σε τέτοιες συνθήκες απαιτείται μια λύση που να εντοπίζει με ακρίβεια τα άτομα στο χώρο, αλλά παράλληλα να διατηρεί υψηλή ταχύτητα και χαμηλή υπολογιστική επιβάρυνση. Τα μοντέλα YOLO έχουν σχεδιαστεί ακριβώς με αυτές τις προτεραιότητες: εντοπίζουν τα αντικείμενα με μία μόνο προώθηση της εικόνας μέσα από το δίκτυο, κάτι που τα καθιστά πολύ πιο γρήγορα σε σχέση με άλλες κατηγορίες ανιχνευτών. Παράλληλα, αξιοποιούν χαρακτηριστικά της εικόνας από διαφορετικά επίπεδα ανάλυσης, γεγονός που τα βοηθά να παραμένουν αποτελεσματικά σε σκηνές με διαφορετικές αποστάσεις και μεγέθη ατόμων, όπως συμβαίνει συνήθως στα μέσα μαζικής μεταφοράς ή σε σταθμούς. Επιπλέον, η συνεχής εξέλιξη της σειράς YOLO έχει οδηγήσει σε πιο σταθερή εκπαίδευση και βελτιωμένη ακρίβεια εντοπισμού, χάρη σε σύγχρονες τεχνικές βελτιστοποίησης που έχουν εισαχθεί στις νεότερες εκδόσεις. Και το σημαντικότερο: η απόδοσή τους παραμένει υψηλή ακόμα και σε φορητές ή χαμηλής ισχύος συσκευές, κάτι που αποτελεί βασική προϋπόθεση για την υλοποίηση και αξιολόγηση του προτεινόμενου συστήματος στο πεδίο[4, 8,12,13].

Ένα ακόμη σημαντικό πλεονέκτημα των μοντέλων YOLO είναι ότι μπορούν να αξιοποιήσουν ελαφριές αρχιτεκτονικές βάσης, όπως το MobileNetV2, ώστε να λειτουργούν χωρίς υπερβολικές απαιτήσεις σε υπολογιστικούς πόρους. Σε συνδυασμό με βιβλιοθήκες επιτάχυνσης που αξιοποιούν την κάρτα γραφικών (όπως οι cuDNN και TensorRT), καθίσταται δυνατή η γρήγορη εκτέλεση του συστήματος ακόμη και σε μικρές ενσωματωμένες πλατφόρμες, όπως το Jetson Orin Nano, που είναι χαρακτηριστικό παράδειγμα συσκευής edge computing. Αντίθετα, οι density-based μέθοδοι, παρότι θεωρούνται εξαιρετικά αποτελεσματικές σε πολύ πυκνά πλήθη, δεν δίνουν άμεση χωρική πληροφορία για τις θέσεις των ατόμων, κάτι που σε πολλές εφαρμογές όπως η δική μας είναι σημαντικό. Παράλληλα, η υψηλή υπολογιστική τους απαίτηση τις καθιστά λιγότερο κατάλληλες για συστήματα που πρέπει να λειτουργούν με χαμηλό latency σε πραγματικές συνθήκες πεδίου. Τέλος, σε ένα edge σύστημα είναι συχνά αναγκαία η προσαρμογή του μοντέλου στις τρέχουσες συνθήκες λειτουργίας, ακόμη και η γρήγορη επανεκπαίδευση πάνω στη συσκευή (on-device). Οι YOLO

ανιχνευτές υποστηρίζουν πιο εύκολα αυτή την ευελιξία, χάρη στον modular σχεδιασμό τους και στη μεγάλη υποστήριξη από εργαλεία βελτιστοποίησης και ανάπτυξης. [19,25,31]. Συνολικά, ο συνδυασμός ταχύτητας, ακρίβειας και προσαρμοστικότητας καθιστά την οικογένεια YOLO την πλέον κατάλληλη επιλογή για το προτεινόμενο σύστημα real-time ανίχνευσης και καταμέτρησης επιβατών σε πλατφόρμες edge. Επί της ουσίας, δε μας νοιάζει η πυκνότητα αλλά η ύπαρξη έκνομων επιβατών, γι' αυτό θέλουμε detection-based method και εξηγήσαμε γιατί το YOLO είναι η καλύτερη επιλογή.

Η YOLOv11 ακολουθεί την τυπική δομή Backbone–Neck–Head, όπως φαίνεται στο Σχήμα 1, ώστε να διατηρεί μια ισορροπία ανάμεσα στην ακρίβεια και την ταχύτητα. Το Backbone είναι το βασικό μέρος του δικτύου που εξάγει τα χαρακτηριστικά από την εικόνα. Εκεί εντοπίζονται τα βασικά μοτίβα, όπως άκρα, σχήματα και υφές, τα οποία σταδιακά συνδυάζονται για να περιγράψουν πιο σύνθετα αντικείμενα. Στο Backbone, η χρήση ειδικών μπλοκ με υπολειμματικές και πιο «πυκνές» συνδέσεις βοηθά το δίκτυο να αξιοποιεί καλύτερα την πληροφορία καθώς αυξάνει το βάθος του. Παράλληλα, η συνάρτηση ενεργοποίησης



Σχήμα 1: Εσωτερική δομή YOLO11

SiLU (Sigmoid Linear Unit) συμβάλλει στη βελτίωση της απόδοσης, επιτρέποντας στο μοντέλο να μαθαίνει πιο εύκολα πολύπλοκα μοτίβα από τα οπτικά δεδομένα. [17].

Το Neck λειτουργεί ως γέφυρα ανάμεσα στο Backbone και στο τελικό στάδιο. Ο ρόλος του είναι να συνδυάζει χαρακτηριστικά από διαφορετικά επίπεδα ανάλυσης (χαμηλά και υψηλά), ώστε το δίκτυο να κατανοεί ταυτόχρονα λεπτομέρειες μικρών αντικειμένων και γενικότερες δομές της εικόνας. Στο Neck, η αρχιτεκτονική αξιοποιεί μια πυραμιδική δομή χαρακτηριστικών, ώστε να συνδυάζει πληροφορία από διαφορετικά επίπεδα ανάλυσης. Πιο συγκεκριμένα, το Feature Pyramid Network συμβάλλει στη μεταφορά πιο σημασιολογικού περιεχομένου από τα βαθύτερα στρώματα προς τα πιο ρηχά, ενώ το Path Aggregation Network επιτρέπει την αντίστροφη ροή λεπτομερειών χαμηλού επιπέδου προς τα ανώτερα στάδια. Με αυτόν τον τρόπο, η YOLOv11 μπορεί να εντοπίζει αποτελεσματικά και μικρότερα αντικείμενα, αλλά και να διαχειρίζεται καλύτερα τις αλλαγές κλίμακας μέσα στη σκηνή [4, 13]. Η μονάδα

SPPF (Spatial Pyramid Pooling – Fast) διευρύνει το «πεδίο όρασης» του μοντέλου, επιτρέποντάς του να συλλαμβάνει καλύτερα τα συνολικά μοτίβα της εικόνας, χωρίς όμως να αυξάνει σημαντικά το υπολογιστικό κόστος. Με αυτόν τον τρόπο ενισχύεται η κατανόηση της σκηνής, ακόμη και όταν υπάρχουν μεγάλα αντικείμενα ή πιο σύνθετες χωρικές σχέσεις [12].

Το Head είναι το τελικό τμήμα του δικτύου που παράγει τις προβλέψεις — δηλαδή τα πλαίσια (bounding boxes), τις κατηγορίες των αντικειμένων και τα ποσοστά εμπιστοσύνης. Εκεί γίνεται η τελική ανίχνευση και ταξινόμηση. Η διαδικασία αυτή μπορεί να βασιστεί είτε σε anchors είτε να είναι anchor-free, ανάλογα με τη διαμόρφωση. Στα anchor-based μοντέλα, η ανίχνευση βασίζεται σε προκαθορισμένα πλαίσια (anchors) διαφορετικών μεγεθών και αναλογιών, τα οποία το δίκτυο προσαρμόζει στα αντικείμενα. Είναι σταθερή αλλά πιο περίπλοκη μέθοδος. Αντίθετα, τα anchor-free μοντέλα προβλέπουν απευθείας το κέντρο ή τα όρια κάθε αντικειμένου, χωρίς τη χρήση anchors. Έτσι γίνονται πιο απλά, γρήγορα και ευέλικτα, ειδικά σε σκηνές με μεγάλη πυκνότητα ή διαφορετικά σχήματα αντικειμένων. Με τον τρόπο αυτό, η YOLOv11 μπορεί να ανιχνεύει άτομα διαφορετικών μεγεθών, αξιοποιώντας τρία βασικά επίπεδα χαρακτηριστικών (π.χ. με βήμα 8, 16 και 32 pixels), ώστε να ανταποκρίνεται τόσο σε κοντινά όσο και σε πιο απομακρυσμένα σημεία της σκηνής [17, 18]. Η εκδοχή YOLO-Seg προσθέτει branch τμηματοποίησης για λεπτομερή χωρική κατανόηση σε σκηνές συνωστισμού.

Η συνάρτηση απώλειας συνδυάζει όρους γεωμετρίας πλαισίων, ύπαρξης αντικειμένου και ταξινόμησης. Οι μετρικές GIoU, DIoU και CIoU συνυπολογίζουν επικάλυψη, απόσταση κέντρων και αναλογία πλευρών, βελτιώνοντας τη σταθερότητα εκπαίδευσης [7, 8]. Μετά την πρόβλεψη εφαρμόζεται καταστολή επικάλυψης· η παραλλαγή Soft-NMS μειώνει βαθμωτά τις βαθμολογίες αντί για δυαδική απόρριψη, ωφελώντας σε σκηνές με πυκνά πλαίσια [6]. Η πολυκλιμακική σύνθεση χαρακτηριστικών με FPN/PANet και η προσαρμογή των thresholds επιτρέπουν την επιθυμητή ισορροπία precision–recall σε πραγματικές ροές βίντεο [4, 12, 13, 18].

2.3 Edge AI, Jetson Orin Nano και Βελτιστοποίηση με TensorRT

Η υλοποίηση συστημάτων καταμέτρησης πλήθους στο edge αποκτά ιδιαίτερη σημασία σε εφαρμογές όπου η καθυστέρηση, το εύρος ζώνης και η ασφάλεια δεδομένων αποτελούν κρίσιμους παράγοντες. Σε περιπτώσεις όπως η παρακολούθηση επιβατών σε μέσα μεταφοράς ή η διαχείριση πλήθους σε απομακρυσμένες υποδομές, η αποστολή δεδομένων σε απομακρυσμένο cloud αυξάνει την latency και επιβαρύνει τη μετάδοση. Το edge Computing επιτρέπει την εκτέλεση της ανίχνευσης και καταμέτρησης τοπικά, μειώνοντας σημαντικά την

καθυστέρηση και τις απαιτήσεις εύρους ζώνης, ενώ ενισχύει την ιδιωτικότητα και την επιχειρησιακή αυτονομία του συστήματος [19,25]. Για το crowd counting, αυτό σημαίνει ότι οι αποφάσεις λαμβάνονται σε πραγματικό χρόνο στο σημείο λήψης, βελτιώνοντας την απόκριση και καθιστώντας εφικτή τη συνεχή παρακολούθηση ακόμη και σε περιβάλλοντα περιορισμένων πόρων. Αντί ολόκληρες εικόνες ή ροές βίντεο να αποστέλλονται σε απομακρμένους διακομιστές για ανάλυση, η επεξεργασία πραγματοποιείται τοπικά στη συσκευή λήψης (π.χ. Jetson Orin Nano), και προς το κεντρικό δίκτυο αποστέλλεται μόνο ένα συμπιεσμένο αποτέλεσμα — όπως μια ειδοποίηση, αποφορτίζοντας το εκάστοτε δίκτυο.

Το edge Computing φέρνει το inference κοντά στην πηγή των δεδομένων, μειώνοντας καθυστέρηση και απαιτήσεις εύρους ζώνης. Οι μετρικές απόδοσης είναι ο χρόνος αιτιολόγησης, ο ρυθμός καρέ, το αποτύπωμα μνήμης και η κατανάλωση ισχύος [19]. Η πλατφόρμα NVIDIA Jetson Orin Nano ενσωματώνει πολυπύρηνη CPU ARM και GPU αρχιτεκτονικής Ampere με Tensor Cores, προσφέροντας υψηλή υπολογιστική ισχύ σε μικρό μέγεθος και χαμηλή κατανάλωση. Το οικοσύστημα JetPack, που περιλαμβάνει βιβλιοθήκες όπως CUDA, cuDNN και TensorRT, παρέχει το απαραίτητο λογισμικό περιβάλλον για την ανάπτυξη και επιτάχυνση εφαρμογών υπολογιστικής όρασης στο άκρο (edge) με βελτιστοποιημένη απόδοση[15, 16].

Η cuDNN είναι μια βιβλιοθήκη της NVIDIA που κάνει πιο γρήγορες τις βασικές λειτουργίες των νευρωνικών δικτύων, όπως οι συνελίξεις και η κανονικοποίηση, χρησιμοποιώντας βελτιστοποιημένο κώδικα για τις GPU. Το TensorRT βελτιώνει ακόμη περισσότερο την απόδοση του μοντέλου πριν από την εκτέλεση, συνδυάζοντας ορισμένα βήματα (layer fusion), απλοποιώντας υπολογισμούς (constant folding) και ρυθμίζοντας αυτόματα τον τρόπο εκτέλεσης (kernel auto-tuning). [14]. Έτσι προκύπτει engine προσαρμοσμένο στη συγκεκριμένη GPU, με σημαντικά χαμηλότερο latency και υψηλότερο FPS σε σύγκριση με τυπική εκτέλεση PyTorch. Η ροή PyTorch → ONNX → TensorRT → Jetson επιτρέπει την ενσωμάτωση μοντέλων YOLO σε embedded εφαρμογές πραγματικού χρόνου [17, 18].

2.4 Θεωρία Επανεκπαίδευσης και Συνεχούς Προσαρμογής

Η επανεκπαίδευση (retraining) είναι η διαδικασία με την οποία ένα ήδη εκπαιδευμένο μοντέλο ενημερώνεται με νέα δεδομένα, ώστε να διατηρεί ή να βελτιώνει την απόδοσή του όταν αλλάζουν οι συνθήκες του περιβάλλοντος. Αντί να εκπαιδεύεται από την αρχή, το μοντέλο «ξεκινά» από τη γνώση που έχει ήδη αποκτήσει και προσαρμόζει τα βάρη του με βάση τα πιο πρόσφατα δείγματα. Με αυτόν τον τρόπο εξοικονομούνται χρόνος και υπολογιστικοί πόροι, ενώ παράλληλα το σύστημα παραμένει επίκαιρο και ικανό να λειτουργεί σε νέες καταστάσεις. Σε πιο απλές μορφές, η επανεκπαίδευση γίνεται περιοδικά: συλλέγονται νέα δεδομένα,

επισημαίνονται και χρησιμοποιούνται για fine-tuning των υπαρχόντων βαρών του δικτύου. Σε πιο προχωρημένες υλοποιήσεις, η διαδικασία είναι συνεχής και βασίζεται σε μικρά, αντιπροσωπευτικά υποσύνολα δεδομένων που ενημερώνουν το μοντέλο σταδιακά, με εκπαιδευτικούς βρόχους χαμηλού κόστους. Ωστόσο, η επανεκπαίδευση ενέχει και τον κίνδυνο υπερπροσαρμογής (overfitting), ιδιαίτερα όταν τα νέα δεδομένα είναι περιορισμένα ή μη αντιπροσωπευτικά του συνολικού φάσματος συνθηκών. Σε αυτή την περίπτωση, το μοντέλο προσαρμόζεται υπερβολικά στα συγκεκριμένα δείγματα του retraining και χάνει την ικανότητά του να γενικεύει σε νέες ή ελαφρώς διαφορετικές εισόδους. Η υπερπροσαρμογή μπορεί να οδηγήσει σε πολύ υψηλά αποτελέσματα στο σύνολο εκπαίδευσης, αλλά σε σημαντική μείωση της απόδοσης κατά την πραγματική λειτουργία του συστήματος. Για τον λόγο αυτό, είναι απαραίτητο ο σχεδιασμός της επανεκπαίδευσης να περιλαμβάνει κατάλληλα μέτρα τακτοποίησης (regularization), επαρκή ποικιλία δεδομένων και ανεξάρτητη αξιολόγηση, ώστε να διατηρείται η ισορροπία ανάμεσα στη μάθηση νέων πληροφοριών και στη γενίκευση της ήδη αποκτημένης γνώσης.

Όταν τα νέα δεδομένα διαφέρουν σημαντικά από τα αρχικά, το μοντέλο μπορεί να «ξεχάσει» τη γνώση που είχε αποκτήσει στο παρελθόν — ένα φαινόμενο γνωστό ως καταστροφική λήθη (catastrophic forgetting). σενάρια

Για να αποφευχθεί η μείωση της απόδοσης του μοντέλου σε παλιότερα σενάρια, έχουν αναπτυχθεί, επίσης, τεχνικές που βοηθούν το μοντέλο να διατηρεί τη σημαντική γνώση που έχει αποκτήσει. Μία από τις πιο γνωστές είναι το Elastic Weight Consolidation (EWC), το οποίο περιορίζει την αλλαγή σε ορισμένα «κρίσιμα» βάρη του δικτύου κατά την επανεκπαίδευση, ώστε να προστατεύεται ό,τι έχει μάθει ήδη [22]. Συμπληρωματικά, υπάρχουν τεχνικές που βοηθούν το μοντέλο να θυμάται όσα έχει μάθει, δίνοντάς του ξανά μερικά από τα παλαιά παραδείγματα κατά την επανεκπαίδευση. Αυτές οι μέθοδοι συχνά βασίζονται στη χρήση μικρών δειγμάτων (exemplar buffers), τα οποία περιέχουν αντιπροσωπευτικά στιγμιότυπα από προηγούμενες σκηνές. Με τον τρόπο αυτό, το μοντέλο εξασκείται ταυτόχρονα σε παλιές και νέες καταστάσεις, μειώνοντας σημαντικά τον κίνδυνο να ξεχάσει την ήδη αποκτημένη γνώση [23].

Η συνεχής μάθηση (continual learning) είναι ένα παράδειγμα εκπαίδευσης τεχνητών νευρωνικών δικτύων όπου το μοντέλο ενημερώνεται σταδιακά με νέα δεδομένα ή εργασίες, χωρίς να απαιτείται εκπαίδευση από την αρχή. Στόχος της είναι να επιτρέψει στο σύστημα να προσαρμόζεται σε μεταβαλλόμενα περιβάλλοντα και να διατηρεί τη γνώση που έχει ήδη αποκτήσει, αποφεύγοντας το φαινόμενο της καταστροφικής λήθης, δηλαδή της απώλειας προηγούμενων πληροφοριών κατά την εκμάθηση νέων. Αναλυτικές επισκοπήσεις του πεδίου παρουσιάζουν το φάσμα στρατηγικών continual learning και τα αντίστοιχα πλεονεκτήματα/μειονεκτήματα [24]. Σε πλατφόρμες του άκρου, η επανεκπαίδευση πρέπει να

σχεδιάζεται με γνώμονα τον χρόνο απόκρισης, την ενέργεια και τη μνήμη, με μικρά παράθυρα προσαρμογής και ρητή στόχευση των κλάσεων που επιδεινώνονται [25].

Σε εφαρμογές όρασης πεδίου, το στατιστικό προφίλ των δεδομένων μεταβάλλεται με τον χρόνο. Η αλλαγή εποχών, οι διαφοροποιήσεις φωτισμού, οι νέες γωνίες λήψης και η εμφάνιση άγνωστων σκηνών δημιουργούν απόκλιση μεταξύ του κατανομικού χώρου εκπαίδευσης και εκείνου της ανάπτυξης, ένα φαινόμενο που έχει μελετηθεί συστηματικά στη βιβλιογραφία για dataset shift [20]. Η μετατόπιση μεταξύ διαφορετικών σετ δεδομένων μπορεί να εμφανιστεί με πολλούς τρόπους, όπως αλλαγές στη μορφή της εικόνας, στις ετικέτες ή ακόμη και στον τρόπο που ορίζονται οι ίδιες οι έννοιες που πρέπει να μάθει το μοντέλο. Στην πράξη, τέτοιες αλλαγές επηρεάζουν πιο έντονα τις «δύσκολες» κατηγορίες, όπως ανθρώπους που εμφανίζονται πολύ μακριά ή μερικώς κρυμμένους. Για τον λόγο αυτό χρειάζονται ειδικές στρατηγικές αντιμετώπισης, ώστε το μοντέλο να μπορεί να προσαρμόζεται ομαλά σε νέα περιβάλλοντα χωρίς να χάνει την ακρίβειά του [21].

Τα epochs είναι καθοριστικός παράγοντας για την ποιότητα ενός μοντέλου: πολύ λίγα οδηγούν σε underfitting, πάρα πολλά αυξάνουν τον κίνδυνο overfitting και μπορούν να μετακινήσουν το λειτουργικό σημείο στις καμπύλες precision–recall (άρα και δείκτες όπως mAP, Best-F1, Max Recall). Γι’ αυτό, η επιλογή τους δεν είναι τυπική ρύθμιση αλλά στρατηγική απόφαση που γίνεται με σάρωση epochs και early stopping βάσει validation, ώστε το μοντέλο να σταθεροποιηθεί εκεί όπου μεγιστοποιεί τη γενίκευση στο δικό μας πρόβλημα[32].

2.5 Αξιολόγηση μοντέλου

Η αξιολόγηση του μοντέλου πραγματοποιείται σε ανεξάρτητο σύνολο δεδομένων, το οποίο δεν χρησιμοποιείται κατά την εκπαίδευση, ώστε να διαπιστωθεί η ικανότητά του να γενικεύει σωστά σε νέες και άγνωστες σκηνές. Στην ανίχνευση αντικειμένων, οι προβλέψεις του συστήματος συσχετίζονται με τις πραγματικές θέσεις των ατόμων μέσω του δείκτη επικάλυψης (IoU). Αν η επικάλυψη υπερβεί ένα προκαθορισμένο κατώφλι, η ανίχνευση θεωρείται επιτυχής. Από αυτή τη σύγκριση υπολογίζονται βασικές μετρικές απόδοσης.

Η ακρίβεια (precision) και η ανάκληση (recall) είναι δύο βασικές μετρικές αξιολόγησης της απόδοσης ενός συστήματος ανίχνευσης. Η ακρίβεια (precision) υπολογίζεται διαιρώντας τον αριθμό των σωστών ανιχνεύσεων (True Positives) με το άθροισμα των σωστών και των λανθασμένων ανιχνεύσεων (True Positives + False Positives). Δείχνει, δηλαδή, ποιο ποσοστό από όλα όσα αναγνώρισε το μοντέλο αντιστοιχούν πραγματικά σε σωστά αντικείμενα. Η ανάκληση (recall) προκύπτει αν διαιρέσουμε τον αριθμό των σωστών ανιχνεύσεων (True Positives) με το άθροισμα των σωστών και των παραλειπόμενων ανιχνεύσεων (True Positives + False Negatives). Μετρά, επομένως, πόσα από τα πραγματικά αντικείμενα κατάφερε να εντοπίσει το σύστημα. Το F1-score συνδυάζει τις δύο αυτές μετρικές μέσω του αρμονικού τους

μέσου, έτσι ώστε να αποτυπώνει σε έναν αριθμό τη συνολική ισορροπία ανάμεσα στην ακρίβεια και την ανάκληση. Ένα υψηλό F1 δείχνει ότι το σύστημα πετυχαίνει ταυτόχρονα λίγες ψευδείς ανιχνεύσεις και ελάχιστες παραλείψεις.

Καθώς μεταβάλλεται το κατώφλι εμπιστοσύνης του μοντέλου, προκύπτουν διαφορετικές τιμές ακρίβειας και ανάκλησης, οι οποίες σχηματίζουν μια καμπύλη precision–recall. Το Average Precision (AP) αντιστοιχεί στο εμβαδό κάτω από αυτή την καμπύλη και δείχνει πόσο καλά το μοντέλο ισορροπεί ανάμεσα στην ακρίβεια και την ανάκληση για μία συγκεκριμένη κατηγορία.Για να εκτιμηθεί η συνολική απόδοση του μοντέλου σε όλες τις κατηγορίες και σε διαφορετικά επίπεδα δυσκολίας, υπολογίζεται ο μέσος όρος των επιμέρους AP, γνωστός ως mean Average Precision (mAP). Ο δείκτης αυτός υπολογίζεται για διάφορα ποσοστά επικάλυψης μεταξύ των προβλέψεων και των πραγματικών αντικειμένων, συνήθως από 0.50 έως 0.95. Με τον τρόπο αυτό, το mAP αποτυπώνει συνολικά πόσο αξιόπιστα το μοντέλο εντοπίζει και αναγνωρίζει αντικείμενα τόσο σε απλές όσο και σε πιο σύνθετες περιπτώσεις.

Παράλληλα, η μελέτη των καμπυλών precision–confidence και recall–confidence επιτρέπει την επιλογή του βέλτιστου λειτουργικού σημείου ανάλογα με τις ανάγκες της εφαρμογής. Για παράδειγμα, αν προτεραιότητα είναι η ανίχνευση όλων των ατόμων, τότε επιλέγεται υψηλή ανάκληση ακόμη και εις βάρος της ακρίβειας. Αντίθετα, όταν είναι σημαντικό να αποφεύγονται ψευδείς ανιχνεύσεις, προτιμάται υψηλότερη ακρίβεια. Συμπληρωματικά, ο πίνακας σφαλμάτων ανά κλάση προσφέρει μια σαφή εικόνα των συστηματικών συγχύσεων του μοντέλου, διευκολύνοντας τον εντοπισμό συγκεκριμένων αδυναμιών και την ιεράρχηση των απαιτούμενων βελτιώσεων.

3

Σχετικές εργασίες

3.1 Βιβλιογραφική Επισκόπηση της Καταμέτρησης Πλήθους

Η καταμέτρηση πλήθους αποτελεί ένα από τα πιο μελετημένα προβλήματα στην υπολογιστική όραση[38], καθώς συνδέεται με εφαρμογές δημόσιας ασφάλειας, διαχείρισης μεταφορών, παρακολούθησης εκδηλώσεων και βιομηχανικής αυτοματοποίησης. Η πρόκληση έγκειται όχι μόνο στην ποικιλία των σκηνών (εσωτερικοί/εξωτερικοί χώροι, μεταβαλλόμενες συνθήκες φωτισμού), αλλά και στα φαινόμενα σύγκλεισης (occlusion), απότομης αλλαγής κλίμακας και παραμόρφωσης προοπτικής, που μεταβάλλουν δραματικά το οπτικό αποτύπωμα των ατόμων μέσα στο κάδρο.

Οι πρώτες προσεγγίσεις στην καταμέτρηση πλήθους στηρίχθηκαν σε παραδοσιακές τεχνικές υπολογιστικής όρασης, όπου η ανάλυση της εικόνας γινόταν με χειροποίητα χαρακτηριστικά. Περιγραφείς όπως οι HOG [36] και SIFT [37] χρησιμοποιούνταν για την απόσπαση πληροφορίας από την εικόνα και στη συνέχεια τροφοδοτούσαν κλασικούς παλινδρομητές. Σε ορισμένα συστήματα, η καταμέτρηση επιχειρούνταν έμμεσα, μέσω ανίχνευσης κεφαλιών ή ώμων, προσπαθώντας να εντοπιστούν τα άτομα με βάση συγκεκριμένα μορφολογικά στοιχεία.

Παρόλα αυτά, η μεγάλη ποικιλία στην εμφάνιση των ανθρώπων, οι αλλαγές κλίμακας, καθώς και οι ιδιαίτερα πυκνές σκηνές, περιόριζαν σημαντικά την ακρίβεια και δυσκόλευαν την προσαρμογή των μεθόδων αυτών σε νέα δεδομένα. Η μετάβαση στα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) σηματοδότησε μια ριζική αλλαγή στο πεδίο. Τα βαθιά μοντέλα κατάφεραν να μάθουν πιο ισχυρές και ανεξάρτητες χαρακτηριστικές αναπαραστάσεις, εμφανίζοντας μεγαλύτερη ανθεκτικότητα σε αλλαγές φωτισμού, οπτικού θορύβου και προοπτικής. Έτσι, έγινε δυνατή η πιο αξιόπιστη εκτίμηση ατόμων ακόμη και σε σκηνές με έντονο συνωστισμό.

Η ανασκόπηση των Sindagi και Patel [26] χαρτογράφησε συστηματικά αυτή τη μετάβαση, παρουσιάζοντας την εξέλιξη από κλασικές μεθόδους βασισμένες σε χειροποίητα χαρακτηριστικά προς αρχιτεκτονικές εκτίμησης πυκνότητας (density estimation) με CNNs. Οι

συγγραφείς έδειξαν ότι η χρήση πολυκλιμακικών χαρακτηριστικών και εξειδικευμένων κεφαλών παλινδρόμησης βελτίωσε αισθητά τα αποτελέσματα σε σκηνές με υψηλό συνωστισμό, αντιμετωπίζοντας πιο αποτελεσματικά ζητήματα όπως η σύγκλειση και οι παραμορφώσεις προοπτικής. Τόνισαν επίσης τον ρόλο των μετρικών αξιολόγησης, όπως οι MAE και MSE, καθώς και την ανάγκη για ποικιλία συνόλων δεδομένων, από πιο «αραιά» (UCSD) έως εξαιρετικά πυκνά (ShanghaiTech, UCF-QNRF), που ανέδειξαν τα όρια των παλαιότερων τεχνικών και συνέβαλαν στη διαμόρφωση των σύγχρονων προσεγγίσεων.

Το multi-column CNN (MCNN) είναι ένα νευρωνικό δίκτυο με πολλές παράλληλες "στήλες", καθεμία με διαφορετικό μέγεθος φίλτρων, ώστε να αναγνωρίζει άτομα σε διαφορετικές κλίμακες και αποστάσεις μέσα στην εικόνα.

Το switch-CNN είναι ένα δίκτυο που διαθέτει πολλαπλές στήλες όπως το MCNN, αλλά χρησιμοποιεί έναν μηχανισμό "διακόπτη" που επιλέγει δυναμικά ποια στήλη είναι καταλληλότερη για κάθε περιοχή, βελτιώνοντας έτσι την ακρίβεια και την αποδοτικότητα.

Συγκεκριμένα, έδειξαν ότι τα multi-column CNNs (MCNN) αξιοποιούν διαφορετικά μεγέθη φίλτρων για να διαχειρίζονται σκηνές με ποικίλες πυκνότητες πλήθους, αλλά είναι βαριά σε υπολογισμούς και δυσκολεύονται να προσαρμοστούν σε πολύπλοκες προοπτικές. Αντίθετα, τα switch-CNNs χρησιμοποιούν μηχανισμό επιλογής του καταλληλότερου υποδικτύου ανάλογα με την τοπική πυκνότητα, επιτυγχάνοντας καλύτερη ακρίβεια με μικρότερο σφάλμα MAE. Οι context-aware και pyramid-based προσεγγίσεις, όπως το CP-CNN, υπερέχουν ακόμη περισσότερο σε σκηνές με πολύ πυκνά πλήθη, επειδή ενσωματώνουν πληροφορίες από τα συμφραζόμενα της εικόνας και προσαρμόζουν δυναμικά την εκτίμηση. Μέσα από αυτές τις συγκρίσεις, οι Sindagi και Patel κατέληξαν ότι η πρόοδος δεν εξαρτάται μόνο από την πολυπλοκότητα των δικτύων, αλλά από το πώς τα μοντέλα μαθαίνουν να προσαρμόζονται σε διαφορετικά είδη πλήθους και προοπτικών. Παράλληλα, τόνισαν ότι η αξιολόγηση των μεθόδων επηρεάζεται σημαντικά από τη διαφορετικότητα των συνόλων δεδομένων και την έλλειψη κοινών προτύπων αξιολόγησης, κάτι που καθιστά δύσκολη τη δίκαιη σύγκριση μεταξύ ερευνών.

Η εξέλιξη των CNN-based μεθόδων έχει βελτιώσει θεαματικά την ακρίβεια στην εκτίμηση πλήθους, ιδιαίτερα σε σκηνές υψηλής πυκνότητας όπου οι παραδοσιακές τεχνικές αποτύγχαναν λόγω σύγκλεισης και παραμορφώσεων προοπτικής. Η επιτυχία των πιο σύγχρονων μοντέλων, όπως τα Switch-CNN, CP-CNN και Cascaded-MTL, καταδεικνύει ότι η βελτίωση της απόδοσης δεν προκύπτει απλώς από την αύξηση του βάθους των δικτύων, αλλά κυρίως από την ικανότητά τους να προσαρμόζονται στην τοπική πυκνότητα και στα συμφραζόμενα της σκηνής (context). Παρά τη σημαντική πρόοδο, οι Sindagi και Patel επισημαίνουν ότι το πεδίο εξακολουθεί να αντιμετωπίζει βασικές προκλήσεις, όπως η περιορισμένη δυνατότητα γενίκευσης των μοντέλων σε νέα περιβάλλοντα, η έλλειψη μεγάλων και ποιοτικών συνόλων

δεδομένων με αξιόπιστες επισημειώσεις (annotations), καθώς και η απουσία ενοποιημένων πρωτοκόλλων αξιολόγησης (benchmarking), τα οποία δυσχεραίνουν τη δίκαιη και άμεση σύγκριση μεταξύ διαφορετικών προσεγγίσεων.

Σε νεότερη ανασκόπηση, οι Gao και συνεργάτες [27] σε συγκριτική ανάλυση των νεότερων τάσεων, εξετάζοντας πώς οι παραδοσιακές CNN-based μέθοδοι διαφέρουν από τις πιο σύγχρονες προσεγγίσεις που βασίζονται σε μηχανισμούς προσοχής (attention) και δομές τύπου transformer. Οι attention μηχανισμοί επιτρέπουν στο δίκτυο να εστιάζει επιλεκτικά σε κρίσιμες περιοχές της εικόνας, μειώνοντας το σφάλμα σε περιπτώσεις σύγκλεισης ή ανομοιόμορφης πυκνότητας, ενώ τα transformer blocks ενισχύουν την κατανόηση της χωρικής συσχέτισης μεταξύ απομακρυσμένων σημείων. Παράλληλα, συγκρίνουν τις lightweight και edge-friendly αρχιτεκτονικές με τα κλασικά "βαριά" μοντέλα, δείχνοντας ότι η υπολογιστική αποδοτικότητα μπορεί να επιτευχθεί χωρίς σημαντική απώλεια ακρίβειας μέσω βελτιστοποιημένων backbone δικτύων. Ιδιαίτερο ενδιαφέρον παρουσιάζει η ανάλυσή τους στις υβριδικές προσεγγίσεις, που συνδυάζουν τεχνικές ανίχνευσης (detection) και εκτίμησης πυκνότητας (density estimation), αξιοποιώντας τα πλεονεκτήματα και των δύο μεθόδων ανάλογα με τη φύση της σκηνής.

Οι συγγραφείς καταλήγουν ότι η πρόοδος στο πεδίο της καταμέτρησης πλήθους εξαρτάται πλέον περισσότερο από την προσαρμοστικότητα και τη γενίκευση των μοντέλων παρά από την απλή αύξηση του μεγέθους ή της πολυπλοκότητάς τους. Η έμφαση μετατοπίζεται στην ανάπτυξη μοντέλων που μπορούν να λειτουργούν αξιόπιστα σε πραγματικές συνθήκες, με διαφορετικό φωτισμό, προοπτική ή βαθμό συνωστισμού. Οι Gao et al. επισημαίνουν ως βασική κατεύθυνση τη δημιουργία ελαφριών, ενεργειακά αποδοτικών δικτύων για εφαρμογές πραγματικού χρόνου σε edge πλατφόρμες, αλλά και την ενσωμάτωση τεχνικών domain adaptation και συνεχούς μάθησης (continual learning) ώστε τα μοντέλα να προσαρμόζονται αυτόματα σε νέα δεδομένα. Συνολικά, συμπεραίνουν ότι το μέλλον της έρευνας στο crowd counting θα καθοριστεί από την ικανότητα εξισορρόπησης ακρίβειας, αποδοτικότητας και προσαρμοστικότητας, καθιστώντας τις τεχνολογίες αυτές ώριμες για πραγματικές, επιχειρησιακές εφαρμογές.

Και οι δύο παραπάνω έρευνες έδειξαν τη σημαντικότητα της προσαρμογής του μοντέλου σε νέα περιβάλλοντα και την ανάγκη για νέο dataset, κάτι με το οποίο θα ασχοληθούμε εκτενώς.

3.2 Μέθοδοι Εκτίμησης Πυκνότητας (Density Estimation)

Η εκτίμηση πυκνότητας εισήγαγε μια νέα οπτική στην καταμέτρηση πλήθους: αντί της ρητής ανίχνευσης κάθε ατόμου, το μοντέλο μαθαίνει να παράγει έναν χάρτη πυκνότητας (density map) που αποτυπώνει τη χωρική κατανομή παρουσιών, με το ολοκλήρωμα του χάρτη να

ισοδυναμεί με το πλήθος. Η προσέγγιση αυτή είναι ιδιαίτερα αποτελεσματική σε σκηνές όπου η ανίχνευση αποτυγχάνει λόγω ακραίας σύγκλεισης ή μικροσκοπικής κλίμακας κεφαλών.

Η πρώτη καθοριστική συμβολή ήρθε από την εργασία των Zhang et al. [9], με το MCNN (Multi-Column CNN). Η αρχιτεκτονική χρησιμοποιεί τρεις παράλληλες στήλες με διαφορετικά μεγέθη πυρήνων, ώστε να «αισθάνεται» αντικείμενα σε ποικίλες κλίμακες. Η πολυκλιμακική αυτή σχεδίαση μετριάζει την εξάρτηση από συγκεκριμένες προοπτικές, βελτιώνει τη γενίκευση σε ποικιλία σκηνών και καθιστά το δίκτυο πιο ανθεκτικό σε απότομες αλλαγές μεγέθους/απόστασης. Στην πράξη, το MCNN σηματοδότησε τη μετάβαση από απλοϊκές backbones σε κλιμακωτές/παράλληλες ροές χαρακτηριστικών, οι οποίες έγιναν πρότυπο για μεταγενέστερες δουλειές.

Ακολούθως, οι Li et al. [11] παρουσίασαν το CSRNet, ένα μοντέλο που στηρίζεται στη χρήση dilated convolutions, δηλαδή συνελίξεων με διάτρηση. Με αυτή την τεχνική, το δίκτυο μπορεί να «κοιτάζει» μεγαλύτερο τμήμα της εικόνας χωρίς να χρειάζεται να μειώσει την ανάλυση μέσω υποδειγματοληψίας. Έτσι διατηρείται η λεπτομέρεια στις περιοχές όπου τα άτομα είναι πολύ κοντά μεταξύ τους, ενώ παράλληλα αξιοποιείται το ευρύτερο περιβάλλον της σκηνής για πιο αξιόπιστη εκτίμηση του πλήθους. Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική σε υπερπυκνά πλήθη, όπου η πληροφορία του άμεσου περιγράμματος ενός ατόμου δεν είναι αρκετή και απαιτούνται συμφραζόμενα από μεγαλύτερη περιοχή. Χάρη σε αυτή τη σχεδίαση, το CSRNet παρουσίασε πολύ υψηλές επιδόσεις σε γνωστά σύνολα δεδομένων, όπως το ShanghaiTech και το UCF-QNRF, και αποτέλεσε σημαντικό βήμα προόδου για το πεδίο. Η εργασία έδειξε ότι το «μεσαίο» τμήμα του δικτύου, όπου εξάγονται και επεξεργάζονται τα χαρακτηριστικά, μπορεί να καθορίσει σε μεγάλο βαθμό την τελική απόδοση, και όχι μόνο η κεφαλή που παράγει τον χάρτη πυκνότητας.

Το μοντέλο SegCrowdNet συνδυάζει τεχνικές εντοπισμού περιοχών και εκτίμησης πυκνότητας για πιο ακριβή καταμέτρηση πλήθους. Αρχικά, το δίκτυο εντοπίζει τις περιοχές όπου υπάρχουν κεφάλια ανθρώπων και απομακρύνει το φόντο, ώστε να επικεντρώνεται μόνο στα σχετικά σημεία. Στη συνέχεια, χρησιμοποιεί έναν μηχανισμό προσοχής (attention) για να δίνει μεγαλύτερη βαρύτητα στις πιο σημαντικές περιοχές της εικόνας. Παράλληλα, αντλεί πληροφορίες από διαφορετικές κλίμακες ώστε να ανταποκρίνεται σε αλλαγές μεγέθους ή απόστασης των ατόμων. Σύμφωνα με τα πειραματικά αποτελέσματα, το SegCrowdNet αποδίδει καλύτερα από προηγούμενα μοντέλα σε διάφορα απαιτητικά σύνολα δεδομένων[33].

Τα τρία δείχνουν πώς εξελίχθηκε η εκτίμηση πλήθους με τη χρήση CNNs. Το MCNN ήταν το πρώτο που χρησιμοποίησε πολλές παράλληλες «στήλες» για να αναγνωρίζει άτομα σε διαφορετικές αποστάσεις και μεγέθη. Το CSRNet βελτίωσε αυτή την ιδέα, επιτρέποντας στο δίκτυο να βλέπει μεγαλύτερο μέρος της εικόνας χωρίς να χάνει λεπτομέρεια, κάτι που το έκανε πιο ακριβές σε πολύ πυκνά πλήθη. Το SegCrowdNet έκανε ένα ακόμη βήμα, συνδυάζοντας

ανίχνευση και εκτίμηση πυκνότητας και χρησιμοποιώντας μηχανισμούς προσοχής για να εστιάζει στα πιο σημαντικά σημεία της εικόνας. Συνολικά, η εξέλιξη από το MCNN στο SegCrowdNet δείχνει τη μετάβαση από πιο απλά μοντέλα σε πιο «έξυπνα» και προσαρμοστικά δίκτυα που κατανοούν καλύτερα τη σκηνή.

Μετά τα ορόσημα MCNN/CSRNet, η βιβλιογραφία εστίασε σε: (α) attention μηχανισμούς για προσαρμοστική εστίαση σε πυκνές περιοχές, (β) transformer-based blocks για μοντελοποίηση μακρινών εξαρτήσεων, (γ) scale-aware/scale-adaptive κεφαλές που προβλέπουν πολλαπλούς χάρτες ή ενσωματώνουν δυναμική επιλογή κλίμακας, και (δ) αντιληπτικές απώλειες (perceptual/SSIM) που βελτιώνουν τη δομή του παραγόμενου density map. Παράλληλα, εφαρμόζονται στρατηγικές data augmentation και perspective normalization με χάρτες προοπτικής, ώστε να αντισταθμιστούν οι γεωμετρικές παραμορφώσεις.

Επειδή η εφαρμογή μας απαιτεί ρητή χωρική εντόπιση και αναγνώριση των επιβατών στο τρένο, είναι προφανές ότι οι density-based μέθοδοι δεν είναι κατάλληλες, καθώς δίνουν μόνο συνολικό πλήθος χωρίς να εντοπίζουν ξεχωριστά άτομα. Επιπλέον, απαιτούν υψηλό υπολογιστικό κόστος και μεγάλη μνήμη, γεγονός που τις καθιστά μη πρακτικές για χρήση σε edge συστήματα, όπου οι πόροι είναι περιορισμένοι και η απόκριση πρέπει να είναι άμεση.

3.3 Αλγόριθμοι Ανίχνευσης Αντικειμένων (YOLO & MobileNet)

Οι ανιχνευτές αντικειμένων αποτελούν τον δεύτερο μεγάλο πυλώνα στην καταμέτρηση πλήθους, ιδίως σε σκηνές με χαμηλή έως μέση πυκνότητα όπου τα άτομα διαχωρίζονται ευκρινώς. Ο Redmon και οι συνεργάτες του [5] παρουσίασαν το 2016 το YOLO (You Only Look Once), μια μέθοδο που άλλαξε τον τρόπο ανίχνευσης αντικειμένων. Αντί να ψάχνει πρώτα πιθανές περιοχές και μετά να τις ελέγχει ξεχωριστά, όπως έκαναν οι παλαιότερες προσεγγίσεις, το YOLO χωρίζει την εικόνα σε ένα πλέγμα και εκπαιδεύει ένα ενιαίο νευρωνικό δίκτυο να προβλέπει απευθείας πού βρίσκονται τα αντικείμενα και τι είναι το καθένα. Με αυτόν τον τρόπο, η ανίχνευση γίνεται σε ένα μόνο βήμα, πολύ πιο γρήγορα, χωρίς να χρειάζονται ξεχωριστά στάδια επεξεργασίας. Αυτή η ενιαία διαδικασία το κάνει πολύ γρήγορο και κατάλληλο για εφαρμογές πραγματικού χρόνου, όπως η παρακολούθηση, η ρομποτική και τα βιομηχανικά συστήματα.

Οι πιο πρόσφατες εκδόσεις του YOLO έχουν βελτιωθεί πολύ, ώστε να δίνουν πιο ακριβείς και σταθερές ανιχνεύσεις. Πλέον, τα μοντέλα αυτά δεν χρειάζονται προκαθορισμένα πλαίσια για να εντοπίσουν αντικείμενα, κάτι που κάνει την εκπαίδευση πιο απλή και αξιόπιστη.

Χρησιμοποιούν επίσης μηχανισμούς προσοχής, που τα βοηθούν να εστιάζουν στα πιο σημαντικά σημεία της εικόνας, ειδικά όταν υπάρχουν πολλά ή μικρά αντικείμενα. Οι βελτιώσεις στη δομή τους τούς επιτρέπουν να αναγνωρίζουν αντικείμενα σε διαφορετικά μεγέθη και αποστάσεις, ακόμη και σε δύσκολες σκηνές με πλήθος. Επιπλέον, χάρη σε πιο έξυπνους τρόπους εκπαίδευσης και επεξεργασίας των δεδομένων, τα νεότερα μοντέλα, όπως τα YOLOv8 και YOLOv11, καταφέρνουν να είναι γρήγορα και ακριβή ταυτόχρονα. Τέλος, εφαρμόζονται τεχνικές που αφαιρούν τα διπλά ή λανθασμένα αποτελέσματα, ώστε η τελική ανίχνευση να είναι όσο το δυνατόν πιο καθαρή και αξιόπιστη.

Για περιβάλλοντα περιορισμένων πόρων, όπως edge συσκευές, έχουν προταθεί ελαφριές αρχιτεκτονικές. Το MobileNetV2 των Sandler et al. [28] αναπτύχθηκε για να κάνει τα νευρωνικά δίκτυα πιο ελαφριά και γρήγορα, ώστε να μπορούν να λειτουργούν σε φορητές συσκευές ή πλατφόρμες με περιορισμένους πόρους. Η αρχιτεκτονική του βασίζεται σε έναν πιο αποδοτικό τρόπο υπολογισμού, όπου κάθε φίλτρο εφαρμόζεται ξεχωριστά σε κάθε «κανάλι» της εικόνας, μειώνοντας έτσι δραστικά τον αριθμό των πράξεων και τις απαιτήσεις μνήμης χωρίς σημαντική απώλεια ακρίβειας. Επιπλέον, το δίκτυο έχει τη δυνατότητα να προσαρμόζει το μέγεθος και τη λεπτομέρεια της επεξεργασίας του, ώστε να ταιριάζει στις δυνατότητες της εκάστοτε συσκευής. Στην πράξη, η χρήση του MobileNetV2 ως backbone σε συνδυασμό με ανιχνευτές τύπου YOLO έχει γίνει πολύ συχνή επιλογή, γιατί προσφέρει υψηλή ταχύτητα, χαμηλή κατανάλωση και σταθερή λειτουργία σε συστήματα που πρέπει να επεξεργάζονται βίντεο σε πραγματικό χρόνο στο πεδίο (edge).

Πρόσφατα, τα foundation μοντέλα όπως το Segment Anything Model (SAM) του Kirillov et al. [29], έχουν ανοίξει νέους δρόμους για semi-supervised annotation και prompt-based segmentation. Χωρίς να περιορίζονται σε προκαθορισμένες κατηγορίες, μπορούν να παράγουν μάσκες υψηλής ποιότητας με ελάχιστη ανθρώπινη παρέμβαση, επιταχύνοντας θεαματικά τον κύκλο σεναριοποίησης/επισημείωσης. Στην περίπτωση της καταμέτρησης πλήθους, το SAM μπορεί να χρησιμοποιηθεί ως βοηθητικό εργαλείο που επιταχύνει την παραγωγή ή την επικαιροποίηση των ετικετών για τα άτομα, κάτι ιδιαίτερα χρήσιμο σε εφαρμογές όπου το περιβάλλον αλλάζει συχνά, όπως σε σταθμούς ή δημόσιους χώρους με εποχιακή κίνηση και διαφορετικές γωνίες λήψης. Έτσι, διευκολύνεται ο κύκλος βελτίωσης των ανιχνευτών τύπου YOLO, καθώς μπορούμε να ανανεώνουμε τα δεδομένα με λιγότερο κόστος και μεγαλύτερη ταχύτητα. Εμείς στο πλαίσιο της διπλωματικής χρησιμοποιούμε το SAM μοντέλο στη διαδικασία της αυτόματης επισημείωσης, όπως θα αναλυθεί περαιτέρω στο κεφάλαιο 4.2.

Παρόλο που το SAM είναι ιδιαίτερα χρήσιμο για επισημείωση και υποβοήθηση μοντέλων, δεν είναι πρακτικό για λειτουργία σε πραγματικό χρόνο σε edge συσκευές. Ο βασικός λόγος είναι ότι έχει σχεδιαστεί ως μεγάλο, γενικού σκοπού μοντέλο, με εκατοντάδες εκατομμύρια παραμέτρους, έτσι ώστε να ανταποκρίνεται σε κάθε είδους αντικείμενο και σκηνή. Αυτό το

κάνει πολύ ακριβό υπολογιστικά: χρειάζεται ισχυρή GPU, αρκετή μνήμη και υψηλή κατανάλωση ενέργειας. Σε μικρές συσκευές, όπως αυτές που χρησιμοποιούνται σε κάμερες ασφαλείας ή σε έξυπνους αισθητήρες, αυτό δεν μπορεί να υποστηριχθεί, ειδικά αν απαιτείται συνεχής ανάλυση video. Γι'αυτό εμείς στο edge χρησιμοποιούμε το μοντέλο τμηματοποίησης του YOLO.

3.4 Ενσωματωμένα Συστήματα AI και Βελτιστοποίηση (Edge AI & Jetson)

Η υλοποίηση συστημάτων βαθιάς μάθησης σε ενσωματωμένες συσκευές είναι κρίσιμο βήμα για τη μεταφορά της ΤΝ από το cloud στο edge, όπου οι απαιτήσεις για χαμηλή καθυστέρηση, ιδιωτικότητα δεδομένων και λειτουργία εκτός σύνδεσης γίνονται καθοριστικές. Οι πλατφόρμες NVIDIA Jetson έχουν καθιερωθεί ως κορυφαίες λύσεις για εφαρμογές όρασης σε πραγματικό χρόνο, χάρη στον συνδυασμό ARM CPUs και GPU πυρήνων/Tensor Cores, αλλά και τη διαθεσιμότητα οικοσυστήματος λογισμικού (CUDA, cuDNN, TensorRT, DeepStream).

Η χρήση του TensorRT βοηθά να τρέχουν τα μοντέλα πιο γρήγορα και πιο αποδοτικά. Το εργαλείο αυτό βελτιστοποιεί το δίκτυο κατά την εκτέλεση, συνδυάζοντας ορισμένα βήματα μεταξύ τους (layer fusion), ρυθμίζοντας αυτόματα τον τρόπο που γίνονται οι υπολογισμοί (kernel auto-tuning) και μειώνοντας το μέγεθος των αριθμών που χρησιμοποιεί (quantization, δηλαδή από υψηλή ακρίβεια FP32 σε FP16 ή INT8). Με αυτόν τον τρόπο μειώνεται ο χρόνος εκτέλεσης και αυξάνεται ο ρυθμός καρέ, συνήθως χωρίς να επηρεάζεται αισθητά η ακρίβεια των αποτελεσμάτων [30]. Η συνηθισμένη διαδικασία περιλαμβάνει τρία βασικά βήματα. Πρώτα, το μοντέλο που έχει εκπαιδευτεί σε PyTorch μετατρέπεται σε μορφή ONNX, ώστε να μπορεί να χρησιμοποιηθεί από άλλα εργαλεία. Στη συνέχεια, το μοντέλο αυτό βελτιστοποιείται και μετατρέπεται σε μια TensorRT engine, η οποία είναι η πιο γρήγορη εκδοχή του για εκτέλεση, με δυνατότητα προσαρμογής σε διαφορετικά μεγέθη εισόδου. Τέλος, το βελτιστοποιημένο μοντέλο εγκαθίσταται σε συσκευές Jetson, που χρησιμοποιούν ειδικούς επιταχυντές υλικού (όπως τους DLA) για ακόμα μεγαλύτερη ταχύτητα.

Στο πλαίσιο αυτό, οι τεκμηριώσεις της NVIDIA [30] συνιστούν ουσιώδη οδηγό για σχεδιασμό end-to-end συστημάτων edge AI: από τη μετατροπή μοντέλων και την επιλογή κατάλληλων backbones μέχρι τις βέλτιστες πρακτικές εκτέλεσης με χαμηλή καθυστέρηση (π.χ. batching/streaming, zero-copy pipelines, pinning μνήμης, και χρήση DeepStream για πολυκάναλες ροές). Πέραν της βελτιστοποίησης inference, η ανάγκη για συνεχή προσαρμογή των μοντέλων σε μεταβαλλόμενα περιβάλλοντα έχει οδηγήσει στην ανάπτυξη πλαισίων όπως το Ekya των Bhardwaj et al. [31]. Το Ekya είναι ένα σύστημα continuous learning σχεδιασμένο για εφαρμογές video analytics στο edge, δηλαδή σε συσκευές με περιορισμένους πόρους όπως

τα Jetson Nano, Xavier ή Orin. Η βασική του ιδέα είναι να ανανεώνει δυναμικά τα μοντέλα μηχανικής μάθησης χωρίς να διακόπτει τη λειτουργία τους. Για να το πετύχει αυτό, το Ekya επιλέγει αυτόματα τα πιο αντιπροσωπευτικά δείγματα από τις ροές βίντεο που λαμβάνει, χρησιμοποιώντας έξυπνους αλγορίθμους επιλογής δεδομένων (data sampling). Στη συνέχεια, εκτελεί στοχευμένο retraining του μοντέλου, δηλαδή μικρές επαναεκπαιδεύσεις μόνο στα τμήματα όπου η απόδοση έχει μειωθεί, ενώ το υπόλοιπο σύστημα συνεχίζει κανονικά το inference (δηλαδή την ανάλυση βίντεο σε πραγματικό χρόνο). Στη δική μας περίπτωση, δεν υλοποιείται τέτοιος μηχανισμός συνεχούς μάθησης· η επανεκπαίδευση των μοντέλων πραγματοποιείται εκτός της edge συσκευής, σε ξεχωριστό περιβάλλον, και τα ενημερωμένα βάρη μεταφέρονται εκ των υστέρων στο σύστημα για χρήση.

Συνολικά, η σύζευξη βέλτιστων μοντέλων (π.χ. YOLO με ελαφρύ backbone ή density-based αρχιτεκτονικές με προσεκτική διάταξη dilations) και συστημικής βελτιστοποίησης (TensorRT, σωστές επιλογές precision, αποδοτικές ροές εισόδου/εξόδου) καθιστά εφικτή την αξιόπιστη καταμέτρηση πλήθους σε πραγματικό χρόνο στο edge. Η ενσωμάτωση πλαισίων continuous learning όπως το Ekya [31] προσθέτει τη διάσταση της διαρκούς προσαρμογής, εξασφαλίζοντας ότι τα συστήματα παραμένουν εύρωστα και ακριβή σε βάθος χρόνου, παρά την αναπόφευκτη μετατόπιση κατανομής (dataset shift) που εμφανίζεται σε πραγματικά περιβάλλοντα.

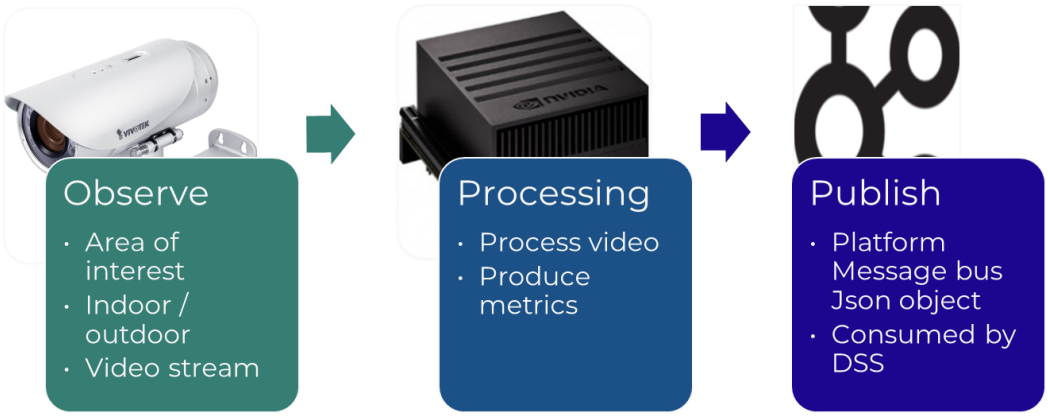
3.5 Σύστημα αναγνώρισης και καταμέτρησης παράνομων επιβατών σε τρένα (ODYSSEUS Project)

Ως μέρος του ερευνητικού έργου ODYSSEUS[1] Horizon, αναπτύχθηκε και υλοποιήθηκε ένα σύστημα με στόχο τη δημιουργία ενός μηχανισμού επιτήρησης των συνοριακών φυλακίων καθώς και ανίχνευσης παρουσίας ατόμων σε μη εξουσιοδοτημένες περιοχές, αξιοποιώντας τεχνολογίες μηχανικής όρασης και τεχνητής νοημοσύνης. Ο σκοπός του συστήματος είναι η επαύξηση της επίγνωσης της κατάστασης στην περιοχή των συνοριακών ελέγχων. Τρία είναι τα βασικά στάδια του συστήματος (Σχήμα 2):

- Παρατήρηση: Συσκευές επιτήρησης , όπως κλειστό κύκλωμα τηλεόρασης, δικτυακές κάμερες ή κάμερες τοποθετημένες σε μη επανδρωμένα εναέρια οχήματα (drones). Οι συσκευές αυτές καλύπτουν τις περιοχές ενδιαφέροντος και τροφοδοτούν το επόμενο βήμα με ροή βίντεο.

[1] : https://odysseusproject.eu/

- Επεξεργασία: Σύστημα επεξεργασίας των ροών βίντεο από το πρώτο βήμα. Στο βήμα γίνεται η ανίχνευση και καταγραφή δεικτών που αφορούν την εκάστοτε περιοχή επιτήρησης (παρουσία μη ατόμων σε περιοχή απαγόρευσης, καταμέτρηση πλήθους κλπ.)
- Αποστολή ευρημάτων: Το τελευταία βήμα είναι η αποστολή των ευρημάτων σε ένα κεντρικό σύστημα αποφάσεων. Ανάλογα με τον τύπο του ευρήματος, η αποστολή μπορεί να είναι με τη μορφή ενός μηνύματος ειδοποίησης ή μηνύματος συναγερμού.



Σχήμα 2:Βασικά στάδια συστήματος επιτήρησης (ODYSSEUS project)

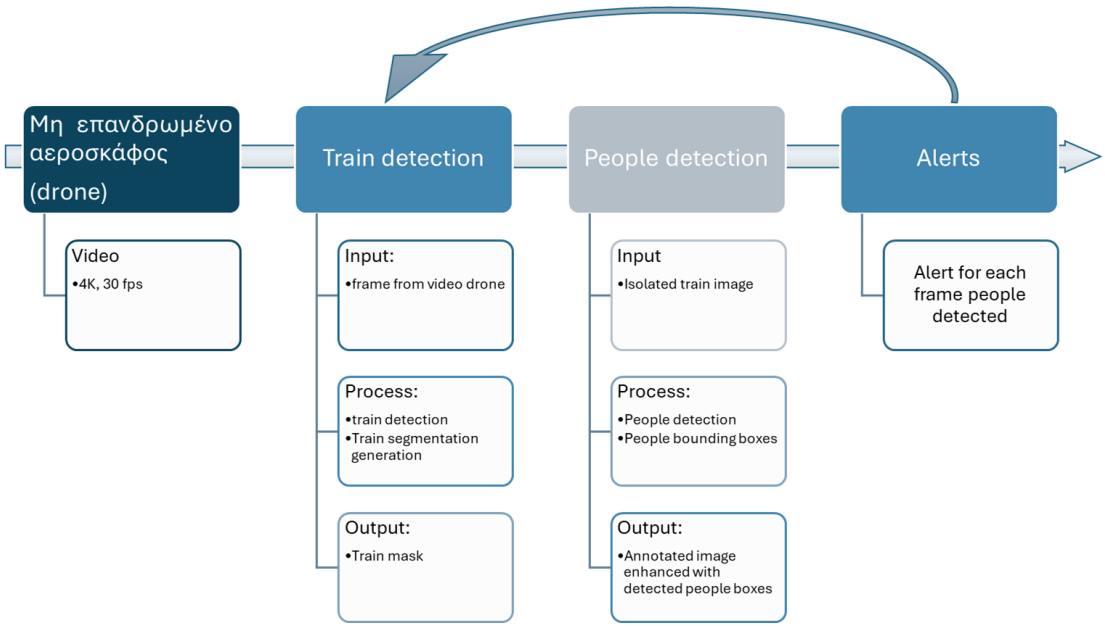
Μέρος του συστήματος αποτελεί η μεθοδολογία εντοπισμού και καταγραφής μη νόμιμων επιβατών κρυμμένων στη οροφή των τρένων. Αν και η βασική αρχιτεκτονική του συστήματος υποστηρίζει οποιονδήποτε τύπο καμερών, η συγκεκριμένη μεθοδολογία είναι αφιερωμένη στη χρήση ροών βίντεο που προέρχονται από τις κάμερες οι οποίες είναι εγκατεστημένες σε μη επανδρωμένα εναέρια οχήματα.

Η πρόκληση σε αυτό το σενάριο είναι ότι η περιοχή ενδιαφέροντος δεν είναι προκαθορισμένη μέσα στο οπτικό πεδίο της κάμερας, αλλά μεταβάλλεται δυναμικά καθώς τόσο η κάμερα όσο και το τρένο (το αντικείμενο ενδιαφέροντος) βρίσκονται σε κίνηση με την πάροδο του χρόνου. Για τον λόγο αυτό, η περιοχής ενδιαφέροντος γίνεται δυναμικά με τη χρήση μοντέλων τμηματοποίησης. Γενικά, ακολουθείται μία προσέγγιση τεσσάρων (4) σταδίων όπως φαίνεται και στην Σχήμα 3:

1. Επιτήρηση χώρου. Στο στάδιο αυτό έχουμε την επιτήρηση της περιοχής με κάμερα, με τη χρήση μη επανδρωμένου εναέριου οχήματος, το οποίο παράγει ροή βίντεο και την αποστέλλει σε πραγματικό χρόνο στο επόμενο στάδιο.
2. Ανίχνευση τρένου: Σε αυτό το στάδιο, χρησιμοποιώντας ένα μοντέλο τμηματοποίησης (segmentation model), εντοπίζονται τα τρένα και τα περιγράμματά τους σε κάθε καρέ του βίντεο. Το μοντέλο τμηματοποίησης που χρησιμοποιείται, βασίζεται στη σειρά YOLO11. Ως τελικό βήμα του σταδίου αυτού, πραγματοποιείται η δημιουργία της

μάσκας του τρένου και τελικά η απομόνωση του τρένου από το υπόλοιπο καρέ. Το στάδιο αυτό είναι ιδιαίτερα κρίσιμο για την ακρίβεια της ανίχνευσης ανθρώπων, καθώς αποτρέπει την εμφάνιση πολλών ψευδώς θετικών εντοπισμών — ειδικά στις περιπτώσεις όπου άνθρωποι που βρίσκονται κοντά στο τρένο θα μπορούσαν εσφαλμένα να χαρακτηριστούν ως επιβάτες πάνω σε αυτό. Ο κρίσιμος παράγοντας στο στάδιο αυτό είναι η ποιότητα του μοντέλου τμηματοποίησης. Υψηλό επίπεδο ακρίβειας στην τμηματοποίηση απομονώνει την περιοχή του τρένου με αντίστοιχη ακρίβεια.

3. Ανίχνευση ανθρώπων πάνω στο Τρένο: Σε αυτό το στάδιο, το απομονωμένο τμήμα του καρέ που αντιστοιχεί στο τρένο υποβάλλεται σε περαιτέρω ανάλυση με στόχο τον εντοπισμό ανθρώπων εντός των ορίων του τρένου. Η διαφορά με το προηγούμενο στάδιο είναι ότι εδώ χρησιμοποιείται ένα απλό μοντέλο εντοπισμού αντικειμένων και όχι μοντέλο τμηματοποίησης. Εδώ δεν μας ενδιαφέρει ο εντοπισμός της ακριβής θέσης του περιγράμματος του αντικειμένου (άνθρωπος) αλλά ο ακριβής εντοπισμός της παρουσίας. Το αποτέλεσμα αυτού του σταδίου είναι ο εντοπισμός της παρουσίας ατόμων εντός του πλαισίου του τρένου.
4. Δημιουργία ειδοποιήσεων συναγερμού. Στο στάδιο αυτό, αν εντοπιστεί ποαρουσία ατόμων τότε αποστέλλεται ειδοποίηση συναγερμού στο κεντρικό σύσταμα αποφάσεων. Το στάδιο αυτό είναι εκτός ενδιαφέροντος της παρούσης διπλωματικής.



Σχήμα 3: Μεθοδολογία εντοπισμού ανθρώπων στην οροφή του τρένου

3.5.1 Συσχέτιση με τη διπλωματική

Στα πλαίσια της διπλωματικής εργασίας, υιοθετήσαμε τη μεθοδολογία ανίχνευσης και εντοπισμού ατόμων στην οροφή του τρένου, που αναπτύχθηκε για το έργο ODYSSEUS. Η διπλωματική εργασία επικεντρώνεται κυρίως στα στάδια όπου με χρήση μηχανικής μάθησης γίνεται αρχικά ο εντοπισμός του περιγράμματος του τρένου (στάδιο στην Σχήμα 3) και στη συνέχεια ο εντοπισμός των ανθρώπων (στάδιο 3 στην Σχήμα 3).

4

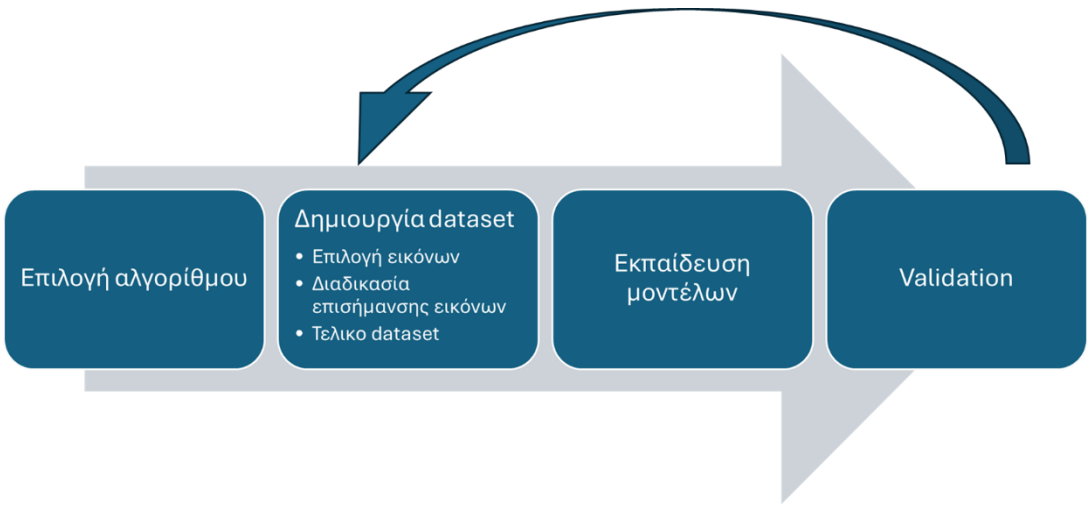
Υλοποίηση και αξιολόγηση

4.1 Μεθοδολογία

Όπως ειπώθηκε και στην παράγραφο 3.5.1, το σύστημα ανίχνευσης ατόμων στην οροφή του τρένου που χρησιμοποιήθηκε στα πλαίσια της διπλωματικής, στηρίχθηκε στο αντίστοιχο του έργου ODYSSEUS. Στο παρόν κεφάλαιο θα αναπτύξουμε τη μεθοδολογία εκπαίδευσης και αξιολόγησης των μοντέλων μηχανικής μάθησης που χρησιμοποιήθηκαν για:

1) Τον εντοπισμό του τρένου και του περιγράμματος του τρένου
2) Τον εντοπισμό ατόμων εντός των ορίων του περιγράμματος του τρένου

Όπως περιγράφεται στην Σχήμα 4 το πρώτο στάδιο μεθοδολογίας αποτελείται από την επιλογή αλγορίθμου για τον εντοπισμό των κλάσεων. Για τις ανάγκες του συστήματος επιλέχθηκε ο YOLOv11.



Σχήμα 4: Μεθοδολογία εκπαίδευσης και αξιολόγησης των μοντέλων

Ο αλγόριθμος YOLO παρέχει τη δυνατότητα εύχρηστης επανεκπαίδευσης (retraining) και αξιοποίησης μέσω των συναρτήσεων της βιβλιοθήκης Ultralytics. Επιπλέον, προσφέρει ιδιαίτερα αποδοτικά μοντέλα τόσο για την ανίχνευση όσο και για την τμηματοποίηση αντικειμένων. Αναλυτική εξήγηση της αρχιτεκτονικής και των πλεονεκτημάτων του παρατίθεται στην παράγραφο 2.2.

Το δεύτερο στάδιο, μετά την επιλογή αλγορίθμου, αφορά τη δημιουργία dataset, όπως φαίνεται στην εικόνα 4. Η διαδικασία δημιουργίας του dataset βασίστηκε σε λήψη βίντεο από μη επανδρωμένο αεροσκάφος (drone), το οποίο αποτυπώνει την άφιξη ενός τρένου στο σημείο ελέγχου καθώς και τη διαδικασία ελέγχου του τρένου. Για τις ανάγκες του έργου ODYSSEUS ένα άτομο ανέβηκε στην οροφή του τρένου προσομοιώνοντας έναν παράνομο ταξιδιώτη. Από το βίντεο επιλέχθηκαν 10 τμήματα των 5 δευτερολέπτων – κυρίως αυτά που αποτυπώνουν τον παράνομο ταξιδιώτη. Το κάθε ένα από αυτά τα τμήματα απαρτίζουν τα 10 επιμέρους dataset. Επιπρόσθετα δημιουργήθηκε ένα άλλο επιμέρους dataset των 292 εικόνων λαμβάνοντας ένα καρέ από τα βίντεο κάθε 5 δευτερόλεπτα Στη συνέχεια, εφαρμόστηκε η διαδικασία επισημείωσης. Η αναλυτική περιγραφή της διαδικασίας αυτής παρουσιάζεται στην παράγραφο 4.2.

Το επόμενο στάδιο της εικόνας 4 είναι αυτό της εκπαίδευσης των μοντέλων. Η εκπαίδευση πραγματοποιήθηκε σε δύο διακριτά μοντέλα σύμφωνα με το στόχο του υποσυστήματος. Το πρώτο μοντέλο αφορά το τον εντοπισμό του τρένου και του περιγράμματος αυτού (segmentation) και επανεκπαιδεύτηκε (retrain) για την κλάση *train*, με στόχο την ακριβή τμηματοποίηση της περιοχής ενδιαφέροντος, δηλαδή του ίδιου του συρμού, ώστε να οριοθετείται με ακρίβεια το πεδίο στο οποίο πραγματοποιείται η ανίχνευση. Το δεύτερο αφορά το μοντέλο εντοπισμού ατόμων (detection) και επανεκπαιδεύτηκε για την κλάση *person*. Η διαδικασία εκπαίδευσης υλοποιήθηκε μέσω του περιβάλλοντος Ultralytics YOLO. Η παραμετροποίηση της διαδικασίας παρατίθεται αναλυτικά στην παράγραφο 4.3.

Στο τελικό στάδιο της εικόνας 4 είναι η αξιολόγηση των μοντέλων που παράχθηκαν από την επανεκπαίδευση του αλγορίθμου. Μέσω διαφόρων μετρικών όπως mAP@0.5 και F1, αξιολογείται ποσοτικά η ικανότητα του μοντέλου να εντοπίζει και να τμηματοποιεί νέα δεδομένα. Σε περίπτωση που δεν είναι ικανοποιητικά τα αποτελέσματα επαναλαμβάνεται η διαδικασία από το στάδιο 2, όπου μπορεί να χρειαστεί να δημιουργηθεί επιπρόσθετο dataset και να γίνει επανεκπαίδευση. Η αξιολόγηση γίνεται στην παράγραφο 4.5.

4.2 Διαδικασία δημιουργίας Dataset

Η δημιουργία του dataset αποτελεί ένα κρίσιμο τμήμα της εκπαίδευσης ενός μοντέλου. Η επιλογή ενός καλού dataset επηρεάζει άμεσα την ποιότητα του εκπαιδευμένου μοντέλου. Ένα σημαντικό μέρος της δημιουργίας ενός dataset είναι η επισήμανση των δεδομένων αυτού. Υπάρχουν ιστότοποι[2,3] όπου μπορούν να βρεθούν dataset για διάφορες εφαρμογές μηχανικής μάθησης. Στην περίπτωση του προβλήματος της διπλωματικής μας δεν εντοπίστηκε σχετικό dataset οπότε δημιουργήσαμε από βίντεο που παρήχθη στα πλαίσια του έργου ODYSSEUS.

Για την εξαγωγή του dataset, σχεδιάστηκε μια διαδικασία annotation, όπως εμφανίζεται και στην Σχήμα 5, που βασίζεται στη λογική της επαναληπτικής μάθησης με ανθρώπινη εποπτεία (human-in-the-loop), έτσι ώστε να μη δημιουργηθεί το dataset εξ ολοκλήρου χειροκίνητα. Η μεθοδολογία αυτή στοχεύει στην προοδευτική ενίσχυση του μοντέλου μέσω επαναλαμβανόμενων κύκλων εκπαίδευσης, όπου το μοντέλο διορθώνει και βελτιώνει τη συμπεριφορά του με βάση τις ανθρώπινες παρεμβάσεις στις επισημάνσεις, έτσι ώστε στο annotation να συμβάλει το μοντέλο στη δημιουργία του dataset.

Για τις ανάγκες της παρούσας εργασίας, στοχεύσαμε στη δημιουργία ενός dataset με εικόνες (1ο και 2ο βήμα στην Σχήμα 5), με το οποίο θα επανεκπαιδεύσουμε τα μοντέλα που χρησιμοποιήθηκαν στο σύστημα εντοπισμού έκνομων επιβατών. Οι εικόνες του dataset προέρχονται από λήψεις βίντεο ενός μη επανδρωμένου εναέριου οχήματος σε πραγματικό περιβάλλον, κατά τη διάρκεια της πιλοτικής εφαρμογής του συστήματος εντοπισμού στο έργο ODYSSEUS. Συνολικά συλλέχθηκαν δύο βίντεο διάρκειας περίπου 12 λεπτών το καθένα. Το κάθε βίντεο αντιστοιχεί σε διαφορετική πιλοτική εφαρμογή του συστήματος.

Συνολικά το dataset αποτελείται από 12 τμήματα. Το κάθε ένα από τα πρώτα 10 τμήματα περιλαμβάνει συνεχόμενα καρέ διάρκειας 5 δευτερολέπτων και προέρχεται από το 1ο βίντεο. Επιλέχθηκαν τα σημεία του βίντεο που εμφανίζεται ο άνθρωπος πάνω στο τρένο. Επίσης επιλέχθηκαν και μερικά σημεία του βίντεο όπου έχουμε λήψεις του τρένου από διαφορετικές γωνίες. Το 11ο τμήμα του dataset περιέχει εικόνες και από τα δύο βίντεο. Εδώ δεν απομονώσαμε τα σημεία του βίντεο που αποτυπώνεται ο άνθρωπος επάνω στο τρένο. Το τμήμα αυτό δημιουργήθηκε κάνοντας δειγματοληψία στο βίντεο με συχνότητα μια εικόνα (καρέ) ανά 5 δευτερόλεπτα βίντεο. Τέλος το 12ο τμήμα του dataset περιέχει αποκλειστικά μερικές εικόνες από το 2ο βίντεο (διαφορετικές λήψεις του τρένου όπου εμφανίζεται ο άνθρωπος πάνω σ' αυτό).

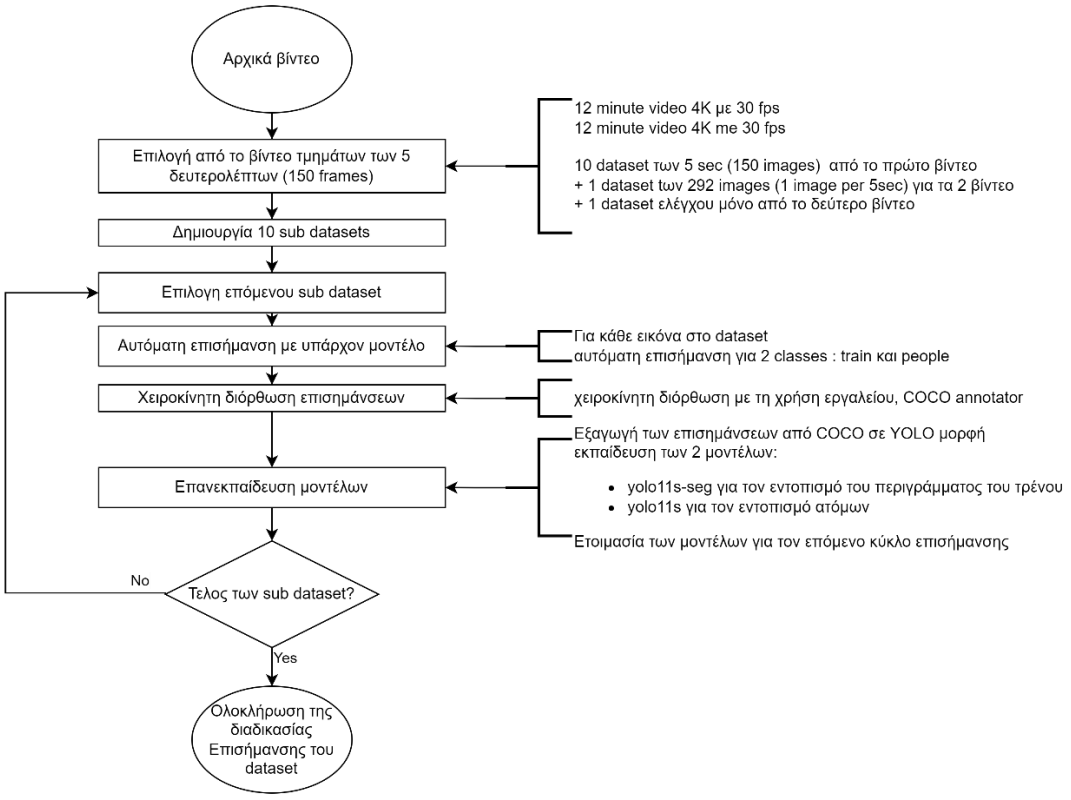
Στο επόμενο βήμα (4ο βήμα στην Σχήμα 5), για κάθε τμήμα του dataset, εφαρμόζεται αυτόματη επισημείωση (auto annotation) των καρέ με τη χρήση προ-εκπαιδευμένων μοντέλων. Το

[2] https://data.worldbank.org/

[3] https://www.kaggle.com/datasets

YOLOv11x χρησιμοποιείται για την ανίχνευση ανθρώπων, ενώ το μοντέλο Segment Anything Model (SAM-L), όπου αναφερθήκαμε στο κεφάλαιο 3.3εφαρμόζεται για την τμηματοποίηση του συρμού. Η χρήση των δύο αυτών μοντέλων επιτρέπει την ταχεία δημιουργία αρχικών επισημάνσεων για τις κατηγορίες «person» και «train», οι οποίες αποτελούν το θεμέλιο για την περαιτέρω επεξεργασία. Παρά την αυτοματοποίηση της διαδικασίας, οι επισημάνσεις αυτές δεν θεωρούνται τελικές, καθώς συχνά περιέχουν ελλείψεις ή ανακρίβειες.

Για την αντιμετώπιση των ανακριβειών της αυτόματης επισημείωσης, ακολουθεί το βήμα της χειροκίνητης επιμέλειας (Σχήμα 5: Βήμα – Χειροκίνητη διόρθωση επισημάνσεων). Τα αποτελέσματα της αυτόματης επισημείωσης εισάγονται στο εργαλείο COCO Annotator, όπου ο ερευνητής ελέγχει και διορθώνει τις ετικέτες. Η ανθρώπινη παρέμβαση είναι κρίσιμη, διότι εξασφαλίζει την ορθότητα των δεδομένων, διορθώνοντας περιπτώσεις όπου ο αλγόριθμος έχει αποτύχει να εντοπίσει σωστά ένα άτομο ή έχει δημιουργήσει ανακριβή όρια.



Σχήμα 5: Διαδικασία δημιουργίας και επισήμανσης του dataset

Μετά την ολοκλήρωση της επιμέλειας, τα επικαιροποιημένα δεδομένα χρησιμοποιούνται για την εκ νέου εκπαίδευση του ανιχνευτή. Το YOLOv11 επανεκπαιδεύεται (fine-tuning) πάνω στα νέα δεδομένα, προσαρμόζοντας τα βάρη του ώστε να ενσωματώνει τις διορθώσεις και να βελτιώνει την ικανότητά του να αναγνωρίζει ανθρώπους σε πραγματικές συνθήκες.

Μετά την ολοκλήρωση κάθε κύκλου, το νέο μοντέλο αντικαθιστά το προηγούμενο στη διαδικασία αυτόματης επισημείωσης, ξεκινώντας έναν νέο γύρο εκπαίδευσης. Με τον τρόπο αυτό δημιουργείται ένας συνεχής κύκλος βελτίωσης, όπου κάθε επανάληψη οδηγεί σε μεγαλύτερη ακρίβεια, καλύτερη γενίκευση και μειωμένα σφάλματα. Το σύστημα ανανεώνει σταδιακά τη γνώση του και προσαρμόζεται στις πραγματικές συνθήκες λήψης, επιτυγχάνοντας έτσι η ανθρώπινη παρέμβαση στο COCO Annotator να γίνεται όλο και πιο λίγη χρονικά. Έτσι καταφέραμε να δημιουργήσουμε dataset 1802 εικόνων από το διαθέσιμο βίντεο.

Στις παρακάτω παραγράφους αναλύονται τα βασικά βήματα τις διαδικασίας που είναι: η επιλογή του dataset (παράγραφος 4.2.1), η αυτόματη επισήμανση δεδομένων (παράγραφος 4.2.2), η χειροκίνητη διόρθωση επισημάνσεων (παράγραφος 4.2.3), η επανεκπαίδευση των μοντέλων (παράγραφος 4.2.4) και τέλος το τελικό παραγόμενο dataset (παράγραφος 4.2.5).

4.2.1 Περιγραφή dataset

Ξεκινώντας με τη δημιουργία dataset, χρησιμοποιήθηκαν δύο βίντεο διάρκειας 12 λεπτών περίπου έκαστο, τα οποίο λήφθηκαν από μη επανδρωμένο αεροσκάφος (drone) εξοπλισμένο με κάμερα υψηλής ανάλυσης σε διαφορετική χρονική στιγμή. Το drone κατέγραψε διάφορες σκηνές από περιοχή πάνω από σταθμό, όπου σε ορισμένες λήψεις εμφανιζόταν άνθρωπος επάνω στο τρένο σε διαφορετικές στάσεις και θέσεις. Το βίντεο διαθέτει χαρακτηριστικά 30 καρέ ανά δευτερόλεπτο (fps) και ανάλυση UHD/4K (3840x2160), εξασφαλίζοντας υψηλή ποιότητα εικόνας για την ανίχνευση και επισημείωση των αντικειμένων. Εδώ θα πρέπει να αναφερθεί ότι τα βίντεο παρήχθησαν σε διαφορετικές στιγμές με διαφορά επτά μηνών – το πρώτο Δεκέμβριος 2024 και το δεύτερο Ιούλιος 2025. Αυτός είναι ο λόγος που τα περισσότερα παραγόμενα dataset προέρχονται από το πρώτο βίντεο.

Κατά την επιλογή των καρέ για τη δημιουργία του dataset, δόθηκε έμφαση στην παρουσία του ανθρώπου υπό διαφορετικές γωνίες λήψης, αποστάσεις και στάσεις, ώστε να επιτευχθεί μεγαλύτερη ποικιλία στο σύνολο που θα εκπαιδευτούν τα μοντέλα. Από το 1ο βίντεο επιλέχθηκαν 1500 καρέ, που χρονικά μεταφράζεται σε βίντεο διαρκείας 50 δευτερολέπτων, δεδομένου ότι κάθε δευτερόλεπτο είναι 30 καρέ. Η εξαγωγή των εικόνων πραγματοποιήθηκε μέσω script σε python, το οποίο απέσπασε τα επιλεγμένα καρέ από το βίντεο με βάση προκαθορισμένα χρονικά διαστήματα και τα αποθήκευσε σε μορφή εικόνων. Για την καλύτερη διαχείριση των εικόνων, οι 1500 εικόνες χωρίστηκαν σε 10 τμήματα (subdataset) των 150 εικόνων που το καθένα αντιστοιχεί σε συνεχόμενο βίντεο των 5 δευτερολέπτων. Επιπρόσθετα με τη διαθεσιμότητα του δεύτερου βίντεο δημιουργήθηκε ένα ακόμη τμήμα που παρήχθη με τη χρήση και των 2 βίντεο επιλέγοντας ένα καρέ ανά 5 δευτερόλεπτα και συνολικά 292 εικόνες. Το subdataset12 (subdataset ελέγχου) δημιουργήθηκε από το βίντεο του δεύτερου οδηγού με

σκοπό να ελεγχθεί η ικανότητα των μοντέλων του πρώτου πειραματικού σεναρίου, το οποίο θα αναλυθεί στο υποκεφάλαιο 4.4, να γενικεύουν σε διαφορετικά δεδομένα από αυτά που εκπαιδεύτηκαν.

Πίνακας 1: Περιγραφή των επιμέρους datasets

Υπό-dataset	Αριθμός εικόνων	Πιλοτικό Βίντεο	Σχόλια
Subdataset01	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset02	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset03	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset04	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset05	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset06	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset07	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset08	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset09	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset10	150	Βίντεο 1	Συνεχόμενα καρέ. Αντιστοιχεί σε βίντεο 5 sec
Subdataset11	292	Βίντεο 1 και 2	Επιλογή ενός καρέ ανά 5 δευτερόλεπτα. Και από τα 2 βίντεο
Subdataset12	10	Βίντεο 2	10 καρέ από το βίντεο 2. Αποτυπώνουν τον άνθρωπο στο τρένο από διαφορετικές λήψεις

4.2.2 Αυτόματη επισήμανση δεδομένων

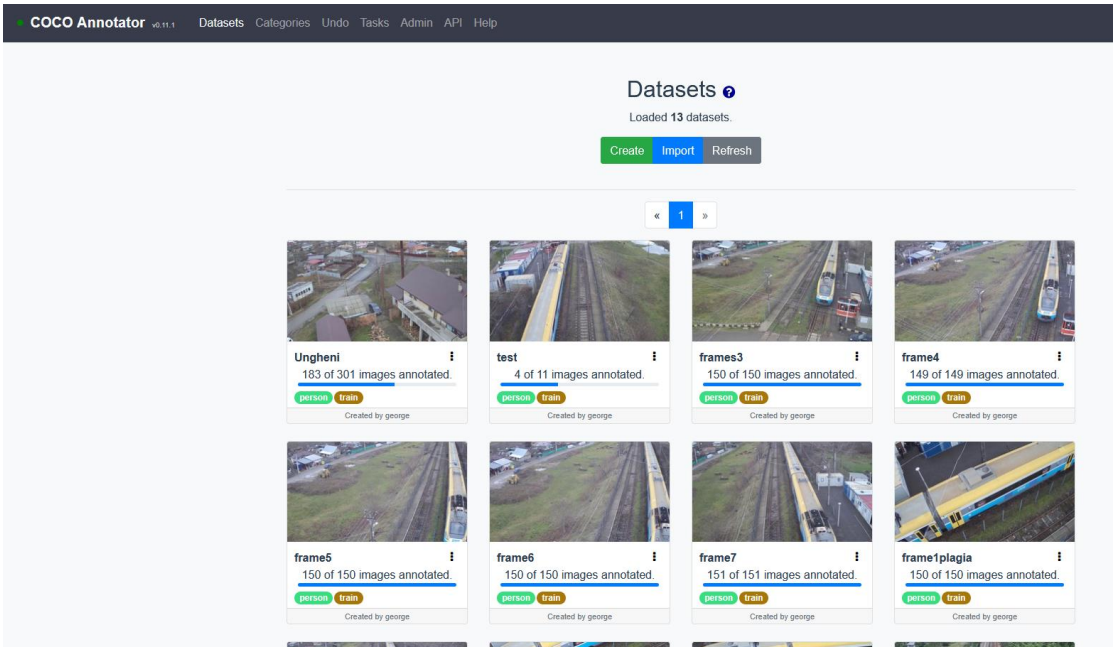
Αρχικά, για κάθε υπο-dataset πραγματοποιήθηκε η ανίχνευση ανθρώπων με τη χρήση του μοντέλου YOLOv11x, ενώ για την τμηματοποίηση του συρμού εφαρμόστηκε το Segment Anything Model (SAM-L), με τη βοήθεια της συνάρτησης auto-annotation της βιβλιοθήκης Ultralytics, η οποία επιτρέπει τον αυτόματο εντοπισμό και τη δημιουργία αρχικών επισημάνσεων χωρίς την ανάγκη χειροκίνητης παρέμβασης. Οι παραγόμενες επισημάνσεις μετατράπηκαν από μορφή YOLO σε COCO format μέσω python script και ενσωματώθηκαν αυτόματα στο περιβάλλον του COCO Annotator για περαιτέρω διαχείριση μέσω των APIs που διαθέτει η πλατφόρμα, όπως περιγράφεται στην επόμενη παράγραφο.

Θα πρέπει να σημειωθεί ότι η επιλογή του YOLO11x αντί του YOLO11s έγινε γιατί το πρώτο έχει καλύτερη απόδοση έναντι του δευτέρου, θυσιάζοντας όμως υπολογιστική ισχύ. Το

μοντέλο που χρησιμοποιείται για τον τελικό εντοπισμό των ατόμων είναι το YOLO11s το οποίο είναι βελτιστοποιημένο για χρήση σε edge devices και επιτρέπει παρακολούθηση σε πραγματικό χρόνο. Κατά τη διάρκεια της επισήμανσης δεν μας ενδιαφέρει η παρακολούθηση σε πραγματικό χρόνο αλλά η καλύτερη ποιότητα των αποτελεσμάτων. Αντίστοιχα συμβαίνει και με το SAM-L όπου παρουσιάζει καλύτερα αποτελέσματα έναντι του YOLO [34].

4.2.3 Χειροκίνητη διόρθωση επισημάνσεων

Για την χειροκίνητη επισήμανση χρησιμοποιήσαμε το εργαλείο COCO Annotator[4]. Το COCO Annotator (Σχήμα 6) είναι ένα διαδικτυακό εργαλείο επισημείωσης (web-based annotation tool) ανοιχτού κώδικα, που έχει αναπτυχθεί για τη δημιουργία, οπτικοποίηση και διαχείριση συνόλων δεδομένων τύπου COCO (Common Objects in Context). Υποστηρίζει ανίχνευση αντικειμένων (object detection), τμηματοποίηση (instance segmentation) και πολυκατηγορική ταξινόμηση (multi-class labeling), επιτρέποντας στους χρήστες να δημιουργούν ή να διορθώνουν ετικέτες απευθείας πάνω στις εικόνες μέσω γραφικού περιβάλλοντος.



Σχήμα 6: COCO Annotator για τη διαχείριση των dataset

Στο στάδιο αυτό, έγιναν χειροκίνητες διορθώσεις στα περιγράμματα των αντικειμένων στο COCO Annotator, ώστε να αποτυπώνονται με μεγαλύτερη ακρίβεια τα όρια των κλάσεων person και train. Στο Σχήμα 7, μπορούμε να παρατηρήσουμε τη διαφορά μεταξύ της αυτόματης

[4] https://madewithvuejs.com/coco-annotator

και επισήμανσης και της χειροκίνητης διόρθωσης. Ενώ η αυτόματη επισήμανση εντοπίζει το τρένο, ωστόσο αποτυγχάνει να αποτυπώσει με ακρίβεια τα όρια της περιοχή του τρένου.

Σχήμα 7: Σύγκριση αυτόματης επισήμανσης (αριστερά) και χειροκίνητης διόρθωσης (δεξιά) των εικόνων

Μετά την ολοκλήρωση των διορθώσεων, τα annotations εξήχθησαν σε YOLO format. Επειδή όμως για την επανεκπαίδευση του YOLOv11x απαιτούνται bounding boxes, δημιουργήθηκαν αυτά μέσω Python script, χρησιμοποιώντας γεωμετρικό υπολογισμό βάσει των ελάχιστων και μέγιστων συντεταγμένων των σημείων κάθε πολυγώνου. Η διαδικασία αυτή βασίζεται στη Shoelace Formula [35], η οποία επιτρέπει τον ακριβή προσδιορισμό των ορίων κάθε αντικειμένου με υπολογιστική ακρίβεια. Η Shoelace Formula (ή Gauss area formula) είναι ένας αλγόριθμος της υπολογιστικής γεωμετρίας που επιτρέπει τον υπολογισμό του εμβαδού και των ορίων ενός απλού πολυγώνου χρησιμοποιώντας μόνο τις συντεταγμένες των κορυφών του. Στην προκειμένη περίπτωση, οι συντεταγμένες αυτές αντιστοιχούν στα σημεία που ορίζουν το segmentation κάθε αντικειμένου, δηλαδή σε όλα τα σημεία του πολυγώνου που προκύπτει από την επεξεργασία των annotations.

4.2.4 Επανεκπαίδευση των μοντέλων

Στο τέλος της χειροκίνητης διόρθωσης και της εξαγωγής των επισημάνσεων σε YOLO format, το επισημασμένο υπο-dataset των 150 καρέ, χρησιμοποιείται για επανεκπαίδευση του YOLO11x, έτσι ώστε το παραγόμενο επανεκπαιδευμένο μοντέλο YOLOv11x-new.pt να χρησιμοποιηθεί για την αυτόματη επισημείωση του επόμενου υποdataset. Σε κάθε κύκλο παρατηρείται βελτιωμένη ικανότητα του YOLOv11 να ανιχνεύει τα θεμιτά αντικείμενα, μειώνοντας έτσι το χρόνο που καταναλώνεται στις χειροκίνητες διορθώσεις στο COCO Annotator.

4.2.5 Περιγραφή παραγόμενου dataset

Το τελικό dataset αποτελείται από 1802 εικόνες, οι οποίες περιλαμβάνουν δύο βασικές κλάσεις: person και train. Η κλάση *train* προβλέπεται να χρησιμοποιηθεί για τη διαδικασία τμηματοποίησης (segmentation), με σκοπό τον ακριβή προσδιορισμό και την οριοθέτηση του

συρμού εντός του πεδίου ενδιαφέροντος, ενώ η κλάση *person* προορίζεται για τη διαδικασία ανίχνευσης (prediction), η οποία στοχεύει στον εντοπισμό και την αναγνώριση ανθρώπινων παρουσιών επί του συρμού.

Στο Σχήμα 8, μπορούμε να παρατηρήσουμε μερικά αποτελέσματα της διαδικασίας annotation όπως αναλύθηκε παραπάνω.

Σχήμα 8: Παραδείγματα εικόνων μετά τη διαδικασία επισήμανσης (αυτόματη και χειροκίνητη) για τις 2 κλάσεις. Πράσινο – άνθρωπος και κίτρινο - τρένο

Για την προετοιμασία του dataset για τη διαδικασία εκπαίδευσης των μοντέλων, το σύνολο δεδομένων χωρίστηκε σε δύο υποσύνολα: 90% για εκπαίδευση (train dataset) και 10% για επικύρωση (validation dataset) για κάθε υπο-dataset. Ο διαχωρισμός αυτός επιλέχθηκε ώστε το μοντέλο να έχει επαρκή ποικιλία παραδειγμάτων από κάθε υπο-dataset για να μάθει τα πρότυπα των αντικειμένων, ενώ παράλληλα το μικρότερο σύνολο επικύρωσης επιτρέπει την αντικειμενική αξιολόγηση της απόδοσης του μοντέλου σε δεδομένα χωρίς το μοντέλο να μην περιέχει στην εκπαίδευση του κάποιο υπο-dataset.

Αξίζει να σημειωθεί ότι αναλογίες διαχωρισμού όπως 80-20 ή 70-30 αποτελούν επίσης ευρέως αποδεκτές πρακτικές στη βιβλιογραφία [40]. Ωστόσο, στην παρούσα εργασία προτιμήθηκε ο διαχωρισμός 90-10 ώστε να μεγιστοποιηθεί το πλήθος δειγμάτων εκπαίδευσης.

4.3 Μεθοδολογία εκπαίδευσης μοντέλων

4.3.1 Αρχιτεκτονική μοντέλου/ων

Αφού, δημιουργήθηκε το dataset, επόμενο είναι η επανεκπαίδευση του μοντέλου πάνω σε αυτό. Σύμφωνα με την παράγραφο 3.5, το προτεινόμενο σύστημα αποτελείται από δύο διακριτά στάδια επεξεργασίας και δύο επίπεδα πρόβλεψης. Στο πρώτο στάδιο πραγματοποιείται η αναγνώριση του τρένου στην τρέχουσα εικόνα και ο ορισμός της αντίστοιχης μάσκας που προσδιορίζει τα όρια του αντικειμένου ενδιαφέροντος. Το δεύτερο στάδιο περιλαμβάνει την αναγνώριση ανθρώπων επάνω στο απομονωμένο τμήμα του τρένου, δηλαδή στη διορθωμένη εικόνα που έχει προκύψει από το προηγούμενο βήμα.

Για την υλοποίηση των επιμέρους λειτουργιών χρησιμοποιούνται δύο διακριτά μοντέλα μάθησης τύπου YOLO της σειράς 11. Το πρώτο μοντέλο, Yolo11s-seg.pt, αξιοποιείται για την αναγνώριση και την εξαγωγή της μάσκας του τρένου μέσω διαδικασίας τμηματοποίησης (segmentation). Η αρχιτεκτονική του περιλαμβάνει χαρακτηριστικά επίπεδα (layers) και στάδια (stages) που επιτρέπουν τον ακριβή εντοπισμό των περιγραμμάτων του τρένου σε κάθε καρέ της εισόδου. Το δεύτερο μοντέλο, Yolo11s.pt, είναι υπεύθυνο για την ανίχνευση ανθρώπων πάνω στην επιφάνεια του τρένου, αξιοποιώντας την απομονωμένη εικόνα που προκύπτει μετά την εφαρμογή της μάσκας. Και τα δύο μοντέλα είναι προσαρμοσμένα ώστε να λειτουργούν αποδοτικά σε πραγματικό χρόνο, λαμβάνοντας υπόψη τη δυναμική φύση των σκηνών που καταγράφονται από UAV.

Και τα δύο μοντέλα βασίζονται στην εκδοχή small (s) της αρχιτεκτονικής YOLO11, η οποία επιλέχθηκε με σκοπό τη μείωση του υπολογιστικού κόστους και την εξασφάλιση λειτουργίας σε πραγματικό χρόνο, ιδίως υπό περιορισμένους υπολογιστικούς πόρους, όπως αυτοί που χαρακτηρίζουν τα συστήματα επιτήρησης UAV. Η επιλογή του μοντέλου small συνεπάγεται μικρή απώλεια ακρίβειας σε σχέση με μεγαλύτερες εκδόσεις (medium, large).

Η αρχιτεκτονική των μοντέλων YOLO δίνεται αναλυτικά στο υποκεφάλαιο 2.2. Επισημαίνεται ότι η εκδοχή YOLO11s-seg.pt ενσωματώνει επιπλέον ένα segmentation head, το οποίο επιτρέπει την εκτίμηση ακριβούς μάσκας για κάθε αναγνωρισμένο τρένο, ενώ το YOLO11s.pt επικεντρώνεται στην ανίχνευση ανθρώπινων παρουσιών μέσω του detection head.

4.3.2 Πλατφόρμα εκτέλεσης

Για την εκτέλεση των πειραμάτων αξιοποιήθηκαν 2 διαφορετικές πλατφόρμες. Στην πρώτη πλατφόρμα εκτελέστηκε η διαδικασία επισημείωσης και εκπαίδευσης, ενώ στη δεύτερη έγινε η εκτέλεση και η αξιολόγηση των μοντέλων.

Η πρώτη πλατφόρμα ήταν μια εικονική μηχανή (Virtual Machine – VM), η οποία παρείχε το απαραίτητο υπολογιστικό περιβάλλον για την επανεκπαίδευση των μοντέλων, την εφαρμογή των συναρτήσεων της βιβλιοθήκης Ultralytics και την εκτέλεση της διαδικασίας της δημιουργίας του dataset. Η πλατφόρμα αυτή χρησιμοποιήθηκε για όλες τις φάσεις annotation, επανεκπαίδευσης σε διάφορες παραμέτρους και αξιολόγησης.

Πίνακας 2: Χαρακτηριστικά εικονικής μηχανής

CPU cores	8
RAM	16
Hard disk	120GB
GPU	NaN

Ο Πίνακας 2 αποτυπώνει τα βασικά χαρακτηριστικά της εικονικής μηχανής. Εδώ θα πρέπει να σημειωθεί ότι δεν υπήρξε διαθέσιμο εικονικό μηχάνημα με κάρτα γραφικών, οπότε το σύνολο της εκπαίδευσης έγινε με τη χρήση CPU.

Ως πλατφόρμα εκτέλεσης και αξιολόγησης των μοντέλων χρησιμοποιήθηκε μέρος της σειράς NVIDIA Jetson, η οποία αναλύεται στην παράγραφο 2.3 και πιο συγκεκριμένα το Jetson Orin Nano development kit. Το Σχήμα 9 περιλαμβάνει τις προδιαγραφές του υλικού της επιλεγμένης πλατφόρμας εκτέλεσης.

NVIDIA Jetson Orin Nano Developer Kit

Technical Specifications	
	Jetson Orin Nano 8GB Module
GPU	NVIDIA Ampere architecture with 1024 CUDA cores and 32 tensor cores
CPU	6-core Arm® Cortex®-A78AE v8.2 64-bit CPU 1.5MB L2 + 4MB L3
Memory	8GB 128-bit LPDDR5 68GB/s
Storage	Supports SD card slot and external NVMe
Video Encode	1080p30 supported by 1-2 CPU cores
Video Decode	1x 4K60 (H.265) 2x 4K30 (H.265) 5x 1080p60 (H.265) 11x 1080p30 (H.265)
Power	7W-15W

Refer to the Software Features section of the latest NVIDIA Jetson Linux Developer Guide for a list of supported features.

	Reference Carrier Board
Camera	2x MIPI CSI-2 22-pin camera connectors
PCIe	M.2 Key M slot with x4 PCIe Gen3 M.2 Key M slot with x2 PCIe Gen3 M.2 Key E slot
USB	USB Type-A connector: 4x USB 3.2 Gen2 USB Type-C connector for UFP
Networking	1xGbE connector
Display	1x DP 1.2 (+MST) connector
Other I/O	40-pin expansion header (UART, SPI, I2S, I2C, GPIO) 12-pin button header 4-pin fan header microSD slot

Σχήμα 9: Προδιαγραφές υλικού Jetson Orin Nano development Kit[5].

4.3.3 Διαδικασία εκπαίδευσης

Η διαδικασία εκπαίδευσης των μοντέλων πραγματοποιήθηκε μέσω του μηχανισμού Ultralytics YOLO [39]. Πριν από την εκπαίδευση, το σύνολο δεδομένων χωρίστηκε σε δύο υποσύνολα: train set και validation set. Το train set χρησιμοποιήθηκε για την ενημέρωση των βαρών του μοντέλου, ενώ το validation set χρησιμοποιήθηκε για την αξιολόγηση της απόδοσης σε δεδομένα που δεν είχαν παρουσιαστεί κατά την εκπαίδευση, ώστε να εκτιμηθεί η ικανότητα γενίκευσης.

Η εκπαίδευση του μοντέλου ρυθμίστηκε μέσω συγκεκριμένων παραμέτρων, οι οποίες καθορίζουν τη συμπεριφορά της διαδικασίας μάθησης. Ο αριθμός των epochs ορίζει πόσες φορές το σύνολο δεδομένων θα χρησιμοποιηθεί πλήρως για ενημέρωση των βαρών· μικρός αριθμός οδηγεί σε ανεπαρκή μάθηση (underfitting), ενώ υπερβολικά μεγάλος μπορεί να

[5] https://files.seeedstudio.com/wiki/Jetson-Orin-Nano-DevKit/jetson-orin-nano-developer-kit-datasheet.pdf

προκαλέσει υπερπροσαρμογή (overfitting). Η παράμετρος batch size καθορίζει πόσα δεδομένα επεξεργάζεται το μοντέλο κάθε φορά. Όσο μεγαλύτερη είναι, τόσο πιο σταθερά μαθαίνει, γιατί η ενημέρωση των βαρών βασίζεται σε περισσότερες πληροφορίες, αλλά αυτό απαιτεί περισσότερη μνήμη και υπολογιστική ισχύ, αφού πρέπει να επεξεργαστεί περισσότερα δεδομένα ταυτόχρονα. Η παράμετρος freeze επιτρέπει τον καθορισμό τμημάτων του δικτύου που δεν θα ενημερωθούν κατά το fine-tuning, κάτι που είναι χρήσιμο όταν επιδιώκεται η προσαρμογή ενός συγκεκριμένου υποσυστήματος, όπως το segmentation head, χωρίς να αλλοιωθούν τα γενικά χαρακτηριστικά του backbone. Ο αρχικός ρυθμός μάθησης (lr0) επηρεάζει την ταχύτητα προσαρμογής των βαρών: υψηλές τιμές μπορεί να οδηγήσουν σε αστάθεια, ενώ χαμηλές σε αργή σύγκλιση. Ο τελικός ρυθμός μάθησης (lrf) εφαρμόζεται μέσω scheduling ώστε η διαδικασία μάθησης να σταθεροποιείται όσο πλησιάζει στη σύγκλιση. Τέλος, η παράμετρος patience καθορίζει πόσες επαναλήψεις χωρίς βελτίωση στο validation επιτρέπονται πριν ενεργοποιηθεί μηχανισμός early stopping, αποτρέποντας έτσι την άσκοπη συνέχιση της εκπαίδευσης και περιορίζοντας το φαινόμενο της υπερπροσαρμογής.

4.4 Πειραματικά σενάρια

Στην παρούσα εργασία πραγματοποιήθηκαν δύο διακριτά πειραματικά σενάρια εκπαίδευσης, με στόχο την αξιολόγηση της ικανότητας του μοντέλου να γενικεύει όταν αλλάζουν οι συνθήκες λήψης και το οπτικό περιβάλλον.

4.4.1 Σενάριο 1

Το πρώτο πείραμα βασίστηκε στο αρχικό σύνολο δεδομένων των 1500 εικόνων, το οποίο προήλθε εξ ολοκλήρου από βίντεο του πρώτου πιλότου. Τα δεδομένα χωρίστηκαν σε αναλογία 90% για εκπαίδευση και 10% για αξιολόγηση, και εφαρμόστηκε εκπαίδευση με transfer learning χωρίς χρήση freeze. Δηλαδή το μοντέλο ενημέρωσε όλα τα βάρη του, περιλαμβάνοντας τόσο το backbone όσο και το detection head. Δοκιμάστηκαν διαφορετικές διάρκειες εκπαίδευσης, με 5, 10, 20, 50 και 100 epochs. Παράλληλα, έγινε αξιολόγηση του μοντέλου και σε δεύτερο dataset 10 των δέκα καρέ, από βίντεο του δεύτερου οδηγού με ελαφρώς διαφορετική γωνία λήψης και στάση του σώματος των έκνομων επιβατών πάνω στο συρμό.

Η αναλυτική παραμετροποίηση της συνάρτησης επανεκπαίδευσης του πρώτου πειράματος παρουσιάζεται στον Πίνακα Πίνακας 3.

Πίνακας 3: Παράμετροι επανεκπαίδευσης πρώτου πειραματικού σεναρίου

Παράμετρος	**Τιμή**
epochs	5, 10, 20, 50, 100 (multiple runs)
freeze	0 (default)
batches	8
lr0	0.01 (default)
lrf	0.01 (default)
patience	100 (default)
subdatasets Που χρησιμοποιήθηκαν	Τα 10 πρώτα subdatasets (150 εικόνες έκαστο) Subdataset01 – subdataset10

Με βάση τα αποτελέσματα του πρώτου πειράματος, τα οποία θα παρουσιαστούν αναλυτικά στο κεφάλαιο 4.5, κρίθηκε αναγκαίο να τροποποιηθεί τόσο η επιλογή δεδομένων όσο και ο τρόπος εκπαίδευσης, ώστε να παραχθούν πιο αποδοτικά μοντέλα.

4.4.2 Σενάριο 2

Στο δεύτερο πείραμα χρησιμοποιήθηκε το subdataset11, το οποίο περιλάμβανε δεδομένα από βίντεο των δύο οδηγών, με δειγματοληψία ενός καρέ ανά πέντε δευτερόλεπτα. Με αυτόν τον τρόπο ενσωματώθηκε μεγαλύτερη ποικιλία ως προς τη γωνία λήψης, τη σχετική θέση του UAV και τη στάση των ανθρώπων πάνω στον συρμό. Επιπλέον, τροποποιήθηκαν οι παράμετροι εκπαίδευσης. Συγκεκριμένα, εφαρμόστηκε freeze στο πρώτο στάδιο του δικτύου (freeze = 11), ώστε το backbone να παραμείνει σταθερό και να μην επανεκπαιδευτεί. Έτσι, η προσαρμογή περιορίστηκε στα υψηλότερα επίπεδα του δικτύου και στο detection head. Ο αρχικός ρυθμός μάθησης μειώθηκε στο 1e-4, ώστε να αποφευχθούν απότομες και ανεπιθύμητες αλλαγές στα προ-εκπαιδευμένα βάρη. Ο μηχανισμός early stopping ορίστηκε με παράμετρο patience ίση με 20, με αποτέλεσμα η εκπαίδευση να ολοκληρωθεί αυτόματα στο epoch 39, ενώ το καλύτερο μοντέλο προέκυψε στο epoch 29. Η αναλυτική παραμετροποίηση της συνάρτησης επανεκπαίδευσης του δεύτερου πειράματος παρουσιάζεται στον Πίνακα Πίνακας 4.

Πίνακας 4: Παράμετροι επανεκπαίδευσης δεύτερου πειραματικού σεναρίου και για τα 2 μοντέλα

Παράμετρος	**Τιμή**

epochs	100 (default)
freeze	11
batches	2
lr0	1e-4
lrf	0.01
patience	20
Subdataset που χρησιμοποιήθηκε	subdataset11 (292 εικόνες)

4.5 Αποτελέσματα και ανάλυση

4.5.1 Ποσοτικά αποτελέσματα

4.5.1.1 Σενάριο 1

Ξεκινώντας με το πρώτο πείραμα κάνουμε την αξιολόγηση στο validation set, που εξηγήσαμε προηγουμένως για τις διάφορες εποχές του μοντέλου. Αναλυτικά τα αποτελέσματα φαίνονται στον Πίνακα Πίνακας 5.

Πίνακας 5: Αποτελέσματα αξιολόγησης μοντέλων στο validation set για το πρώτο πειραματικό σενάριο

Model	**epochs**	**Mean Average precision**	**Average precision per class**		**F1**	**recall**
		mAP@0.5	AP_person	AP_train		
YOLO11s	0	0.629	0.274	0.984	0.62	0.69
YOLO11s-5	5	0.976	0.956	0.995	0.96	0.98
YOLO11s-10	10	0.97	0.948	0.995	0.96	0.97
YOLO11s-20	20	0.98	0.964	0.995	0.97	0.98
YOLO11s-50	50	0.982	0.969	0.995	0.96	0.99
YOLO11s-100	100	0.982	0.969	0.995	0.97	0.98

Η βελτίωση της απόδοσης του μοντέλου πραγματοποιείται κυρίως στις πρώτες πέντε εποχές. Το mAP@0.5 αυξάνεται από περίπου 0.63 σε 0.976, ενώ το AP_person από 0.274 σε 0.956, δείχνοντας ότι το μοντέλο προσαρμόζεται γρήγορα στα νέα δεδομένα. Η κατηγορία train

εμφανίζει ήδη υψηλή απόδοση (AP≈0.995) και παραμένει σταθερή, γεγονός που υποδηλώνει χαμηλή οπτική μεταβλητότητα.

Στις δέκα εποχές, οι επιδόσεις σταθεροποιούνται (mAP≈0.972, AP_people≈0.948), κάτι που σημαίνει ότι το μεγαλύτερο μέρος της πληροφορίας έχει ήδη απορριφθεί.

Η καλύτερη επίδοση επιτυγχάνεται περίπου στις είκοσι εποχές (mAP≈0.980, AP_people≈0.964), με μικρή αλλά μετρήσιμη βελτίωση τόσο σε precision όσο και σε recall. Μετά τις είκοσι εποχές, η διαδικασία εισέρχεται σε φάση κορεσμού: τα βάρη σταθεροποιούνται και η απόδοση δεν μεταβάλλεται ουσιαστικά.

Στις πενήντα και εκατό εποχές, οι μετρικές παραμένουν πρακτικά αμετάβλητες (mAP≈0.982, AP_people≈0.969, AP_train≈0.995). Το μοντέλο έχει ολοκληρώσει την προσαρμογή του και οποιαδήποτε περαιτέρω εκπαίδευση δεν προσφέρει ουσιαστική βελτίωση.

Στη συνέχεια, όπως αναφέρθηκε και στην περιγραφή των πειραματικών σεναρίων τα παραγόμενα μοντέλα των διαφόρων εποχών ελέγχθηκαν ως προς την ικανότητα τους στο subdataset ελέγχου (subdataset12). Αναλυτικά τα αποτελέσματα φαίνονται στον Πίνακα 6.

Πίνακας 6: Αποτελέσματα αξιολόγησης μοντέλων στο με βάση το dataset ελέγχου (subdataset12) για το πρώτο πειραματικό σενάριο

Model	**epochs**	**Mean Average precision**	**Average precision per class**		**F1**	**recall**
		mAP@0.5	AP_person	AP_train		
YOLO11s	0	0.648	0.302	0.995	0.62	0.7
YOLO11s-5	5	0.575	0.155	0.995	0.6	0.6
YOLO11s-10	10	0.525	0.056	0.995	0.54	0.54
YOLO11s-20	20	0.483	0.0	0.966	0.47	0.49
YOLO11s-50	50	0.476	0.0	0.995	0.36	0.44
YOLO11s-100	100	0.032	0.06	0.0	0.01	0.0

Το αρχικό μοντέλο ξεκινά με μέτρια επίδοση (mAP@0.5 ≈ 0.648), αλλά η επανεκπαίδευση δεν οδηγεί σε βελτίωση. Αντιθέτως, η συνολική απόδοση μειώνεται σταθερά όσο αυξάνονται οι εποχές, έως την πλήρη κατάρρευση στις εκατό εποχές (mAP ≈ 0.03). Η συμπεριφορά αυτή δείχνει ότι το μοντέλο προσαρμόζεται υπερβολικά στα δεδομένα του αρχικού dataset, χωρίς να αποκτά ικανότητα γενίκευσης.

Η ανάλυση ανά κλάση εξηγεί την υποβάθμιση. Η κατηγορία train διατηρεί υψηλό AP για αρκετές εποχές, λόγω χαμηλής οπτικής μεταβλητότητας, αλλά αρχίζει να υποχωρεί στις πολύ

μεγάλες διάρκειες εκπαίδευσης. Αντίθετα, η κατηγορία person εμφανίζει άμεση και σοβαρή πτώση, με το AP να μειώνεται προοδευτικά έως και τον πλήρη μηδενισμό της ανίχνευσης. Το μοντέλο μαθαίνει να αναγνωρίζει ανθρώπους μόνο στις αρχικές συνθήκες λήψης και αδυνατεί να μεταφέρει αυτή τη γνώση σε διαφορετικό οπτικό περιβάλλον. Παρότι το precision παραμένει υψηλό, αυτό οφείλεται στο ότι το μοντέλο επιλέγει να κάνει λίγες και σίγουρες προβλέψεις, χάνοντας όμως μεγάλο μέρος των πραγματικών εμφανίσεων της κατηγορίας person.

Επιπλέον, αν και το αρχικό dataset περιλάμβανε μεγάλο αριθμό εικόνων, αυτές προέρχονταν από συνεχόμενα καρέ του ίδιου βίντεο, με ελάχιστη διακύμανση στη γωνία λήψης, στην κίνηση και στη σκηνή. Η επαναληπτικότητα αυτή μειώνει την ποικιλία του οπτικού περιεχομένου, κάνοντας το μοντέλο να «αποστηθίζει» τη συγκεκριμένη σκηνή αντί να μαθαίνει γενικεύσιμα χαρακτηριστικά. Η επιλογή freeze = 0 και αρχικού learning rate lr0 = 0.01 ενισχύει αυτή τη συμπεριφορά, καθώς επιτρέπει ταχύτερες και εκτεταμένες ενημερώσεις σε όλα τα επίπεδα του δικτύου. Γι’ αυτό και στις πρώτες εποχές παρατηρείται εντυπωσιακή αύξηση της απόδοσης στο training domain, όμως αυτή η ταχεία προσαρμογή οδηγεί στη συνέχεια σε υπερπροσαρμογή και βαθμιαία υποβάθμιση της γενίκευσης.

Συνολικά, η εκπαίδευση χωρίς προσαρμογή στη δομή των δεδομένων και χωρίς έλεγχο του ρυθμού ενημέρωσης των βαρών οδηγεί ταυτόχρονα σε υπερπροσαρμογή στο αρχικό σύνολο και σε καταστροφική λήθη της προηγούμενης γενικής γνώσης του μοντέλου.

4.5.1.2 Σενάριο 2

Εξαιτίας της κακής απόδοσης των μοντέλων στην ανίχνευση του πρώτου πειράματος, δεν προχωρήσαμε στην αξιολόγηση της τμηματοποίησης των μοντέλων. Αντί αυτού θα προχωρήσουμε κατευθείαν στο δεύτερο πείραμα. Τα αποτελέσματα της τμηματοποίησης παρατίθενται στον Πίνακα Πίνακας 7. Στο δεύτερο πείραμα το μοντέλο ανίχνευσης επανεκπαιδεύτηκε

Στο δεύτερο πείραμα η επανεκπαίδευση του μοντέλου ανίχνευσης πραγματοποιήθηκε μόνο για την κλάση *person*, ενώ αντίστοιχα η επανεκπαίδευση του μοντέλου τμηματοποίησης έγινε αποκλειστικά για την κλάση *train*. Για τον λόγο αυτό δεν υπολογίζεται *average precision per class*, καθώς δεν υπάρχει αξιολόγηση πολλαπλών κατηγοριών αλλά μονή στόχευση ανά μοντέλο.

Επιπλέον, για κάθε κλάση επιλέχθηκε η τιμή του *precision* σε συγκεκριμένο threshold που παρείχε τη βέλτιστη απόδοση και όχι ο μέσος όρος των τιμών σε διαφορετικά thresholds. Αυτή η επιλογή έγινε με σκοπό τη σαφέστερη αποτίμηση της πραγματικής συμπεριφοράς του μοντέλου στο εύρος λειτουργίας που μας ενδιαφέρει.

Τέλος, προστέθηκε η μετρική *mAP50-95*, ώστε να αξιολογηθεί η απόδοση του μοντέλου σε υψηλό εύρος τιμών IoU. Η συγκεκριμένη μετρική είναι σημαντική διότι επιτρέπει την πιο αξιόπιστη εξέταση της ποιότητας τμηματοποίησης, ειδικά σε περιβάλλοντα όπου απαιτείται υψηλή ακρίβεια στον εντοπισμό ορίων των αντικειμένων

Πίνακας 7: Αποτελέσματα ανίχνευσης ατόμων στο δεύτερο πειραματικό σενάριο

model	mAP50	mAP50-95	precision	recall	f1_score
yolo11s.pt	0.651	0.406	0.770	0.590	0.668
Updated.pt	0.852	0.687	0.884	0.824	0.853

Ο πίνακας παρουσιάζει την ποσοτική βελτίωση της απόδοσης του μοντέλου μετά την επανεκπαίδευση στο νέο, ποικιλόμορφο dataset. Το αρχικό προεκπαιδευμένο μοντέλο (yolo11s.pt) εμφανίζει mAP@0.5 = 0.651 και mAP@0.5–0.95 = 0.406, με μέτρια ισορροπία ανάμεσα σε precision (0.770) και recall (0.590), γεγονός που οδηγεί σε συνολικό f1-score = 0.668. Μετά τη στοχευμένη προσαρμογή (freeze στο backbone και χαμηλό learning rate), το ενημερωμένο μοντέλο (updated.pt) επιτυγχάνει σημαντικά υψηλότερες επιδόσεις σε όλες τις μετρικές: mAP@0.5 = 0.852 και mAP@0.5–0.95 = 0.687, ενώ τόσο το precision (0.884) όσο και το recall (0.824) αυξάνονται, με συνέπεια την άνοδο του f1-score σε 0.853. Η συνολική βελτίωση υποδεικνύει ότι το μοντέλο όχι μόνο μειώνει τις λανθασμένες ανιχνεύσεις, αλλά εντοπίζει και μεγαλύτερο ποσοστό πραγματικών εμφανίσεων, αναπτύσσοντας ουσιαστικά καλύτερη ικανότητα γενίκευσης.

Εδώ πρέπει να επισημανθεί ότι η αξιολόγηση έγινε σε όλες τις επισημειωμένες παρουσίες ανθρώπων στο dataset, και όχι μόνο σε όσες βρίσκονται πάνω στον συρμό. Οι περισσότερες αποτυχίες εντοπισμού σχετίζονται με άτομα που βρίσκονται πλάι στον συρμό, όποτε η πραγματική επίδοση είναι καλύτερη.

Στο Πίνακα 8 βλέπουμε την απόδοση των δύο μοντέλων, ένα το αρχικό και δεύτερο το επανεκπαιδευμένο τμηματοποίησης.

Στο μοντέλο τμηματοποίησης παρατηρείται ουσιαστική βελτίωση μετά την επανεκπαίδευση. Το αρχικό προεκπαιδευμένο μοντέλο (yolo11s-seg.pt) επιτυγχάνει mAP50= 0.932 και mAP0.5-0.95 = 0.754, τιμές που δείχνουν ήδη υψηλή ικανότητα εντοπισμού του συρμού με καλή αντιστοίχιση των μασκών σε βασικό επίπεδο επικάλυψης. Ωστόσο, το mAP0.5-0.95, το οποίο αποτυπώνει την γεωμετρική ακρίβεια της μάσκας σε πολλαπλά thresholds του Intersection over Union, κατώφλι ιδιαίτερης σημασίας στη τμηματοποίηση, παραμένει σχετικά χαμηλό σε σχέση με τις απαιτήσεις της εφαρμογής. Η παράμετρος αυτή είναι κρίσιμη, διότι η μάσκα του συρμού χρησιμοποιείται για την αυστηρή οριοθέτηση της περιοχής ενδιαφέροντος

στο δεύτερο στάδιο εντοπισμού ανθρώπων πάνω στον συρμό. Όταν τα όρια της μάσκας δεν αποδίδονται με ακρίβεια, περιοχές εκτός συρμού μπορεί να συμπεριληφθούν, οδηγώντας σε ψευδείς θετικές ανιχνεύσεις ατόμων.

Πίνακας 8: Αποτελέσματα εντοπισμού και τμηματοποίησης τρένου στο δεύτερο πειραματικό σενάριο

model	mAP50	mAP50-95	precision	recall	f1_score
yolo11s-seg.pt	0.932	0.754	0.862	0.961	0.888
Updated-seg.pt	0.960	0.940	0.967	0.941	0.954

Μετά την επανεκπαίδευση στο νέο, πιο ποικιλόμορφο dataset και την εφαρμογή freeze στα κατώτερα επίπεδα του δικτύου, το ενημερωμένο μοντέλο (updated-seg.pt) παρουσιάζει σαφή βελτίωση σε όλες τις μετρικές: mAP0.5= 0.960, mAP0.5-0.95 = 0.940, precision = 0.967, recall = 0.941 και f1-score = 0.954. Η αύξηση του mAP0.5-0.95 κατά περίπου 25 ποσοστιαίες μονάδες είναι ιδιαίτερα σημαντική, καθώς υποδηλώνει πολύ πιο καθαρές, ακριβείς και σταθερές μάσκες, οι οποίες αντανακλούν το πραγματικό περίγραμμα του συρμού με υψηλή γεωμετρική συνέπεια.

4.5.2 Ποιοτική ανάλυση

4.5.2.1 Σενάριο 1

Η σύγκριση των δύο πειραματικών προσεγγίσεων αναδεικνύει ότι η δομή και ποικιλία του dataset καθώς και ο τρόπος προσαρμογής των βαρών του δικτύου αποτελούν κρίσιμους παράγοντες για την επίδοση του συστήματος. Στο πρώτο πείραμα, η εκπαίδευση βασίστηκε σε μεγάλο αλλά χαμηλής ποικιλίας σύνολο δεδομένων, αποτελούμενο από συνεχόμενα καρέ του ίδιου βίντεο. Σε αυτήν τη συνθήκη, το μοντέλο προσαρμόστηκε πολύ γρήγορα και έντονα στην οπτική δομή της συγκεκριμένης σκηνής. Η γρήγορη αυτή σύγκλιση είχε ως αποτέλεσμα να «αποστηθίσει» την εικόνα του συρμού και των ανθρώπων από τη συγκεκριμένη γωνία λήψης, χωρίς να μάθει χαρακτηριστικά που να γενικεύονται. Έτσι, εμφάνισε υψηλή απόδοση στο validation του ίδιου βίντεο, αλλά κατέρρευσε πλήρως όταν κλήθηκε να λειτουργήσει σε νέο περιβάλλον, επιβεβαιώνοντας τόσο υπερπροσαρμογή όσο και καταστροφική λήθη.

Από την αξιολόγηση στο validation set του πρώτου βίντεο έχει ενδιαφέρον να δείξουμε τις διαφορές στις ανίχνευσης για διάφορες εποχές μέσα από ένα επιλεγμένο δείγμα του dataset.

Στο Σχήμα 10 βλέπουμε στη πρώτη γραμμή τις επισημασμένες ετικέτες, στη δεύτερη τις ανιχνεύσεις του αρχικού μοντέλου που δεν έχει επανεκπαιδευτεί, όπου είναι με κόκκινο η ανίχνευση του συρμού, έτσι ώστε να είναι διακριτή από τις υπόλοιπες, στη τρίτη γραμμή, φαίνεται η ανίχνευση για 5 εποχές και στη τελευταία για 100 για το ίδιο σύνολο 4 εικόνων. Ενδιαφέρον παρουσιάζει το ότι το αρχικό μοντέλο (δεύτερη στήλη) δεν εντοπίζει τον άνθρωπο πάνω στο συρμό στο μακρινό πλάνο, ενώ στη μια που το εντοπίζει είναι σε κοντινό. Στις επισημασμένες ετικέτες, στους εντοπισμούς του μοντέλου 5 εποχών και 100 εποχών, παρατηρούμε ταύτιση ως προς τις επισημάνσεις, με διαφοροποίηση το ποσοστό βεβαιότητας (confidence score) όπου στις ετικέτες προφανώς δεν υπάρχει, στις 5 είναι πιο χαμηλό για κάποια πιο δυσδιάκριτα χωρία της εικόνας, ενώ στις 100 εποχές πλέον το μοντέλο εντοπίζει πιο άνετα τις ετικέτες.

Σχήμα 10: Σύγκριση μοντέλων πρώτου πειραματικού σεναρίου στο validation set

Όπως φαίνεται στο Σχήμα 11 στην αξιολόγηση του μοντέλου στο καινούργιο και λίγο διαφορετικό dataset επιλέξαμε 4 εικόνες σε κάθε στήλη, όπου στη πρώτη στήλη φαίνονται οι ετικέτες των εικόνων στη δεύτερη οι ανιχνεύσεις του μοντέλου των 5 εποχών, στη τρίτη των 10, στη τέταρτή των 20 και στη τελευταία το μοντέλο των 100 εποχών. Εδώ φαίνεται καθαρά η καταστροφική λήθη, δηλαδή ότι το μοντέλο στις 100 εποχές αδυνατεί να εντοπίσει το οτιδήποτε. Ακόμα και ο συρμός που είναι ελαφρώς πανομοιότυπος ως προς τη λήψη με το προηγούμενο dataset δεν εντοπίζεται καθόλου. Ενώ παρατηρούμε για τους ανθρώπους ότι το μοντέλο ήδη από τις 5 εποχές χάνει πολύ πληροφορία, δεν εντοπίζει σε 3 εικόνες τον άνθρωπο πάνω στο συρμό και δεν εντοπίζει κανέναν άνθρωπο εκτός συρμού. Στις 10 εποχές το μοντέλο εντοπίζει σε δύο εικόνες άνθρωπο πάνω από το συρμό, δηλαδή έναν παραπάνω από τις 5

εποχές, κάτι το οποίο δε θα περίμενε κανείς λόγω της σταδιακής μείωσης της επίδοσης των μοντέλων. Αυτό οφείλεται στον παράγοντα τυχαιότητας επιλογής 4 εικόνων. Στις 20 εποχές το μοντέλο εντοπίζει μόνο το συρμό, έχοντας χάσει πλήρως την ικανότητα να εντοπίζει ανθρώπους.

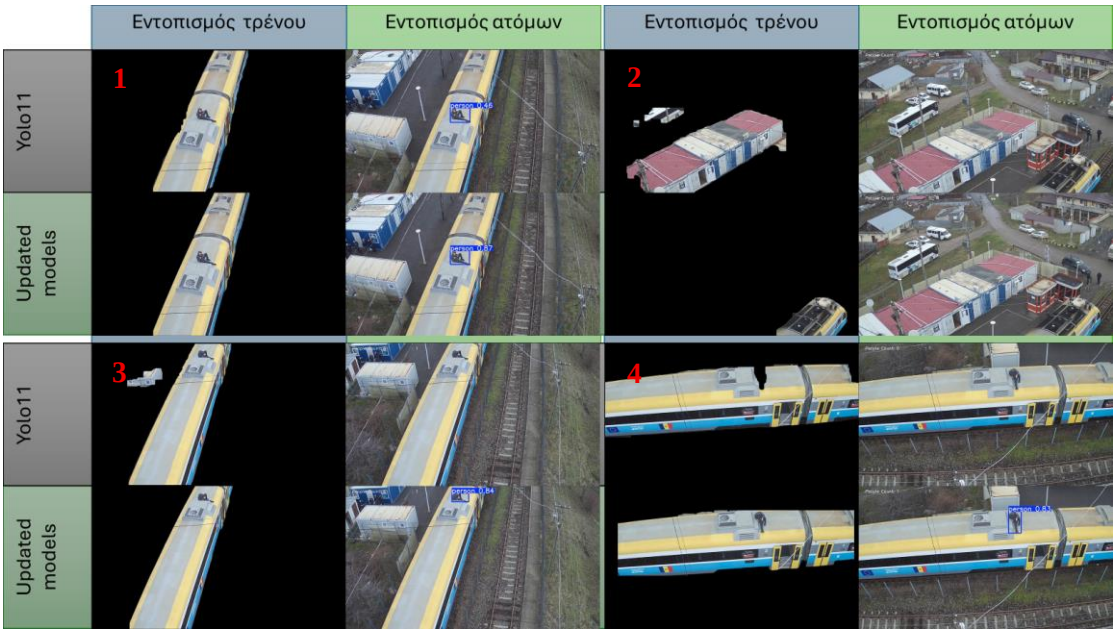
Σχήμα 11: Σύγκριση μοντέλων πρώτου πειραματικού σεναρίου στο

4.5.2.2 Σενάριο 2

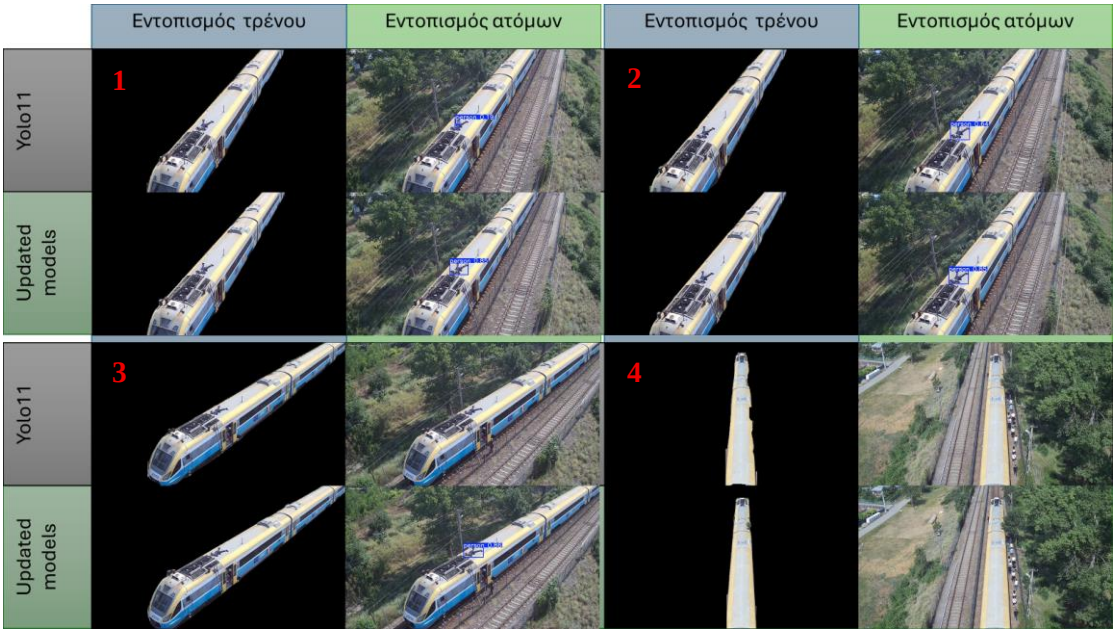
Στο δεύτερο πείραμα, η ίδια διαδικασία εκπαίδευσης σχεδιάστηκε εκ νέου με διαφορετικό dataset και παραμετροποίηση. Το dataset αποτέλεσε συνδυασμό δεδομένων δύο διαφορετικών οδηγών, με χρονικά αραιή δειγματοληψία ώστε να προκύψει μεγαλύτερη ποικιλία στις σκηνές, στις αποστάσεις και στις στάσεις του ανθρώπου. Επιπλέον, με την εφαρμογή freeze στο backbone, το μοντέλο διατήρησε τις βασικές αναπαραστάσεις που είχε μάθει από την αρχική εκπαίδευση και επικεντρώθηκε αποκλειστικά στην προσαρμογή των ανώτερων επιπέδων στα νέα δεδομένα. Ο χαμηλός ρυθμός μάθησης επέτρεψε σταδιακές, ελεγχόμενες αλλαγές, αποφεύγοντας την απότομη αναδιαμόρφωση των βαρών.

Το αποτέλεσμα ήταν η ανάπτυξη μοντέλου με σταθερότερη συμπεριφορά και σαφώς βελτιωμένη ικανότητα γενίκευσης. Το σύστημα δεν περιορίζεται πλέον σε μία συγκεκριμένη οπτική συνθήκη, αλλά διατηρεί την απόδοσή του και σε διαφορετική γωνία λήψης, φωτισμό και ανθρώπινη στάση. Η σημαντική αύξηση του mAP@0.5-0.95 στο segmentation επιβεβαιώνει την βελτίωση της ποιότητας των μασκών, η οποία είναι κρίσιμη για τη σωστή απομόνωση της περιοχής ενδιαφέροντος και την αποφυγή λανθασμένων ανιχνεύσεων ανθρώπων δίπλα στον συρμό. Αυτό αποτυπώνεται και στις εικόνες από το Σχήμα 12 και Σχήμα 13. Το κάθε σχήμα αποτυπώνει τέσσερεις εικόνες (καρέ) από το πρώτο και δεύτερο βίντεο

αντίστοιχα. Η κάθε εικόνα χωρίζεται σε τέσσερα μέρη. Το πάνω αριστερά μέρος της κάθε εικόνας είναι η έξοδος του πρώτου σταδίου (εντοπισμός τρένου και δημιουργία μάσκας) με τη χρήση του YOLO11s, το οποίο αποτελεί και την είσοδο του δεύτερου σταδίου. Το πάνω δεξιά μέρος είναι η έξοδος του δεύτερου σταδίου (εντοπισμός ατόμων εντός ορίων του τρένου). Αντίστοιχα έχουμε και για τα τμήματα κάτω αριστερά (εντοπισμός τρένου και δημιουργία μάσκας) και κάτω δεξιά (εντοπισμός ατόμων εντός ορίων του τρένου) αλλά πλέον με τη χρήση των νέων εκπαιδευμένων μοντέλων. Όπως φαίνεται στο Σχήμα 12 και εικόνα 1, και τα 2 μοντέλα εντοπίζουν το τρένο και δημιουργούν τη μάσκα με αρκετή ακρίβεια. Επίσης εντοπίζουν και το άτομο εντός των ορίων του τρένου. Η διαφορά είναι ότι το εκπαιδευμένο μοντέλο τον εντοπίζει με μεγαλύτερο ποσοστό εμπιστοσύνης σε σχέση με το αρχικό μοντέλο (87% έναντι 46%). Στο ίδιο σχήμα (Σχήμα 12) και εικόνες 3 και 4, το αρχικό μοντέλο τμηματοποίησης (YOLO11s-seg) αποτυγχάνει να αποτυπώσει με ακρίβεια το περίγραμμα του τρένου: στην περίπτωση της εικόνας 3 περιλαμβάνει στο περίγραμμα του τρένου και μέρος του container ενώ στην εικόνα 4 μέρος του τρένου χάνεται από το περίγραμμα. Τα πράγματα είναι ακόμη χειρότερα στην εικόνα 2, όπου όχι μόνο αστοχεί να βρει το τρένο, αλλά αναγνωρίζει ως τρένο τη συστοιχία των container με ένα μέρος του λεωφορείου. Από την άλλη πλευρά, το νέο μοντέλο τμηματοποίησης αναγνωρίζει το τρένο και σχηματίζει το περίγραμμά του με αρκετά καλή ακρίβεια. Αντίστοιχα το νέο μοντέλο εντοπισμού αναγνωρίζει τον άνθρωπο στις εικόνες 3 και 4 από το Σχήμα 12. Τέλος έντονο ενδιαφέρον έχει η εικόνα 4, όπου το μοντέλο εντοπισμού (YOLO11s) δεν μπορεί να εντοπίσει τον άνθρωπο διότι ήδη το σημείο που βρίσκεται ο άνθρωπος είναι εκτός μάσκας. Ουσιαστικά είναι αδυναμία που προκύπτει από το μοντέλο τμηματοποίησης.



Σχήμα 12: Δεύτερο σενάριο – Σύγκριση μοντέλων σε (4) εικόνες από το πρώτο βίντεο.



Σχήμα 13: Δεύτερο σενάριο – Σύγκριση μοντέλων σε (4) εικόνες από το δεύτερο βίντεο.

Προχωρώντας στο Σχήμα 13, μπορούμε να δούμε στιγμιότυπα από το δεύτερο βίντεο. Το Σχήμα 13 είναι δομημένο με τον ίδιο τρόπο όπως το Σχήμα 12. Εδώ δεν παρατηρήσουμε μεγάλες διαφορές στο αρχικό και στο εκπαιδευμένο μοντέλο τμηματοποίησης, αν και το νέο μοντέλο αποτυπώνει με μεγαλύτερη ακρίβεια τα όρια του τρένου και κυρίως στην εικόνα 4. Εδώ όμως μπορούμε να παρατηρήσουμε διαφορές στο στάδιο του εντοπισμού των ατόμων όπου το νέο μοντέλο αναγνωρίζει τον άνθρωπο με μεγαλύτερο ποσοστό εμπιστοσύνης (85%, 85% και 86% για τις εικόνες 1, 2 και 3 αντίστοιχα) σε σχέση με το αρχικό μοντέλο (19% και 64% για τις εικόνες 1 και 2 αντίστοιχα) ενώ στην εικόνα 3 αποτυγχάνει να τον εντοπίσει. Τέλος στην εικόνα 4 κανένα μοντέλο δεν καταφέρνει να αναγνωρίσει τον άνθρωπο.

Συνολικά, το δεύτερο πείραμα επιβεβαιώνει ότι η ποικιλία των δεδομένων και η ελεγχόμενη προσαρμογή του μοντέλου αποτελούν τον καθοριστικό παράγοντα για τη λειτουργική αξιοπιστία του συστήματος σε πραγματικές συνθήκες πεδίου. Με άλλα λόγια, η βελτίωση δεν επιτυγχάνεται με περισσότερη εκπαίδευση, αλλά με καλύτερη εκπαίδευση.

4.5.3 Συζήτηση ευρημάτων

Η σύγκριση των δύο πειραματικών σεναρίων δείχνει με σαφήνεια ότι η απόδοση του συστήματος εξαρτάται περισσότερο από την ποικιλία του dataset και τον τρόπο προσαρμογής των βαρών του μοντέλου, παρά από την ποσότητα της εκπαίδευσης. Στο πρώτο πείραμα, όπου η εκπαίδευση βασίστηκε σε συνεχόμενα καρέ του ίδιου βίντεο, παρατηρήθηκε ότι το μοντέλο σημείωσε εντυπωσιακή βελτίωση στο validation set του αρχικού περιβάλλοντος, με το

mAP@0.5 να αυξάνεται από 0.629 σε 0.976 μέσα στις πρώτες τέσσερις εποχές. Η αύξηση αυτή όμως δεν συνοδεύτηκε από αντίστοιχη ικανότητα γενίκευσης. Όταν το μοντέλο αξιολογήθηκε στο ανεξάρτητο σύνολο ελέγχου (subdataset12), η απόδοση μειώθηκε σταδιακά καθώς αυξάνονταν οι εποχές, καταλήγοντας σε πλήρη κατάρρευση στις εκατό εποχές (mAP ≈ 0.03). Η πτώση του AP_person από 0.302 σε σχεδόν 0.0 επιβεβαιώνει ότι το μοντέλο έχασε την ικανότητα αναγνώρισης ανθρώπων σε διαφοροποιημένες συνθήκες. Η συμπεριφορά αυτή αποτελεί χαρακτηριστική ένδειξη υπερπροσαρμογής και καταστροφικής λήθης, καθώς το μοντέλο «αποστήθισε» τη συγκεκριμένη σκηνή του αρχικού βίντεο, χωρίς να αναπτύξει γενικεύσιμα οπτικά χαρακτηριστικά.

Αντιθέτως, στο δεύτερο πείραμα, όπου το dataset αποτέλεσε συνδυασμό οπτικά διαφορετικών σκηνών και εφαρμόστηκε μερικό πάγωμα των κατώτερων επιπέδων του δικτύου σε συνδυασμό με χαμηλό ρυθμό μάθησης, η απόδοση βελτιώθηκε σημαντικά και σταθερά. Το ενημερωμένο μοντέλο εντοπισμού ατόμων παρουσίασε αύξηση του mAP@0.5 από 0.651 σε 0.852 και του recall από 0.590 σε 0.824, γεγονός που αντικατοπτρίζεται και στην άνοδο του f1-score από 0.668 σε 0.853. Παράλληλα, στο στάδιο τμηματοποίησης, η βελτίωση της γεωμετρικής ακρίβειας της μάσκας του συρμού ήταν ιδιαίτερα έντονη, καθώς το mAP@0.5–0.95 αυξήθηκε από 0.754 σε 0.940. Η αύξηση αυτή είναι κρίσιμη, δεδομένου ότι η ποιότητα της μάσκας καθορίζει την ακριβή οριοθέτηση της περιοχής ενδιαφέροντος για τον εντοπισμό των ατόμων και επηρεάζει την τελική λειτουργική αξιοπιστία του συστήματος.

Τα παραπάνω αποτελέσματα δείχνουν ότι η βελτίωση της απόδοσης δεν συνδέεται με την αύξηση των εποχών εκπαίδευσης, αλλά με τον τρόπο και το περιεχόμενο της εκπαίδευσης. Η χρήση μεγαλύτερης ποικιλίας οπτικών συνθηκών και η σταδιακή, ελεγχόμενη προσαρμογή των βαρών επιτρέπουν στο μοντέλο να διατηρεί χρήσιμες γενικεύσιμες αναπαραστάσεις, αποφεύγοντας την υπερπροσαρμογή σε συγκεκριμένες σκηνές. Συνεπώς, η ποιότητα του dataset και η στοχευμένη προσαρμογή των βαρών προκύπτουν ως οι δύο κρίσιμοι παράγοντες για την επιτυχή γενίκευση των μοντέλων σε πραγματικές συνθήκες πεδίου.

5

Συμπεράσματα & Μελλοντικές Εργασίες

5.1 Συμπεράσματα

Οι μετρικές αξιολόγησης έδειξαν ότι τα διαθέσιμα προεκπαιδευμένα μοντέλα YOLO δεν αποδίδουν ικανοποιητικά στη συγκεκριμένη εφαρμογή, καθώς δεν είναι προσαρμοσμένα στα ιδιαίτερα οπτικά χαρακτηριστικά των εναέριων λήψεων και της γεωμετρίας της σκηνής. Η εφαρμογή επανεκπαίδευσης κρίθηκε επομένως απαραίτητη, επιτρέποντας στο μοντέλο να προσαρμοστεί στο συγκεκριμένο πεδίο και να βελτιώσει σημαντικά την απόδοσή του.

Στοιχείο της μεθοδολογίας αποτέλεσε η ημιαυτόματη επισημείωση των δεδομένων, η οποία αξιοποίησε υπάρχοντα μοντέλα για την αρχική παραγωγή ετικετών και στη συνέχεια ανθρώπινη επιμέλεια για τη διόρθωσή τους. Η διαδικασία αυτή συνέβαλε στη διευκόλυνση της δημιουργίας του dataset, διατηρώντας παράλληλα την ποιότητα των επισημάνσεων.

Η συγκριτική αξιολόγηση βασίστηκε σε δύο πειραματικά σενάρια. Στο πρώτο σενάριο, όπου το σύνολο δεδομένων παρουσίαζε μικρή ποικιλία και σταθερές συνθήκες λήψης, το μοντέλο εμφάνισε υπερπροσαρμογή, με περιορισμένη δυνατότητα γενίκευσης σε νέες εικόνες. Στο δεύτερο σενάριο, η χρήση ενός πιο ποικιλόμορφου και χρονικά αραιωμένου dataset, σε συνδυασμό με προσεκτική παραμετροποίηση της διαδικασίας επανεκπαίδευσης (όπως το μερικό freeze του backbone), οδήγησε σε βελτίωση των μοντέλων. Με αυτόν τον τρόπο αναδείχθηκε ότι ο κρίσιμος παράγοντας δεν είναι ο απόλυτος αριθμός των δεδομένων, αλλά η ποικιλία και η αντιπροσωπευτικότητά τους.

Συνολικά, η εργασία καταδεικνύει ότι η ποιότητα και η ποικιλία των δεδομένων, σε συνδυασμό με τη σωστή παραμετροποίηση της διαδικασίας επανεκπαίδευσης, αποτελούν καθοριστικούς παράγοντες για την επίτευξη αξιόπιστης και σταθερής απόδοσης σε συστήματα καταμέτρησης πλήθους που λειτουργούν σε πραγματικές συνθήκες.

5.2 Μελλοντικές Εργασίες

Μια κατεύθυνση για μελλοντική έρευνα αφορά την εξαντλητική διερεύνηση των παραμέτρων της επανεκπαίδευσης, με στόχο την πιο ακριβή κατανόηση της επίδρασης κάθε επιλογής στη συμπεριφορά του μοντέλου. Η εφαρμογή μιας συστηματικής ανάλυσης ευαισθησίας (sensitivity analysis) σε παραμέτρους όπως ο ρυθμός μάθησης, ο βαθμός παγώματος του backbone, το batch size κα άλλοι παράμετροι μπορεί να οδηγήσει στον εντοπισμό βέλτιστων σημείων λειτουργίας για διαφορετικές συνθήκες εισόδου. Μια τέτοια ανάλυση θα επιτρέψει όχι μόνο τη βελτίωση της απόδοσης, αλλά και τη βαθύτερη κατανόηση των μηχανισμών προσαρμογής του μοντέλου στο συγκεκριμένο είδος δεδομένων.

Επιπλέον, μια σημαντική μελλοντική εξέλιξη αποτελεί η διερεύνηση τεχνικών continual learning, ώστε το σύστημα να μπορεί να ενημερώνει σταδιακά τις παραμέτρους του καθώς συλλέγονται νέα δεδομένα από διαφορετικά περιβάλλοντα λειτουργίας. Με τον τρόπο αυτό, το μοντέλο μπορεί να βελτιώνει συνεχώς την απόδοσή του χωρίς να εμφανίζει το φαινόμενο της καταστροφικής λήθης. Η προσέγγιση αυτή θα επιτρέψει τη μακροχρόνια διατήρηση της αξιοπιστίας του συστήματος σε πραγματικές επιχειρησιακές συνθήκες όπου οι οπτικές και λειτουργικές παράμετροι μεταβάλλονται δυναμικά.

Μια εναλλακτική προσέγγιση είναι η εξής διαδικασία επανεκπαίδευσης: Πρώτα εφαρμόζεται το μοντέλο τμηματοποίησης για να παραχθεί η μάσκα του συρμού, από την οποία προκύπτει η περιοχή ενδιαφέροντος (ROI). Έπειτα, η ανίχνευση ανθρώπων εκτελείται μόνο μέσα στο ROI, είτε μέσω μάσκας είτε μέσω αποκοπής της εικόνας.

Με αυτόν τον τρόπο περιορίζεται η αναζήτηση σε μια καλά ορισμένη περιοχή, μειώνοντας τα false positives από άτομα ή αντικείμενα που βρίσκονται εκτός συρμού. Επιπλέον, η αξιολόγηση του detection μπορεί επίσης να γίνεται μόνο εντός του ROI, ώστε οι μετρικές (mAP, precision, recall) να αντανακλούν αποκλειστικά την απόδοση στην πραγματική περιοχή ενδιαφέροντος. Αυτό καθιστά την αξιολόγηση πιο ρεαλιστική και συνεπή με τον τελικό στόχο της εφαρμογής.

6

Παράρτημα

6.1 Ανάλυση Υλικοτεχνικής Υποδομής & Ρύθμιση Jetson Orin Nano για YOLO

Η επιτυχής υλοποίηση του συστήματος καταμέτρησης πλήθους απαιτεί κατάλληλη υλικοτεχνική υποδομή, τόσο στο edge (επεξεργασία βίντεο σε πραγματικό χρόνο) όσο και σε server (επανεκπαίδευση μοντέλων). Στην παρούσα εργασία χρησιμοποιείται η ενσωματωμένη πλατφόρμα NVIDIA Jetson Orin Nano για επιτόπια επεξεργασία εικόνας/βίντεο και ένας τοπικός εξυπηρετητής με GPU για την επανεκπαίδευση των YOLOv11 και MobileNet. Το Orin Nano, νεότερης γενιάς σύστημα edge AI (εξέλιξη των Jetson Nano/Xavier NX), προσφέρει έως ~40 TOPS σε συμπαγές και ενεργειακά αποδοτικό form factor, κατάλληλο για ρομποτική, βιομηχανική αυτοματοποίηση, παρακολούθηση και real-time video analytics. Σε επίπεδο υλικού αξιοποιεί αρχιτεκτονική NVIDIA Ampere με 512 CUDA cores και 16 Tensor Cores, και CPU ARM Cortex-A78AE (6 πυρήνες), ενώ συνοδεύεται από 4GB ή 8GB LPDDR5 για υψηλό ρυθμό μεταφοράς μεταξύ CPU/GPU/I/O. Σε επίπεδο λογισμικού λειτουργεί με JetPack SDK 6.0/6.1, το οποίο περιλαμβάνει CUDA Toolkit, cuDNN, TensorRT, OpenCV και υποστήριξη για PyTorch/TensorFlow, επιτρέποντας απαιτητικές εφαρμογές deep learning χωρίς ανάγκη απομακρυσμένης επεξεργασίας δεδομένων.

Η ρύθμιση του Jetson Orin Nano ξεκινά από την προετοιμασία αποθηκευτικού μέσου, την εγκατάσταση JetPack και συνεχίζει έως τη διαμόρφωση περιβάλλοντος PyTorch/Ultralytics και εξαγωγών TensorRT. Η εγκατάσταση βασίζεται στα JetPack 6.0/6.1 με πλήρη υποστήριξη CUDA, cuDNN, PyTorch 2.5 και ONNX Runtime για ARM64. Πρώτο βήμα είναι το flashing της εικόνας JetPack 6.0 (Ubuntu 20.04 LTS) σε microSD από το Jetson Download Center (π.χ. με balenaEtcher) και η αρχική ρύθμιση (χρήστης, γλώσσα, δίκτυο). Στη συνέχεια, μέσω NVIDIA SDK Manager (NVIDIA Developer Zone) επιλέγεται η σωστή έκδοση JetPack 6.0/6.1

και εγκαθίστανται τα πακέτα CUDA Toolkit, cuDNN, TensorRT, OpenCV, Python κ.ά., με αυτόματο εντοπισμό της συσκευής και καθοδηγούμενη διαδικασία μέχρι την ολοκλήρωση.

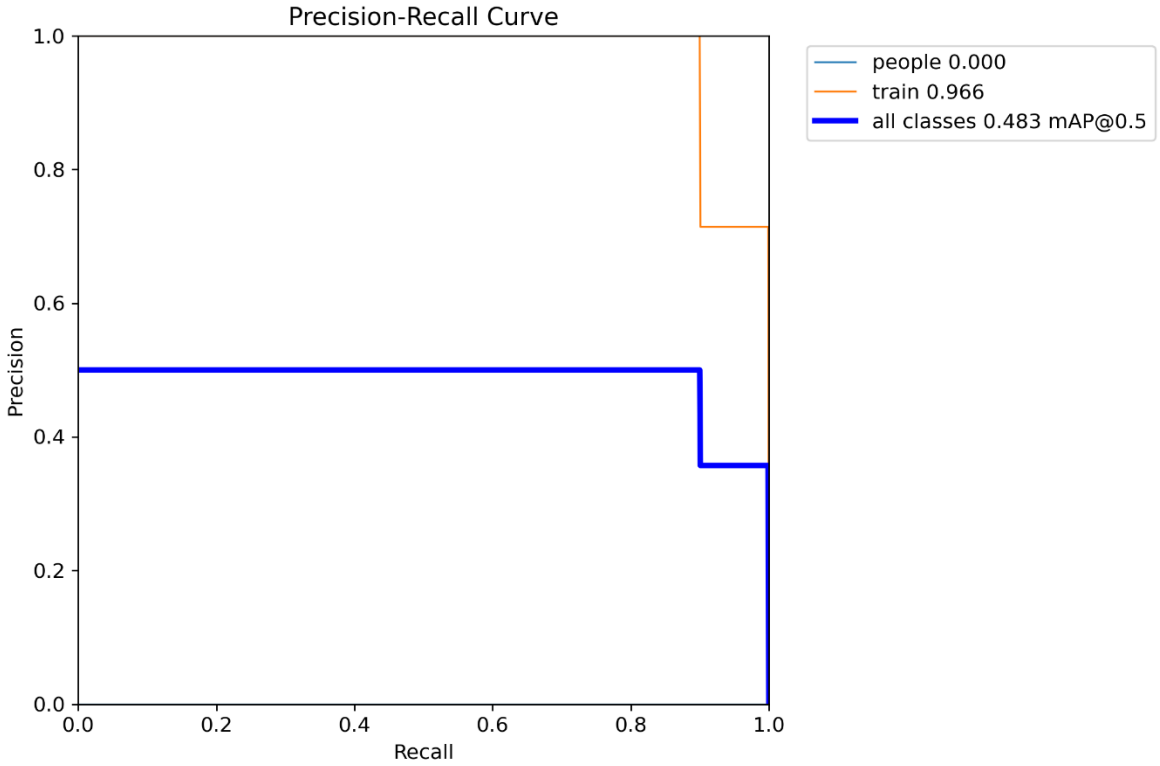
Μετά την εγκατάσταση του JetPack, απαιτείται επιβεβαίωση λειτουργίας CUDA και συμβατότητας του PyTorch με την εγκατεστημένη CUDA 12.x (π.χ. έλεγχος torch.cuda.is_available()). Επειδή οι προεπιλεγμένες εκδόσεις pip δεν είναι πάντα συμβατές με ARM64, γίνεται χειροκίνητη εγκατάσταση προκατασκευασμένων πακέτων PyTorch 2.5.0 και TorchVision 0.20.0 για Python 3.10, βελτιστοποιημένων για Jetson/JetPack 6.1 (με επιτάχυνση CUDA και συμβατότητα TensorRT). Με αυτόν τον συνδυασμό το Ultralytics YOLO μπορεί να εκπαιδεύει/τρέχει μοντέλα και να τα εξάγει σε ONNX/TensorRT, επιτρέποντας real-time αναλύσεις στην άκρη, ενώ ο GPU server αναλαμβάνει τα πιο βαριά workloads fine-tuning. Έτσι, το Jetson Orin Nano λειτουργεί ως αυτόνομος επεξεργαστής εικόνας/βίντεο για την εφαρμογή καταμέτρησης πλήθους, με υψηλή ακρίβεια και χαμηλή κατανάλωση, ενώ η υποδομή server εξασφαλίζει γρήγορες επανεκπαιδεύσεις και βελτιστοποιήσεις των μοντέλων.

6.2 Επεξήγηση Διαγραμμάτων Αξιολόγησης Μοντέλου

Η αξιολόγηση ενός μοντέλου αποσκοπεί στο να εκτιμήσει πόσο καλά γενικεύει, δηλαδή πόσο αξιόπιστα μπορεί να εντοπίζει αντικείμενα σε ανεξάρτητα, μη γνωστά δεδομένα. Στην ανίχνευση αντικειμένων, οι προβλέψεις του μοντέλου συγκρίνονται με τις πραγματικές θέσεις (ground truth) μέσω ενός κατωφλίου επικάλυψης (IoU threshold). Από αυτή τη σύγκριση υπολογίζονται οι δείκτες ακρίβειας (precision), ανάκλησης (recall) και ο αρμονικός τους μέσος, το F1-score, που εκφράζει τη συνολική ισορροπία μεταξύ των δύο. Με τη σάρωση διαφορετικών τιμών του κατωφλίου προκύπτουν οι καμπύλες precision–recall, των οποίων το εμβαδό (Average Precision, AP) συνοψίζει την απόδοση κάθε κατηγορίας. Ο μέσος όρος όλων των AP σε πολλαπλά κατώφλια (από 0.5 έως 0.95 IoU) δίνει τον δείκτη mAP@[.5:.95], ο οποίος θεωρείται το πιο ολοκληρωμένο και αξιόπιστο μέτρο συνολικής επίδοσης.

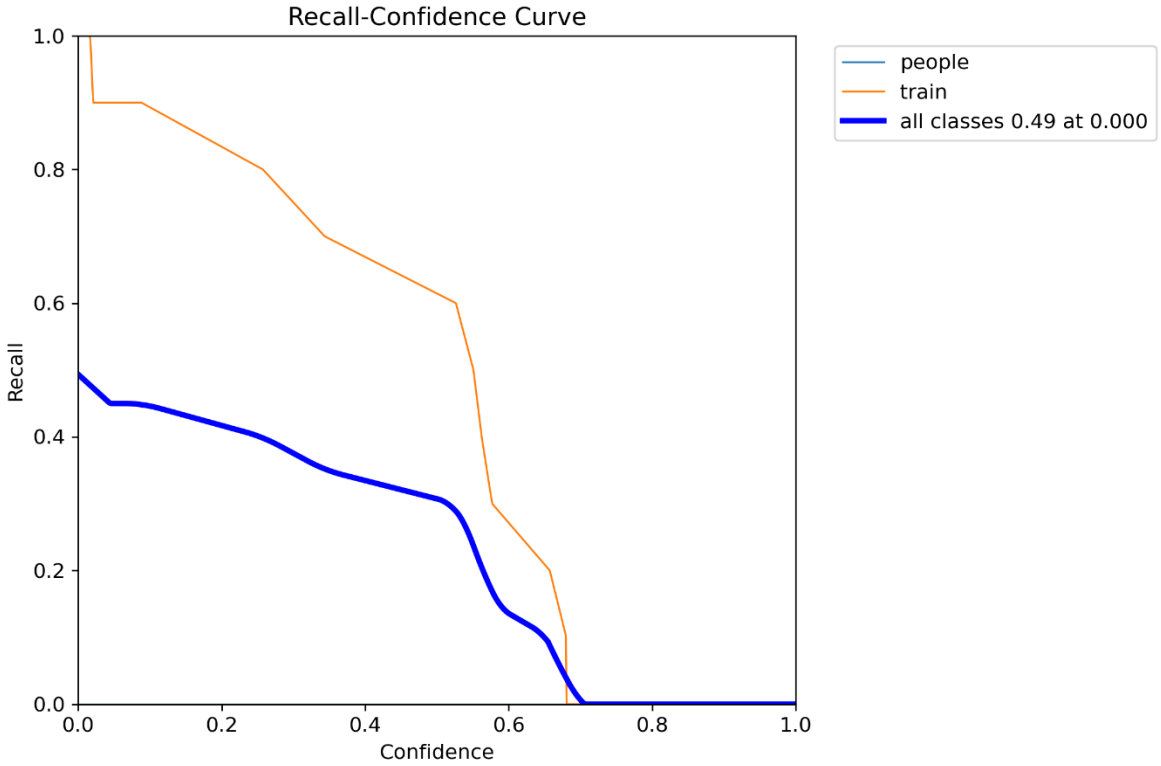
Στην τμηματοποίηση αντικειμένων, η αξιολόγηση πραγματοποιείται σε επίπεδο εικονοστοιχείων, μετρώντας τον βαθμό επικάλυψης μεταξύ των προβλεπόμενων και των πραγματικών μασκών. Έτσι προκύπτουν οι δείκτες AP^mask και mAP^mask, που αποτυπώνουν την ακρίβεια των προβλέψεων σε επίπεδο pixel. Παράλληλα, οι καμπύλες precision–confidence και recall–confidence βοηθούν στην επιλογή του κατάλληλου λειτουργικού σημείου του μοντέλου, ανάλογα με το ζητούμενο προφίλ λειτουργίας — για παράδειγμα, αν προτιμάται υψηλή ανάκληση (πληρότητα εντοπισμών) ή υψηλή ακρίβεια (περιορισμός ψευδών θετικών). Ο πίνακας σφαλμάτων ανά κατηγορία, τόσο σε απόλυτους αριθμούς όσο και ομαλοποιημένος, αναδεικνύει συστηματικές συγχύσεις μεταξύ κατηγοριών και καθοδηγεί τις στοχευμένες βελτιώσεις του μοντέλου.

Τα διαγράμματα αυτά λειτουργούν σαν «χάρτης» της συμπεριφοράς του μοντέλου σε όλο το εύρος των κατωφλιών εμπιστοσύνης και βοηθούν να καταλάβουμε όχι μόνο αν είναι καλό, αλλά και πότε είναι καλό. Το Precision–Recall δείχνει την ανταλλαγή ανάμεσα στην καθαρότητα των προβλέψεων και την κάλυψη των στόχων: όσο η καμπύλη αγγίζει την πάνω δεξιά γωνία, τόσο πετυχαίνουμε ταυτόχρονα λίγα ψευδώς θετικά και λίγα χαμένα αντικείμενα.



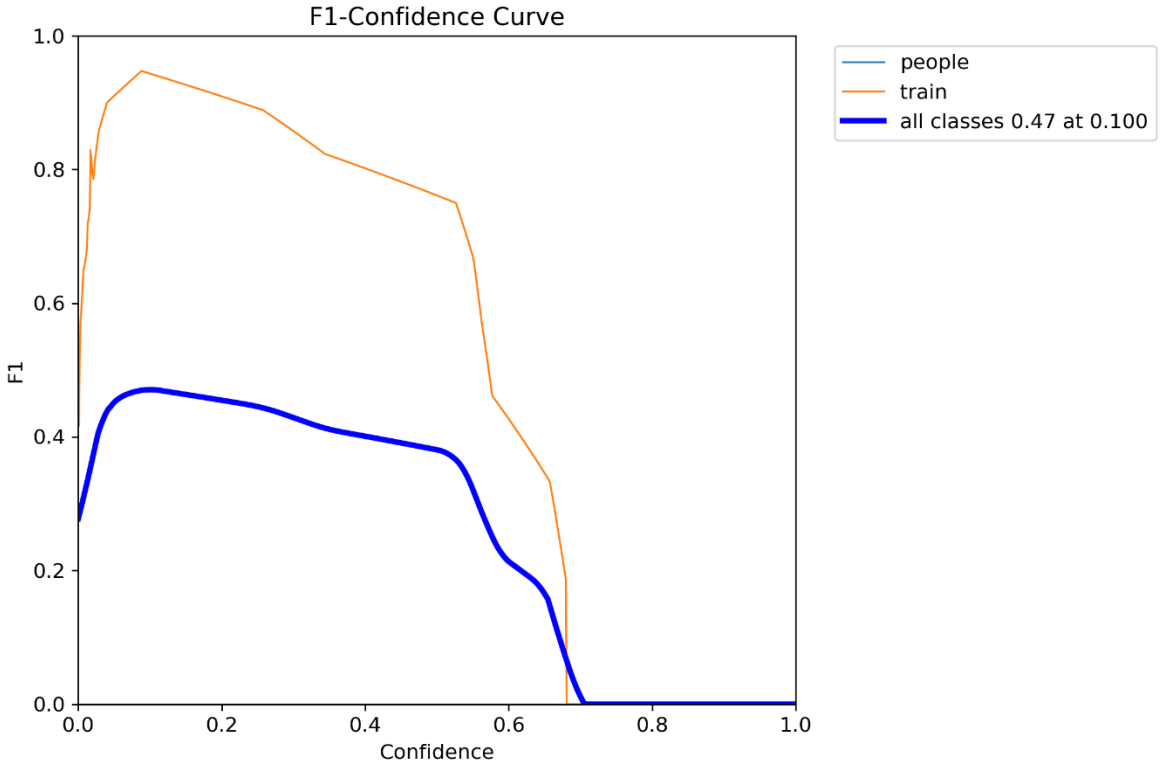
Σχήμα 14: Ενδεικτικό διάγραμμα Precision-Recall Curve

Η καμπύλη Recall–Confidence απεικονίζει τη μεταβολή της ανάκτησης σε σχέση με την αύξηση του κατωφλίου εμπιστοσύνης. Σε χαμηλές τιμές κατωφλίου, το μοντέλο παράγει μεγάλο αριθμό ανιχνεύσεων, ενώ η αυστηροποίηση του κατωφλίου οδηγεί σε απόρριψη ανιχνεύσεων με χαμηλή εμπιστοσύνη και συνεπακόλουθη μείωση της ανάκτησης. Μια αργή μείωση της ανάκτησης υποδεικνύει ότι το μοντέλο διατηρεί υψηλή απόδοση ακόμη και υπό αυστηρά κατώφλια, ενώ μια απότομη πτώση δείχνει ότι μεγάλο ποσοστό ορθών ανιχνεύσεων αντιστοιχεί σε μέτρια επίπεδα εμπιστοσύνης. Συνήθως, οι κλάσεις με σαφέστερα ή μεγαλύτερα αντικείμενα παρουσιάζουν υψηλή ανάκτηση σε υψηλά κατώφλια, ενώ οι πιο δύσκολες κλάσεις (π.χ. μικρά ή μερικώς καλυμμένα αντικείμενα) εμφανίζουν ταχύτερη μείωση.



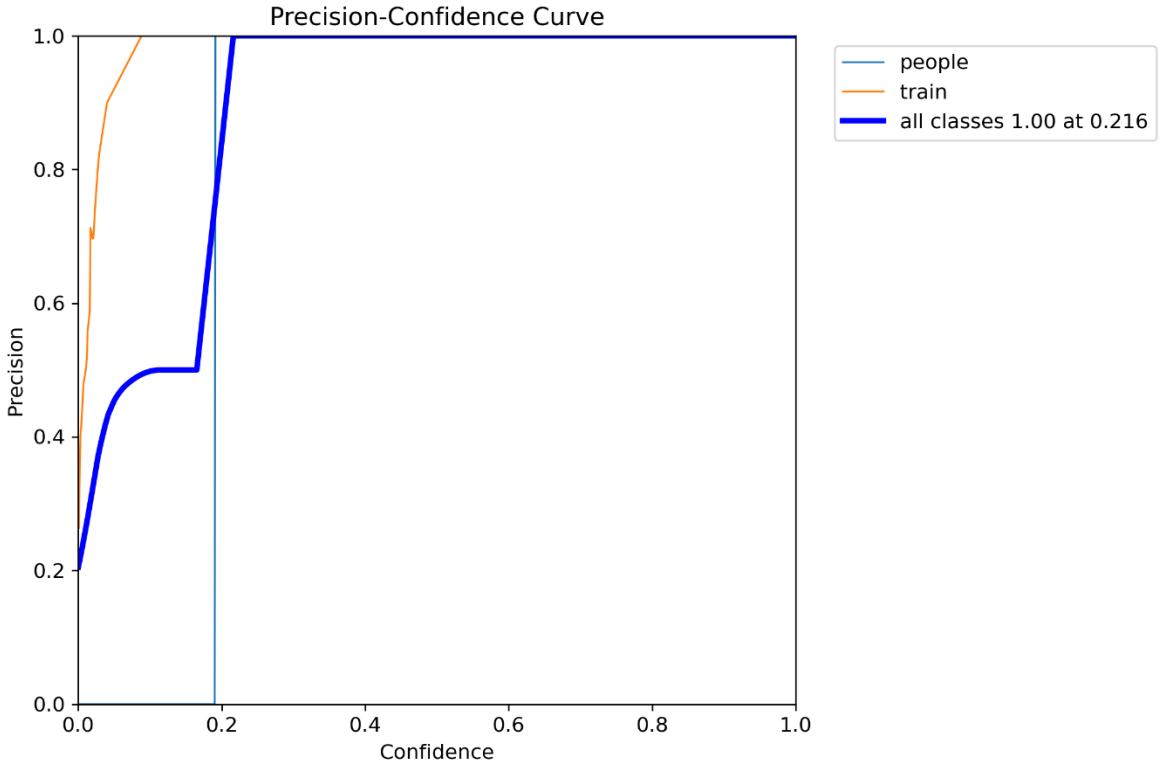
Σχήμα 15: Ενδεικτικό διάγραμμα Recall-Condifence Curve

Στο F1–Confidence αποτυπώνεται το σημείο που ισορροπείται η ακρίβεια και η ανάκτηση. Αν η καμπύλη σχηματίζει ένα πλατύ «πλατώ», υπάρχει ευελιξία στην επιλογή κατωφλίου χωρίς να χάνεται απόδοση· αν κορυφώνει στενά, η επιλογή κατωφλίου γίνεται πιο κρίσιμη.



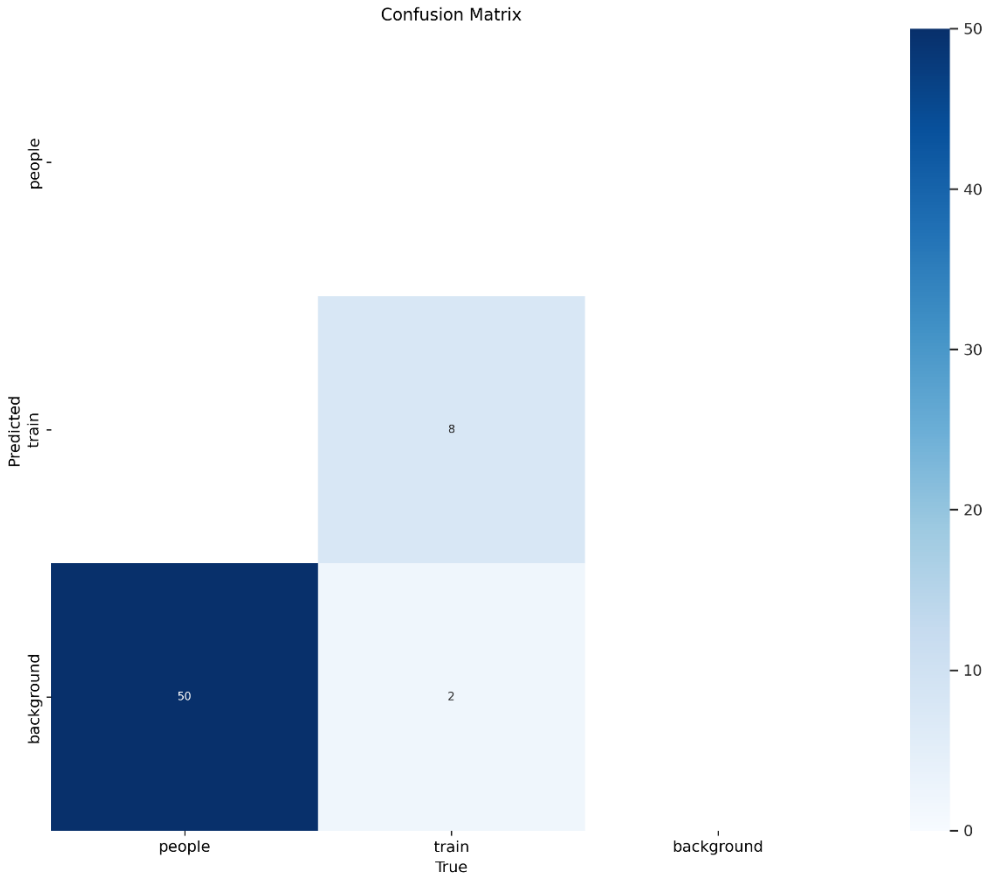
Σχήμα 16: Ενδεικτικό διάγραμμα F1-Confidence Curve

Το Precision–Confidence εξηγεί γιατί συμβαίνουν τα παραπάνω: όσο ανεβάζουμε το κατώφλι, κρατάμε μόνο τις πιο βέβαιες προβλέψεις και η ακρίβεια αυξάνει. Αυτό είναι θεμιτό όταν προέχει η αποφυγή ψευδών συναγερμών (π.χ. σε συστήματα ειδοποίησης), αλλά συνοδεύεται από απώλειες στην ανάκτηση. Αν η ακρίβεια φτάνει πολύ γρήγορα σε πολύ υψηλές τιμές, σημαίνει ότι μπορούμε να πετύχουμε έξοδους χωρίς σφάλματα ακόμη και με σχετικά χαλαρά thresholds· αν ανεβαίνει αργά, χρειάζεται πιο αυστηρή ρύθμιση για να καθαρίσουν τα σφάλματα.

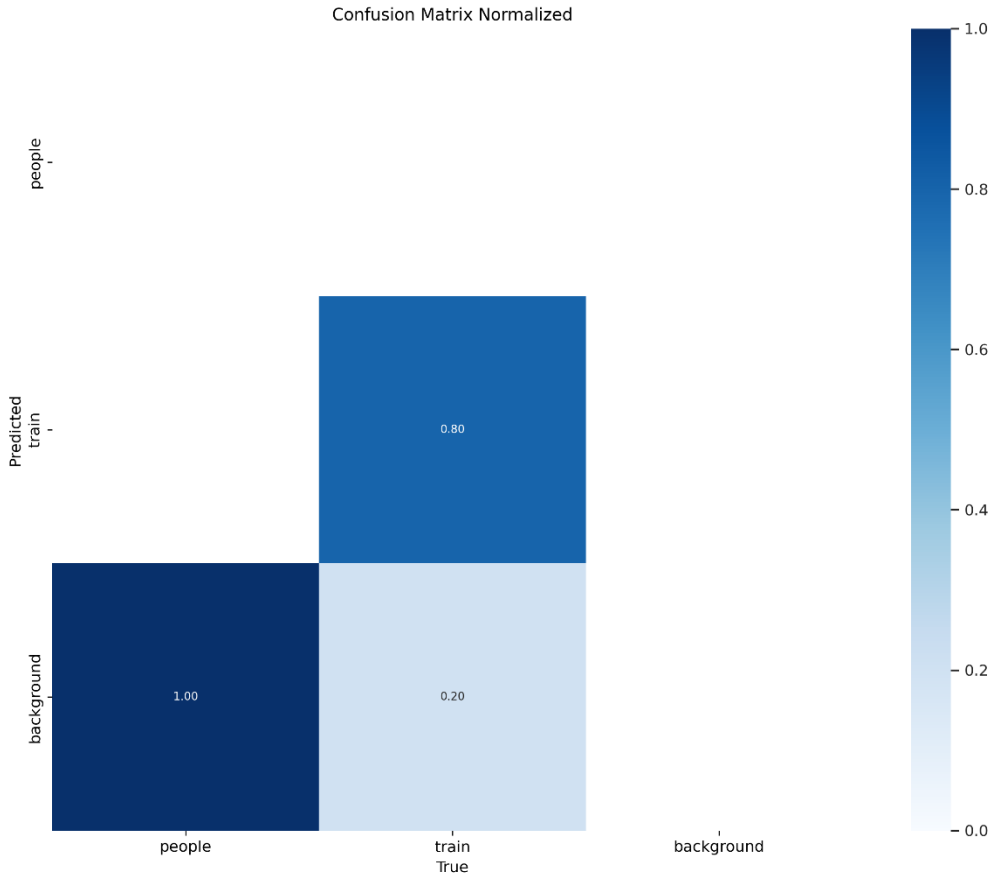


Σχήμα 17: Ενδεικτικό διάγραμμα Precision-Confidence Curve

Τέλος, τα confusion matrices συγκεντρώνουν σε μία εικόνα το πού γίνονται τα λάθη. Η «διαγώνιος» αποτυπώνει τα σωστά, ενώ τα εκτός διαγωνίου κελιά δείχνουν πού μπερδεύονται οι κλάσεις μεταξύ τους ή με το υπόβαθρο. Ένα έντονο τετράγωνο πάνω στη διαγώνιο σημαίνει ότι η αντίστοιχη κλάση αναγνωρίζεται αξιόπιστα· επισήμανση προς τη γραμμή του background σημαίνει ότι χάνουμε πραγματικά αντικείμενα (χαμηλή ανάκτηση), ενώ επισήμανση από το background προς μια κλάση σημαίνει ψευδώς θετικά (χαμηλή καθαρότητα). Όταν οι δύο κλάσεις δεν μπερδεύονται μεταξύ τους αλλά μια από τις δύο εμφανίζει πολλά λάθη με το background, το ζήτημα είναι κυρίως ανίχνευση (να τη βρούμε) και λιγότερο διάκριση (να τη ξεχωρίσουμε από την άλλη κλάση).



Σχήμα 18: Ενδεικτικό confusion matrix normalized



Σχήμα 19: Ενδεικτικό confusion matrix

Συνολικά, ο συνδυασμός των πέντε γραφημάτων μάς εξηγεί αν το μοντέλο είναι αξιόπιστο, σε ποια ζώνη κατωφλιών αποδίδει καλύτερα, ποια κλάση είναι «απαιτητική» ή «όχι», και αν τα λάθη οφείλονται περισσότερο σε χαμένες ανιχνεύσεις ή σε ψευδώς θετικά. Με αυτά, μπορούμε να επιλέξουμε σωστά το λειτουργικό κατώφλι, να αποφασίσουμε αν χρειαζόμαστε διαφορετικά thresholds ανά κλάση και να κατευθύνουμε στοχευμένες βελτιώσεις στα δεδομένα ή στο fine-tuning.

7

Βιβλιογραφία

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2012.

[2] K. Simonyan and A. Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," Apr. 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

[4] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection." Available: https://openaccess.thecvf.com/content_cvpr_2017/papers/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.pdf

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf

[6] N. Bodla, B. Singh, R. Chellappa, and L. Davis, "Soft-NMS -Improving Object Detection With One Line of Code." Available: https://openaccess.thecvf.com/content_ICCV_2017/papers/Bodla_Soft-NMS_--_Improving_ICCV_2017_paper.pdf

[7] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression.” Available: https://openaccess.thecvf.com/content_CVPR_2019/papers/Rezatofighi_Generalized_Intersection_Over_Union_A_Metric_and_a_Loss_for_CVPR_2019_paper.pdf

[8] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression,” *arXiv:1911.08287 [cs]*, Nov. 2019, Available: https://arxiv.org/abs/1911.08287

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network,” 2016. Available: https://openaccess.thecvf.com/content_cvpr_2016/papers/Zhang_Single-Image_Crowd_Counting_CVPR_2016_paper.pdf

[10] V. Lempitsky and A. Zisserman, “Learning To Count Objects in Images.” Accessed: Nov. 16, 2025. [Online]. Available: https://papers.neurips.cc/paper/4043-learning-to-count-objects-in-images.pdf

[11] R. Wang, F. Tan, K. Yan, Y. Hao, F. Li, and X. Yu, “LCNNet: Light-weight convolutional neural networks for understanding the highly congested scenes,” *Journal of Intelligent & Fuzzy Systems*, pp. 1–14, May 2023, doi: https://doi.org/10.3233/jifs-224081.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” *Computer Vision – ECCV 2014*, pp. 346–361, 2014, doi: https://doi.org/10.1007/978-3-319-10578-9_23.

[13] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. (2018). *Path Aggregation Network for Instance Segmentation.* CVPR.

[14] “TensorRT Documentation — NVIDIA TensorRT Documentation,” *Nvidia.com*, 2021. https://docs.nvidia.com/deeplearning/tensorrt/latest/index.html

[15] “NVIDIA cuDNN — NVIDIA cuDNN,” *Nvidia.com*, 2024. https://docs.nvidia.com/deeplearning/cudnn/latest/ [16] “Jetson Orin Nano Super Developer Kit,” *NVIDIA*, 2025. https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/nano-super-developer-kit/

[17] Ultralytics, “YOLO11,” *Ultralytics.com*, 2024. https://docs.ultralytics.com/models/yolo11/

[18] Ultralytics, “Detect,” *docs.ultralytics.com*. https://docs.ultralytics.com/tasks/detect/

[19] M. Satyanarayanan, “The Emergence of Edge Computing,” *Computer*, vol. 50, no. 1, pp. 30–39, Jan. 2017, doi: https://doi.org/10.1109/mc.2017.9.

[20] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, “DATASET SHIFT IN MACHINE LEARNING.” Accessed: Nov. 16, 2025. [Online]. Available: https://ndl.ethernet.edu.et/bitstream/123456789/47647/1/Joaquin%20Quinonero-Candela_2009.pdf

[21] M. Ataei *et al.*, “Understanding Dataset Shift and Potential Remedies A Vector Institute Industry Collaborative Project Technical Report Authors (in alphabetical order) Vector Project Team Focus Area 2: Time Series Focus Area 1: Cross Sectional Focus Area 3: Computer Vision.” Available: https://vectorinstitute.ai/wp-content/uploads/2021/08/ds_project_report_final_august9.pdf

[22] J. Kirkpatrick *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017, doi: https://doi.org/10.1073/pnas.1611835114

[23] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural Networks*, vol. 113, pp. 54–71, May 2019, doi: https://doi.org/10.1016/j.neunet.2019.01.012.

[24] M. Delange *et al.*, “A continual learning survey: Defying forgetting in classification tasks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021, doi: https://doi.org/10.1109/TPAMI.2021.3057446.

[25] J. Chen and X. Ran, “Deep Learning With Edge Computing: A Review,” *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019, doi: https://doi.org/10.1109/jproc.2019.2921977. [26] V. A. Sindagi and V. M. Patel, “A survey of recent advances in CNN-based single image crowd counting and density estimation,” *Pattern Recognition Letters*, vol. 107, pp. 3–16, May 2018, doi: https://doi.org/10.1016/j.patrec.2017.07.007.

[27] hui gao, miaolei deng, wenjun zhao, and dexian zhang, “A Survey of Crowd Counting CNN-based,” Oct. 06, 2021. https://www.researchgate.net/publication/355347302_A_Survey_of_Crowd_Counting_CNN-based/download

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” 2018. Available: https://openaccess.thecvf.com/content_cvpr_2018/papers/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.pdf [29] A. Kirillov *et al.*, “Segment Anything.” Available: https://openaccess.thecvf.com/content/ICCV2023/papers/Kirillov_Segment_Anything_ICCV_2023_paper.pdf

[30] “TensorRT Documentation — NVIDIA TensorRT Documentation,” *Nvidia.com*, 2021. https://docs.nvidia.com/deeplearning/tensorrt/latest/index.html

[31] R. Berta and Alessandro De Gloria, *Applications in Electronics Pervading Industry, Environment and Society*. Springer Nature, 2023.

[32] B. M. Hussein and S. M. Shareef, “An Empirical Study on the Correlation between Early Stopping Patience and Epochs in Deep Learning,” *ITM Web of Conferences*, vol. 64, p. 01003, 2024, doi: https://doi.org/10.1051/itmconf/20246401003.

[33] J. Chen and Z. Wang, “Crowd counting with segmentation attention convolutional neural network,” *IET Image Processing*, vol. 15, no. 6, pp. 1221–1231, Jan. 2021, doi: https://doi.org/10.1049/ipr2.12099.

[34] S. Pandey, K.-F. Chen, and E. B. Dam, “Comprehensive Multimodal Segmentation in Medical Imaging: Combining YOLOv8 with SAM and HQ-SAM Models,” *arXiv.org*, 2023. https://arxiv.org/abs/2310.12995 (accessed Nov. 16, 2025).

[35] E. W. Weisstein, “Shoelace Formula,” *mathworld.wolfram.com*. https://mathworld.wolfram.com/ShoelaceFormula.html

[36] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 886–893, 2005, doi: https://doi.org/10.1109/cvpr.2005.177

[37] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: https://doi.org/10.1023/b:visi.0000029664.99615.94.

[38] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, “CNN-based Density Estimation and Crowd Counting: A Survey,” *arXiv:2003.12783 [cs]*, Mar. 2020, Available: https://arxiv.org/abs/2003.12783

[39] “Ultralytics | Revolutionizing the World of Vision AI,” *www.ultralytics.com*. https://www.ultralytics.com

[40] H. Bichri, A. Chergui, and M. Hain, “Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets,” *International journal of advanced computer science and applications/International journal of advanced computer science & applications*, vol. 15, no. 2, Jan. 2024, doi: https://doi.org/10.14569/ijacsa.2024.0150235.