Διπλωματική Εργασία

Αναγνώριση Αντικειμένων και των Χαρακτηριστικών τους από Οπτικά Δεδομένα Μη Επανδρωμένων Εναέριων Οχημάτων

Βασίλειος Γιοβάνογλου

Υποβλήθηκε για την μερική εκπλήρωση των απαιτήσεων για τη λήψη του

Διπλώματος Ηλεκτρολόγου Μηχανικού και Μηχανικών Υπολογιστών

από τη

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εξεταστική Επιτροπή

Καθ. Μιχαήλ Ζερβάκης (Επιβλέπων)

Καθ. Ευριπίδης Πετράκης,

Καθ. Μιχαήλ Γ. Λαγουδάκης

Χανιά, Μάιος 2025

Object Detection, Localization and Feature Characterization of Image Data from UAV

Vasileios Giovanoglou

Submitted in partial fulfilment of the requirements for the

Integrated Master's Degree in Electrical and Computer Engineering

in the

School of Electrical and Computer Engineering

Examination Committee

Prof. Michael Zervakis (Advisor)

Prof. Evripidis Petrakis

Prof. Michail G. Lagoudakis

Chania, May 2025

Περίληψη

Τα τελευταία χρόνια οι εφαρμογές που βασίζονται στην χρήση μη επανδρωμένων εναέριων οχημάτων (UAVs) εξαπλώνονται όλο και περισσότερο, αξιοποιώντας τις τεχνολογικές δυνατότητές τους σε συνδυασμό με άλλα αναπτυσσόμενα πεδία. Η παρούσα διπλωματική εργασία εξετάζει την ανίχνευση αντικειμένων σε αστικά περιβάλλοντα και την αναγνώριση του χρώματος τους εστιάζοντας σε τρεις μεγάλες κλάσεις αντικειμένων: αυτοκίνητα, λεωφορεία, πεζοί. Αξιοποιούνται οπτικά δεδομένα αεροφωτογραφιών που συλλέγονται από UAVs, ενώ δίνεται ιδιαίτερη έμφαση στην μελέτη και χρήση του νευρωνικού δικτύου YOLOv5 για ακριβή αναγνώριση αντικειμένων από αεροφωτογραφίες με κατακόρυφη γωνία λήψης. Ο πυρήνας της έρευνας αφορά την μελέτη και εφαρμογή του YOLOv5 καθώς και παραλλαγών του στο Stanford Drone Dataset, την αξιολόγηση της απόδοσής του με διάφορες μετρικές, καθώς και πειραματικές μελέτες με διαφορετικές διαμορφώσεις για τη βελτίωση της ακρίβειας και της αποδοτικότητας. Το μη τροποποιημένο YOLOv5 μοντέλο πέτυχε mAP@.5 89%, απόδοση υψηλότερη των παραλλαγών που αναπτύχθηκαν όπως του YOLOv5 με Softpool και Squeeze-and-Excitation mAP@.5 71% και του YOLOv5 με Softpool και CoordAttention mAP@.5 75%. Τα αποτελέσματα αναδεικνύουν την εξαιρετική ικανότητα του μοντέλου YOLOv5x στην ανίχνευση αντικειμένων, επιβεβαιώνοντας τη δυναμική του για πρακτικές εφαρμογές στην επιτήρηση, τη γεωργία και την ασφάλεια. Με την ενσωμάτωση τεχνικών απομόνωσης υποβάθρου (background) και ανίχνευσης χαρακτηριστικών, όπως η ανάλυση χρώματος, η εργασία αυτή συμβάλλει στον τομέα της μηχανικής όρασης, παρουσιάζοντας μια προηγμένη προσέγγιση στην ανίχνευση αντικειμένων από UAVs, με σημαντικές προεκτάσεις, τόσο για την ακαδημαϊκή έρευνα, όσο και για τις βιομηχανικές πρακτικές.

Abstract

This thesis explores object detection and feature characterization from visual data collected by unmanned aerial vehicles (UAVs), with a primary focus on leveraging the YOLOv5 model for accurate object recognition and classification in aerial imagery focusing on three classes of objects: cars, buses, pedestrians. The core of the research involves applying YOLOv5 and its variations to the Stanford Drone Dataset, evaluating their performance using various metrics, and conducting experimental studies with different configurations to enhance accuracy and efficiency. The unmodified YOLOv5 model achieved a mAP@.5 89%, outperforming variations, proposed in the thesis, such as YOLOv5 with Softpool and Squeeze-and-Excitation mAP@.5 71% and YOLOv5 with Softpool and CoordAttention mAP@.5 75%. Results highlight the exceptional capability of the YOLOv5x model in object detection, demonstrating its potential for practical applications in surveillance, agriculture, and security. By integrating background isolation techniques and feature detection methods, such as color analysis, this study contributes to the field of computer vision, presenting an advanced approach to UAV-based object detection with significant implications for both academic research and industrial practice.

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον Καθηγητή Μιχαήλ Ζερβάκη, επιβλέποντα της παρούσας διπλωματικής εργασίας, για την πολύτιμη καθοδήγηση, την υποστήριξη και την εμπιστοσύνη που μου έδειξε καθ' όλη τη διάρκεια πραγματοποίησης της διπλωματικής μου εργασίας. Επίσης, ευχαριστώ θερμά τα μέλη της εξεταστικής επιτροπής, Καθηγητή Ευριπίδη Πετράκη και Καθηγητή Μιχαήλ Γ. Λαγουδάκη, για τον χρόνο τους, τις παρατηρήσεις τους και τη συμβολή τους στην αξιολόγηση αυτής της εργασίας. Ένα ιδιαίτερο ευχαριστώ οφείλω στον Μάριο Αντωνακάκη για τη διαρκή καθοδήγηση, την τεχνική υποστήριξη και τις ουσιαστικές του συμβουλές που με βοήθησαν σε κάθε στάδιο αυτής της διαδρομής. Τέλος, θέλω να ευχαριστήσω από καρδιάς την οικογένειά μου για την αμέριστη στήριξη, την υπομονή και την αγάπη τους, που αποτέλεσαν το σταθερό μου σημείο αναφοράς καθ' όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Λίστα Εικόνων

Λίστα Πινάκων

Κεφάλαιο 1 Εισαγωγή

Τωρινές Καινοτομίες και Προκλήσεις

Η ανίχνευση αντικειμένων σε εικόνες που συλλέγονται από μη επανδρωμένα εναέρια οχήματα (UAVs) αποτελεί μία σημαντική πρόκληση στον τομέα της μηχανικής όρασης. Οι εικόνες αυτές συχνά χαρακτηρίζονται από πολυπλοκότητα, καθώς περιλαμβάνουν αντικείμενα διαφόρων μεγεθών, σχήματος και χρωμάτων, ενώ η ποιότητά τους επηρεάζεται από παράγοντες, όπως οι συνθήκες φωτισμού, η κίνηση του UAV, η απόσταση από τα αντικείμενα και η γωνία λήψης. Η ακριβής και αποτελεσματική ανίχνευση αντικειμένων σε τέτοιες συνθήκες είναι κρίσιμη για εφαρμογές, όπως η επιτήρηση, η γεωργία ακριβείας, η χαρτογράφηση και η αντιμετώπιση καταστροφών. Ωστόσο, οι υπάρχουσες μέθοδοι συχνά δυσκολεύονται να ανταπεξέλθουν στις απαιτήσεις πραγματικού χρόνου και στην ανάγκη για υψηλή ακρίβεια, ειδικά σε αεροφωτογραφίες, όπου τα αντικείμενα μπορεί να είναι μικρά ή μερικώς κρυμμένα.

Η γενίκευση των αλγορίθμων ανίχνευσης, λόγω της απόστασης, αλλά και της γωνίας λήψης των αντικειμένων, σε διαφορετικά δεδομένα αποτελεί σημαντική πρόκληση. Τα UAVs συλλέγουν δεδομένα από ποικίλα περιβάλλοντα, όπως αστικές περιοχές, αγροτικές εκτάσεις ή δασικές περιοχές, όπου η ποικιλομορφία και η δυναμική του περιβάλλοντος καθιστούν δύσκολη την ανίχνευση. Οι υπάρχοντες αλγόριθμοι απαιτούν συχνά εξειδικευμένη προσαρμογή για κάθε νέα εφαρμογή, κάτι που αυξάνει το κόστος και τον χρόνο ανάπτυξης. Οι περισσότεροι αλγόριθμοι αναγνώρισης αντικειμένων με χρήση βαθιάς μάθησης έχουν εκπαιδευτεί με φωτογραφίες / αεροφωτογραφίες, οι οποίες φέρουν συγκεκριμένες γωνίες λήψης – συνήθως πλάγια ή και πανοραμική λήψη – γεγονός που δυσχεραίνει την διαδικασία αναγνώρισης. Οι εικόνες που λαμβάνονται από «απότομες» γωνίες ή από μεγάλα ύψη μπορεί να στερούνται της ανάλυσης ή της καθαρότητας που απαιτείται για αποτελεσματική αναγνώριση. Τα αντικείμενα μακριά από τον αισθητήρα μπορεί να φαίνονται συμπιεσμένα, ενώ οι πλάγιες γωνίες παραμορφώνουν τα σχήματα των αντικειμένων, περιπλέκοντας την ακριβή ανίχνευση και αναγνώριση.

Κίνητρα και Συμβολή

Για την αντιμετώπιση των παραπάνω προκλήσεων, είναι απαραίτητη η ανάπτυξη αποδοτικών αλγορίθμων ανίχνευσης αντικειμένων που συνδυάζουν ακρίβεια, ταχύτητα και δυνατότητα γενίκευσης, ώστε να ανταποκρίνονται στις απαιτήσεις εφαρμογών πραγματικού χρόνου και να αξιοποιούν στο έπακρο τις δυνατότητες των UAVs. Η ανίχνευση αντικειμένων από δεδομένα UAV αποτελεί ένα πεδίο με σημαντικές προοπτικές και εφαρμογές σε διάφορους τομείς, όπως η επιτήρηση, η γεωργία, η χαρτογράφηση, και η ασφάλεια. Τα UAVs, χάρη στην ευελιξία και την ικανότητά τους να συλλέγουν δεδομένα από διαφορετικά ύψη και γωνίες, προσφέρουν ένα εξαιρετικό εργαλείο για την ανάλυση μεγάλων περιοχών σε σύντομο χρόνο. Παράλληλα, οι τεχνικές ανίχνευσης αντικειμένων μπορούν να ενισχύσουν την ακρίβεια αυτών των εφαρμογών,

επιτρέποντας τον αυτόματο εντοπισμό και την παρακολούθηση αντικειμένων, χωρίς την ανάγκη χειροκίνητης ανάλυσης.

Δεδομένου ότι οι περισσότερες λύσεις βασίζονται σε αεροφωτογραφίες με γωνία λήψης πλάγια ή οριζόντια, ανακύπτει η ανάγκη μελέτης και ανάπτυξης ενός συστήματος βαθιάς μάθησης για την αναγνώριση αντικειμένων από αεροφωτογραφίες με κατακόρυφη γωνία λήψης. Σκοπός αυτής της εργασίας είναι η ανάπτυξη και αξιολόγηση ενός προηγμένου μοντέλου ανίχνευσης αντικειμένων που μπορεί να λειτουργήσει αποτελεσματικά σε πραγματικό χρόνο. Η αξιοποίηση του YOLOv5, ενός από τα πιο σύγχρονα και αποδοτικά μοντέλα βαθιάς μάθησης, και η προσαρμογή του σε δεδομένα από UAV, δηλαδή να βασίζεται σε αεροφωτογραφίες από κατακόρυφη γωνία λήψης, αποτελούν τον πυρήνα της προσέγγισης. Η εργασία αυτή εστιάζει στη δημιουργία λύσεων που συνδυάζουν ταχύτητα, ακρίβεια και δυνατότητα γενίκευσης με κατακόρυφη γωνία λήψης, ώστε να είναι εφαρμόσιμες σε ένα ευρύ φάσμα πραγματικών σεναρίων, όπως η έγκαιρη διάγνωση καταστροφών ή η διαχείριση γεωργικών καλλιεργειών. Τα αποτελέσματα της εργασίας αυτής, όχι μόνο επιβεβαιώνουν την αποτελεσματικότητα της προσέγγισης, αλλά προσφέρουν και κατευθυντήριες γραμμές για τη βελτίωση των υπαρχόντων συστημάτων. Η δυνατότητα εντοπισμού αντικειμένων υψηλής ακρίβειας, ακόμα και σε σύνθετα περιβάλλοντα, ανοίγει τον δρόμο για την ανάπτυξη νέων εφαρμογών που απαιτούν ακριβή ανίχνευση και παρακολούθηση αντικειμένων, ενισχύοντας έτσι τη χρήση των UAV σε κρίσιμους τομείς.

Για τις ανάγκες της παρούσας μελέτης, η ανίχνευση επικεντρώνεται σε τρεις βασικές κατηγορίες αντικειμένων με ιδιαίτερο ενδιαφέρον σε αστικά περιβάλλοντα: πεζούς (pedestrians), αυτοκίνητα (cars) και λεωφορεία (buses). Οι κατηγορίες αυτές επιλέχθηκαν λόγω της σημασίας τους σε εφαρμογές επιτήρησης, παρακολούθησης κυκλοφορίας και ανάλυσης κινητικότητας.

Τέλος, η συμβολή της παρούσας εργασίας εκτείνεται και στη διερεύνηση νέων τεχνικών (μηχανισμών προσοχής – attention mechanisms) που βελτιώνουν τη γενίκευση και την αποδοτικότητα των μοντέλων βαθιάς μάθησης, με έμφαση στις ανάγκες της σύγχρονης βιομηχανίας και έρευνας. Η εφαρμογή αυτών των τεχνικών προσφέρει λύσεις που μπορούν να υιοθετηθούν από διάφορους κλάδους, διευρύνοντας τη χρήση της τεχνολογίας αυτής και ενισχύοντας τη βιωσιμότητα της ανάπτυξης εφαρμογών βασισμένων σε UAV.

Οργάνωση της Εργασίας

Η εργασία αυτή οργανώνεται σε δέκα κεφάλαια, κάθε ένα από τα οποία ασχολείται με διαφορετικές πτυχές της ανίχνευσης και χαρακτηρισμού αντικειμένων από δεδομένα εικόνων UAV. Στο πρώτο κεφάλαιο, παρέχεται μια εισαγωγή στα μη επανδρωμένα εναέρια οχήματα (drones) και στις βασικές τεχνικές ανίχνευσης αντικειμένων. Το δεύτερο κεφάλαιο εξετάζει το θεωρητικό υπόβαθρο της μηχανικής μάθησης και της βαθιάς μάθησης, ενώ το τρίτο κεφάλαιο εστιάζει στο βασικό μοντέλο YOLOv5 που χρησιμοποιείται σε αυτή την εργασία. Στα επόμενα κεφάλαια, περιγράφεται το σύνολο δεδομένων που χρησιμοποιήθηκε (Stanford Drone Dataset), η επεξεργασία και οι τροποποιήσεις του. Ακολουθεί η παρουσίαση των μετρικών αξιολόγησης, η

περιγραφή των τροποποιήσεων που εφαρμόστηκαν στην αρχιτεκτονική του YOLOv5x μοντέλου αλλά και η πειραματική μελέτη με την παρουσίαση των αποτελεσμάτων. Στη συνέχεια παρουσιάζεται η διαδικασία ανίχνευσης του χρώματος των εντοπισμένων αντικειμένων ενώ το τελευταίο κεφάλαιο συγκρίνει τα αποτελέσματα με άλλα μοντέλα και παρέχει τα τελικά συμπεράσματα της εργασίας.

Κεφάλαιο 2 – Θεωρητικό Υπόβαθρο: Στο δεύτερο κεφάλαιο πραγματοποιείται μια συνοπτική παρουσίαση των αεροφωτογραφιών και των βασικών τύπων της, όπως οι κάθετες και πλάγιες αεροφωτογραφίες. Στη συνέχεια, παρουσιάζονται συνοπτικά τα μη επανδρωμένα αεροσκάφη και drones και περιγράφεται ο ρόλος τους στην αεροφωτογράφηση. Τέλος, παρουσιάζεται η θεωρία της τεχνητής νοημοσύνης, με έμφαση στη μηχανική μάθηση (machine learning) και τις state-of-the-art μεθόδους, διαδικασίες και αρχιτεκτονικές που χρησιμοποιούνται στη βιβλιογραφία, εστιάζοντας ιδιαίτερα στα συνελικτικά νευρωνικά δίκτυα.

Κεφάλαιο 3 – Βασικό Μοντέλο (YOLOv5): Εξηγεί την αρχιτεκτονική και τα μέρη του YOLOv5, όπως το δίκτυο κορμού και το δίκτυο ταξινόμησης.

Κεφάλαιο 4 – Σύνολο Δεδομένων: Περιγράφει τα δεδομένα που χρησιμοποιήθηκαν, τη διαδικασία επεξεργασίας τους και τις αλλαγές που έγιναν.

Κεφάλαιο 5 – Μετρικές Αξιολόγησης: Παρουσιάζονται και αναλύονται συνοπτικά οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων και τον μοντέλων που αναπτύχθηκαν σε αυτή την εργασία.

Κεφάλαιο 6 – Παραλλαγές του YOLOv5: Στο κεφάλαιο αυτό παρουσιάζονται οι τροποποιήσεις που εφαρμόστηκαν στο βασικό μοντέλο YOLOv5, με στόχο τη βελτίωση της ακρίβειας στην αναγνώριση αντικειμένων από εναέριες εικόνες. Περιγράφονται αναλυτικά τρεις παραλλαγές του μοντέλου: η αντικατάσταση του MaxPooling με Softpool, η ενσωμάτωση του μηχανισμού Squeeze-and-Excitation (SE), και η προσθήκη του μηχανισμού Coord Attention (CA).

Κεφάλαιο 7 – Πειραματική Μελέτη – Μεθοδολογία – Αποτελέσματα: Παρουσιάζει τη μεθοδολογία των πειραμάτων και τα αποτελέσματά τους. Στο κεφάλαιο αυτό παρουσιάζεται η μεθοδολογία που χρησιμοποιήθηκε. Περιγράφονται τα μοντέλα που αναπτύχθηκα για την αναγνώριση αντικειμένων από οπτικά δεδομένα μη επανδρωμένων εναέριων οχημάτων.

Κεφάλαιο 8 – Εντοπισμός χρώματος: Στο κεφάλαιο αυτό παρουσιάζεται η διαδικασία εντοπισμού χρώματος των αντικειμένων αναγνώρισης στα δεδομένα μελέτης.

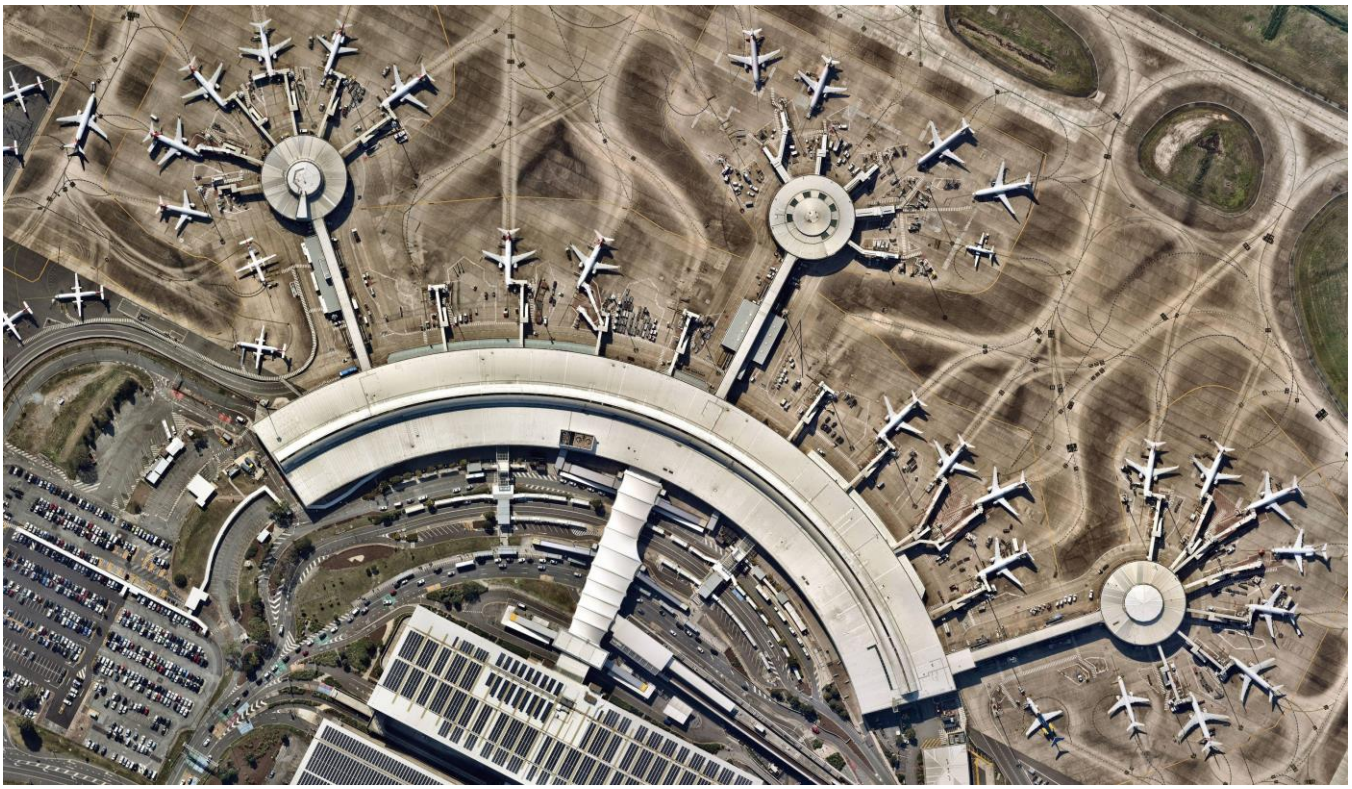
Κεφάλαιο 9 – Σύγκριση με άλλα μοντέλα: Στο κεφάλαιο αυτό τα αποτελέσματα και η απόδοση των μοντέλων που αναπτύχθηκαν αξιολογούνται και συγκρίνονται με τα αποτελέσματα από άλλα μοντέλα ανίχνευσης αντικειμένων της βιβλιογραφίας για το ίδιο σύνολο δεδομένων.

Κεφάλαιο 10 – Συζήτηση και Συμπεράσματα: Στο τελευταίο κεφάλαιο, εξάγονται σημαντικά συμπεράσματα για την χρήση της μηχανικής μάθησης και συγκεκριμένα της χρήσης του μοντέλου YOLOv5 και των παραλλαγών του στην αναγνώριση αντικειμένων και των χαρακτηριστικών τους από οπτικά δεδομένα μη επανδρωμένων εναέριων οχημάτων.

Κεφάλαιο 2 Θεωρητικό Υπόβαθρο

Αεροφωτογραφίες

Οι εναέριες εικόνες, γνωστές και ως αεροφωτογραφίες, είναι φωτογραφίες που λαμβάνονται από υπερυψωμένες πλατφόρμες, όπως αεροπλάνα, δορυφόρους, drones ή αερόστατα [1]. Ιστορικά, οι πρώτες εναέριες εικόνες λήφθηκαν με τη χρήση αερόστατων στα μέσα του 19ου αιώνα, ενώ κατά τον 20ό αιώνα, ιδιαίτερα κατά τη διάρκεια του Α' Παγκοσμίου Πολέμου, η χρήση τους εξελίχθηκε με την εφαρμογή της φωτογραφίας από αεροπλάνα για στρατιωτικούς σκοπούς. Σήμερα, οι τεχνολογικές εξελίξεις στην τηλεπισκόπηση (remote sensing) περιλαμβάνουν δορυφορικές πλατφόρμες και διάφορα οπτικο-μηχανικά συστήματα σάρωσης, επιτρέποντας ακριβή λήψη δεδομένων σε ένα ευρύ φάσμα του ηλεκτρομαγνητικού φάσματος, από το υπεριώδες έως το εγγύς υπέρυθρο.



Εικόνα 2.1. Εναέρια εικόνα αεροδρομίου - Μια υψηλής ανάλυσης εναέρια φωτογραφία ενός αεροδρομίου, που δείχνει αεροσκάφη σταθμευμένα στις πύλες, διαδρόμους τροχοδρόμησης και υποδομές αεροπορικών εγκαταστάσεων [1].

Οι αεροφωτογραφίες αποτελούν πολύτιμο εργαλείο σε πολλούς τομείς, όπως ο αστικός σχεδιασμός, η γεωλογία, η περιβαλλοντική παρακολούθηση και η διαχείριση φυσικών καταστροφών. Παρέχουν εκτενή χωρική πληροφόρηση, διευκολύνοντας τη συστηματική καταγραφή μεταβολών της επιφάνειας της Γης, την ανάλυση βλάστησης και τη χαρτογράφηση της δομής αστικών περιοχών. Στο σύγχρονο πλαίσιο, οι εναέριες εικόνες διαδραματίζουν σημαντικό

ρόλο στην ανίχνευση και αναγνώριση αντικειμένων, όπου αλγόριθμοι εντοπίζουν και αναγνωρίζουν αντικείμενα, όπως οχήματα, κτίρια, πλοία κ.α. Τέτοιες εφαρμογές επεκτείνονται σε τομείς όπως η γεωργία ακριβείας, η διαχείριση κυκλοφορίας και η στρατιωτική αναγνώριση.

Εικόνα 2.2. Μεγάλης κλίμακας εναέρια φωτογραφία παραλιακής περιοχής: Οι εναέριες εικόνες μεγάλης κλίμακας λαμβάνονται από χαμηλό υψόμετρο, επιτρέποντας λεπτομερή καταγραφή τοπικών χαρακτηριστικών. Αν και καλύπτουν μικρότερη έκταση εδάφους, παρέχουν μεγαλύτερη ανάλυση και είναι ιδιαίτερα χρήσιμες για χαρτογράφηση παράκτιων ζωνών, περιβαλλοντική ανάλυση και μετρήσεις γεωμορφολογικών αλλαγών [2].

Είδη Αεροφωτογραφιών

Οι αεροφωτογραφίες μπορούν να ταξινομηθούν με βάση διάφορα κριτήρια, όπως η γωνία λήψης, το τη φασματική ζώνη καταγραφής αλλά και το οπτικό πεδίο της κάμερας [1], [2], [3]. Κάθε τύπος έχει συγκεκριμένες εφαρμογές, από τη χαρτογράφηση και τη γεωμορφολογική ανάλυση έως την περιβαλλοντική παρακολούθηση και την αστική ανάπτυξη.

1. Ταξινόμηση με βάση τη γωνία λήψης

Οι αεροφωτογραφίες μπορούν να κατηγοριοποιηθούν ανάλογα με τη γωνία υπό την οποία λαμβάνονται οι εικόνες:

- Κάθετες αεροφωτογραφίες (vertical aerial imaging): Η κάμερα είναι τοποθετημένη κατακόρυφα προς το έδαφος (με απόκλιση μικρότερη των 3°). Αυτή η τεχνική παρέχει εικόνες με γεωμετρική ακρίβεια και ομοιόμορφη κλίμακα, γεγονός που τις καθιστά ιδανικές για χαρτογραφικές και τοπογραφικές εφαρμογές, καθώς και για τη δημιουργία ορθοφωτοχαρτών.

- Πλάγιες αεροφωτογραφίες (oblique aerial imaging): Η κάμερα είναι κεκλιμένη σε σχέση με το έδαφος, προσφέροντας εικόνες με προοπτική. Αυτή η κατηγορία διακρίνεται περαιτέρω σε αεροφωτογραφίες οι οποίες μπορούν να έχουν:
    - *Ήπια πλάγια λήψη (low oblique):* Ο ορίζοντας δεν εμφανίζεται στο κάδρο και χρησιμοποιείται κυρίως για μελέτες αρχιτεκτονικής και πολεοδομίας.
    - *Έντονα πλάγια λήψη (high oblique):* Ο ορίζοντας περιλαμβάνεται στο πεδίο λήψης, επιτρέποντας την αποτύπωση ευρύτερων τοπίων και μεγάλων εδαφικών εκτάσεων.

2. Ταξινόμηση με βάση τη φασματική ζώνη

- Ορατές αεροφωτογραφίες (visible spectrum aerial imaging): Απεικονίζουν το τοπίο όπως γίνεται αντιληπτό από το ανθρώπινο μάτι και χρησιμοποιούνται κυρίως για χαρτογράφηση και πολεοδομικές μελέτες.

- Υπέρυθρες και πολυφασματικές αεροφωτογραφίες (infrared & multispectral aerial imaging): Καταγράφουν δεδομένα σε μη ορατά μήκη κύματος, παρέχοντας πληροφορίες για την υγεία της βλάστησης, την υγρασία του εδάφους, τη ρύπανση των υδάτων κ.α..

- Θερμικές αεροφωτογραφίες (thermal aerial imaging): Ανιχνεύουν τις θερμοκρασιακές διαφορές αντικειμένων και επιφανειών, γεγονός που τις καθιστά ιδανικές για αναζητήσεις διάσωσης, ενεργειακές επιθεωρήσεις και ανίχνευση πυρκαγιών.

3. Ταξινόμηση με βάση το οπτικό πεδίο της κάμερας

Το εύρος του οπτικού πεδίου της κάμερας επηρεάζει την κλίμακα και τις παραμορφώσεις των αεροφωτογραφιών. Με βάση το εύρος του οπτικού πεδίου, οι φωτογραφικές κάμερες μπορούν να έχουν [3]:

- Κανονικό οπτικό πεδίο (50° – 75°): Παρέχει εικόνες με ελάχιστες γεωμετρικές παραμορφώσεις και χρησιμοποιείται για ακριβείς μετρήσεις.
- Ευρύ οπτικό πεδίο (75° – 100°): Επιτρέπει την αποτύπωση μεγαλύτερων περιοχών χωρίς μεγάλη απώλεια λεπτομέρειας.
- Πολύ ευρύ οπτικό πεδίο (100° – 125°): Καλύπτει ευρείες περιοχές, αλλά μπορεί να προκαλέσει σημαντικές παραμορφώσεις στα απεικονιζόμενα αντικείμενα.

Αξίζει να σημειωθεί ότι το πιο συχνά χρησιμοποιούμενο οπτικό πεδίο στις αεροφωτογραφίες είναι αυτό των 90°. Οι κάμερες με πολύ ευρύ οπτικό πεδίο καλύπτουν σχεδόν ολόκληρο τον ορίζοντα, γεγονός που οδηγεί σε σημαντικές γεωμετρικές παραμορφώσεις των αντικειμένων που απεικονίζονται. Αυτό έχει ως αποτέλεσμα οι φωτογραφίες αυτές να μην είναι κατάλληλες για

εφαρμογές που απαιτούν υψηλή ακρίβεια μετρήσεων. Στο παρελθόν, οι πανοραμικές κάμερες χρησιμοποιούνταν για την αποτύπωση μεγάλων γεωγραφικών περιοχών. Ωστόσο, με την πρόοδο της δορυφορικής τηλεπισκόπησης, η οποία επιτρέπει την απεικόνιση εκτεταμένων περιοχών με μειωμένες γεωμετρικές αλλοιώσεις, η χρήση των πανοραμικών αεροφωτογραφιών έχει περιοριστεί σημαντικά.

Προκλήσεις

Παρά τα πλεονεκτήματά τους, οι αεροφωτογραφίες παρουσιάζουν συγκεκριμένες προκλήσεις [4]:

1. **Μεταβολές κλίμακας και προσανατολισμού**: Τα αντικείμενα στις αεροφωτογραφίες εμφανίζουν συχνά σημαντικές διακυμάνσεις στο μέγεθος και τον προσανατολισμό λόγω της προοπτικής από ψηλά, καθιστώντας πιο δύσκολη την ανίχνευση και ταξινόμησή τους.
2. **Παραμορφώσεις από τη γωνία και την απόσταση λήψης**: Η γωνία της κάμερας και το ύψος της πλατφόρμας λήψης προκαλούν γεωμετρικές παραμορφώσεις. Για παράδειγμα, οι πλάγιες λήψεις μπορεί να οδηγήσουν σε ασυνέπειες κλίμακας, όπου τα αντικείμενα κοντά στο κέντρο της εικόνας απεικονίζονται με μεγαλύτερη ακρίβεια από αυτά στα άκρα. Παρομοίως, η λήψη από μεγάλο ύψος μειώνει τη χωρική ανάλυση, καθιστώντας δύσκολη τη διάκριση μικρότερων χαρακτηριστικών. Επιπλέον, η γωνία λήψης επηρεάζει τη σαφήνεια των αντικειμένων—αντικείμενα που είναι κεκλιμένα σε σχέση με τον αισθητήρα μπορεί να κρύβουν σημαντικές λεπτομέρειες, δυσχεραίνοντας περαιτέρω την ανίχνευση και αναγνώριση.
3. **Συνθήκες περιβάλλοντος και φωτός**: Νεφοκάλυψη, ατμοσφαιρική σκέδαση και διαφοροποιήσεις στον φωτισμό, όπως σκιές και χαμηλές ηλιακές γωνίες, επηρεάζουν την ποιότητα της εικόνας και συνεπώς την διαύγεια των αντικειμένων.
4. **Περιορισμοί στην επεξεργασία δεδομένων**: Η επεξεργασία μεγάλων τέτοιων εικόνων (π.χ., 20.000 x 20.000 pixel) απαιτεί σημαντικούς υπολογιστικούς πόρους. Μέθοδοι όπως η κατάτμηση (segmentation) των εικόνων σε μικρότερα τμήματα είναι απαραίτητες, αλλά μπορούν να οδηγήσουν σε προβλήματα κατακερματισμού δεδομένων.

Οι εναέριες εικόνες προσφέρουν μια μοναδική και σημαντική προοπτική για την ανάλυση και τη διαχείριση της επιφάνειας της Γης. Είναι απαραίτητες σε πολλούς τομείς, από την παρακολούθηση φυσικών καταστροφών έως τη χαρτογράφηση της αστικής ανάπτυξης. Οι κατακόρυφες εναέριες εικόνες, που λαμβάνονται με τον οπτικό άξονα σχεδόν κάθετα προς το έδαφος, είναι ιδιαίτερα σημαντικές λόγω της ομοιόμορφης κλίμακας και της γεωμετρικής ακρίβειας που προσφέρουν, καθιστώντας τις ιδανικές για τη δημιουργία χαρτών, τη διεξαγωγή μετρήσεων και τη συστηματική ανάλυση.

Ωστόσο, οι προκλήσεις παραμένουν, ιδιαίτερα όσον αφορά τις παραμορφώσεις γωνίας και απόστασης. Οι εικόνες που λαμβάνονται από απότομες γωνίες ή από μεγάλα ύψη μπορεί να στερούνται της ανάλυσης ή της καθαρότητας που απαιτείται για αποτελεσματική ανάλυση. Τα αντικείμενα μακριά από τον αισθητήρα μπορεί να φαίνονται συμπιεσμένα, ενώ οι πλάγιες γωνίες παραμορφώνουν τα σχήματα των αντικειμένων, περιπλέκοντας την ακριβή ανίχνευση και

αναγνώριση. Αυτές οι προκλήσεις υπογραμμίζουν την ανάγκη για προηγμένες μεθόδους προεπεξεργασίας, εξελιγμένους αλγορίθμους και τεχνολογία αισθητήρων για τη διόρθωση γεωμετρικών και προοπτικών παραμορφώσεων.

Μη Επανδρωμένα Αεροσκάφη

Τα μη επανδρωμένα ιπτάμενα οχήματα (Unmanned Aerial vehicles - UAV), ονομάζονται τα κάθε είδους ιπτάμενα οχήματα που κινούνται στον αέρα (πάνω από την επιφάνεια της Γης) αυτόνομα (χωρίς πιλότο ή κυβερνήτη), προγραμματισμένα ή τηλεκατευθυνόμενα [5], σε μορφή μικρού αεροπλάνου ή ελικοπτέρου με έναν ή περισσότερους κινητήρες και έλικες συντονισμένους για πλήρως ελεγχόμενη πτήση από ειδικό λογισμικό ή χειριστήριο εδάφους (Εικόνα 2.3). Άλλες ονομασίες που κατά καιρούς έχουν αποδοθεί στα μη επανδρωμένα αεροσκάφη είναι εξής:

- Unmanned Aerial System (UAS)
- Remotely Piloted Aircraft System (RPAS)
- Drones

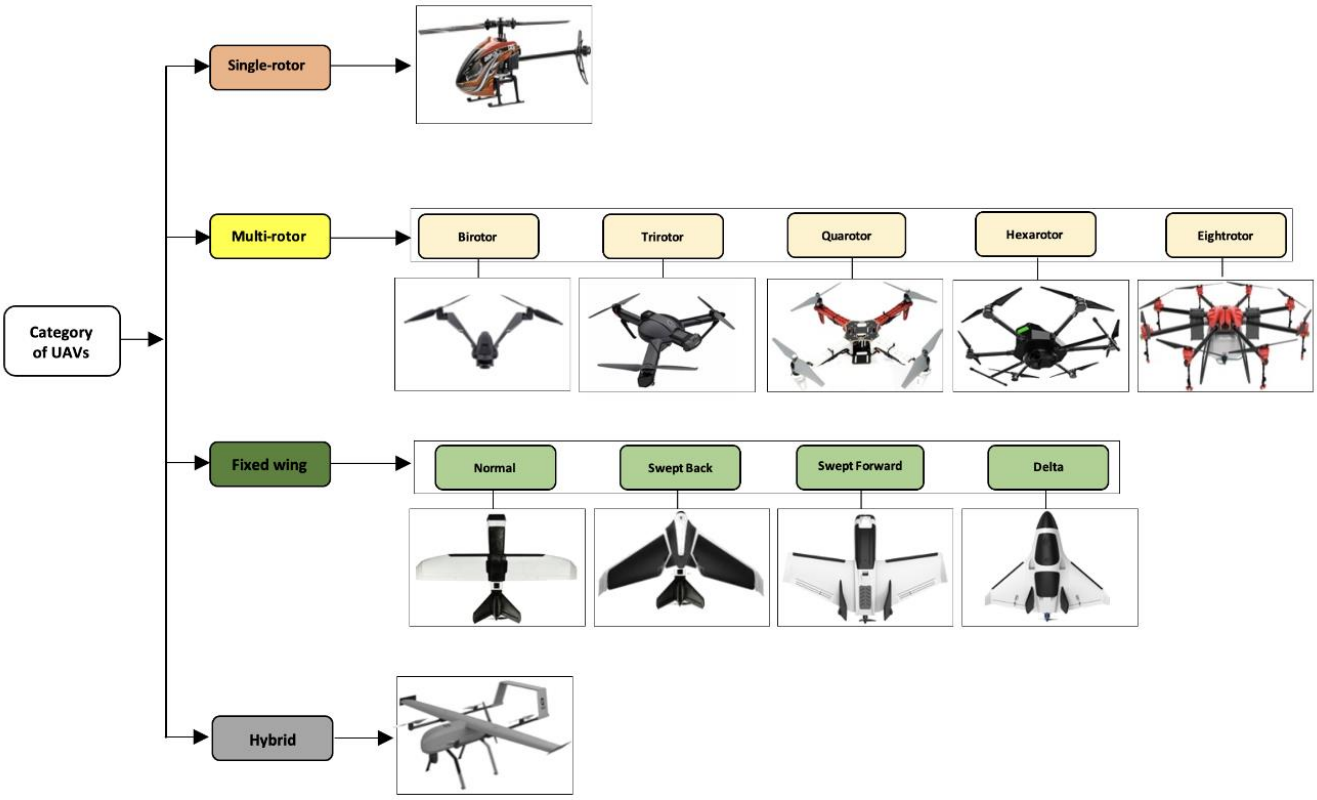
Εικόνα 2.3. Παραδείγματα μη επανδρωμένων ιπτάμενων αντικειμένων (UAVs): Αριστερά, αεροσκάφος σταθερών πτερύγων και δεξιά, τετρακόπτερο (quadcopter).

Κατά την εργασία αυτή θα αναφερόμαστε στα μη επανδρωμένα αεροσκάφη με τον όρο “drones”. Τα drones χρησιμοποιούν τέσσερις ισομεγέθεις έλικες συμμετρικά τοποθετημένους πάνω σε έναν σκελετό και στο κέντρο του οποίου βρίσκεται το ωφέλιμο φορτίο. Στη θέση αυτή βρίσκεται και ο αυτόματος πιλότος. Η περιστροφή των ελίκων γίνεται ανά δύο αντίστροφα ούτως ώστε στο κέντρο του να εφαρμόζεται μηδενική ροπή. Οι γωνιακές ταχύτητες είναι αυτές που ελέγχουν την κίνηση των drones. Όταν οι έλικες έχουν την ίδια ακριβώς γωνιακή ταχύτητα, τότε το drone ανυψώνεται

διατηρώντας σταθερή την κλίση του με αποτέλεσμα να μην περιστρέφεται γύρω από το κέντρο μάζας. Η περιστροφή γύρω από τον κάθετο άξονα του επιτυγχάνεται όταν η ταχύτητα δύο ομοίως περιστρεφόμενων κινητήρων αυξομειωθεί. Η παράλληλη κίνηση προς το έδαφος επιτυγχάνεται αποκτώντας κάποια κλίση ως προς το έδαφος.

Υπάρχουν όμως διάφορες κατηγορίες από drones, οι οποίες διαφοροποιούνται ανάλογα με την δομή και την χρησιμοποίησή τους, όπως επίσης και διάφοροι τρόποι ελέγχου ενός UAV. Οι βασικές κατηγορίες στις οποίες μπορούν να χωριστούν τα drones είναι τέσσερις και βασίζονται στην κατασκευή τους αλλά και τις διαφορετικές τεχνικές πτήσης και ανύψωσης (Εικόνα 2.4):

- Απλού Έλικα – Single-rotor
- Πολλαπλών Ελίκων – Multi-rotor
- Σταθερής Πτέρυγας – Fixed wing
- Υβριδικά – Hybrid



Εικόνα 2.4: Βασικές κατηγορίες των drones [6].

Στις παραπάνω κύριες κατηγορίες, υπάρχουν και πιο σπάνιες περιπτώσεις οι οποίες περιλαμβάνουν για παράδειγμα drones με 12 ή 16 έλικες ή με 8 έλικες των οποίων ωστόσο η διάταξη σχηματίζει το γράμμα V του λατινικού αλφαβήτου.

Η ταχύτατα αναπτυσσόμενη αγορά των drones μοιάζει με ένα οικοσύστημα καινούργιων λογισμικών και ήδη ετοιμάζεται να ικανοποιήσει μια μακρά λίστα ενδιαφερόμενων από πολλούς χώρους [5]:

- της περιβαλλοντικής διαχείρισης,
- της γεωργίας,
- της ενέργειας,
- του real estate και των κατασκευών με σκοπό την έρευνα,
- τη διάσωση, την ασφάλεια και παρατήρηση.

Μικρές ιδιωτικές εταιρείες, είτε νεοφυείς επιχειρήσεις αποτελούν ορισμένους από τους κατασκευαστές τους. Ωστόσο, τα τελευταία χρόνια μεγάλες εταιρείες τεχνολογίας και αλλά και όμιλοι βιομηχανιών έχουν πλέον αρχίσει να επενδύουν σημαντικά ποσά στην έρευνα και την ανάπτυξη τεχνολογιών αυτού του είδους. Στην Ελλάδα η χρήση των drones εστιάζεται κυρίως στη κάλυψη αθλητικών γεγονότων, όπως οι ποδοσφαιρικοί αγώνες, στην επιτήρηση δασικών περιοχών, αλλά και στον χώρο της επαγγελματικής φωτογραφίας και της καταγραφής δεξιώσεις σε εξωτερικούς χώρους. Παράλληλα, διερευνάται η προοπτική χρήσης μη επανδρωμένων αεροσκαφών για ειρηνικές εφαρμογές, όπως η μεταφορά φαρμάκων, βιολογικών δειγμάτων για ιατρικούς ελέγχους και τροφίμων προς και από απομακρυσμένες περιοχές, χρησιμοποιώντας την τεχνολογία και τις μελέτες που προωθούνται υπό την ονομασία Matternet [7] (παραλλαγή του Internet), η οποία αφορά την αυτόματη μετακίνηση υλικών.

Καθώς τα μη επανδρωμένα αεροσκάφη γίνονται όλο και πιο διαδεδομένα, τα ρυθμιστικά πλαίσια έχουν προσαρμοστεί και εξελιχθεί ώστε να διασφαλιστεί η ασφαλής ενσωμάτωση στον χώρο της αεροπορίας αλλά και να αντιμετωπίσουν πιθανά προβλήματα και ανησυχίες που σχετίζονται με προστασία της ιδιωτικής ζωής. Αναδύονται επίσης ηθικοί προβληματισμοί, ιδίως σε σχέση με την επιτήρηση και τις αυτόνομες λειτουργίες, αναδεικνύοντας την ανάγκη για υπεύθυνη χρήση τους.

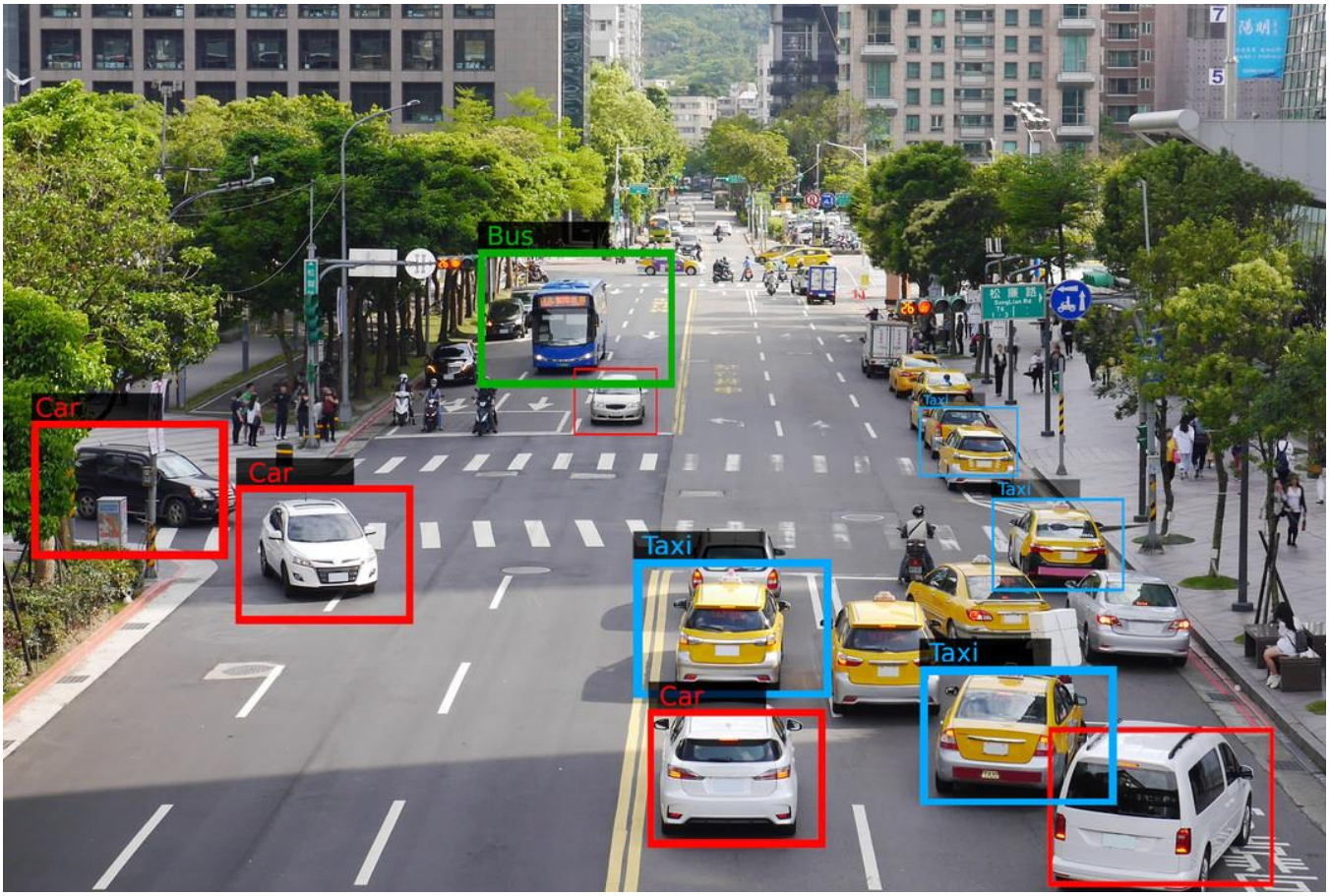
Το μέλλον της τεχνολογίας UAV είναι έτοιμο για περαιτέρω καινοτομία, με την έρευνα να επικεντρώνεται κυρίως στην αυτόνομη πτήση (autonomous flight), τη νοημοσύνη σμήνους (swarm intelligence) αλλά και τη βελτίωση της αντοχής τους. Ωστόσο, πρέπει να αντιμετωπιστούν προκλήσεις όπως η διαχείριση του εναέριου χώρου, η προστασία της ιδιωτικής ζωής και τα ζητήματα ασφάλειας προκειμένου να αξιοποιηθούν στο έπακρο οι δυνατότητές τους.

Τα UAVs αντιπροσωπεύουν ένα δυναμικό και ταχέως εξελισσόμενο τεχνολογικό πεδίο με τη δύναμη να μετασχηματίσει τις στρατιωτικές επιχειρήσεις, να ενισχύσει τις εμπορικές δραστηριότητες και να αντιμετωπίσει κρίσιμες κοινωνικές προκλήσεις. Καθώς ατενίζουμε το μέλλον, η συνεχής εξέλιξη της τεχνολογίας UAV υπόσχεται να ξεκλειδώσει ακόμη μεγαλύτερες δυνατότητες.

Η ανάπτυξη των μη επανδρωμένων αεροσκαφών έχει μεταμορφώσει την αεροφωτογράφηση, επιτρέποντας τη συλλογή δεδομένων με υψηλή ανάλυση και χαμηλότερο κόστος. Τα drones μπορούν να πετούν σε χαμηλά ύψη, μειώνοντας τις παραμορφώσεις λόγω απόστασης και γωνίας λήψης, ενώ προσφέρουν μεγαλύτερη ευελιξία στη συλλογή δεδομένων για εφαρμογές όπως η γεωργία ακριβείας, η επιτήρηση υποδομών και η περιβαλλοντική διαχείριση.

Ανίχνευση Αντικειμένων και Αλγοριθμικές Λύσεις

Η ανίχνευση αντικειμένων (object detection) αποτελεί ένα πρόβλημα το οποίο απασχολεί την ερευνητική κοινότητα και συγκεκριμένα τους τομείς της μηχανικής όρασης υπολογιστών και της επεξεργασίας εικόνας από τα τέλει του 20ου αιώνα. Πιο συγκεκριμένα αφορά τον εντοπισμό και την αναγνώριση αντικειμένων, που ανήκουν σε μια κλάση ή συγκεκριμένη κατηγορία (π.χ. «άνθρωπος», «αυτοκίνητο»), σε βίντεο ή ψηφιακές εικόνες. Δηλαδή, δεδομένης μίας εικόνας ή ενός βίντεο, ένας αλγόριθμος μπορεί να εντοπίσει και να αναγνωρίσει και κατηγοριοποιήσει αντικείμενα συγκεκριμένου ενδιαφέροντος και να παρέχει πληροφορίες σχετικές με τη θέση τους μέσα στην εικόνα. Ένα παράδειγμα αναγνώρισης αντικειμένων φαίνεται στην Εικόνα 2.5. Τα αναγνωρισμένα αντικείμενα – στην συγκεκριμένη περίπτωση οχήματα όπως αυτοκίνητα (Car), λεωφορεία (Bus) και ταξί (Taxi) – περικλείονται από ένα πλαίσιο το χρώμα του οποίου υποδηλώνει και την κατηγορία που ανήκει.



Εικόνα 2.5. Παράδειγμα αναγνώρισης αντικειμένων. Τα αναγνωρισμένα αντικείμενα – οχήματα περικλείονται από χρωματιστά πλαίσια.

Βέβαια, τα τελευταία χρόνια η ραγδαία ανάπτυξη στον τομέα των GPUs, από εταιρείες όπως η NVIDIA, και η υπολογιστική βοήθεια που προσφέρουν, σε συνδυασμό με τη ανάπτυξη της τεχνητής νοημοσύνης και συγκεκριμένα νέων και πιο αποτελεσματικών μεθόδων και τεχνικών βαθιάς μάθησης (deep learning) ενίσχυσε σημαντικά την πρόοδο στον τομέα αυτό. Τα παραπάνω σε συνδυασμό με την πρόοδο που έχει επιτευχθεί στις κάμερες αλλά την ευκολία συλλογής και

επεξεργασίας δεδομένων έχουν οδηγήσει στην χρήση μοντέλων και μεθόδων αναγνώρισης αντικειμένων σε διάφορους τομείς όπως του αυτοκινήτου και της αυτόνομης οδήγησης, οι αποστολές έρευνας και διάσωσης, της ασφάλειας και επιτήρησης αλλά και του εντοπισμού ανωμαλιών.

Σημείο αναφοράς αποτελεί η παρουσίαση του AlexNet το 2012 [8] η οποία αποτέλεσε ουσιαστικά την αρχή της εποχής των τεχνικών βαθιάς μάθησης στην επιστήμη της όρασης υπολογιστών. Η συγκεκριμένη μέθοδος κάνοντας χρήση συντελεστικών νευρωνικών δικτύων κατάφερε, πετυχαίνοντας ακρίβεια ~85%, να ταξινομήσει της 1000 εικόνες του ImageNet κατά τη διάρκεια του διαγωνισμού ImageNet LSVRC-2012. Το ποσοστό αυτό ήταν πολύ καλύτερο από κάθε μοντέλο/αρχιτεκτονική που είχε δοκιμαστεί μέχρι εκείνη την χρονική στιγμή. Η τεχνική αυτή χρησιμοποιήθηκε μετέπειτα για τη δημιουργία νέων καλύτερων μεθόδων στον τομέα της μηχανικής μάθησης.

Ορισμένοι από τους πιο γνωστούς detectors είναι οι παρακάτω:

4. Single-shot Detector (SDD)

Ο SSD είναι ένας δημοφιλής ανιχνευτής ενός σταδίου (single-stage detector) που μπορεί να προβλέψει πολλαπλές κλάσεις [9]. Η μέθοδος ανιχνεύει αντικείμενα σε εικόνες χρησιμοποιώντας ένα ενιαίο βαθύ νευρωνικό δίκτυο (single deep learning network), διακριτοποιώντας τον χώρο εξόδου των περιγραμμάτων (bounding boxes) σε ένα σύνολο προεπιλεγμένων κουτιών σε διάφορες αναλογίες διαστάσεων και κλίμακες ανά feature map location. Ο detector υπολογίζει βαθμολογίες για την παρουσία κάθε κατηγορίας αντικειμένου σε κάθε προεπιλεγμένο πλαίσιο και προσαρμόζει το πλαίσιο για να ταιριάζει καλύτερα στο σχήμα του αντικειμένου. Επίσης, το δίκτυο συνδυάζει προβλέψεις από πολλαπλούς feature maps με διαφορετικές αναλύσεις για να χειρίζεται αντικείμενα διαφορετικών μεγεθών. Είναι σημαντικό να αναφερθεί ότι ο SSD είναι εύκολο να εκπαιδευτεί και συνεπώς να ενσωματωθεί σε συστήματα λογισμικού που απαιτούν ανίχνευση αντικειμένων. Σε σύγκριση με άλλες μεθόδους ενός σταδίου, ο SSD έχει πολύ καλύτερη ακρίβεια, ακόμη και με μικρότερα μεγέθη εικόνας εισόδου.

5. Region-based Convolutional Neural Networks (R-CNN)

Τα νευρωνικά δίκτυα που βασίζονται σε περιοχές (region-based) ή περιοχές με χαρακτηριστικά CNN (R-CNN) είναι προσεγγίσεις που εφαρμόζουν βαθιά μοντέλα στην ανίχνευση αντικειμένων [10]. Τα R-CNN μοντέλα επιλέγουν πρώτα διάφορες προτεινόμενες περιοχές από μια εικόνα (για παράδειγμα, τα κουτιά αγκύρωσης (anchor boxes) είναι ένας τύπος μεθόδου επιλογής) και στη συνέχεια επισημαίνουν (label) τις κατηγορίες και τα πλαίσια οριοθέτησής τους (π.χ. μετατοπίσεις). Τα labels αυτά δημιουργούνται με βάση προκαθορισμένες κατηγορίες που δίνονται. Στη συνέχεια, χρησιμοποιούν ένα συνελικτικό νευρωνικό δίκτυο (CNN) για να εκτελέσουν υπολογισμούς προς τα εμπρός για την εξαγωγή χαρακτηριστικών από κάθε προτεινόμενη περιοχή. Στο R-CNN, η εισαγόμενη εικόνα διαιρείται πρώτα σε σχεδόν δύο χιλιάδες τμήματα περιοχών και στη συνέχεια εφαρμόζεται

ένα CNN για κάθε περιοχή, αντίστοιχα. Το μέγεθος των περιοχών υπολογίζεται και η σωστή περιοχή εισάγεται στο νευρωνικό δίκτυο. Όπως είναι όμως λογικό μια τόσο αναλυτική μέθοδος μπορεί να σημαντικούς χρονικούς περιορισμούς στην όλη διαδικασία. Ο χρόνος εκπαίδευσης είναι σημαντικά μεγαλύτερος σε σύγκριση με το YOLO. Αυτό ουσιαστικά συμβαίνει διότι ταξινομεί και δημιουργεί bounding boxes αλλά και από το γεγονός ότι ένα νευρωνικό δίκτυο εφαρμόζεται σε μία περιοχή κάθε φορά.

Το 2015, αναπτύχθηκε το Fast R-CNN [10] με σκοπό την μειώσει του χρόνου εκπαίδευσης του δικτύου. Ενώ το αρχικό R-CNN υπολόγιζε ανεξάρτητα τα χαρακτηριστικά του νευρωνικού δικτύου σε κάθε μία από τις δύο χιλιάδες περιοχές ενδιαφέροντος, το Fast R-CNN εκτελεί το νευρωνικό δίκτυο μία φορά σε ολόκληρη την εικόνα. Κάτι τέτοιο είναι συγκρίσιμο με την αρχιτεκτονική του YOLO. Ωστόσο, το YOLO παραμένει μια ταχύτερη εναλλακτική λύση του Fast R-CNN λόγω της απλότητας του. Επιπλέον, στο τέλος του Fast-R-CNN υπάρχει μια νέα μέθοδος γνωστή ως Region of Interest (ROI) Pooling, η οποία αποκόπτει κάθε περιοχή ενδιαφέροντος από τον τανυστή εξόδου (output tensor) του δικτύου, την αναδιαμορφώνει και την ταξινομεί (Image Classification). Αυτή η διαφοροποίηση καθιστά το Fast R-CNN πιο ακριβές από το αρχικό R-CNN.

6. You Only Look Once (YOLO)

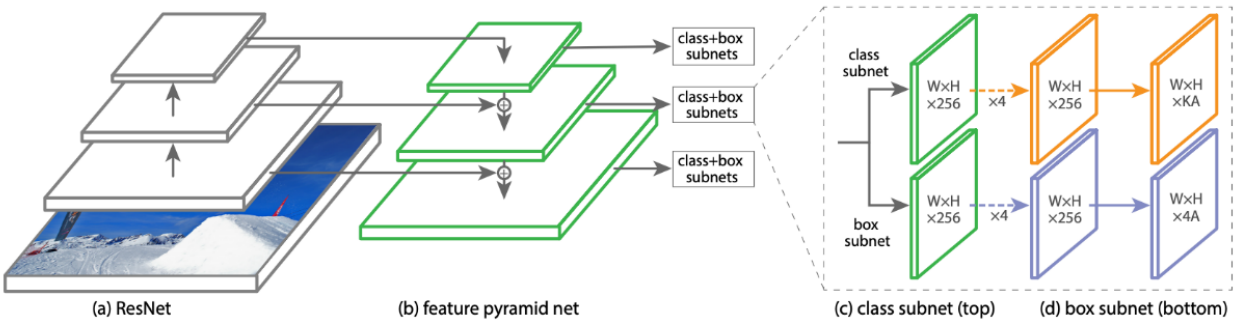
Το YOLO αποτελεί έναν από τους δημοφιλέστερους αλγορίθμους ανίχνευσης αντικειμένων ο οποίος χρησιμοποιείται σε πληθώρα εμπορικών προϊόντων που κάνουν χρήση μηχανικής όρασης. Ο πρώτος YOLO detector παρουσιάστηκε για πρώτη φορά το 2016 [11], προτείνοντας μια νέα αρχιτεκτονική η οποία ήταν σημαντικά γρηγορότερη από οποιαδήποτε άλλη μέχρι εκείνη τη χρονική στιγμή. Η μέθοδος δεν απαιτεί την ύπαρξη δύο ξεχωριστών δικτύων στην αρχιτεκτονική της σε αντίθεση με της two-stage μεθόδους, αντιμετωπίζοντας το πρόβλημα της ανίχνευσης αντικειμένων ως πρόβλημα παλινδρόμησης. Αποτέλεσε σημείο αναφοράς στην ανίχνευση αντικειμένων αφού ήταν η πρώτη single-stage μέθοδος.

Από τότε παρουσιάστηκαν πολλές νέες εκδόσεις και παραλλαγές του YOLO, κάθε μια από τις οποίες ενίσχυε σημαντικά την απόδοση και την επίδοση στην ανίχνευση αντικειμένων συγκριτικά με τις προηγούμενες. Ωστόσο, παρά την πολύ καλή απόδοση της μεθόδου τόσο σε ταχύτητα και ακρίβεια πρόβλεψης το YOLO είναι λιγότερο ακριβές από τις two-stage μεθόδους, κυρίως στο κομμάτι της αναγνώρισης πολύ μικρών αντικειμένων ή αντικειμένων με ακανόνιστο σχήμα. Το συγκεκριμένο πρόβλημα προσπαθούν να επιλύσουν οι νεότερες εκδόσεις του YOLO.

7. Retina-Net

Το RetinaNet είναι ένα ακόμη μοντέλο ανίχνευσης αντικειμένων ενός σταδίου [12]. Χρησιμοποιεί μια συνάρτηση εστιακής απώλειας (focal loss function) για την αντιμετώπιση της ανισορροπίας των κλάσεων κατά τη διάρκεια της εκπαίδευσης. Η εστιακή απώλεια εφαρμόζει έναν διαμορφωτικό όρο (modulating term) στην απώλεια διασταυρούμενης

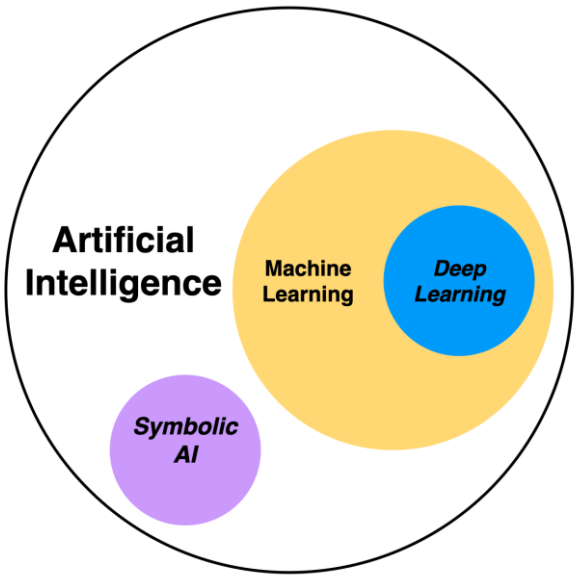
εντροπίας (cross entropy loss) προκειμένου να εστιάζει τη μάθηση σε “σκληρά” αρνητικά παραδείγματα. Το RetinaNet αποτελεί ένα ενιαίο, ενοποιημένο δίκτυο που αποτελείται από ένα δίκτυο κορμού (backbone) και δύο υποδίκτυα (subnets) ειδικών εργασιών (task-specific subnets). Το backbone είναι υπεύθυνο για τον υπολογισμό ενός χάρτη συνελικτικών χαρακτηριστικών (convolutional feature map) σε ολόκληρη την εικόνα εισόδου ενώ αποτελεί ένα off-the-self συνελικτικό δίκτυο. Το πρώτο subnet εκτελεί συνελικτική ταξινόμηση αντικειμένων στην έξοδο του backbone – το δεύτερο subnet εκτελεί συνελικτική παλινδρόμηση οριοθετημένου πλαισίου (convolutional bounding box regression).



Εικόνα 2.6. Σχηματική αναπαράσταση της αρχιτεκτονικής του RetinaNet [12].

Μηχανική Μάθηση

Τα τελευταία χρόνια, ο τομέας της Μηχανικής Μάθησης (Machine Learning), ο οποίος αποτελεί τμήμα ενός ευρύτερου πεδίου (Εικόνα 2.7), αυτού της τεχνητής νοημοσύνης (Artificial Intelligence) εμφανίζει ραγδαία αύξηση.



Εικόνα 2.7. Artificial Intelligence - Machine Learning - Deep Learning και Symbolic AI [13].

Σύμφωνα με τον Mitchell [14] ένα πρόγραμμα υπολογιστή μπορεί να μάθει από μία εμπειρία Ε ως προς κάποια κλάση εργασιών Τ και μέτρο απόδοσης Α, όταν η απόδοση του από το Τ η οποία μετράται από το Α, βελτιώνεται μέσω της εμπειρίας Ε. Ένας δεύτερος αλλά πιο γενικός ορισμός περιγράφει τη μηχανική μάθηση ως το πεδίο μελέτης που δίνει στους ηλεκτρονικούς υπολογιστές την ικανότητα να μαθαίνουν, χωρίς όμως αυτοί να έχουν ρητά προγραμματιστεί [15].

Στην πράξη, η μηχανική μάθηση ενσωματώνει βασικές αρχές της επιστήμης των υπολογιστών και της στατιστικής για τη δημιουργία στατιστικών μοντέλων, τα οποία χρησιμοποιούνται για την πρόβλεψη και την εξαγωγή συμπερασμάτων (inference). Αυτά τα μοντέλα περιλαμβάνουν σύνολα μαθηματικών σχέσεων που συνδέουν τις εισόδους και τις εξόδους ενός δεδομένου συστήματος που αναλύεται.

Ανάλογα με τη μέθοδο που χρησιμοποιείται για τη ”μάθηση”, η μηχανική μάθηση μπορεί να χωριστεί σε τρεις βασικές κατηγορίες [14]:

1. Επιτηρούμενη ή Επιβλεπόμενη Μάθηση (Supervised Learning).
2. Μη Επιτηρούμενη ή Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning).
3. Ενισχυτική Μάθηση (Reinforcement Learning).

Βαθιά Μάθηση

Η βαθιά μάθηση (deep learning) αποτελεί ένα σύνολο μεθόδων οι οποίες επιχειρούν να μοντελοποιήσουν δεδομένα συσχετίζοντας διαφορετικούς μη γραμμικούς μετασχηματισμούς. Αποτελεί μία υποκατηγορία της μηχανικής μάθησης, η οποία βασίζεται στα νευρωνικά δίκτυα. Η κύρια διαφορά της βαθιάς μάθησης από άλλες μεθόδους μηχανικής μάθησης είναι η ικανότητά της να εκτελεί αυτόματα τη διαδικασία της εξαγωγής χαρακτηριστικών (feature extraction) χρησιμοποιώντας πολλαπλά επίπεδα αφαίρεσης. Αυτό επιτυγχάνεται μέσω της εφαρμογής σειρών μη γραμμικών μετασχηματισμών στα δεδομένα εισόδου, επιτρέποντας την ανακάλυψη περίπλοκων δομών και σχέσεων των δεδομένων. Αυτές οι τεχνικές επέτρεψαν σημαντική πρόοδο στους τομείς της επεξεργασίας ήχου και εικόνας, συμπεριλαμβανομένης της αναγνώρισης του προσώπου (face recognition), της αναγνώρισης ομιλίας, της μηχανικής όρασης (machine vision), της αυτοματοποιημένης επεξεργασίας γλώσσας, της ταξινόμησης κειμένων (για παράδειγμα, της αναγνώρισης ανεπιθύμητων μηνυμάτων).

Υπάρχουν διάφορα είδη αρχιτεκτονικών για νευρωνικά δίκτυα:

- **Πολυεπίπεδα Στρώματα (Multi-level Perspectors - MLPs)**

  Τα πολυεπίπεδα στρώματα που είναι τα παλαιότερα και απλούστερα μοντέλα νευρωνικών δικτύων. Αποτελείται από πολλαπλά επίπεδα νευρώνων που συνδέονται πλήρως μεταξύ τους.

- **Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs)**

Ιδανικά για την επεξεργασία οπτικών δεδομένων, χάρη στην ικανότητά τους να αναγνωρίζουν patterns σε εικόνες, όπως άκρες ή γωνίες.

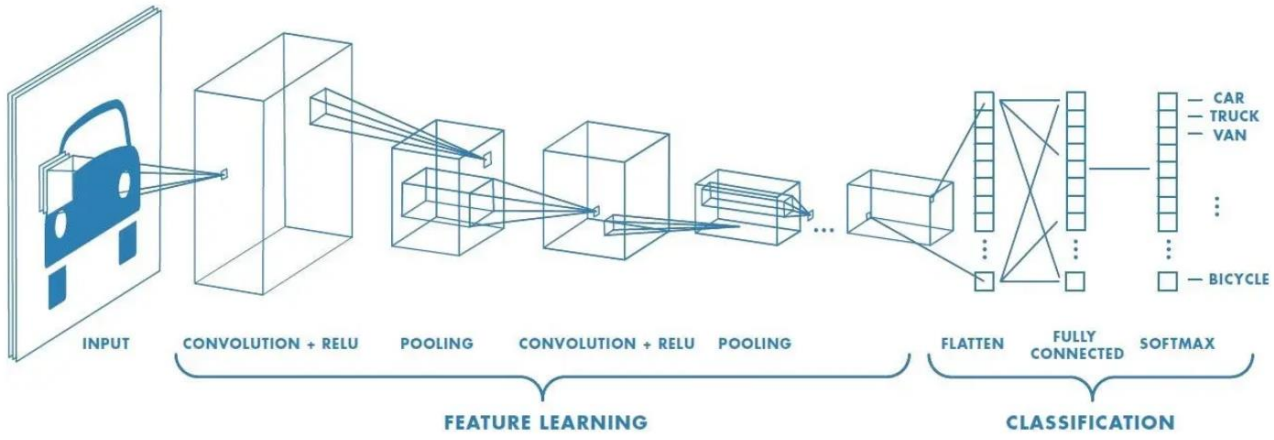
- **Επαναληπτικά νευρωνικά δίκτυα (Recurrent Neural Networks - RNNs)**

  Χρησιμοποιούνται για δεδομένα σειράς (sequential data), όπως κείμενο ή χρονοσειρές, λόγω της δυνατότητάς τους να διατηρούν μνήμη των προηγούμενων εισόδων στην αλυσίδα δεδομένων.

Πρόσφατες εξελίξεις επίσης περιλαμβάνουν την ανάπτυξη των **Δικτύων Βαθιάς Ενισχυτικής Μάθησης (Deep Reinforcement Learning Networks)**, τα οποία συνδυάζουν τη βαθιά μάθηση με τεχνικές reinforcement learning, επιτρέποντας τη δημιουργία συστημάτων που μπορούν να μαθαίνουν πώς να λαμβάνουν αποφάσεις βασισμένες σε περίπλοκα περιβάλλοντα.

Συνελικτικό Νευρωνικό Δίκτυο (CNN)

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks) πήραν το όνομά τους από τη μαθηματική γραμμική λειτουργία μεταξύ πινάκων που ονομάζεται συνέλιξη. Στην πραγματικότητα, περιλαμβάνουν συνελικτικά επίπεδα (convolutional layers) τα οποία λαμβάνουν ως είσοδο κάποιο διάνυσμα και εκτελούν τη πράξη της συνέλιξης μεταξύ της εισόδου και του φίλτρου τους. Περιλαμβάνουν και άλλα πολλαπλά επίπεδα, μεταξύ των οποίων βρίσκονται τα επίπεδα υποδειγματοληψίας (pooling layers) και τα πλήρως συνδεδεμένα επίπεδα (fully connected layer) [16]. Μια τυπική μορφή ενός συνελικτικού νευρωνικού δικτύου που δέχεται ως είσοδο μια εικόνα παρουσιάζεται στο διάγραμμα της παρακάτω εικόνας [16]:



Εικόνα 2.8. Τυπική δομή ενός συνελικτικού νευρονικού δικτύου [16].

Τα συνελικτικά δίκτυα χρησιμοποιούνται κυρίως σε προβλήματα που σχετίζονται με χωρική πληροφορία και το μεγαλύτερο πλεονέκτημά τους συγκριτικά με τα απλά νευρωνικά δίκτυα είναι η μείωση του αριθμού των παραμέτρων που απαιτούνται να υπολογιστούν μέσω της διαδικασίας εκπαίδευσης. Ετσι, τα δίκτυα αυτά έδωσαν τη δυνατότητα καλύτερης προσέγγισης πολύ μεγαλύτερων μοντέλων προκειμένου να επιλυθούν πολύπλοκες διεργασίες, οι οποίες δεν ήταν δυνατές με χρήση των κλασικών ANNs. ΄Εχουν εξαιρετική απόδοση σε εφαρμογές που

ασχολούνται με δεδομένα εικόνων, μιας και η αρχιτεκτονική τους είναι ιδανική για την εύρεση μοτίβων και βασικών χαρακτηριστικών σε χωρικά δεδομένα, αλλά και σε εφαρμογές που σχετίζονται με επεξεργασία φυσικής γλώσσας (NLP – Natural Language Processing). Επιπλέον, χρησιμοποιούνται αποτελεσματικά και για την επεξεργασία άλλων δεδομένων, όπως για παράδειγμα οι χρονοσειρές, ο ήχος, καθώς και άλλα είδη σημάτων. Μια δεδομένη συνθήκη που πρέπει να πληρούν τα προβλήματα που επιλύονται με χρήση συνελικτικών νευρωνικών δικτύων είναι πως είναι αναγκαίο να μην έχουν χαρακτηριστικά που έχουν εξάρτηση από το χώρο. Χαρακτηριστικό παράδειγμα αποτελεί πως σε μια εφαρμογή που έχει ως σκοπό την ανίχνευση προσώπου δεν θα δοθεί έμφαση στη θέση που βρίσκεται το πρόσωπο μέσα στην κάθε εικόνα, αλλά στην ανίχνευσή του στην δεδομένη εικόνα.

Ένα τέτοιο νευρωνικό δίκτυο αποτελείται από ένα ή περισσότερα επίπεδα συνέλιξης (convolution layers) τα οποία ακολουθούνται από επίπεδα υποδειγματοληψίας (pooling layers). Τα επίπεδα αυτά μειώνουν τις διαστάσεις των feature maps που προκύπτουν από τα συνελικτικά στρώματα, ενώ τις περισσότερες φορές, στο τέλος του δικτύου εντοπίζονται πλήρη συνδεδεμένα επίπεδα (fully connected layers). Στα αρχικά επίπεδα, το δίκτυο μαθαίνει να αναγνωρίζει κάποιες βασικές δομές των εισόδων, ενώ στα βαθύτερα layers αυξάνεται η πολυπλοκότητα και εξάγονται συμπεράσματα για τα πιο σύνθετα χαρακτηριστικά. Τα πλήρως συνδεδεμένα επίπεδα στο τέλος του δικτύου βοηθούν στην τελική ταξινόμηση της εισόδου σε συγκεκριμένες κλάσεις.

Για συγκεκριμένους τύπους δεδομένων, και ιδιαίτερα για τις εικόνες, τα πολυστρωματικά νευρωνικά δίκτυα (multi-layer neural networks) δεν είναι ιδανικά. Αυτά τα δίκτυα απαιτούν τη μετατροπή των εικόνων σε διανύσματα, χάνοντας έτσι χωρικές πληροφορίες. Τα CNNs, από την άλλη, μπορούν να ενεργούν απευθείας σε πίνακες ή ακόμα και σε πολυδιάστατες συστοιχίες (tensor) για εικόνες με τρία κανάλια χρώματος RGB, διατηρώντας την χωρική πληροφορία και βελτιώνοντας την ακρίβεια. Πριν την εμφάνιση της βαθιάς μάθησης, η υπολογιστική όραση βασιζόταν σε χειροκίνητη εξαγωγή χαρακτηριστικών, αλλά τα CNNs έχουν αυτοματοποιήσει αυτή τη διαδικασία, επιτρέποντας την ευρεία χρήση τους σε ταξινόμηση, καταμερισμό, αναγνώριση αντικειμένων και προσώπων.

Ένα CNN αποτελείται από διάφορα είδη στρωμάτων, τα οποία περιγράφονται συνοπτικά στις παρακάτω ενότητες: στρώματα συνέλιξης, στρώματα συγκέντρωσης και πλήρως συνδεδεμένα στρώματα.

Στρώμα Συνέλιξης (Convolution layer)

Το στρώμα συνέλιξης (convolutional kernel or layer) αποτελεί τον πυρήνα των CNNs. Η βασική λειτουργία του συνελικτικού επιπέδου είναι η εκτέλεση μιας συνέλιξης (convolution) μεταξύ των εισερχόμενων δεδομένων και ενός συνόλου φίλτρων ή πυρήνων (kernels), με σκοπό την εξαγωγή χαρακτηριστικών από τα δεδομένα. Η συνέλιξη είναι μια μαθηματική πράξη που περιγράφει τον τρόπο με τον οποίο δύο σήματα (ή συναρτήσεις) συνδυάζονται για να παράγουν ένα τρίτο σήμα. Στο πλαίσιο των CNNs, η συνέλιξη χρησιμοποιείται για να εφαρμοστούν φίλτρα πάνω στις εικόνες με σκοπό την ανίχνευση συγκεκριμένων χαρακτηριστικών, όπως άκρες, γωνίες ή άλλες υφές.

Η πράξη της συνέλιξης περιγράφεται από την ακόλουθη μαθηματική έκφραση:

$$y[k] = (x * h)[k][n] = \sum h[i][j]x[k-i][n-j]$$

Σε ένα συνελικτικό επίπεδο, κάθε φίλτρο είναι ένας πυρήνας (kernel), ο οποίος "σαρώνει" την εικόνα εισόδου ή το feature map από το προηγούμενο επίπεδο του δικτύου, εφαρμόζοντας την πράξη της συνέλιξης σε κάθε τοπική περιοχή των δεδομένων. Αυτή η διαδικασία παράγει έναν νέο χάρτη χαρακτηριστικών που αποτελείται από τις αποκρίσεις του φίλτρου σε κάθε τοπική περιοχή, καταδεικνύοντας πού και πόσο έντονα εμφανίζονται τα χαρακτηριστικά που αναγνωρίζει το φίλτρο στα δεδομένα.

Το μέγεθος του φίλτρου (kernel size) αποτελεί ένα πολύ μικρό κλάσμα των δεδομένων εισόδου και υλοποιεί τμηματικά την πράξη της συνέλιξης κατά μήκος του συνολικού εύρους των δεδομένων. Η μετάβαση του φίλτρου από το ένα τμήμα των δεδομένων εισόδου στο διπλανό του ρυθμίζεται από το βήμα του φίλτρου (kernel stride). Το βήμα αυτό καθορίζει πόσες χωρικές μονάδες θα μετακινείται το φίλτρο του συνελικτικού επιπέδου πάνω στα δεδομένα εισόδου οριζόντια και κάθετα. Όταν η τιμή του είναι μεγάλη μειώνεται ο συνολικός χρόνος της πράξης της συνέλιξης. Για τον λόγο αυτό το βήμα προτιμάται να έχει τιμή μεγαλύτερη της μονάδας στις περιπτώσεις ύπαρξης μεγάλου όγκου δεδομένων και διαστάσεων. Στις περιπτώσεις που είναι αναγκαία η διατήρηση των διαστάσεων της εισόδου και της εφαρμογής δεδομένου μεγέθους φίλτρου για τη σάρωση των χαρακτηριστικών της τότε εφαρμόζονται τεχνικές padding. Συγκεκριμένα, χρησιμοποιείται zero-padding υπερπαράμετρος με σκοπό την προσθήκη μηδενικών στο περίγραμμα της εισόδου.

Τελικό αποτέλεσμα της συνελικτικής διαδικασίας των δεδομένων εισόδου με το φίλτρο αποτελεί ο χάρτης χαρακτηριστικών ο οποίος περιλαμβάνει συσχετίσεις χωρικής πολυπλοκότητας, η οποία αποικοδομείται από τα συνελικτικά στρώματα στα κανάλια εξόδου. Η ίδια διαδικασία επαναλαμβάνεται με διαφορετικά φίλτρα ώστε να παραχθεί ένα σύνολο χαρτών που αναπαριστούν διαφορετικά χαρακτηριστικά της εισόδου. Είναι σημαντικό να αναφερθεί ότι συνήθως η είσοδος ενός συνελικτικού επιπέδου είναι τρισδιάστατη ενώ το μέγεθος της εξόδου του επιπέδου ελέγχεται από τις υπερπαραμέτρους (kernel size, kernel stride και zero-padding). Το μέγεθος της εξόδου ενός συνελικτικού επιπέδου δίνεται από την ακόλουθη σχέση:

$$O = \frac{I - K + 2P}{S} + 1$$

Όπου O το μέγεθος της εξόδου (output size), I το μέγεθος της εισόδου (input size), K το μέγεθος του φίλτρου (kernel size), S το kernel stride και P το padding.

Οι πυρήνες των φίλτρων στα συνελικτικά επίπεδα δεν προγραμματίζονται με συγκεκριμένες τιμές. Αντίθετα, οι τιμές τους (βάρη) προσαρμόζονται μέσω της διαδικασίας της εκπαίδευσης του

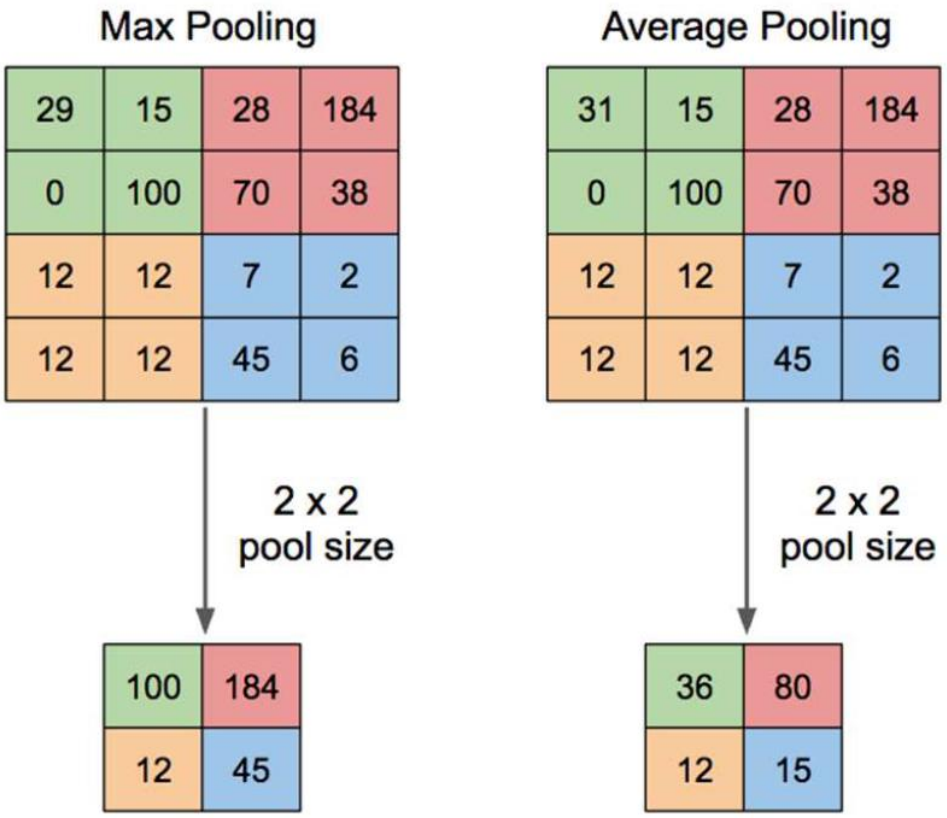
δικτύου, επιτρέποντας την αυτόματη μάθηση των βέλτιστων φίλτρων για την εξαγωγή χαρακτηριστικών από τα δεδομένα.

Στρώμα Συγκέντρωσης (Pooling layer)

Το συνελικτικό στρώμα περιλαμβάνει επίσης στρώματα συγκέντρωσης ή αλλιώς υποδειγματοληψίας (pooling layers), τα οποία επιτρέπουν τη μείωση της διάστασης μέσω υποδειγματοληψίας, λαμβάνοντας το μέσο όρο ή το μέγιστο από τα εμβαδά της εικόνας (average pooling ή max pooling), όπως φαίνεται στην εικόνα παρακάτω [17]. Παρόμοια με τα συνελικτικά στρώματα, τα στρώματα συγκέντρωσης λειτουργούν σε μικρές περιοχές της εικόνας και περιλαμβάνουν επικάλυψη - stride. Για παράδειγμα, αν εξετάσουμε 2×2 εμβαδά (patches) και λάβουμε τη μέγιστη τιμή για να καθορίσουμε την έξοδο, με βήμα $s = 2$, το πλάτος και το ύψος της εικόνας μειώνονται κατά το ήμισυ. Επιπλέον, η διάσταση μπορεί να μειωθεί και με το συνελικτικό στρώμα, χρησιμοποιώντας βήμα μεγαλύτερο από 1 και χωρίς μηδενικό γέμισμα (zero padding). Ένα επιπλέον πλεονέκτημα του pooling είναι ότι καθιστά το δίκτυο λιγότερο ευαίσθητο σε μικρές μετατοπίσεις των εικόνων εισόδου.

Ένα επίπεδο υποδειγματοληψίας λαμβάνει ως είσοδο τον χάρτη χαρακτηριστικών (feature map) από το προηγούμενο συνελικτικό επίπεδο και έχει σκοπό τη μείωση των διαστάσεών του χωρίς σημαντική απώλεια χρήσιμης πληροφορίας. Έτσι, βελτιώνεται ο χρόνος επεξεργασίας διατηρώντας την απαραίτητη πληροφορία του χάρτη. Κατά τη διαδικασία υποδειγματοληψίας, ένα φίλτρο που αντιστοιχεί σε μια μαθηματική συνάρτηση εφαρμόζεται σε όλο τον χάρτη για να υπολογιστεί η υποδειγματοληπτημένη εκδοχή του. Οι δύο κύριες συναρτήσεις που χρησιμοποιούνται είναι η max pooling και η average pooling. Η max pooling επιλέγει το μέγιστο στοιχείο του χάρτη όπου εφαρμόζεται το φίλτρο, ενώ η average pooling επιλέγει τον μέσο όρο των στοιχείων του.

Στην παρακάτω εικόνα (Εικόνα 2.9) παρουσιάζεται ένα παράδειγμα max και average pooling:



Εικόνα 2.9. Παραδείγματα max pooling και average pooling [17].

Η έξοδος του επιπέδου ελέγχεται τόσο από το μέγεθος του φίλτρου (pool size) όσο και από το άλμα του φίλτρου (pool stride). Μαθηματικά περιγράφεται από την παρακάτω σχέση:

$$O = \frac{I - P_s}{S} + 1$$

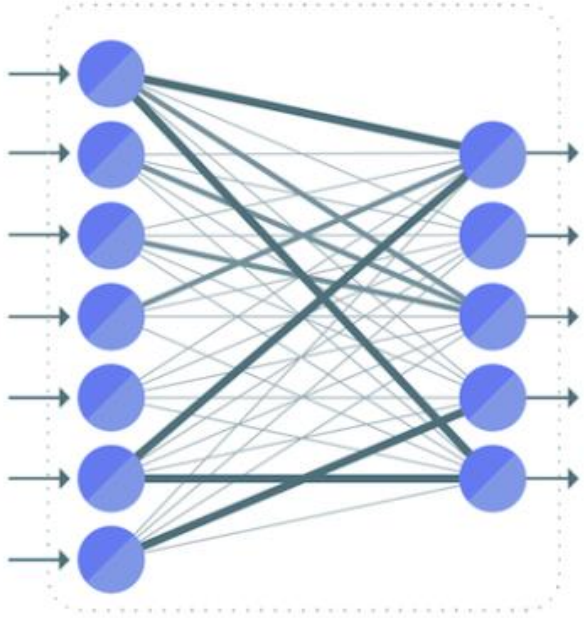
Όπου O το μέγεθος της εξόδου του επιπέδου (output size), I το μέγεθος της εισόδου του επιπέδου (input size), P_S το μέγεθος του φίλτρου (pool size) και S το kernel stride.

Πλήρως Συνδεδεμένο Στρώμα (Fully Connected Layer)

Μετά από αρκετά στρώματα συνέλιξης και συγκέντρωσης, το CNN συνήθως τελειώνει με αρκετά πλήρως συνδεδεμένα στρώματα (fully connected layers). Η πολυδιάστατη συστοιχία (tensor) που έχουμε στην έξοδο αυτών των στρωμάτων μετατρέπεται σε διάνυσμα και στη συνέχεια προσθέτουμε αρκετά στρώματα perceptron [18].

Στα πλήρως συνδεδεμένα επίπεδα, κάθε κόμβος (νευρώνας) του προηγούμενου επιπέδου συνδέεται απευθείας με κάθε κόμβο του επόμενου. Είναι συχνό φαινόμενο να τοποθετούνται στο τέλος του συνελικτικού νευρωνικού δικτύου για να επιτευχθεί ταξινόμηση της εισόδου σε κάποια από τις πιθανές κλάσεις, χρησιμοποιώντας τους χάρτες χαρακτηριστικών (feature maps) που προέκυψαν από τα προηγούμενα επίπεδα. Το τελευταίο τέτοιο επίπεδο περιέχει αριθμό κόμβων ίσο με τις αρχικές κλάσεις που είχε το πρόβλημα ταξινόμησης και καθορίζει το μοντέλο απόφασης του δικτύου. Με άλλα λόγια, το μοντέλο απόφασης του δικτύου καθορίζεται στα πλήρως

συνδεδεμένα επίπεδα. Ένα παράδειγμα πλήρως συνδεδεμένου επίπεδου απεικονίζεται στη παρακάτω εικόνα (Εικόνα 2.10) [18]:



Εικόνα 2.10. Παράδειγμα ενός πλήρους συνδεμένου επιπέδου (fully connected layer) που υλοποιεί ταξινόμηση σε πέντε κλάσεις [18].

Συνάρτηση Ενεργοποίησης (Activation Layer)

Οι συναρτήσεις ενεργοποίησης (activation functions) είναι «σταθερές» συναρτήσεις οι οποίες αποτελούν δομικά στοιχεία στην αρχιτεκτονική των νευρωνικών δικτύων καθώς στόχος τους είναι ο προσδιορισμός της κατάστασης ενεργοποίησης των νευρώνων του δικτύου. Δέχονται ως είσοδο το άθροισμα των γινωμένων εισόδων-βαρών ενώ η έξοδος της συνάρτησης μπορεί να είναι η τελική έξοδος του δικτύου ή μία ενδιάμεση έξοδος η οποία χρησιμοποιείται στη συνέχεια ως είσοδος από κάποιον άλλο νευρώνα του νευρωνικού δικτύου. Επομένως, είναι συναρτήσεις οι οποίες χαρακτηρίζουν το δίκτυο στο σύνολό του και χρησιμοποιούνται για τον έλεγχο των εξόδων του.

Μία τέτοια συνάρτηση μπορεί να είναι είτε γραμμική είτε μη γραμμική. Η μορφή της εξαρτάται από τον τύπο της συνάρτησης που την αντιπροσωπεύει. Υπάρχουν διάφορες μορφές τέτοιων συναρτήσεων. Η επιλογή μεταξύ αυτών εξαρτάται τόσο από τον τύπο της εφαρμογής που θα χρησιμοποιηθεί το εκάστοτε νευρωνικό δίκτυο, αλλά και από την αρχιτεκτονική του δικτύου που έχει επιλεχτεί. Η εισαγωγή μη γραμμικότητας στις εξόδους των νευρώνων του δικτύου αποτελεί έναν ακόμη βασικό λόγο χρησιμοποίησης τέτοιων συναρτήσεων [19]. Η μη γραμμικότητα κυριαρχεί κατά κόρων στη φύση και αυτό έχει ως αποτέλεσμα σχεδόν όλα τα πραγματικά δεδομένα να είναι μη γραμμικά. Συνεπώς, κάνοντας χρήση συναρτήσεων ενεργοποίησης η διαδικασία της μάθησης μπορεί να ενισχυθεί και ταυτόχρονα να αυξηθεί σημαντικά η ακρίβεια των αποτελεσμάτων.

Μερικές κοινές συναρτήσεις ενεργοποίησης που ενσωματώνονται συχνά στα νευρωνικά δίκτυα είναι οι παρακάτω:

Σιγμοειδής ή Λογιστική συνάρτηση (Logistic ή Sigmoid)

Η σιγμοειδής ή λογιστική συνάρτηση (Logistic or Sigmoid function) είναι μια από τις πιο διαδεδομένες μη γραμμικές συναρτήσεις ενεργοποίησης, ιδιαίτερα στα feedforward μοντέλα [19]. Χαρακτηριστικό της είναι ότι λαμβάνει οποιαδήποτε πραγματική τιμή και την «περιορίζει» σε μια τιμή μεταξύ 0 και 1. Κάτι τέτοιο είναι ιδιαίτερα χρήσιμο σε προβλήματα ταξινόμησης δύο κλάσεων, όπου η έξοδος του νευρωνικού δικτύου ερμηνεύεται ως η πιθανότητα η δεδομένη είσοδος να ανήκει στην θετική κλάση.

Μαθηματικά, η λογιστική συνάρτηση ορίζεται ως:

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

όπου x είναι η είσοδος στην συνάρτηση.

Η συνάρτηση αυτή ωστόσο δεν είναι εξαιρετική για όλου του είδους τις εφαρμογές, καθώς παρουσιάζει κάποιες προκλήσεις όπως το φαινόμενο της εξαφάνισης των gradient (vanishing gradients), ιδίως σε βαθιά δίκτυα (deep networks), λόγω της πολύ μικρής κλίσης της συνάρτησης στα άκρα της. Αυτό μπορεί να καταστήσει δύσκολη την εκπαίδευση των νευρωνικών δικτύων μέσω backpropagation, καθώς οι διορθώσεις που πρέπει να γίνουν στα βάρη κατά την εκπαίδευση μπορεί να είναι πολύ μικρές για να είναι αποτελεσματικές. Παρ' όλα αυτά, η σιγμοειδής συνάρτηση παραμένει μια καλή επιλογή σε πολλές εφαρμογές λόγω της ικανότητάς της να μοντελοποιεί πιθανοτικά αποτελέσματα και να επιτυγχάνει μη-γραμμικότητα στα νευρωνικά δίκτυα.

Υπερβολική Εφαπτομένη (Hyperbolic Tangent - tanh)

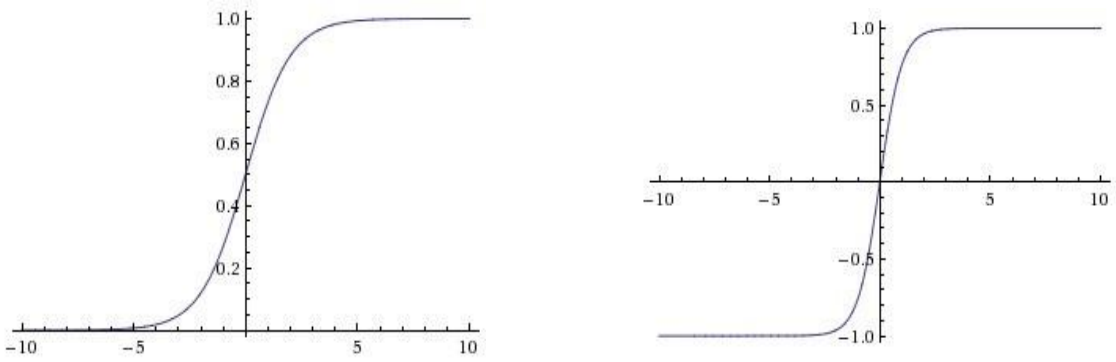
Η υπερβολική εφαπτομένη (tanh) αποτελεί μία έκδοση της λογιστικής σιγμοειδούς συνάρτησης αφού η έξοδος της βρίσκεται εντός του διαστήματος $[-1,1]$. Συναρτήσεις αυτού το είδους χρησιμοποιούνται κυρίως σε αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks - RNNs) [20]. Είναι προτιμητέα από τη σιγμοειδή καμπύλη καθώς το σύνολο των τιμών της έχει κέντρο το μηδέν. Το γεγονός αυτό διευκολύνει την εκπαίδευση του δικτύου καθώς αποτρέπει την ανανέωση των παραμέτρων σε μια μόνο κατεύθυνση.

Μαθηματικά, η συνάρτηση tanh ορίζεται ως:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Η tanh μοιράζεται κάποιες ιδιότητες με τη σιγμοειδή, όπως τη μη-γραμμικότητα. Ωστόσο, επειδή το σύνολο τιμών της έχει κέντρο το μηδέν, συνήθως οδηγεί σε ταχύτερη σύγκλιση κατά τη διάρκεια της εκπαίδευσης και μειώνει την πιθανότητα εμφάνισης του φαινομένου vanishing gradients. Παρά

τα πλεονεκτήματα αυτά, η tanh έχει επίσης περιορισμούς. Για παράδειγμα, σε εισόδους με μεγάλες απόλυτες τιμές, μπορεί να οδηγήσει σε κλίσεις κοντά στο μηδέν. Αυτό μπορεί να επιβραδύνει ή ακόμα και να σταματήσει τη διαδικασία εκπαίδευσης για ορισμένα layers του δικτύου.



Εικόνα 2.11. (Αριστερά) Σιγμοειδής συνάρτηση, (Δεξιά) Υπερβολική εφαπτομένη [22].

ReLU (Rectified Linear Units)

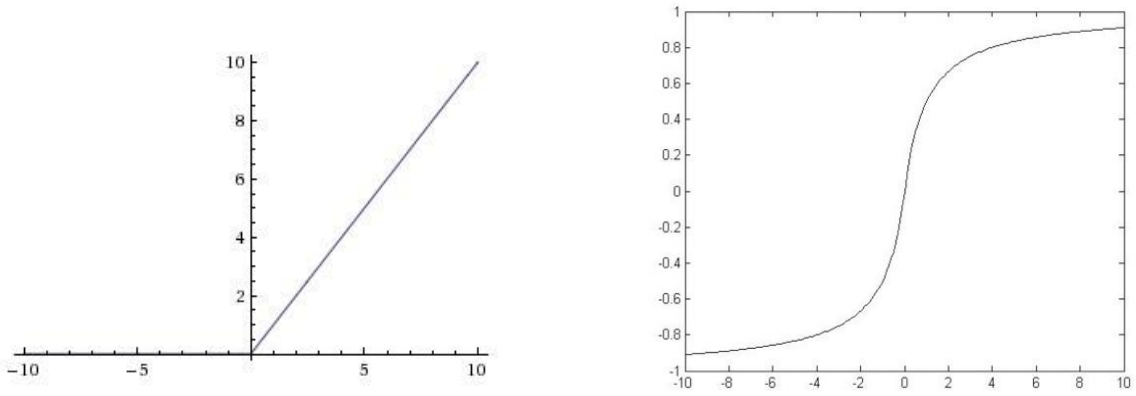
Η ReLU αποτελεί ίσως την πιο συχνά χρησιμοποιούμενη συνάρτηση ενεργοποίησης τόσο λόγω της αποτελεσματικότητας, αφού μπορεί να αποδώσει καλύτερα στις περισσότερες των περιπτώσεων όσο και της απλότητας της. Αποτελεί μία γραμμική συνάρτηση [21], [22] η οποία δεν είναι υπολογιστικά ακριβή και επιπλέον συγκλίνει γρήγορα. Η ReLU επιστρέφει την τιμή εισόδου για κάθε θετική είσοδο και μηδέν διαφορετικά. Αυτή η απλοϊκή λειτουργία επιταχύνει τη διαδικασία εκπαίδευσης και βοηθά στην αντιμετώπιση του κορεσμού. Ο κορεσμός αποτελεί το πιο σημαντικό πρόβλημα των σιγμοειδών συναρτήσεων. Συναρτήσεις ενεργοποίησης όπως η υπερβολική εφαπτομένη και η λογιστική συνάρτηση κορέζονται στο -1, το 0 ή στο 1, γεγονός που είναι πολλές φορές ανεπιθύμητο. Ωστόσο, η ReLU δεν είναι ιδανική και αυτό διότι στις περιπτώσεις που οι είσοδοι έχουν αρνητικές τιμές η έξοδός της θα είναι πάντα μηδενική με αποτέλεσμα οι αντίστοιχοι νευρώνες να μην ενεργοποιούνται ποτέ, εμποδίζοντας τη μάθηση. Παραλλαγές όπως η Leaky ReLU και η Parametric ReLU αντιμετωπίζουν αυτό το πρόβλημα, διατηρώντας έτσι τους νευρώνες ενεργούς [23].

$$\mathrm{ReLU} = \max(x, 0)$$

Soft sign

Η συνάρτηση αυτή βρίσκει εφαρμογή κυρίως σε δίκτυα βαθιάς μάθησης (deep learning) [24]. Οι ιδιότητές της είναι παρόμοιες με αυτής της υπερβολικής εφαπτομένης αφού οι έξοδοι και των δύο περιορίζονται μεταξύ -1 και 1. Οι διαφορές τους έγκειται κυρίως στο ότι η soft-sign έχει τετραγωνικά πολυώνυμα αντί για εκθετικά. Έτσι, δεν πλησιάζει ασυμπτωτικά τα όριά της τόσο απότομα όσο η σιγμοειδής ή η tanh, κάτι που μπορεί να βοηθήσει στην μετρίαση του προβλήματος της εξαφάνισης των κλίσεων σε κάποιο βαθμό.

$$f(x) = \frac{x}{1 + |x|}$$



Εικόνα 2.12. (Αριστερά) ReLU, (Δεξιά) Soft-sign.

Συναρτήσεις Απώλειας (Loss Functions)

Οι συναρτήσεις απώλειας (loss functions), γνωστές επίσης και ως συναρτήσεις κόστους αποτελούν βασικό στοιχείο στην εκπαίδευση και την αξιολόγηση της απόδοσης των συνελικτικών νευρωνικών δικτύων (CNNs), καθώς παρέχουν ένα μέτρο της διαφοράς μεταξύ των προβλέψεων του δικτύου και των πραγματικών labels των δεδομένων. Η συνάρτηση κόστους, μετρά την διαφορά μεταξύ της πραγματικής εξόδου και της προβλεπόμενης εξόδου του δικτύου. Η ελαχιστοποίηση του κόστους κατά την διαδικασία εκπαίδευσης βοηθά το δίκτυο να βελτιώσει την ακρίβεια των προβλέψεών του.

Η επιλογή της σωστής συνάρτησης απώλειας επηρεάζει τον τρόπο εκμάθησης του δικτύου και την ικανότητά του να γενικεύει σε νέα δεδομένα. Επομένως, είναι σημαντικό να γίνεται με βάση την φύση του προβλήματος, τα χαρακτηριστικά του υπό ανάλυση dataset και τους στόχους που έχουν τεθεί.

Δύο από τις βασικές συναρτήσεις κόστους που χρησιμοποιούνται στα CNN στο πρόβλημα ανίχνευσης αντικειμένων είναι η Διασταυρούμενη Εντροπία (Cross-Entropy Loss) και η συνάρτηση κόστους IoU (Intersection over Union).

Διασταυρούμενη Εντροπία (Cross-Entropy Loss)

Η Διασταυρούμενη Εντροπία είναι μια ευρέως χρησιμοποιούμενη συνάρτηση κόστους για ταξινόμηση. Μετράει το πόσο διαφορετικές είναι οι πραγματικές και οι προβλεπόμενες κατανομές πιθανότητας για κάθε κλάση. Σε πολυκατηγορική ταξινόμηση, χρησιμοποιείται η μορφή της Softmax Cross-Entropy Loss, η οποία συνδυάζει την συνάρτηση Softmax και την Cross-Entropy σε ένα βήμα.

IoU (Intersection over Union)

Η συνάρτηση κόστους IoU είναι πιο ειδική για την ανίχνευση αντικειμένων και μετρά το ποσοστό της επικάλυψης μεταξύ της πραγματικής και της προβλεπόμενης περιοχής περιβάλλοντος πλαισίου

(bounding box). Μια παραλλαγή της IoU είναι η συνάρτηση κόστους GIoU (Generalized IoU), η οποία λαμβάνει υπόψη και τη γεωμετρία των πλαισίων.

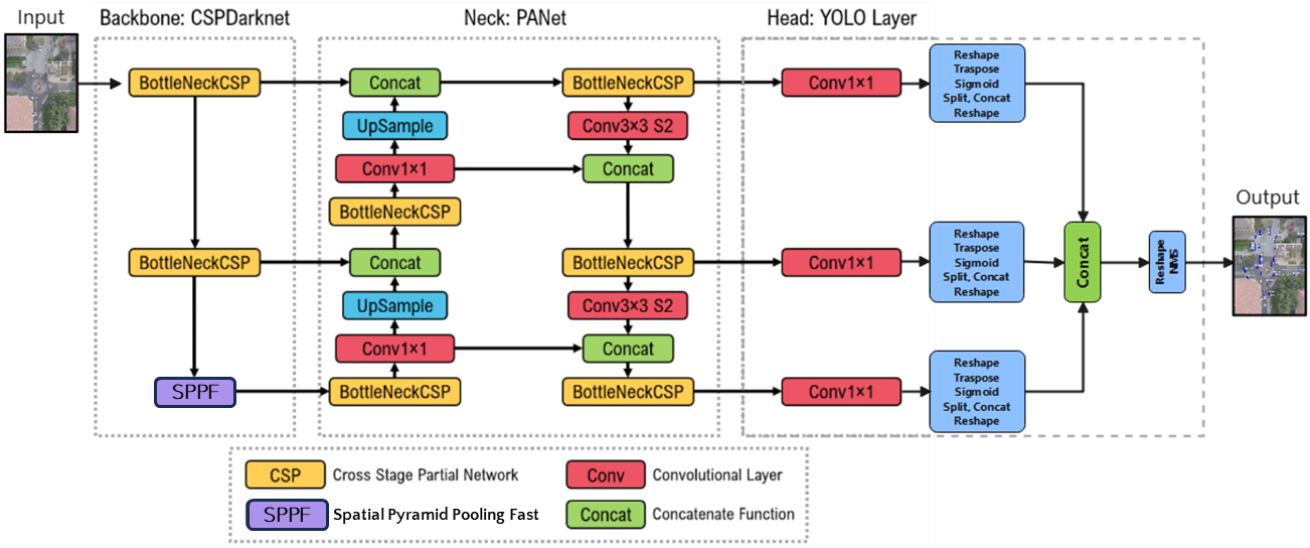
Συνδυασμός Συναρτήσεων Κόστους

Στην πράξη, για πιο περίπλοκες εργασίες ανίχνευσης, συχνά χρησιμοποιείται ένας συνδυασμός πολλαπλών συναρτήσεων κόστους. Για παράδειγμα, μπορεί να χρησιμοποιηθεί η Διασταυρούμενη Εντροπία για την ταξινόμηση του αντικειμένου και η IoU για την ακριβή τοποθέτηση του περιβάλλοντος πλαισίου.

Κεφάλαιο 3 Βασικό Μοντέλο - YOLOv5

Το YOLO (You only Look Once) είναι ένας state-of-the-art αλγόριθμος ο οποίος χρησιμοποιεί νευρωνικά δίκτυα για τα παρέχει αναγνώριση αντικειμένων σε πραγματικό χρόνο [25]. Ο συγκεκριμένος αλγόριθμος είναι γνωστός λόγω της ταχύτητας και της ακρίβειάς του. Έχει χρησιμοποιηθεί σε ποικίλες εφαρμογές για τον εντοπισμό αντικειμένων, ανθρώπων, αλλά και ζώων. Το YOLOv5 είναι μία από τις πιο πρόσφατες εκδόσεις αυτού του αλγορίθμου και αποτέλεσε το κύριο μοντέλο που χρησιμοποιήθηκε σε αυτήν την εργασία. Το συγκεκριμένο μοντέλο προσφέρει μια σειρά ανιχνευτών που έχουν προεκπαιδευτεί στο dataset MS COCO.

Η αρχιτεκτονική του YOLOv5 μοντέλου παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 3.1).



Εικόνα 3.1. Αρχιτεκτονική του μοντέλου YOLOv5.

Το μοντέλο αποτελείται από τρία βασικά μέρη:

1. *Δίκτυο κορμού (backbone) - CSPDarknet με Spatial Pyramid Pooling Fast (SPPF):*

   Χρησιμοποιείται για την εξαγωγή χαρακτηριστικών από την προ εκπαίδευση του δικτύου σε κάποιο σύνολο δεδομένων όπως το ImageNet.

2. *Neck – PANet:*

   Βρίσκεται μεταξύ του backbone και του head και χρησιμοποιείται για τον συνδυασμό πολύ-επίπεδων χαρακτηριστικών από τα διαφορετικά στάδια του backbone. Για τον σκοπό αυτό έχει υιοθετήσει τη χρήση των feature pyramid network (FPN) και pixel aggregation network (PAN) δομών.

3. *Head – YOLO layer:*

   Χρησιμοποιείται για την εξαγωγή των labels των κλάσεων και των περιεγραμμένων τετραγώνων.

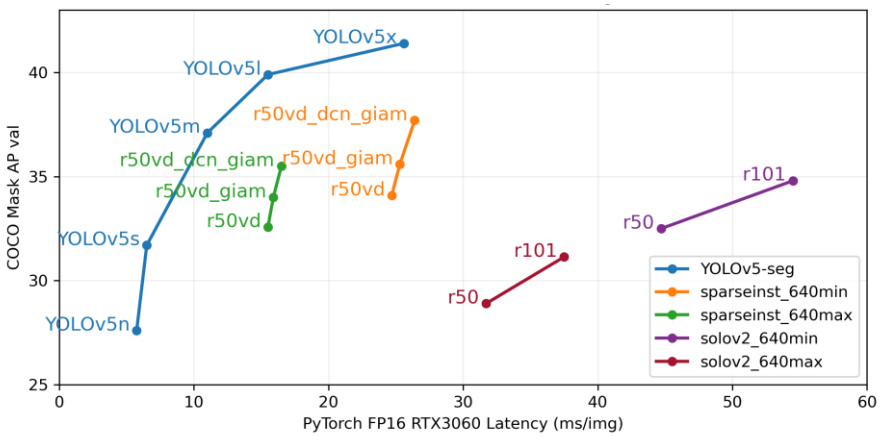
Το δίκτυο κορμού ή αλλιώς backbone είναι ένα συνελικτικό νευρωνικό δίκτυο (CNN) το οποίο εξάγει χάρτες χαρακτηριστικών (feature maps) διάφορων μεγεθών από την εικόνα εισόδου κάνοντας χρήση πολλαπλής συνέλιξης και pooling [29].

Όπως φαίνεται και από το διάγραμμα της εικόνας, υπάρχουν τέσσερα επίπεδα από feature maps που παράγονται στο δίκτυο κορμού. Τα μεγέθη τους είναι τα εξής:

- 152 × 152 pixels,
- 76 × 76 pixels,
- 38 × 38 pixels και
- 19 × 19 pixels.

Χρησιμοποιώντας αυτούς τους διαφορετικού μεγέθους feature maps, το neck συγχωνεύει (fuses) feature maps διαφορετικών επιπέδων για να αποκτήσει περισσότερες πληροφορίες και να μειώσει την απώλεια πληροφορίας. Κατά τη διάρκεια αυτής της διαδικασίας συγχώνευσης χρησιμοποιούνται οι δομές πυραμίδας χαρακτηριστικών (feature pyramid structures) του FPN και του PAN. Η δομή FPN μεταφέρει ισχυρά σημασιολογικά χαρακτηριστικά από τους κορυφαίους χάρτες χαρακτηριστικών στους χαμηλότερους χάρτες χαρακτηριστικών. Ταυτόχρονα, η δομή PAN μεταφέρει ισχυρά χαρακτηριστικά εντοπισμού από τους χαμηλότερους χάρτες χαρακτηριστικών στους υψηλότερους χάρτες χαρακτηριστικών. Οι δύο δομές ενισχύουν από κοινού την ικανότητα σύντηξης χαρακτηριστικών του δικτύου λαιμού. Ειδικότερα, μπορεί να παρατηρηθεί ότι υπάρχουν τρία επίπεδα σύντηξης χαρακτηριστικών που δημιουργούν τρεις κλίμακες νέων χαρτών χαρακτηριστικών με μεγέθη 76 × 76 × 255, 38 × 38 × 255 και 19 × 19 × 255, όπου το 255 υποδηλώνει τον αριθμό των καναλιών. Όσο μικρότερο είναι το μέγεθος των χαρτών χαρακτηριστικών, τόσο μεγαλύτερη είναι η περιοχή της εικόνας στην οποία αντιστοιχεί κάθε μονάδα πλέγματος στον χάρτη χαρακτηριστικών.

Αξίζει να σημειωθεί ότι το YOLOv5 έχει πέντε διαφορετικά μοντέλα τα οποία περιλαμβάνουν τα YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5n, and YOLOv5x ενώ έχουν αρχικοποιηθεί με βάρη που προέκυψαν από προ-εκπαίδευση με το MS COCO dataset. Από τα παραπάνω το YOLOv5x αποτελεί τον «καλύτερο» ανιχνευτή όπως φαίνεται και από το παρακάτω γράφημα (Εικόνα 3.2) [26].



Εικόνα 3.2. Διάγραμμα αποδόσεων των μοντέλων YOLOv5 στο COCO validation set [26].

Δίκτυο Κορμού (Backbone)

Το δίκτυο κορμού αποτελεί το θεμέλιο του αλγορίθμου YOLOv5 και παίζει κρίσιμο ρόλο στην εξαγωγή χαρακτηριστικών από τις εικόνες εισόδου. Συγκεκριμένα, το YOLOv5 χρησιμοποιεί το CSPDarknet53 ως backbone, ένα συνελικτικό νευρωνικό δίκτυο (CNN) που έχει σχεδιαστεί για να βελτιώνει την αποδοτικότητα και την ακρίβεια της εξαγωγής χαρακτηριστικών [27]. Το CSPDarknet53 ενσωματώνει τεχνικές όπως το Cross Stage Partial Networks (CSPNet) και το Spatial Pyramid Pooling Fast (SPPF), επιτρέποντας την αποτελεσματική εξαγωγή πολυεπίπεδων χαρακτηριστικών με υψηλή ακρίβεια [28], [29].

Η δομή του backbone στο YOLOv5 περιλαμβάνει πολλαπλά συνελικτικά επίπεδα που εφαρμόζουν συνέλιξη και pooling στις εισόδους, παράγοντας χάρτες χαρακτηριστικών (feature maps) σε διάφορα επίπεδα ανάλυσης. Αυτοί οι χάρτες χαρακτηριστικών έχουν μεγέθη όπως 152 × 152 pixels, 76 × 76 pixels, 38 × 38 pixels και 19 × 19 pixels. Αυτά τα πολλαπλά επίπεδα επιτρέπουν την ανίχνευση αντικειμένων διαφόρων μεγεθών και κλιμάκων.

Δίκτυο Ταξινόμησης/Κατάταξης (Classification Network)

Το δίκτυο ταξινόμησης στο YOLOv5, γνωστό και ως YOLO head, είναι υπεύθυνο για την τελική ανίχνευση και ταξινόμηση των αντικειμένων στις εικόνες. Αποτελείται από πολλαπλά επίπεδα συνελικτικών και πλήρως συνδεδεμένων στρώσεων που λαμβάνουν τους χάρτες χαρακτηριστικών από το backbone και το neck και παράγουν τις τελικές προβλέψεις.

Το neck του YOLOv5 χρησιμοποιεί τη δομή Path Aggregation Network (PANet) για τη συγχώνευση πολυεπίπεδων χαρακτηριστικών από το backbone, ενσωματώνοντας πληροφορίες από διάφορα επίπεδα ανάλυσης. Αυτή η διαδικασία συγχώνευσης περιλαμβάνει ανωδική και καθοδική διαδρομή, χρησιμοποιώντας μηχανισμούς προσοχής για τη βελτίωση της απόδοσης της ανίχνευσης αντικειμένων [27], [28], [30].

Το YOLO head παράγει τις τελικές προβλέψεις, συμπεριλαμβανομένων των πιθανών κλάσεων των αντικειμένων, των συντεταγμένων των περιγραμμάτων (bounding boxes) και των πιθανότητων. Χρησιμοποιεί προκαθορισμένα anchor boxes για την πρόβλεψη της κλάσης, της θέσης και του μεγέθους κάθε αντικειμένου. Κατά τη διάρκεια της ανάλυσης, το YOLOv5 χρησιμοποιεί την τεχνική Non-Maximum Suppression (NMS) για την εξάλειψη των επικαλυπτόμενων περιγραμμάτων και την παροχή των τελικών αποτελεσμάτων ανίχνευσης.

Συναρτήσεις Απωλειών

Όπως και οι προηγούμενες εκδόσεις του YOLO, η έκδοση 5 εστιάζει στην ταχύτητα και την ακρίβεια, παρέχοντας ταυτόχρονα βελτιώσεις στην απόδοση της ανίχνευσης αντικειμένων. Η συνάρτηση απώλειας (loss function) που χρησιμοποιείται στο YOLOv5 είναι ζωτικής σημασίας για την εκπαίδευση του μοντέλου, καθώς καθορίζει πώς το μοντέλο μετράει το σφάλμα μεταξύ των προβλεπόμενων και των πραγματικών ετικετών κατά τη διαδικασία της εκπαίδευσης και ουσιαστικά οδηγεί στην βελτιστοποίηση των παραμέτρων του.

Στην περίπτωση του YOLOv5, η συνάρτηση αυτή είναι σύνθετη αφού αποτελείται από ένα συνδυασμό από διαφορετικά στοιχεία, καθένα από τα οποία σχεδιάζεται για να βελτιστοποιεί διαφορετικές πτυχές της ανίχνευσης αντικειμένων [27]. Συγκεκριμένα, η συνολική απώλεια υπολογίζεται ως άθροισμα των εξής τμημάτων:

1. **Box Regression Loss**: Αυτό το μέρος της συνάρτησης απώλειας είναι υπεύθυνο για την ακριβή πρόβλεψη της θέσης και του μεγέθους των περιγραμμάτων γύρω από τα ανιχνευμένα αντικείμενα. Το YOLOv5 χρησιμοποιεί μια παραλλαγή του mean squared error (MSE), γνωστή ως CIoU (Complete Intersection over Union) loss, η οποία λαμβάνει υπόψη την επικάλυψη μεταξύ των προβλεπόμενων και των πραγματικών πλαισίων, την απόσταση μεταξύ των κέντρων τους και την αναλογία διαστάσεων, οδηγώντας σε πιο ακριβείς προβλέψεις περιγραμμάτων.
2. **Objectness Loss**: Αυτό το στοιχείο μετράει πόσο καλά το μοντέλο προβλέπει την παρουσία ενός αντικειμένου μέσα σε ένα πλαίσιο. Βασικά, είναι μια δυαδική συνάρτηση διασταυρούμενης εντροπίας (binary cross-entropy loss) που διακρίνει μεταξύ πλαισίων που περιέχουν αντικείμενα και φόντου.
3. **Classification Loss**: Για τα πλαίσια που περιέχουν αντικείμενα, το classification loss υπολογίζει πόσο ακριβώς το μοντέλο ταξινομεί αυτά τα αντικείμενα στις αντίστοιχες κατηγορίες τους χρησιμοποιώντας διασταυρούμενη εντροπία (cross-entropy loss). Ουσιαστικά, συγκρίνει τις προβλεπόμενες πιθανότητες του μοντέλου για κάθε sample με την πραγματική κλάση στην οποία ανήκουν και στη συνέχεια υπολογίζει πόσο μακριά είναι η πραγματική αξία από την προβλεπόμενη, η οποία είναι 0 ή 1.
4. **Label Smoothing**: Αυτή είναι μια τεχνική κανονικοποίησης που χρησιμοποιείται στο classification loss για να αποτρέψει το μοντέλο από το να γίνει υπερβολικά βέβαιο για τις προβλέψεις του, κάτι που μπορεί να βελτιώσει τη γενίκευση.

Η συνολική συνάρτηση απώλειας υπολογίζεται ως το άθροισμα αυτών των τριών συνιστωσών, με την κάθε μία να έχει το δικό της βάρος στον υπολογισμό.

Γιατί YOLO;

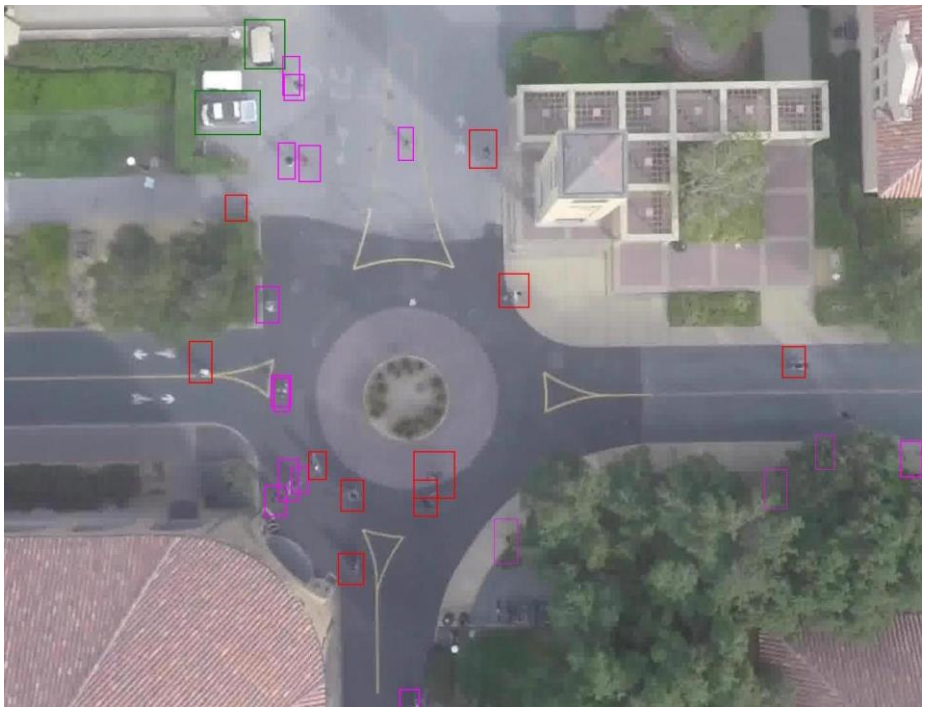
Όπως αναφέρθηκε, στην συγκεκριμένη εργασία χρησιμοποιήθηκε το μοντέλο YOLOv5. Ωστόσο οι επιλογή του κατάλληλου μοντέλου μηχανικής μάθησης για την ολοκλήρωση μιας συγκεκριμένης διαδικασίας και την επίλυση ενός προβλήματος είναι αρκετά απαιτητική. Η επιλογή αυτή εξαρτάται από πολλούς παράγοντες, όπως το πρόβλημα που χρειάζεται να επιλυθεί, τα διαθέσιμα δεδομένα και το είδος τους, την εφαρμογή του μοντέλου κ.α.. Για την περίπτωση της ανίχνευσης αντικειμένων το YOLO μοιάζει ιδανικό.

Η επιλογή του για εργασίες ανίχνευσης αντικειμένων προσφέρει έναν μοναδικό συνδυασμό ταχύτητας και ακρίβειας, συγκριτικά με άλλα μοντέλα, γεγονός ιδιαίτερα σημαντικό για εφαρμογές πραγματικού χρόνου [26], [31], [32]. Ο σχεδιασμός του επιτρέπει την γρήγορη επεξεργασία εικόνων χωρίς να θυσιάζεται σημαντικά η ακρίβεια ανίχνευσης, καθιστώντας το ιδανικό για σενάρια όπου απαιτείται ταυτόχρονα γρήγορη και αξιόπιστη ταυτοποίηση αντικειμένων, όπως σε αυτόνομα οχήματα, συστήματα επιτήρησης και παρακολούθηση σε πραγματικό χρόνο. Επιπρόσθετα, η ευελιξία που προσφέρει το PyTorch, διευκολύνει την εκπαίδευση σε προσαρμοσμένα σύνολα δεδομένων, πράγμα πολύ σημαντικό για προβλήματα ανίχνευσης αντικειμένων. Ο συνδυασμός όλων των παραπάνω χαρακτηριστικών—ταχύτητα, ακρίβεια, ευκολία χρήσης, και ευελιξία— αλλά και η ισχυρή κοινότητα που το υποστηρίζει καθιστά το YOLO μια αποτελεσματική και πολύ αποδοτική λύση σε προβλήματα ανίχνευσης αντικειμένων.

Κεφάλαιο 4 Σύνολο Δεδομένων

Περιγραφή του Stanford Drone Dataset

Το Stanford Drone Dataset (SDD) [33] αποτελεί ένα τεράστιο σύνολο δεδομένων από αεροφωτογραφίες που έχουν ληφθεί με την βοήθεια drones στην περιοχή του Stanford University Campus. Το συγκεκριμένο dataset είναι ιδανικό για περιπτώσεις μηχανικές όρασης (computer vision), όπως η αναγνώριση αντικειμένων ή η παρακολούθηση στόχων (target tracking). Αποτελείται από περισσότερα από 60 βίντεο (aerial videos) τα οποία αντιστοιχούν σε περίπου 70GB δεδομένων. Σε κάθε βίντεο, ένα μοντέλο μπορεί να αναγνωρίσει έξι διαφορετικούς πράκτορες (agents) – πεζούς (pedestrians), ποδηλάτες (bikers), skaters, carts, αυτοκίνητα (cars) και λεωφορεία (bus). Δυστυχώς, το συγκεκριμένο dataset είναι «προκατειλημμένο» (biased), αφού οι κλάσεις των πεζών και των ποδηλατών αποτελούν περισσότερο από 80% του συνόλου των annotations.



Εικόνα 4.1. Παράδειγμα εικόνας από το SDD [33].

Επεξεργασία Δεδομένων

Για την χρήση των δεδομένων (βίντεο) για τον σκοπό της εργασίας ήταν απαραίτητη η μετατροπή τους σε εικόνες. Η εξαγωγή των frames από τα βίντεο συνδυάστηκε με τον έλεγχο των αντίστοιχων annotations. Ωστόσο κατά την παραπάνω διαδικασία απαιτήθηκε χώρος στον δίσκο μεγαλύτερος του διαθέσιμου με αποτέλεσμα την μη ολοκλήρωση της. Για την επίλυση του παραπάνω προβλήματος κρίθηκε απαραίτητη η μείωση του μέγεθος του χώρου που καταλαμβάνουν οι εξαγόμενες εικόνες. Για τον σκοπό αυτό αποφασίστηκε η εξαγωγή μίας εικόνας κάθε 30 frames.

Συνεπώς το τελικό dataset αποτελείται συνολικά από 16.189 εικόνες μεγέθους 16,1GB και βίντεο των 30fps (3 εικόνες ανά δευτερόλεπτο).

Αλλαγές στο Σύνολο των Δεδομένων

Για την πραγματοποίηση των πειραμάτων που πραγματοποιήθηκαν κατά την διάρκεια αυτής της εργασίας πραγματοποιήθηκαν ορισμένες αλλαγές στο σύνολο των δεδομένων. Συγκεκριμένα, πραγματοποιήθηκε αντικατάσταση του ονόματος κάθε κλάσης από έναν ακέραιο δείκτη, όπως φαίνεται στον παρακάτω πίνακα:

Πίνακας 4.1. Κλάσεις και αντίστοιχοι δείκτες των αντικειμένων στο dataset.

Pedestrian	Skater	Biker	Car	Bus	Cart
0	1	2	3	4	5

Επιπλέον της παραπάνω αλλαγής πραγματοποιήθηκε διαγραφή όλων των αντικειμένων που υπάρχουν στο annotation στις περιπτώσεις που επικαλύπτονται ή έχουν φύγει από την εικόνα (**lost, occurred = 1**).

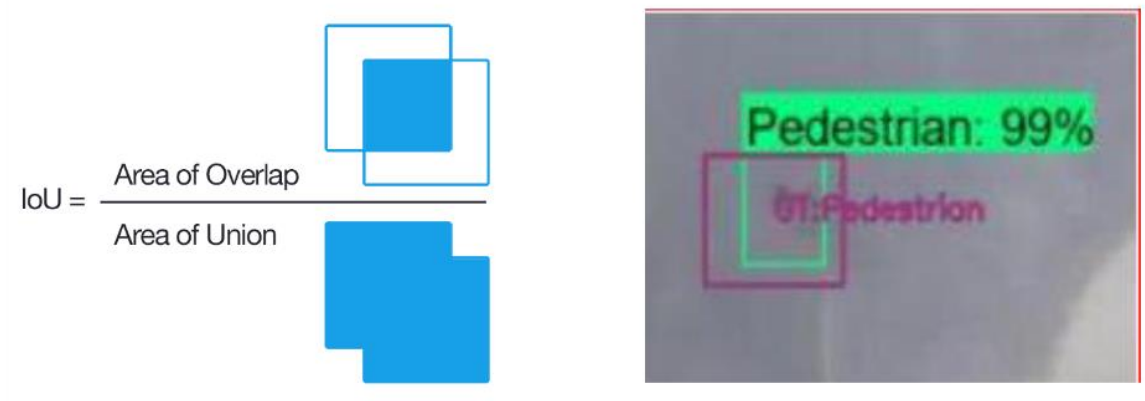
Η επιλογή του διαχωρισμού (splitting) του συνόλου δεδομένων παίζει καθοριστικό ρόλο στην ανάπτυξη και αξιολόγηση μοντέλων μηχανικής μάθησης και βαθιάς μάθησης. Για την επίτευξη μιας ισσοροποιμένης κατανομή η οποία θα βηθήσει στην εκπαίδευση των μοντέλων αλλα και στην αξιολόγηση τους το dataset χωρίστικε σε τρία τμήματα. Το 70% του συνόλου των δεδομένων δεσμεύτηκε και χρησιμοποιήθηκε για την εκπαίδευση και τη δημιουργία των μοντέλων (training set). Το σύνολο αυτό περιλαμβάνει την πλειονότητα των δεδομένων διευκολύνει την απόκτηση πλούσιων αναπαραστάσεων χαρακτηριστικών και τη γενίκευση του μοντέλου. Το 20% για την επικύρωση (validation set) τους και την βελτίωση του μοντέλου σε κάθε εποχή. Το σύνολο επικύρωσης χρησιμεύει ως κρίσιμος ενδιάμεσος παράγοντας, προσφέροντας γνώσεις σχετικά με τη σύγκλιση του μοντέλου και τη ρύθμιση των υπερπαραμέτρων χωρίς τον κίνδυνο υπερβολικής προσαρμογής στα δεδομένα δοκιμής. Τέλος το 10% των δεδομένων χρησιμοποιήθηκε για την αξιολόγηση των μοντέλων μετά την ολοκλήρωση της εκπαίδευσης (testing set). Το σύνολο αυτό, αντιπροσωπευτικό των αθέατων περιπτώσεων, επιτρέπει την αξιολόγηση της απόδοσης του μοντέλου σε πραγματικές συνθήκες.

Κεφάλαιο 5 Μετρικές Αξιολόγησης

Διατομή επί της Ένωσης (IoU)

Η διατομή επί της ένωσης ή Intersection over Union (IoU) είναι μια μετρική στην αναγνώριση αντικειμένων που περιγράφει πως αλληλεπικαλύπτονται το πραγματικό πλαίσιο σε σχέση με το πλαίσιο πρόβλεψης. Σε προβλήματα ταξινόμησης εικόνας στα οποία το αποτέλεσμα είναι δυαδικό (true ή false) η χρήση αυτού του δείκτη είναι πολύ συχνή. Παρόλα αυτά, χρησιμοποιείται και σε προβλήματα αναγνώρισης αντικειμένων σε συνδυασμό με τον καθορισμό νέων μετρικών όπως το Recall και το Precision. Ουσιαστικά χρησιμοποιείται ως ένα όριο για τον προσδιορισμό του εάν ένα προβλεπόμενο αποτέλεσμα είναι αληθώς θετικό ή ψευδώς θετικό.

Το YOLO χρησιμοποιεί το IoU για να παρέχει ένα πλαίσιο το οποίο καλύπτει περιμετρικά τέλεια τα αντικείμενα. Κάθε πλεγματικό κελί είναι υπεύθυνο για να προβλέπει τα περιμετρικά πλαίσια και το ποσοστό σιγουριάς τους. Ειδικότερα, το IoU είναι ίσο με 1 εάν το περιμετρικό πλαίσιο που υπολογίστηκε είναι το ίδιο με το πραγματικό. Με τη χρήση αυτού του μηχανισμού απορρίπτονται περιμετρικά πλαίσια που δεν ταυτίζονται με τα πραγματικά. Ένα παράδειγμα το οποίο δείχνει πως δουλεύει το IoU παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 5.1).



Εικόνα 5.1. Δείκτης Intersection over Union (αριστερά) και παράδειγμα χρήσης του στην αναγνώρηση αντικειμένων (δεξιά) [34].

Μετρικές

Στην ανίχνευση αντικειμένων, ο κύριος σκοπός είναι η ανίχνευση όλων των αντικειμένων στην εικόνα που παρουσιάζονται. Αυτό επιτυγχάνεται με την τοποθέτηση οριοθετημένων πλαισίων (bounding boxes) γύρω από τα αντικείμενα. Τα πλαίσια αυτά χαρακτηρίζονται:

1. από τη θέση του κέντρου και τις διαστάσεις του οριοθετημένου πλαισίου,
2. την κλάση του αντικειμένου και

3. την πιθανότητα (με τιμές από 0 έως 1) που δείχνει πόσο σίγουρος είναι ο αλγόριθμος για την πρόβλεψη.

Υπάρχουν 2 διαφορετικά σύνολα οριοθετημένων πλαισίων. Ένα σύνολο από ground truth bounding boxes, τα οποία είναι αυτά που δίνονται στο μοντέλο στο σύνολο δεδομένων εκπαίδευσης, και ένα σύνολο από προβλεπόμενα bounding boxes, τα οποία είναι αυτά που το αντικείμενο εξάγει ως τις προβλέψεις του. Είναι προφανές ότι κάθε bounding box που προβλέπει το μοντέλο δεν είναι πάντα σωστό. Συνεπώς είναι σημαντικό να γνωρίζουμε την απόδοση του κάθε μοντέλου.

Σε αυτό το σημείο θα αναφέρουμε μερικές από τις σημαντικότερες μετρικές που χρησιμοποιούνται στην μηχανική όραση για την αξιολόγηση της απόδοσης των μοντέλων. Για τον ορισμό αυτών των μετρικών είναι απαραίτητη η χρήση των ορισμών που παρουσιάστηκαν στις προηγούμενες παραγράφους.

Χρησιμοποιώντας τον IoU δείκτη μπορούν να οριστούν οι παρακάτω ορισμοί:

- **Αληθώς Θετικά - True Positive (TP):** Το μοντέλο έχει προβλέψει κάτι ως Α και στην πραγματικότητα είναι το Α. Στην περίπτωση αυτή ισχύει ότι IoU ≥ threshold. Το κατώφλι (threshold) είναι μια σταθερά η οποία συνήθως λαμβάνει τιμές πάνω από 0,5 ή 50%.
- **Ψευδώς Αρνητικά - False Negatives (FN):** Το μοντέλο ΔΕΝ έχει προβλέψει κάτι ως Α και στην πραγματικότητα είναι ορθώς το Α. Ισχύει ότι IoU < threshold.
- **Ψευδώς Θετικά - False Positives (FP)**: Το μοντέλο έχει προβλέψει κάτι ως Α και στην πραγματικότητα είναι το Β.
- **Αληθώς Αρνητικά - True Negatives (TN):** Το μοντέλο ΔΕΝ έχει προβλέψει κάτι ως Α και στην πραγματικότητα είναι το Β.

Πρέπει μα σημειωθεί ότι στην ανίχνευση αντικειμένων η μετρική TN δεν έχει καμία χρησιμότητα, αφού TN θα ήταν όλα τα πιθανά bounding boxes που δεν ανιχνεύτηκαν σωστά σε μία εικόνα (σε μία εικόνα υπάρχουν πολλά bounding boxes που δεν πρέπει να προβλεφθούν).

Αφού ολοκληρωθεί η ταξινόμηση, μπορούν να υπολογιστούν μετρικές όπως η ακρίβεια (Precision) του μοντέλου, η ανάκληση (Recall) και η βαθμολογία F1 (F1 score) [35].

Ακρίβεια (Precision)

Η ακρίβεια είναι η ικανότητα του μοντέλου να προβλέπει μόνο τα σχετικά αντικείμενα και είναι το ποσοστό των σωστών προβλέψεων. Είναι δηλαδή ο λόγος του αριθμού των αληθών θετικών προς τον συνολικό αριθμό των θετικών προβλέψεων. Για παράδειγμα, εάν το μοντέλο ανίχνευσε 100 αυτοκίνητα και οι 90 ήταν σωστές τότε η ακρίβεια είναι θα είναι 90%.

$$Precision = \frac{(True\ Positive)}{(True\ Positive\ +\ False\ Positive)}$$

Ανάκληση (Recall)

Η ανάκληση (Recall ή R) είναι η ικανότητα του μοντέλου να προβλέπει όλα τα πλαίσια οριοθέτησης της βασικής αλήθειας (ground truth) και ορίζεται ως το ποσοστό των σωστών προβλέψεων μεταξύ των πλαισίων οριοθέτησης της βασικής αλήθειας (). Είναι δηλαδή ο δείκτης των σωστών θετικών προβλέψεων επί τον συνολικό αριθμό των πραγματικών (σχετικών) αντικειμένων. Για παράδειγμα, εάν το μοντέλο ανιχνεύσει σωστά 75 αυτοκίνητα σε μια εικόνα και στην πραγματικότητα υπάρχουν 100 αυτοκίνητα στην εικόνα, η ανάκληση είναι 75%.

$$Recall = \frac{(True\ Positive)}{(True\ Positive\ +\ False\ Negative)}$$

Βαθμολογία F1 (F1 score)

Η βαθμολογία F1 είναι ένας σταθμισμένος μέσος των μετρήσεων ακρίβειας και ανάκλησης, προσφέροντας μία πιο ολοκληρωμένη μέτρηση της αποδοτικότητας. Το μέγεθος του F1 κυμαίνεται από 0 έως 1, με το 1 να συμβολίζει την τέλεια ακρίβεια. Ο τύπος της βαθμολογίας F1 είναι:

$$F1\ score = (Precision \times Recall)\ /\ [(Precision + Recall)/2]$$

Αξίζει να σημειωθεί ότι η συνολική ακρίβεια ενός μοντέλου ανίχνευσης αντικειμένων στηρίζεται στην ποιότητα και τον όγκο των δειγμάτων εκπαίδευσης, την ποιότητα των εικόνων εισόδου, τις παραμέτρους του μοντέλου καθώς και στο καθορισμένο όριο ακρίβειας που έχει οριστεί.

Κεφάλαιο 6 Παραλλαγές του YOLOv5

Το YOLOv5 αποτελεί μία από τις πιο διαδεδομένες και ευρέως χρησιμοποιούμενες εκδόσεις της οικογένειας μοντέλων "You Only Look Once" (YOLO), τα οποία έχουν καθιερωθεί για την ταχύτητα και την ακρίβειά τους στην ανίχνευση αντικειμένων σε πραγματικό χρόνο. Ως μια ευέλικτη και αξιόπιστη αρχιτεκτονική, το YOLOv5 έχει αξιοποιηθεί σε πληθώρα εφαρμογών μηχανικής όρασης, από την αυτόνομη οδήγηση έως την επιτήρηση.

Στην παρούσα εργασία, το μοντέλο YOLOv5x επιλέχθηκε ως βασικό σημείο αναφοράς, πάνω στο οποίο πραγματοποιήθηκαν και αξιολογήθηκαν περαιτέρω βελτιώσεις. Πιο συγκεκριμένα, πέρα από την βασική έκδοση του εξετάστηκαν δύο τροποποιήσεις της αρχικής αρχιτεκτονικής με στόχο την ενίσχυση της ακρίβειας εντοπισμού αντικειμένων (πεζοί, λεωφορεία, αυτοκίνητα) σε κατακόρυφες εναέριες εικόνες.

Η πρώτη τροποποίηση αφορά την αντικατάσταση του max-pooling με Softpool στο SPPF block, με σκοπό τη διατήρηση περισσότερης πληροφορίας σε συνδιασμό με τη προσθήκη του μηχανισμού Squeeze-and-Excitation (SE) στο backbone του μοντέλου, ώστε να ενισχυθούν επιλεκτικά τα πιο σημαντικά χαρακτηριστικά. Η δεύτερη τροποποίηση αφορά την προσθήκη του μηχανισμού Coord Attention (CA) στο backbone, προσδίδοντας στο μοντέλο τη δυνατότητα να λαμβάνει υπόψη όχι μόνο την παρουσία αλλά και τη χωρική θέση των χαρακτηριστικών εντός της εικόνας. Στις επόμενες ενότητες περιγράφονται αναλυτικά οι αλλαγές αυτές και ο τρόπος ενσωμάτωσής τους στο δίκτυο YOLOv5.

YOLOv5x με Softpool και Squeeze and Excitation Module

Η SoftPool είναι μια μέθοδος υποδειγματοληψίας που έχει αναπτυχθεί για να βελτιώσει τις επιδόσεις των νευρωνικών δικτύων, ειδικά σε εφαρμογές αναγνώρισης αντικειμένων όπως ο YOLOv5 [37] [38]. Η τυπική διαδικασία pooling, όπως το max-pooling και το average pooling, μπορεί να οδηγήσει σε απώλεια σημαντικών χαρακτηριστικών της εικόνας. Το SoftPool στοχεύει στη μείωση αυτής της απώλειας μέσω μιας πιο "μαλακής" διαδικασίας που διατηρεί περισσότερες πληροφορίες από την αρχική εικόνα. Σε αντίθεση με το max pooling που διατηρεί μόνο τη μέγιστη τιμή ενός παράθυρου και το average pooling που υπολογίζει τον μέσο όρο όλων των τιμών, το SoftPool χρησιμοποιεί μια συνάρτηση που ζυγίζει όλες τις τιμές σε ένα παράθυρο pooling και παράγει μια "μαλακή" τιμή που αντιπροσωπεύει καλύτερα τα χαρακτηριστικά του αρχικού παραθύρου.

Η βασική ιδέα πίσω από το SoftPool είναι η χρήση μιας καμπύλης πιθανότητας για τον υπολογισμό του βάρους κάθε τιμής σε ένα παράθυρο pooling. Αυτή η καμπύλη εξασφαλίζει ότι όλες οι τιμές συμβάλλουν στην τελική έξοδο, αλλά με διαφορετικούς βαθμούς βαρύτητας.

Η εξίσωση που περιγράφει το SoftPool είναι η εξής [37]:

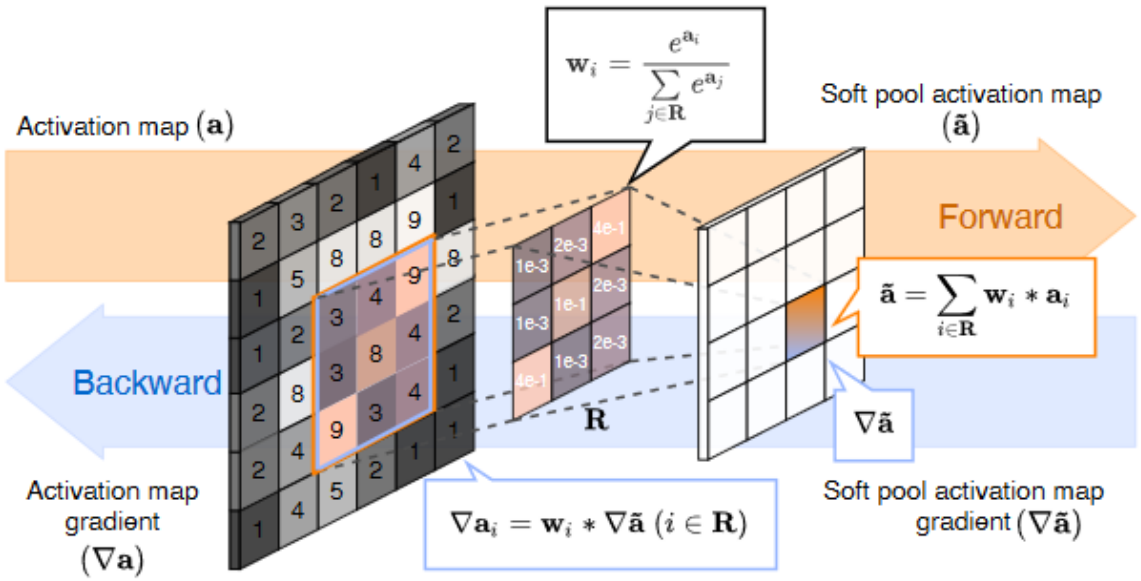
$$W_i = \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}}$$

$$\tilde{a} = \sum_{i \in R} W_i \, a_i$$

Όπου:

- W_i είναι το βάρος του i-οστού στοιχείου στο παράθυρο pooling,
- a_i είναι η ενεργή τιμή του iii-οστού στοιχείου και
- $\tilde{a}$ είναι η έξοδος του SoftPool για το παράθυρο pooling.

Το βάρος κάθε τιμής στο παράθυρο υπολογίζεται χρησιμοποιώντας τη συνάρτηση softmax, εξασφαλίζοντας ότι όλες οι τιμές συμβάλλουν στην τελική έξοδο ανάλογα με τη σημαντικότητά τους.



Εικόνα 6.1. Υπολογισμός SoftPool. Κατά το forward operation – πορτοκαλί χρώμα – χρησιμοποιείται η εκθετική softmax τιμή του κάθε activation ως βάρος και υπολογίζεται το σταθμισμένο άθροισμα για την περιοχή R. Αυτά τα βάρη χρησιμοποιούνται επίσης για τα gradients (κλίσεις) – με μπλέ χρώμα. Τα activation gradients είναι ανάλογα με τα υπολογισμένα softmax βάρη [37].

Στην αναγνώριση αντικειμένων όπως στην περίπτωση αυτή που μελετάμε, το SoftPool μπορεί να βελτιώσει την απόδοση του YOLOv5 με τους εξής τρόπους:

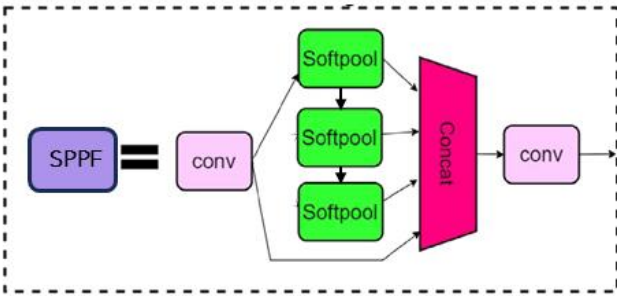
1. **Διατήρηση Χαρακτηριστικών**: Με το να μην απορρίπτονται τα χαρακτηριστικά κατά τη διάρκεια της διαδικασίας pooling, το Softpool μπορεί να διατηρεί περισσότερες πληροφορίες από την αρχική εικόνα. Αυτό είναι ιδιαίτερα σημαντικό για την ανίχνευση μικρών ή λεπτομερών αντικειμένων.

2. **Βελτίωση της Γενίκευσης**: Η χρήση μιας “softer” προσέγγισης για το pooling μπορεί να βοηθήσει το μοντέλο να γενικεύσει καλύτερα σε νέα δεδομένα. Η διατήρηση περισσότερων χαρακτηριστικών μπορεί να οδηγήσει σε καλύτερη αναγνώριση σε ποικίλες συνθήκες φωτισμού και φόντου.
3. **Αποδοτικότητα**: Παρά τη διατήρηση περισσότερων πληροφοριών, το η μέθοδος αυτή δεν απαιτεί σημαντικά περισσότερους υπολογιστικούς πόρους από τις παραδοσιακές μεθόδους pooling, καθιστώντας το μια πρακτική λύση για πραγματικές εφαρμογές.

Το YOLOv5 χρησιμοποιεί το SPPF (Spatial Pyramid Pooling – Fast) για να αυξήσει το οπτικό πεδίο του μοντέλου και να ενισχύσει την πληροφορία που λαμβάνει από διαφορετικές περιοχές της εικόνας. Αυτό επιτυγχάνεται με διαδοχικές εφαρμογές της λειτουργίας pooling στο ίδιο feature map, επιτρέποντας στο μοντέλο να «βλέπει» μεγαλύτερα χωρικά πλαίσια, χωρίς να αυξάνει το computational cost. Το SPPF συνήθως εφαρμόζει πολλαπλά pooling με ίδιο kernel size (π.χ. 5x5) και τα αποτελέσματα αυτών των pooling συνδυάζονται μέσω concatenation με το αρχικό feature map.

Στην αρχική του υλοποίηση, το SPPF χρησιμοποιεί max-pooling, δηλαδή κρατά την μέγιστη τιμή από κάθε περιοχή (π.χ. 5x5 pixels). Ωστόσο, στην παρούσα εργασία προτείνεται η αντικατάσταση του max-pooling με Softpool. Το Softpool δεν κρατά απλώς τη μέγιστη τιμή, αλλά υπολογίζει έναν σταθμισμένο μέσο όρο, δίνοντας μεγαλύτερη έμφαση στις υψηλές τιμές του παραθύρου pooling. Η σταθμισμένη αυτή προσέγγιση διατηρεί περισσότερη πληροφορία από τοπικές περιοχές και επιτρέπει πιο ήπια και σταθερή ροή πληροφορίας στο δίκτυο, ενισχύοντας τη δυνατότητα γενίκευσης του μοντέλου σε δύσκολες συνθήκες εικόνας.

Η θέση του Softpool στο SPPF του YOLOv5 φαίνεται στην παρακάτω εικόνα (Εικόνα 6.2):



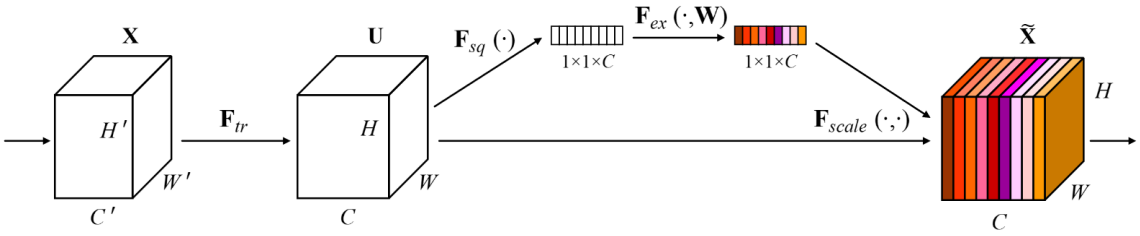
Εικόνα 6.2. Η μέθοδος Softpool στο SPPF module του μοντέλο YOLOv5 [38].

Επιπλέον η αρχιτεκτονική του μοντέλου ενισχύθηκε κάνοντας χρήση στοιχείων όπως το Squeeze-and-Excitation (SE). Στην πραγματικότητα το SE module προσθέτει έναν μηχανισμό «προσοχής» (attention mechanism) που επιτρέπει στο μοντέλο να ενισχύει τα σημαντικότερα χαρακτηριστικά και να καταστέλλει τα λιγότερο σημαντικά, βελτιώνοντας έτσι την ικανότητα εξαγωγής χαρακτηριστικών. Η προσθήκη του SE module στοχεύει στη βελτίωση της απόδοσης μέσω καλύτερης ανάλυσης και προσοχής στις σημαντικές περιοχές των εικόνων.

Η SE προτάθηκε από τους Hu et al. το 2018 [39], και εισάγει μια καινοτόμα δομική μονάδα γνωστή ως Squeeze-and-Excitation block. Αυτό το module αναπροσαρμόζει τις τιμές ενός τρισδιάστατου διανύσματος χαρακτηριστικών ανά επίπεδο βάθους ή ανά κανάλι, μέσω της αναλυτικής

μοντελοποίησης των σχέσεων μεταξύ των καναλιών. Τα κύρια στοιχεία της περιλαμβάνουν ένα επίπεδο global average pooling ανά κανάλι, που ακολουθείται από δύο διαδοχικά fully connected layers με συναρτήσεις ενεργοποίησης ReLU και Sigmoid. Το πρώτο layer επιτελεί τη λειτουργία της συμπίεσης (squeeze), εξάγοντας στατιστική πληροφορία καθολικού χαρακτήρα ανά κανάλι, ενώ τα επόμενα δύο layers επιτελούν τη λειτουργία της διέγερσης (excitation), αναδεικνύοντας τις εσωτερικές σχέσεις και εξαρτήσεις μεταξύ των καναλιών. Η έξοδος του τελευταίου πλήρως συνδεδεμένου επιπέδου πολλαπλασιάζεται με το διάνυσμα εισόδου της δομικής μονάδας, προσαρμόζοντας το με βάση την εξαγόμενη πληροφορία.

Η δομή αυτού του module φαίνεται την παρακάτω εικόνα (Εικόνα 6.3):



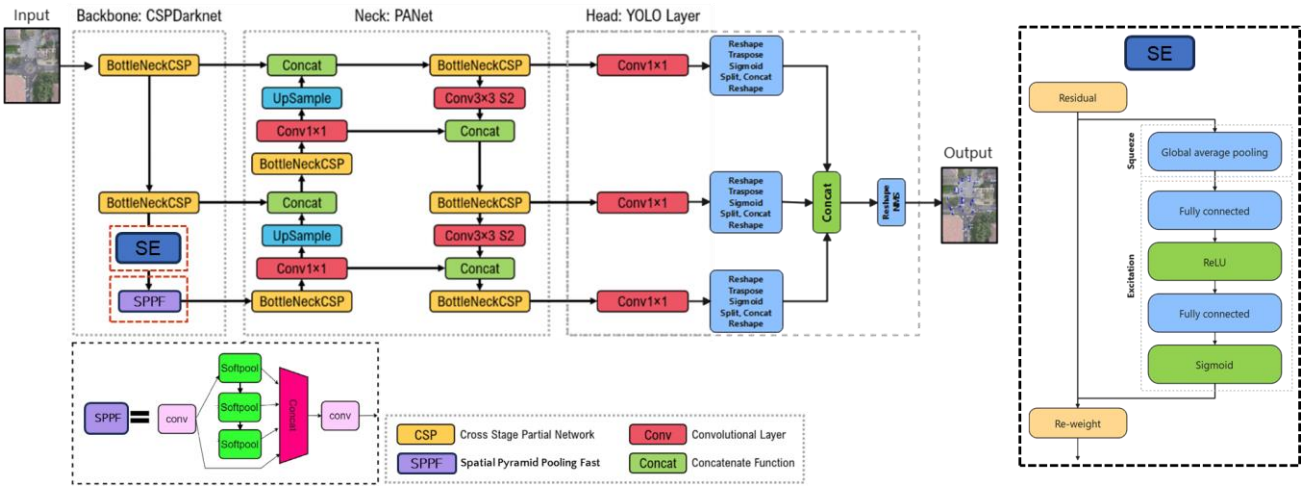
Εικόνα 6.3. Η δομή του Squeeze-and-Excitation Module. F_{sq}, F_{ex}, F_{scale} συμβολίζουν τις λειτουργίες συμπίεσης, διέγερσης και πολλαπλασιασμού αντίστοιχα και με H, W, C, οι διαστάσεις μήκους, πλάτους και βάθους αντίστοιχα [39].

Η λειτουργία του SE block περιλαμβάνει τρία στάδια: Squeeze, Excitation και Re-weighting. Στο στάδιο Squeeze, εφαρμόζεται Global Average Pooling σε κάθε κανάλι (feature map) του εισερχόμενου σήματος. Αυτό «συμπιέζει» τη χωρική πληροφορία κάθε χαρακτηριστικού σε μία μόνο τιμή, η οποία εκφράζει τη συνολική του ενεργοποίηση στο πλαίσιο της εικόνας.

Το παραγόμενο διάνυσμα περνά στο στάδιο Excitation, όπου υποβάλλεται σε δύο πλήρως συνδεδεμένα (fully connected) layers. Το πρώτο layer μειώνει τη διάσταση του διανύσματος (bottleneck) για να περιορίσει το υπολογιστικό κόστος, και εφαρμόζεται η συνάρτηση ReLU για μη-γραμμικότητα και ενίσχυση θετικών συσχετίσεων. Στο δεύτερο layer, η διάσταση επανέρχεται στο αρχικό μέγεθος και χρησιμοποιείται η συνάρτηση ενεργοποίησης Sigmoid, ώστε να παραχθεί ένα σύνολο συντελεστών "προσοχής" με τιμές μεταξύ 0 και 1.

Στο τελικό στάδιο, Re-weighting, οι συντελεστές αυτοί χρησιμοποιούνται για να πολλαπλασιάσουν τα αρχικά χαρακτηριστικά (features) ανά κανάλι. Έτσι, τα σημαντικά χαρακτηριστικά ενισχύονται, ενώ τα λιγότερο κρίσιμα καταστέλλονται. Το τελικό output έχει το ίδιο σχήμα με την είσοδο, αλλά με προσαρμοσμένη και πιο «φιλτραρισμένη» πληροφορία, βελτιώνοντας την ικανότητα του δικτύου να επικεντρώνεται σε πληροφορία ουσιαστικής σημασίας για την αναγνώριση αντικειμένων.

Στην περίπτωση του YOLOv5, το Squeeze-and-Excitation Module (SE) ενσωματώθηκε πριν από το SPPF block στο backbone του μοντέλου, όπως απεικονίζεται στην Εικόνα 6.4.



Εικόνα 6.4. Αρχιτεκτονική του μοντέλου YOLOv5 με χρήση Softpool αντί για maxpool και με προσθήκη Squeeze-and-Excitation Module στο backbone layer.

YOLOv5x με Softpool και CoordAttention

πραγματοποιήθηκε ενσωμάτωση του μηχανισμού Coord Attention (CA) στο backbone του YOLOv5, κατά παρόμοιο τρόπο με την προηγούμενη περίπτωση του SE Module. Το block Softpool διατηρήθηκε στο SPPF, με σκοπό να αξιοποιηθεί η συνδυαστική ενίσχυση τόσο των σημαντικών χαρακτηριστικών όσο και της χωρικής πληροφορίας.

Ο μηχανισμός Coord Attention διαφοροποιείται από άλλες τεχνικές προσοχής, καθώς δεν επικεντρώνεται μόνο στο ποια χαρακτηριστικά υπάρχουν στην εικόνα, αλλά και πού ακριβώς βρίσκονται. Αυτό επιτυγχάνεται με τη διατήρηση της χωρικής πληροφορίας κατά μήκος των δύο κύριων αξόνων της εικόνας.

Συγκεκριμένα, εφαρμόζεται μέσος όρος (average pooling) κατά μήκος δύο κατευθύνσεων:

- Οριζόντια (κατά τον άξονα X), και
- Κάθετα (κατά τον άξονα Y).

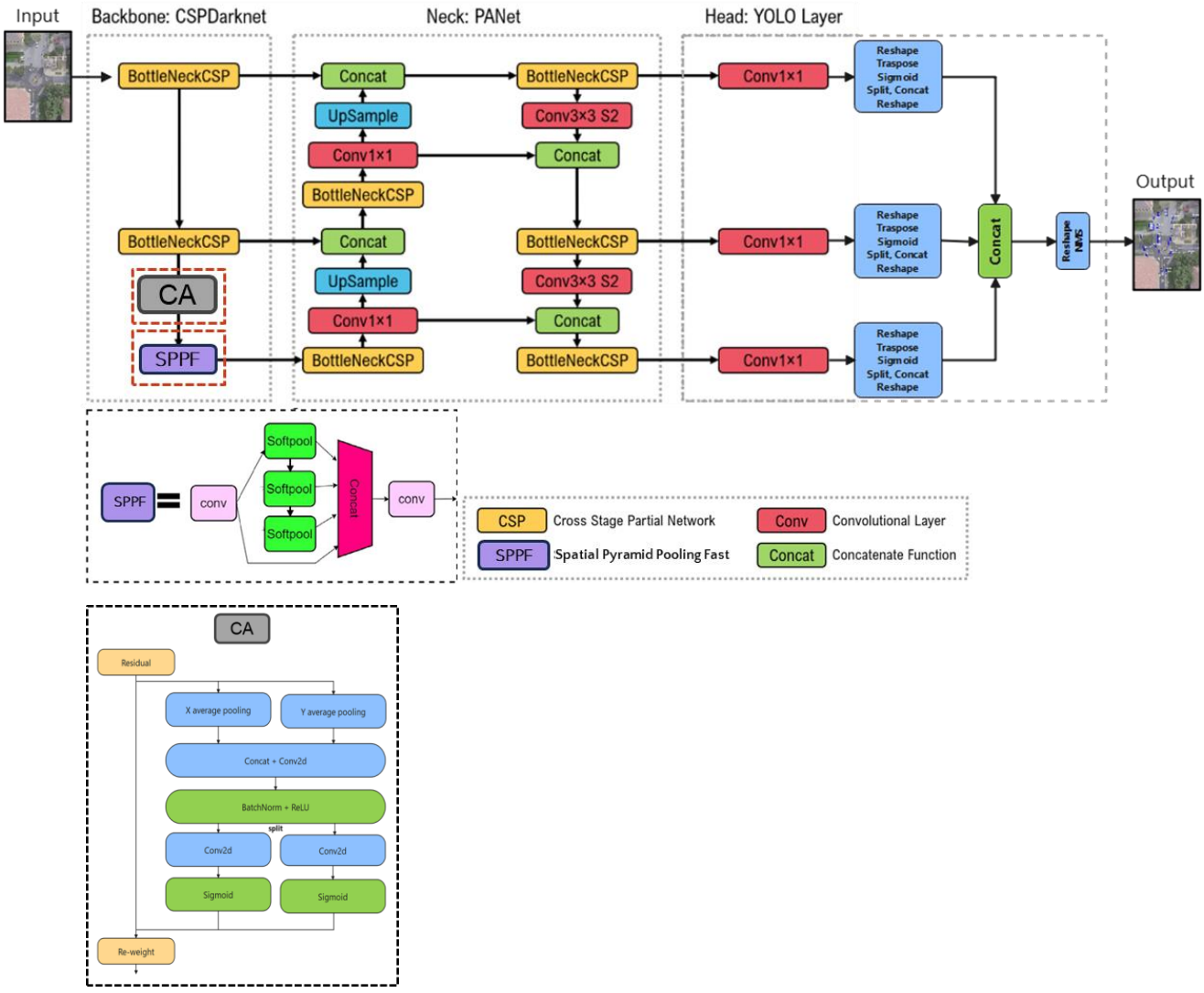
Αυτό το διπλό pooling επιτρέπει στο δίκτυο να αποκτήσει πληροφορία για τη θέση των αντικειμένων σε κάθε άξονα, διατηρώντας τις σχετικές εξαρτήσεις. Οι παραγόμενες δύο ροές (X και Y) συνδυάζονται μέσω concatenation και στη συνέχεια περνούν από ένα συνελικτικό επίπεδο (Conv2D), το οποίο συνθέτει τη χωρική πληροφορία των δύο κατευθύνσεων. Ακολουθεί Batch Normalization, το οποίο σταθεροποιεί την εκπαίδευση, και ReLU ενεργοποίηση, η οποία εισάγει μη γραμμικότητα και ενισχύει θετικές συσχετίσεις.

Η ροή στη συνέχεια διαχωρίζεται σε δύο ανεξάρτητες διαδρομές:

- Μία που διαχειρίζεται την πληροφορία του άξονα X, και
- Μία που διαχειρίζεται την πληροφορία του άξονα Y.

Κάθε διαδρομή περνά από ένα ακόμη Conv2D layer, και στη συνέχεια εφαρμόζεται Sigmoid συνάρτηση ενεργοποίησης ώστε να παραχθούν μάσκες προσοχής (attention masks) για κάθε κατεύθυνση. Οι μάσκες αυτές εκφράζουν τη σχετική σημαντικότητα περιοχών της εικόνας ανάλογα με την κατεύθυνση, και χρησιμοποιούνται για ανακαθορισμό των αρχικών χαρακτηριστικών (re-weighting). Έτσι παράγεται ένα ενισχυμένο feature map, όπου κάθε περιοχή της εικόνας έχει τροποποιηθεί κατάλληλα βάσει της χωρικής της σημασίας. Το Coord Attention επιτρέπει στο YOLOv5 να κατανοεί καλύτερα τη χωρική διάταξη των αντικειμένων εντός της εικόνας, προσδίδοντας στο δίκτυο ικανότητα πιο ακριβούς εντοπισμού, ιδιαίτερα σε εικόνες όπου τα αντικείμενα εμφανίζονται με διαφορετικούς προσανατολισμούς ή αποστάσεις.

Η ενσωμάτωση του μηχανισμού Coord Attention (CA) στο backbone του μοντέλου, σε συνδυασμό με το Softpool στο SPPF block, απεικονίζεται στην Εικόνα 6.5.



Εικόνα 6.5. Αρχιτεκτονική του μοντέλου YOLOv5 με χρήση Softpool αντί για maxpool και με προσθήκη Coord Attention (CA) block στο backbone layer.

Κεφάλαιο 7 Πειραματική Μελέτη – Μεθοδολογία – Αποτελέσματα

Το YOLOv5 αποτελεί την πιο συχνά χρησιμοποιούμενη έκδοση στη σειρά των μοντέλων "You Only Look Once", τα οποία έχουν χρησιμοποιηθεί ευρέως για την ανίχνευση αντικειμένων σε πραγματικό χρόνο λόγω της ταχύτητας και της ακρίβειάς τους. Ως μια αξιόπιστη και ευέλικτη αρχιτεκτονική, το YOLOv5 έχει γίνει ευρέως δεκτό για διάφορες εφαρμογές. Αυτή η εργασία χρησιμοποίησε το μοντέλο YOLOv5, ως βάση σύγκρισης με τις βελτιώσεις που θα μελετήθηκαν.

Συγκεκριμένα εξετάστηκαν και αξιολογήθηκαν ως προς την απόδοση τους να αναγνωρίζουν τα αντικείμενα ενδιαφέροντος από εναέριες λήψεις τρία διαφορετικά μοντέλα:

1. **Πείραμα 1 – YOLOv5x**
2. **Πείραμα 2 – YOLOv5x με Softpool και Squeze-and-Excitation module**
3. **Πείραμα 3 – YOLOv5x με Softpool και CoordAttention**

Αξίζει στο σημείο αυτό να αναφερθεί ότι σε κάθε βίντεο, ένα μοντέλο μπορεί να αναγνωρίσει έξι διαφορετικά αντικείμενα – πεζούς (pedestrians), ποδηλάτες (bikers), skaters, carts, αυτοκίνητα (cars) και λεωφορεία (bus). Ωστόσο, επειδή το Stanford Drone Dataset είναι, όπως ήδη έχει αναφερθεί, είναι biased στις κλάσεις των πεζών (pedestrians) και των ποδηλατών (bikes), οι κλάσεις skater, biker συγχωνεύτηκαν με αυτή των πεζών και η κλάση και carts συγχωνεύτηκε με αυτή των αυτοκινήτων. Ένας επιπλέον λόγος που έγινε αυτή η συγχώνευση είναι το γεγονός ότι η λήψη των εικόνων έγινε κατακόρυφα από εναέριο μη επανδρωμένο μέσο. Κάτι τέτοιο έχει ως αποτέλεσμα να είναι δύσκολο να διαχωριστεί ένας πεζός από ένα ποδηλάτη και ένα skater αφού υπάρχει επικάλυψη του ανθρώπου με το ποδήλατο ή το skate. Αυτό έχει ως αποτέλεσμα την εκπαίδευση του μοντέλων και την αξιολόγηση τους μόνο για αντικείμενα που ανήκουν στις κλάσεις car, bus και pedestrian. Η συγχώνευση αυτή είχε ως αποτέλεσμα η κλάση των πεζών να αποτελεί το 86% του συνόλου με τα αυτοκίνητα να έχουν το 12% και τα λεωφορεία το 2%.

Προκειμένου να αξιολογηθούν σωστά τα αποτελέσματα από όλες τις περιπτώσεις μελέτης ορίστηκαν οι παρακάτω παράμετροι:

1. **Διάσταση Εικόνας Εισόδου**: 1020 εικονοστοιχεία (pixels), προσφέροντας επαρκή ανάλυση για την ανίχνευση αντικειμένων. Παράλληλα, ο αριθμός αυτός συμβάλει στην ισορροπία της απαιτούμενης υπολογιστικής απόδοσης.
2. **Μέγεθος Σειράς (Batch Size)**: Ορίστηκε στο 4, με σκοπό την αποτελεσματική διαχείριση της μνήμης της GPU.
3. **Βάρη (Weights)**: Σε όλα τα πειράματα επιλέχθηκε η παραλλαγή YOLOv5x λόγω του μεγαλύτερου μεγέθους της και της ικανότητα που έχει να μαθαίνει πιθανώς πιο περίπλοκα χαρακτηριστικά από τα δεδομένα εκπαίδευσης [26].

4. **Εποχές (Epochs)**: Η διαδικασία της εκπαίδευσης διεξήχθη για 500 εποχές. Ο αριθμός αυτός είναι ικανός για να εξασφαλιστεί η πλήρης μάθηση χωρίς την παρουσία φαινομένων υπερπροσαρμογής (overfitting) του εκάστοτε μοντέλου στα δεδομένα εκπαίδευσης.

Για την εκπαίδευση των μοντέλων αλλα και την αξιολόγηση τους το dataset χωρίστικε τυχαία σε τρία τμήματα. Το 70% των δεδομένων δεσμεύτηκε και χρησιμοποιήθηκε για την εκπαίδευση και τη δημιουργία των μοντέλων (training set) ενώ το 20% για την επικύρωση (validation set) τους και την βελτίωση του μοντέλου σε κάθε εποχή. 10% του συνόλου των δεδομένων χρησιμοποιήθηκε για την αξιολόγηση των μοντέλων μετά την ολοκλήρωση της εκπαίδευσης (testing set).

Πίνακας 7.1. Αριθμός εικόνων που χρησιμοποιήθηκαν στις φάσεις ανάπτυξης και αξιολόγησης των μοντέλων.

	Αριθμός εικόνων (Num. of frames)
Train set	11.332
Test set	1.619
Validation set	3.238

Για τη δημιουργία αυτών των συνόλων χρησιμοποιήθηκε η συνάρτηση train_test_split της βιβλιοθήκης scikit-learn [36].

Η ανάπτυξη των μοντέλων πραγματοποιήθηκε σε γλώσσα προγραμματισμού Python.

Πείραμα 1 – YOLOv5x

Κατά την πρώτο πείραμα χρησιμοποιήθηκε η αρχική έκδοση του YOLOv5x. Για την εκπαίδευση του μοντέλου, χρησιμοποιήθηκαν προεκπαιδευμένα βάρη από το COCO dataset, ώστε να επιτευχθεί ταχύτερη σύγκλιση και καλύτερη αρχική απόδοση. Ακολούθησε fine-tuning στα δεδομένα του Stanford Drone Dataset, με στόχο την προσαρμογή του μοντέλου στις ιδιαιτερότητες των εναέριων εικόνων.

Παρά τις προκλήσεις που θέτει η κατακόρυφη γωνία λήψης και το μικρό μέγεθος πολλών αντικειμένων, το YOLOv5x πέτυχε πολύ υψηλή απόδοση, φτάνοντας σε mAP@.5 89% για όλες τις κλάσεις. Η ανάλυση του confusion matrix δείχνει ότι το μοντέλο λειτουργεί πολύ αποτελεσματικά, με 93% ακρίβεια εντοπισμού για τα λεωφορεία, 91% για τα αυτοκίνητα και 80% για τους πεζούς. Ωστόσο, παρατηρείται ένα μικρό ποσοστό ταξινομήσεων ως "Unknown", κυρίως στην κατηγορία των πεζών, όπου το μοντέλο δυσκολεύεται περισσότερο λόγω της ποικιλομορφίας των σχημάτων και του μικρού μεγέθους των αντικειμένων σε εναέριες εικόνες.

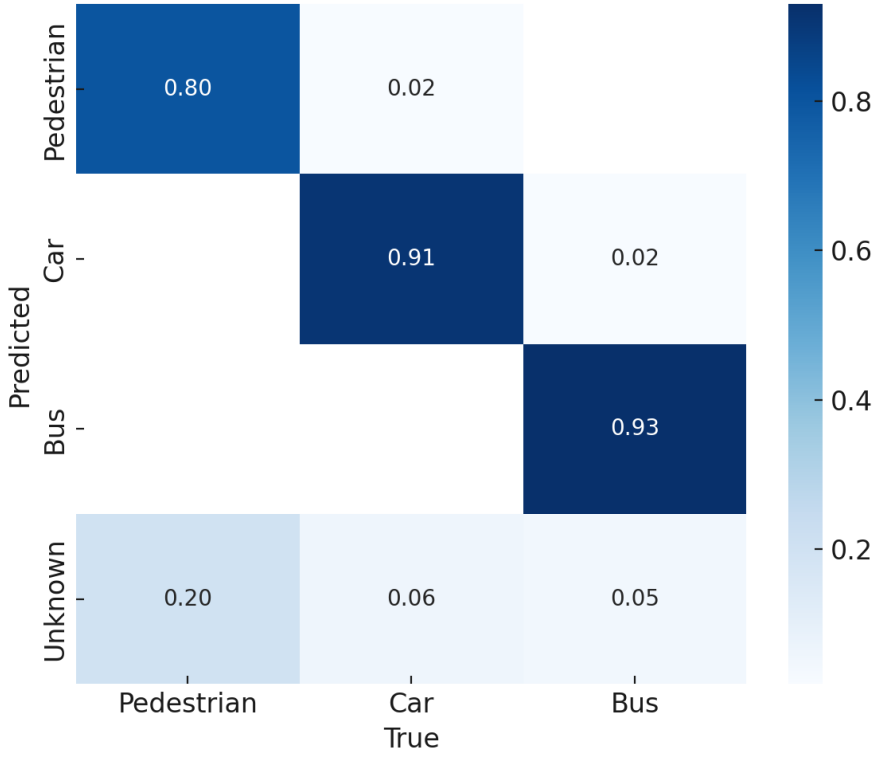
Πίνακας 7.2. Μετρικές απόδοσης του YOLOv5 original model.

Class	Labels	P	R	mAP@.5	mAP@.5:.95
all	42889	0.907	0.862	0.896	0.663
Pedestrian	37532	0.815	0.74	0.785	0.406
Car	5061	0.943	0.907	0.944	0.821
Bus	296	0.962	0.938	0.959	0.763

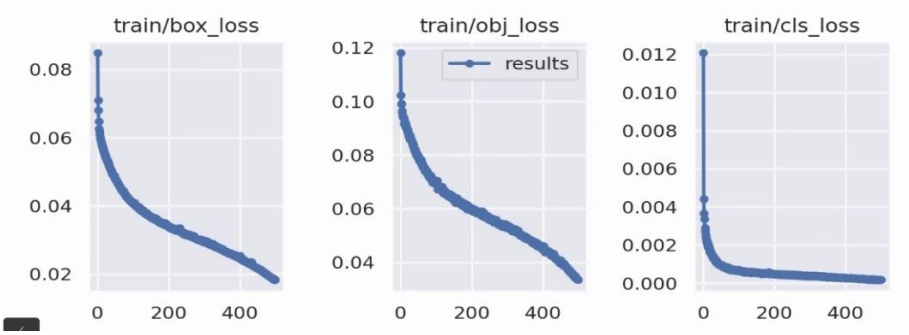
*P: Precision, R: Recall, mAP@.5: Mean Average Precision at IoU threshold 0.5, mAP@.5:.95: Mean Average Precision at IoU thresholds from 0.5 to 0.95.

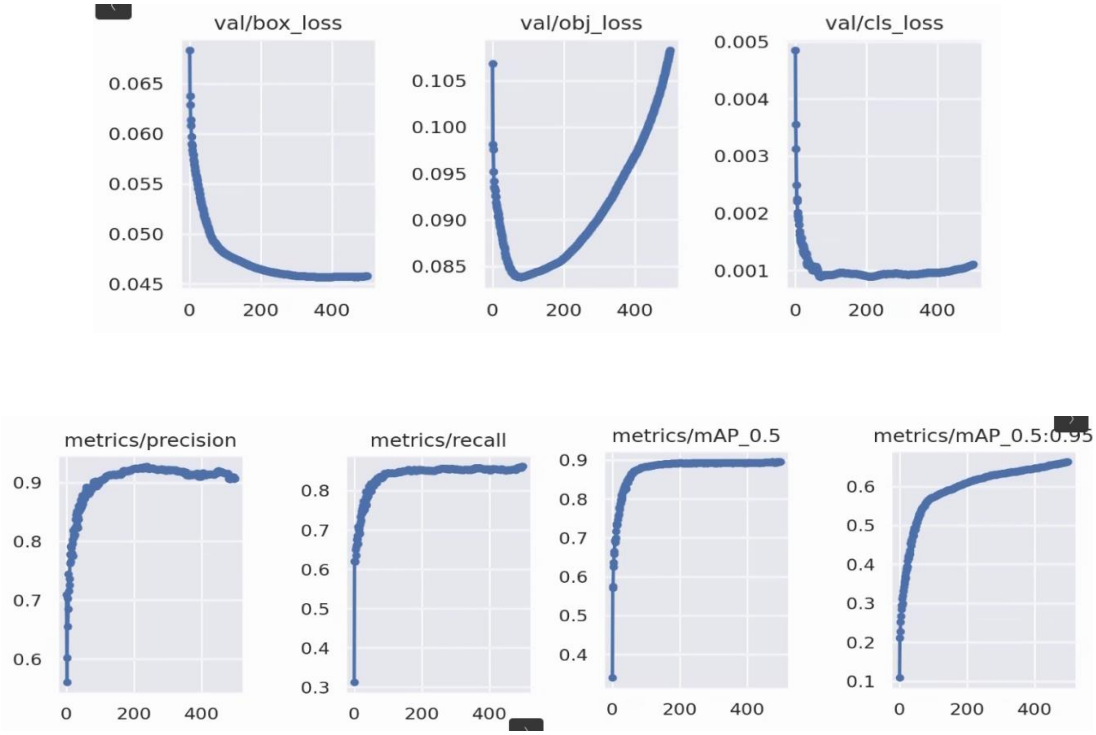
Όπως φαίνεται και από τον παραπάνω πίνακα η αρχική έκδοση του YOLOv5 (χωρίς καμία παραμετροποίηση) προσφέρει ισορροπημένα αποτελέσματα με υψηλή ακρίβεια (P) και ανάκληση (R), καθώς και υψηλές τιμές mAP@.5 και mAP@.5:.95. Αυτό υποδηλώνει ότι το μοντέλο είναι αποτελεσματικό στην αναγνώριση των κατηγοριών ενδιαφέροντος των αντικειμένων με καλή ακρίβεια.

Στις παρακάτω εικόνες, παρουσιάζονται ο confusion matrix, τα γραφήματα απωλειών κατά την εκπαίδευση και την επικύρωση, οι μετρικές ακρίβειας και ανάκλησης, αλλά και την καμπύλη ROC για την το μοντέλο YOLOv5.

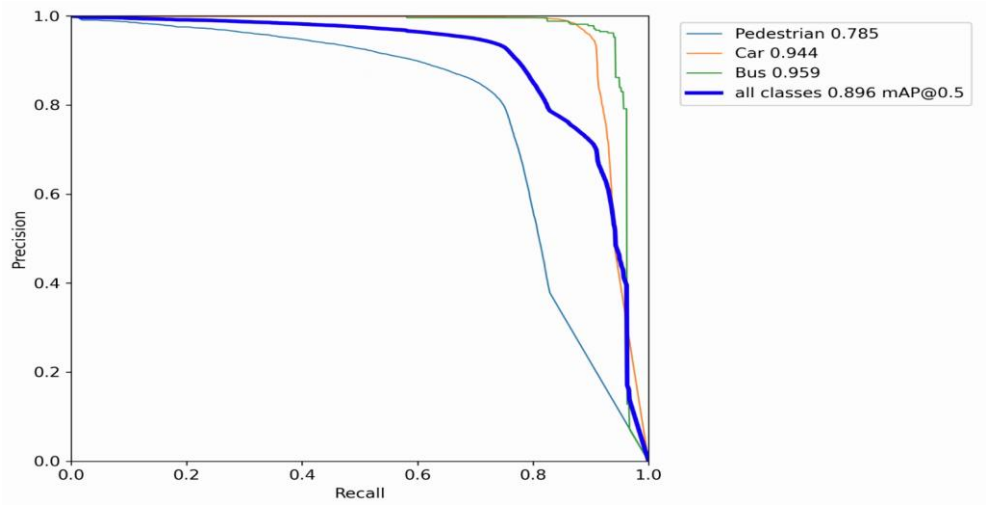


Εικόνα 7.1. YOLOv5x – Confusion matrix.





Εικόνα 7.2. Γραφήματα απωλειών για την εκπαίδευση (training) και την επικύρωση (validation) της αρχικής έκδοσης του YOLOv5 μοντέλου γραφήματα μετρικών όπως το precision και το recall.



Εικόνα 7.3. Καμπύλη ROC για το πρωτότυπο YOLOv5 μοντέλο.

Πείραμα 2 – YOLOv5x με Softpool και Squeeze-and-Excitation Module

Στην αναγνώριση αντικειμένων, όπως στην περίπτωση που μελετάται στην εργασία αυτή είναι σημαντική η απόδοση των μοντέλων. Συνεπώς, εκτός από τη βασική αρχιτεκτονική του YOLOv5 δοκιμάστηκαν εναλλακτικές αρχιτεκτονικές κάνοντας χρήση διαφόρων μεθόδων. Αρχικά αντικαταστάθηκε η κλασσική μέθοδος pooling από μία τεχνική με πιο ευέλικτη προσέγγιση, την SoftPool σε συνδιασμό με την προσθήκη ενός Squeeze and Excitation module πριν από το SPPF.

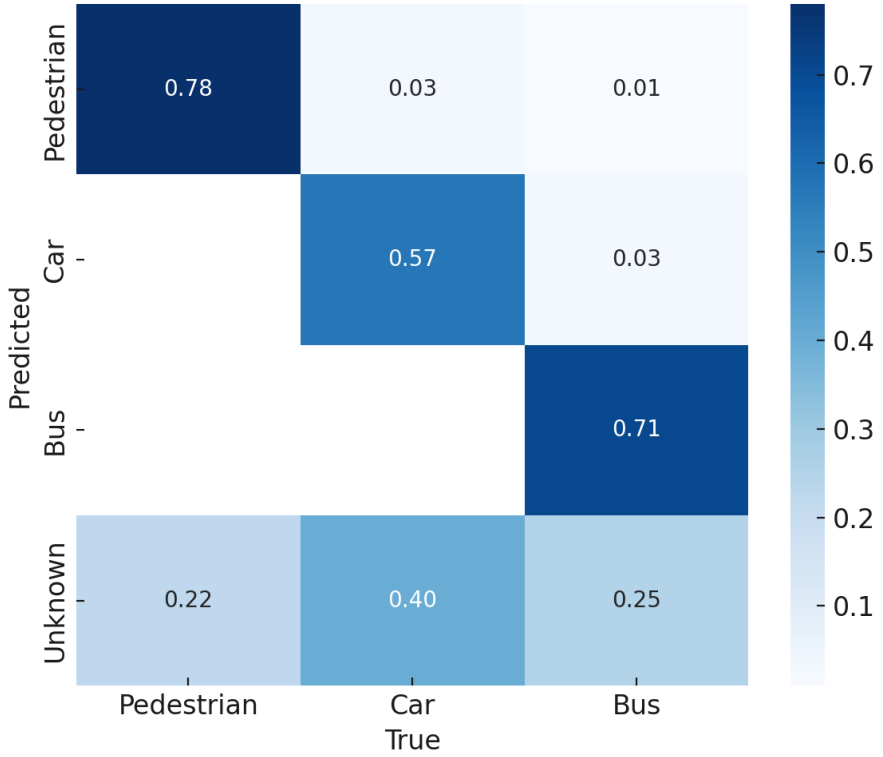
Παρότι θεωρητικά τα modules SPPF και Squeeze-and-Excitation (SE) προσφέρουν σημαντικά πλεονεκτήματα, όπως η ενίσχυση της πληροφορίας και η δυναμική ενίσχυση σημαντικών χαρακτηριστικών, στην πράξη δεν οδήγησαν σε βελτίωση της απόδοσης, καθώς το mAP μειώθηκε στο 71%, σε σύγκριση με το baseline πείραμα. Σύμφωνα με το confusion matrix, η ακρίβεια αναγνώρισης για τους πεζούς μειώθηκε στο 78%, ενώ στα αυτοκίνητα παρατηρήθηκε σημαντική πτώση, με μόνο 57% σωστές προβλέψεις και αύξηση των περιπτώσεων που ταξινομήθηκαν ως "Unknown". Παρόμοια τάση παρατηρήθηκε και στα λεωφορεία, με ακρίβεια εντοπισμού 71%. Τα αποτελέσματα αυτά υποδεικνύουν ότι, παρά τη θεωρητική τους αξία, η διατήρηση περισσότερων λεπτομερειών μέσω των SE modules ενδεχομένως δεν είναι ωφέλιμη όταν εφαρμόζεται σε εικόνες χαμηλής ανάλυσης ή με κατακόρυφη γωνία λήψης, όπως αυτές του Stanford Drone Dataset.

Η ακρίβεια είναι 0.902, και η mAP@.5:.95 είναι 0.34.

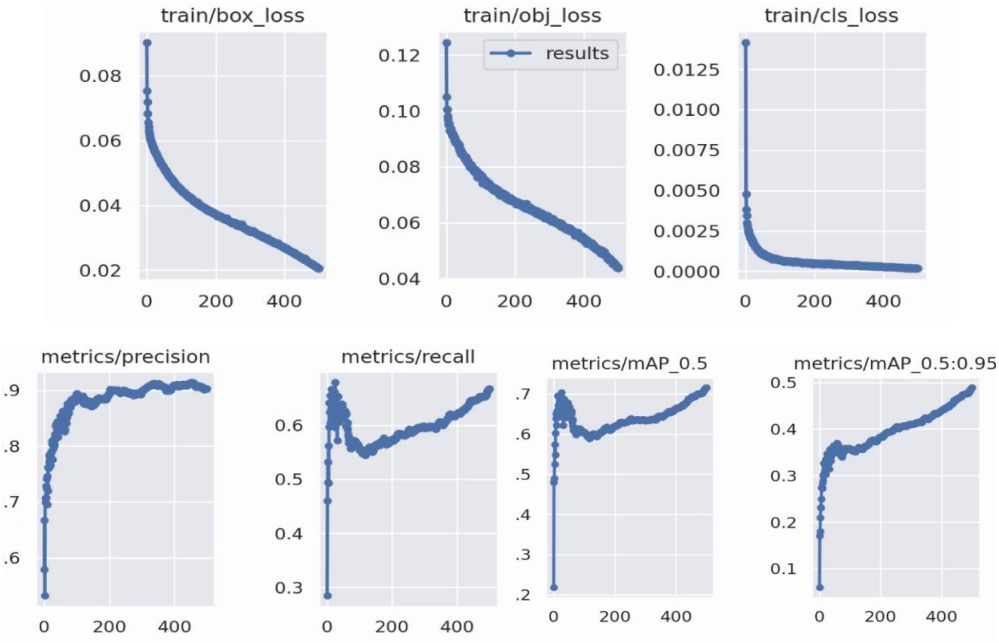
Πίνακας 7.3. Μετρικές απόδοσης του YOLOv5 model με Softpool και Squeeze-and-Excitation Module.

Class	Labels	P	R	mAP@.5	mAP@.5:.95
all	42889	0.902	0.667	0.716	0.34
Pedestrian	37532	0.812	0.712	0.731	0.342
Car	5061	0.942	0.757	0.661	0.456
Bus	296	0.955	0.722	0.755	0.58

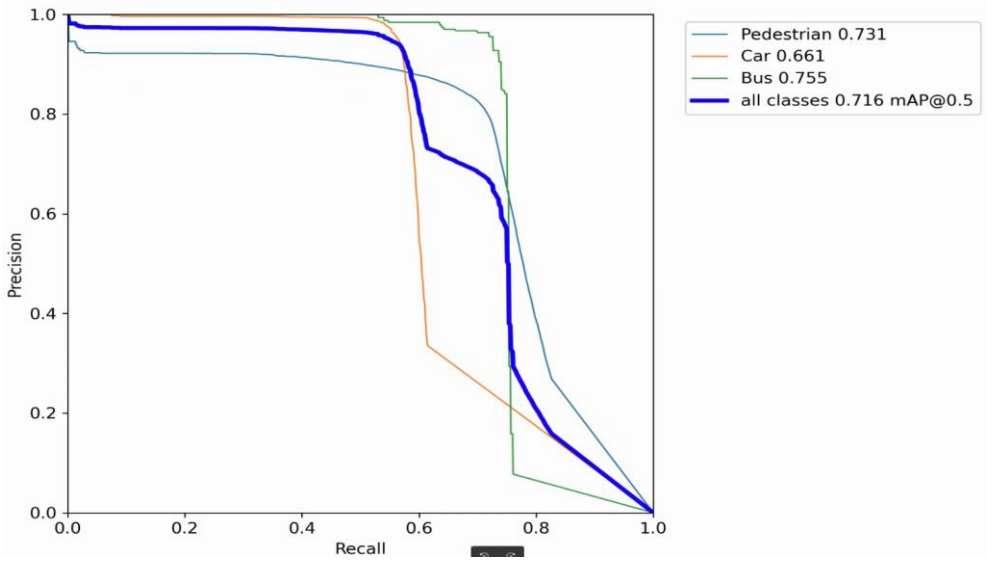
*P: Precision, R: Recall, mAP@.5: Mean Average Precision at IoU threshold 0.5, mAP@.5:.95: Mean Average Precision at IoU thresholds from 0.5 to 0.95.



Εικόνα 7.4. YOLOv5x με Softpool και Squeeze-and-Excitation Module – Confusion matrix.



Εικόνα 7.5. Γραφήματα απωλειών για την εκπαίδευση (training) και γραφήματα μετρικών όπως το precision και το recall για την περίπτωση του μοντέλου YOLOv5 με Softpool και Squeeze-and-Excitation Module.



Εικόνα 7.6. Καμπύλη Ακρίβειας-Ανάκλησης (PR) για το πρωτότυπο YOLOv5 με Softpool και Squeeze-and-Excitation Module.

Πείραμα 3 – YOLOv5x με Softpool και Coord Attention

Τέλος, δοκιμάστηκε ακόμη μία παραλλαγή. Συγκεκριμένα αυτή η παραλλαγή συνδυάζει το Softpool με τον μηχανισμό CoordAttention. Το CoordAttention εισάγει χωρική προσοχή (space attention), βελτιώνοντας την ικανότητα του μοντέλου να εντοπίζει ακριβώς τα αντικείμενα και τις θέσεις τους [40]. Όπως μπορεί να παρατηρήσει κανείς από τον πίνακα που φαίνεται παρακάτω, τα αποτελέσματα δείχνουν σημαντική βελτίωση τόσο στην ακρίβεια (0.908) όσο και στην ανάκληση

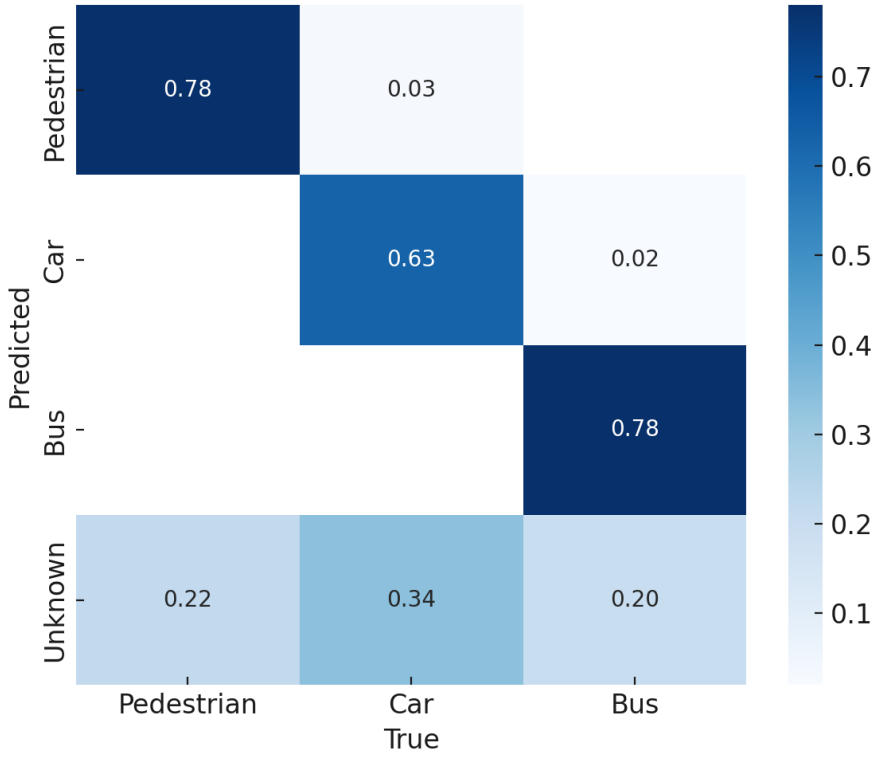
(0.7), καθώς και στις τιμές mAP, με το mAP@.5 να φτάνει το 75% και το mAP@.5:.95 το 52%. Αυτή η βελτίωση μπορεί να αποδοθεί στην ικανότητα του CoordAttention να εστιάζει καλύτερα στα χαρακτηριστικά που είναι σημαντικά για την αναγνώριση αντικειμένων, καθιστώντας το μοντέλο πιο ακριβές και αποδοτικό. Ακόμη όπως και σε αυτή την περίπτωση, δηλαδή με την εισαγωγή χωρικής προσοχής, η αρχική έκδοση του YOLOv5 παραμένει αυτή με τα καλύτερα αποτελέσματα. Ένας πιθανός λόγος γι' αυτό είναι ότι τα βίντεο - frames, έχουν κατακόρυφο προσανατολισμό, με αποτέλεσμα η χωρική διαφοροποίηση εντός των αντικειμένων να είναι περιορισμένη και η χωρική προσοχή να μην προσφέρει ουσιαστικό πλεονέκτημα.

Πίνακας 7.4. Μετρικές απόδοσης του YOLOv5 model με Softpool και Coord Attention.

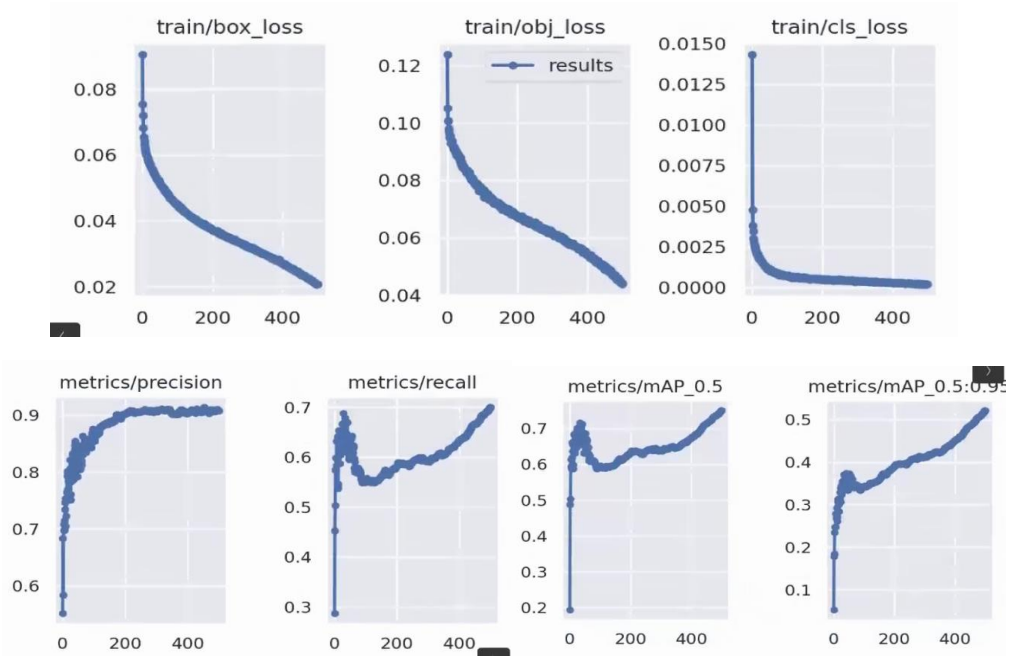
Class	Labels	P	R	mAP@.5	mAP@.5:.95
all	42889	0.908	0.7	0.75	0.521
Pedestrian	37532	0.824	0.7	0.731	0.338
Car	5061	0.95	0.618	0.711	0.591
Bus	296	0.95	0.78	0.807	0.632

*P: Precision, R: Recall, mAP@.5: Mean Average Precision at IoU threshold 0.5, mAP@.5:.95: Mean Average Precision at IoU thresholds from 0.5 to 0.95.

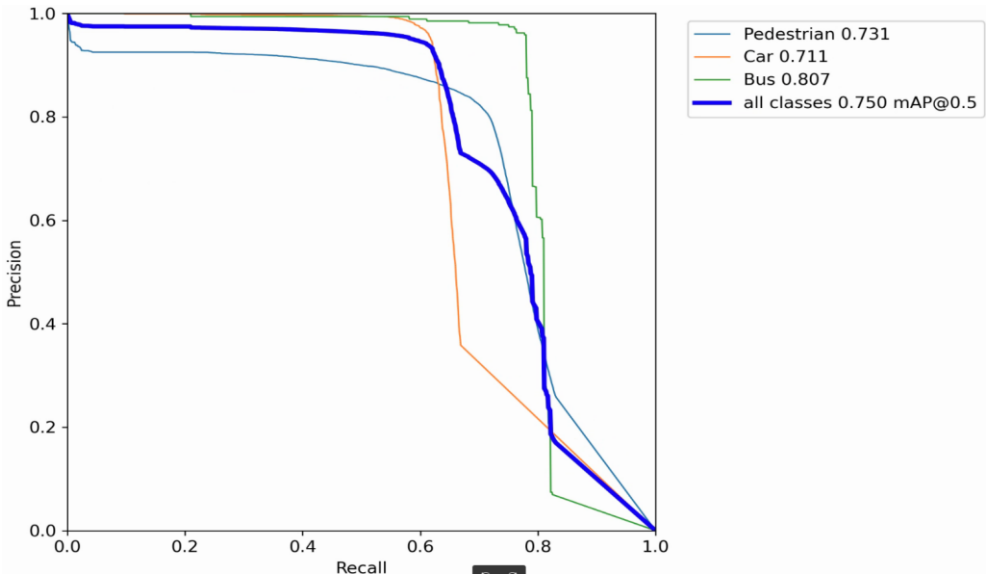
Στις παρακάτω εικόνες, μπορούμε να δούμε τον confusion matrix, τα γραφήματα απωλειών κατά την εκπαίδευση και την επικύρωση, τις μετρικές ακρίβειας και ανάκλησης, αλλά και την καμπύλη ROC.



Εικόνα 7.7. YOLOv5x με Softpool και CoordAttention – Confusion matrix.



Εικόνα 7.8. Γραφήματα απωλειών για την εκπαίδευση (training) και γραφήματα μετρικών όπως το precision και το recall για την περίπτωση του μοντέλου YOLOv5 με Softpool και CoordAttention.

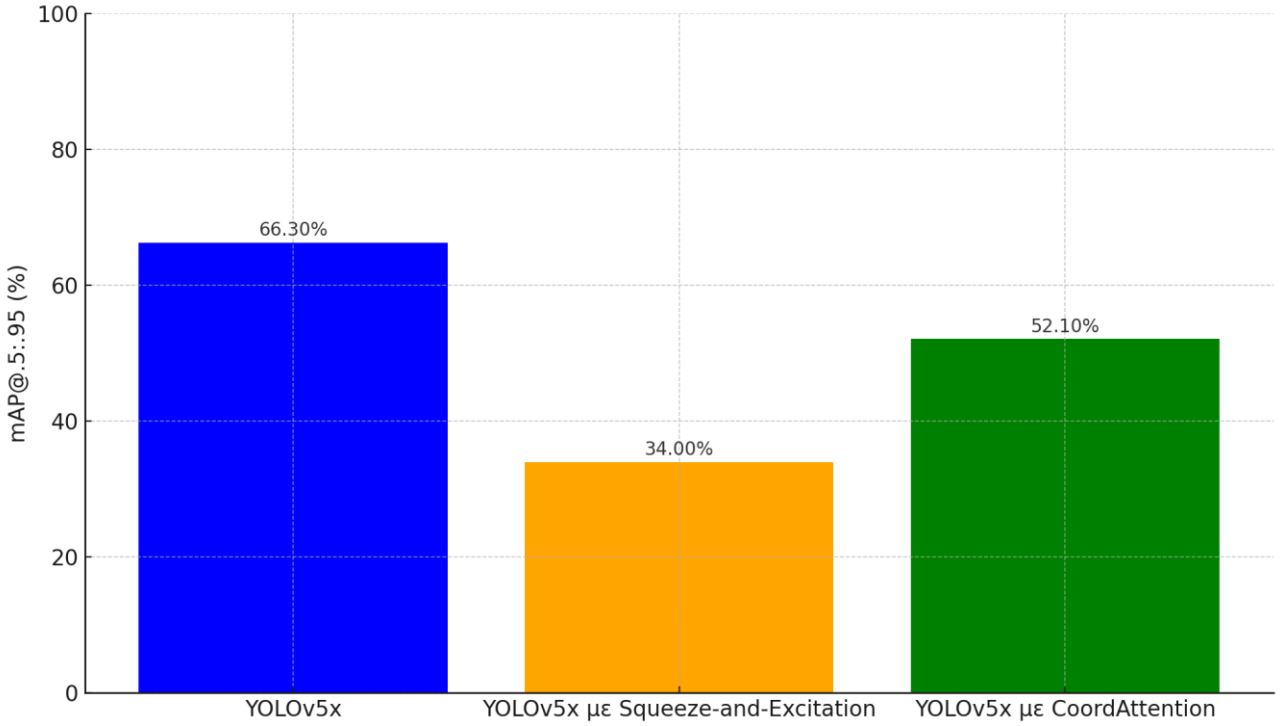


Εικόνα 7.9. Καμπύλη Ακρίβειας-Ανάκλησης (PR) για το πρωτότυπο YOLOv5 με Softpool και CoordAttention.

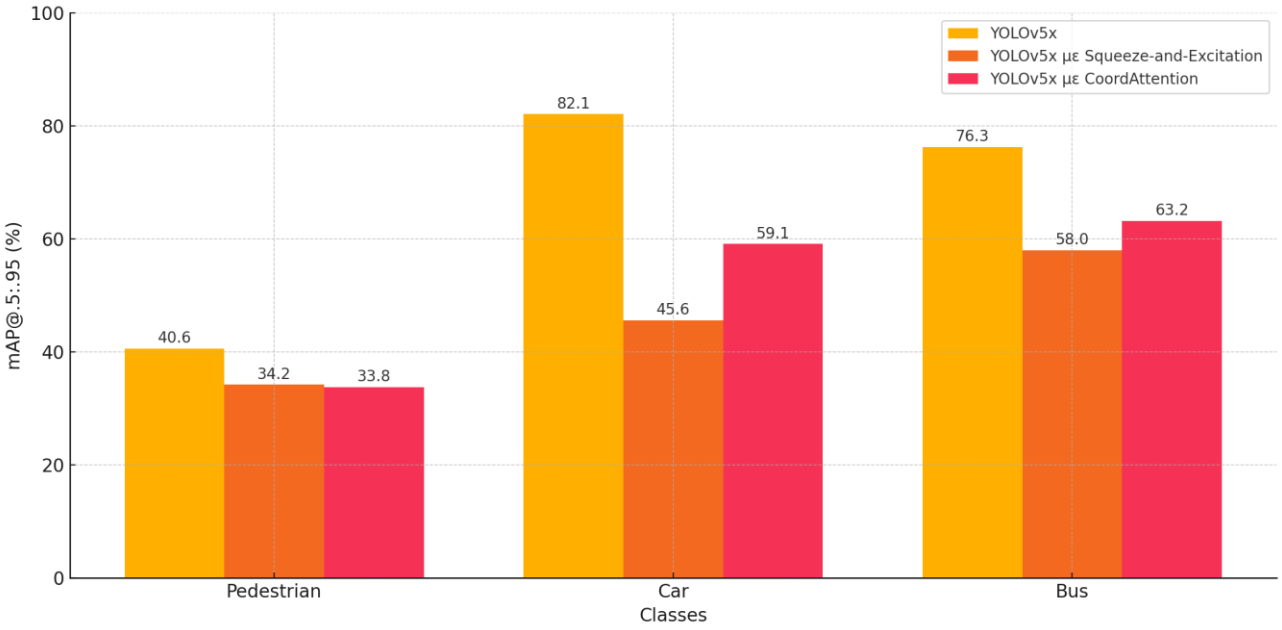
Ανάλυση των Επιδόσεων των Μοντέλων YOLOv5x και των Παραλλαγών του

Τα μοντέλα YOLOv5x, YOLOv5x με Softpool και Squeeze-and-Excitation και YOLOv5x με Softpool και CoordAttention παρουσιάζουν διαφορετικά επίπεδα απόδοσης ως προς την ακρίβεια εντοπισμού αντικειμένων (mAP@.5:.95) συνολικά αλλά και ανά κλάση ενδιαφέροντος (Pedestrian - πεζοί, Car - αυτοκίνητα και Bus - λεωφορεία). Το YOLOv5x διατηρεί σαφώς την υψηλότερη απόδοση συνολικά, επιτυγχάνοντας mAP 40.6% για πεζούς, 82.1% για αυτοκίνητα και 76.3% για λεωφορεία. Αντίθετα,

το YOLOv5x με Squeeze-and-Excitation εμφανίζει σημαντική πτώση στις επιδόσεις του, ειδικά στις κατηγορίες πεζών (34.2%) και αυτοκινήτων (45.6%), πιθανώς λόγω αυξημένου κόστους υπολογισμών ή αδυναμίας του μηχανισμού Squeeze-and-Excitation να ενισχύσει τις κρίσιμες χωρικές πληροφορίες. Το YOLOv5x με CoordAttention, παρότι υπερέχει του Squeeze-and-Excitation, δεν φτάνει το βασικό YOLOv5x, με επιδόσεις 33.8% για πεζούς, 59.1% για αυτοκίνητα και 63.2% για λεωφορεία. Συνολικά, τα αποτελέσματα δείχνουν ότι οι προσθήκες όπως Squeeze-and-Excitation και CoordAttention δεν ενισχύουν την απόδοση του YOLOv5x στην αναγνώριση αντικειμένων στο παρόν dataset. Τα αποτελέσματα αυτά αποτυπώνονται στα διαγράμματα των εικόνων 7.10 και 7.11 παρακάτω, όπου φαίνεται ξεκάθαρα η υπεροχή του βασικού YOLOv5x έναντι των εκδοχών με Squeeze-and-Excitation και CoordAttention, τόσο συνολικά όσο και στις επιμέρους κατηγορίες (Πεζός, Αυτοκίνητο, Λεωφορείο).



Εικόνα 7.10. Σύγκριση των συνολικών επιδόσεων (mAP@.5:.95) για τα μοντέλα YOLOv5x, YOLOv5x με Squeeze-and-Excitation και YOLOv5x με CoordAttention.

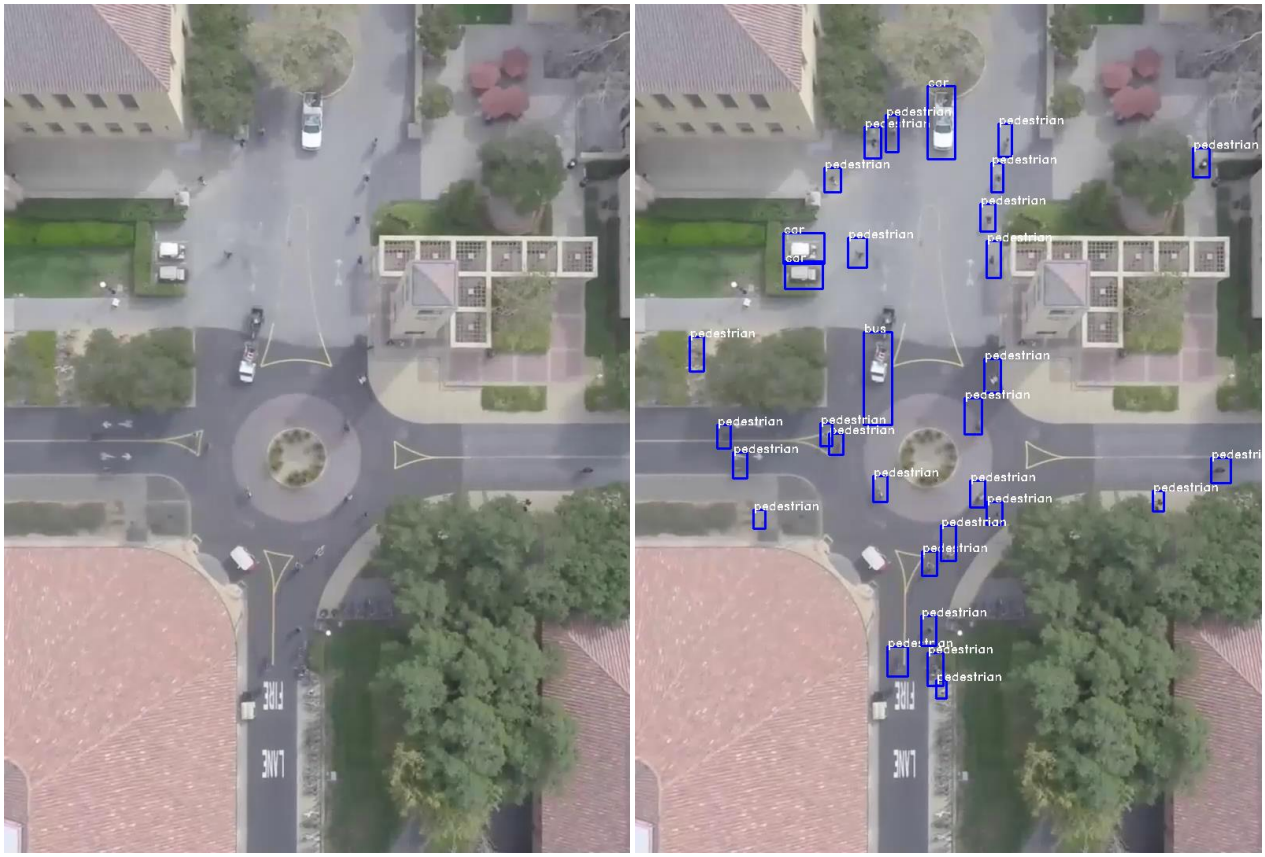


Εικόνα 7.11. Σύγκριση των επιδόσεων (mAP@.5:.95) ανά κλάση (Pedestrian-Πεζός, Car-Αυτοκίνητο και Bus-Λεωφορείο) για τα μοντέλα YOLOv5x, YOLOv5x με Squeeze-and-Excitation και YOLOv5x με CoordAttention.

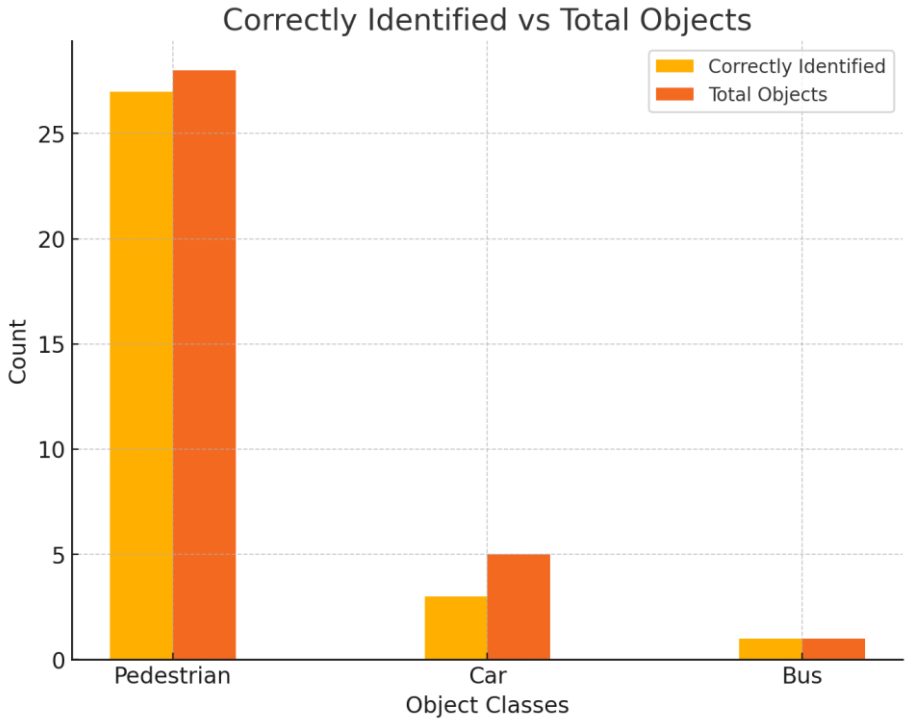
Παραδείγματα Αναγνώρισης Αντικειμένων με το YOLOv5x

Από τα παραπάνω πειράματα και συγκρίνοντας τα αποτελέσματα και τις μετρικές αξιολόγησης, η αρχική υλοποίηση του YOLOv5x που αναπτύχθηκε κατά το Πείραμα 1 αποτελεί το καλύτερο μοντέλο. Παρακάτω μπορούμε να δούμε ορισμένες εικόνες καθώς και τα αποτελέσματα της εφαρμογής του YOLOv5x σε αυτές (Εικόνες 7.12 – 7.16). Στην πρώτη φωτογραφία παρουσιάζεται ένας κυκλικός κόμβος σε αστική περιοχή. Στην φωτογραφία διακρίνονται οικήματα, δέντρα καθώς και αυτοκίνητα, λεωφορεία και πεζοί. Τα αναγνωρισμένα αντικείμενα εμφανίζουν περιμετρικά τους πλαίσιο μπλε χρώματος καθώς και ένα label το οποίο υποδεικνύει την κλάση στην οποία ανήκει το καθένα (αυτοκίνητα – car, πεζοί – pedestrian και λεωφορεία - bus).

Όπως μπορεί κανείς να παρατηρήσει από τις παρακάτω εικόνες αλλά και το διάγραμμα που παρουσιάζεται στην Εικόνα 7.13, ο αλγόριθμος αναγνώρισε με μεγάλη επιτυχία τα αντικείμενα ενδιαφέροντος στην συγκεκριμένη περίπτωση. Ειδικότερα, 27/28 πεζοί αναγνωρίστηκαν σωστά ενώ 3/5 αυτοκίνητα και 1/1 λεωφορεία.



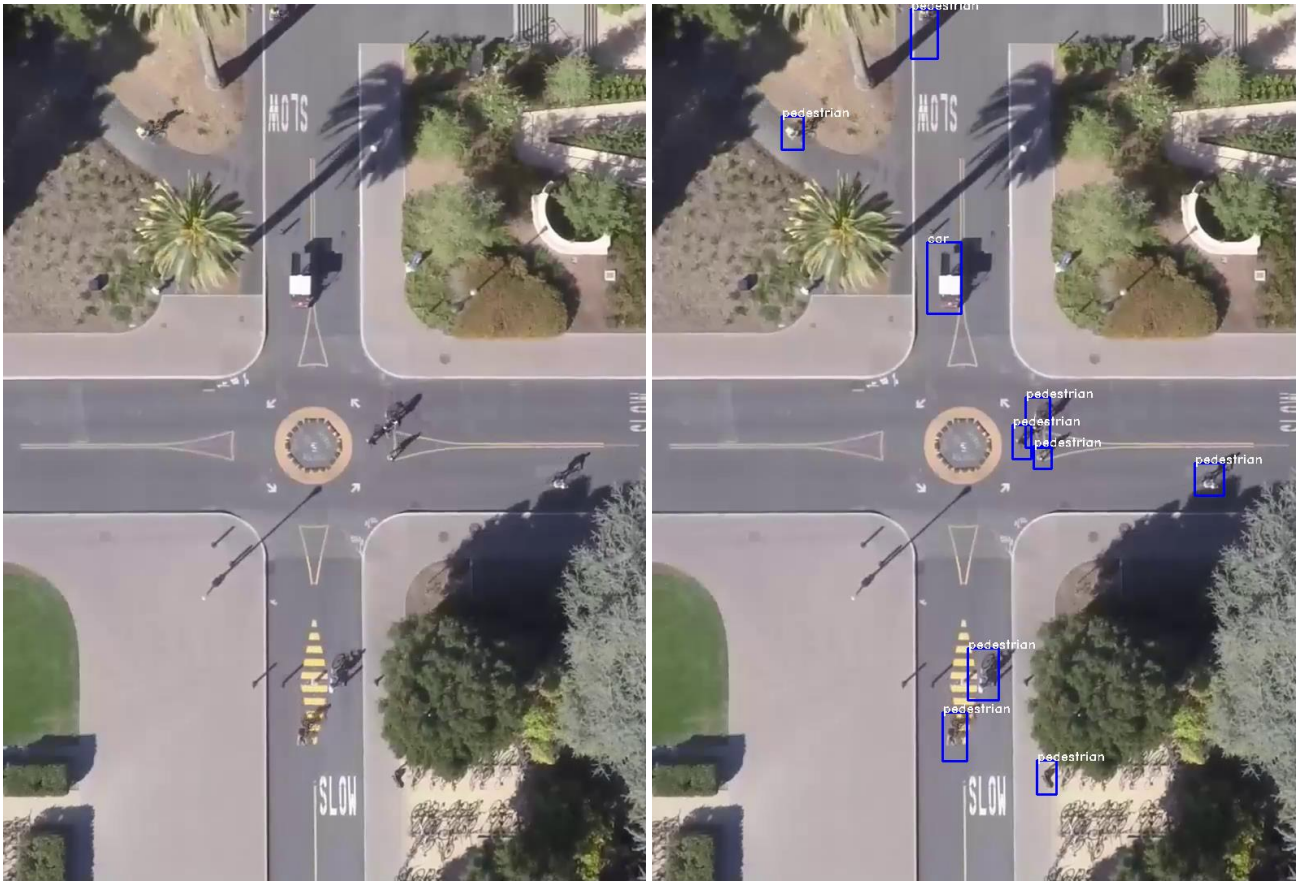
Εικόνα 7.12 Περίπτωση 1 - Αναγνώριση των αντικειμένων με χρήση του αρχικού μοντέλου YOLOv5x. Αρχική φωτογραφία (Αριστερά) και φωτογραφία με τα αναγνωρισμένα αντικείμενα σε πλαίσια μπλε χρώματος (Δεξιά).



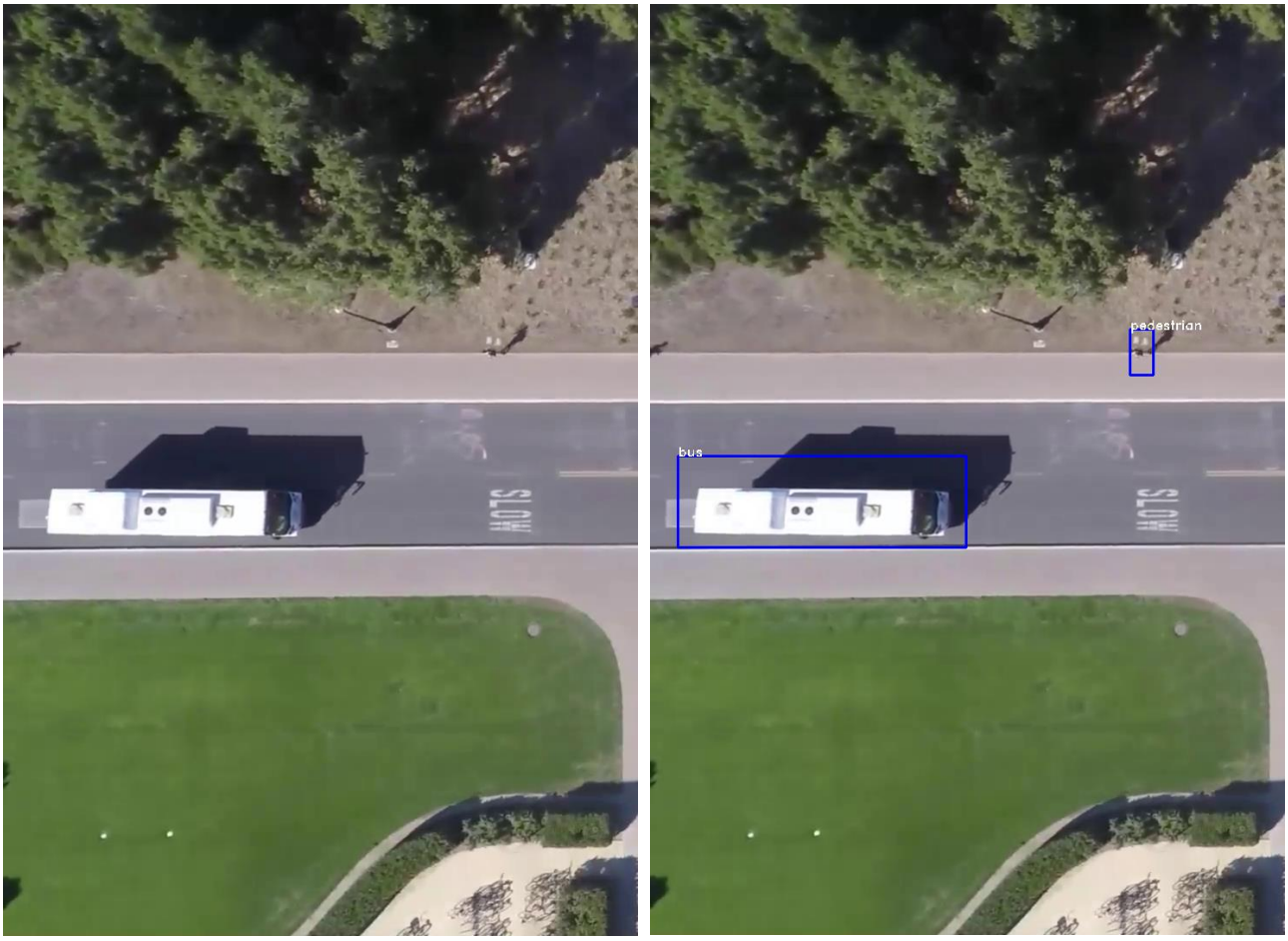
Εικόνα 7.13. Περίπτωση 1 - Αναγνώριση των αντικειμένων με χρήση του αρχικού μοντέλου YOLOv5x. Σύγκριση μεταξύ των σωστά αναγνωρισμένων και του συνολικού αριθμού αντικειμένων ενδιαφέροντος για κάθε κατηγορία.

Μερικά ακόμη παραδείγματα της ικανότητας του αλγορίθμου να αναγνωρίζει με επιτυχία τα αντικείμενα ενδιαφέροντος παρουσιάζονται στις παρακάτω εικόνες, όπου σε κάθε περίπτωση η αρχική εικόνα βρίσκεται σε αντιπαραβολή με την «τελική» εικόνα όπου τα αναγνωρισμένα αντικείμενα περικλείονται από πλαίσια μπλε χρώματος. Όπως μπορεί κανείς να παρατηρήσει το μοντέλο καταφέρνει να εντοπίσει και να κατηγοριοποιήσει τα αντικείμενα ενδιαφέροντος σωστά σε όλες τις περιπτώσεις.

Εικόνα 7.14. Περίπτωση 2 - Αναγνώριση των αντικειμένων με χρήση του αρχικού μοντέλου YOLOv5x. Αρχική φωτογραφία (Πάνω) και φωτογραφία με τα αναγνωρισμένα αντικείμενα σε πλαίσια μπλε χρώματος (Κάτω).



Εικόνα 7.15. Περίπτωση 3 - Αναγνώριση των αντικειμένων με χρήση του αρχικού μοντέλου YOLOv5x. Αρχική φωτογραφία (Αριστερά) και φωτογραφία με τα αναγνωρισμένα αντικείμενα σε πλαίσια μπλε χρώματος (Δεξιά).



Εικόνα 7.16. Περίπτωση 4 - Αναγνώριση των αντικειμένων με χρήση του αρχικού μοντέλου YOLOv5x. Αρχική φωτογραφία (Αριστερά) και φωτογραφία με τα αναγνωρισμένα αντικείμενα σε πλαίσια μπλε χρώματος (Δεξιά).

Κεφάλαιο 8 Εντοπισμός Χρώματος

Ωστόσο, η αναγνώριση αντικειμένων από μόνη της δεν αρκεί για την επίτευξη ολοκληρωμένων λύσεων σε πολλές εφαρμογές υπολογιστικής όρασης και επεξεργασίας εικόνας. Για τον λόγο αυτό, κατά τη διάρκεια αυτής της εργασίας, πραγματοποιήθηκε παράλληλα και ο προσδιορισμός των χαρακτηριστικών των αντικειμένων. Αυτή η διαδικασία είναι κρίσιμη για την εξαγωγή λεπτομερέστερων πληροφοριών που μπορούν να βελτιώσουν την απόδοση και την ακρίβεια των συστημάτων αναγνώρισης.

Ειδικότερα, τα χαρακτηριστικά που αναλύθηκαν και προσδιορίστηκαν περιλαμβάνουν:

1. **Αφαίρεση του Υποβάθρου (Background)**: Η απομόνωση του προσκηνίου από το υπόβαθρο είναι ένα κρίσιμο βήμα για την ενίσχυση της ακρίβειας στην αναγνώριση αντικειμένων. Αυτό επιτρέπει την εστίαση στα ουσιώδη χαρακτηριστικά του αντικειμένου, μειώνοντας τον θόρυβο και τις παρεμβολές από το υπόβαθρο. Στην εργασία αυτή χρησιμοποιήθηκε ο αλγόριθμος GrabCut [41], ο οποίος παρέχει ένα αποτελεσματικό και ακριβές εργαλείο για την τμηματοποίηση εικόνων, διαχωρίζοντας το προσκήνιο από το υπόβαθρο μέσω επαναληπτικών περικοπών γραφημάτων (graph cuts).
2. **Αναγνώριση του Χρώματος (Color Detection)**: Η αναγνώριση χρώματος είναι ένα σημαντικό χαρακτηριστικό που μπορεί να προσφέρει πολύτιμες πληροφορίες για την ταυτοποίηση και την κατηγοριοποίηση αντικειμένων. Η ανάλυση των χρωματικών χαρακτηριστικών επιτρέπει την ταξινόμηση και αναγνώριση αντικειμένων βάσει των χρωματικών τους προφίλ, ενισχύοντας την ικανότητα των συστημάτων να διακρίνουν μεταξύ αντικειμένων με παρόμοια σχήματα αλλά διαφορετικά χρώματα.

Αυτός ο συνδυασμός της αφαίρεσης του background και της αναγνώρισης χρώματος προσφέρει μια ολοκληρωμένη μέθοδο για την βελτιστοποίηση της αναγνώρισης αντικειμένων. Η αφαίρεση του background εξασφαλίζει ότι τα χαρακτηριστικά που εξάγονται ανήκουν αποκλειστικά στο αντικείμενο ενδιαφέροντος, ενώ η αναγνώριση χρώματος παρέχει επιπρόσθετη διάσταση πληροφορίας, καθιστώντας δυνατή την ακριβέστερη και λεπτομερέστερη ανάλυση των εικόνων. Συνεπώς, οι τεχνικές που αναπτύχθηκαν και εφαρμόστηκαν κατά τη διάρκεια αυτής της εργασίας συντελούν στη βελτίωση των συστημάτων αναγνώρισης αντικειμένων, προσφέροντας μεγαλύτερη ακρίβεια και αξιοπιστία στις εφαρμογές τους.

Αλγόριθμος GrabCut

Ο αλγόριθμος GrabCut είναι ένα ισχυρό εργαλείο για την τμηματοποίηση εικόνων, το οποίο αποσκοπεί στην απομόνωση του προσκηνίου από το υπόβαθρο σε μια εικόνα. Αναπτύχθηκε από τους Rother, Kolmogorov, και Blake το 2004 [41] και βασίζεται σε επαναληπτικές περικοπές γραφημάτων (graph cuts), παρέχοντας έναν αποτελεσματικό τρόπο εξαγωγής αντικειμένων από

σύνθετα υπόβαθρα. Σε αυτή την ενότητα θα εξετάσουμε τον μηχανισμό λειτουργίας του αλγορίθμου GrabCut, τις εφαρμογές του και τις λεπτομέρειες υλοποίησής του.

Ουσιαστικά χρησιμοποιεί έναν συνδυασμό περικοπών γραφημάτων (graph cuts) και μοντέλων Gaussian Mixture Models (GMMs) [42] για να επιτύχει την τμηματοποίηση εικόνων. Η διαδικασία ξεκινά με μια αρχική εισαγωγή από τον χρήστη, συνήθως ένα ορθογώνιο πλαίσιο γύρω από το αντικείμενο ενδιαφέροντος. Αυτή η εισαγωγή βοηθά στον διαχωρισμό του προσκηνίου από το υπόβαθρο, κάτι που είναι κρίσιμο για την απόδοση του αλγορίθμου.

Ο αλγόριθμος μπορεί να διαχωριστεί στα εξής βήματα:

1. **Αρχικοποίηση**: Ο χρήστης ορίζει ένα ορθογώνιο πλαίσιο γύρω από το αντικείμενο ενδιαφέροντος. Τα pixel μέσα στο πλαίσιο σημειώνονται ως πιθανό προσκήνιο, ενώ τα pixel έξω από το πλαίσιο σημειώνονται ως βέβαιο υπόβαθρο.

2. **Αρχικοποίηση Gaussian Mixture Model (GMM)**: Δημιουργούνται δύο GMMs: ένα για το προσκήνιο και ένα για το υπόβαθρο. Οι κατανομές χρωμάτων των pixel που σημειώνονται ως προσκήνιο και υπόβαθρο μοντελοποιούνται χρησιμοποιώντας αυτά τα GMMs.

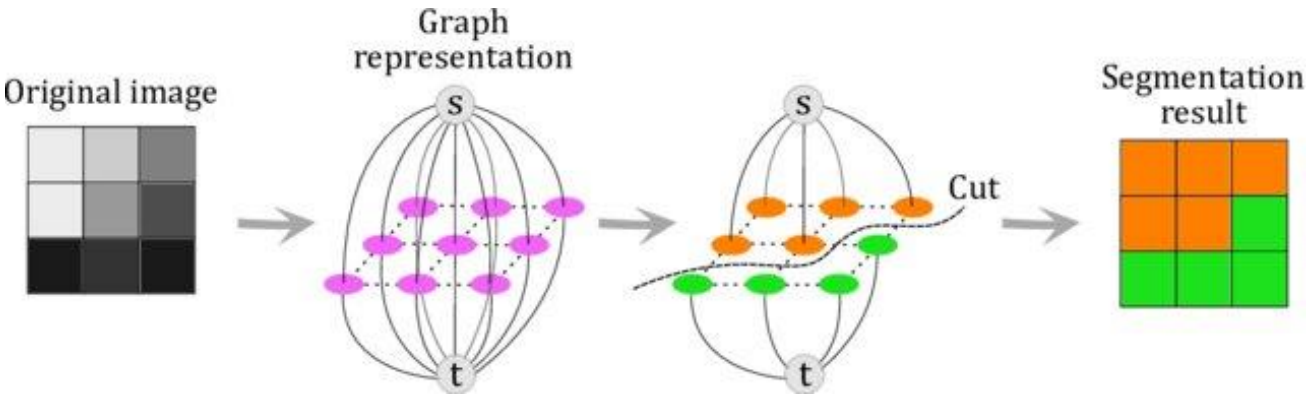
3. **Κατασκευή Γραφήματος**: Ένα γράφημα κατασκευάζεται όπου κάθε pixel είναι ένας κόμβος συνδεδεμένος με μια πηγή και ένα άκρο. Οι ακμές μεταξύ των κόμβων είναι σταθμισμένες βάσει της χρωματικής ομοιότητας (χρησιμοποιώντας τις πιθανότητες των GMMs), και οι ακμές μεταξύ των pixel και της πηγής/άκρου σταθμίζονται βάσει του εάν το pixel είναι πιθανό προσκήνιο ή υπόβαθρο.

4. **Βελτιστοποίηση Min-Cut**: Ο αλγόριθμος εκτελεί μια περικοπή min-cut στο γράφημα, διαχωρίζοντας τα pixel σε προσκήνιο και υπόβαθρο βρίσκοντας την βέλτιστη περικοπή που ελαχιστοποιεί το κόστος.

5. **Επαναληπτική Βελτίωση**: Τα αποτελέσματα της τμηματοποίησης βελτιώνονται επαναληπτικά. Τα GMMs ενημερώνονται βάσει της τρέχουσας τμηματοποίησης και η περικοπή γραφήματος επαναϋπολογίζεται. Αυτή η διαδικασία επαναλαμβάνεται μέχρι τη σύγκλιση.

6. **Τελική Τμηματοποίηση**: Μετά από αρκετές επαναλήψεις, ο αλγόριθμος συγκλίνει, παράγοντας μια δυαδική μάσκα που τμηματοποιεί ακριβώς το προσκήνιο από το υπόβαθρο.

Η παραπάνω διαδικασία απεικονίζεται στην παρακάτω εικόνα:



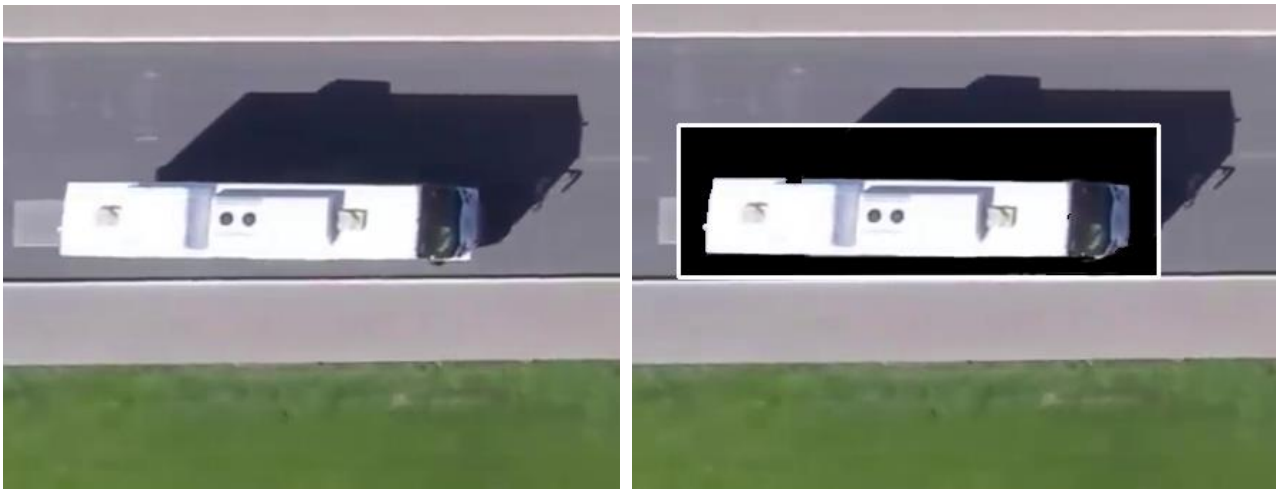
Εικόνα 8.1. Αλγόριθμος Grabcut [43].

Ο αλγόριθμος GrabCut έχει ευρεία γκάμα εφαρμογών σε διάφορους τομείς, όπως η επεξεργασία εικόνας με την εξαγωγή αντικειμένων από φωτογραφίες για αντικατάσταση ή αφαίρεση background αλλά και η υπολογιστική όραση στην οποία ο αλγόριθμος αποτελεί ένα βήμα προ επεξεργασίας στην αναγνώριση και παρακολούθηση αντικειμένων. Αποτελεί μια ισχυρή και αποδοτική μέθοδος για την τμηματοποίηση εικόνων, αξιοποιώντας την ισχύ των graph cuts και των Gaussian Mixture Models. Η επαναληπτική φύση του εξασφαλίζει ακριβή τμηματοποίηση, καθιστώντας τον κατάλληλο για ποικίλες εφαρμογές.

Η διαθεσιμότητα του GrabCut σε βιβλιοθήκες όπως το OpenCV [43] απλοποιεί περαιτέρω την υλοποίησή του, επιτρέποντας στους προγραμματιστές και ερευνητές να ενσωματώσουν εύκολα προηγμένες τεχνικές τμηματοποίησης στα έργα τους. Δύο παραδείγματα χρήσης του αλγορίθμου φαίνεται στις παρακάτω εικόνες (Εικόνα 8.2 και Εικόνα 8.3).

Συγκεκριμένα, στην πρώτη περίπτωση της εικόνας με τον ποδοσφαιριστή ο αλγόριθμος έχει αφαιρέσει επιτυχώς το background ενώ όπως βλέπουμε έχει διατηρηθεί το σώμα του. Στη δεύτερη περίπτωση, παρουσιάζεται ένα ακόμη παράδειγμα τμηματοποίησης σε εικόνα και αφαίρεσης του υπόβαθρου. Σε αυτή την περίπτωση, χρησιμοποιήθηκε μία εικόνα από το dataset του Stanford. Στην εικόνα αυτή, μπορούμε να αναγνωρίσουμε ένα λεωφορείο, του οποίου το υπόβαθρο έχει αφαιρεθεί επιτυχώς με τη χρήση του αλγορίθμου GrabCut. Η επιτυχημένη εφαρμογή του αλγορίθμου αναδεικνύει την ευελιξία του σε πολύπλοκες σκηνές και διαφορετικά είδη αντικειμένων, καθιστώντας τον ιδανικό για εφαρμογές σε ποικίλα πεδία.

Εικόνα 8.2. Παράδειγμα segmented εικόνας με αφαίρεση του background κάνοντας χρήση του αλγορίθμου GrabCut. Αρχική εικόνα (αριστερά) και επεξεργασμένη εικόνα (δεξιά).

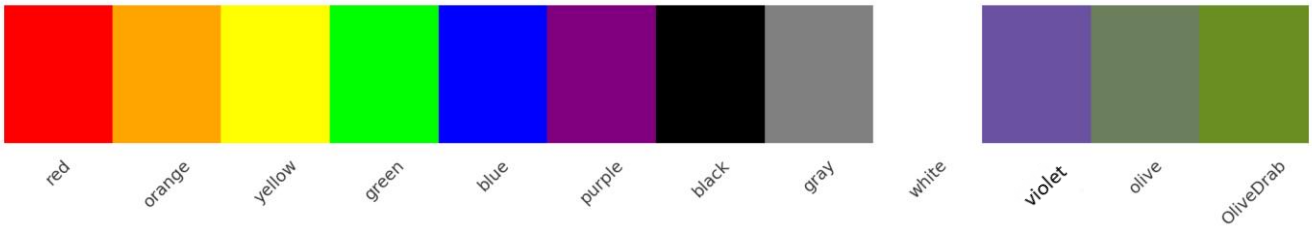


Εικόνα 8.3. Παράδειγμα segmented εικόνας από το dataset του Stanford, όπου έχει αφαιρεθεί το background κάνοντας χρήση του αλγορίθμου GrabCut. Στην αρχική εικόνα (αριστερά) διακρίνεται ένα λεωφορείο, ενώ στην επεξεργασμένη εικόνα (δεξιά) το background έχει αφαιρεθεί επιτυχώς.

Εντοπισμός Χρώματος σε Αντικείμενα του Stanford Drone Dataset

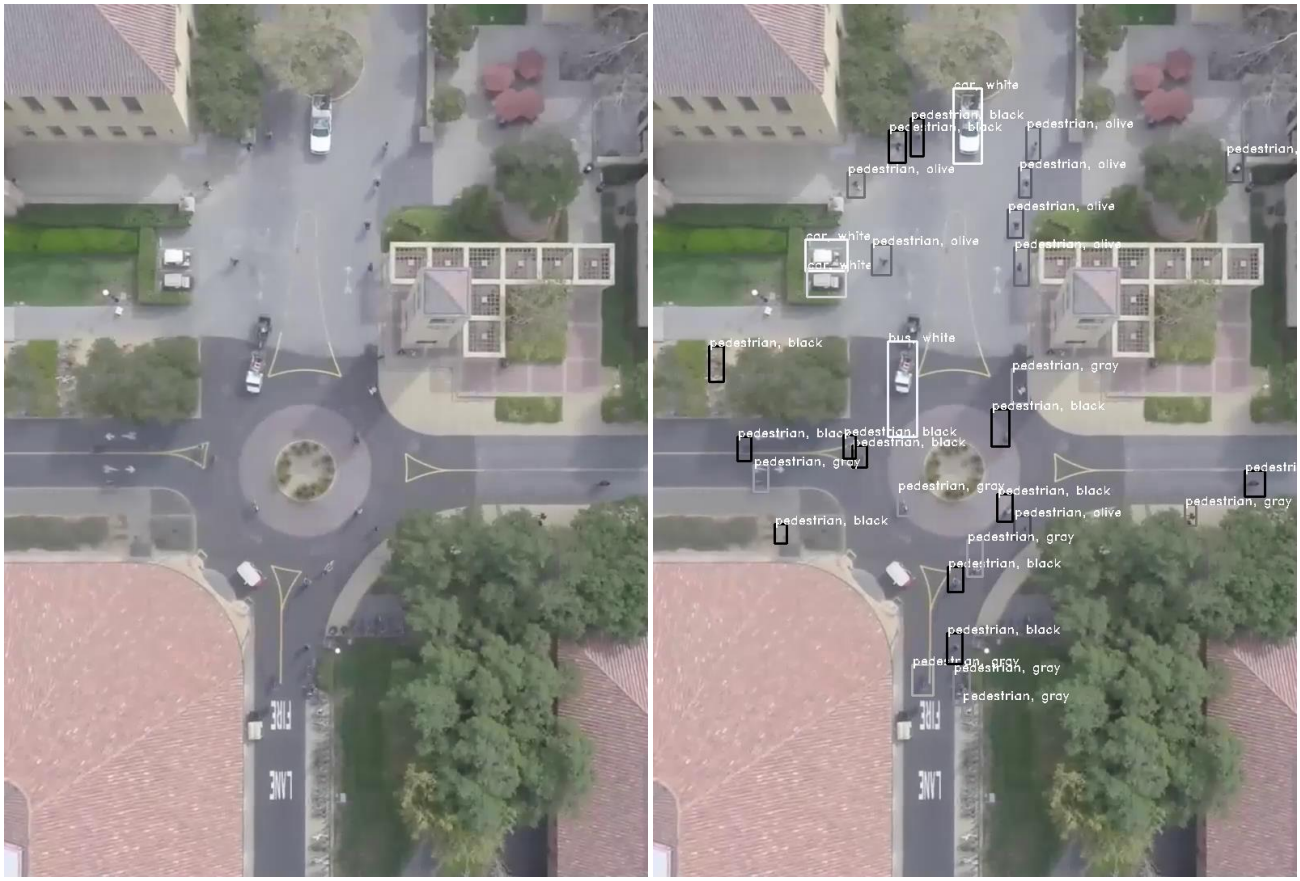
Η χρήση του Grabcut είναι ιδιαίτερα χρήσιμη για την απομόνωση και αναγνώριση αντικειμένων. Ο αλγόριθμος χρησιμοποιήθηκε σε συνδυασμό με το μοντέλο YOLOv5x με σκοπό την αφαίρεση του υποβάθρου και στη συνέχεια τον εντοπισμό του χρώματος του κάθε αντικειμένου. Πιο συγκεκριμένα αφού εντοπιστεί το αντικείμενο με την βοήθεια του μοντέλου YOLOv5x, αφαιρείται το υπόβαθρο μέσα από το bounding box, κάνοντας χρήση του OpenCV και συγκεκριμένα του GrabCut. Για τον εντοπισμό του χρώματος, αρχικά εξάγονται όλα τα pixels του εντοπισμένου αντικειμένου και δημιουργείται ένα ιστόγραμμα χρώματος στο χρωματικό μοντέλο RGB. Από το ιστόγραμμα, υπολογίζεται η επικρατούσα τιμή των χρωμάτων, η οποία αντιπροσωπεύει το κυρίαρχο χρώμα του αντικειμένου. Στη συνέχεια, για να αντιστοιχιστεί το προκύπτον χρώμα σε μια προκαθορισμένη παλέτα βασικών χρωμάτων (webcolors), υπολογίζεται η Ευκλείδεια απόσταση μεταξύ του μέσου χρώματος του αντικειμένου και κάθε χρώματος της παλέτας. Το χρώμα με τη μικρότερη απόσταση θεωρείται το πιο κοντινό και αντιστοιχεί στο αντικείμενο.

Η παλέτα RGB (12 διαφορετικά χρώματα) που χρησιμοποιήθηκε στην εργασία φαίνεται στην παρακάτω εικόνα (Εικόνα 8.4):



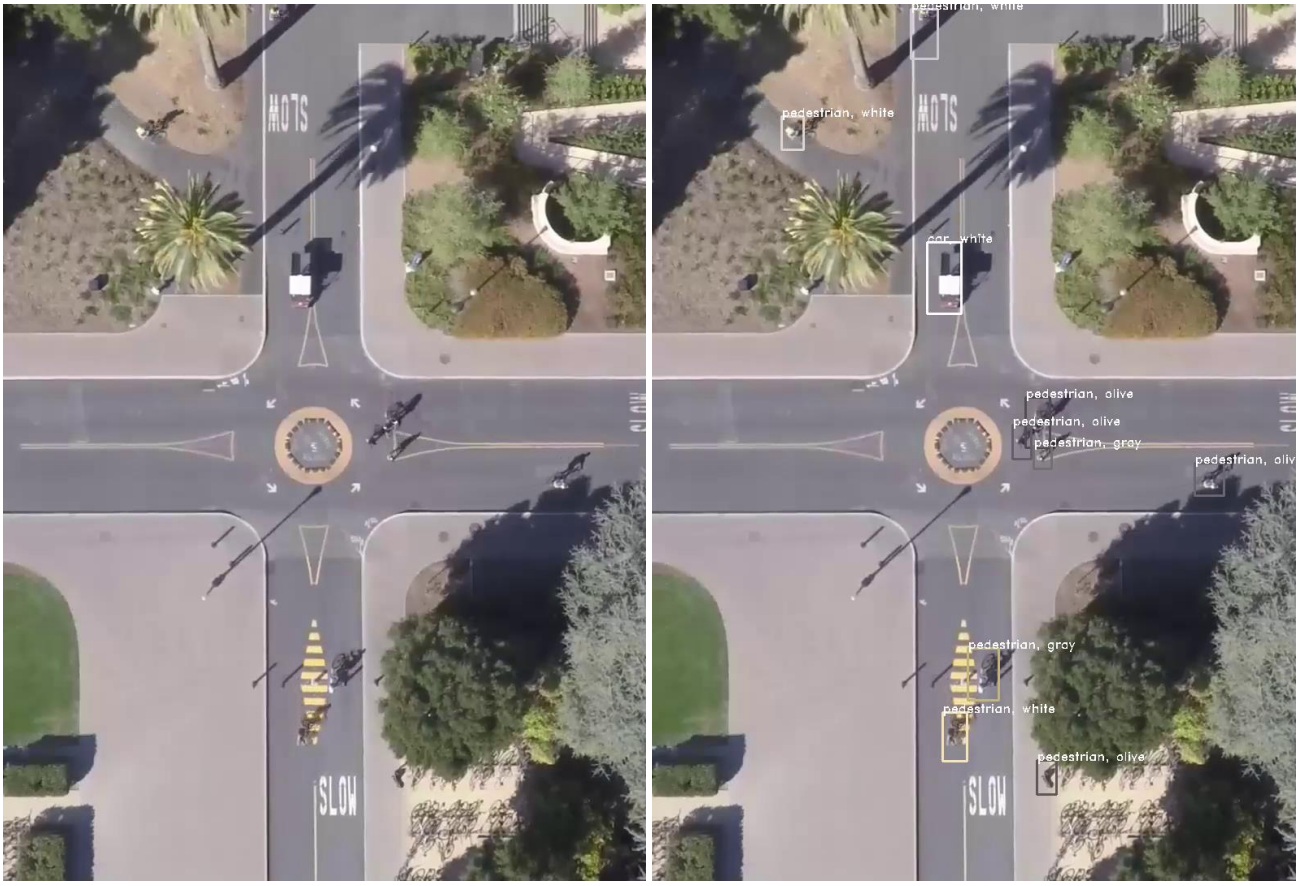
Εικόνα 8.4 Χρωματική παλέτα για τον εντοπισμό χρώματος σε αντικείμενα του Stanford Drone Dataset.

Στις παρακάτω εικόνες παρουσιάζονται ενδεικτικά τέσσερεις περιπτώσεις αναγνώρισης αντικειμένων και εντοπισμού του χρώματος τους από το Stanford Drone Dataset. Όπως φαίνεται από τις εικόνες παρακάτω, σε όλες τις περιπτώσεις τα αντικείμενα (bus, pedestrian και car) εντοπίστηκαν, αναγνωρίστηκαν και το χρώμα τους εντοπίστηκε επιτυχώς.

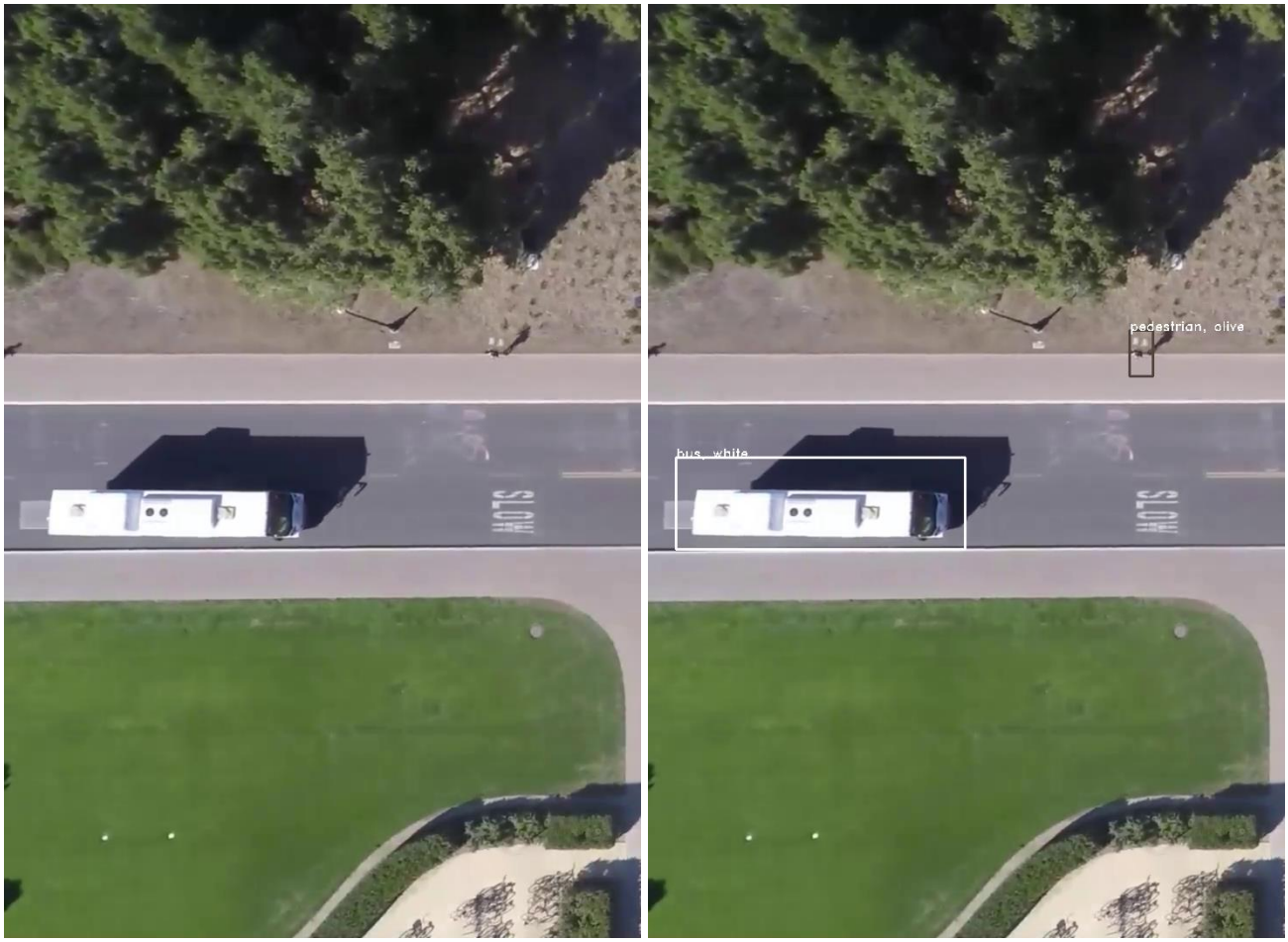


Εικόνα 8.5 Περίπτωση 1 - Αναγνώριση των αντικειμένων και εντοπισμός χρώματος χρησιμοποιώντας το μοντέλο YOLOv5x και GrabCut. Αρχική εικόνα (Αριστερά) και εικόνα με τα αναγνωρισμένα αντικείμενα σε πλαίσιο περιμετρικά καθώς και το εντοπισμένο χρώμα (Δεξιά).

Εικόνα 8.6. Περίπτωση 2 - Αναγνώριση των αντικειμένων και εντοπισμός χρώματος χρησιμοποιώντας το μοντέλο YOLOv5x και GrabCut. Αρχική εικόνα (Πάνω) και εικόνα με τα αναγνωρισμένα αντικείμενα σε πλαίσιο περιμετρικά καθώς και το εντοπισμένο χρώμα (Κάτω).



Εικόνα 8.7. Περίπτωση 3 - Αναγνώριση των αντικειμένων και εντοπισμός χρώματος χρησιμοποιώντας το μοντέλο YOLOv5x και GrabCut. Αρχική εικόνα (Αριστερά) και εικόνα με τα αναγνωρισμένα αντικείμενα σε πλαίσιο περιμετρικά καθώς και το εντοπισμένο χρώμα του αντικειμένου (Δεξιά).



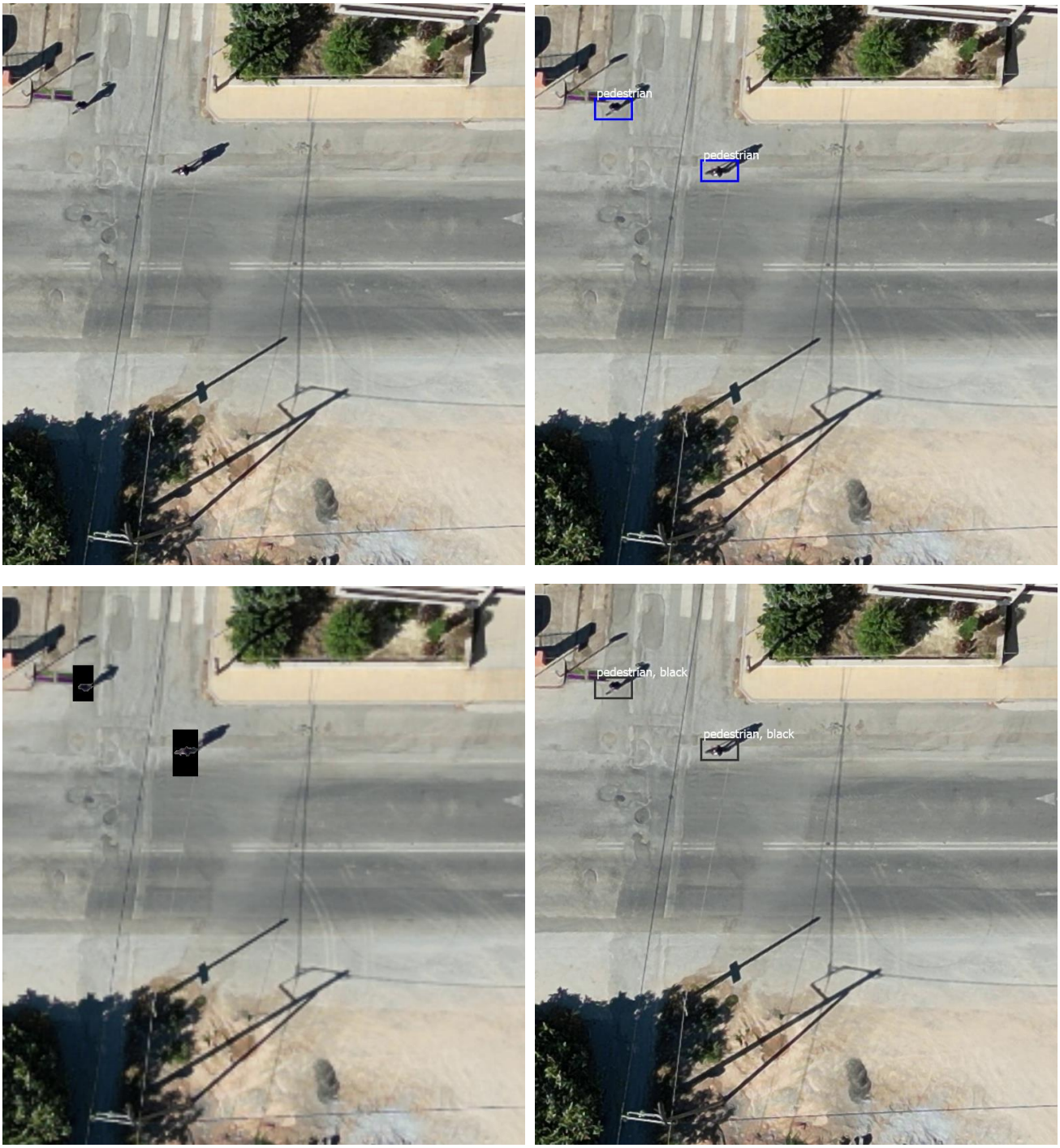
Εικόνα 8.8. Περίπτωση 4 - Αναγνώριση των αντικειμένων με χρήση του αρχικού μοντέλου YOLOv5x. Αρχική φωτογραφία (Αριστερά) και φωτογραφία με τα αναγνωρισμένα αντικείμενα σε πλαίσια μπλε χρώματος (Δεξιά).

Τέλος, αξίζει να σημειωθεί ότι το σύστημα που αναπτύχθηκε κατά την εργασία αυτή δοκιμάστηκε και σε διαφορετικό dataset εικόνων προκειμένου να ελέγξουμε την ικανότητα του να αναγνωρίζει αντικείμενα και να εντοπίζει το χρώμα τους. Πιο συγκεκριμένα χρησιμοποιήθηκαν αεροφωτογραφίες από το εργαστήριο «Εργαστήριο Ψηφιακής Επεξεργασίας Σήματος και Εικόνας» του Πολυτεχνείου Κρήτης για να ελέγξει κατά πόσο μπορούν να εντοπιστούν και να αναγνωριστούν σωστά αντικείμενα από φωτογραφίες UAV. Οι αεροφωτογραφίες λήφθηκαν από μία κάμερα 12MP και από ύψος 120 μέτρων. Για τον σκοπό αυτό χρησιμοποιήθηκε drone DJI Mini Pro 3.

Πίνακας 8.1. Τρόπος συλλογής των αεροφωτογραφιών από το «Εργαστήριο Ψηφιακής Επεξεργασίας Σήματος και Εικόνας» του Πολυτεχνείου Κρήτης.

Drone	DJI Mini Pro 3
Κάμερα	12MP
Ύψος λήψης	120 μέτρα
Γωνία λήψης	Κατακόρυφη

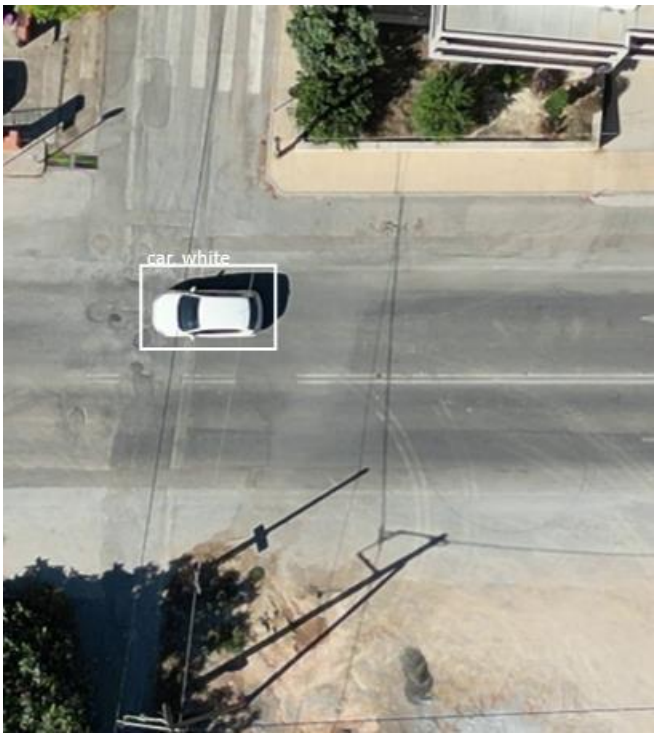
Παρακάτω παρουσιάζονται δύο τέτοια παραδείγματα εφαρμογής του αλγορίθμου και εντοπισμού του χρώματος τους (Εικόνες 8.9 και 8.10).



Εικόνα 8.9. Κατακόρυφη λήψη αστικής περιοχής. Αεροφωτογραφία από UAV του εργαστηρίου «Εργαστήριο Ψηφιακής Επεξεργασίας Σήματος και Εικόνας» του Πολυτεχνείου Κρήτης. Αναγνώριση αντικειμένων (πεζών) – τα αναγνωρισμένα αντικείμενα (πεζοί) διακρίνονται με μπλέ πλαίσια – και εντοπισμός του χρώματος τους.

Χρησιμοποιώντας το YOLOv5x, αρχικά αναγνωρίζονται και εντοπίζονται τα αντικείμενα ενδιαφέροντος (πεζοί), τα οποία εμφανίζονται με μπλε πλαίσια (bounding boxes). Στη συνέχεια, απομονώνεται το υπόβαθρο εντός των ορίων (bounding boxes) των αναγνωρισμένων αντικειμένων, ώστε να διευκολυνθεί η περαιτέρω ανάλυση. Τέλος, η διαδικασία ολοκληρώνεται με επιτυχή ανίχνευση και αναγνώριση του χρώματος των εντοπισμένων πεζών.

Στο δεύτερο παράδειγμα (Εικόνα 8.10), μπορούμε να δούμε τον επιτυχή εντοπισμό και την αναγνώριση ενός αυτοκινήτου σε εικόνα από το ίδιο dataset, επιβεβαιώνοντας την ικανότητα του μοντέλου να ανιχνεύει διαφορετικές κατηγορίες αντικειμένων με ακρίβεια.



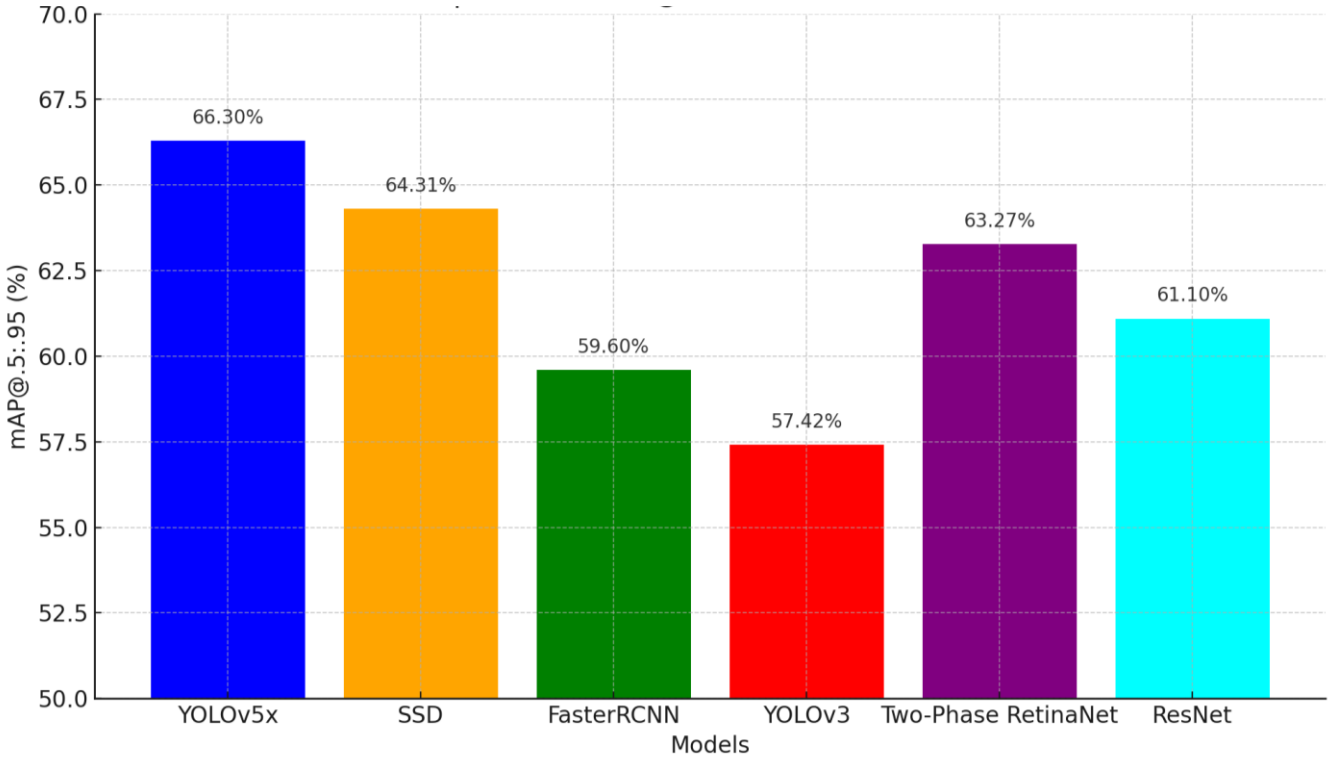
Εικόνα 8.10. Αναγνώριση αντικειμένων (αυτοκίνητο) και εντοπισμός χρώματος σε φωτογραφία από UAV του εργαστηρίου «Εργαστήριο Ψηφιακής Επεξεργασίας Σήματος και Εικόνας» του Πολυτεχνείου Κρήτης.

Όπως μπορεί να παρατηρήσει κανείς από τις παραπάνω αεροφωτογραφίες, το σύστημα εντόπισε και απομόνωσε με επιτυχία τους δύο πεζούς που φαίνονται στην πρώτη περίπτωση καθώς και το αυτοκίνητο στην εικόνα της δεύτερης περίπτωσης. Επιπλέον αναγνώρισε επιτυχώς το χρώμα τους, γεγονός που επιβεβαιώνει την ικανότητα που έχει να εντοπίζει επιτυχώς το χρώμα των αντικειμένων.

Κεφάλαιο 9 Σύγκριση με Άλλα Μοντέλα

Η σύγκριση μεταξύ διαφορετικών αλγορίθμων ανίχνευσης αντικειμένων είναι απαραίτητη, καθώς παρέχει πολύτιμες πληροφορίες για την απόδοση και την καταλληλόλητα κάθε μοντέλου σε συγκεκριμένα δεδομένα. Ειδικά για εφαρμογές σε αεροφωτογραφίες, όπως αυτές του Stanford Drone Dataset, η ακρίβεια και η δυνατότητα γενίκευσης των μοντέλων είναι καθοριστικοί παράγοντες. Για τον λόγο αυτό, παρακάτω συγκρίνουμε το μοντέλο που αναπτύχθηκε κατά την εργασία αυτή με μοντέλα από την βιβλιογραφία, όπως τα YOLOv3, Faster R-CNN, SSD, Two-Phase RetinaNet και ResNet ώστε να διαμορφωθεί μια ολοκληρωμένη εικόνα για την απόδοση του αλγορίθμου.

Από το Πείραμα 1 αυτής της εργασίας, το αρχικό μοντέλο YOLOv5x απέδειξε την καλύτερη απόδοσή σε σύγκριση με τα υπόλοιπα μοντέλα, όπως αυτά αναφέρονται στις εργασίες [44], [45]. Συγκεκριμένα, πέτυχε mAP@.5:.95: 66.3%, τιμή σημαντικά υψηλότερη από τις αντίστοιχες τιμές των YOLOv3 (57.42%), Faster R-CNN (59.60%), SSD (64.31%), Two-Phase RetinaNet (63.27%) και ResNet (61.10%) – βλέπε Εικόνα 8.1 παρακάτω.



Εικόνα 9.1. Σύγκριση της απόδοσης (mAP@.5:.95) του μοντέλου YOLOv5x που αναπτύχθηκε σε σχέση με μοντέλα που αναφέρονται στη βιβλιογραφία όπως τα SSD, Faster R-CNN, ResNet, Two-Phase ResNet και YOLOv3.

Η παραπάνω σύγκριση υπογραμμίζει την ικανότητα του YOLOv5x να γενικεύει καλύτερα. Το YOLOv5x όχι μόνο υπερέχει σε ακρίβεια, αλλά καταφέρνει να προσαρμοστεί καλύτερα στις

απαιτήσεις της ανίχνευσης αντικειμένων σε πολύπλοκα δεδομένα, όπως αυτά που συλλέγονται από UAVs.

Κεφάλαιο 10 Συζήτηση και Συμπεράσματα

Η παρούσα εργασία επικεντρώθηκε στην ανάπτυξη και αξιολόγηση ενός συστήματος ανίχνευσης αντικειμένων και χαρακτηρισμού των ιδιοτήτων τους (βασισμένο στην αρχιτεκτονική YOLOv5) από οπτικά δεδομένα που συλλέγονται μέσω UAVs. Στο πλαίσιο της εργασίας, χρησιμοποιήθηκε το Stanford Drone Dataset, το οποίο περιλαμβάνει αντικείμενα ενδιαφέροντος (κλάσεις), όπως πεζούς (pedestrians), αυτοκίνητα (cars) και λεωφορεία (buses). Η εκπαίδευση και αξιολόγηση του YOLOv5 έγινε με τη χρήση διαφόρων μετρικών απόδοσης, όπως το mAP, η ακρίβεια (precision), η ανάκληση (recall) και η καμπύλη Precision-Recall. Επιπλέον, εφαρμόστηκαν τεχνικές data augmentation, όπως περιστροφή και αλλαγή κλίμακας, για τη βελτίωση της γενίκευσης του μοντέλου.

Η αρχική έκδοση του YOLOv5 αποδείχθηκε η πιο αξιόπιστη επιλογή συνολικά, παρουσιάζοντας τις υψηλότερες τιμές ακρίβειας (precision) και ανάκλησης (recall). Οι τροποποιήσεις που εισήχθησαν με τη χρήση του μηχανισμού SoftPool επιδίωξαν να βελτιώσουν την απόδοση μέσω καλύτερης γενίκευσης, αλλά τα αποτελέσματα έδειξαν ότι μόνο ο συνδυασμός SoftPool με το CoordAttention μπορεί να προσφέρει μια ουσιαστική βελτίωση. Συγκεκριμένα, το βασικό YOLOv5x πέτυχε απόδοση με mAP 66.3%, ξεπερνώντας παραλλαγές, όπως το YOLOv5 με SoftPool και Squeeze-and-Excitation (34%) και το YOLOv5 με SoftPool και CoordAttention (52.1%). Παρά τη χαμηλότερη απόδοση των παραλλαγών, ο συνδυασμός SoftPool με CoordAttention έδειξε δυνατότητες για συγκεκριμένες βελτιώσεις, ιδιαίτερα σε κατηγορίες όπως τα λεωφορεία. Τα αποτελέσματα ανέδειξαν επίσης την ικανότητα του YOLOv5x να ανιχνεύει αντικείμενα διαφορετικών μεγεθών με υψηλή ακρίβεια

Η σύγκριση του YOLOv5x με άλλα μοντέλα από τη βιβλιογραφία, όπως τα SSD, Faster R-CNN, RetinaNet και Two-Fold ResNet, ανέδειξε τη σημαντική υπεροχή του YOLOv5x όσον αφορά την ακρίβεια και την ταχύτητα του. Το YOLOv5x κατέγραψε υψηλότερες επιδόσεις, αποδεικνύοντας την αποτελεσματικότητά του σε εφαρμογές πραγματικού χρόνου.

Στο δεύτερο μέρος της εργασίας, ο εντοπισμός του χρώματος των αντικειμένων επιτεύχθηκε μέσω τεχνικών, όπως το GrabCut για την αφαίρεση του υποβάθρου, ακολουθούμενο από την αναγνώριση/εντοπισμό του χρώματος. Το σύστημα που αναπτύχθηκε, αξιολογήθηκε περαιτέρω σε διαφορετικό dataset, προερχόμενο από αεροφωτογραφίες που συλλέχθηκαν από UAVs/drones του Εργαστηρίου Ψηφιακής Επεξεργασίας Σήματος και Εικόνας του Πολυτεχνείου Κρήτης, επιδεικνύοντας τη δυνατότητα προσαρμογής του μοντέλου και της χρήσης του σε πραγματικά σενάρια.

Ωστόσο, η παρούσα εργασία είχε ορισμένους περιορισμούς. Η χρήση μόνο του Stanford Drone Dataset μπορεί να περιορίζει τη γενίκευση των αποτελεσμάτων σε πραγματικές συνθήκες. Η επέκταση του υπάρχοντος dataset, με δεδομένα από ποικίλα περιβάλλοντα και συνθήκες φωτισμού αποτελεί μια κρίσιμη βελτίωση που δυνητικά θα ενισχύσει τη γενίκευση του μοντέλου. Μελλοντικές εργασίες μπορούν να εστιάσουν στη χρήση πιο πρόσφατων μοντέλων, όπως το RT-DETR και YOLOv8, το οποίο εισάγει περαιτέρω βελτιώσεις σε ταχύτητα και ακρίβεια. Επιπλέον, η

ενσωμάτωση δεδομένων από ποικίλες πηγές και περιβάλλοντα, καθώς και η χρήση συνθετικών δεδομένων με τεχνικές όπως τα Generative Adversarial Networks (GANs), θα μπορούσε να ενισχύσει τη γενίκευση και την αποτελεσματικότητα του μοντέλου. Τέλος, η υιοθέτηση μηχανισμών, όπως το ECA (Efficient Channel Attention) μπορεί να βελτιώσει περαιτέρω την ανίχνευση κρίσιμων χαρακτηριστικών, επιτρέποντας στο μοντέλο να αποδώσει ακόμα καλύτερα σε εφαρμογές, όπως η επιτήρηση και η γεωργία ακριβείας.

Βιβλιογραφία

[1] "The difference between aerial and satellite imagery | Nearmap US." Accessed: Feb. 02, 2025. [Online]. Available: https://www.nearmap.com/blog/aerial-maps-versus-satellite-maps?utm_source=google&utm_medium=organic

[2] C. Gaither, "Types of Aerial Photography and Its Applications," Great Big Photography World. Accessed: Feb. 02, 2025. [Online]. Available: https://greatbigphotographyworld.com/types-of-aerial-photography/

[3] R. P. Gupta, *Remote Sensing Geology*, 3rd ed. 2018 edition. Berlin, Heidelberg: Springer, 2017.

[4] J. Ding *et al.*, "Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022, doi: 10.1109/TPAMI.2021.3117983.

[5] "What is a Drone? - Definition from WhatIs.com," IoT Agenda. Accessed: Jun. 12, 2024. [Online]. Available: https://www.techtarget.com/iotagenda/definition/drone

[6] W. Y. H. Adoni, S. Lorenz, J. S. Fareedh, R. Gloaguen, and M. Bussmann, "Investigation of Autonomous Multi-UAV Systems for Target Detection in Distributed Environment: Current Developments and Open Challenges," *Drones*, vol. 7, no. 4, Art. no. 4, Apr. 2023, doi: 10.3390/drones7040263.

[7] R. Boyle, "Introducing the Matternet, A Network of Drones For Deliveries In Remote Locations," Popular Science. Accessed: Jun. 12, 2024. [Online]. Available: https://www.popsci.com/technology/article/2011-08/introducing-matternet-quadcopter-network-deliveries-remote-locations/

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Feb. 2017, doi: 10.1145/3065386.

[9] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," vol. 9905, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Oct. 22, 2014, *arXiv*: arXiv:1311.2524. doi: 10.48550/arXiv.1311.2524.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Oct. 2020, doi: 10.1109/TPAMI.2018.2858826.

[13] "Introduction to Artificial Intelligence and Machine Learning," Community.aws. Accessed: Jun. 12, 2024. [Online]. Available: https://community.aws/content/2drbbXokwrIXivItJ8ZeCk3gT5F/introduction-to-artificial-intelligence-and-machine-learning

[14] T. M. Mitchell, *Machine Learning*, 1st edition. New York: McGraw-Hill Education, 1997.

[15] P. Simon, *Too Big to Ignore: The Business Case for Big Data*. Wiley, 2013.

[16] "A Guide to Convolutional Neural Networks — the ELI5 way | Saturn Cloud Blog." Accessed: Jun. 12, 2024. [Online]. Available: https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way

[17] M. Yani, M. T. B. I. S Si., and M. T. C. S. S.T., "Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail," *J. Phys. Conf. Ser.*, vol. 1201, no. 1, p. 012052, Feb. 2019, doi: 10.1088/1742-6596/1201/1/012052.

[18] "4. Fully Connected Deep Networks - TensorFlow for Deep Learning [Book]." Accessed: Jun. 12, 2024. [Online]. Available: https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html

[19] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," Nov. 08, 2018, *arXiv*: arXiv:1811.03378. doi: 10.48550/arXiv.1811.03378.

[20] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and Understanding Recurrent Networks," Nov. 16, 2015, *arXiv*: arXiv:1506.02078. doi: 10.48550/arXiv.1506.02078.

[21] G. E. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds., Berlin, Heidelberg: Springer, 2012, pp. 599–619. doi: 10.1007/978-3-642-35289-8_32.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[23] "ReLU vs. LeakyReLU vs. PReLU | Baeldung on Computer Science." Accessed: Jun. 12, 2024. [Online]. Available:https:/www.baeldung.com/cs/relu-vs-leakyrelu-vs-prelu

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence*

and Statistics, JMLR Workshop and Conference Proceedings, Mar. 2010, pp. 249–256. Accessed: Jun. 12, 2024. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a.html

[25] A. Sharma, "Introduction to the YOLO Family," PyImageSearch. Accessed: Jun. 12, 2024. [Online]. Available: https://pyimagesearch.com/2022/04/04/introduction-to-the-yolo-family/

[26] "Releases · ultralytics/yolov5," GitHub. Accessed: Jun. 12, 2024. [Online]. Available: https://github.com/ultralytics/yolov5/releases

[27] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, Art. no. 4, Dec. 2023, doi: 10.3390/make5040083.

[28] M. Jani, J. Fayyad, Y. Al-Younes, and H. Najjaran, "Model Compression Methods for YOLOv5: A Review," Jul. 21, 2023, *arXiv*: arXiv:2307.11904. doi: 10.48550/arXiv.2307.11904.

[29] J. Liu, Q. Cai, F. Zou, Y. Zhu, L. Liao, and F. Guo, "BiGA-YOLO: A Lightweight Object Detection Network Based on YOLOv5 for Autonomous Driving," *Electronics*, vol. 12, no. 12, Art. no. 12, Jan. 2023, doi: 10.3390/electronics12122745.

[30] Z. Li, J. Song, K. Qiao, C. Li, Y. Zhang, and Z. Li, "Research on efficient feature extraction: Improving YOLOv5 backbone for facial expression detection in live streaming scenes," *Front. Comput. Neurosci.*, vol. 16, Aug. 2022, doi: 10.3389/fncom.2022.980063.

[31] E. Reswara, S. Suakanto, and S. A. Putra, "Comparison of Object Detection Algorithm using YOLO vs Faster R-CNN : A Systematic Literature Review," in *Proceedings of the 2023 6th International Conference on Big Data Technologies*, in ICBDT '23. New York, NY, USA: Association for Computing Machinery, Sep. 2023, pp. 419–424. doi: 10.1145/3627377.3627443.

[32] S. Srivastava, A. V. Divekar, C. Anilkumar, I. Naik, V. Kulkarni, and V. Pattabiraman, "Comparative analysis of deep learning image detection algorithms," *J. Big Data*, vol. 8, no. 1, p. 66, May 2021, doi: 10.1186/s40537-021-00434-w.

[33] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, Eds., "Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes," *Comput. Vis. – ECCV 2016*, 2016, doi: 10.1007/978-3-319-46484-8_33.

[34] K. Βασίλη and K. Vasili, "Αναγνώριση αντικειμένων σε εναέρια υψηλής ανάλυσης δεδομένα βίντεο με συνελικτικά νευρωνικά δίκτυα," Mar. 2019, doi: 10.26240/heal.ntua.15630.

[35] R. Padilla, S. Netto, and E. da Silva, *A Survey on Performance Metrics for Object-Detection Algorithms*. 2020. doi: 10.1109/IWSSIP48289.2020.

[36] “train_test_split,” scikit-learn. Accessed: Jun. 12, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.model_selection.train_test_split.html

[37] A. Stergiou, R. Poppe, and G. Kalliatakis, “Refining activation downsampling with SoftPool,” presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Oct. 2021, pp. 10337–10346. doi: 10.1109/ICCV48922.2021.01019.

[38] Z. Li, “Road Aerial Object Detection Based on Improved YOLOv5,” *J. Phys. Conf. Ser.*, vol. 2171, no. 1, p. 012039, Jan. 2022, doi: 10.1088/1742-6596/2171/1/012039.

[39] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[40] Q. Hou, D. Zhou, and J. Feng, “Coordinate Attention for Efficient Mobile Network Design,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 13708–13717. doi: 10.1109/CVPR46437.2021.01350.

[41] C. Rother, V. Kolmogorov, and A. Blake, “‘GrabCut’: interactive foreground extraction using iterated graph cuts,” in *ACM SIGGRAPH 2004 Papers*, in SIGGRAPH ’04. New York, NY, USA: Association for Computing Machinery, Dec. 2004, pp. 309–314. doi: 10.1145/1186562.1015720.

[42] “2.1. Gaussian mixture models,” scikit-learn. Accessed: Jun. 12, 2024. [Online]. Available: https://scikit-learn/stable/modules/mixture.html

[43] “OpenCV: Interactive Foreground Extraction using GrabCut Algorithm.” Accessed: Jun. 12, 2024. [Online]. Available: https://docs.opencv.org/3.4/d8/d83/tutorial_py_grabcut.html

[44] M. Antonakakis, C. Trimas, and M. Zervakis, “A Two-Phase ResNet for Object Detection in Aerial Images,” in *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*, Copenhagen, Denmark: IEEE, Oct. 2023, pp. 1–5. doi: 10.1109/IST59124.2023.10355709.

[45] M. Maktab Dar Oghaz, M. Razaak, and P. Remagnino, “Enhanced Single Shot Small Object Detector for Aerial Imagery Using Super-Resolution, Feature Fusion and Deconvolution,” *Sensors*, vol. 22, no. 12, p. 4339, Jun. 2022, doi: 10.3390/s22124339.