ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΤΩΝ ΚΑΤΕΡΓΑΣΙΩΝ

Αναγνώριση σκηνής και ανίχνευση προσωπικού και μηχανολογικού εξοπλισμού με χρήση συνθετικών εικόνων από μοντέλα διάχυσης

Χρήστος Λύτρας

Διπλωματική Εργασία

Επιβλέπων:

Πανώριος Μπενάρδος, Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Μηχανολόγων Μηχανικών
Τομέας Τεχνολογίας των Κατεργασιών

Αναγνώριση σκηνής και ανίχνευση προσωπικού και μηχανολογικού εξοπλισμού με χρήση συνθετικών εικόνων από μοντέλα διάχυσης

Χρήστος Λύτρας

Διπλωματική Εργασία

Επιβλέπων:

Πανώριος Μπενάρδος, Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Οκτωβρίου, 2023.

Πανώριος Μπενάρδος	Γ.–Χ. Βοσνιάκος	Δημήτριος Ναθαναήλ
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτης. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

Περίληψη

Το ολοένα και αυξανόμενο ενδιαφέρον για αποτελεσματικές μεθόδους ανίχνευσης αντικειμένων στο επίπεδο παραγωγής έχει ωθήσει την αναζήτηση καινοτόμων λύσεων για την αντιμετώπιση της έλλειψης διαθέσιμων στο κοινό συνόλων δεδομένων σχετικών με αυτόν τον τομέα. Η παρούσα διπλωματική εργασία ερευνά εκτενώς τη χρήση καινοτόμων μοντέλων διάχυσης κειμένου-σε-εικόνα για τη παραγωγή συνθετικών δεδομένων, για να αντισταθμίσει την απουσία πραγματικών δεδομένων, με σκοπό την εκπαίδευση μοντέλων ανίχνευσης αντικειμένων στο επίπεδο παραγωγής.

Η μελέτη αντλεί από τους τομείς της μηχανικής μάθησης, της μηχανικής όρασης και της μηχανολογίας για την ανάπτυξη μιας σύνθετης προσέγγισης δημιουργίας συνθετικού συνόλου δεδομένων. Η διαδικασία περιλαμβάνει την τριδιάστατη μοντελοποίηση των υπό μελέτη αντικειμένων, την εισαγωγή μεταβλητότητας μέσω της τυχαιοποίησης των παραμέτρων μιας εικονικής σκηνής και την επακόλουθη σύνθεση αληθοφανών εικόνων μέσω του Stable Diffusion, ενός μοντέλου βαθιάς μάθησης τελευταίας τεχνολογίας που εισήχθη το 2022, το οποίο αξιοποιεί μεθόδους διάχυσης στον λανθάνοντα χώρο για τη σύνθεση αληθοφανών εικόνων βάσει περιγραφών κειμένου. Το ControlNet χρησιμοποιείται για τον περαιτέρω έλεγχο της σύνθεσης των εικόνων, χρησιμοποιώντας εικόνες βάθους της εικονικής σκηνής ως συνθήκες για την απόδοση εικόνων που ακολουθούν πιστά τα χαρακτηριστικά των αντικειμένων και της εικονικής σκηνής. Το παραγόμενο συνθετικό σύνολο δεδομένων χαρακτηρίζεται από υψηλό βαθμό μεταβλητότητας και αποτελείται από συντεθειμένες εικόνες που περιλαμβάνουν μηχανολογικό εξοπλισμό μαζί με το εργαζόμενο προσωπικό. Επιπλέον, συλλέγεται ένα μικρό σύνολο δεδομένων από αληθινές εικόνες των υπό μελέτη αντικειμένων.

Εκπαιδεύονται τρία μοντέλα ανίχνευσης αντικειμένων τύπου YOLO (You Only Look Once) αντίστοιχα με συνθετικά δεδομένα, με αληθινά δεδομένα και με συνδυασμό των δύο και στη συνέχεια αξιολογούνται ως προς ένα σύνολο επικύρωσης που αποτελείται από αληθινές εικόνες. Μέσω της σύγκρισης της επίδοσης των μοντέλων, εξετάζεται η αποτελεσματικότητα της προτεινόμενης προσέγγισης ως προς την ικανότητα ανίχνευσης αντικειμένων σε πραγματικές συνθήκες. Το μοντέλο που εκπαιδεύτηκε αρχικά με συνθετικά δεδομένα και έπειτα προσαρμόστηκε με αληθινές εικόνες επιτυγχάνει την υψηλότερη επίδοση, εμφανίζοντας αύξηση 3.1% στη συνολική μέση ακρίβειας – mAP σε σχέση με το μοντέλο που εκπαιδεύτηκε αποκλειστικά με αληθινά δεδομένα.

Λέξεις-κλειδιά — Ανίχνευση Αντικειμένων, Συνθετικό Σύνολο Δεδομένων, Μοντέλα Διάχυσης, Κείμενο-σε-Εικόνα, Έλεγχος υπό Όρους, Προσαρμογή Τομέα, Μεταφορά Μάθησης, Εικόνες Βάθους

Abstract

The increasing interest in effective object detection methods in manufacturing floors has prompted the exploration of innovative solutions to address the scarcity of publicly available datasets relevant to this domain. This thesis presents a comprehensive investigation into the utilization of novel text-to-image diffusion models for synthetic data generation, to mitigate the absence of real-world data for training models for object detection in the production floor.

This study draws from the domains of computer vision, machine learning, and mechanical engineering to develop a multifaceted approach for the generation of a synthetic dataset. The approach includes the 3D modeling of the objects of interest, the introduction of variability through the randomization of parameters of a virtual scene, and the subsequent generation of photorealistic images using Stable Diffusion, a cutting-edge deep learning model introduced in 2022 which leverages latent diffusion techniques to synthesize photorealistic images based on textual descriptions. The image generation process is fine tuned by ControlNet, utilizing previously rendered depth maps as conditions to yield images closely aligned with visual cues of the objects and the virtual scene. The resulting dataset has a high degree of variability and is composed of collated images which include production equipment and personnel simultaneously. Additionally, a small dataset consisting of real images of the objects of interest is collected.

Three YOLO (You Only Look Once) type object detection models are trained respectively on synthetic data, real data and a combination of the two, and subsequently evaluated on a validation set consisting of real-world images. By comparing the performance of each model, the efficacy of the proposed approach for object detection on real world conditions is examined. The model which was pretrained on synthetic data and later fine-tuned with real images achieves the highest performance, showing a 3.1% increase in mean average precision – mAP in comparison to the model trained exclusively on real data.

Keywords — Object Detection, Synthetic Dataset, Diffusion Models, Text-to-Image, Conditional Control, Domain Adaptation, Transfer Learning, Depth Map

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Επίκουρο Καθηγητή ΕΜΠ κ. Πανώριο Μπενάρδο, υπό την επίβλεψη του οποίου εκπονήθηκε η παρούσα έρευνα, για την ευκαιρία που μου προσέφερε να ασχοληθώ με το συγκεκριμένο αντικείμενο καθώς και για την εμπιστοσύνη που μου έδειξε.

Ευχαριστώ από τα βάθη της καρδιάς μου την Αρετή, έναν υπέροχο άνθρωπο, η οποία έχει σταθεί δίπλα μου σε χαρούμενες και δύσκολες στιγμές και με εμπνέει να είμαι καλύτερος άνθρωπος.

Η εργασία αυτή είναι αφιερωμένη στην Νιαούρω, την Αθηνά, την Αλίκη, τη Βούλα, τον Μπάμπη, τον Φούλη, την Μπέλλα και τη Ρόξυ.

Εις μνήμην της Τιτίκας και του Φοίβου, που μας αποχαιρέτησαν πολύ νωρίς.

Χρήστος Λύτρας

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1. Αυτοματοποίηση στη Βιομηχανία

Η αυτοματοποίηση έχει καθοριστικό ρόλο στις σύγχρονες βιομηχανικές εφαρμογές και έχει επιφέρει ρηξικέλευθες εξελίξεις στον τρόπο με τον οποίο διεξάγονται οι διεργασίες της παραγωγής [1]. Εντός του επιπέδου παραγωγής συναντάται ένα μεγάλο εύρος δραστηριοτήτων που περιλαμβάνει διαδικασίες κατεργασιών και γραμμές συναρμολόγησης [2], διακίνηση φορτίων και εφοδιασμό [3], ασφάλεια και επιτήρηση [4]. Οι δραστηριότητες αυτές αποτελούν συλλογικά τα θεμέλια των βιομηχανικών λειτουργιών και διαμορφώνουν την αποδοτικότητα, την παραγωγικότητα και τα πρότυπα ασφάλειας σε ένα περιβάλλον παραγωγής.

Η ενσωμάτωση συστημάτων αυτοματοποίησης σε βιομηχανικές εφαρμογές καλείται να εκπληρώσει αρκετούς κρίσιμους στόχους. Αρχικά, στοχεύει στη βελτιστοποίηση διάφορων διεργασιών παραγωγής με σκοπό την ενίσχυση της παραγωγικότητας και τη βέλτιστη αξιοποίηση πόρων διατηρώντας παράλληλα αυστηρά πρότυπα ποιότητας [5]. Επιπλέον, η αυτοματοποίηση επικουρεί στην αύξηση της ασφάλειας στον χώρο εργασίας καθώς μειώνει την άμεση έκθεση των εργαζομένων σε πιθανώς επικίνδυνες καταστάσεις [6]. Τέλος, καθιστά δυνατή την ομαλή διεξαγωγή των λειτουργιών, τη μείωση των ανθρώπινων σφαλμάτων και την εκτέλεση διάφορων εργασιών σχετικές με τη βιομηχανία με μεγαλύτερη ακρίβεια [7].

Το κίνητρο για την ένταξη της αυτοματοποίησης πηγάζει από τα άμεσα πλεονεκτήματα που προσφέρει σχετικά με την ταχύτητα, την ακρίβεια και την σχέση κόστους-αποτελεσματικότητας [5]. Εξαιτίας αυτών, η διερεύνηση και ανάπτυξη συστημάτων αυτοματοποίησης βιομηχανικών εφαρμογών αποτελεί αναγκαία κατεύθυνση για την διατήρηση βιωσιμότητας και κυρίως ανταγωνιστικού πλεονεκτήματος στο σημερινό εμπόριο.

1.2. Μηχανική Όραση

Η μηχανική όραση είναι ένα πεδίο της τεχνητής νοημοσύνης το οποίο εστιάζει στην ανάπτυξη αλγορίθμων και μεθοδολογιών που επιτρέπουν στους υπολογιστές να εξάγουν σημαντικές πληροφορίες από οπτικά δεδομένα, όπως εικόνες και βίντεο. Ο διεπιστημονικός αυτός κλάδος έχει αναδειχθεί ως ένα κομβικό πεδίο στον τομέα της μηχανικής μάθησης, επιτρέποντας σε μηχανές να ερμηνεύουν και να κατανοούν οπτικές πληροφορίες αντίστοιχα με την ανθρώπινη όραση. Την τελευταία δεκαετία, έχουν υπάρξει αξιοσημείωτες εξελίξεις στη μηχανική όραση οι οποίες της έχουν επιτρέψει να μεταμορφώσει πολλούς τομείς, αυτοματοποιώντας πολύπλοκες εργασίες που εκτελούνταν κάποτε αποκλειστικά από ανθρώπους.

Η εξέλιξη της μηχανικής όρασης έχει ενισχυθεί σημαντικά από τις εξελίξεις στην τεχνολογία υλισμικού των υπολογιστών. Η αύξηση της επεξεργαστικής ισχύος την οποία έχει επιφέρει η άνοδος των μονάδων επεξεργασίας γραφικών ή graphics processing units (GPUs) και εξειδικευμένου υλικού όπως οι μονάδες επεξεργασίας τανυστών ή tensor processing units (TPUs), έχει οδηγήσει στην επιτάχυνση των υπολογισμών και στον χειρισμό μεγάλου όγκου δεδομένων. Η τεχνολογική αυτή πρόοδος έχει συμβάλει ουσιαστικά στις τρέχουσες δυνατότητες των συστημάτων μηχανικής όρασης, επιτρέποντας αυτά να εκτελούν περίπλοκες εργασίες με βελτιωμένη ταχύτητα και ακρίβεια.

Η μηχανική όραση έχει εκτεταμένο πεδίο εφαρμογής. Μερικά παραδείγματα των ποικίλων εφαρμογών της αποτελούν η αυτόνομη πλοήγηση οχημάτων σε πολύπλοκα αστικά περιβάλλοντα [8], η ανάλυση ιατρικών απεικονίσεων οδηγώντας σε πιο ακριβή διάγνωση [9], η ενίσχυση συστημάτων ασφαλείας

μέσω αναγνώρισης προσώπου [10] και η ανίχνευση αντικειμένων σε αποθήκες που βελτιστοποιεί τη διαχείριση τους [11]. Ωστόσο, ενώ οι πρόσφατες εξελίξεις στοχεύουν να καταστήσουν την ικανότητα των συστημάτων μηχανικής όρασης εφάμιλλη αυτής της ανθρώπινης όρασης, ορισμένοι παράγοντες περιορίζουν την αύξηση της απόδοσης τους.

Σε αντίθεση με την εγγενή ανθρώπινη ικανότητα γενίκευσης και προσαρμογής σε νέα δεδομένα, οι περισσότερες μέθοδοι μηχανικής όρασης δεν αποδίδουν αποτελεσματικά αντιμέτωπες με συνθήκες που διαφέρουν από εκείνες στις οποίες έχουν εκπαιδευτεί. Το ανθρώπινο μάτι είναι ικανό να προσαρμοστεί γρήγορα σε διαφορετικές συνθήκες φωτισμού, προοπτικές και παραμορφώσεις αντικειμένων [12], ενώ τα μοντέλα μηχανικής όρασης τείνουν να εμφανίζουν πτώση απόδοσης αντιμέτωπα με τέτοιες αλλαγές πλαισίου [13]. Τα μοντέλα αυτά εξειδικεύονται σε συγκεκριμένες εργασίες και βασίζονται σε δεδομένα προσαρμοσμένα στην εκάστοτε εργασία για εκπαίδευση, γεγονός που περιορίζει την απόδοσή τους σε ένα στενό εύρος το οποίο καθορίζεται από τη μεταβλητότητα αυτού του συνόλου δεδομένων.

1.3. Εφαρμογές Μεθόδων Μηχανικής Όρασης στην Παραγωγή

Η μηχανική όραση κατέχει εξέχοντα ρόλο στην βιομηχανική αυτοματοποίηση και συμβάλλει στην αύξηση της αποδοτικότητας της παραγωγής. Οι εφαρμογές της διευκολύνουν την αυτοματοποίηση διάφορων διαδικασιών που λαμβάνουν χώρα στο επίπεδο παραγωγής, ενισχύοντας, κατά συνέπεια, την αποτελεσματικότητα και την ακρίβεια τους.

Η μηχανική όραση χρησιμοποιείται στην παραγωγή για την αυτοματοποίηση και τη βελτιστοποίηση διαφόρων εργασιών. Μία από τις κύριες εφαρμογές της είναι ο ποιοτικός έλεγχος στις γραμμές παραγωγής [14] [15] [16]. Τα αυτόματα συστήματα οπτικής επιθεώρησης χρησιμοποιούν αλγορίθμους μηχανικής όρασης για να ανιχνεύσουν ελαττώματα ή ανωμαλίες στα προϊόντα σε πραγματικό χρόνο, διασφαλίζοντας ότι μόνο τα προϊόντα που πληρούν τα προκαθορισμένα κριτήρια ποιότητας μεταβαίνουν στην επόμενη φάση.

Επιπλέον, η μηχανική όραση συμβάλλει στην προγνωστική συντήρηση [17], μια σημαντική πτυχή της αποδοτικής παραγωγής. Τα συστήματα προγνωστικής συντήρησης που αξιοποιούν μεθόδους μηχανικής όρασης έχουν την ικανότητα να προβλέψουν αξιόπιστα πιθανές δυσλειτουργίες ή βλάβες σε μηχανήματα αναλύοντας δεδομένα από αισθητήρες και οπτικές εισόδους. Η προληπτική αυτή προσέγγιση επιτρέπει τον έγκαιρο προγραμματισμό συντήρησης με αποτέλεσμα τη μείωση των απροσδόκητων διαστημάτων διακοπής λειτουργίας και στην αύξηση της συνολικής παραγωγικότητας.

Τα συστήματα αυτοματοποίησης που ενσωματώνουν αλγορίθμους μηχανικής όρασης ενισχύουν τον τομέα της ασφάλειας του εργαζόμενου προσωπικού [6]. Τέτοια συστήματα παρακολουθούν τις δραστηριότητες των εργαζομένων και να ανιχνεύουν μη ασφαλείς συμπεριφορές ή καταστάσεις. Η εποψία σε πραγματικό χρόνο επιτρέπει την άμεση επέμβαση ή ειδοποίηση για την πρόληψη ατυχημάτων και την τήρηση των πρωτοκόλλων ασφαλείας, όπως η χρήση των κατάλληλων μέσων ατομικής προστασίας.

Τα πλεονεκτήματα της χρήσης της μηχανικής όρασης στην παραγωγή περιλαμβλανουν τη χαρακτηριστική ικανότητα των συστημάτων αυτών να επεξεργάζονται και να αναλύουν μεγάλους όγκους οπτικών δεδομένων γρήγορα και με ακρίβεια. Οι παραδοσιακές μέθοδοι συχνά υπολείπονται στον χειρισμό του τεράστιου όγκου και της πολυπλοκότητας των δεδομένων που παράγονται σε βιομηχανικά περιβάλλοντα. Η μηχανική όραση, εκτός από την ενίσχυση της αποτελεσματικότητας, συμβάλλει και στην ελαχιστοποίηση των ανθρώπινων σφαλμάτων [7], με αποτέλεσμα την επίτευξη υψηλότερου επίπεδου ακρίβειας και αξιοπιστίας στις διάφορες λειτουργίες.

Επιπλέον, τα συστήματα μηχανικής όρασης προσαρμόζονται σε νέα δεδομένα, βελτιώνοντας την απόδοσή τους με την πάροδο του χρόνου και επιτρέποντας την εξειδίκευσή τους μέσω αλγορίθμων μηχανικής μάθησης. Η προσαρμοστικότητα και η συνεχής βελτίωση έχουν ζωτική σημασία σε

δυναμικά περιβάλλοντα παραγωγής όπου οι εξελισσόμενες συνθήκες και απαιτήσεις απαιτούν ευέλικτες λύσεις.

Συμπεραίνεται ότι οι εφαρμογές μηχανικής όρασης στην παραγωγή παρέχουν απαραίτητη υποστήριξη για την αυτοματοποίηση, αντιμετωπίζοντας κρίσιμα προβλήματα όπως ο ποιοτικός έλεγχος, η συντήρηση και η ασφάλεια. Τα πλεονεκτήματα της έγκεινται στην αποτελεσματικότητα, την ακρίβεια, την προσαρμοστικότητα και την ικανότητα επεξεργασίας μεγάλου όγκου δεδομένων, και την καθιστούν εξαιρετικά χρήσιμο εργαλείο για τη σύγχρονη βιομηχανία.

1.3.1. Προκλήσεις

Η εφαρμογή ανίχνευσης αντικειμένων και άλλων τεχνικών μηχανικής όρασης σε βιομηχανικές εγκαταστάσεις στο επίπεδο παραγωγής αποτελεί ιδιαίτερη πρόκληση λόγω της πολύπλοκης και δυναμικής φύσης που επικρατεί σε τέτοια περιβάλλοντα. Οι εγκαταστάσεις αποτελούνται από μια ποικιλία μηχανημάτων, εξοπλισμού και προσωπικού, τα οποία συνίστανται σε ένα ξεχωριστό σύνολο εμποδίων που καλούνται να υπερνικήσουν τα μοντέλα μηχανικής όρασης. Μία από τις κύριες προκλήσεις είναι η ανάγκη για ακριβή και αξιόπιστη ανίχνευση διαφόρων αντικειμένων μέσα στο συγκεχυμένο και συνεχώς μεταβαλλόμενο περιβάλλον. Τα επίπεδα παραγωγής χαρακτηρίζονται συχνά από κακές συνθήκες φωτισμού, εμπόδια που κρύβουν τα αντικείμενα, αντανακλάσεις και αμέτρητους απρόβλεπτους παράγοντες που μπορούν να επηρεάσουν σημαντικά την απόδοση των αλγορίθμων ανίχνευσης αντικειμένων.

Κύρια πρόκληση για την εφαρμογή μεθόδων ανίχνευσης αντικειμένων αποτελεί η έλλειψη δημοσίως διαθέσιμων συνόλων δεδομένων που περιέχουν συγκεκριμένα αντικείμενα που υπάρχουν σε βιομηχανικά περιβάλλοντα. Η απουσία τέτοιων πολύτιμων και αναγκαίων δεδομένων δυσχεραίνει την προσπάθεια ανάπτυξης αξιόπιστων μοντέλων μηχανικής όρασης. Σε αντίθεση με τομείς όπως η αυτόνομη οδήγηση ή η αναγνώριση κοινών αντικειμένων, τα βιομηχανικά περιβάλλοντα είναι εξαιρετικά εξειδικευμένα και ενδέχεται να περιλαμβάνουν ιδιόκτητα μηχανήματα ή εξαρτήματα. Αυτό έχει ως αποτέλεσμα την έλλειψη ολοκληρωμένων και ποικίλων δεδομένων που αντιπροσωπεύουν με ακρίβεια τις συνθήκες που επικρατούν τα οποία μπορούν να χρησιμοποιηθούν για την εκπαίδευση μοντέλων ανίχνευσης αντικειμένων. Κατά συνέπεια, αυτή γίνεται δύσκολη, καθώς τα μοντέλα πρέπει να εκπαιδεύονται σε περιορισμένα ή ακόμα και συνθετικά δεδομένα που μπορεί να μην αποτυπώνουν την πλήρη πολυπλοκότητα των πραγματικών συνθηκών.

Η ποικιλομορφία των αντικειμένων εντός των εγκαταστάσεων παραγωγής συμβάλλει επίσης στην πολυπλοκότητα της εφαρμογής τεχνικών μηχανικής όρασης. Το εύρος των αντικειμένων που καλούνται να εντοπίσουν και να παρακολουθήσουν είναι ευρύ και εκτείνεται από περίπλοκα εξαρτήματα εργαλειομηχανών έως εξοπλισμό ασφαλείας και προσωπικό. Κάθε κατηγορία αντικειμένων αναμένεται να έχει ξεχωριστή εμφάνιση, σχήμα, μέγεθος και σημασιολογική σχέση με γειτονικά αντικείμενα. Η προσαρμογή ενός ενιαίου μοντέλου ανίχνευσης αντικειμένων για τον χειρισμό αυτού του ποικίλου πλήθους αντικειμένων απαιτεί την ανάπτυξη εξελιγμένων αλγορίθμων ικανών χειρισμού πολλαπλών κατηγοριών αντικειμένων με διάφορους βαθμούς ορατότητας.

Η ανάγκη λειτουργίας σε πραγματικό χρόνο των βιομηχανικών διαδικασιών προσθέτει ένα ακόμα επίπεδο πολυπλοκότητας. Οι διεργασίες παραγωγής απαιτούν συχνά στιγμιαίες αποφάσεις που βασίζονται στα αποτελέσματα των συστημάτων μηχανικής όρασης [15]. Τυχούσες καθυστερήσεις στον εντοπισμό κρίσιμων αντικειμένων ή ψευδώς θετικές ή αρνητικές ανιχνεύσεις μπορούν να οδηγήσουν σε διακοπές της γραμμής παραγωγής, κινδύνους ασφαλείας ή προβλήματα στον ποιοτικό έλεγχο των προϊόντων [14]. Προκύπτει η απαίτηση μοντέλων ανίχνευσης αντικειμένων τα οποία χαρακτηρίζονται από συνδυασμό υψηλής ακρίβειας και χαμηλών υπολογιστικών χρόνων που επιτρέπουν τη λειτουργία σε πραγματικό χρόνο.

Για την αντιμετώπιση αυτών των προκλήσεων, η συνεργασία μεταξύ μελών της βιομηχανίας, ερευνητών και προμηθευτών καθίσταται ζωτικής σημασίας. Οι συμφωνίες κοινής χρήσης δεδομένων

που σέβονται δικαιώματα ιδιοκτησίας μπορούν να επιτρέψουν τη δημιουργία πιο εξειδικευμένων συνόλων δεδομένων με μεγάλο βαθμό μεταβλητότητας για την εκπαίδευση και αξιολόγηση μοντέλων ανίχνευσης αντικειμένων προσαρμοσμένων σε βιομηχανικά περιβάλλοντα. Επιπλέον, η ανάπτυξη μεθόδων μεταφοράς μάθησης που θα επιτρέπουν σε μοντέλα που έχουν εκπαιδευτεί σε πιο γενικά σύνολα δεδομένων να προσαρμόζονται σε συγκεκριμένες βιομηχανικές συνθήκες αξιοποιώντας περιορισμένο αριθμό διαθέσιμων δεδομένων μπορεί να αποδειχθεί ανεκτίμητη.

Συμπερασματικά, η εφαρμογή ανίχνευσης αντικειμένων και άλλων μεθόδων μηχανικής όρασης σε βιομηχανικά περιβάλλοντα είναι μια σύνθετη πρόκληση με πολλά μέτωπα. Η ποικιλομορφία των μηχανημάτων, του εξοπλισμού και του προσωπικού που συναντώνται στο επίπεδο παραγωγής, σε συνδυασμό με την έλλειψη εξειδικευμένων δεδομένων για εκπαίδευση, απαιτεί καινοτόμες λύσεις που αντιστοιχούν στην πολυπλοκότητα του πραγματικού κόσμου. Η υπέρβαση αυτών των προκλήσεων απαιτεί έναν συνδυασμό ανάπτυξης προηγμένων αλγορίθμων, συμφωνιών για διάθεση δεδομένων και τεχνογνωσίας στον τομέα για να διασφαλιστεί η επιτυχής ανάπτυξη αξιόπιστων συστημάτων μηχανικής όρασης σε βιομηχανικά περιβάλλοντα.

1.4. Στόχοι Μελέτης

Η παρούσα διατριβή αποσκοπεί στην διερεύνηση και στην ανάπτυξη μοντέλου μηχανικής όρασης για την αναγνώριση σκηνής στο επίπεδο παραγωγής. Στόχο αποτελεί η αυτόματη ανίχνευση του εργαζόμενου προσωπικού και μηχανολογικού εξοπλισμού εντός εικόνων βιομηχανικών εγκαταστάσεων. Η τρέχουσα μελέτη θέτει τις βάσεις επίλυσης ενός μείζονος ζητήματος στην εφαρμογή μεθόδων μηχανικής όρασης στο επίπεδο παραγωγής, εστιάζοντας στην πρόκληση που θέτει η περιορισμένη διαθεσιμότητα συνόλων δεδομένων που περιλαμβάνουν εξοπλισμό παραγωγής όπως εργαλειομηχανές, εξοπλισμό μηχανουργείου κ.ά. Όπως αναλύθηκε παραπάνω, η απουσία τέτοιων συνόλων δεδομένων στέκεται εμπόδιο στην εκπαίδευση αξιόπιστων μοντέλων μηχανικής όρασης εξειδικευμένα σε εργασίες ανίχνευσης και αναγνώρισης αντικειμένων που βρίσκονται αποκλειστικά σε περιβάλλοντα παραγωγής. Για την παράκαμψη αυτού του περιορισμού, αναπτύχθηκε μια πρωτοποριακή προσέγγιση για τη δημιουργία ενός συνθετικού συνόλου δεδομένων, επιτρέποντας την εκπαίδευση μοντέλων μηχανικής όρασης με μια αληθοφανή αναπαράσταση των υπό μελέτη κατηγοριών αντικειμένων.

Η εκπόνηση της διπλωματικής εργασίας περιλαμβάνει τη δημιουργία ενός συνθετικού συνόλου δεδομένων που περιλαμβάνει πολλαπλές κατηγορίες αντικειμένων. Οι κατηγορίες αυτές είναι CNC τόρνοι, συμβατικοί τόρνοι, CNC φρέζες, συμβατικές φρέζες, περονοφόρα οχήματα, παλέτες και άνθρωποι. Επιπλέον, συλλέγεται ένα ξεχωριστό σύνολο δεδομένων, το οποίο αποτελείται από πραγματικές φωτογραφίες των ίδιων κατηγοριών αντικειμένων που έχουν ληγθεί σε διάφορες τοποθεσίες, συμπεριλαμβανομένων των εγκαταστάσεων του Εργαστηρίου Τεχνολογίας των Κατεργασιών (ΕΤΚ) του Εθνικού Μετσόβιου Πολυτεχνείου. Το συνθετικό και το αληθινό σύνολο δεδομένων χρησιμοποιούνται για την εκπαίδευση τριών πανομοιότυπων μοντέλων ανίχνευσης αντικειμένων της οικογένειας μοντέλων YOLO (You Only Look Once) [18]. Το πρώτο μοντέλο εκπαιδεύεται μόνο με τα συνθετικά δεδομένα, το δεύτερο μόνο με αληθινά ενώ το τρίτο εκπαιδεύεται σε πρώτη φάση με συνθετικά δεδομένα και στη συνέχεια προσαρμόζεται δεχόμενο επιπλέον εκπαίδευση με αληθινά δεδομένα. Τα τρία μοντέλα αξιολογούνται ως προς ένα σύνολο επικύρωσης που αποτελείται από αληθινές φωτογραφίες με σκοπό την εκτίμηση της επίδοσής τους σε αληθινές καταστάσεις. Οι επιδόσεις των μοντέλων συγκρίνονται και εξάγονται συμπεράσματα για την αποτελεσματικότητα της μεθόδου χρήσης συνθετικών δεδομένων.

1.5. Δομή της Διατριβής

Η διατριβή απαρτίζεται από έξι κεφάλαια. Το Κεφάλαιο 1 αποτελεί την παρούσα εισαγωγή στην οποία αναλύθηκαν τα εμπόδια στην εφαρμογή μεθόδων μηχανικής όρασης σε επίπεδα παραγωγής και παρουσιάστηκε η προτεινόμενη λύση και συνεισφορά της διατριβής στον κλάδο της μηχανικής όρασης. Στο Κεφάλαιο 2 γίνεται βιβλιογραφική ανασκόπηση μεθόδων μηχανικής όρασης και παρουσιάζονται τα μοντέλα που αξιοποιούνται για την υλοποίηση της σύνθεσης του συνόλου δεδομένων και το μοντέλο ανίχνευσης αντικειμένων που επιλέχθηκε. Στο Κεφάλαιο 3 αναλύονται ενδελεχώς η διαδικασία δημιουργίας του συνθετικού συνόλου δεδομένων και διάφορες προκαταρκτικές προσεγγίσεις οι οποίες τελικά απορρίφθηκαν. Στο Κεφάλαιο 4 παρουσιάζεται η διαδικασία εκπαίδευσης του μοντέλου ανίχνευσης αντικειμένων και αναλύονται οι επιλογές υπερπαραμέτρων του δικτύου καθώς και άλλων σχετικών παραμέτρων εκπαίδευσης. Στο Κεφάλαιο 5 γίνεται επικύρωση του μοντέλου στο σύνολο που αποτελείται από αληθινές εικόνες και γίνεται ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων και των περιορισμών της προτεινόμενης προσέγγισης. Τέλος, το Κεφάλαιο 6 ολοκληρώνει την παρούσα μελέτη παρουσιάζοντας τα συμπεράσματα που προέκυψαν από την έρευνα που πραγματοποιήθηκε καθώς και πιθανές μελλοντικές κατευθύνσεις.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1. Μέθοδοι Ανίχνευσης Αντικειμένων

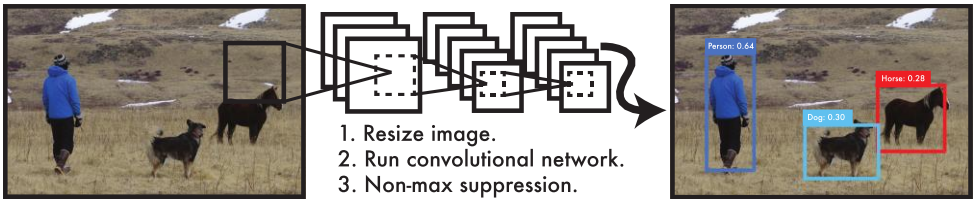
Η ανίχνευση αντικειμένων αποτελεί θεμελιώδη εφαρμογή της μηχανικής όρασης που περιλαμβάνει την αναγνώριση και την τοποθέτηση ορισμένων αντικειμένων ενδιαφέροντος εντός μιας εικόνας. Μέσα σε δεκαετίες έρευνας, έχουν αναπτυχθεί πολυάριθμες προσεγγίσεις για την αντιμετώπιση αυτής της πρόκλησης, η καθεμία με τα δικά της πλεονεκτήματα και περιορισμούς.

Παλαιότερα, τα συστήματα ανίχνευσης αντικειμένων προσάρμοζαν ταξινομητές για την εκτέλεση της ανίχνευσης. Τέτοια συστήματα χρησιμοποιούσαν έναν ταξινομητή εκπαιδευμένο να αναγνωρίζει κάποιο αντικείμενο και τον αξιολογούσαν σε διάφορες θέσεις και κλίμακες σε μια εικόνα με σκοπό τον εντοπισμό του συγκεκριμένου αντικειμένου εντός αυτής. Για παράδειγμα, συστήματα όπως τα μοντέλα παραμορφώσιμων μερών (deformable parts models - DPM) χρησιμοποιούν την προσέγγιση των ολισθαινόντων παράθυρων κατά την οποία ο ταξινομητής εκτελείται σε πολλές, ομοιόμορφα κατανεμημένες θέσεις σε ολόκληρη την εικόνα [19].

Επακόλουθες προσεγγίσεις όπως το συνελικτικό νευρωνικό δίκτυο περιοχής ή regional convolutional neural network (R-CNN) χρησιμοποιούν μεθόδους πρότασης περιοχής για να δημιουργήσουν σε πρώτο στάδιο πιθανά πλαίσια οριοθέτησης σε μια εικόνα και στη συνέχεια να εκτελέσουν έναν ταξινομητή σε αυτά τα προτεινόμενα πλαίσια. Μετά την ταξινόμηση, λαμβάνει χώρα μετεπεξεργασία για τη βελτίωση των πλαισίων οριοθέτησης, την εξάλειψη των πολλαπλών όμοιων ανιχνεύσεων και την επαναβαθμολόγηση των πλαισίων βάσει άλλων αντικειμένων εντός της εικόνας [20]. Αυτές οι περίπλοκες διαδικασίες χαρακτηρίζονται από μεγάλους χρόνους εκτέλεσης και δυσκολία βελτιστοποίησης, καθώς κάθε ξεχωριστό τμήμα πρέπει να εκπαιδευτεί ξεχωριστά

2.2. YOLO, You Only Look Once

Το You Only Look Once (YOLO) [18] είναι ένα καινοτόμο μοντέλο ανίχνευσης αντικειμένων το οποίο επέφερε σημαντική εξέλιξη σε αυτό το πεδίο της μηχανικής όρασης. Σε αντίθεση με τις παραδοσιακές μεθόδους που αναφέρθηκαν παραπάνω, η προσέγγιση του YOLO για την ανίχνευση αντικειμένων έχει θεμελιώδεις διαφορές στις οποίες οφείλονται η εντυπωσιακή ταχύτητα, ακρίβεια και αποτελεσματικότητά του. Συγκεκριμένα, επαναπλαισίωσε την ανίχνευση αντικειμένων ως απλό πρόβλημα παλινδρόμησης αντί για μια σειρά διακριτών εφαρμογών ταξινόμησης και εντοπισμού, απλοποιώντας σημαντικά τη διαδικασία ανίχνευσης. Το YOLO αξιοποιεί μια προσέγγιση ενός πάσου, όπου ένα μοναδικό νευρωνικό δίκτυο προβλέπει ταυτόχρονα την κατηγορία και την τοποθεσία κάθε αντικειμένου σε μια δεδομένη εικόνα, όπως φαίνεται στο Σχήμα 2.1.



Σχήμα 2.1. Το σύστημα ανίχνευσης του YOLO [18].

2.2.1. Ενοποίηση Διαδικασίας Ανίχνευσης

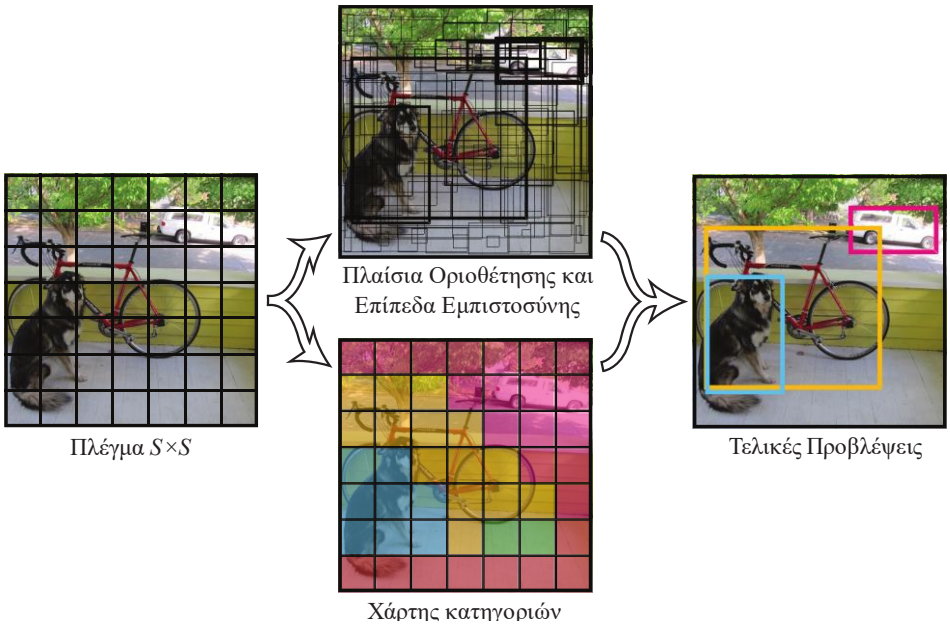
Η προσέγγιση του YOLO ενοποιεί τα διακριτά τμήματα της ανίχνευσης αντικειμένων σε μια ενιαία αρχιτεκτονική νευρωνικού δικτύου. Το δίκτυο επιτυγχάνει την πρόβλεψη κάθε πλαισίου οριοθέτησης αξιοποιώντας χαρακτηριστικά από όλη την εικόνα για όλα τα πλαίσια οριοθέτησης για όλες τις διάφορες κατηγορίες ταυτόχρονα. Η συγκεκριμένη προσέγγιση επιτρέπει στο δίκτυο να αναλύει καθολικά την εικόνα και τα αντικείμενα που εμπεριέχονται σε αυτή.

Το σύστημα υποδιαιρεί την εικόνα εισόδου σε ένα πλέγμα $S \times S$ κελιών, κάθε ένα από τα οποία είναι υπεύθυνο για την ανίχνευση οποιουδήποτε αντικειμένου του οποίου το κέντρο βρίσκεται εντός αυτού. Κάθε κελί προβλέπει B σε πλήθος πλαίσια οριοθέτησης και τα αντίστοιχα επίπεδα εμπιστοσύνης τους [18]. Τα επίπεδα αυτά χαρακτηρίζουν την εμπιστοσύνη του δικτύου όσον αφορά την ύπαρξη αντικειμένου εντός του πλαισίου και της ακρίβειας αυτού. Το επίπεδο εμπιστοσύνης ορίζεται ως

$$\text{Εμπιστοσύνη} = \text{P}(\text{Ύπαρξη Αντικειμένου}) \cdot \text{IOU}^{\text{αληθινό}}_{\text{πρόβλεψη}} \quad (1)$$

Στην περίπτωση όπου δεν υπάρχει κανένα αντικείμενο εντός του πλαισίου το επίπεδο εμπιστοσύνης οφείλει να μηδενίζεται. Εναλλακτικά, το επίπεδο εμπιστοσύνης οφείλει να ισούται με τον λόγο τομής–ένωσης (intersection over union - IOU) μεταξύ του προβλεπόμενου και του αληθινού πλαισίου οριοθέτησης.

Σε κάθε πλαίσιο οριοθέτησης αντιστοιχούν πέντε προβλέψεις: x, y, w, h και το επίπεδο εμπιστοσύνης. Το ζεύγος (x, y) αναπαριστά τις συντεταγμένες του κέντρου του πλαισίου σχετικά με το κελί του πλέγματος ενώ το ζεύγος (w, h) αναπαριστά τα κανονικοποιημένα ως προς τις διαστάσεις της εικόνας πλάτος και ύψος, αντίστοιχα. Επιπλέον, το επίπεδο εμπιστοσύνης αναπαριστά το IOU μεταξύ του προβλεπόμενου και του αληθινού πλαισίου οριοθέτησης. Τέλος, για κάθε κελί του πλέγματος προβλέπονται C σε πλήθος πιθανότητες, μια για κάθε κατηγορία αντικειμένου, ανεξάρτητα του πλήθους B πλαισίων. Οι δεσμευμένες αυτές πιθανότητες είναι δεδομένες της ύπαρξης αντικειμένου εντός του κελιού, δηλαδή $P(\textit{Κατηγορία}_i|\textit{Ύπαρξη Αντικειμένου})$. Όλες οι προβλέψεις εμπεριέχονται εντός ενός τανυστή διαστάσεων $S \times S \times (5 \cdot B + C)$. Η διαδικασία ανίχνευσης του μοντέλου απεικονίζεται στο Σχήμα 2.2.



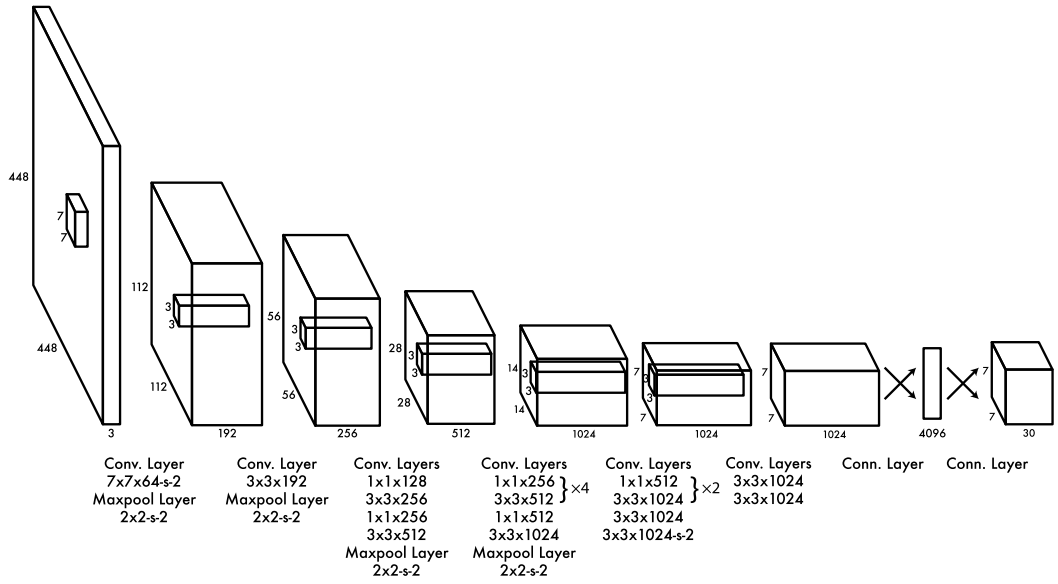
Σχήμα 2.2. Η διαδικασία ανίχνευσης αντικειμένων του YOLO μέσω υποδιαίρεσης της εικόνας εισόδου σε πλέγμα $S \times S$ [18].

Κατά την εκτέλεση της ανίχνευσης, υπολογίζονται ξεχωριστά επίπεδα εμπιστοσύνης για κάθε κατηγορία αντικειμένου εντός των οποίων εμπεριέχεται πληροφορία για την πιθανότητα της ύπαρξης της εκάστοτε κατηγορίας εντός του πλαισίου καθώς και για το πόσο καλά αυτό περιβάλλει το αντικείμενο. Τα συγκεκριμένα επίπεδα εμπιστοσύνης υπολογίζονται ως το γινόμενο της κάθε δεσμευμένης πιθανότητας με το επίπεδο εμπιστοσύνης κάθε πλαισίου οριοθέτησης, δηλαδή:

$$\begin{gathered} \text{P}(\text{Κατηγορία}_i|\text{Ύπαρξη Αντικειμένου}) \cdot \text{P}(\text{Ύπαρξη Αντικειμένου}) \cdot \text{IOU}_{\text{πρόβλεψη}}^{\text{αληθινό}} \\ = \\ \text{P}(\text{Κατηγορία}_i) \cdot \text{IOU}_{\text{πρόβλεψη}}^{\text{αληθινό}} \end{gathered} \tag{2}$$

2.2.2. Η Αρχιτεκτονική του Νευρωνικού Δικτύου

Η αρχιτεκτονική του μοντέλου YOLO έχει εμπνευστεί από τον ταξινομητή εικόνων GoogLeNet [21]. Το YOLO υλοποιείται μέσω ενός συνελικτικού νευρωνικού δικτύου (convolutional neural network – CNN) το οποίο αποτελείται από 24 συνελικτικά επίπεδα τα οποία ακολουθούν 2 πλήρως συνδεδεμένα επίπεδα. Η πλήρης αρχιτεκτονική του δικτύου αποτυπώνεται στο Σχήμα 2.3.



Σχήμα 2.3. Η αρχιτεκτονική του συνελικτικού νευρωνικού δικτύου του YOLO για RGB εικόνα εισόδου ανάλυσης 448×448 [18].

Το δίκτυο χρησιμοποιεί γραμμική συνάρτηση ενεργοποίησης για το τελευταίο επίπεδο, ενώ στα υπόλοιπα επίπεδα χρησιμοποιείται η συνάρτηση διαρρέουσας ανορθωμένης γραμμικής μονάδας (leaky rectified linear unit – leaky ReLU) η οποία δίνεται από τον τύπο

$$\varphi(x) = \begin{cases} x, & x \geq 0 \\ 0.1x, & x < 0 \end{cases} \tag{3}$$

2.2.3. Συνάρτηση Απώλειας

Όπως αναφέρθηκε προηγουμένως, η πρωτοποριακή προσέγγιση του YOLO οφείλεται στην ενοποίηση της διαδικασίας ανίχνευσης μέσω ενός ενιαίου νευρωνικού δικτύου. Η προσέγγισή της ως απλό πρόβλημα παλινδρόμησης επιτρέπει την καθολική εκπαίδευση του μοντέλου μέσω της ελαχιστοποίησης μιας και μόνο συνάρτησης απώλειας. Οι [18] χρησιμοποίησαν την μέθοδο των αθροισμάτων των τετραγώνων των σφαλμάτων για χάρη ευκολίας βελτιστοποίησης και έκαναν απαραίτητες τροποποιήσεις έτσι ώστε να εξασφαλίσουν τη μεγιστοποίηση της μέσης ακρίβειας.

Η συνάρτηση απώλειας αποτελείται από επιμέρους όρους που σχετίζονται με το μέγεθος και τη θέση των πλαισίων οριοθέτησης και με την ύπαρξη ή μη αντικειμένων εντός αυτών. Κάθε όρος διατυπώνεται μαθηματικά με έκφραση που δίνεται στην εξίσωση (4).

Για την αντιστάθμιση της δυσανάλογης επιρροής του όρου σφάλματος εμπιστοσύνης για μη ύπαρξη αντικειμένου στη συνάρτηση απώλειας, η οποία προκαλεί αστάθεια στο μοντέλο κατά την εκπαίδευση καθώς δυσχεραίνει τη σύγκλισής της, οι όροι πολλαπλασιάζονται με εμπειρικά επιλεγμένα βάρη. Κατά συνέπεια, δίνεται μεγαλύτερη έμφαση στο σφάλμα που αφορά το σχήμα και τη θέση των πλαισίων οριοθέτησης, με αποτέλεσμα την ακριβέστερη ανίχνευση των αντικειμένων. Τελικά, η συνάρτηση απώλειας δίνεται από τη σχέση:

$$
\begin{aligned}
\mathcal{L} = \lambda_{συντεταγμενων} & \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{αντικ.} \left[(x_i - \hat{x}_i)^2 + \left(y_i - \hat{y}_i\right)^2 \right] \\
& + \lambda_{συντεταγμενων} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{αντικ.} \left[\left(\sqrt{w_i} - \sqrt{\widehat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\widehat{h}_i}\right)^2 \right] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{αντικ.} \left(C_i - \widehat{C}_i\right)^2 + \lambda_{μη\ υπαρξη} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{οχι\ αντικ.} \left(C_i - \widehat{C}_i\right)^2 \\
& + \sum_{i=0}^{S^2} \mathbb{1}_{i}^{αντικ.} \sum_{c \in κατηγορίες} \left(p_i(c) - \hat{p}_i(c)\right)^2
\end{aligned}
\tag{4}
$$

όπου ο συντελεστής $\mathbb{1}_{i}^{αντικ.}$ υποδηλώνει την ύπαρξη αντικειμένου εντός του κελιού i. Οι συντελεστές $\mathbb{1}_{ij}^{αντικ.}$ και $\mathbb{1}_{ij}^{οχι\ αντικ.}$ υποδηλώνουν την αντίστοιχη ύπαρξη και μη ύπαρξη αντικειμένου εντός του κελιού i δεδομένου του j-οστού πλαισίου οριοθέτησης που είναι υπεύθυνο για την ανίχνευση του, καθώς έχει το μεγαλύτερο IOU με το αληθινό πλαίσιο από ότι οι υπόλοιπες προβλέψεις. Οι συντελεστές ισούνται με μονάδα όταν ισχύει η αντίστοιχη συνθήκη ειδάλλως μηδενίζονται. Συνεπώς, η συνάρτηση απώλειας επιβάλλει ποινή για λάθος ταξινόμηση μόνο στην περίπτωση ύπαρξης αντικειμένου εντός του κελιού, ενώ επιβάλλει ποινή για σφάλμα της θέσης πλαισίου μόνο για το καταλληλότερο υποψήφιο πλαίσιο, αποφεύγοντας έτσι την εισαγωγή πλεονάζουσων απωλειών.

2.2.4. Εξέλιξη του YOLO

Από τη δημοσίευση της πρώτης έκδοσης του YOLO το 2015, η αρχιτεκτονική του μοντέλου έχει δεχθεί πολλές αλλαγές με σκοπό τη βελτίωση της ακρίβειας, της ταχύτητας και της ευρωστίας του. Την πρώτη έκδοση του YOLO ακολούθησαν οι YOLOv2, YOLOv3 και YOLOv4 οι οποίες αντιμετώπισαν διάφορους περιορισμούς του αρχικού μοντέλου και εισήγαγαν νέες τεχνικές υλοποίησης ανίχνευσης αντικειμένων σε πραγματικό χρόνο.

Το YOLOv2, γνωστό και ως YOLO9000, σημείωσε αξιοσημείωτη πρόοδο σε σχέση με τον προκάτοχό του, εισάγοντας αρκετές εύστοχες τροποποιήσεις [22]. Η αρχιτεκτονική του αντικατέστησε αυτή της πρώτης έκδοσης που βασιζόταν στο GoogLeNet με ένα προσαρμοσμένο δίκτυο ονόματι Darknet-19, το οποίο συνιστάται από 19 συνελικτικά επίπεδα και μείωσε σημαντικά τους απαιτούμενους υπολογισμούς, διατηρώντας ή ακόμα και ενισχύοντας την απόδοση σε ορισμένες περιπτώσεις [23]. Επιπλέον, το YOLOv2 εισήγαγε τα λεγόμενα πλαίσια αγκύρωσης (anchor boxes), επιτρέποντας στο μοντέλο να προβλέπει αντικείμενα διαφορετικών μεγεθών και αναλογιών με μεγαλύτερη ακρίβεια. Η εισαγωγή των πλαισίων αγκύρωσης επέτρεψε τον καλύτερο εντοπισμό αντικειμένων εντός των κελιών του πλέγματος, βελτιώνοντας τη συνολική ακρίβεια ανίχνευσης.

Το YOLOv3, βασιζόμενο στις καινοτομίες της δεύτερης έκδοσης, ενίσχυσε σταδιακά την απόδοση του μοντέλου [24]. Ένα από τα καθοριστικά χαρακτηριστικά της έκδοσης αυτής ήταν η εισαγωγή πολλαπλών συνελικτικών επιπέδων που αποσκοπούν στην εξαγωγή χαρακτηριστικών σε τρεις διαφορετικές κλίμακες. Ο μηχανισμός αυτός είναι παρόμοιος με τις λεγόμενες πυραμίδες χαρακτηριστικών (feature pyramids) [25] και επέτρεψε την ακόμη αποτελεσματικότερη ανίχνευση αντικειμένων σε διαφορετικά μεγέθη. Επιπλέον, η αρχιτεκτονική του μοντέλου εξελίχθηκε ξανά με τη χρήση ενός πιο περίπλοκου δικτύου, που ονομάστηκε Darknet-53 λόγω των 53 συνελικτικών επιπέδων του , το οποίο βελτίωσε περαιτέρω την εξαγωγή χαρακτηριστικών από την εικόνα.

Η τέταρτη έκδοση του YOLO δημοσιεύτηκε το 2020 και έφερε σημαντικές βελτιώσεις οι οποίες το κατέστησαν τεχνολογία αιχμής όσον αφορά την ανίχνευση αντικειμένων [26]. Η έκδοση αυτή ενσωμάτωσε μια πληθώρα καινοτομιών που στοχεύουν στη βελτίωση τόσο της ακρίβειας όσο και της απόδοτικότητας του μοντέλου. Μια αξιοσημείωτη βελτίωση ήταν η ενσωμάτωση του δικτύου CSPDarknet53 [27], η οποία βελτιστοποίησε την εξαγωγή χαρακτηριστικών χωρίζοντας το δίκτυο σε δύο κλάδους, μειώνοντας αποτελεσματικά την υπολογιστική πολυπλοκότητα. Επιπλέον, η υλοποίηση συνδέσεων Cross-Stage Partial (CSP) και η σύντηξη χαρακτηριστικών μέσω του PANet [28] εμπλούτισαν περαιτέρω την ικανότητα του μοντέλου να ερμηνεύει χαρακτηριστικά και σημασιολογικές σχέσεις εντός της εικόνας. Το YOLOv4 εισήγαγε επίσης τεχνικές επαύξησης εικόνας (image augmentation) [29] οι οποίες όχι μόνο επέκτειναν την ποικιλομορφία των δεδομένων εκπαίδευσης αλλά συνέβαλαν επίσης στην αύξηση της ευρωστίας του μοντέλου εκθέτοντάς το σε πιο ποικίλες διαμορφώσεις αντικειμένων. Ως αποτέλεσμα, το YOLOv4 έφερε σημαντική πρόοδο στην ακρίβεια ανίχνευσης αντικειμένων διατηρώντας την ικανότητα λειτουργίας σε πραγματικό χρόνο.

Εν κατακλείδι, η εξέλιξη της αρχιτεκτονικής YOLO από την πρώτη έως και την τέταρτη έκδοση του δείχνει μια συνεχή προσπάθεια για την επίτευξη ταχύτερης και ακριβέστερης ανίχνευσης αντικειμένων. Κάθε έκδοση έχει βασιστεί στις δυνατότητες των προκατόχων της, ενσωματώνοντας καινοτόμες τεχνικές που έχουν καταστήσει συλλογικά το YOLO ως ένα από τα πλέον βέλτιστα μοντέλα ανίχνευσης αντικειμένων σε πραγματικό χρόνο, με το YOLOv4 να αποτελεί τεχνολογία αιχμής στο πεδίο της μηχανικής όρασης.

2.3. Σύνθεση Εικόνας με Μεθόδους Μηχανικής Μάθησης

Η γένεση συνθετικών εικόνων πρόκειται για ένα πεδίο της μηχανικής όρασης το οποίο έχει συγκεντρώσει σημαντική προσοχή λόγω του ευρέος φάσματος εφαρμογών του, που εκτείνεται από την τέχνη [30] και την ψυχαγωγία έως τις ιατρικές απεικονίσεις [31] [32] [9]. Το έργο της δημιουργίας συνθετικών εικόνων περιλαμβάνει τη δημιουργία εικόνων με αλγοριθμικά μέσα, συχνά με στόχο την μίμηση των οπτικών χαρακτηριστικών των εικόνων του πραγματικού κόσμου. Η παραγωγή αληθοφανών εικόνων δύναται να ενισχύσει διάφορες τεχνικές μηχανικής όρασης, όπως η επαύξηση δεδομένων [33] [34], η μεταφορά στυλ (style transfer) [35] [36], ο εντοπισμός ανωμαλιών [37], ακόμη και η δημιουργία εντελώς νέου οπτικού περιεχομένου [38].

Η κύρια απαίτηση για τη δημιουργία συνθετικών εικόνων είναι η επινόηση αλγορίθμων που μπορούν να δημιουργήσουν αληθοφανείς εικόνες, διατηρώντας παράλληλα έναν βαθμό ελέγχου του περιεχομένου τους. Η επίτευξη αυτού του στόχου είναι μια πολύπλευρη πρόκληση που απαιτεί τη βαθιά κατανόηση τόσο των υποβοσκόντων δεδομένων μιας εικόνας όσο και των περίπλοκων στατιστικών μοτίβων που διέπουν την εμφάνισή της. Επιπλέον, η ζήτηση για αληθοφανείς συνθετικές εικόνες έχει οδηγήσει στην εξερεύνηση μιας ποικιλίας μεθόδων μηχανικής μάθησης, η καθεμία με τα δικά της πλεονεκτήματα και περιορισμούς.

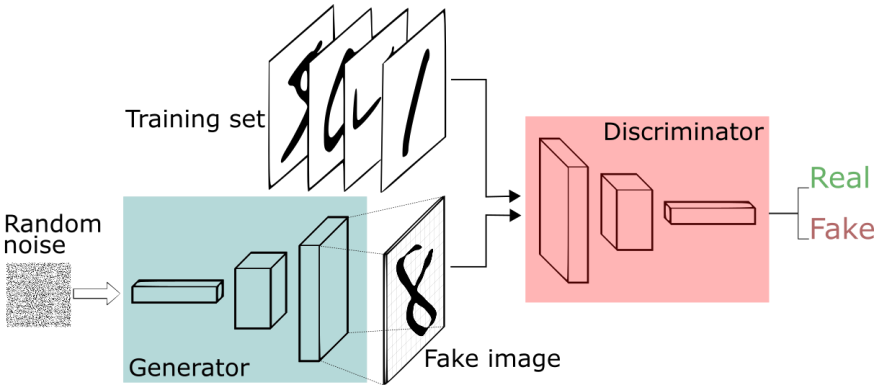
2.3.1. Παλαιότερες Μέθοδοι Σύνθεσης Εικόνας

Οι πρώιμες μέθοδοι για τη δημιουργία συνθετικών εικόνων βασίζονταν κυρίως σε διαδικαστικές προσεγγίσεις που χρησιμοποιούσαν σαφείς κανόνες και αλγορίθμους για τον καθορισμό της εμφάνισης διαφορετικών στοιχείων εντός μιας εικόνας. Μολονότι αυτές οι μέθοδοι παρείχαν ένα βασικό πλαίσιο για τη σύνθεση εικόνων, τα αποτελέσματα που παρήγαγαν συχνά δεν είχαν την ποικιλομορφία και τον πλούτο που απαιτούνταν για να μιμηθούν την πολυπλοκότητα που χαρακτηρίζει τις σκηνές του πραγματικού κόσμου.

Η έλευση της βαθιάς μάθησης επέφερε ρηξικέλευθες εξελίξεις στη γένεση συνθετικών εικόνων [39]. Τα συνελικτικά νευρωνικά δίκτυα, που αναπτύχθηκαν αρχικά για εφαρμογές ταξινόμησης ή ανίχνευσης αντικειμένων σε εικόνες, προσαρμόστηκαν έτσι ώστε να δημιουργιούν εικόνες. Οι μεταβλητοί αυτοκωδικοποιητές (Variational Autoencoders - VAEs) [40] αναδείχθηκαν ως μια σημαντική εξέλιξη, συνδυάζοντας βαθιά νευρωνικά δίκτυα με πιθανοτική μοντελοποίηση για τη δημιουργία εικόνων που τηρούν χαρακτηριστικά των δεδομένων εκπαίδευσης. Οι VAE καινοτόμησαν στη δημιουργία ποικιλόμορφων εικόνων και επέτρεψαν την ελεγχόμενη εξερεύνηση του λανθάνοντος χώρου. Στο Σχήμα 2.4 φαίνονται μερικά παραδείγματα σύνθεσης εικόνων ανθρώπινων προσώπων με χρήση μεταβλητών αυτοκωδικοποιητών.

Σχήμα 2.4. Συνθετικές εικόνες ανθρώπων παραγόμενες με χρησή μεταβλητού αυτοκωδικοποιητή [41].

Οι μέθοδοι για τη δημιουργία συνθετικών εικόνων συνέχισαν να εξελίσσονται με την έλευση των παραγωγικών αντιπαραθετικών δικτύων (Generative Adversarial Networks – GANs) το 2014 [42]. Τα GAN απέκτησαν τεράστια δημοτικότητα για την ικανότητά τους να παράγουν άκρως αληθοφανείς εικόνες προσεγγίζοντας την διαδικασία παραγωγής εικόνων ως ένα παίγνιο αντίπαλων παικτών, ενός παραγωγικού και ενός διαχωριστικού δικτύου. Η αντιπαράθεση μεταξύ αυτών των δικτύων οδηγεί το μεν να παράγει όλο και πιο αυθεντικές εικόνες και το δε να διακρίνει τις αληθινές και τις παραγόμενες εικόνες με όλο και μεγαλύτερη ακρίβεια. Το Σχήμα 2.5 αποτυπώνει την αρχιτεκτονική ενός παραγωγκού αντιπαραθετικού δικτύου.



Σχήμα 2.5. Η δομή ενός παραγωγικού αντιπαραθετικού δικτύου [43].

Επακόλουθες εξελίξεις ενσωμάτωσαν μηχανισμούς προσοχής, ιεραρχικές δομές και μηχανισμούς αυτοπροσοχής για να ενισχύσουν τη συνοχή και τις λεπτομέρειες των παραγόμενων εικόνων. Ένα αξιοσημείωτο παράδειγμα αποτελεί το StyleGAN [44] το οποίο έδειξε ικανότητα σύνθεσης εικόνων με καταπληκτική πιστότητα και έλεγχο συγκεκριμένων οπτικών χαρακτηριστικών, όπως το στυλ και η ανάλυση. Το Σχήμα 2.6 αποτυπώνει την ποιότητα των εικόνων που παράγει το StyleGAN2 [45].

Σχήμα 2.6. Συνθετικές εικόνες γατών παραγόμενες από το StyleGAN2 (πηγή: https://thesecatsdonotexist.com/).

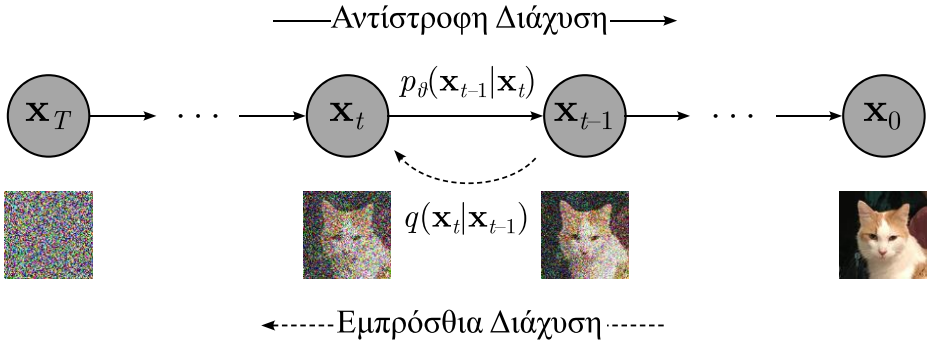
Ωστόσο, παρά τις σημαντικές προόδους στη δημιουργία συνθετικών εικόνων, υπάρχουν αρκετές προκλήσεις. Η εξασφάλιση ποικιλομορφίας και ελέγχου του παραγόμενου περιεχομένου, η αντιμετώπιση του λεγόμενου mode collapse, το φαινόμενο όπου ένα GAN παράγει περιορισμένη

ποικιλία εικόνων, και η δυσκολία σύγκλισης της εκπαίδευσης των μοντέλων είναι μεταξύ των σημαντικών ζητημάτων που συνεχίζουν να αντιμετωπίζουν οι ερευνητές.

2.3.2. Σύνθεση Εικόνας με Μοντέλα Διάχυσης

Τα μοντέλα διάχυσης [46] αντιπροσωπεύουν ένα υποσύνολο παραγωγικών μοντέλων που χαρακτηρίζονται από την ιδιαίτερη προσέγγισή τους στη γένεση και την ανακατασκευή δεδομένων. Σε αντίθεση με τους προκατόχους τους, που συχνά βασίζονται στην αντιπαραθετική μάθηση ή πιθανοτικές μεθόδους, τα μοντέλα διάχυσης καινοτομούν εισάγοντας μια διαδικασία εμπρόσθιας και αντίστροφης διάχυσης. Η προσέγγιση αυτή, αν και εννοιολογικά απλή, προσφέρει αξιοσημείωτη ευελιξία και αποτελεσματικότητα στη σύνθεση δεδομένων.

Η θεμελιώδης αρχή των μοντέλων διάχυσης βασίζεται στην σταδιακή προσθήκη θορύβου σε αρχικά δεδομένα, αλλοιώνοντας την πληροφορία που φέρουν ολοένα και περισσότερο σε κάθε βήμα. Η εμπρόσθια διαδικασία διάχυσης συνεχίζεται έως ότου τα δεδομένα να συγκλίνουν ασυμπτωτικά σε καθαρό θόρυβο Gauss. Η εκπαίδευση ενός μοντέλου διάχυσης περιλαμβάνει την εκμάθηση της ικανότητας αντιστροφής αυτής της διαδικασίας, δηλαδή την ελεγχόμενη αποθορυβοποίηση των δεδομένων μέσω αυτοκωδικοποιητών (autoencoders). Η εφαρμογή της αντίστροφης διάχυσης στην ουσία παράγει νεα δεδομένα από θορυβώδεις εισόδους που τηρούν τα μοτίβα που κωδικοποιούνται εντός των παραμέτρων των μοντέλων κατά τη διάρκεια της εκπαίδευσης. Ο μηχανισμός αυτός, ο οποίος φαίνεται στο Σχήμα 2.7, αποτελεί μια αλυσίδα Markov με κάθε βήμα αυτής να οδηγεί στην παραγωγή ενός δείγματος από μια επιθυμητή κατανομή [46].



Σχήμα 2.7. Βασική αρχή λειτουργίας των μοντέλων διάχυσης.

Τα μοντέλα διάχυσης υπερτερούν σε πολλές εργασίες, γεγονός το οποίο έχει οδηγήσει στην ταχεία άνοδό τους στον τομέα της παραγωγής εικόνων [47]. Η προσέγγιση τους διαφέρει από αυτή των GAN, εξαλείφοντας την ανάγκη για αντιπαραθετική μάθηση και συνεπώς παρακάμπτοντας την πολυπλοκότητα που αυτή επιφέρει. Αντίθετα, τα μοντέλα διάχυσης προσφέρουν μια βελτιωμένη και πιο αποδοτική μέθοδο εκπαίδευσης η οποία δέχεται παραλληλοποίηση και επεκτείνεται αποτελεσματικά. Η αξιοποίηση υπολογιστικών πόρων μέσω παραλληλοποίησης τους επιτρέπει να χειριστούν μεγάλους όγκους δεδομένων πολύ αποδοτικά. Η απλή αρχιτεκτονική και μεθοδολογία εκπαίδευσής τους τα καθιστούν εξαιρετικά προσαρμόσιμα σε μια ποικιλία εφαρμογών, τόσο εντός όσο και εκτός της μηχανικής όρασης.

Την τελευταία διετία έχουν κυκλοφορήσει μοντέλα με εξαιρετική ικανότητα παραγωγής εικόνων υψηλής ποιότητας απο περιγραφές κειμένου που βασίζονται σε μηχανισμούς διάχυσης όπως το DALL-E [48] [49], το Midjourney [50] και το Stable Diffusion [51]. Το Σχήμα 2.7 περιλαμβάνει εικόνες που έχουν δημιουργηθεί με τα παραπάνω μοντέλα.

DALL-E 2

"a teddy bear on a skateboard in times square"

"a shiba inu wearing a beret and black turtleneck"

"a close up of a handpalm with leaves growing from it"

Midjourney

"Metalpoint drawing, vintage, abstract, fashion art, colorful, artistic, expressive, sketch, Crosshatching, in the style lillian bassman, Sophie Calle, Coco Chanel , Dior, Prada, Sharan Ranshi, RENÉ BOUCHÉ, ALFREDO BOURET, RENÉ GRUAU,BIL DONOVAN"

"halloween t shirt, cute design cat with guirland, flat illustration"

"a woman in a white dress holding a large pink flower, in the style of surreal fashion photography, chinese cultural themes, serene oceanic vistas, made of plastic, light red and red, colorful costumes, pop inspo"

Σχήμα 2.7. Συνθετικές εικόνες δημιουργημένες με το DALL-E και το Midjourney (πηγή: https://www.midjourney.com/showcase/recent/).

Συνοψίζοντας, τα μοντέλα διάχυσης, παρόλο που αποτελούν πρόσφατη ανακάλυψη στον τομέα της σύνθεσης εικόνας, έχουν ήδη αποδειχθεί ικανά παραγωγής εικόνων που χαρακτηρίζονται από υψηλό βαθμό αληθοφάνειας και υψηλή ποιότητα. Η υλοποίηση τους μέσω αρχών διάχυσης, σε αντίθεση με αντιπαραθετική μάθηση, σε συνδυασμό με την δεκτικότητα τους σε παραλληλοποίηση, τα καθιστά χρήσιμα εργαλεία με ποικίλες εφαρμογές. Παρακάτω, θα αναλυθεί λεπτομερώς ένα πρόσφατο μοντέλο διάχυσης ανοικτού κώδικα το οποίο χρησιμοποιήθηκε για την υλοποίηση της παρούσας διατριβής.

2.4. Μοντέλα Λανθάνουσας Διάχυσης

Στην προηγούμενη ενότητα, αναλύθηκαν οι βασικές αρχές των μοντέλων διάχυσης και η εφαρμογή τους στη σύνθεση εικόνας. Παρόλο που τα μοντέλα διάχυσης σηματοδοτούν μια σημαντική εξέλιξη για τη δημιουργία εικόνων υψηλής ποιότητας, είναι υπολογιστικά απαιτητικά και χρειάζονται σημαντικούς υπολογιστικούς πόρους τόσο για εκπαίδευση όσο και για την εκτέλεση τους [52] [53]. Η παρούσα ενότητα εστιάζει σε μια πρόσφατη ανακάλυψη στον τομέα της σύνθεσης εικόνας, τα μοντέλα λανθάνουσας διάχυσης (latent diffusion models - LDMs) [51].

2.4.1. Εισαγωγή

Τα μοντέλα λανθάνουσας διάχυσης, τα οποία θα αναφέρονται στη συνέχεια ως LDMs, εισάγουν μια εναλλακτική προσέγγιση που αντιμετωπίζει τις υπολογιστικές προκλήσεις που σχετίζονται με τα μοντέλα διάχυσης διατηρώντας, ή ακόμη και βελτιώνοντας, την ποιότητα των παραγόμενων εικόνων. Η θεμελιώδης αρχή των LDM συνίσταται στην εφαρμογή της διαδικασίας αντίστροφης διάχυσης στον λανθάνοντα χώρο, με μια διαδικασία δύο σταδίων. Η παραλλαγή αυτή αξιοποιεί τα πλεονεκτήματα της βαθιάς αντιληπτικής συμπίεσης (deep perceptual compression) μέσω αυτοκωδικοποιητών για τη μείωση της υπολογιστικής πολυπλοκότητας διατηρώντας παράλληλα τις βασικές λεπτομέρειες της εικόνας [54].

2.4.2. Ελάττωση Διάστασης Χώρου

Το πρώτο στάδιο λειτουργίας των LDM περιλαμβάνει την συμπίεση εικόνας διατηρώντας τα κρίσιμα αντιληπτά χαρακτηριστικά. Αυτή επιτυγχάνεται με τη χρήση ενός αυτοκωδικοποιητή $\mathcal{E}, \mathcal{D}$ με στόχο την κωδικοποίηση εικόνων υψηλών διαστάσεων $x \in \mathbb{R}^{H \times W x 3}$ σε αναπαραστάσεις χαμηλότερης διάστασης $z \in \mathbb{R}^{h \times w x c}$ που ανήκουν σε έναν λανθάνοντα χώρο. Ο μετασχηματισμός από τον χώρο εικόνας στον λανθάνοντα χώρο επιτυγχάνεται μέσω του κωδικοποιητή, $z = \mathcal{E}(x)$, ενώ η ανάκτηση μιας προσέγγισης της αρχικής εικόνας γίνεται με τον αποκωδικοποιητή, $\tilde{x} = \mathcal{D}(z)$. Ο λανθάνοντας αυτός χώρος σχεδιάζεται έτσι ώστε να είναι αντιληπτικά ισοδύναμος με τον αρχικό χώρο της εικόνας, διασφαλίζοντας την διατήρηση των σημαντικών σημασιολογικών πληροφορών, αφαιρώντας, ωστόσο, χαρακτηριστικά υψηλής συχνότητας τα οποία δεν γίνονται αντιληπτά. Ο βαθμός υποδειγματοληψίας (downsampling) $f = H/h = W/w$ οφείλει να επιλεχθεί κατάλληλα έτσι ώστε να επιτευχθεί μια ισορροπία μεταξύ της υπολογιστικής απόδοσης και της πιστής ανακατασκευής των εικόνων στο στάδιο της αποκωδικοποίησης.

2.4.3. Εφαρμογή Αντίστροφης Διάχυσης

Ο σχεδιασμός του αποδοτικότερου λανθάνοντα χώρου επιτρέπει την υλοποίηση του δεύτερου σταδίου, το οποίο περιλαμβάνει την εφαρμογή της αντίστροφης διάχυσης σε διακριτά βήματα με σκοπό τη σύνθεση εικόνας. Η βασική καινοτομία των LDM συνίσταται στη χρήση αυτού του συμπαγούς λανθάνοντα χώρου για την πιθανοτική παραγωγική μοντελοποίηση των μοντέλων διάχυσης.

Η αρχιτεκτονική του νευρωνικού δικτύου που οφείλεται για την αντίστροφη διάχυση ακολουθεί μια δομή U-Net, η οποία έχει αποδειχθεί ιδιαίτερα αποτελεσματική για τη μοντελοποίηση δεδομένων με χωρική δομή [55]. Η αρχιτεκτονική αυτή αξιοποιεί την εγγενή μεροληψία του δικτύου για να επιτρέψει στα LDM να εστιάσουν στις πιο σημασιολογικά σημαντικές πτυχές των δεδομένων.

Όπως περιγράφηκε στην ενότητα 2.3, η εκπαίδευση των LDM βασίζεται σε μια διαδικασία διάχυσης η οποία περιλαμβάνει τη σταδιακή αποθορυβοποίηση μιας κανονικά κατανεμημένης μεταβλητής. Η διαδικασία αυτή αποτελεί την αντιστροφή της πορείας μιας αλυσίδας Markov μήκους T. Η αποθορυβοποίηση υπό συνθήκες (conditional denoising) επιτυγχάνεται μέσω μιας ακολουθίας

αυτοκωδικοποιητών, που συμβολίζονται ως $\epsilon_\theta(z_t, y, t); t = 1, \dots, T$, όπου ως z_t συμβολίζεται η λανθάνουσα αναπαράσταση της εικόνας για βήμα t της αλυσίδας Markov, υπό τη συνθήκη y. Η συνθήκη αυτή μπορεί να είναι κείμενο [56], σημασιολογικός χάρτης [57] ή να σχετίζεται με μετασχηματισμό εικόνας-σε-εικόνα [58]. Η ένταξη της συνθήκης αυτής υλοποιείται μέσω της επάυξησης του U-Net με μηχανισμό διασταυρούμενης προσοχής (cross-attention) [59].

2.4.4. Προεπεξεργασία Συνθηκών

Ο αυτοκωδικοποιητής που πραγματοποιεί την αποθορυβοποίηση υπό συνθήκες δέχεται ως είσοδο και συνθήκη y η οποία δύναται να είναι, μεταξύ άλλων, κάποια κωδικοποιημένη μορφή περιγραφής κειμένου. Για τον μετασχηματισμό της συνθήκης χρησιμοποιείται ένας ειδικός κωδικοποιητής τ_θ ο οποίος μετατρέπει τη συνθήκη y σε μια ενδιάμεση αναπαράσταση $\tau_\theta(y) \in \mathbb{R}^{M \times d_t}$ η οποία με τη σειρά της προβάλλεται στα ενδιάμεσα επίπεδα του U-Net μέσω ενός επιπέδου διασταυρούμενης προσοχής [51]. Το επίπεδο υλοποιεί τον μηχανισμό διασταυρούμενης προσοχής:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{6}$$

με

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t) \tag{7}$$

$$K = W_K^{(i)} \cdot \tau_\theta(y) \tag{8}$$

$$V = W_V^{(i)} \cdot \tau_\theta(y) \tag{9}$$

όπου $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ συμβολίζει μια πεπλατυσμένη, ενδιάμεση αναπαράσταση του ϵ_θ και τα $W_V^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}$, $W_Q^{(i)} \in \mathbb{R}^{d \times d_t}$ και $W_K^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}$ είναι μητρώα που υπόκεινται σε εκπαίδευση [60].

2.4.5. Συνάρτηση Απώλειας Αυτοκωδικοποιητή Αποθορυβοποίησης

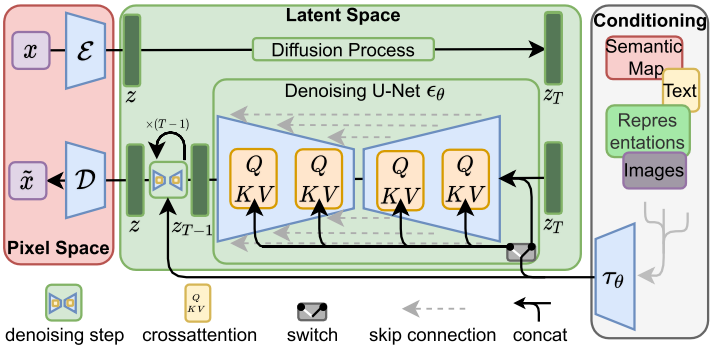
Η τελική συνάρτηση απώλειας που χρησιμοποιείται για την εκπαίδευση των αυτοκωδικοποιητών αποθορυβοποίησης των LDM ορίζεται ως η αναμενόμενη τιμή του τετραγώνου του L2 μέτρου του σφάλματος μεταξύ της αληθινής και της αποθορυβοποιημένης πρόβλεψης της λανθάνουσας αναπαράστασης της εικόνας για βήμα t:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(z_t, \tau_\theta(y), t)\|_2^2\right] \tag{10}$$

Η σχέση (10) [51] χρησιμοποιείται για την εκπαίδευση του αυτοκωδικοποιητή ϵ_θ και του κωδικοποιητή τ_θ.

2.4.6. *Αρχιτεκτονική των LDM*

Η τελική δομή του μοντέλου, περιλαμβάνοντας όλα τα επιμέρους μέλη που συνεργάζονται για την παραγωγή εικόνας, μπορεί να φανεί στο Σχήμα 2.8.



Σχήμα 2.8. Η αρχιτεκτονική ενός μοντέλου λανθάνουσας διάχυσης [51].

Το σχήμα αποτυπώνει όλα τα στάδια λειτουργίας ενός LDM, περιλαμβάνοντας την αρχική κωδικοποίηση της εικόνας x μέσω του κωδικοποιητή $\mathcal{E}$ στη λανθάνουσα μεταβλητή z, τη διαδικασία διαδοχικής αποθορυβοποίησης μέσω του αυτοκωδικοποιητή ϵ_θ, την κωδικοποίηση της συνθήκης y μέσω του κωδικοποιητή τ_θ καθώς και την ένταξη της εντός του U-Net μέσω του μηχανισμού διασταυρούμενης προσοχής και, τέλος, την αποκωδικοποίηση της αποθορυβοποιημένης λανθάνουσας μεταβλητής z μέσω του αποκωδικοποιητή $\mathcal{D}$ και την επιστροφή στον χώρο της εικόνας. Η τελική έξοδος του δικτύου είναι η παραγόμενη εικόνα $\tilde{x}$.

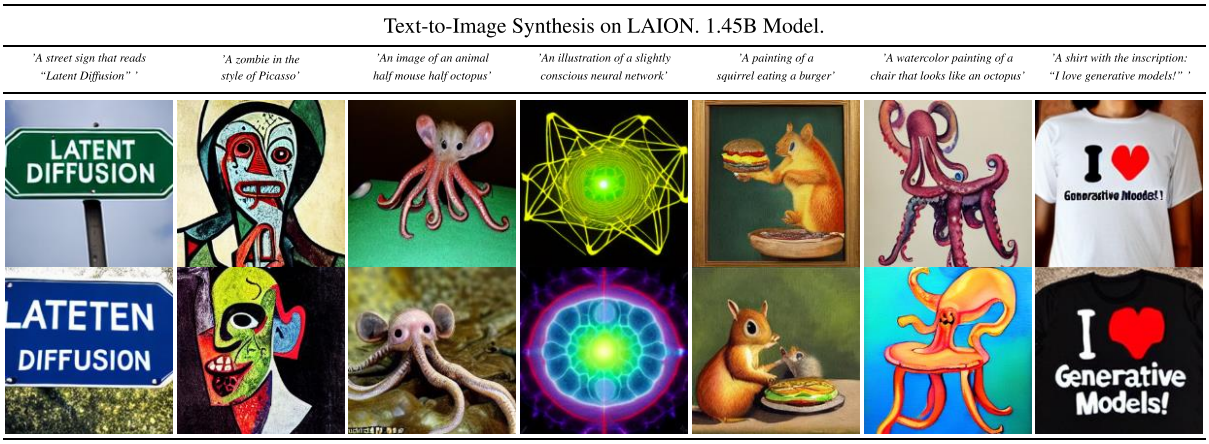
2.4.7. Παραδείγματα Παραγόμενων Εικόνων

Τα πειραματικά αποτελέσματα χρήσης LDM αναδεικνύει τις δυνατότητες των μοντέλων αυτών στην επίτευξη σύνθεσης εικόνων υψηλής ποιότητας. Η απόδοση των LDM ξεπερνάει αυτή των παλαιότερων μοντέλων, όπως των GAN και άλλων πιθανοτικών προσεγγίσεων. Επιπλέον, έχουν δείξει αξιοσημείωτες δυνατότητες σε εργασίες μεγέθυνσης εικόνας και αφαίρεσης αντικειμένων εντός μιας σκηνής. Στο Σχήμα 2.9 φαίνονται παραδείγματα παραγόμενων εικόνων από μοντέλα λανθάνουσας διάχυσης που έχουν εκπαιδευτεί σε γνωστά σύνολα δεδομένων.

Σχήμα 2.9. Παραδείγματα συνθετικών εικόνων παραγόμενες από LDM εκπαιδευμένα στα σύνολα δεδομένων CelebAHQ [61], FFHQ [44], LSUN-Churches [62], LSUN-Beds [62] και ImageNet [63]. Πηγή: [51] .

Τα δείγματα που παράγονται από τα μοντέλα λανθάνουσας διάχυσης χαρακτηρίζονται από εξαιρετική αληθοφάνεια, ποιότητα και ποικιλία, καθώς αποφεύγουν την κατάρρευση λειτουργίας των GAN, αξιοποιώντας την τυχαιότητα που εισάγεται μέσω του τυχαίου θορύβου Gauss κατά τη διαδικασια διάχυσης.

Το Stable Diffusion είναι ένα μοντέλο λανθάνουσας διάχυσης ανοικτού κώδικα, 1.45 δισ. παραμέτρων, που προέκυψε από την παραπάνω μελέτη [51] και εκπαιδεύτηκε στο σύνολο δεδομένων LAION-400M [64]. Στο Σχήμα 2.10 παρατίθενται παραδείγματα σύνθετικών εικόνων παραγόμενων από το Stable Diffusion με συνθήκη περιγραφής κειμένου. Η διαδικασία παραγωγής εικόνας βάσει περιγραφής κειμένου έχει αποκτήσει την ονομασία text-to-image.



Σχήμα 2.10. Παραδείγματα συνθετικών εικόνων παραγόμενες από το Stable Diffusion με χρήση συνθήκης περιγραφής κειμένου [51].

Η δυνατότητα παραγωγής εικόνας βάσει περιγραφής κειμένου επιτρέπει την σύνθεση εικόνων υψηλής ποιότητας δημιουργικού και ποικίλου περιεχομένου. Παρατηρείται πως οι εικόνες ακολουθούν πιστά την περιγραφή κειμένου διατηρώντας όμως υψηλό βαθμό μεταβλητότητας. Εικόνες που έχουν παραχθεί σύμφωνα με την ίδια περιγραφή κειμένου διαφέρουν σημαντικά στην τοποθέτηση των αντικειμένων και στο στυλ. Η ιδιότητα αυτή επιτρέπει την σύνθεση μεγάλου και ποικιλόμορφου πλήθους εικόνων, γεγονός το οποίο δύναται να αξιοποιηθεί στη σύνθεση συνόλων δεδομένων για εφαρμογές μηχανικής όρασης.

2.5. Έλεγχος των LDM με Συνθήκες

Η ικανότητα λεπτομερούς ελέγχου στη σύνθεση των εικόνων των μοντέλων λανθάνουσας διάχυσης έχει αποτελεί πρόκληση ακολουθώντας την κυκλοφορία τους στο κοινό. Σε αυτή την ενότητα εισάγεται το ControlNet, ένα ειδικό νευρωνικό δίκτυο, που έχει αναδειχθεί ως ένα ισχυρό εργαλείο για την προσθήκη χωρικών συνθηκών ελέγχου σε μεγάλα προεκπαιδευμένα μοντέλα διάχυσης κειμένου-σε-εικόνα [56].

2.5.1. Η Ανάγκη Για Έλεγχο Με Συνθήκες

Τα LDM έχουν επιδείξει αξιοσημείωτες ικανότητες στη δημιουργία εικόνων από περιγραφές κειμένου, ωστόσο, παραμένει η πρόκληση για την επίτευξη ακριβούς χωρικού ελέγχου στη σύνθεση των εικόνων αυτών. Η αποτύπωση πολύπλοκων διατάξεων, στάσεων, σχημάτων και μορφών αποκλειστικά μέσω περιγραφών κειμένου είναι περιοριστική και συχνά καταλήγει σε μια πληκτική, επαναλαμβανόμενη διαδικασία πειραματισμού μέσω τροποποίησης των περιγραφών αυτών μέχρι τη

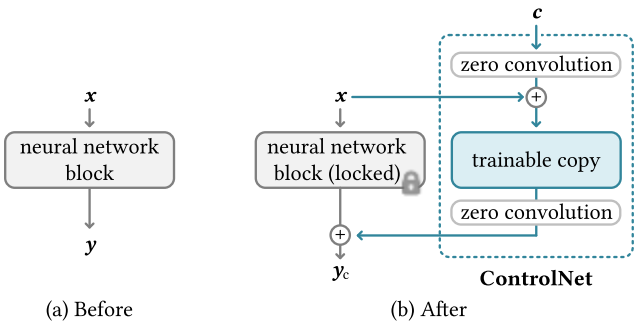
λήψη του επιθυμητού αποτελέσματος. Το ControlNet, με σκοπό την εξάλειψη αυτού του περιορισμού, επιδιώκει την αξιοποίηση πρόσθετων εικόνων που περιλαμβάνουν πληροφορίες όπως ακμές, βάθος, σημασιολογική τμηματοποίηση, στάση σώματος και άλλα, για τον άμεσο έλεγχο της επιθυμητής σύνθεσης της εικόνας. Οι εικόνες αυτές εισάγονται ως συνθήκες που καθοδηγούν τη διαδικασία δημιουργίας εικόνας.

2.5.2. Η Αρχιτεκτονική του ControlNet

Το ControlNet επαυξάνει τη δομή προεκπαιδευμένων μοντέλων διάχυσης κειμένου-σε-εικόνα, ενισχύοντας αποτελεσματικά τις δυνατότητές τους με την εισαγωγή επιπλέον συνθηκών. Η αρχιτεκτονική του χαρακτηρίζεται από μια δομή νευρωνικού δικτύου που λειτουργεί παράλληλα με το αρχικό μοντέλο, αξιοποιώντας τα βαθιά και εύρωστα επίπεδα κωδικοποίησης του που έχουν εκπαιδευτεί σε δισεκατομμύρια εικόνες.

Το βασικό στοιχείο σχεδιασμού του ControlNet συνίσταται στην εισαγωγή «μηδενικών συνελίξεων», δηλαδή, συνελικτικών επιπέδων που συμβολίζονται ως $\mathcal{Z}(\cdot;\cdot)$ που αρχικοποιούνται με μηδενικά βάρη και κατώφλια, για τη σύνδεση του αρχικού μοντέλου και ενός εκπαιδεύσιμου αντιγράφού του. Η σύνδεση αυτή διασφαλίζει τη μη εισαγωγή επιβλαβούς θορύβου στα βαθιά χαρακτηριστικά του προεκπαιδευμένου μοντέλου κατά τη διάρκεια της εκπαίδευσης, και κατά επέκταση την ποιότητα της σύνθεσης.

Έστω αρχικό μοντέλο που δέχεται είσοδο x και υπολογίζει έξοδο y. Η προσθήκη ελέγχου μέσω ControlNet στο μοντέλο αυτό επιτυγχάνεται με το «κλείδωμα» του, τη δημιουργία του εκπαιδεύσιμου αντιγράφου του και σύνδεση των δύο με μηδενικά συνελικτικά επίπεδα. Η εισαγωγή της συνθήκης c πραγματοποιείται στο αντίγραφο μοντέλο. Μια απλοποιημένη αναπαράσταση της διαδικασίας αυτής φαίνεται στο Σχήμα 2.11.



Σχήμα 2.11. Απλοποιημένη αναπαράσταση της ένταξης μιας συνθήκης c σε ένα νευρωνικό δίκτυο μέσω του ControlNet [56].

2.5.3. Εκπαίδευση του ControlNet

Η διαδικασία εκπαίδευσης του ControlNet είναι εύρωστη και αποτελεσματική. Σε αντιστοιχία με τα μοντέλα διάχυσης, το δίκτυο δέχεται σαν είσοδο μια εικόνα z_0, το βήμα διάχυσης t και την συνθήκη c_t, η οποία είναι μια περιγραφή κειμένου. Το ControlNet δέχεται μια επιπλέον συνθήκη c_f η διαφέρει για κάθε μορφή εισόδου. Κατά την εκπαίδευση, καλείται να ελαχιστοποιηθεί η τροποποιημένη συνάρτηση απώλειας [56]:

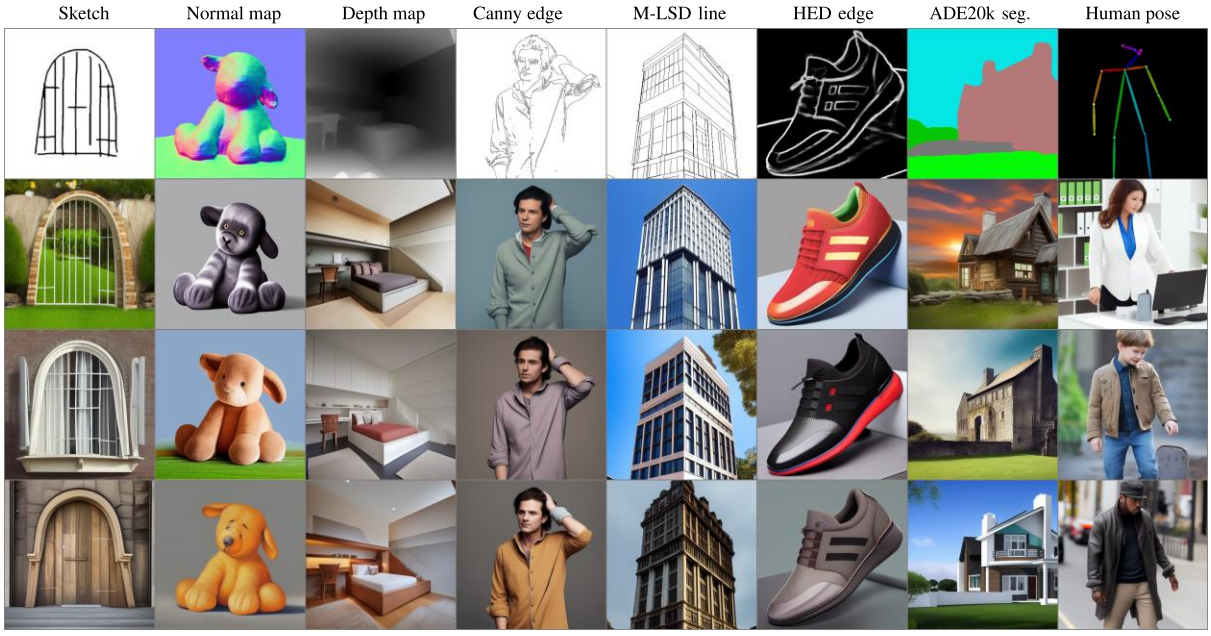
$$\mathcal{L} = \mathbb{E}_{z_0, c_t, c_f, \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, c_t, c_f, t\right)\right\|_2^2\right] \tag{11}$$

η οποία περιλαμβάνει και την ειδική συνθήκη c_f. Με σκοπό την άυξηση της εγγενούς ικανότητας σημασιολογικής αναγνώρισης του ControlNet, το 50% των περιγραφών κειμένου στη συνθήκη c_t αντικαθίσταται με κενό κείμενο. Αυτή η τροποποίηση επιτρέπει στο ControlNet να αποκτήσει σημασιολογική κατανόηση του περιεχομένου της εικόνας μόνο από της συνθήκες εισόδου, δηλαδή, ακμές, βάθος, στάση σώματος κλπ..

Κατά την εκπαίδευση του ControlNet, παρατηρείται ένα φαινόμενο ξαφνικής σύγκλισης της απόδοσης του μοντέλου ύστερα από έναν αριθμό κύκλων βελτιστοποίησης, όπου αυτό αποκτά σχεδόν ακαριαία την ικανότητα σύνθεσης εικόνων που τηρούν πιστά την συνθήκη εισόδου. Το φαινόμενο αυτό έρχεται σε αντίθεση με την αναμενόμενη συμπεριφορά της σταδιακής βελτίωσης της απόδοσης του μοντέλου κατά τη διαδικασία της εκπαίδευσης.

2.5.4. Αποτελέσματα

Με το πέρας της εκπαίδευσης, το ControlNet παρέχει τη δυνατότητα λεπτομερούς ελέγχου της σύνθεσης εικόνων βάσει ποικίλων συνθηκών. Στο Σχήμα 2.12, φαίνονται μερικά παραδείγματα σύνθεσης εικόνων για διαφορετικές μορφές εισόδου ελέγχου.



Σχήμα 2.12. Παραδείγματα σύνθεσης εικόνων με το Stable Diffusion και επιπλέον έλεγχο με διάφορες συνθήκες μέσω του ControlNet [56].

Το ControlNet υλοποιείται για συνεργασία με το Stable Diffusion για διάφορες συνθήκες εισόδου όπως ακμές Canny [65], χάρτες βάθους [66], χάρτες κάθετων επιφανειών [67], γραμμές M-LSD [68], ακμές HED [69], τμηματοποίηση ADE20K [70], στάση σώματος Openpose [71] και σκαριφημάτων του χρήστη.

2.5.5. Συμπεράσματα

Το ControlNet αντιπροσωπεύει μια σημαντική πρόοδο στον τομέα των μοντέλων διάχυσης καθώς αποτελεί ένα ισχυρό εργαλείο που προσφέρει μεγάλο βαθμό ελέγχου σε προεκπαιδευμένα μοντέλα κειμένου-σε-εικόνα. Το ControlNet πρωτοπορεί στον ακριβή χωρικό έλεγχο στη σύνθεση εικόνας, το οποίο το καθιστά πολύτιμο εργαλείο για διάφορες εφαρμογές στη δημιουργία και χειρισμό εικόνων, όπως η σύνθεση συνόλων δεδομένων.

Κεφάλαιο 3

Δημιουργία Συνθετικού Συνόλου Δεδομένων

3.1. Κίνητρο

Η ανάγκη ανάπτυξης εύρωστων μοντέλων ανίχνευσης αντικειμένων για εφαρμογή σε βιομηχανικά περιβάλλοντα υποστηρίζεται από την αναμφισβήτητη αξία της μηχανικής όρασης στην παραγωγή. Ωστόσο, η επιδίωξη της υλοποίησης τέτοιων μοντέλων αποκαλύπτει πολλές σημαντικές προκλήσεις που καλούν για έναν εναλλακτικό τρόπο προσέγγισης, ο οποίος συνίσταται στη δημιουργία συνθετικών δεδομένων. Οι κύριες προκλήσεις που οδηγούν σε αυτή την επιλογή είναι οι εξής:

- Απουσία δημοσίως διαθέσιμων συνόλων δεδομένων – Η έλλειψη συνόλων δεδομένων που καλύπτουν την ποικιλομορφία του εξειδικευμένου βιομηχανικού εξοπλισμού αποτελεί μια θεμελιώδη πρόκληση. Σε αντίθεση με συνήθη αντικείμενα, για τα οποία υπάρχουν εκτεταμένα σύνολα δεδομένων που αξιοποιούνται για πιο συμβατικές εργασίες, η σπάνια αλληλεπίδραση του γενικού πληθυσμού με αντικείμενα που σχετίζονται με το επίπεδο παραγωγής καθώς και η τεχνική φύση αυτών οφείλονται για το εμφανές κενό σε αυτόν τον τομέα. Η απουσία κατάλληλων δεδομένων χρησιμεύει ως πρωταρχικό κίνητρο για την αναζήτηση εναλλακτικών λύσεων.

- Πνευματικά δικαιώματα – Η συμβατική προσφυγή της άντλησης δεδομένων από το διαδίκτυο για σχετικές εικόνες, που προέρχονται από κατασκευαστές και άλλες πηγές, είναι απαγορευτική λόγω περιορισμών αδειοδότησης. Πολλές εικόνες βιομηχανικού εξοπλισμού που βρίσκονται στο διαδίκτυο υπόκεινται σε πνευματικά δικαιώματα και άδειες χρήσης που αποκλείουν τη χρήση τους για εκπαίδευση μοντέλων μηχανικής μάθησης. Ο νομικός αυτός περιορισμός υπογραμμίζει την ανάγκη για μια πιο δημιουργική και νομικά ορθή προσέγγιση.

- Αδυναμία λήψης φωτογραφιών – Η δημιουργία ενός συνόλου δεδομένων με τη λήψη εικόνων εντός βιομηχανικών εγκαταστάσεων δεν αποτελεί πρακτική λύση καθώς αντιμετωπίζει μια σειρά από ανυπέρβλητες προκλήσεις. Η πολυπλοκότητα που σχετίζεται με τη φυσική παρουσία σε πολυάριθμες εγκαταστάσεις, καθεμία από τις οποίες βρίσκεται σε διαφορετικές γεωγραφικές τοποθεσίες, καθιστά μια τέτοια προσέγγιση ακριβή και χρονοβόρα. Επιπλέον, οι βιομηχανικές εγκαταστάσεις περιλαμβάνουν ιδιόκτητο εξοπλισμό και συστήματα, καθιστώντας τη λήψη εικόνων απαγορευτική. Ακόμη, η λήψη εικόνων εντός τέτοιων εγκαταστάσεων απαιτεί τη συγκατάθεση του εργαζόμενου προσωπικού που χειρίζεται τα μηχανήματα. Η κατάσταση περιπλέκεται περισσότερο αν ληφθεί υπόψιν και η ικανοποίηση των λειτουργικών απαιτήσεων και περιορισμών των εργαζομένων που μπορεί να μην έχουν την ευελιξία να παραμερίσουν την εργασία τους για μια φωτογραφία.

Συλλογικά, οι παραπάνω λόγοι οδήγησαν στην εξερεύνηση της εναλλακτικής προσέγγισης δημιουργίας συνθετικών δεδομένων ως μια πρακτική και καινοτόμος λύση. Στις επόμενες ενότητες αυτού του κεφαλαίου, αναλύεται ενδελεχώς η διαδικασία δημιουργίας συνθετικών δεδομένων των υπό μελέτη κατηγοριών αντικειμένων.

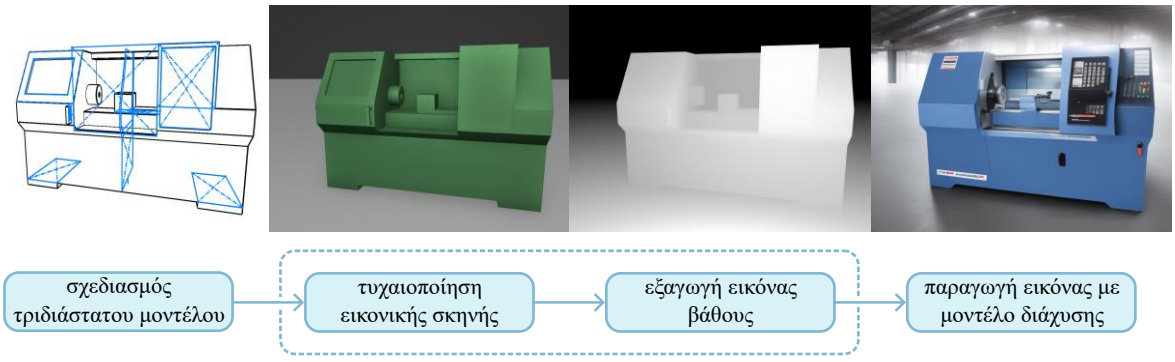
3.2. Δημιουργία Συνθετικών Εικόνων

Λόγω της έλλειψης διαθέσιμων αληθινών εικόνων των υπό μελέτη κατηγοριών αντικειμένων, επιλέχθηκε η προσέγγιση της σύνθεσης τεχνητών εικόνων χρησιμοποιώντας καινοτόμα εργαλεία μηχανικής μάθησης. Το πρωτοπόρο μοντέλο λανθάνουσας διάχυσης Stable Diffusion, σε συνδυασμό με τη χρήση εικόνων βάθους για τον περαιτέρω έλεγχο της σύνθεσης των παραγόμενων εικόνων μέσω το ControlNet, αξιοποιήθηκαν για την παραγωγή μεγάλου πλήθους αληθοφανών συνθετικών εικόνων βιομηχανικού εξοπλισμού και εργαλειομηχανών καθώς και εργαζόμενου προσωπικού εντός βιομηχανικών εγκαταστάσεων.

Η παρούσα μέθοδος σύνθεσης δεδομένων αποτελεί μια προσέγγιση προσαρμογής πεδίου (domain adaptation) [72] [73] [74], η οποία περιλαμβάνει την εκπαίδευση ενός μοντέλου μηχανικής μάθησης με δεδομένα από έναν αρχικό τομέα με στόχο την εφαρμογή του σε έναν διαφορετικό τομέα, μετριάζοντας τις επιπτώσεις της αλλαγής αυτής στην επίδοση του. Ο στόχος της προσέγγισης αυτής είναι η επίτευξη ικανής απόδοσης και γενίκευσης στον τομέα-στόχο, ακόμη και όταν η κατανομή των δεδομένων στον τομέα-στόχο διαφέρει από εκείνη του τομέα προέλευσης. Αυτό επιτυγχάνεται με την εκμάθηση χαρακτηριστικών που παραμένουν αμετάβλητα από τομέα σε τομέα (domain-invariant features) ή προσαρμόζοντας το μοντέλο στα χαρακτηριστικά του τομέα-στόχου. Η τρέχουσα μελέτη αντλεί συνθετικά δεδομένα με σκοπό την εφαρμογή ενός μοντέλου ανίχνευσης αντικειμένων σε αληθινό περιβάλλον.

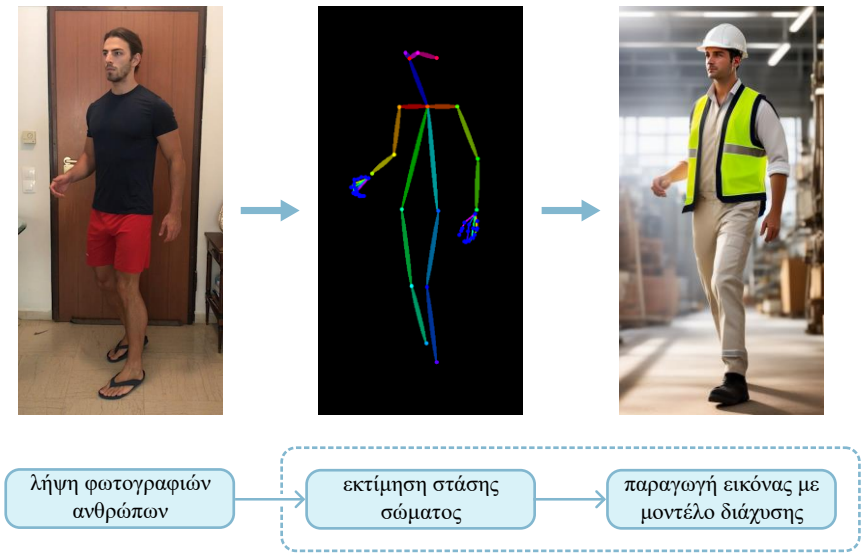
Η διαδικασία παραγωγής των εικόνων αποτελείται από τρεις διαδοχικές φάσεις με τις πρώτες δύο να διαφέρουν για τις εικόνες μηχανημάτων και τις εικόνες εργαζομένων.

Όσον αφορά τις εικόνες μηχανημάτων και εξοπλισμού, σε πρώτο στάδιο, σχεδιάζονται προσεγγιστικές τριδιάστατες αναπαραστάσεις των αντικειμένων. Έπειτα, το κάθε τριδιάστατο μοντέλο τοποθετείται σε εικονική σκηνή όπου παράμετροι που αφορούν την διάταξή της τυχαιοποιούνται και εξάγονται εικόνες βάθους. Τέλος, οι εικόνες βάθους εισάγονται στο Stable Diffusion μέσω του ControlNet για την παραγωγή συνθετικών εικόνων που τηρούν τη δεδομένη γεωμετρία των αντικειμένων. Η ροή εργασίας η οποία περιγράφει την παραγωγή εικόνων των κατηγοριών αντικειμένων εξοπλισμού αποτυπώνεται στο Σχήμα 3.1.



Σχήμα 3.1. Η ροή εργασίας που ακολουθείται για την παραγωγή συνθετικών εικόνων μηχανημάτων.

Όσον αφορά τις εικόνες εργαζομένων, οι πρώτες δύο φάσεις διαφέρουν από την προηγούμενη διαδικασία. Σε πρώτο στάδιο, γίνεται λήψη φωτογραφιών ανθρώπων σε ελεγχόμενο περιβάλλον χωρίς την ανάγκη να φέρουν ειδική ενδυμασία και μέσα ατομικής προστασίας (personal protective equipment - PPE). Έπειτα, η στάση του σώματος των ανθρώπων στις φωτογραφίες εκτιμάται με τη χρήση του προεπεξεργαστή OpenPose και, τέλος, εισάγεται στο Stable Diffusion μέσω του ControlNet για την παραγωγή συνθετικών εικόνων εργαζομένων βιομηχανικών εγκαταστάσεων που τηρούν την ίδια στάση σώματος. Η ροή εργασίας η οποία περιγράφει την παραγωγή εικόνων των εργαζομένων αποτυπώνεται στο Σχήμα 3.2.



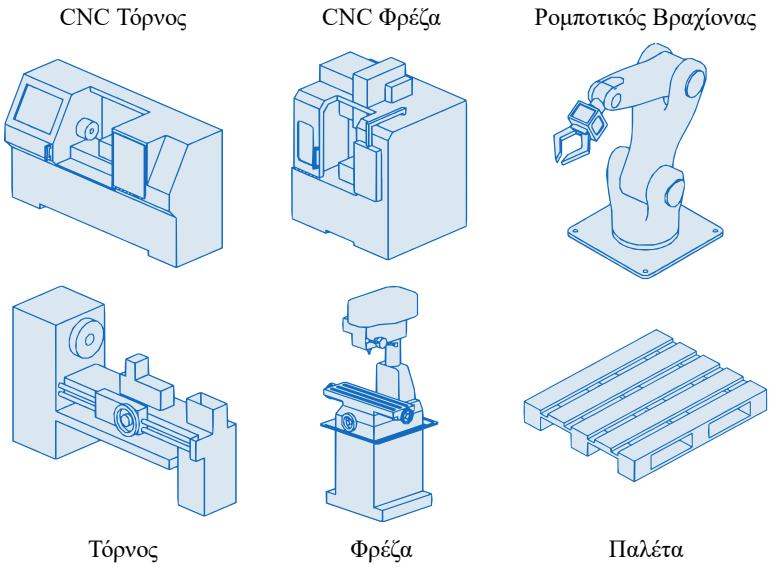
Σχήμα 3.2. Η ροή εργασίας που ακολουθείται για την παραγωγή συνθετικών εικόνων εργαζομένων.

Οι παραπάνω διαδικασίες χαρακτηρίζονται από μεγάλο βαθμό αυτοματοποίησης, επεκτασιμότητας και παραλληλοποίησης. Οι δυνατότητες του ControlNet για τον ακριβή έλεγχο της σύνθεσης της εικόνας μέσω διάφορων μορφών εισόδου αξιοποιούνται για τη σύνθεση μεγάλου πλήθους αληθοφανών εικόνων μηχανημάτων και ανθρώπων, χρησιμοποιώντας, σε κάθε περίπτωση, την πιο κατάλληλη ροή εργασίας. Η αναλυτική διαδικασία παραγωγής συνθετικών εικόνων ανθρώπων βρίσκεται στο Παράρτημα Α. Παρακάτω, αναλύεται ενδελεχώς κάθε φάση της διαδικασίας παραγωγής των εικόνων μηχανημάτων και εξοπλισμού.

3.2.1. Δημιουργία Τριδιάστατων Μοντέλων

Η πρώτη φάση της διαδικασίας παραγωγής των συνθετικών εικόνων περιλαμβάνει την τριδιάστατη μοντελοποίηση των υπό μελέτη κατηγοριών αντικειμένων με χρήση λογισμικού CAD. Το στάδιο αυτό δεν απαιτεί λεπτομερή σχεδιασμό των μικρότερων χαρακτηριστικών των αντικειμένων. Αντίθετα, τα αντικείμενα μοντελοποιούνται πρόχειρα και χονδρικά με σκοπό την αποτύπωση του γενικού τους σχήματος και των πιο βασικών χαρακτηριστικών τους. Η μεταβλητότητα των εικόνων που απαιτείται για την εκπαίδευση ενός εύρωστου μοντέλου ανίχνευσης αντικειμένων, καθώς και η λεπτομέρεια που καθιστά τις εικόνες αληθοφανείς, εισάγονται στις επόμενες φάσεις και δεν αποτελούν αντικείμενο ενδιαφέροντος κατά τη διάρκεια του τριδιάστατου σχεδιασμού.

Στο Σχήμα 3.3 παρατίθενται μερικές όψεις των τριδιάστατων μοντέλων που σχεδιάστηκαν έτσι ώστε να δράσουν ως βάση για τα επόμενα στάδια.



Σχήμα 3.3. Τριδιάστατα μοντέλα των υπό μελέτη αντικειμένων.

Παρατηρείται πως οι λεπτομέρειες καθώς και άλλες ιδιότητες των αντικειμένων όπως το χρώμα και το υλικό απουσιάζουν από τα μοντέλα καθώς δεν απαιτούνται για την αποτύπωση των βασικών γεωμετρικών χαρακτηριστικών τους.

3.2.2. Τυχαιοποίηση Εικονικής Σκηνής

Σε αυτή τη φάση της δημιουργίας του συνθετικού συνόλου δεδομένων τα μοντέλα των αντικειμένων που σχεδιάστηκαν προηγουμένως τοποθετούνται και διατάσσονται εντός μιας εικονικής σκηνής. Μέσω της τυχαιοποίησης διάφορων παραμέτρων εντός της σκηνής, το λεγόμενο Domain Randomization [75], αποτυπώνονται πολλές και ποικίλες δυνατές διατάξεις των αντικειμένων σε μια προσπάθεια προσομοίωσης της τυχαιότητας που χαρακτηρίζει τις αληθινές φωτογραφίες. Το στάδιο αυτό είναι κρίσιμο καθώς εισάγει μεταβλητότητα στο συνθετικό σύνολο δεδομένων, η οποία επιτρέπει την εκπαίδευση ενός πιο εύρωστου μοντέλου ανίχνευσης αντικειμένων, αυξάνοντας την ικανότητα γενίκευσής του, όπως θα δειχθεί στη συνέχεια της εργασίας μέσω σχετικής σύγκριση.

Οι εικονικές σκηνές υλοποιούνται εντός του λογισμικού τριδιάστατης μοντελοποίησης Blender. Η κάθε σκηνή περιλαμβάνει ένα αντικείμενο, το έδαφος στο οποίο αυτό είναι τοποθετημένο και μια εικονική κάμερα η οποία καταγράφει τη σκηνή. Δημιουργείται μια χρονική αλληλουχία στιγμιοτύπων για $t \in [t_0, t_f]$, σε κάθε ένα από τα οποία τροποποιούνται διάφορες παράμετροι της σκηνής, είτε μέσω θορύβου είτε ακολουθώντας κάποια χρονική συνάρτηση. Με την κάμερα να εστιάζει στο γεωμετρικό κέντρο του αντικειμένου μέσω κινηματικού περιορισμού, τροποποιούνται οι κάτωθι παράμετροι:

- Αζιμούθιο κάμερας – Η κάμερα περιστρέφεται γύρω από κατακόρυφο άξονα, ο οποίος διέρχεται από το κέντρο του αντικειμένου, με γωνία η οποία μεταβάλλεται γραμμικά από μια αρχική έως μια τελική τιμή σύμφωνα με τον ψευδοχρόνο. Για τα περισσότερα αντικείμενα επιλέχθηκε $\varphi_0 = -45°, \varphi_f = 45°$. Η εξίσωση η οποία περιγράφει την περιστροφή της κάμερας γύρω από το αντικείμενο είναι η εξής:

$$\varphi = \varphi_0 + (\varphi_f - \varphi_0) \cdot \frac{t - t_0}{t_f - t_0} \tag{12}$$

- Απόσταση κάμερας από το αντικειμένο – Η απόσταση d μεταξύ της κάμερας και του αντικειμένου μεταβάλλεται λαμβάνοντας σε κάθε στιγμιότυπο τυχαία τιμή που ακολουθεί ομοιόμορφη κατανομή μεταξύ d_{min} και d_{max}. Οι ακραίες τιμές επιλέγονται κατάλληλα για κάθε αντικείμενο δεδομένης της διαφοράς μεγέθους που έχουν αυτά μεταξύ τους.

- Εστιακή απόσταση φακού – Η εστιακή απόσταση (focal length) f της εικονικής κάμερας μεταβάλλεται μεταξύ δύο ακραίων τιμών, έτσι ώστε να αντισταθμίσει τη μεταβολή της φυσικής απόστασης μεταξύ της κάμερας και του αντικειμένου, διασφαλίζοντας τη διατήρηση της κλίμακας του αντικειμένου εντός της εικόνας. Η τιμή της εστιακής απόστασης υπολογίζεται ως εξής:

$$f = f_{min} + (f_{max} - f_{min}) \cdot \frac{d - d_{min}}{d_{max} - d_{min}} \tag{13}$$

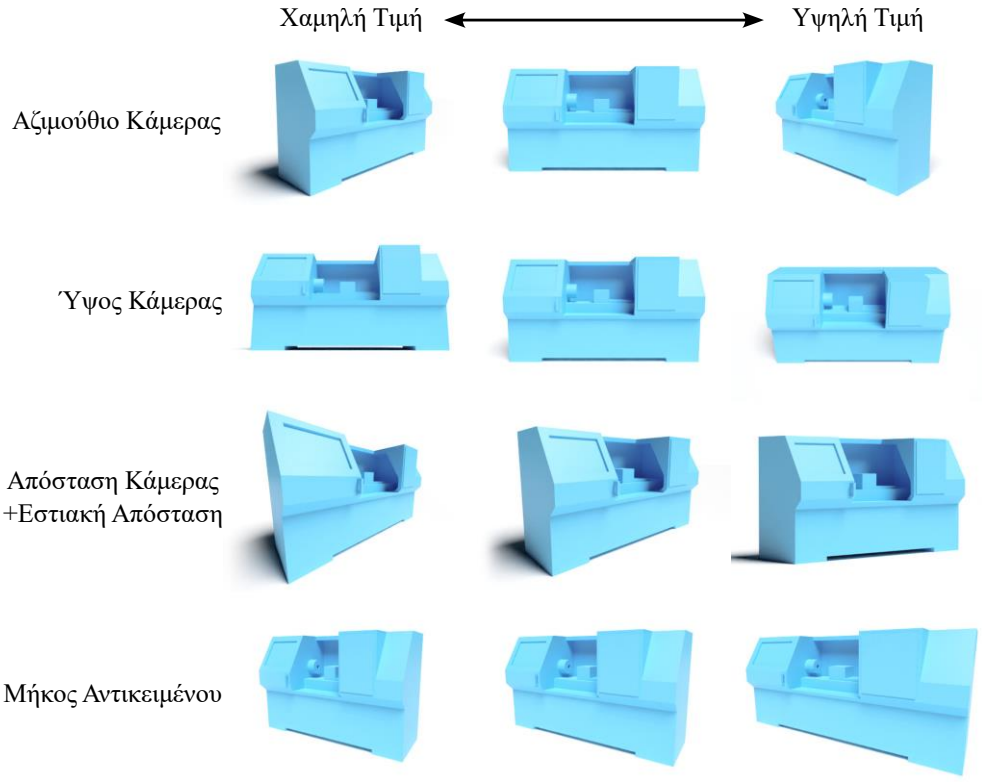
  Η εξάρτηση της εστιακής απόστασης από τη φυσική απόσταση μεταξύ της κάμερας και του αντικειμένου έχει ως αποτέλεσμα τη φαινομενική μεταβολή της προοπτικής της σκηνής. Μεγαλύτερες φυσικές αποστάσεις αντιστοιχούν στη χρήση τηλεφακού και την εξάλειψη της προοπτικής από την εικόνα προσεγγίζοντας ορθογραφική προβολή της σκηνής.

- Αναλογίες αντικειμένου – Η κλίμακα του αντικειμένου κατά τον διαμήκη άξονα του μεταβάλλεται λαμβάνοντας σε κάθε στιγμιότυπο τυχαία τιμή που ακολουθεί ομοιόμορφη κατανομή μεταξύ 0.8 και 1.2 . Το αποτέλεσμα αυτής της τροποποίησης είναι η μεταβολή των αναλογιών του αντικειμένου, με μεγαλύτερες τιμές να δίνουν πιο μακρόστενα αντικείμενα.

- Απόσταση κάμερας από το έδαφος – Η απόσταση h μεταξύ της κάμερας και του εδάφους μεταβάλλεται λαμβάνοντας σε κάθε στιγμιότυπο τυχαία τιμή που ακολουθεί ομοιόμορφη κατανομή μεταξύ h_{min} και h_{max}. Οι ακραίες τιμές επιλέγονται κατάλληλα για κάθε αντικείμενο δεδομένης της διαφοράς μεγέθους που έχουν αυτά μεταξύ τους. Καθώς η πρόνευση της κάμερας υπόκειται σε κινηματικό περιορισμό έτσι ώστε αυτή να εστιάζει πάντα στο αντικείμενο, η μεταβολή του ύψους της κάμερας αποδίδει λήψεις από χαμηλές και ψηλές γωνίες.

Οι αναλυτικές τιμές όλων των παραμέτρων για κάθε κατηγορία βρίσκονται στον Πίνακα Β.1 στο Παράρτημα Β. Η ταυτόχρονη τροποποίηση όλων των παραμέτρων σε κάθε στιγμότυπο παράγει αμέτρητους συνδυασμούς συνθηκών λήψης εικόνας και καταστάσεων των αντικειμένων. Η ποικιλομορφία αυτή παράγει σκηνές που διαφέρουν δραματικά από στατικές ή προκαθορισμένες λήψεις των υπό μελέτη αντικειμένων και εισάγει υψηλή μεταβλητότητα στο παραγόμενο σύνολο δεδομένων.

Για την ευκολότερη κατανόηση των παραπάνω παραμέτρων και της οπτικής επίδρασης της τροποποίησής τους στην εικόνα, παρατίθεται το Σχήμα 3.4.



Σχήμα 3.4. Μεταβολή παραμέτρων σκηνής κατά την τυχαιοποίηση σκηνής.

Παρατηρείται πως η τυχαιοποίηση των παραπάνω παραμέτρων είναι επαρκής για την επαύξηση του συνόλου δεδομένων και τη δημιουργία πολλών ποικίλων εικόνων από μια μόνο γεωμετρία.

3.2.3. Εξαγωγή Εικόνων Βάθους και Αυτόματη Επισήμανση

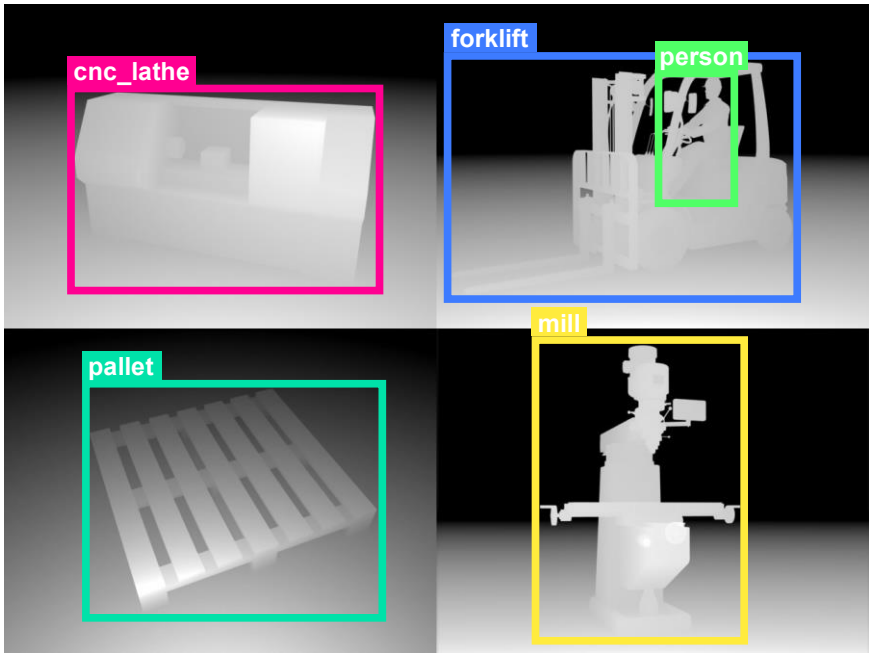
Ακολουθώντας την παραγωγή ενός ικανού πλήθους σκηνών για κάθε αντικείμενο μέσω τυχαιοποίησης, εξάγονται εικόνες βάθους για κάθε ένα από τα στιγμιότυπα. Οι εικόνες βάθους (depth map) είναι ασπρόμαυρες εικόνες με το χρώμα εντός αυτών να αντιστοιχεί στην απόσταση των αντικειμένων στον χώρο από την κάμερα. Για την παραγωγή των εικόνων βάθους, εφαρμόζεται ένας πολύκλαδος μετασχηματισμός που παρεμβάλει την απόχρωση στην εικόνα μεταξύ λευκού και μαύρου, ή 255 και 0 σε σύστημα 8bit, σύμφωνα με την απόσταση. Η απόχρωση κάθε pixel, το οποίο αποτυπώνει επιφάνεια που απέχει απόσταση d_c από την κάμερα, δίνεται από τη σχέση:

$$\text{βάθος}(d_c) = \begin{cases} 255, & d_c < d_{min} \\ 255\dfrac{d_{max} - d_c}{d_{max} - d_{min}}, & d_{min} \le d_c \le d_{max} \\ 0, & d_{max} < d_c \end{cases} \tag{14}$$

Το βάθος φράσσεται μεταξύ 0 και 255, για αποστάσεις μεταξύ d_{min} και d_{max} έτσι ώστε να παράγεται εικόνα που να είναι συμβατή με τα γραφικά υπολογιστών, ειδάλλως, τα pixel που αντιστοιχούν στον ορίζοντα και έχουν άπειρη απόσταση από την κάμερα θα υπερνικούσαν και η τελική εικόνα θα ήταν μαύρη.

Οι παραγόμενες εικόνες βάθους λειτουργούν ως συνθήκες ελέγχου για την παραγωγή αληθοφανών εικόνων των αντικειμένων στις οποίες τηρούνται οι παράμετροι της σκηνής. Στο Σχήμα 3.5 παρατίθενται ενδεικτικές εικόνες βάθους των αντικειμένων.

Για κάθε εικόνα, υπολογίζονται και εξάγονται τα πλαίσια οριοθέτησης των αντικειμένων εντός αυτών, έτσι ώστε να χρησιμοποιηθούν για την εκπαίδευση μοντέλων ανίχνευσης αντικειμένων. Το βήμα αυτό είναι εξαιρετικά σημαντικό καθώς εξαλείφει την ανάγκη ανθρώπινης εργασίας και χειρονακτικής επισήμανσης κάθε εικόνας, μια πληκτική και χρονοβόρα δραστηριότητα η οποία απαγορεύει την αυτόματη και αποδοτική παραγωγή δεδομένων για εφαρμογές ανίχνευσης αντικειμένων σε μεγάλη κλίμακα. Η αυτόματη επισήμανση των εικόνων επιτυγχάνεται μέσω του υπολογισμού των ακραίων τιμών των συντεταγμένων των σημείων των αντικειμένων εντός της εικόνας και επιτρέπει την αυτοματοποίηση της διαδικασίας παραγωγής ολόκληρων συνόλων δεδομένων απαιτώντας μόνο την προϋπαρξη πρόχειρων τριδιάστατων μοντέλων.



Σχήμα 3.5. Εξαγωγή εικόνων βάθους και αυτόματη επισήμανση μέσω παραγωγής πλαισίων οριοθέτησης για εφαρμογές ανίχνευσης αντικειμένων.

3.2.4. Παραγωγή Εικόνων με το Stable Diffusion

Η έλλειψη δεδομένων για την εκπαίδευση μοντέλων μηχανικής μάθησης αποτελεί σημαντικό ζήτημα σε πολλούς επιστημονικούς κλάδους. Η δημιουργία συνθετικών δεδομένων για την επαύξηση συνόλων είναι μέχρι στιγμής μια από τις δημοφιλέστερες μεθόδους αντιμετώπισης της πρόκλησης αυτής. Ακολουθώντας την έλευση των μοντέλων διάχυσης, μερικές προσπάθειες έχουν γίνει για την παραγωγή συνθετικών συνόλων δεδομένων για χρήση σε εφαρμογές μηχανικής όρασης. Οι περισσότερες προσεγγίσεις [76] [77] [78] [79] χρησιμοποιούν σαν βάση κάποιο μοντέλο διάχυσης, το οποίο έχει προεκπαιδευτεί σε γενικά σύνολα δεδομένων και προσαρμόζεται με επιπλέον εκπαίδευση σε εξειδικευμένα σύνολα δεδομένων που σχετίζονται την εκάστοτε εφαρμογή. Άλλες, πιο απλές προσεγγίσεις, παράγουν εικόνες συνήθων αντικειμένων χρησιμοποιώντας μόνο το βασικό μοντέλο χωρίς επιπλέον εκπαίδευση [80]. Τέλος, περισσότερες προσπάθειες παράγουν συνθετικά δεδομένα για εφαρμογές ταξινόμησης και εντοπισμού ανωμαλιών, καθώς η ελεγχόμενη σύνθεση των εικόνων για τον ακριβή προσδιορισμό της τοποθεσίας των αντικειμένων εντός αυτών είναι πολύπλοκη.

Η αναζήτηση της σχετικής βιβλιογραφίας δείχνει πως η παρούσα μελέτη αποτελεί την πρώτη προσπάθεια δημιουργίας συνθετικών δεδομένων με μεθόδους διάχυσης, δίχως επιπλέον εκπαίδευση σε ειδικά δεδομένα και ελέγχοντας τη θέση και τον προσανατολισμό των αντικειμένων εντός της εικόνας, για εφαρμογή ανίχνευσης αντικειμένων.

Οι εικόνες βάθους που παράχθηκαν στην προηγούμενη φάση χρησιμοποιούνται, μέσω του ControlNet, ως συνθήκες ελέγχου για την σύνθεση εικόνων με την έκδοση v1.5 του Stable Diffusion η οποία έχει εκπαιδευτεί στο σύνολο δεδομένων LAION-5B [81].

Για κάθε κατηγορία αντικειμένου χρησιμοποιήθηκε λεπτομερώς σχεδιασμένη περιγραφή κειμένου για τη διασφάλιση της βέλτιστης ποιότητας και μεταβλητότητας των παραγόμενων εικόνων. Εντός της περιγραφής προσδιορίζονται η κατηγορία του αντικειμένου, καθώς και λεπτομερή χαρακτηριστικά αυτού, λεπτομέρειες για τον χώρο εντός του οποίου βρίσκεται το αντικείμενο και διάφορες επιπλέον παράμετροι που εστιάζουν το μοντέλο με αποτέλεσμα τη σύνθεση εικόνων υψηλής ποιότητας. Παράλληλα, χρησιμοποιούνται και αρνητικές περιγραφές κειμένου (negative prompts), οι οποίες προσδιορίζουν ανεπιθύμητα χαρακτηριστικά της εικόνας, δίνοντας περαιτέρω έλεγχο στη διαδικασία σύνθεσης και βελτιώνοντας επιπλέον την ποιότητα των παραγόμενων εικόνων. Στον Πίνακα 3.1 βρίσκονται συγκεντρωμένες οι περιγραφές που χρησιμοποιήθηκαν για την παραγωγή των εικόνων μηχανημάτων και εξοπλισμού. Επιπλέον περιγραφές που χρησιμοποιήθηκαν για την παραγωγή εικόνων εργαζομένων καθώς και παραδείγματα συνθετικών εικόνων αυτών μπορούν να βρεθούν στο Παράρτημα Α.

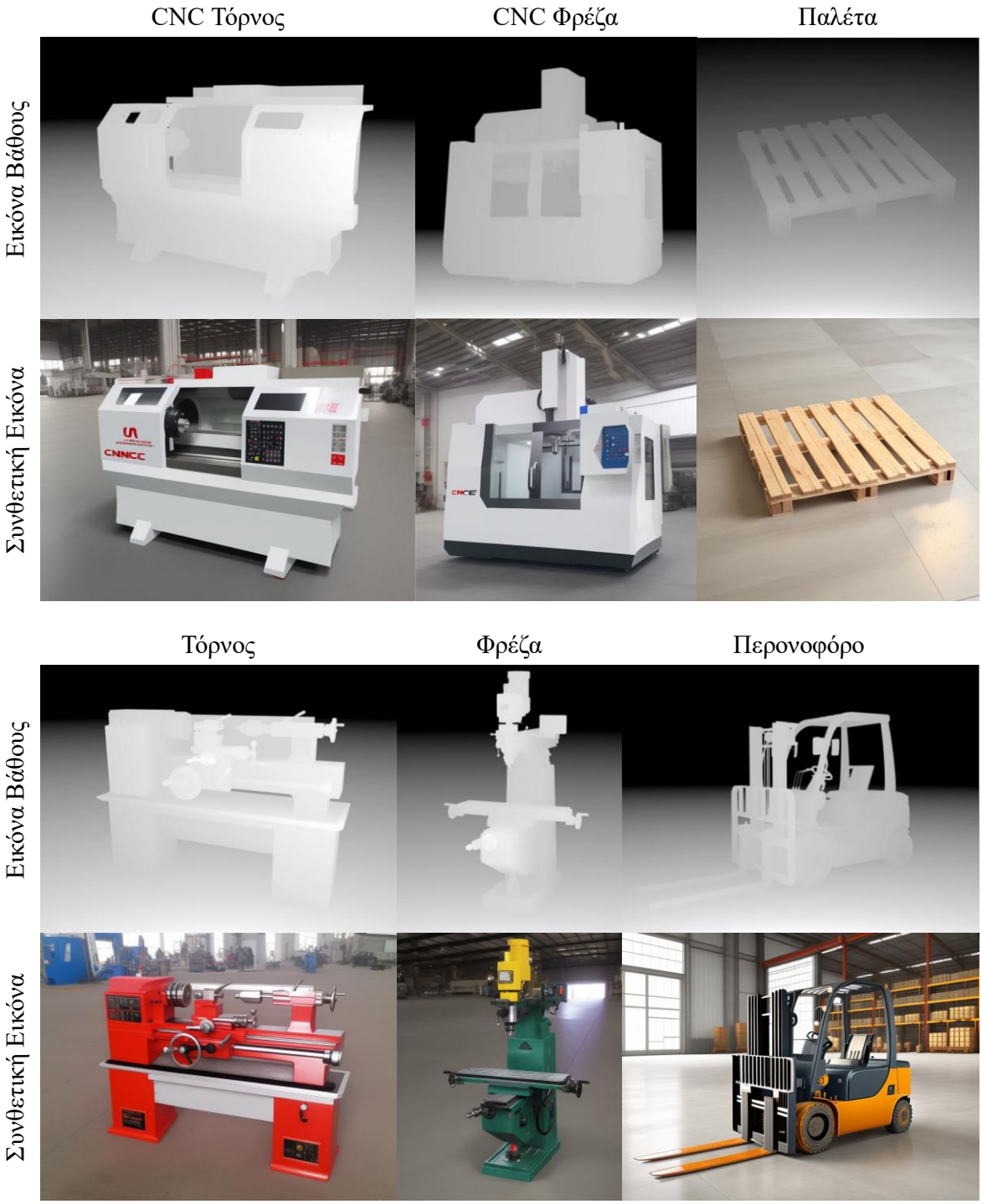
Πίνακας 3.1

Οι περιγραφές κειμένου που χρησιμοποιήθηκαν για την παραγωγή εικόνων με το Stable Diffusion.

Κατηγορία Αντικειμένου	Περιγραφή Κειμένου	Αρνητική Περιγραφή Κειμένου
CNC Τόρνος	*((cnc lathe)) inside a (factory), ((control panel)), best quality, masterpiece, perfect geometry, straight lines, clear edges, realistic, fine detail, [[Haas CNC lathe, Okuma CNC lathe]], (windows in background), bright room*	*depth of field, blurry, out of focus, ((bad geometry)), poorly drawn, ((duplicate, multiple machines, two machines)), person, human, worker, watercolor, blurry, painting, smudge, repeating pattern, extension, (dark room), watermark, [[getty images]], [[grayscale, b&w]], ((((flat background, studio backdrop, empty room))), [outside, outdoors, sky, nature]*
CNC Φρέζα	*((cnc vertical mill)) inside a (factory), ((control panel)), best quality, masterpiece, perfect geometry, straight lines, clear edges, realistic, fine detail, [[Haas CNC mill, Okuma CNC machining center]], (windows in background), bright room*	*depth of field, blurry, out of focus, ((bad geometry)), poorly drawn, ((duplicate, multiple machines, two machines)), person, human, worker, watercolor, blurry, painting, smudge, repeating pattern, extension, (dark room), watermark, [[getty images]], [[grayscale, b&w]], ((((flat background, studio backdrop, empty room))), [outside, outdoors, sky, nature]*
Περονοφόρο	*((forklift)) inside factory, [indoors], best quality, masterpiece, (perfect geometry, straight lines), clear edges, realistic, fine detail, ((well lit))*	*depth of field, blurry background, out of focus, motion blur, moving vehicle, (((dim lighting, rim lighting, harsh shadows, dark room)), ((bad geometry))), (duplicate, multiple forklifts, two forklifts, forklifts in background), poorly drawn, watercolor, blurry, painting, smudge, repeating pattern, extension, logo, watermark, getty images, [grayscale, b&w], outdoors, outside, parking lot, sky, sunset, windows, light coming from windows, [art, painting, illustration, cartoon, render]*
Τόρνος	*[[Summit Machine Tool]] ((manual lathe)) inside a factory, chuck, ((metal hand wheel)), best quality, masterpiece, (perfect geometry, straight lines)*	*depth of field, blurry, out of focus, [smudge, watercolor painting], watermark, logo, ((bad geometry)), poorly drawn, (((duplicate, multiple machines))), person, human, worker, ((studio backdrop, flat background)), [[old factory]], photo taken with flash, flash photography, [[[[green machine]]]], outside, outdoors*
Φρέζα	*((milling machine)) inside a factory, best quality, masterpiece, (perfect geometry, straight lines), well lit, [[[Bridgeport, Baileigh Industrial, Bolton Tools, Erie Tools, Palmgren, Central Machinery]]]*	*depth of field, blurry, out of focus, dim lighting, ((bad geometry)), poorly drawn, (((duplicate, multiple machines))), person, human, worker, [outside, outdoors, sky, parking lot], white machine, desaturated machine, studio backdrop*
Παλέτα	*wooden pallet on ((concrete floor))*	*desaturated image*

Παρατηρείται πως οι αρνητικές περιγραφές είναι εκτενέστερες από τις θετικές. Οι αρνητικές περιγραφές προέκυψαν μέσω πειραματισμού όπου, μέσω μιας επαναληπτικής διαδικασίας, διάφορα ανεπιθύμητα χαρακτηριστικά αποκλείονταν από τη σύνθεση της εικόνας. Για όλες τις εικόνες χρησιμοποιήθηκε ο δειγματολήπτης διάχυσης (sampler) Euler a [82] με 15 βήματα δειγματοληψίας.

Στο Σχήμα 3.6 παρατίθενται παραδείγματα συνθετικών εικόνων που παράχθηκαν για κάθε κατηγορία αντικειμένου.



Σχήμα 3.6. Παραδείγματα συνθετικών εικόνων κάθε κατηγορίας αντικειμένου.

Συνολικά, παράχθηκαν 1000 συνθετικές εικόνες για κάθε κατηγορία μηχανημάτων και εξοπλισμού. Οι εικόνες παράχθηκαν σε παρτίδες των δύο με χρόνο παραγωγής κάθε παρτίδας περίπου 7 δευτερόλεπτων χρησιμοποιώντας κάρτα γραφικών RTX 3070 με 8GB VRAM. Οι συνθετικές εικόνες χαρακτηρίζονται από μεγάλο βαθμό αληθοφάνειας και τηρούν πιστά τα γεωμετρικά χαρακτηριστικά και τον προσανατολισμό των αντικειμένων εντός των εικονικών σκηνών. Η διατήρηση της διάταξης της σκηνής είναι απαραίτητη έτσι ώστε το παραγόμενο σύνολο δεδομένων να διατηρήσει τη μεταβλητότητα που εισάχθηκε κατά την τυχαιοποίηση της σκηνής. Συνεπώς, οι συνθετικές εικόνες απεικονίζουν τα αντικείμενα από διάφορες γωνίες με μεταβαλλόμενη προοπτική και αναλογίες.

3.3. Σύνθεση Σκηνής

Στο προηγούμενο στάδιο, σε κάθε συνθετική εικόνα που παράχθηκε απεικονίζεται μόνο ένα αντικείμενο, με σταθερή σχετική κλίμακα εντός της εικόνας, εν απουσία άλλων αντικείμενων ή ανθρώπων. Η προσέγγιση αυτή, αν και παράγει εικόνες όπου τα αντικείμενα αποτυπώνονται σε υψηλή ανάλυση, αξιοποιώντας ολόκληρη την έκταση της εικόνας, δεν προσφέρει μεταβλητότητα ως προς το σχετικό μέγεθος και τη θέση των αντικειμένων εντός αυτής. Επιπλέον, αποκλείει τη συνύπαρξη πολλαπλών αντικειμένων εντός της ίδιας εικόνας, η οποία είναι απαραίτητη για την εκπαίδευση ενός εύρωστου μοντέλου ανίχνευσης αντικειμένων με υψηλή ικανότητα γενίκευσης. Η λύση σε αυτόν τον περιορισμό συνίσταται στη δημιουργία συνθετικών σκηνών μέσω επεξεργασίας εικόνων.

Στην παρούσα και τελευταία φάση της δημιουργίας του συνθετικού συνόλου δεδομένων, οι συνθετικές εικόνες που παράχθηκαν ενσωματώνονται σε νέες εικόνες σε τυχαίες θέσεις και μεγέθη σχηματίζοντας συνθέσεις οι οποίες περιλαμβάνουν μαζί πολλαπλά αντικείμενα και ανθρώπους. Η προσέγγιση αυτή υλοποιείται μέσω συμβατικών μεθόδων επεξεργασίας εικόνας και δεν απαιτεί πολλούς υπολογιστικούς πόρους. Η διαδικασία περιλαμβάνει την επιλογή τυχαίων εικόνων αντικειμένων και ανθρώπων και την επικόλληση τους σε τυχαία μεγέθη και θέσεις μπροστά από ένα τυχαία επιλεγμένο συνθετικό φόντο. Οι εικόνες φόντου παράχθηκαν χρησιμοποιώντας το Stable Diffusion, με περιγραφές κειμένου εσωτερικού χώρου βιομηχανικών εγκαταστάσεων ή εργοταξίων. Στο Σχήμα 3.7 φαίνονται μερικά παραδείγματα εικόνων που χρησιμοποιήθηκαν ως φόντο καθώς και οι περιγραφές κειμένου που χρησιμοποιήθηκαν για την παραγωγή τους.

Επίπεδο Παραγωγής Εργοτάξιο

Σχήμα 3.7. Παραδείγματα συνθετικών εικόνων φόντων.

Πίνακας 3.2

Οι περιγραφές κειμένου που χρησιμοποιήθηκαν για την παραγωγή εικόνων φόντου με το Stable Diffusion.

Επίπεδο Παραγωγής	*an empty industrial warehouse, best quality, masterpiece, (perfect geometry, straight lines), clear edges, realistic, fine detail, detailed background, ((well lit)), (wide shot), shot from a distance, perspective*	*depth of field, blurry, out of focus, dim lighting, ((bad geometry)), poorly drawn, watercolor, blurry, painting, smudge, repeating pattern, extension, logo, dark room, watermark, getty images, [grayscale, b&w], concept art, flat background, outside*
Εργοτάξιο	*a construction site, concrete, best quality, masterpiece, perfect geometry, realistic, realism, HD, high resolution, trending on artstation, realistic textures, photography, POV, [[[[[[cranes]]]]]], 4k textures, [overcast]*	*((humans, people, person, workers, hard hat, safety vest)), (vehicle, excavator), blurry, bad drawing, smudge, watercolor, artwork, paint, bad geometry, deformed, 3d render, dry, desert, [abandoned], futuristic, sunset, ((aerial view, drone shot)), city, wide angle, sketch, child drawing, b&w filter, grayscale, sunny, rain*

Η αναλυτική διαδικασία δημιουργίας κάθε συνθετικής σκηνής συνίσταται στα κάτωθι βήματα:

1. Επιλογή εικόνων – Αρχικά, γίνεται τυχαία επιλογή μιας κύριας εικόνας αντικειμένου. Για κάθε κύρια εικόνα αντικειμένου παράγονται τέσσερεις διαφορετικές παραλλαγές σκηνών. Σε κάθε παραλλαγή γίνεται τυχαία επιλογή 3 εικόνων κενών χώρων και μιας συνοδευτικής εικόνας αντικειμένου από τα αντίστοιχα σύνολα εικόνων φόντου και μηχανημάτων ή εξοπλισμού. Μια από τις κενές εικόνες αποτελεί το φόντο της συνθετικής σκηνής.

2. Επιλογή μεγέθους και θέσης – Για κάθε παραλλαγή, υπολογίζονται τυχαία ένας συντελεστής μεγέθυνσης s και η θέση $\Delta x, \Delta y$ της πάνω αριστερά γωνίας της κύριας εικόνας αντικειμένου εντός της σύνθεσης, με το μέγεθος να ελαττώνεται από εκδοχή σε εκδοχή. Αυτό γίνεται έτσι ώστε να δημιουργηθούν εικόνες που απεικονίζουν το ίδιο αντικείμενο σε διάφορες κλίμακες, επαυξάνοντας σημαντικά το παραγόμενο σύνολο δεδομένων. Η κύρια εικόνα τοποθετείται στην προσδιορισμένη θέση επάνω στο φόντο.

3. Αποφυγή επικάλυψης – Στη συνέχεια, εξαιρώντας την πρώτη παραλλαγή, όπου το μέγεθος του αντικειμένου ταυτίζεται με αυτό του φόντου, γίνεται προσπάθεια τοποθέτησης επιπλέον εικόνων εντός της σκηνής, σε τυχαίες θέσεις και μεγέθη. Η τοποθέτησή τους γίνεται ελέγχοντας για επικαλύψεις μεταξύ αυτών και των εικόνων αντικειμένων που είχαν τοποθετηθεί προηγουμένως ως εξής:

$$\text{Επικάλυψη}(Y, \Pi_i) = \frac{A(Y \cap \Pi_i)}{A(\Pi_i)} \tag{15}$$

Όπου Y η υποψήφια εικόνα, Π_i η i-οστή προϋπάρχουσα εικόνα και A το εμβαδό. Οι υποψήφιες εικόνες με επικάλυψη με κάθε προηγούμενη εικόνα μικρότερη από μια τιμή κατωφλίου ίση με 5% γίνεται δεκτή. Η επιλογή τυχαίου μεγέθους και θέσης επαναλαμβάνεται για έναν πεπερασμένο αριθμό προσπαθειών έως ότου ικανοποιηθεί το κριτήριο αυτό.

4. Τοποθέτηση εικόνας – Η κύρια εικόνα αντικειμένου και κάθε εικόνα που ικανοποιεί το κριτήριο μη επικάλυψης τοποθετούνται εντός της σκηνής και, αν πρόκειται για εικόνα αντικειμένου, υπολογίζεται το νέο πλαίσιο οριοθέτησης αυτού μέσω κατάλληλου μετασχηματισμού:

$$\begin{aligned} x^{\text{new}} &= \frac{x^{\text{original}} \cdot (s \cdot w) + \Delta x}{W} & w^{\text{new}} &= \frac{w^{\text{original}} \cdot (s \cdot w)}{W} \\ y^{\text{new}} &= \frac{y^{\text{original}} \cdot (s \cdot h) + \Delta y}{H} & h^{\text{new}} &= \frac{h^{\text{original}} \cdot (s \cdot h)}{H} \end{aligned} \tag{16}$$

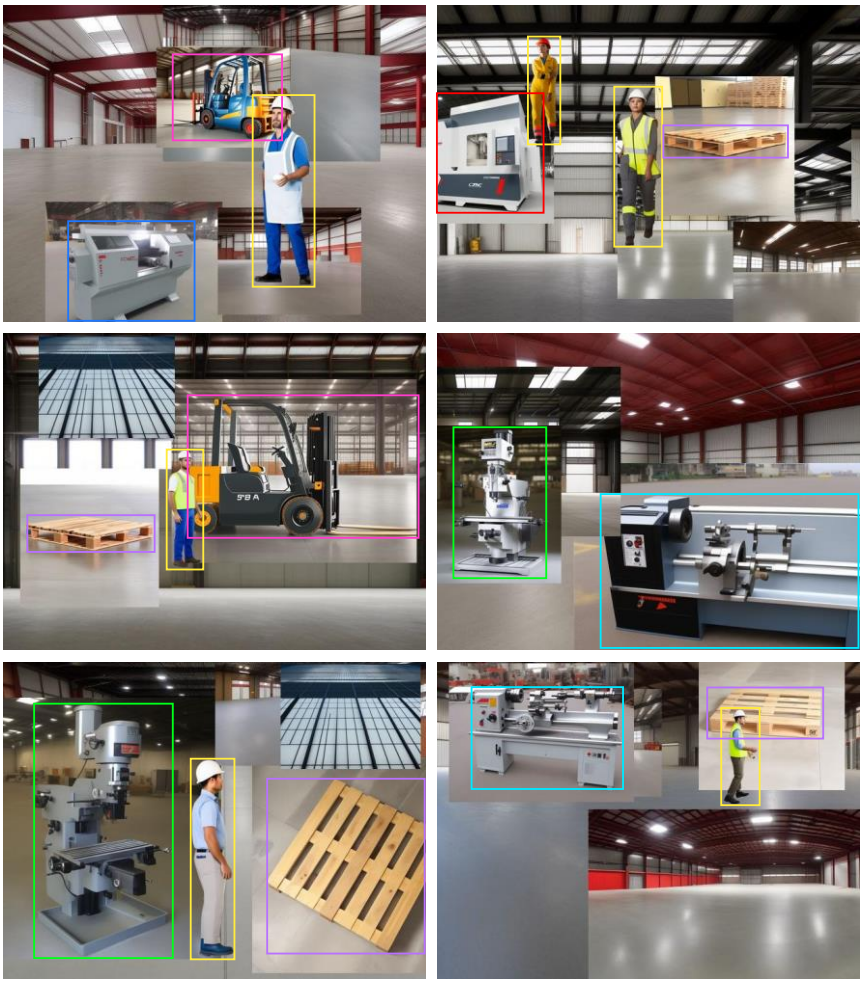
όπου $(x, y, w, h)^i$ είναι το πλαίσιο οριοθέτησης σε μορφή YOLO, $w \times h$ οι διαστάσεις της εικόνας προς τοποθέτηση και $W \times H$ οι διαστάσεις της εικόνας φόντου.

5. Τοποθέτηση ανθρώπων – Σε αυτό το βήμα, επιλέγονται δύο τυχαίες εικόνες ανθρώπων και τοποθετούνται στη σύνθεση σε τυχαίες θέσεις και μεγέθη, επικαλύπτοντας τα υπάρχοντα αντικείμενα, παρόμοια με τους flying distractors των [83]. Η μερική επικάλυψη των αντικειμένων με ανθρώπους λειτουργεί σαν το dropout για νευρωνικά δίκτυα, καθώς, κατά την εκπαίδευση, αναγκάζει το μοντέλο να αγνοήσει τοπικά μοτίβα και ενισχύει την ικανότητα γενίκευσής του.

6. Εξαγωγή εικόνας – Τέλος, οι παραγόμενες συνθετικές σκηνές αποθηκεύονται σαν νέες εικόνες και μαζί τους εξάγονται τα νέα πλαίσια οριοθέτησης.

Η παραπάνω διαδικασία είναι πλήρως αυτοματοποιημένη και μπορεί να εφαρμοσθεί για την παραγωγή οποιουδήποτε επιθυμητού αριθμού συνδυασμών σκηνών, με αποτέλεσμα την δημιουργία μεγάλων και πλουσίων συνόλων δεδομένων με μεγάλη μεταβλητότητα. Το τελικό σύνολο δεδομένων αποτελείται από 22000 εικόνες.

Στο Σχήμα 3.8 μπορούν να φανούν παραδείγματα εικόνων του συνθετικού συνόλου δεδομένων που παράγονται με την παραπάνω μέθοδο.



Σχήμα 3.8. Παραδείγματα εικόνων του συνθετικού συνόλου δεδομένων με τα πλαίσια οριοθέτησης τους.

Παρατηρείται πως οι παραγόμενες συνθετικές σκηνές δεν διατηρούν τον ίδιο βαθμό αληθοφάνειας με τις μεμονωμένες εικόνες που παράχθηκαν στην προηγούμενη φάση. Αυτό οφείλεται κυρίως σε δύο παράγοντες.

Ο πρώτος έγκειται στο ότι η τυχαία μεγέθυνση και τοποθέτηση των επιμέρους εικόνων δεν εξασφαλίζει λογική συνοχή εντός της εικόνας. Για παράδειγμα, λόγω της απουσίας κάποιου περιορισμού στην τοποθέτηση των εικόνων των ανθρώπων στο φαινομενικό έδαφος της σκηνής, σε πολλές σκηνές οι άνθρωποι φαίνονται να βρίσκονται στον αέρα, όπως φαίνεται στο Σχήμα 3.9. Αντίστοιχα, λόγω της τυχαίας μεγέθυνσης τους, σε άλλες εικόνες οι άνθρωποι βρίσκονται φαινομενικά κοντά στην κάμερα, ωστόσο το σχετικό μέγεθός τους στην εικόνα είναι πολύ μικρό. Το αντίστροφο φαινόμενο συμβαίνει επίσης. Γενικά, δεν υπάρχει κάποια συσχέτιση μεταξύ των μεγεθών ή των θέσεων των διάφορων αντικειμένων που συνυπάρχουν εντός μιας σκηνής.

Ο δεύτερος παράγοντας έγκειται στην τοποθέτηση των εικόνων των μηχανημάτων στη σκηνή δίχως κάποια επεξεργασία ή αφαίρεση του φόντου τους. Το γεγονός αυτό καθιστά εμφανή τα όρια της εικόνας εντός της σκηνής, όπως φαίνεται στο Σχήμα 3.9, πράγμα που εξαλείφει οποιαδήποτε συνολική αληθοφάνεια της σκηνής. Ωστόσο, η πιο σημαντική επίπτωση της παρουσίας των ακμών των επιμέρους εικόνων στην σκηνή αφορά την εκπαίδευση του μοντέλου ανίχνευσης αντικειμένων. Χαρακτηριστικά όπως καθαρά οριζόντιες ή κατακόρυφες γραμμές εντός μιας εικόνας είναι πολύ εμφανή και ανιχνεύονται ακόμα και από απλοϊκά εργαλεία μηχανικής όρασης όπως τα φίλτρα. Λόγω της απλότητας τέτοιων χαρακτηριστικών, το δίκτυο τείνει να συγκλίνει προς την εκμάθηση αυτών και όχι των αληθινών χαρακτηριστικών των αντικειμένων, για την πρόβλεψη της θέσης και της κατηγορίας τους. Κατά συνέπεια, το δίκτυο, έχοντας εγκλωβιστεί σε τοπικό ακρότατο, εμφανίζει χαμηλή ακρίβεια σε αληθινές εικόνες στις οποίες απουσιάζουν οι ατέλειες (artifacts) αυτές.

Αναφορικά με τον πρώτο παράγοντα, αντίθετα στην κοινή λογική, η έλλειψη λογικής συνοχής δεν φέρει επιπτώσεις στην ακρίβεια μοντέλων ανίχνευσης αντικειμένων [75] [84] [85]. Αντιθέτως, η σκόπιμη εισαγωγή μη αληθοφανών χαρακτηριστικών εντός της εικόνας μέσω τροποποίησης της συνθετικής σκηνής αναγκάζει το δίκτυο να μάθει τα ουσιώδη χαρακτηριστικά των υπό μελέτη αντικειμένων, αυξάνοντας την επίδοσή του [83]. Συμπεραίνεται ότι η έλλειψη σχετικών με την εκάστοτε εφαρμογή δεδομένων μπορεί να παρακαμφθεί με τη σύνθεση και χρήση συνθετικών δεδομένων μεγάλης μεταβλητότητας για την εκπαίδευση μοντέλων, με αποτέλεσμα την ικανότητα γενίκευσης σε αληθινά σύνολα δεδομένων.

Για την αντιμετώπιση του δεύτερου παράγοντα, γίνεται χρήση κενών εικόνων, που απεικονίζουν άδεια επίπεδα παραγωγής ή εργοτάξια, που δρουν ως αρνητικά δείγματα για την εκπαίδευση του μοντέλου. Η τοποθέτηση κενών εικόνων εντός της συνθετικής σκηνής, όπως φαίνεται στο Σχήμα 3.9, εισάγει εμφανή περιγράμματα, όπως και οι εικόνες των αντικειμένων, δίχως να δίνει κίνητρο στο δίκτυο να τα ερμηνεύει σαν χρήσιμα χαρακτηριστικά. Το πλήθος των κενών εικόνων ισοσταθμίζεται με αυτό των εικόνων αντικειμένων έτσι ώστε να αντιπροσωπεύονται σε ίδιο βαθμό. Ενσωματώνοντας τις κενές εικόνες, η παρουσία ορατών ακμών εικόνων παύει να συσχετίζεται με την παρουσία κάποιου αντικειμένου εντός του περιγράμματος της εικόνας. Συνεπώς, η προσβολή της αληθοφάνειας της εικόνας λόγω των ορατών ακμών παύει να αποτελεί αιτία για χαμηλή απόδοση του μοντέλου.

Σχήμα 3.9. Οι ατέλειες που εμφανίζονται στις συνθετικές εικόνες.

Γενικότερα, η καθολική αληθοφάνεια, αν και επιθυμητό χαρακτηριστικό συνθετικών εικόνων, δεν είναι πάντα απαραίτητη για την εκπαίδευση μοντέλων ανίχνευσης αντικειμένων. Οι [86] χρησιμοποιούν τον όρο «τμηματική αληθοφάνεια» (patch-level realism) για να χαρακτηρίσουν το φαινόμενο όπου τμήματα μιας συνθετικής εικόνας μπορεί να φαίνονται αληθινά ενώ η ολική σύνθεση φαίνεται τεχνητή. Υποστηρίζεται ότι τα μοντέλα ανίχνευσης αντικειμένων εστιάζουν περισσότερο σε τοπικά χαρακτηριστικά παρά στην καθολική διάταξη των εικόνων για τον εντοπισμό και την ταξινόμηση των αντικειμένων εντός αυτών.

Συμπεραίνοντας, οι συνθετικές εικόνες μηχανημάτων, εξοπλισμού, ανθρώπων και παρασκηνίων που παράγονται με το Stable Diffusion ενσωματώνονται σε συνθετικές σκηνές μέσω μιας αυτόματης διαδικασίας. Η αυτοματοποίηση της παραγωγής συνθετικών σκηνών επιτρέπει τη δημιουργία πολύ μεγάλου πλήθους τελικών εικονων που εμπεριέχουν πολλά διαφορετικά αντικείμενα. Η συνύπαρξη των αντικειμένων εντός των εικόνων και η μεταβλητότητα που εισάγεται μέσω της τυχαιοποίησης της θέσης και τους μεγέθους τους αποτελούν ικανές ιδιότητες για τη συλλογή ενός ποιοτικού συνόλου δεδομένων. Επιπλέον, η μέθοδος παραγωγής των συνθετικών σκηνών εισάγει στοιχεία που βλάπτουν την αληθοφάνεια τους. Ωστόσο, έχει αποδειχθεί πως η έλλειψη αληθοφάνειας δεν μειώνει την ακρίβεια ανίχνευσης του μοντέλου, ή απαιτεί κατάλληλη αντιμετώπιση έτσι ώστε να μην αποτελεί επιβλαβή παράγοντα για αυτή. Τέλος, η προτεινόμενη μέθοδος συμβαδίζει με προσεγγίσεις που συναντώνται στη βιβλιογραφία και εμφανίζουν βελτίωση της ακρίβειας.

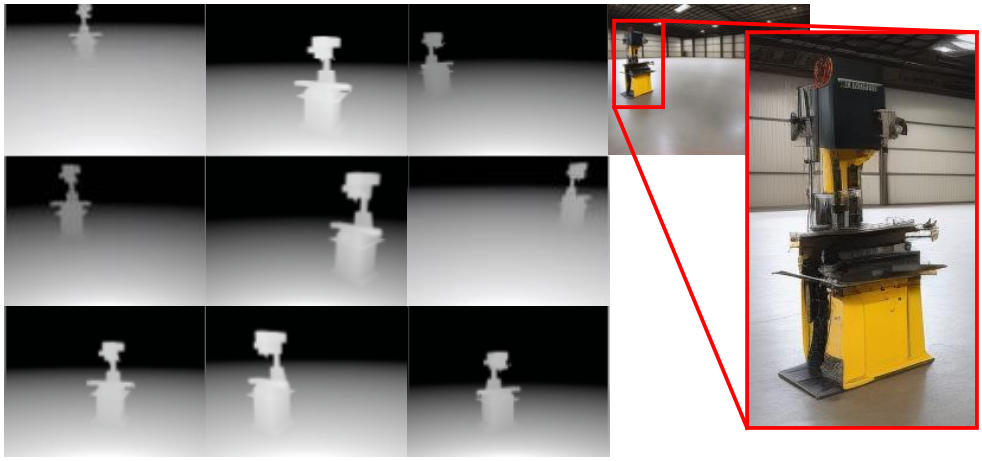
3.4. Απορριφθείσες Προσεγγίσεις

Η επικρατέστερη μέθοδος σύνθεσης του συνόλου δεδομένων που αναλύθηκε παραπάνω και εφαρμόζεται στην παρούσα μελέτη, απαρτίζεται από δύο διακριτές φάσεις παραγωγής εικόνων αντικειμένων και εικόνων σκηνών. Όπως περιγράφηκε στην υποενότητα 3.2.2, οι εικονικές σκηνές που παράγονται εντός του Blender και από τις οποίες εξάγονται οι εικόνες βάθους, διατηρούν το αντικείμενο ενδιαφέροντος στο κέντρο της εικόνας με σταθερό σχετικό μέγεθος, καταλαμβάνοντας όλη την έκταση της εικόνας. Το βήμα αυτό εισάγει μεταβλητότητα ως προς την οπτική γωνία, την προοπτική και τη γεωμετρία του αντικειμένου αλλά όχι ως προς το μέγεθος ή τη θέση εντός της εικόνας. Η μεταβλητότητα ως προς αυτές τις παραμέτρους εισάγεται αργότερα κατά τη δημιουργία συνθετικών σκηνών, με τοποθέτηση επιμέρους εικόνων μπροστά από ένα άδειο παρασκήνιο. Ωστόσο, όπως αναλύθηκε, η προσέγγιση αυτή εισάγει ατέλειες στη μορφή ορατών ακμών. Εξαιτίας αυτού, η αρχική πειραματική προσέγγιση απέφευγε την διακριτή σύνθεση σκηνής σε δεύτερο στάδιο, ακολουθώντας την παραγωγή εικόνων από τις εικόνες βάθους.

Η αρχική προσέγγιση της δημιουργίας του συνθετικού συνόλου δεδομένων περιλάμβανε την ενιαία τυχαιοποίηση όλων των παραμέτρων κατά την παραγωγή της εικονικής σκηνής στο τριδιάστατο περιβάλλον. Πέρα από την τυχαιοποίηση του αζιμουθίου και της ανύψωσης της εικονικής κάμερας και της απόστασης της από το αντικείμενο, μεταβαλλόντουσαν επίσης και η οριζόντια και η κατακόρυφη κλίση αυτής ενώ η εστιακή απόσταση παρέμενε σταθερή, με αποτέλεσμα την μεταβολή της θέσης και του μεγέθους του αντικειμένου μέσα στην εικόνα. Συνεπώς, οι εικόνες βάθους των σκηνών απεικόνιζαν τα αντικείμενα με τέτοιον τρόπο έτσι ώστε η επιθυμητή μεταβλητότητα να ήταν παρούσα κατά την παραγωγή των εικόνων μέσω ControlNet και Stable Diffusion, χωρίς την ανάγκη περαιτέρω επεξεργασίας. Το τελευταίο στάδιο της αρχικής προσέγγισης περιλάμβανε μόνο την ενσωμάτωση των εικόνων ανθρώπων απευθείας εντός των εικόνων αντικειμένων.

Η χρήση εικόνων βάθους οι οποίες αποτυπώνουν τα αντικείμενα σε μεγάλες αποστάσεις από την κάμερα και σε θέση με απόκλιση από το κέντρο της εικόνας για την ελεγχόμενη παραγωγή των συνθετικών εικόνων μέσω ControlNet και Stable Diffusion, έχει ανεπιθύμητες επιπτώσεις στην ποιότητα των παραγόμενων εικόνων. Συγκεκριμένα, η χρήση του ControlNet για την τοποθέτηση αντικειμένων στη σύνθεση με μικρά μεγέθη και σε ακραίες θέσεις εντός της εικόνας, έχει ως αποτέλεσμα την δημιουργία εικόνων όπου τα αντικείμενα αυτά αποτυπώνονται με χαμηλή σημασιολογική ποιότητα. Το φαινόμενο αυτό επικρατεί ακόμη και αν η ανάλυση της εικόνας στην περιοχή που καταλαμβάνουν τα αντικείμενα είναι επαρκής για την ορθή απεικόνισή τους. Στο

Σχήμα 3.10 μπορεί να παρατηρηθεί η χαμηλή σημασιολογική ποιότητα της παραγόμενης εικόνας ενός αντικειμένου που πληροί τις παραπάνω προϋποθέσεις.



Σχήμα 3.10. Η ενιαία τυχαιοποίηση της σκηνής και η επίπτωση στην σημασιολογική ποιότητα της εικόνας που αυτή επιφέρει. Το αντικείμενο που απεικονίζεται είναι μια συμβατική φρέζα.

Η πτώση στην ποιότητα της σύνθεσης έγκειται σε εγγενείς περιορισμούς που αφορούν τα μοντέλα λανθάνουσας διάχυσης και τον έλεγχό τους με εργαλεία εισαγωγής επιπλέον συνθηκών, όπως το ControlNet. Όπως αναλύθηκε στο Κεφάλαιο 2, η αρχή λειτουργίας των μοντέλων διάχυσης βασίζεται στη διαδοχική αφαίρεση θορύβου από μια εικόνα, με σκοπό την ανάκτηση ή τη δημιουργία πληροφορίας. Η αλλοίωση των εικόνων με υψίσυχνο θόρυβο εξαλείφει μικρές λεπτομέρειες, καθιστώντας την σύνθεση εικόνων που αποτυπώνουν μικρά χαρακτηριστικά δύσκολη. Επιπλέον, η διάχυση λαμβάνει χώρα εντός του λανθάνοντα χώρου, ο οποίος συμπιέζει και κωδικοποιεί τις εικόνες αφαιρώντας υψίσυχνα χαρακτηριστικά. Τέλος, η εκπαίδευση των μοντέλων με εικόνες που αποτυπώνουν αντικείμενα σε μεγάλη κλίμακα δύναται να έχει περιορίσει την επίδοσή τους όσον αφορά την παραγωγή εικόνων μικρότερων αντικειμένων.

Οι περιορισμοί και οι προκλήσεις που αντιμετωπίζουν τα μοντέλα διάχυσης στη δημιουργία εικόνων που αποτυπώνουν αντικείμενα σε μικρότερες κλίμακες είναι ενδεικτικές του γεγονότος ότι αυτή η τεχνολογία βρίσκεται ακόμη σε πρώιμο στάδιο. Τα μοντέλα διάχυσης αντιπροσωπεύουν μια εξαιρετικά πρόσφατη πρόοδο στον τομέα των παραγωγικών μοντέλων μηχανικής μάθησης, συνεπώς, αντίστοιχα με κάθε αναδυόμενη τεχνολογία, απαιτείται χρόνος για περαιτέρω έρευνα με σκοπό την αντιμετώπιση των αρχικών αυτών εμποδίων. Με το πέρασμα του χρόνου, αναμένονται σημαντικές πρόοδοι και βελτιώσεις στα μοντέλα διάχυσης μέσω της συνεχούς έρευνας και καινοτομίας, που αναμφίβολα θα αντιμετωπίσουν αυτά τα ζητήματα και θα επιτρέψουν την πιο ευέλικτη και ακριβέστερη δημιουργία εικόνων.

Κεφάλαιο 4

Εκπαίδευση Μοντέλων Ανίχνευσης Αντικειμένων

4.1. Επιλογή Μοντέλου

4.1.1. Επιλογή Έκδοσης YOLO

Για την παρούσα μελέτη επιλέχθηκε η εκδοχή YOLOv4-Tiny-3L του YOLO, η οποία πρόκειται για μια μορφή της τέταρτης έκδοσης του μοντέλου με λιγότερες παραμέτρους και τρία επίπεδα YOLO. Η χρήση του μοντέλου αυτού απαιτεί λιγότερους υπολογιστικούς πόρους κατά την εκπαίδευση και αργότερα κατά τη λειτουργία του, επιτρέποντας την χρήση του σε συσκευές υπολογιστικής παρυφής (edge devices) μικρής υπολογιστικής ισχύος. Ενδεικτικά, για απλή ακρίβεια 32bit (float), το μέγεθος του αρχείου που περιλαμβάνει τις παραμέτρους της πλήρους έκδοσης είναι 246 MB ενώ το μέγεθος του αντίστοιχου αρχείου της εκδοχής Tiny-3L είναι περίπου 24 MB, δηλαδή μια τάξη μεγέθους μικρότερο.

4.1.2. Επιλογή Υπερπαραμέτρων Νευρωνικού Δικτύου

Η αρχιτεκτονική του νευρωνικού δικτύου του μοντέλου έχει δομικό ρόλο στην ικανότητα εξαγωγής σημαντικών χαρακτηριστικών από τις εικόνες εισόδου. Η αρχιτεκτονική του YOLOv4-Tiny-3L συνιστάται από διαδοχικά συνελικτικά επίπεδα και μηχανισμούς παράκαμψης (skip connections) που συμβάλλουν στην ανίχνευση αντικειμένων σε διάφορες κλίμακες, όπως αναλύθηκε στην ενότητα 2.2.4. Οι κύριες υπερπαράμετροι που καθορίζουν τη διάταξη του δικτύου είναι οι εξής:

- Διαστάσεις εισόδου – Οι διαστάσεις των εικόνων όταν εισέρχονται στο δίκτυο ορίζονται στα 512×384 pixels. Η ανάλυση αυτή είναι επαρκής για να εξασφαλίζει τη δυνατότητα ανίχνευσης αντικειμένων σε διάφορες κλίμακες, συμπεριλαμβανομένων των μικρών αντικειμένων ή αντικειμένων σε μακρινή απόσταση, αλλά όχι τόσο μεγάλη έτσι ώστε να απαιτούνται υπερβολικά πολλοί υπολογιστικοί πόροι για την εκπαίδευση και τη λειτουργία του μοντέλου. Επιπλέον, ο αριθμός καναλιών ορίζεται ίσος με 3, καθώς το δίκτυο καλείται να επεξεργαστεί έγχρωμες εικόνες.

- Αριθμός φίλτρων – Ο αριθμός των φίλτρων των συνελικτικών επιπέδων που προηγούνται των τριών επιπέδων YOLO του δικτύου, τα οποία είναι υπεύθυνα για την πρόβλεψη των πλαισίων οριοθέτησης και των κατηγοριών, επιλέγεται έτσι ώστε να είναι συμβατός με το πλήθος των κατηγοριών αντικειμένων που υπάρχουν στο συνθετικό σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση. Ο αριθμός φίλτρων υπολογίζεται ως

$$N_{\text{φίλτρων}} = (C + 5) \cdot 3 \tag{16}$$

όπου C το πλήθος των κατηγοριών αντικειμένων. Τελικά, για 7 κατηγορίες αντικειμένων, τα συνελικτικά επίπεδα που προηγούνται των επιπέδων YOLO έχουν 36 φίλτρα.

Η προσαρμογή των παραπάνω υπερπαραμέτρων είναι απαραίτητη για τη συμβατότητα του δικτύου με το παραγόμενο σύνολο δεδομένων. Παρακάτω, ρυθμίζονται επιπλέον παράμετροι που αφορούν την επίδοση του μοντέλου.

4.1.3. Επιλογή Λοιπών Παραμέτρων Δικτύου

Πέρα από τις υπερπαραμέτρους, οι οποίες αφορούν καθαρά την αρχιτεκτονική του δικτύου, επιλέγονται επιπλέον παράμετροι που σχετίζονται με τις ενεργοποιήσεις των νευρώνων, και διαδικασίες που αφορούν την εκπαίδευση και την λειτουργία του μοντέλου YOLO. Παρακάτω, αναλύονται αυτές οι παράμετροι:

- Συνάρτηση Ενεργοποίησης – Για τη διάδοση του σήματος από επίπεδο σε επίπεδο εντός του συνελικτικού δικτύου χρησιμοποιείται η συνάρτηση διαρρέουσας ανορθωμένης γραμμικής μονάδας (leaky rectified linear unit – leaky ReLU) με κλίση του αρνητικού τμήματος ίση με 0.1. Η επιλογή της συνάρτησης αυτής ενισχύει την ικανότητα εκμάθησης χαρακτηριστικών.

- Πλαίσια αγκύρωσης – Για κάθε επίπεδο YOLO χρησιμοποιούνται 3 πλαίσια αγκύρωσης (anchor boxes) για την κάλυψη ενός εύρους μεγέθους αντικειμένων. Συνολικά, 9 πλαίσια αγκύρωσης υπολογίζονται με βάση τα πλαίσια οριοθέτησης των εικόνων του συνόλου δεδομένων και της διάστασης της εισόδου:

  (51,159), (193,61), (128,115)
  (76,224), (117,278), (217,170)
  (316,192), (243,292), (382,280)

  Κάθε σειρά αντιστοιχεί σε ένα επίπεδο YOLO.

- Μέθοδοι επαύξησης δεδομένων – Όπως αναφέρθηκε στην ενότητα 2.2.4, το YOLOv4 έχει διάφορες μεθόδους επαύξησης εικόνας οι οποίες εφαρμόζονται κατά τη διάρκεια της εκπαίδευσης στις εικόνες που εισέρχονται στο δίκτυο. Η επαύξηση γίνεται μέσω της τροποποίησης της εικόνας με διάφορες μεθόδους. Για την εκπαίδευση του μοντέλου της τρέχουσας μελέτης, για κάθε βήμα της εκπαίδευσης γίνεται τυχαία επιλογή από την παρτίδα εικόνων που εισέρχεται στο δίκτυο και σε αυτές εφαρμόζονται οι κάτωθι τροποποιήσεις:

  - Περιστροφή – Οι εικόνες που επιλέγονται δέχονται μια τυχαία περιστροφή από -7.5° έως 7.5°.

  - Κορεσμός – Ο κορεσμός του χρώματος των εικόνων που επιλέγονται αυξομειώνεται έως ±50%.

  - Έκθεση – Η φωτεινότητα των εικόνων που επιλέγονται αυξομειώνεται έως ±50%.

  - Απόχρωση – Η απόχρωση του χρώματος των εικόνων που επιλέγονται αυξομειώνεται έως ±20%.

  Η επάυξηση λαμβάνει χώρα κατά τη διάρκεια της εκπαίδευσης και ενισχύει την ικανότητα γενίκευσης του δικτύου.

4.1.4. Επιλογή Παραμέτρων Αλγορίθμου Εκπαίδευσης

Η αποτελεσματική εκπαίδευση των μοντέλων μηχανικής μάθησης έγκειται στην κατάλληλη επιλογή των παραμέτρων και των μεθόδων εκπαίδευσης. Το YOLO επιχειρεί την ελαχιστοποίηση της συνάρτησης απώλειας χρησιμοποιώντας τη μέθοδο ορμής (momentum method), γνωστή και ως μέθοδος βαριάς μπάλας (heavy ball method). Η μέθοδος ορμής επαυξάνει τη μέθοδο στοχαστικής απότομης καθόδου (stochastic gradient descent – SGD) με έναν όρο ορμής [87]. Σε κάθε βήμα της εκπαίδευσης, η μέθοδος αυτή ανανεώνει τις μεταβλητές βελτιστοποίησης $\boldsymbol{w}$ χρησιμοποιώντας τη

διαφορά $\Delta \boldsymbol{w}$ του προηγούμενου βήματος. Η νέα ανανέωση των μεταβλητών υπολογίζεται ως ο γραμμικός συνδυασμός της προηγούμενης διαφοράς $\Delta \boldsymbol{w}$ και της κλίσης της συνάρτησης απώλειας $\boldsymbol{Q}$:

$$\Delta \boldsymbol{w} := \alpha \Delta \boldsymbol{w} - \eta \nabla Q_i(\boldsymbol{w}) \tag{17}$$

όπου α είναι ο συντελεστής απόσβεσης (decay), ο οποίος παίρνει τιμές μεταξύ 0 και 1 και καθορίζει τη σχετική συμβολή της προηγούμενης ανανέωσης, και η το βήμα βελτιστοποίησης ή ρυθμός εκπαίδευσης (learning rate). Τελικά, οι νέες μεταβλητές υπολογίζονται ως:

$$\boldsymbol{w} := \boldsymbol{w} + \Delta \boldsymbol{w} \tag{18}$$

ή

$$\boldsymbol{w} := \boldsymbol{w} - \eta \nabla Q_i(\boldsymbol{w}) + \alpha \Delta \boldsymbol{w} \tag{19}$$

Για την τρέχουσα εφαρμογή, χρησιμοποιήθηκε συντελεστής απόσβεσης ορμής $\alpha = 0.9$ και ρυθμός ή βήμα εκπαίδευσης $\eta = 0.001$.

Για την εκπαίδευση του μοντέλου, χρησιμοποιείται L2 κανονικοποίηση (L2 regularization) όπου η συνάρτηση απώλειας επιβάλλει ποινή για μεγάλες τιμές βαρών. Η συμβολή της κανονικοποίησης στην συνάρτηση απώλειας γίνεται με τον λεγόμενο συντελεστή απόσβεσης βαρών λ. Ο συντελεστής αυτός παίρνει συνήθως μικρές τιμές, έτσι ώστε ο όρος L2 να μην υπερισχύσει, οδηγώντας σε αναποτελεσματική εκπαίδευση. Στην τρέχουσα εφαρμογή επιλέχθηκε $\lambda = 0.0005$.

Έπειτα, για την εξασφάλιση της ομαλής πορείας εκπαίδευσης και της σύγκλισης της συνάρτησης απώλειας, το βήμα βελτιστοποίησης η αυξάνεται σταδιακά από το μηδέν έως την κανονική του τιμή σε ένα πεπερασμένο αριθμό επαναλήψεων. Το τέχνασμα αυτό, λεγόμενο ως “burn-in”, επιβραδύνει τον ρυθμό της εκπαίδευσης για τις πρώτες επαναλήψεις έτσι ώστε να αποφευχθεί η απόκλιση της βελτιστοποίησης ή ο εγκλωβισμός σε τοπικό ακρότατο. Ο αριθμός επιλέχθηκε έτσι ώστε το βήμα βελτιστοποίησης να αποκτήσει την πλήρη τιμή του μετά από 1000 επαναλήψεις. Συνολικά. το βήμα εκπαίδευσης υπολογίζεται συναρτήσει της επανάληψης i ως εξής:

$$\eta(i) = \begin{cases} \eta_0 \left(\frac{i}{1000}\right)^4, & i < 1000 \\ \eta_0, & i \geq 1000 \end{cases} \tag{20}$$

Τέλος, επιλέχθηκε ο μέγιστος αριθμός επαναλήψεων της εκπαίδευσης καθώς και δύο στάδια όπου ρυθμός της επιβραδύνεται έτσι ώστε να υπάρξει πιο ομαλή σύγκλιση. Ο μέγιστος αριθμός επαναλήψεων ορίστηκε ίσος με 16000.

4.2. Υλισμικό και Λογισμικό Εκπαίδευσης

Η εκπαίδευση του μοντέλου YOLOv4-Tiny-3L υλοποιήθηκε αξιοποιώντας την υπολογιστική ισχύ μιας κάρτας γραφικών NVIDIA RTX 3090 με 24 GB VRAM και ενός επεξεργαστή Intel® Xeon® E5-2696 v3. Το Darknet, η βιβλιοθήκη ανίχνευσης αντικειμένων ανοικτού κώδικα που περιλαμβάνει μοντέλα όπως οι διάφορες παραλλαγές του YOLO, μεταγλωττίστηκε (compiled) έτσι ώστε να επιτρέπει την επιτάχυνση λειτουργίας με κάρτα γραφικών μέσω της πλατφόρμας παράλληλου υπολογισμού CUDA® και της βιβλιοθήκης βαθιάς μάθησης cuDNN της NVIDIA. Ο ισχυρός συνδυασμός λογισμικού και υλικού επιτρέπει την αποδοτική εκπαίδευση του μοντέλου, αξιοποιώντας πλήρως τις ικανότητες παράλληλου υπολογισμού που παρέχονται.

Κατά την εκπαίδευση, οι εικόνες του συνόλου δεδομένων επιλέγονται τυχαία σε παρτίδες των 128 (batch size) και εισέρχονται παράλληλα στο δίκτυο. Το μέγεθος των παρτίδων επιλέχθηκε με γνώμονα τη διαθέσιμη μνήμη VRAM της κάρτας γραφικών έτσι ώστε να αξιοποιηθεί πλήρως. Τα μεγαλύτερα μεγέθη παρτίδων ευνοούν την καθολική επίδοση του μοντέλου, καθώς, σε κάθε βήμα εκπαίδευσης, το μοντέλο προσαρμόζεται σε ένα πιο αντιπροσωπευτικό υποσύνολο του συνόλου δεδομένων.

4.3. Παραλλαγές Μοντέλου

Συνολικά, στην παρούσα μελέτη εκπαιδεύτηκαν τρία μοντέλα ανίχνευσης αντικειμένων YOLOv4-Tiny-3L. Οι τρεις παραλλαγές διαφέρουν στα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευσή τους..

1. Το μοντέλο Σ εκπαιδεύτηκε αποκλειστικά με το συνθετικό σύνολο δεδομένων που δημιουργήθηκε σύμφωνα με το Κεφάλαιο 3.

2. Το μοντέλο Α εκπαιδεύτηκε με αληθινές εικόνες οι οποίες ελήφθησαν με κάμερα κινητού τηλεφώνου σε διάφορες τοποθεσίες.

3. Το μοντέλο Π προέκυψε μέσω της προσαρμογής (fine-tuning) του μοντέλου Α, το οποίο είχε προεκπαιδευτεί με συνθετικές εικόνες, με περαιτέρω εκπαίδευση με αληθινές εικόνες.

Η εκπαίδευση πολλαπλών μοντέλων έγινε έτσι ώστε να αξιολογηθεί η επίδραση της μεθόδου εκπαίδευσης με συνθετικά δεδομένα και να συγκριθεί με συμβατικές εναλλακτικές. Όλα τα μοντέλα αξιολογούνται ως προς το ίδιο σύνολο επικύρωσης που αποτελείται από αληθινές εικόνες. Στο Σχήμα 4.1 παρατίθενται μερικές λήψεις από το αληθινό σύνολο δεδομένων που συλλέχθηκε, το οποίο συμπεριλαμβάνει εικόνες από τις εγκαταστάσεις του Εργαστηρίου Τεχνολογίας των Κατεργασιών (ΕΤΚ) του Εθνικού Μετσόβιου Πολυτεχνείου.

Σχήμα 4.1. Παραδείγματα αληθινών εικόνων των υπο μελέτη κατηγοριών.

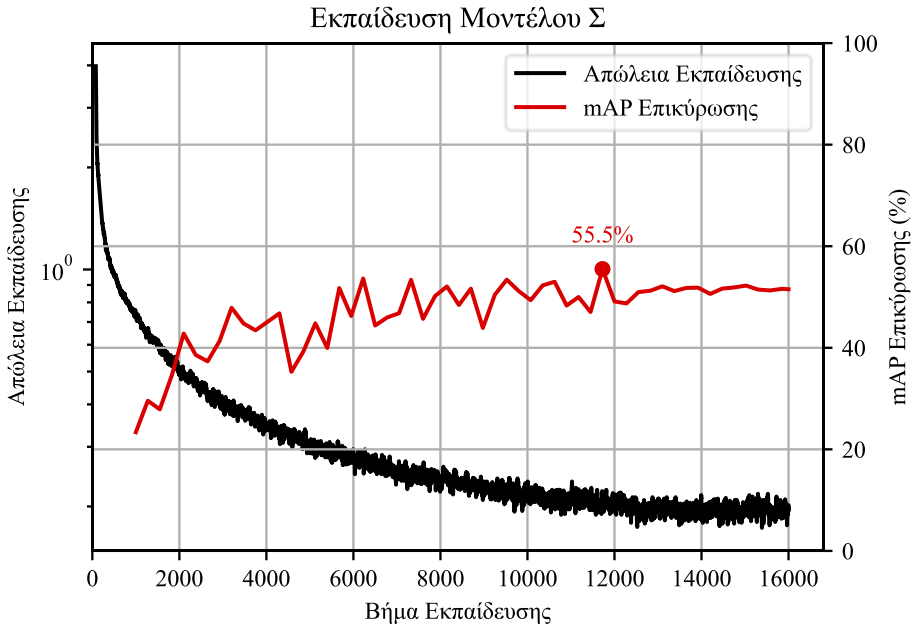
Μόνο πέντε εικόνες από κάθε κατηγορία χρησιμοποιήθηκαν για την εκπαίδευση των δύο τελευταίων μοντέλων. Η επιλογή αυτή έγινε λόγω του περιορισμένου πλήθους αληθινών δεδομένων και με σκοπό τη διερεύνηση της ικανότητας γενίκευσης των μοντέλων έχοντας πρόσβαση σε πολύ περιορισμένο δείγμα δεδομένων.

4.4. *Παρακολούθηση Εκπαίδευσης*

Η εξέλιξη της εκπαίδευσης του κάθε μοντέλου παρακολουθείται μέσω της καταγραφής διάφορων μετρικών σε κάθε βήμα εκπαίδευσης. Οι πιο κρίσιμες μετρήσεις αφορούν την τιμή της συνάρτησης απώλειας και την ακρίβεια του μοντέλου. Ο υπολογισμός της συνάρτησης απώλειας για ολόκληρο το σύνολο δεδομένων εκπαίδευσης είναι ανώφελος και μη αποδοτικός μπορεί να επιτευχθεί παράλληλα λόγω της μνήμης που απαιτείται για την εκτέλεσή του. Συνεπώς, η τιμή της συνάρτησης απώλειας που καταγράφεται, υπολογίζεται για την παρτίδα τυχαίων εικόνων που εισέρχονται στο δίκτυο σε κάθε επανάληψη και πρόκειται για μια εκτίμηση του συνολικού σφάλματος του μοντέλου. Η στοχαστική αυτή προσέγγιση αποτελεί ισχυρό υποκατάστατο καθώς επαρκεί για την αποτελεσματική εκπαίδευση του μοντέλου. Η ακρίβεια του μοντέλου ως προς το αληθινό σύνολο επικύρωσης υπολογίζεται ανά τακτά διαστήματα της εκπαίδευσης για ολόκληρο το σύνολο επικύρωσης που χρησιμοποιείται. Υπολογίζεται ξεχωριστά η ακρίβεια ανίχνευσης κάθε κατηγορίας και έπειτα η μέση ακρίβεια ή mean average precision (mAP), η οποία χαρακτηρίζει την καθολική επίδοση του μοντέλου.

4.5. *Αποτελέσματα Εκπαίδευσης*

Το μοντέλο Σ εκπαιδεύτηκε για 16000 επαναλήψεις. Η εκπαίδευση διήρκεσε 160’ με χρόνο εκτέλεσης κάθε επανάληψης περίπου 550 ms. Στο Σχήμα 4.2 αποτυπώνεται η πορεία της εκπαίδευσής του.

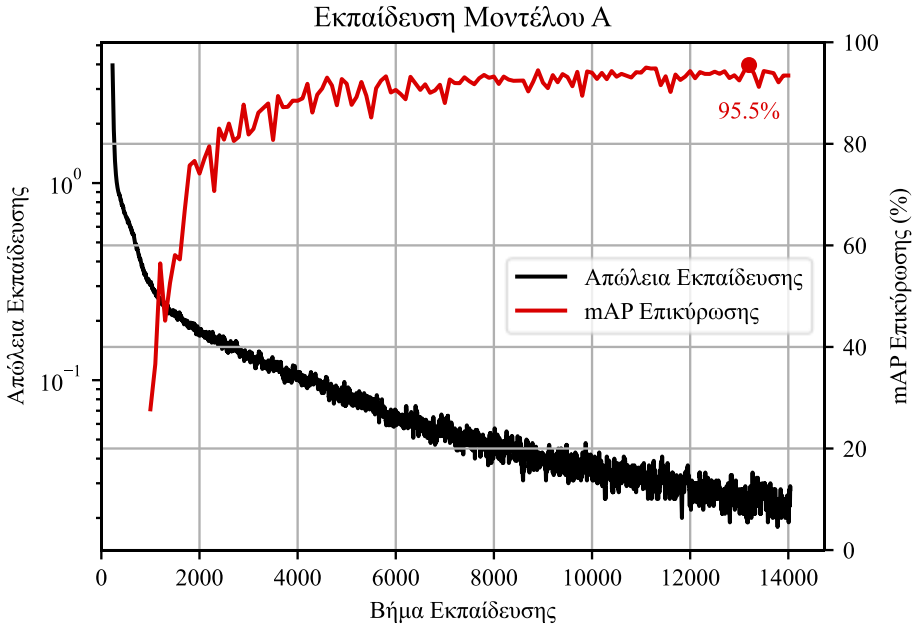


Σχήμα 4.2. Πορεία εκπαίδευσης του μοντέλου Σ.

Η τιμή της απώλειας καθώς και η μέση ακρίβεια καταγράφονται για κάθε βήμα της εκπαίδευσης. Η απώλεια παρίσταται σε λογαριθμική κλίμακα αναδεικνύοντας την πτώση και τη σύγκλιση του σφάλματος του μοντέλου. Σημειώνεται πώς, παρόλο που η συνάρτηση απώλειας εκτιμάται για ένα μικρό υποσύνολο των δεδομένων εκπαίδευσης, το σφάλμα μειώνεται αποτελεσματικά και ομαλά, δίχως να αποκλίνει. Επιπλέον, παρατηρείται η σταδιακή αύξηση της μέσης ακρίβειας για το αληθινό σύνολο επικύρωσης, φθάνοντας σε μέγιστη τιμή 55.5%, έως την τελική τιμή στην οποία συγκλίνει. Η διακύμανση της ακρίβειας που εμφανίζεται κατά τη διάρκεια της εκπαίδευσης οφείλεται στη διαφορά μεταξύ των κατανομών των δεδομένων του συνθετικού συνόλου εκπαίδευσης και του αληθινού

συνόλου επικύρωσης. Ωστόσο, η αύξηση της ακρίβειας για πτώση στο σφάλμα σηματοδοτεί επαρκή συσχέτιση μεταξύ των δύο κατανομών.

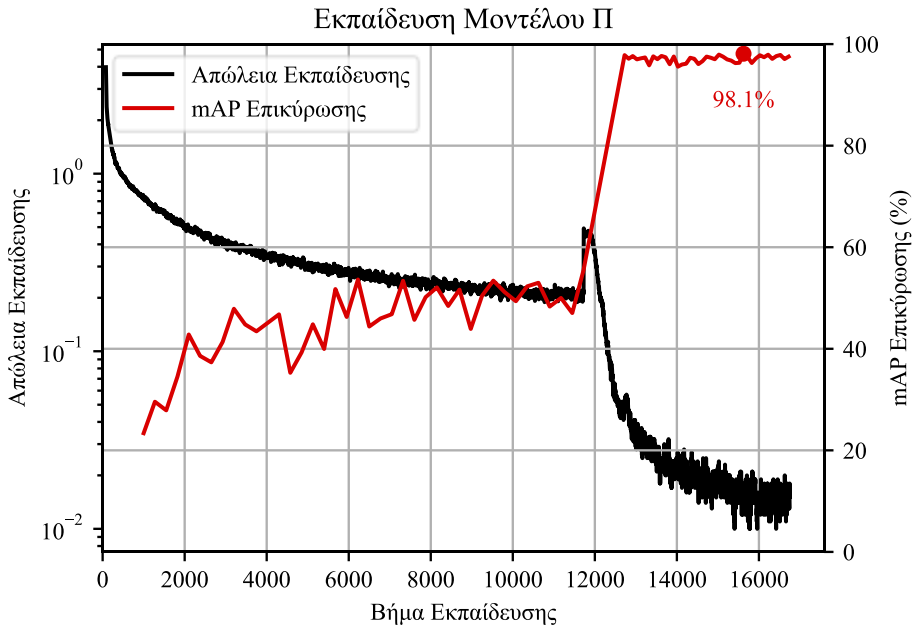
Το μοντέλο Α εκπαιδεύτηκε για περίπου 14000 επαναλήψεις. Στο Σχήμα 4.3 αποτυπώνεται η πορεία της εκπαίδευσής του.



Σχήμα 4.3. Πορεία εκπαίδευσης του μοντέλου Α.

Το σφάλμα του μοντέλου Α, σε σύγκριση με το μοντέλο Σ, μειώνεται περισσότερο και δεν συγκλίνει κατά το πέρας της εκπαίδευσης. Επιπλέον, επιτυγχάνεται πολύ μεγαλύτερη μέση ακρίβεια, φθάνοντας σε μέγιστη τιμή 95.5%, λόγω των αληθινών δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση, η κατανομή των οποίων ταιριάζει περισσότερο με αυτή του συνόλου επικύρωσης. Επίσης, λόγω της μεγαλύτερης συσχέτισης, εμφανίζεται μικρότερη διακύμανση στην ακρίβεια κατά την πορεία της εκπαίδευσης. Τέλος, λόγω της σύγκλισης της μέσης ακρίβειας, και προς αποφυγή της υπερπροσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης, η εκπαίδευση διακόπτεται πριν από τον μέγιστο αριθμό επαναλήψεων.

Για το μοντέλο Π, συνεχίστηκε η εκπαίδευση του μοντέλου Α από το σημείο με την καλύτερη μέση ακρίβεια για περίπου 5000 ακόμα επαναλήψεις, εκπαιδεύοντάς το με αληθινά δεδομένα. Στο Σχήμα 4.4 αποτυπώνεται η πορεία της εκπαίδευσής του.



Σχήμα 4.4. Πορεία εκπαίδευσης του μοντέλου Π.

Η εκπαίδευση με αληθινά δεδομένα ξεκινάει μετά από περίπου 11600 βήματα εκπαίδευσης με συνθετικά δεδομένα. Όσον αφορά την τιμή της απώλειας, παρατηρείται μια αρχική ακαριαία αύξηση λόγω της εισαγωγής νέων δεδομένων, την οποία ακολουθεί μια ραγδαία πτώση. Το σφάλμα του μοντέλου ελαττώνεται σε τιμές χαμηλότερες κατά δύο τάξεις μεγέθους από αυτές που προηγούνταν της εισαγωγής των αληθινών δεδομένων. Στη συνέχεια, παρατηρείται απότομη αύξηση και έπειτα σύγκλιση της μέσης ακρίβειας ακολουθώντας την προσαρμογή στα νέα δεδομένα, επιτυγχάνοντας μέγιστη τιμή 98.1%, ξεπερνώντας τα άλλα δύο μοντέλα.

Κεφάλαιο 5

Αξιολόγηση Μοντέλων Ανίχνευσης Αντικειμένων

5.1. Μετρικές Ανίχνευσης

Η αξιολόγηση μοντέλων ανίχνευσης αντικειμένων περιλαμβάνει την ανάλυση ορισμένων μετρικών που αφορούν την ευστοχία τους. Οι μετρικές αυτές είναι οι αριθμοί αληθινά θετικών (true positives – TP), ψευδώς θετικών (false positive – FP) και ψευδώς αρνητικών (false negatives – FN) προβλέψεων. Η επεξεργασία των παραπάνω μετρήσεων επιτρέπει τον υπολογισμό σημαντικών στατιστικών μεγεθών όπως η ακρίβεια, η ανάκληση και η συνολική ορθότητα ενός μοντέλου. Οι μετρικές αναλύονται ως εξής:

- Αληθινά Θετικά – Πρόκειται για περιπτώσεις όπου το μοντέλο αναγνωρίζει και εντοπίζει ένα αντικείμενο εύστοχα. Η πρόβλεψη του μοντέλου ταυτίζεται με τα αληθινά δεδομένα.

- Ψευδώς Θετικά – Πρόκειται για περιπτώσεις όπου το μοντέλο ανιχνεύει ψευδώς κάποιο αντικείμενο το οποίο δεν βρίσκεται πραγματικά εντός της εικόνας. Η πρόβλεψη του μοντέλου δεν ταυτίζεται με τα αληθινά δεδομένα.

- Αληθινά Αρνητικά – Πρόκειται για περιπτώσεις όπου δεν υπάρχει κάποιο αντικείμενο εντός της εικόνας και το μοντέλο προβλέπει ορθά την απουσία αντικειμένου, συμβαδίζοντας με τα αληθινά δεδομένα.

- Ψευδώς Αρνητικά – Πρόκειται για περιπτώσεις όπου το μοντέλο αποτυγχάνει να ανιχνεύσει ένα αντικείμενο το οποίο βρίσκεται εντός της εικόνας. Και σε αυτή την περίπτωση, η πρόβλεψη του μοντέλου δεν ταυτίζεται με τα αληθινά δεδομένα.

Οι μετρικές συλλέχθηκαν για τα μοντέλα Σ, Α και Π στο Πίνακα 5.1.

Πίνακας 5.1

Πλήθη αληθινά θετικών, ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων κάθε μοντέλου.

	Μετρικές		
Μοντέλο	Αληθινά Θετικά ↑	Ψευδώς Θετικά ↓	Ψευδώς Αρνητικά ↓
Σ	1032	631	264
Α	1424	205	147
Π	**1513 (+6.3%)**	**67 (-67.3%)**	**58 (-60.5%)**

Τα βέλη ↑ και ↓ συμβολίζουν αν η επίδοση αυξάνεται με αύξηση ή μείωση της τιμής, αντίστοιχα.
Η ποσοστιαία μεταβολή υπολογίζεται με βάση τα αποτελέσματα του μοντέλου Α.

Το μοντέλο Π, το οποίο προσαρμόστηκε σε αληθινά δεδομένα ακολουθώντας εκπαίδευση στο συνθετικό σύνολο δεδομένων, εμφανίζει τις καλύτερες μετρικές σε σύγκριση με τα άλλα δύο μοντέλα. Το μοντέλο Α, το οποίο εκπαιδεύτηκε με αληθινές εικόνες εμφανίζει αρκετά καλύτερες μετρικές από το μοντέλο Σ, το οποίο εκπαιδεύτηκε με αποκλειστικά συνθετικά δεδομένα. Αναφορικά με το μοντέλο Π, παρατηρείται αύξηση των αληθινά θετικών προβλέψεων κατά 46.6% και 6.25%, πτώση στις ψευδώς θετικές προβλέψεις κατά 89.3% και 67.3% και στις ψευδώς αρνητικές προβλέψεις κατά 78% και 60.5% σε σχέση με τα μοντέλα Σ και Α, αντίστοιχα. Σημειώνεται πως η αύξηση της επίδοσης του μοντέλου Π σε σχέση με το μοντέλο Α έγκειται κυριώς στην σημαντική μείωση των ψευδών προβλέψεων.

Επιπλέον, οι μετρικές υπολογίζονται ξεχωριστά για κάθε κατηγορία. Η ανάλυση της επίδοσης στην ανίχνευση των επιμέρους κατηγοριών αποτελεί χρήσιμο εργαλείο διάγνωσης και μελέτης της λειτουργίας του μοντέλου και του συνθετικού συνόλου δεδομένων. Στον Πίνακα 5.2 αναλύονται οι μετρικές ξεχωριστά για κάθε κατηγορία για όλα τα μοντέλα.

Πίνακας 5.2

Ξεχωριστά πλήθη αληθινά θετικών, ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων των μοντέλων Σ, Α και Π για κάθε κατηγορία.

	Αληθινά Θετικά ↑			Ψευδώς Θετικά ↓			Ψευδώς Αρνητικά ↓		
Κατηγορία	Σ	Α	**Π**	Σ	Α	**Π**	Σ	Α	**Π**
CNC Τόρνος	8	93	**96 (+3%)**	9	24	**11 (-54%)**	108	23	**20 (-13%)**
CNC Φρέζα	0	**31**	**31**	1	30	**3 (-90%)**	31	**0**	**0**
Τόρνος	42	**78**	**78**	216	24	**17 (-29%)**	38	**2**	**2**
Φρέζα	38	81	**83 (+2%)**	35	31	**6 (-81%)**	45	2	**0 (-100%)**
Περονοφόρο	363	407	**468 (+15%)**	278	75	**22 (-71%)**	135	91	**30 (-67%)**
Παλέτα	212	288	**298 (+3%)**	52	18	**7 (-61%)**	92	16	**6 (-63%)**
Άνθρωπος	360	446	**459 (+3%)**	40	3	**1 (-67%)**	99	13	**0 (-100%)**

Το μοντέλο Σ εμφανίζει εξαιρετικά χαμηλή επίδοση για την ανίχνευση των εργαλειομηχανών CNC, αποτυγχάνοντας να ανιχνεύσει έστω και μια CNC φρέζα. Η επίδοσή του για την ανίχνευση των συμβατικών εργαλειομηχανών είναι επίσης χαμηλή. Το μοντέλο εμφανίζει καλύτερη επίδοση για τις κατηγορίες των περονοφόρων, των παλετών και των ανθρώπων, ωστόσο οι ψευδώς θετικές ανιχνεύσεις παραμένουν υψηλές. Το πλήθος των ψευδώς αρνητικών προβλέψεων είναι πολύ μεγάλο, γεγονός που σημαίνει ότι το μοντέλο αποτυγχάνει να ανιχνεύσει τα αντικείμενα στις περισσότερες εικόνες.

Το μοντέλο Α εντοπίζει επιτυχώς τα αντικείμενα όλων των κατηγοριών σε μεγαλύτερο βαθμό, ωστόσο και αυτό χαρακτηρίζεται από μεγάλο πλήθος ψευδώς θετικών προβλέψεων. Οι ψευδώς αρνητικές προβλέψεις του είναι λιγότερες από αυτές του μοντέλου Σ όμως αποτελούν αρνητικό παράγοντα για την επίδοση του.

Το μοντέλο Π εμφανίζει την καλύτερη επίδοση μεταξύ όλων των μοντέλων. Η βέλτιστη επίδοση του έγκειται εν μέρει στη μικρή αύξησης στο πλήθος των εύστοχων προβλέψεων σε σχέση με το μοντέλο Α αλλά κυρίως στη μεγάλη πτώση των ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων. Το μοντέλο Π εντοπίζει ορθά σχεδόν όλα τα αντικείμενα εντός των εικόνων με λίγα σφάλματα.

Παρακάτω, οι μετρικές που αναλύθηκαν επεξεργάζονται περαιτέρω για τον υπολογισμό σημαντικών μεγεθών που προσφέρουν πιο εύστοχη εκτίμηση της αποτελεσματικότητας των μοντέλων.

5.2. Ακρίβεια και Ανάκληση

Όπως αναλύθηκε στην ενότητα 2.2, η τομή μεταξύ του προβλεπόμενου και του πραγματικού πλαισίου οριοθέτησης αποτελεί καθοριστικό μέγεθος για την εκπαίδευση και την αξιολόγηση μοντέλων ανίχνευσης αντικειμένων. Κατά την αξιολόγηση ενός μοντέλου, οι αληθινά θετικές ανιχνεύσεις καθορίζονται από το intersection over union (IOU) μεταξύ του προβλεπόμενου και του πραγματικού πλαισίου οριοθέτησης και του επιπέδου εμπιστοσύνης της πρόβλεψης. Οι προβλέψεις οι οποίες έχουν IOU και επίπεδο εμπιστοσύνης μεγαλύτερα από αντίστοιχες τιμές κατωφλίου θεωρούνται αληθινά θετικές ενώ άλλες, με τουλάχιστον ένα από τα μεγέθη να έχει τιμή κάτω από το κατώφλι, καταγράφονται ως ψευδώς θετικές. Συνεπώς, για δεδομένη τιμή κατωφλίου IOU, η τιμή κατωφλίου εμπιστοσύνης επηρεάζει άμεσα την κατανομή των αληθινά θετικών, των ψευδώς θετικών, των αληθινά αρνητικών και των ψευδώς αρνητικών.

Μέσω των μετρικών αυτών υπολογίζονται δύο επιπλέον ποσότητες οι οποίες συνεισφέρουν στην ανάλυση της επίδοσης ενός μοντέλου ανίχνευσης αντικειμένων. Οι ποσότητες αυτές είναι η ακρίβεια και η ανάκληση του μοντέλου. Συγκεκριμένα, η ακρίβεια (precision) προσδιορίζει πόσες από τις συνολικές θετικές προβλέψεις είναι πραγματικά θετικές. Πρόκειται για τον λόγο των αληθινά θετικών προβλέψεων προς το άθροισμα των αληθινά θετικών και των ψευδώς θετικών προβλέψεων του μοντέλου και υπολογίζεται ως εξής:

$$\text{Ακρίβεια} = \frac{\text{Αληθινά Θετικές}}{\text{Αληθινά Θετικές} + \text{Ψευδώς Θετικές}} \tag{21}$$

Συνεπάγεται πως η ακρίβεια ενός μοντέλου επηρεάζεται αρνητικά από την ύπαρξη ψευδώς θετικών προβλέψεων που αυτό κάνει.

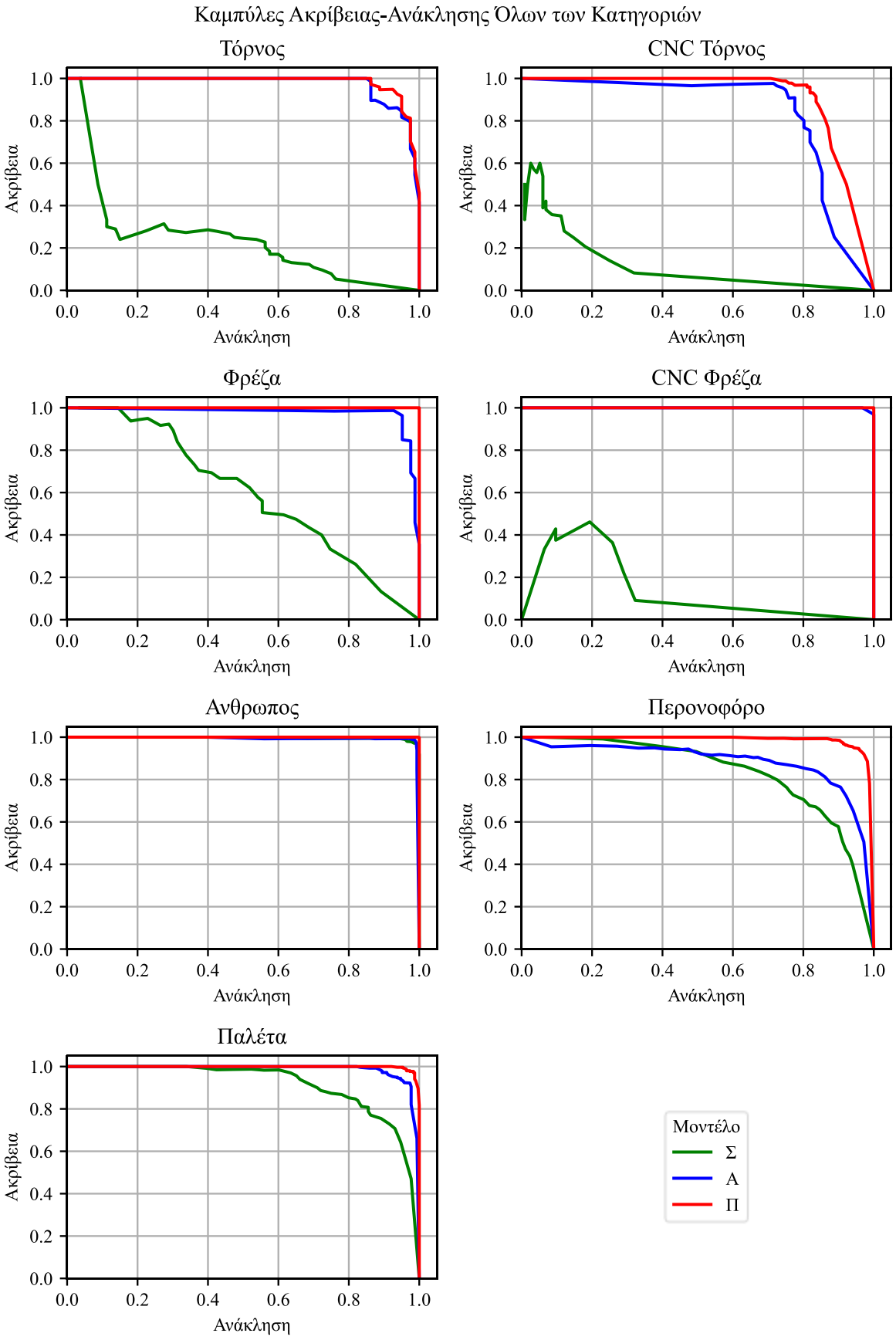
Η ανάκληση (recall) ενός μοντέλου ανίχνευσης αντικειμένων προσδιορίζει την ικανότητα του μοντέλου να ανιχνεύσει τα υπάρχοντα αντικείμενα εντός των εικόνων. Η ανάκληση υπολογίζεται ως ο λόγος των αληθινά θετικών προβλέψεων προς το συνολικό πλήθος αντικειμένων που υπάρχει εντός των εικόνων του συνόλου δεδομένων.

$$\text{Ανάκληση} = \frac{\text{Αληθινά Θετικές}}{\text{Αληθινά Θετικές} + \text{Ψευδώς Αρνητικές}} \tag{22}$$

Το πραγματικό συνολικό πλήθος των αντικειμένων που υπάρχει εντός των εικόνων μπορεί να προκύψει από το άθροισμα των αληθινά θετικών και ψευδώς αρνητικών προβλέψεων. Συμπεραίνεται ότι η ανάκληση επηρεάζεται αρνητικά από την αδυναμία του μοντέλου να ανιχνεύσει αντικείμενα που υπάρχουν σε μια εικόνα, δηλαδή να κάνει ψευδώς αρνητικές προβλέψεις.

Καθώς αυξάνεται το κατώφλι εμπιστοσύνης, το κριτήριο που καθιστά μια πρόβλεψη ως αληθινά θετική γίνεται πιο αυστηρό με αποτέλεσμα την μείωση των θετικών προβλέψεων, κυριώς των ψευδών, καθώς τα επίπεδα εμπιστοσύνης τους τείνουν να είναι χαμηλότερα. Ταυτόχρονα, η μείωση των αληθινά θετικών προβλέψεων επιφέρει την αύξηση των ψευδώς αρνητικών προβλέψεων, καθώς οι δύο ποσότητες αθροίζουν σε σταθερό αριθμό. Συνεπώς, η αύξηση του κατωφλίου έχει ως αποτέλεσμα την αύξηση της ακρίβειας και τη μείωση της ανάκλησης.

Οι μετρικές που αναλύθηκαν προηγουμένως υπολογίστηκαν για κατώφλι IOU ίσο με 0.5 και κατώφλι εμπιστοσύνης ίσο με 0.25. Ωστόσο, η καταγραφή των μετρικών και ο υπολογισμός της ακρίβειας και της ανάκλησης για διάφορες τιμές του κατωφλίου εμπιστοσύνης προσφέρει χρήσιμες πληροφορίες για την επίδοση του μοντέλου. Για σταθερό κατώφλι IOU ίσο με 0.5 και εύρος τιμών του κατωφλίου εμπιστοσύνης μεταξύ 0 και 1, καταγράφονται οι μετρικές και υπολογίζονται ζεύγη ακρίβειας-ανάκλησης ξεχωριστά για κάθε κατηγορία, τα οποία παρίστανται γραφικά για όλα τα μοντέλα που εκπαιδεύτηκαν, στο Σχήμα 5.1.



Σχήμα 5.1. Καμπύλες ακρίβειας-ανάκλησης για κάθε κατηγορία αντικειμένου για όλα τα μοντέλα.

Σε ένα ιδανικό μοντέλο ανίχνευσης αντικειμένων, όπου δεν υπάρχει καμία ψευδή πρόβλεψη, η ακρίβεια και η ανάκληση του μοντέλου είναι τέλειες και η καμπύλη ακρίβειας-ανάκλησης εκφυλίζεται σε ένα σημείο (1,1). Για απουσία ψευδών προβλέψεων, οι λόγοι που εκφράζονται από την ακρίβεια και την ανάκληση ισούνται με μονδάδα, ανεξάρτητα της τιμής του κατωφλίου εμπιστοσύνης. Αντιθέτως, ένα αφελές μοντέλο δίχως καμία ικανότητα ανίχνευσης αντικειμένων έχει μηδενική ακρίβεια και ανάκληση, και η καμπύλη του εκφυλίζεται στο σημείο (0.0). Συνεπώς, η επίδοση ενός μοντέλου μπορεί να εκτιμηθεί οπτικά από την καμπύλη ακρίβειας-ανάκλησης που σχηματίζεται για ζεύγη τιμών ακρίβειας και ανάκλησης για διάφορες τιμές του κατωφλίου εμπιστοσύνης.

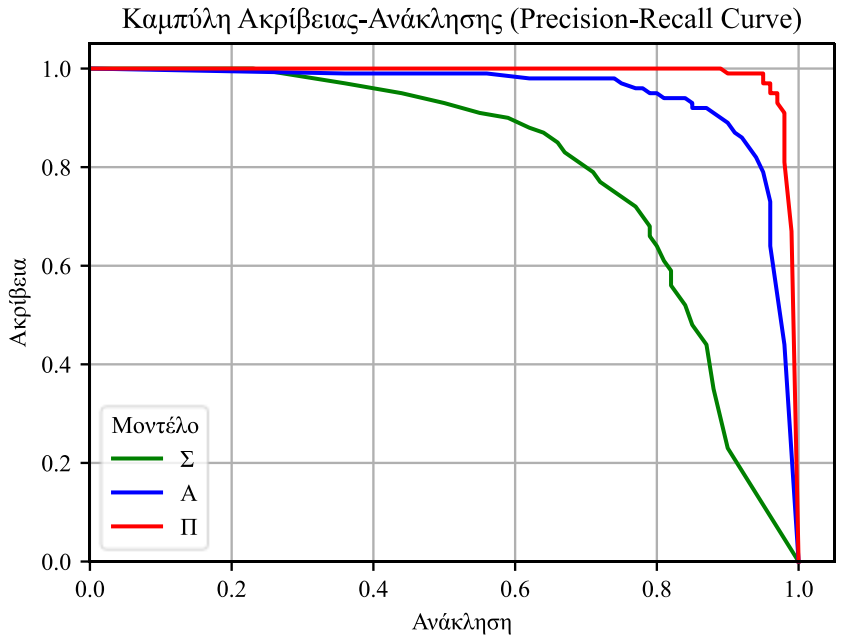
Παρατηρώντας τα διαγράμματα των ξεχωριστών καμπυλών κάθε κατηγορίας για κάθε μοντέλο, εξάγονται σημαντικά συμπεράσματα για την επίδοση του καθενός:

- Το μοντέλο Σ, το οποίο εκπαιδεύτηκε με συνθετικά δεδομένα, εμφανίζει τη χαμηλότερη επιδοση, καθώς η καμπύλη ακρίβειας-ανάκλησής του απέχει από την πάνω δεξιά γωνία (1,1) περισσότερο από τις καμπύλες των υπόλοιπων μοντέλων. Αντίστοιχα με την ανάλυση των μετρικών στους Πίνακες 5.1 και 5.2, παρατηρείται χαμηλότερη επίδοση για τις κατηγορίες των εργαλειομηχανών, με το μοντέλο να εχει την χειρότερη επίδοση για την κατηγορία των CNC φρεζών.

- Το μοντέλο Α, το οποίο εκπαιδεύτηκε με αληθινά δεδομένα, εμφανίζει υψηλή επίδοση η οποία κρίνεται με βάση την καμπύλη ακρίβειας-ανάκλησής του, η οποία αποτελείται από υψηλές τιμές ακρίβειας και ανάκλησης για όλες τις υπό μελέτη κατηγορίες.

- Το μοντέλο Π, το οποίο εκπαιδεύτηκε αρχικά στο συνθετικό σύνολο δεδομένων και αργότερα προσαρμόστηκε σε αληθινά δεδομένα, εμφανίζει την καλύτερη επίδοση με γνώμονα την καμπύλη ακρίβειας-ανάκλησής του, η οποία βρίσκεται πάνω από των άλλων δύο μοντέλων και προσεγγίζει αυτή του ιδανικού μοντέλου.

Στο Σχήμα 5.1 φαίνονται οι ενιαίες καμπύλες ακρίβειας-ανάκλησης για όλες τις κατηγορίες για κάθε μοντέλο.



Σχήμα 5.1. Καμπύλες ακρίβειας-ανάκλησης κάθε μοντέλου.

Γενικά, κατάταξη της επίδοσης των μοντέλων βάσει της καμπύλης ακρίβειας-ανάκλησης ταυτίζεται με την κατάταξη βάσει των μετρικών. Αυτό οφείλεται στο γεγονός ότι οι δύο ποσότητες προκύπτουν άμεσα από τις μετρικές για διάφορες τιμές του κατωφλίου IOU. Ωστόσο, η αξία της μέθοδου έγκειται στην εύκολότερη ερμηνεία των αποτελεσμάτων μέσω της οπτικοποίησής τους.

5.3. Μέση Ακρίβεια

Η μέση ακρίβεια (average precision) αποτελεί βασική μετρική αξιολόγησης της επίδοσης των μοντέλων ανίχνευσης αντικειμένων. Η μέση ακρίβεια συμπεριλαμβάνει την πληροφορία που εμπεριέχεται στους υπολογισμούς της ακρίβειας και της ανάκλησης και παρέχει μια ενιαία ποσότητα η οποία χαρακτηρίζει την επίδοση του μοντέλου για κάθε κατηγορία. Επίσης, σχετίζεται άμεσα με την καμπύλη ακρίβειας-ανάκλησης καθώς πρόκειται για το εμβαδό που σχηματίζεται μεταξύ αυτής και των αξόνων ή το λεγόμενο Area Under the Curve (AUC). Η μέση ακρίβεια ορίζεται ως εξής:

$$\text{Μέση Ακρίβεια} = \int_{r=0}^{1} p(r)dr \tag{23}$$

όπου *r* και *p* συμβολίζουν την ανάκληση και την ακρίβεια, αντίστοιχα. Επιπλέον, υπολογίζεται ο μέσος όρος των τιμών της μέσης ακρίβειας κάθε κατηγορίας και προκύπτει η συνολική ακρίβεια (mean average precision – mAP) του μοντέλου. Ο Πίνακας 5.3 περιέχει τις τιμές της μέσης ακρίβειας κάθε κατηγορίας και της συνολικής ακρίβειας για όλα τα μοντέλα για κατώφλι IOU ίσο με 0.5.

Πίνακας 5.3

Ανάλυση της μέσης ακρίβειας για κάθε κατηγορία και συνολικά για κάθε μοντέλο.

	Μέση Ακρίβεια για IOU ≥ 0.5 (%)		
Κατηγορία	Μοντέλο Σ	Μοντέλο Α	**Μοντέλο Π**
Τόρνος	23.9	97.0	**97.8 (+0.8)**
CNC Τόρνος	11.9	85.0	**91.2 (+6.2)**
Φρέζα	59.9	97.4	**98.8 (+1.4)**
CNC Φρέζα	16.9	**98.4**	**98.4**
Άνθρωπος	99.6	99.3	**99.9 (+0.6)**
Περονοφόρο	83.5	88.4	**98.6 (+10.2)**
Παλέτα	91.5	98.7	**99.7 (+1.0)**
Μέσος Όρος (mAP@0.5)	55.3	94.7	**97.8 (+3.1)**

Η ανάλυση του παρακάτω πίνακα εξάγει αρκετά σημαντικά συμπεράσματα για την επίδοση του κάθε μοντέλου:

- Το μοντέλο Σ εμφανίζει τη χαμηλότερη μέση ακρίβεια σε όλες τις κατηγορίες εκτός από αυτή των ανθρώπων. Παρόμοια με την ανάλυση των μετρικών και των καμπυλών ακρίβειας-ανάκλησης, παρατηρείται πολύ χαμηλή επίδοση στην ανίχνευση των εργαλειομηχανών, με τα μηχανήματα CNC να έχουν τις δύο χαμηλότερες τιμές μέσης ακρίβειας. Το μοντέλο Σ εμφανίζει σχετικά καλύτερη επίδοση για την ανίχνευση των συμβατικών εργαλειομηχανών με την μέση ακρίβεια για την κατηγορία της φρέζας να προσεγγίζει το 60%. Το μοντέλο έχει καλύτερη επίδοση ανίχνευσης για τις κατηγορίες των παλετών, των περονοφόρων και των ανθρώπων, με τη μέση ακρίβεια της τελευταίας να ξεπερνάει αυτή του μοντέλου Α. Η κατανομή της επίδοσης του μοντέλου είναι ανομοιόμορφη για τις διάφορες κατηγορίες, με αποτέλεσμα η συνολική μέση ακρίβεια να έχει τιμή 55.3%.

- Το μοντέλο Α εμφανίζει αρκετά υψηλή μέση ακρίβεια σε όλες τις κατηγορίες, παρόλο του μικρού πλήθους εικόνων που χρησιμοποιήθηκε για την εκπαίδευσή του. Η χαμηλότερη μέση ακρίβεια αφορά την κατηγορία των CNC τόρνων ωστόσο είναι αρκετά υψηλή με τιμή 85%. Η συνολική μέση ακρίβεια του μοντέλου Α είναι υψηλή με τιμή που προσεγγίζει το 95%.

- Το μοντέλο Π έχει την καλύτερη μέση ακρίβεια σε όλες τις κατηγορίες και τη μεγαλύτερη συνολική μέση ακρίβεια με τιμή 97.8%, η οποία αποτελεί βελτίωση 3.1 ποσοστιαίων μονάδων από το μοντέλο Α. Επιπλέον, η επίδοση του μοντέλου είναι πιο ισοκατανομημένη για όλες τις κατηγορίες καθώς μέση ακρίβεια κάθε κατηγορίας έχει τιμή άνω του 90%.

Με βάση τα αποτελέσματα του παρόντος κεφαλαίου, τα τρία μοντέλα κατατάσσονται ως προς την επίδοσή τους ως εξής:

$$\text{Μοντέλο } \Sigma < \text{Μοντέλο } \text{Α} < \text{Μοντέλο } \Pi$$

Τελικά, η αξιολόγηση της μέσης ακρίβειας των τριών μοντέλων εξάγει αντίστοιχα συμπεράσματα με την ανάλυση των μετρικών των αληθινά θετικών, ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων και των καμπυλών ακρίβειας-ανάκλησης. Ο υπολογισμός της συνολικής μέσης ακρίβειας ή mAP αποτελεί πολύτιμο εργαλείο για την αξιολόγηση και της καθολικής επίδοσης των μοντέλων και επιτρέπει την άμεση σύγκρισή τους.

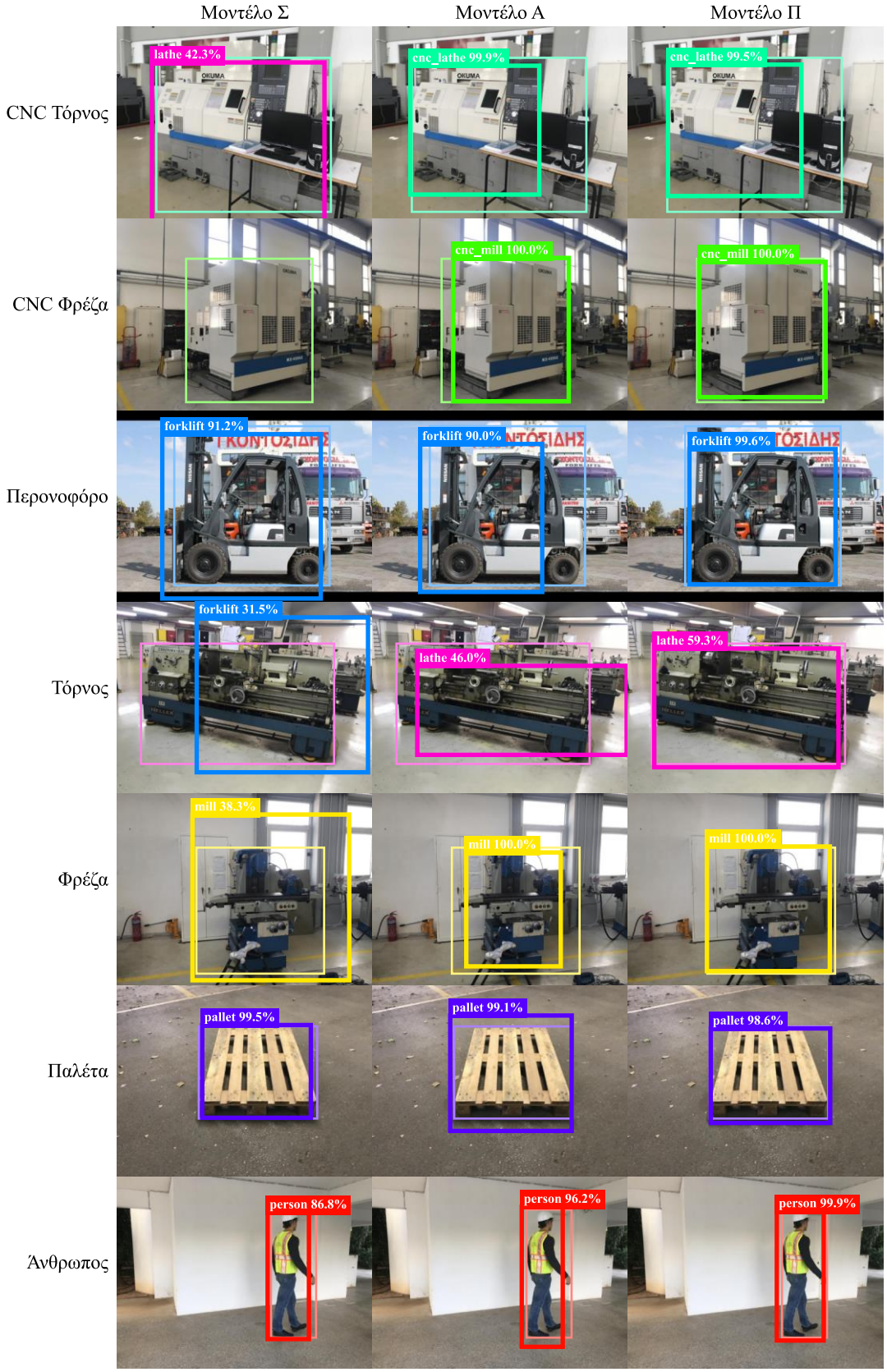
Η προσαρμογή του μοντέλου που προεκπαιδεύτηκε με συνθετικά δεδομένα οδηγεί σε καλύτερη επίδοση από αυτή των μοντέλων που εκπαιδεύτηκαν με αποκλειστικά συνθετικά ή αληθινά δεδομένα.

5.4. Παραδείγματα Ανίχνευσης Αντικειμένων σε Αληθινές Εικόνες

Η ανάλυση των προβλεπόμενων πλαισίων οριοθέτησης και η σύγκρισή τους με τα πραγματικά δεδομένα των διάφορων εικόνων του συνόλου επικύρωσης αποτελεί εξαιρετικά χρήσιμη μέθοδο εκτίμησης της επίδοσης των μοντέλων ανίχνευσης αντικειμένων. Η σύγκριση των πλαισίων οριοθέτησης που εξάγονται από τα μοντέλα με τα πλαίσια που αντιστοιχούν στην πραγματική θέση των αντικειμένων επιτρέπει τον οπτικό εντοπισμό αδυναμιών και παραγόντων που οδηγούν σε χαμηλή επίδοση του κάθε μοντέλου, όπως ψευδώς θετικές ή αρνητικές προβλέψεις ή σφάλματα στην τοποθέτηση των πλαισίων. Η ανάλυση των παραγόντων αυτών βοηθάει στην διάγνωση του μοντέλου και μπορεί να χρησιμοποιηθεί για περαιτέρω βελτίωση του.

Στην περίπτωση πολλαπλών μοντέλων ανίχνευσης αντικειμένων, η οπτική σύγκριση των προβλέψεων του καθενός επιτρέπει την συνειρμική αξιολόγηση της σχετικής επίδοσής τους. Η οπτική εξέταση της συμπεριφοράς του κάθε μοντέλου μπορεί να συνδυαστεί με την ανάλυση των μετρικών, όπως η ακρίβεια και η ανάκληση, για την πιο εύκολη εξαγωγή συμπερασμάτων, η οποία με τη σειρά της βοηθάει στην επιλογή του ακριβέστερου και πιο αποδοτικού μοντέλου.

Στο Σχήμα 5.2, παρατίθενται μερικά τυχαία επιλεγμένα παραδείγματα εικόνων του συνόλου επικύρωσης και οι προβλέψεις του κάθε μοντέλου σε συνδυασμό με τα πραγματικά πλαίσια οριοθέτησης και τις αντίστοιχες κατηγορίες.



Σχήμα 5.2. Οι προβλέψεις των πλαισίων οριοθέτησης και της κατηγορίας του κάθε μοντέλου ανίχνευσης αντικειμένων για τυχαία επιλεγμένες αληθινές εικόνες του συνόλου επικύρωσης. Τα πραγματικά πλαίσια απεικονίζονται με λεπτή γραμμή.

Η παρατήρηση των προβλέψεων του κάθε μοντέλου για τις διάφορες κατηγορίες αντικειμένων καθιστά πιο εύκολη την ερμηνεία των αποτελεσμάτων των προηγούμενων υποκεφαλαίων. Χαρακτηριστικά, η χαμηλή επίδοση ανίχνευσης των εργαλειομηχανών του μοντέλου Σ γίνεται αντιληπτή λόγω της απουσίας έγκυρων προβλέψεων, της λάθος ταξινόμησης των μηχανημάτων και των πλαισίων οριοθέτησης τα οποία έχουν μικρό IOU με τα πραγματικά. Η σύγκριση των προβλέψεων των μοντέλων Α και Π φανερώνει τη βελτίωση της επίδοσης του δεύτερου από το πρώτο. Το μοντέλο Π εντοπίζει τα αντικείμενα στον χώρο με μεγαλύτερη ακρίβεια και οι προβλέψεις του έχουν κατά μέσο όρο μεγαλύτερο επίπεδο εμπιστοσύνης.

5.5. Περιορισμοί Μεθόδου

Το μοντέλο Σ εμφάνισε τη χαμηλότερη ακρίβεια στην ανίχνευση των εργαλειομηχανών. Αντίθετα, η επίδοση του μοντέλου στην ανίχνευση ανθρώπων, παλετών και περονοφόρων προσεγγίζει αυτή των μοντέλων που εκπαιδεύτηκαν σε αληθινά δεδομένα. Η απόκλιση στην επίδοση μεταξύ κατηγοριών αλλά και μεταξύ μοντέλων οφείλεται σε παράγοντες που σχετίζονται με τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του κάθε μοντέλου και με τον τρόπο με τον οποίο αυτά αξιολογήθηκαν.

5.5.1. Περιορισμοί Χρήσης Συνθετικών Δεδομένων

Η εκπαίδευση ενός μοντέλου ανίχνευσης αντικειμένων αποκλειστικά με συνθετικά δεδομένα προσφέρει σημαντικά πλεονεκτήματα και αποτελεί αναγκαία προσέγγιση όταν υπάρχει έλλειψη αληθινών δεδομένων. Ωστόσο, η μέθοδος αυτή έχει διάφορους περιορισμούς οι οποίοι χρήζουν ανάλυσης και διάγνωσης. Η χρήση αμιγώς συνθετικών εικόνων για την εκπαίδευση ενός μοντέλου ανίχνευσης αντικειμένων, όπως πρόκειται για την περίπτωση του μοντέλου Σ, εισάγει ορισμένους παράγοντες που οδηγούν στην επίτευξη χαμηλής επίδοσης.

Η μεθοδολογία δημιουργίας του συνθετικού συνόλου δεδομένων περιλαμβάνει την παραγωγή αληθοφανών εικόνων με βάση ορισμένες γεωμετρικές προδιαγραφές οι οποίες ορίζονται εντός μιας τριδιάστατης σκηνής. Τα αντικείμενα που απεικονίζονται εντός των παραγόμενων εικόνων καλούνται να τηρήσουν τη μορφή των τριδιάστατων μοντέλων που σχεδιάστηκαν στην πρώτη φάση της διαδικασίας. Συνεπάγεται πως η μεταβλητότητα της γεωμετρίας της κάθε κατηγορίας αντικειμένου του συνθετικού συνόλου δεδομένων εξαρτάται άμεσα από τα τριδιάστατα μοντέλα που χρησιμοποιήθηκαν ως βάση για την παραγωγή των εικόνων.

Ορισμένες κατηγορίες αντικειμένων, όπως οι εργαλειομηχανές και συγκεκριμένα τα μηχανήματα CNC, έχουν μεγάλη ποικιλομορφία όσον αφορά την τοπολογία και άλλα γεωμετρικά χαρακτηριστικά τους. Σε αντίθεση με τις παλέτες, η μορφή των οποίων είναι απλούστερη και δεν ποικίλλει σε μεγάλο βαθμό, μιας και η γεωμετρία τους προσεγγίζεται από ένα παραλληλεπίπεδο, οι εργαλειομηχανές έχουν πολύ περισσότερο σύνθετη γεωμετρία η οποία δεν είναι προκαθορισμένη. Μολονότι διάφορα επιμέρους χαρακτηριστικά και εξαρτήματα είναι κοινά μεταξύ διαφορετικών μοντέλων της ίδιας κατηγορίας μηχανήματος, η ύπαρξη βασικών μορφολογικών διαφορών μεταξύ μηχανημάτων που προέρχονται από διαφορετικούς κατασκευαστές καθιστά την ενιαία προσέγγιση της γεωμετρίας τους αδύνατη. Συνεπώς, για την ακριβέστερη αναπαράσταση μιας κατηγορίας μηχανηματων απαιτείται ο σχεδιασμός πολλαπλών τριδιάστατων μοντέλων τα οποία καλύπτουν μεγαλύτερο εύρος χαρακτηριστικών.

Κατά την παρούσα μελέτη, το πλήθος των τριδιάστατων μοντέλων που χρησιμοποιήθηκε για την αναπαράσταση κάθε κατηγορίας ήταν περιορισμένο με αποτέλεσμα το παραγόμενο συνθετικό σύνολο δεδομένων να μην χαρακτηρίζεται από μεγάλο βαθμό μεταβλητότητας όσον αφορά τη γεωμετρία των αντικειμένων.

Ένας επιπλέον περιορισμός που προκύπτει λόγω της προσέγγισης παραγωγής συνθετικών δεδομένων αφορά την ποιότητα των εικόνων που παράγεται μέσω του μοντέλου διάχυσης. Το LAION-5B, το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του Stable Diffusion, πρόκειται για ένα σύνολο γενικής χρήσης το οποίο περιλαμβάνει διαθέσιμες εικόνες όλων των ειδών από πηγές του διαδικτύου. Οι εργαλειομηχανές είναι εξειδικευμένα μηχανήματα, τα οποία δεν είναι όσο κοινότυπα όσο άλλα αντικείμενα τα οποία συναντώνται στην καθημερινή ζωή. Επομένως, δεν απεικονίζονται σε εικόνες τόσο συχνά όσο, για παράδειγμα, οι άνθρωποι. Κατά συνέπεια, τα δεδομένα εκπαίδευσης του Stable Diffusion δεν περιέχουν τόσες εικόνες εργαλειομηχανών όσες άλλων, πιο συνηθισμένων αντικειμένων. Η ανομοιομορφία αυτή οφείλεται για την διαφορά στην επίδοση του μοντέλου και στην ποιότητα των παραγομενων εικόνων που απεικονίζουν τις διαφορετικές κατηγορίες αντικειμένων. Ως αποτέλεσμα, οι εικόνες των λιγότερο συνηθισμένων αντικειμένων, όπως οι εργαλειομηχανές, του συνθετικού συνόλου δεδομένων της τρέχουσας μελέτης δεν προσεγγίζουν επαρκώς την πραγματικότητα και δεν αποτυπώνουν ικανοποιητικά την αληθινή μορφή των αντικειμένων αυτών.

5.5.2. Περιορισμοί Συνόλου Επικύρωσης

Μέχρι στιγμής, έχει δωθεί έμφαση στην έλλειψη διαθέσιμων δεδομένων που μπορούν να χρησιμοποιηθούν για την εκπαίδευση μοντέλων μηχανικής όρασης. Ωστόσο, η έλλειψη δεδομένων, η οποία οφείλεται στους λόγους που συζητήθηκαν στην ενότητα 3.1, αφορά εξίσου την εκπαίδευση αλλά και την αξιολόγηση ενός μοντέλου ανίχνευσης αντικειμένων. Η αξιολόγηση των τριών μοντέλων που διεξήχθη εντός του παρόντος κεφαλαίου βασίστηκε στη χρήση ενός συνόλου επικύρωσης που πρόκειται για υποσύνολο των αληθινών εικόνων που συλλέχθηκαν. Πέρα από την επίδραση του συνόλου στην εκπαίδευση, η κατανομή των δεδομένων που χρησιμοποιούνται για επικύρωση επηρεάζει άμεσα τις προβλέψεις του μοντέλου και την εκτίμησης της επίδοσής του.

Οι φωτογραφίες των εργαλειομηχανών ελήφθησαν αποκλειστικά στις εγκαταστάσεις του ΕΤΚ. Το πλήθος των διαφορετικών μηχανημάτων της ίδιας κατηγορίας που εμπεριέχονται στις εικόνες του συνόλου είναι περιορισμένο από τα διαθέσιμα μηχανήματα που βρίσκονται εντός των εγκαταστάσεων. Για ορισμένα μηχανήματα, όπως οι εργαλειομηχανές CNC, υπάρχουν μόνο ένα ή δύο μοντέλα της ίδιας κατηγορίας. Αντίστοιχα, η μόνιμη διάταξη του εξοπλισμού εντός των εγκαταστάσεων έχει ως αποτέλεσμα το κάθε μηχάνημα να βρίσκεται πάντα στο ίδιο σημασιολογικό πλαίσιο, δηλαδή να περιτριγυρίζεται από τα ίδια αντικείμενα και να έχει το ίδιο φόντο. Συνεπώς, οποιαδήποτε μεταβλητότητα υπάρχει στο σύνολο προέρχεται από μεταβολές της οπτικής γωνίας και της προοπτικής της κάθε λήψης.

Οι ιδιότητες αυτές έχουν αρνητικό αντίκτυπο στην επίδοση του μοντέλου Σ. Η μεταβλητότητα που εισάγεται κατά τις διάφορες φάσεις της δημιουργίας του συνθετικού συνόλου δεδομένων, όπως αναλύθηκε στο τρίτο κεφάλαιο, αποσκοπεί στην έκθεση του μοντέλου σε ποικίλα δεδομένα επιτρέποντάς το να γενικεύει. Ωστόσο, το σύνολο επικύρωσης χαρακτηρίζεται από χαμηλό βαθμό μεταβλητότητας, επομένως, η επίτευξη υψηλής επίδοσης ως προς αυτό εξαρτάται από την ικανότητα του μοντέλου να προβλέψει με ακρίβεια μόνο τις συγκεκριμένες καταστάσεις τις οποίες αυτό περιλαμβάνει. Συνεπάγεται ότι η ικανότητα γενίκευσης του μοντέλου Σ δεν μεταφράζεται σε μεγάλη επίδοση για το δεδομένο σύνολο επικύρωσης. Αντίθετα, ένα μοντέλο που έχει εκπαιδευτεί με λίγα δεδομένα, τα οποία όμως ταυτίζονται σε μεγάλο βαθμό με αυτά του συνόλου επικύρωσης, θα εμφανίσει μεγαλύτερη επίδοση ως προς αυτό.

5.6. Αξία Προσαρμογής Μοντέλου με Αληθινά Δεδομένα

Με βάση την ανάλυση της επίδοσης των μοντέλων που προηγήθηκε, συμπεραίνεται ότι το μοντέλο Π, το οποίο εκπαιδεύτηκε αρχικά με συνθετικά δεδομένα και έπειτα προσαρμόστηκε με αληθινά δεδομένα, αποτελεί βελτίωση από τα υπόλοιπα μοντέλα. Το μοντέλο αυτό πέτυχε μεγαλύτερη ακρίβεια από τα μοντέλα Σ και Α, έχοντας την ίδια αρχιτεκτονική με αυτά και δίχως να εκπαιδευτεί με νέα δεδομένα ή διαφορετικές παραμέτρους. Η βέλτιστη επίδοσή του οφείλεται στην απλή διαδικασία προσαρμογής του.

Η εκπαίδευση του μοντέλου Π αξιοποιεί τη μεταβλητότητα του συνθετικού συνόλου δεδομένων και την εγκυρότητα των αληθινών εικόνων, συνδυάζοντας τη συνεισφορά της καθεμίας για την ακριβέστερη ανίχνευση των αντικειμένων στο σύνολο επικύρωσης. Η εκπαίδευση με τα συνθετικά δεδομένα καλείται να καλύψει τα κενά που εισάγονται από το περιορισμένο μέγεθος του αληθινού συνόλου εκπαίδευσης και να επιτρέψει την αποτελεσματική γενίκευση στην αληθινή κατανομή. Αντίστοιχα, οι αληθινές εικόνες χρησιμοποιούνται για την προσαρμογή του μοντέλου στις συγκεκριμένες μορφές αντικειμένων που συναντώνται στο σύνολο επικύρωσης και στον αληθινό κόσμο. Η συμβολή των δύο προσεγγίσεων έχει αποτέλεσμα την επίτευξη καλύτερης επίδοσης από την ανεξάρτητη εφαρμογή της καθεμίας.

Συμπεραίνεται ότι η δημιουργία και η επακόλουθη χρήση του συνθετικού συνόλου δεδομένων για την εκπαίδευση ενός μοντέλου ανίχνευσης αντικειμένων αποτελεί πρακτική και ωφέλιμη προσέγγιση για περιπτώσεις όπου υπάρχει έλλειψη αληθινών δεδομένων. Παρόλο που τα συνθετικά δεδομένα δεν επαρκούν από μόνα τους για την επίτευξη υψηλής ακρίβειας, η αξία της μεθόδου έγκειται στην προεκπαίδευση ενός μοντέλου με αυτά και την επακόλουθη προσαρμογή του με εξειδικευμένα, αληθινά δεδομένα.

Κεφάλαιο 6

Συμπεράσματα

6.1. Αναγνώριση Σκηνής στο Επίπεδο Παραγωγής

Η ενσωμάτωση μεθόδων μηχανικής όρασης σε βιομηχανικές εφαρμογές έχει τη δυνατότητα να συμβάλλει στην αυτοματοποίηση διάφορων εργασιών του επιπέδου παραγωγής. Η παρούσα μελέτη, ακολουθώντας προηγούμενες εφαρμογές μεθόδων μηχανικής όρασης βιομηχανικού χαρακτήρα οι οποίες σχετίζονται κυρίως με τον ποιοτικό έλεγχο στη γραμμή παραγωγής, επεκτείνει τις δυνατότητες του κλάδου σε ευρύτερη κλίμακα εστιάζοντας στην αναγνώριση σκηνής στο επίπεδο παραγωγής. Η εφαρμογή αυτή περιλαμβάνει την ανίχνευση εργαζόμενου προσωπικού και μηχανολογικού εξοπλισμού.

Για την επίτευξη του στόχου της αναγνώρισης σκηνής χρησιμοποιήθηκαν μέθοδοι μηχανικής μάθησης. Η επιλογή αυτή έγινε με σκοπό την αξιοποίηση των δυνατοτήτων τέτοιων αλγορίθμων, οι οποίοι επιτρέπουν την εξειδίκευση και την προσαρμογή τους μέσω δεδομένων. Η μηχανική μάθηση επιτρέπει στα μοντέλα μηχανικής όρασης να αναγνωρίζουν διάφορα οπτικά μοτίβα και χαρακτηριστικά από διαφορετικά είδη αντικειμένων, γεγονός το οποίο την καθιστά κατάλληλη για εφαρμογές σε πολύπλοκες καταστάσεις όπως οι βιομηχανικές εγκαταστάσεις οι οποίες χαρακτηρίζονται από μεγάλο βαθμό μεταβλητότητας.

Η ανίχνευση αντικειμένων επιτεύχθηκε με την εκπαίδευση τριών μοντέλων YOLO λόγω των εξαιρετικών δυνατοτήτων του. Η παραλλαγή YOLOv4Tiny-3L επιλέχθηκε για την ικανότητα αναγνώρισης πολλαπλών αντικειμένων σε πραγματικό χρόνο με μεγάλη ακρίβεια. Η υψηλή επίδοση του μοντέλου απαιτείται σε ένα δυναμικό περιβάλλον παραγωγής καθώς η έγκαιρη και ακριβής ανίχνευση αντικειμένων έχει πρωταρχική σημασία για τη βελτιστοποίηση των ροών εργασίας, τη βελτίωση της ασφάλειας της συνολικής λειτουργικής απόδοσης.

6.2. Δημιουργία Συνθετικού Συνόλου Δεδομένων

Η χρήση αληθινών εικόνων για την εκπαίδευση των μοντέλων ανίχνευσης αντικειμένων συναντά αρκετές προκλήσεις, οι οποίες την καθιστούν μη βιώσιμη. Λόγω της εξειδικευμένης φύσης του υπό μελέτη μηχανολογικού εξοπλισμού, υπάρχει έλλειψη διαθέσιμων συνόλων δεδομένων τα οποία περιλαμβάνουν εικόνες των αντίστοιχων μηχανημάτων. Η απομένουσα λύση είναι η συλλογή ή η δημιουργία ενός εξειδικευμένου συνόλου δεδομένων από διάφορες πηγές. Ωστόσο, παράγοντες όπως οι άδειες χρήσης των εικόνων που βρίσκονται στο διαδίκτυο και λειτουργικοί και νομικοί περιορισμοί που αφορούν τη λήψη φωτογραφιών εντός πραγματικών βιομηχανικών εγκαταστάσεων περιορίζουν σημαντικά τη δυνατότητα συλλογής αληθινών δεδομένων.

Για την αντιμετώπιση αυτών των προκλήσεων, στην παρούσα μελέτη παρουσιάστηκε μια καινοτόμα διαδικασία παραγωγής συνθετικού συνόλου δεδομένων για εφαρμογές μηχανικής όρασης όπως η ανίχνευση αντικειμένων. Η διαδικασία συνιστάται από διαφορετικές φάσεις και αξιοποιεί μεθόδους μηχανικής μάθησης για την παραγωγή συνθετικών δεδομένων. Σε πρώτη φάση, σχεδιάζονται τριδιάστατα μοντέλα των υπό μελέτη αντικειμένων με χρήση λογισμικού σχεδίασης. Έπειτα, τα αντικείμενα εντάσσονται σε εικονικές σκηνές, όπου διάφορες παράμετροι παραλάσσονται με σκοπό την αύξηση της ποικιλομορφίας, και εξάγονται εικόνες που αποτυπώνουν την τριδιάστατη μορφή του χώρου μέσω πληροφορίας βάθους. Στη συνέχεια, χρησιμοποιώντας τις εικόνες αυτές ως συνθήκες ελέγχου, παράγονται μεγάλα πλήθη συνθετικών εικόνων χρησιμοποιώντας εξελιγμένα μοντέλα μηχανικής όρασης. Τέλος, οι συνθετικές εικόνες διάφορων αντικειμένων συνδυάζονται τυχαία για τη δημιουργία εικόνων που προσομοιώνουν τη συνύπαρξη πολλαπλών αντικειμένων.

6.3. Εκπαίδευση και Αξιολόγηση Μοντέλων Ανίχνευσης Αντικειμένων YOLO

Για την υλοποίηση της αναγνώρισης σκηνής χρησιμοποιούνται μοντέλα μηχανικής μάθησης YOLO με ικανότητα ανίχνευσης αντικειμένων με μεγάλη ακρίβεια σε πραγματικό χρόνο, ιδιότητες οι οποίες το καθιστούν ιδανικό για χρήση εντός του επιπέδου παραγωγής. Συνολικά, εκπαιδεύονται τρία μοντέλα πανομοιότυπων παραμέτρων με διαφορετικά σύνολα δεδομένων, με σκοπό την εξαγωγή συμπερασμάτων μέσω της σύγκρισής τους για τη μέθοδο δημιουργίας των συνθετικών δεδομένων. Ένα μοντέλο εκπαιδεύεται με το σύνολο δεδομένων που παράχθηκε αποτελούμενο από 22 χιλιάδες συνθετικές εικόνες. Ένα μοντέλο εκπαιδεύτηκε με αληθινές εικόνες των υπό μελέτη αντικειμένων που ελήφθησαν σε διάφορες τοποθεσίες, συμπεριλαμβανομένων το ΕΤΚ του Εθνικού Μετσόβιου Πολυτεχνείου. Το τελευταίο μοντέλο εκπαιδεύτηκε αρχικά με συνθετικά δεδομένα και έπειτα προσαρμόστηκε με εκπαίδευση σε αληθινά δεδομένα. Η εκπαίδευση του κάθε μοντέλου διήρκησε λιγότερες από 3 ώρες.

Τα τρία μοντέλα αξιολογήθηκαν ως προς ένα σύνολο επικύρωσης το οποίο αποτελείται από αληθινές εικόνες, έτσι ώστε να εκτιμηθεί η επίδοση του κάθε μοντέλου σε αληθινές συνθήκες. Υπολογίστηκαν και αναλύθηκαν διάφορες μετρικές που αφορούν την ακρίβεια των προβλέψεων των μοντέλων για τις εικόνες του συνόλου επικύρωσης. Η σύγκριση της επίδοσης του κάθε μοντέλου αποδεικνύει ότι η εκπαίδευση με αποκλειστικά συνθετικά δεδομένα προσφέρει αρκετά χαμηλότερη ακρίβεια από τη χρήση αληθινών δεδομένων. Η χαμηλή επίδοση της πρώτης μεθόδου έγκειται στη χαμηλή ακρίβεια ανίχνευσης εργαλειομηχανών. Ωστόσο, η προσαρμογή με αληθινά δεδομένα ενός μοντέλου που έχει προεκπαιδευτεί με μεγάλο όγκο συνθετικών δεδομένων έχει ως αποτέλεσμα την επίτευξη της υψηλότερης επίδοσης και ακρίβειας. Συγκεκριμένα, η μέθοδος αυτή πετυχαίνει αύξηση 3.1% της συνολική μέσης ακρίβειας ή mAP σε σύγκριση με την χρήση μόνο αληθινών δεδομένων.

6.4. Μελλοντική Μελέτη

Η χρήση αποκλειστικά συνθετικών δεδομένων για την εκπαίδευση ενός μοντέλου ανίχνευσης αντικειμένων εμφάνισε σημαντικά χαμηλότερη επίδοση σε σχέση με τη χρήση αληθινών δεδομένων. Η χαμηλή ακρίβεια του μοντέλου οφείλεται στη διαφορά των κατανομών των συνθετικών δεδομένων και των αληθινών δεδομένων του συνόλου επικύρωσης. Οι συνθετικές εικόνες δεν προσομοιώνουν επαρκώς τις πραγματικές συνθήκες που επικρατούν στο επίπεδο παραγωγής, με αποτέλεσμα την αδυναμία του μοντέλου να επιδόσει με ακρίβεια σε αληθινές καταστάσεις. Η αύξηση της επίδοσης δύναται να επιτευχθεί με χρήση πιο ποιοτικών συνθετικών δεδομένων που ταιριάζουν περισσότερο στην αληθινή κατανομή. Παρακάτω, προτείνονται μερικές πιθανές μέθοδοι για την δημιουργία συνθετικών συνόλων δεδομένων μεγαλύτερης ποιότητας και αληθοφάνειας.

6.4.1. Ποιοτικός Έλεγχος Παραγόμενων Εικόνων

Μια απλή μέθοδος για την αύξηση της συνολικής ποιότητας του συνθετικού συνόλου δεδομένων περιλαμβάνει την απόρριψη παραγόμενων εικόνων κακής ποιότητας. Η μέθοδος δημιουργίας του συνθετικού συνόλου που εφαρμόστηκε στην παρούσα μελέτη δεν περιλαμβάνει κανένα στάδιο ποιοτικού ελέγχου των εικόνων που παράγονται με το μοντέλο διάχυσης. Για κάθε κατηγορία, οι περιγραφές κειμένου σχεδιάστηκαν λεπτομερώς και μέσω πολλών δοκιμών έως που κρίθηκαν ικανοποιητικές για ένα μικρό δείγμα εικόνων. Ωστόσο, λόγω της στοχαστικής φύσης της αρχής λειτουργίας των μοντέλων διάχυσης και του μεγάλου πλήθους εικόνων που παράγονται, μερικές από τις εικόνες που δημιουργούνται δεν τηρούν πιστά τα γεωμετρικά χαρακτηριστικά των αντικειμένων ή έχουν χαμηλή σημασιολογική ποιότητα. Η εκπαίδευση ενός μοντέλου με δεδομένα χαμηλής ποιότητας επιδεινώνει την τελική ακρίβεια του και την επίδοσή του σε αληθινές συνθήκες.

Ο ποιοτικός έλεγχος των παραγόμενων εικόνων μπορεί να εφαρμοσθεί με διάφορους τρόπους. Για μικρότερα σύνολα δεδομένων, η διαδικασία απόρριψης των εικόνων χαμηλής ποιότητας μπορεί να γίνει από άνθρωπο με βάση την κρίση του. Η προσέγγιση αυτή είναι απλούστατη και δεν απαιτεί

επιπλέον εργασία πέρα από τον έλεγχο κάθε εικόνας. Ωστόσο, η ανάγκη ανθρώπινης παρέμβασης περιορίζει το μέγεθος του συνόλου για το οποίο μια τέτοια μέθοδος είναι πρακτική καθώς ο απαιτούμενος χρόνος είναι ανάλογος του πλήθους εικόνων. Επιπλέον, μια τέτοια προσέγγιση δεν επιτρέπει την αυτοματοποίηση της διαδικασίας δημιουργίας του συνθετικού συνόλου, η οποία αποτελεί ένα από τα πιο σημαντικά πλεονεκτήματα της. Συνεπώς, απαιτείται μια πιο αποδοτική εναλλακτική.

Μια δυνατή προσέγγιση έγκειται στην αυτοματοποίηση της παραπάνω διαδικασίας με τη χρήση μοντέλων μηχανικής μάθησης. Η διάκριση των εικόνων από άνθρωπο με βάση τη ποιότητά τους μπορεί να διεξαχθεί δοκιμαστικά για ένα συνθετικό σύνολο δεδομένων με σκοπό την καταγραφή της αντίστοιχης βαθμολογίας ποιότητας κάθε εικόνας. Τα ζεύγη εικόνων-βαθμολογιών μπορούν να χρησιμοποιηθούν για την εκπαίδευση ενός ταξινομητή ο οποίος διακρίνει τις εικόνες εισόδου σε αποδεκτές και μη αποδεκτές με βάση την ποιότητά τους. Ένα τέτοιο μοντέλο θα μπορούσε να λειτουργεί αυτόματα και με πολύ μεγαλύτερη ταχύτητα από έναν άνθρωπο, επιτρέποντας την ενσωμάτωσή του ποιοτικού ελέγχου εντός της διαδικασίας δημιουργίας συνθετικών συνόλων δεδομένων χωρίς βλάβη στην αυτόματη εκτέλεσή της.

6.4.2. Προσαρμογή Στυλ Μέσω GAN

Η παραπάνω λογική μπορεί να επεκταθεί με την εφαρμογή μεθόδων προσαρμογής στυλ μέσω παραγωγικών αντιπαραθετικών δικτύων. Ένα GAN μπορεί να εκπαιδευτεί έτσι ώστε να μετασχηματίζει συνθετικές εικόνες, αυξάνοντας την αληθοφάνεια τους, σε εικόνες που μοιάζουν περισσότερο με τις διαθέσιμες αληθινές εικόνες. Το παραγωγικό δίκτυο συνιστάται από ένα U-Net το οποίο δέχεται ως είσοδο συνθετικές εικόνες και προσπαθεί να εξάγει εικόνες οι οποίες ταξινομούνται ως αληθινές από ένα διαχωριστικό δίκτυο. Το διαχωριστικό δίκτυο πρόκεται για δυαδικό ταξινομητή ο οποίος διακρίνει τις εισόδους σε αληθινές ή συνθετικές. Η αποτελεσματική εκπαίδευση ενός τέτοιου μοντέλου επιτρέπει την μεταμόρφωση των συνθετικών εικόνων σε πιο αληθοφανείς εκδοχές τους. Αντίστοιχες προσεγγίσεις έχουν εφαρμοσθεί επιτυχώς για τη δημιουργία συνθετικών συνόλων δεδομένων για εφαρμογές ανίχνευσης αντικειμένων και σημασιολογικής κατάτμησης [88] [89] [35]. Η μέθοδος αυτή, σε αντίθεση με τον ποιοτικό έλεγχο, διατηρεί όλες τις παραγόμενες εικόνες αλλά απαιτεί τη χρήση αληθινών δεδομένων για εκπαίδευση.

6.4.3. Προσαρμογή Μοντέλου Διάχυσης με Αληθινές Εικόνες

Στην παρούσα μελέτη χρησιμοποιήθηκε το βασικό μοντέλο v1.5 του Stable Diffusion, το οποίο έχει εκπαιδευτεί στο σύνολο γενικής χρήσης LAION-5B, δίχως επιπλέον εκπαίδευση σε ειδικό μηχανολογικό εξοπλισμό. Ως εκ τούτου, όπως αναλύθηκε στην ενότητα 5.5.1, η ποιότητα των εικόνων μηχανολογικού εξοπλισμού είναι χαμηλότερη λόγω της ανομοιόμορφης κατανομής των κατηγοριών στα δεδομένα εκπαίδευσης του μοντέλου διάχυσης. Μια λύση σε αυτό το πρόβλημα έγκειται στην προσαρμογή του μοντέλου διάχυσης με εξειδικευμένα δεδομένα [78] [77] [79]. Η προσέγγιση αυτή, αντίστοιχα με αυτή των GAN, απαιτεί ένα αντιπροσωπευτικό σύνολο αληθινών δεδομένων για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Το πλεονέκτημα αυτής της μεθόδου έγκειται στην ενσωμάτωση της αύξησης της ποιότητας των παραγόμενων εικόνων εντός της φάσης της σύνθεσης. Η προσαρμογή του μοντέλου διάχυσης ενσωματώνει την βελτίωση της ποιότητας των συνθετικών εικόνων των υπό μελέτη αντικειμένων στο στάδιο της παραγωγής τους, δίχως την ανάγκη περαιτέρω ελέγχου ή επεξεργασιας.

Βιβλιογραφία

[1] M. N. O. Sadiku, A. J. Ajayi-Majebi and P. O. Adebo, "Robotic Automation in Manufacturing," in *Emerging Technologies in Manufacturing*, Cham, Springer International Publishing, 2023, p. 33–47.

[2] S. J. Hu, "Evolving Paradigms of Manufacturing: From Mass Production to Mass Customization and Personalization," *Procedia CIRP*, vol. 7, pp. 3-8, 2013.

[3] M. Zafarzadeh, M. Wiktorsson and J. B. Hauge, "A Systematic Review on Technologies for Data-Driven Production Logistics: Their Role from a Holistic and Value Creation Perspective," *Logistics*, vol. 5, p. 24, April 2021.

[4] V. Leso, L. Fontana and I. Iavicoli, "The occupational health and safety dimension of Industry 4.0," *La Medicina del Lavoro*, vol. 110, pp. 327-338, 2018.

[5] Y. Lu, X. Xu and L. Wang, "Smart manufacturing process and system automation – A critical review of the standards and envisioned scenarios," *Journal of Manufacturing Systems*, vol. 56, p. 312–325, July 2020.

[6] M. Wang, P. K.-Y. Wong, H. Luo, S. Kumar, V.-S. K. Delhi and J. C.-P. Cheng, "Predicting Safety Hazards Among Construction Workers and Equipment Using Computer Vision and Deep Learning Techniques," in *Proceedings of the International Symposium on Automation and Robotics in Construction (IAARC)*, 2019.

[7] M. M. Adrita, A. Brem, D. O'Sullivan, E. Allen and K. Bruton, "Methodology for Data-Informed Process Improvement to Enable Automated Manufacturing in Current Manual Processes," *Applied Sciences*, vol. 11, p. 3889, April 2021.

[8] B. Kanchana, R. Peiris, D. Perera, D. Jayasinghe and D. Kasthurirathna, "Computer Vision for Autonomous Driving," in *2021 3rd International Conference on Advancements in Computing (ICAC)*, 2021.

[9] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proceedings of the IEEE (2021)*, vol. 109, p. 820–838, 2 May 2020.

[10] S. Das, A. Nag, D. Adhikary, R. J. Ram, B. R. Aravind, S. K. Ojha and G. M. Hegde, "Computer Vision-based Social Distancing Surveillance Solution with Optional Automated Camera Calibration for Large Scale Deployment," April 2021.

[11] A. Naumann, F. Hertlein, L. Dörr, S. Thoma and K. Furmans, "Literature Review: Computer Vision Applications in Transportation Logistics and Warehousing," April 2023.

[12] E. R. Kandel, J. H. Schwartz and T. M. Jessell, Principles of Neural Science, McGraw-Hill Education, 2012.

[13] H. Jin, "Seeing the Unseen: Errors and Bias in Visual Datasets," November 2022.

[14] T. Wang, Y. Chen, M. Qiao and H. Snoussi, "A fast and robust convolutional neural network-based defect detection model in product quality control," *The International Journal of Advanced Manufacturing Technology*, vol. 94, p. 3465–3471, August 2017.

[15] A. Kazemian, X. Yuan, O. Davtalab and B. Khoshnevis, "Computer vision for real-time extrusion quality monitoring and control in robotic construction," *Automation in Construction*, vol. 101, pp. 92-98, 2019.

[16] M. Ferguson, R. Ak, Y.-T. T. Lee and K. H. Law, "Detection and Segmentation of Manufacturing Defects with Convolutional Neural Networks and Transfer Learning," August 2018.

[17] D. Wu, C. Jennings, J. Terpenny, R. X. Gao and S. Kumara, "A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests," *Journal of Manufacturing Science and Engineering*, vol. 139, April 2017.

[18] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, *You Only Look Once: Unified, Real-Time Object Detection*, 2016.

[19] P. Felzenszwalb, R. Girshick, D. Mcallester and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, pp. 1627-45, September 2010.

[20] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," November 2013.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions," September 2014.

[22] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," December 2016.

[23] J. Redmon, *Darknet: Open Source Neural Networks in C,* 2013–2016.

[24] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," April 2018.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," December 2016.

[26] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," April 2020.

[27] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen and J.-W. Hsieh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," November 2019.

[28] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," March 2018.

[29] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao and F. Shen, "Image Data Augmentation for Deep Learning: A Survey," April 2022.

[30] T. Ni, Y. Shen and F. Yu, "The Influence of Artificial Intelligence on Art Design in the Digital Age," *Scientific Programming*, vol. 2021, p. 4838957, 2021.

[31] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan and H. Fu, "Transformers in Medical Imaging: A Survey," January 2022.

[32] N. K. Singh and K. Raza, "Medical Image Generation Using Generative Adversarial Networks: A Review," in *Health Informatics: A Computational Perspective in Healthcare*, R. Patgiri, A. Biswas and P. Roy, Eds., Singapore, Springer Singapore, 2021, p. 77–96.

[33] J. W. Anderson, M. Ziolkowski, K. Kennedy and A. W. Apon, "Synthetic Image Data for Deep Learning," December 2022.

[34] J. M. Rožanec, P. Zajec, S. Theodoropoulos, E. Koehorst, B. Fortuna and D. Mladenić, "Synthetic Data Augmentation Using GAN For Improved Automated Visual Inspection," December 2022.

[35] W. Xu, C. Long, R. Wang and G. Wang, "DRB-GAN: A Dynamic ResBlock Generative Adversarial Network for Artistic Style Transfer," August 2021.

[36] P. N. Deelaka, "Neural Artistic Style Transfer with Conditional Adversaria," February 2023.

[37] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," March 2017.

[38] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] S. Wang, Y. Du, X. Guo, B. Pan, Z. Qin and L. Zhao, "Controllable Data Generation by Deep Learning: A Review," July 2022.

[40] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends in Machine Learning: Vol. 12 (2019): No. 4, pp 307-392,* vol. 12, p. 307–392, June 2019.

[41] W. Mormul, *Variational Autoencoder implementation with Tensorflow.,* GitHub, 2013.

[42] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks," June 2014.

[43] T. S. Silva, "A Short Introduction to Generative Adversarial Networks," *https://sthalles.github.io,* 2017.

[44] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," December 2018.

[45] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," December 2019.

[46] J. Ho, A. Jain and P. Abbeel, "Denoising Diffusion Probabilistic Models," June 2020.

[47] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," May 2021.

[48] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever, "Zero-Shot Text-to-Image Generation," February 2021.

[49] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," April 2022.

[50] J. Oppenlaender, "The Creativity of Text-to-Image Generation," 13 November 2022.

[51] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," December 2021.

[52] E. Strubell, A. Ganesh and A. McCallum, "Energy and Policy Considerations for Modern Deep Learning Research," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 34, p. 13693–13696, April 2020.

[53] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier and J. Dean, "Carbon Emissions and Large Neural Network Training," April 2021.

[54] Y. Patel, S. Appalaraju and R. Manmatha, "Deep Perceptual Compression," July 2019.

[55] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015.

[56] L. Zhang and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," February 2023.

[57] T. Park, M.-Y. Liu, T.-C. Wang and J.-Y. Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," *CVPR 2019,* March 2019.

[58] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," November 2016.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," June 2017.

[60] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals and J. Carreira, "Perceiver: General Perception with Iterative Attention," March 2021.

[61] T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," October 2017.

[62] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser and J. Xiao, "LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop," June 2015.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[64] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev and A. Komatsuzaki, "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," November 2021.

[65] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vols. PAMI-8, pp. 679-698, 1986.

[66] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler and V. Koltun, *Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer*, arXiv, 2019.

[67] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter and G. Shakhnarovich, "DIODE: A Dense Indoor and Outdoor DEpth Dataset," August 2019.

[68] G. Gu, B. Ko, S. Go, S.-H. Lee, J. Lee and M. Shin, "Towards Light-weight and Real-time Line Segment Detection," June 2021.

[69] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," April 2015.

[70] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, "Scene Parsing through ADE20K Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[71] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," December 2018.

[72] M. Wang and W. Deng, "Deep Visual Domain Adaptation: A Survey," *Neurocomputing, 2018, 312: 135-153*, vol. 312, p. 135–153, 10 October 2018.

[73] A. Farahani, S. Voghoei, K. Rasheed and H. R. Arabnia, "A Brief Review of Domain Adaptation," October 2020.

[74] X. Liu, C. Yoo, F. Xing, H. Oh, G. E. Fakhri, J.-W. Kang and J. Woo, "Deep Unsupervised Domain Adaptation: A Review of Recent Advances and Perspectives," August 2022.

[75] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba and P. Abbeel, "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World," March 2017.

[76] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi and D. J. Fleet, "Synthetic Data from Diffusion Models Improves ImageNet Classification," April 2023.

[77] H. Ali, S. Murad and Z. Shah, "Spot the Fake Lungs: Generating Synthetic Medical Images Using Neural Diffusion Models," in *Communications in Computer and Information Science*, Springer Nature Switzerland, 2023, p. 32–39.

[78] P. A. Moghadam, S. Van Dalen, K. C. Martin, J. Lennerz, S. Yip, H. Farahani and A. Bashashati, "A Morphology Focused Diffusion Probabilistic Model for Synthesis of Histopathology Images," September 2022.

[79] F. Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather and D. Truhn, "Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation," November 2022.

[80] A. Stöckl, "Evaluating a Synthetic Image Dataset Generated with Stable Diffusion," November 2022.

[81] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," October 2022.

[82] F. Vargas, W. Grathwohl and A. Doucet, "Denoising Diffusion Samplers," *In The Eleventh In The Eleventh International Conference on Learning Representations, 2023*, February 2023.

[83] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon and S. Birchfield, "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization," April 2018.

[84] F. Sadeghi and S. Levine, "CAD2RL: Real Single-Image Flight without a Single Real Image," November 2016.

[85] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy and T. Brox, "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation," 7 June 2015.

[86] D. Dwibedi, I. Misra and M. Hebert, "Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection," August 2017.

[87] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics,* vol. 4, pp. 1-17, 1964.

[88] A. Dundar, M.-Y. Liu, T.-C. Wang, J. Zedlewski and J. Kautz, "Domain Stylization: A Strong, Simple Baseline for Synthetic to Real Image Domain Adaptation," July 2018.

[89] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros and T. Darrell, "CyCADA: Cycle-Consistent Adversarial Domain Adaptation," November 2017.

[90] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane and M. Jagersand, "U²-Net: Going Deeper with Nested U-Structure for Salient Object Detection," *Pattern Recognition,* vol. 106, p. 107404, 18 October 2020.

[91] J. Yaffe, Lathe, 2011,
https://www.3dcontentcentral.com/Download-Model.aspx?catalogid=171&id=225591
Accessed on August 8, 2023.

[92] tasraven, Forklift,
https://skfb.ly/oFI7G,
Licensed under Creative Commons Attribution 4.0 (CC BY 4.0).

[93] SusiePhilpott, Milling Machine,
https://skfb.ly/6ZEWZ,
Licensed under Creative Commons Attribution 4.0 (CC BY 4.0).

[94] nemapets, Lathe,
https://thangs.com/designer/nemapet/3d-model/MyLathe.stl-8290,
Accessed on August 8, 2023.

[95] Haas Automation, Inc., DT-2 3D Model,
https://www.haascnc.com/machines/vertical-mills/drill-tap-mill/models/dt-2.html,
Accessed on August 5, 2023.

[96] Haas Automation, Inc., ST-40 3D Model,
https://www.haascnc.com/machines/lathes/st/models/large-through-bore/st-40.html,
Accessed on August 5, 2023.

[97] Haas Automation, Inc., TL-1 3D Model,
https://www.haascnc.com/machines/lathes/toolroom-lathe/models/tl-1.html,
Accessed on July 26, 2023.

[98] Haas Automation, Inc., VF-1 3D Model,
https://www.haascnc.com/machines/vertical-mills/vf-series/models/small/vf-1.html,
Accessed on August 5, 2023.

[99] E. Gidoni, Bridgeport-Type Milling Machine, 2023,
https://grabcad.com/library/bridgeport-type-milling-machine-1/details,
Accessed on August 8, 2023.

Παράρτημα

Α. Δημιουργία Συνθετικών Εικόνων Ανθρώπων

Στο Κεφάλαιο 3 εξηγήθηκε η διαδικασία δημιουργίας του συνθετικού συνόλου δεδομένων και αναλύθηκε λεπτομερώς η διαδικασία παραγωγής των εικόνων του μηχανολογικού εξοπλισμού. Η παραγωγή συνθετικών εικόνων ανθρώπων αποτελείται από διαφορετικές φάσεις οι οποίες έχουν ομοιότητες με τα διάφορα βήματα παραγωγής των εικόνων εξοπλισμού. Η ροή εργασίας περιγράφεται από τα κάτωθι βήματα:

1. Λήψη φωτογραφιών ανθρώπων – Κατά το πρώτο στάδιο της διαδικασίας, λαμβάνονται φωτογραφίες ανθρώπων με διάφορες στάσεις σώματος από πολλές οπτικές γωνίες. Για την τρέχουσα μελέτη, ελήφθησαν ξεχωριστά εικόνες δύο ανθρώπων, ενός άνδρα και μιας γυναίκας, σε τρεις στάσεις σώματος:

    - Στάση προσοχής, με το σώμα τεντωμένο σε όρθια θέση, τα πόδια ενωμένα και τα χέρια τεντωμένα προς τα κάτω να εφάπτονται στα πλευρά και τους μηρούς.

    - Στάση βαδίσματος, με το σώμα σε όρθια θέση με τα πόδια και τα χέρια να βρίσκονται σε έκταση, αντίστοιχα με τη στάση σώματος κατά το περπάτημα.

    - Καθιστή στάση, όπου ο άνθρωπος κάθεται σε καρέκλα με τα χέρια σε έκταση μπροστά, σαν να αναπαύονται σε μια επιφάνεια όπως ένα γραφείο.

    Η λήψη εικόνων γίνεται για μια πλήρη περιστροφή του ανθρώπου, με σκοπό την καταγραφή της στάσης του σώματος από όλες τις δυνατές γωνίες. Η φάση αυτή έχει παρόμοιο σκοπό με την τυχαιοποίηση της εικονικής σκηνής στη δημιουργία των εικόνων μηχανημάτων. Μέσω της μεταβολής διάφορων παραμέτρων, όπως η στάση σώματος και η οπτική γωνία, εισάγεται μεταβλητότητα η οποία δύναται να αυξήσει την ευρωστία του μοντέλου ανίχνευσης αντικειμένων. Επιπλέον, η λήψη εικόνων για ανδρικά και γυναικεία σώματα συμβάλλει στην ισότιμη εκπροσώπηση των δύο φύλων στο τελικό σύνολο δεδομένων. Στο Σχήμα Α.1 μπορούν να φανούν παραδείγματα εικόνων των διάφορων στάσεων.

Σχήμα Α.1. Παραδείγματα εικόνων ανθρώπου με διάφορες στάσεις σώματος.

2. Εκτίμηση στάσης σώματος – Η στάση του σώματος των ανθρώπων εντός των εικόνων εκτιμάται με τη χρήση του OpenPose [71]. Το στάδιο αυτό είναι απαραίτητο καθώς εκτελεί προεπεξεργασία των εικόνων με σκοπό την εξαγωγή της χρήσιμης πληροφορίας της στάσης του σώματος για χρήση ως συνθήκη για την παραγωγή εικόνας μέσω του ControlNet.

3. Παραγωγή εικόνων βάσει περιγραφών κειμένου – Η στάση του σώματος εισάγεται μέσω ControlNet ως συνθήκη για τον έλεγχο της παραγωγής των εικόνων μέσω του Stable Diffusion. Για κάθε στάση σώματος, επιλέγονται διαφορετικές περιγραφές κειμένου οι οποίες περιγράφουν με μεγαλύτερη ακρίβεια την επιθυμητή σύνθεση της εικόνας. Η διαδικασία επαναλαμβάνεται ξεχωριστά και για τα δύο φύλα, έτσι ώστε να προσδιοριστεί σαφώς το φύλο του ανθρώπου στην παραγόμενη εικόνα. Τελικά, δεν παράχθηκαν εικόνες για καθιστή στάση σώματος. Οι αναλυτικές περιγραφές κειμένου για κάθε κατηγορία βρίσκονται στον Πίνακα Α.1.

Πίνακας Α.1.

Οι περιγραφές κειμένου που χρησιμοποιήθηκαν για την παραγωγή των εικόνων ανθρώπων μέσω του Stable Diffusion.

	Περιγραφή Κειμένου	Αρνητική Περιγραφή Κειμένου
Άνδρας	raw photo of a male factory worker **(walking)**, *facing away from camera, [[back side, back turned against the camera, shot from behind]]*, ((well lit, soft lighting)), (flat simple background), ((neutral expression)), [[[hard hat, safety vest]]]	(((deformed))), blurry, bad anatomy, disfigured, poorly drawn face, mutation, mutated, (extra_limb), (ugly), (poorly drawn hands), fused fingers, messy drawing, broken legs, (mutated hands and fingers:1.5), (long body :1.3), (mutation, poorly drawn :1.2), black-white, bad anatomy, disfigured, malformed, mutated, anatomical nonsense, text font ui, error, malformed hands, long neck, blurred, lowers, low res, bad anatomy, bad proportions, bad shadow, uncoordinated body, unnatural body, bad hands, fused hand, missing hand, disappearing arms, disappearing thigh, disappearing calf, disappearing legs, fused ears, bad ears, poorly drawn ears, extra ears, missing ears, old photo, low res, black and white, black and white filter, colorless, (((objects in background, detailed background, harsh shadows, intense lighting, moody lighting, windows, studio lighting, professional lighting, illustration, art, professional photography, depth of field, female, woman, sexy, smiling, holding item in hands, beige clothes)))
Γυναίκα	raw photo of a female factory worker **(walking)** in an empty room, *[[facing away from camera, back side, back turned against the camera, shot from behind]]*, ((well lit, soft lighting)), (flat simple background), ((neutral expression)), [[[hard hat, safety vest]]]	(((deformed))), blurry, bad anatomy, disfigured, poorly drawn face, mutation, mutated, (extra_limb), (ugly), (poorly drawn hands), fused fingers, messy drawing, broken legs, (mutated hands and fingers:1.5), (long body :1.3), (mutation, poorly drawn :1.2), bad anatomy, disfigured, malformed, mutated, anatomical nonsense, text font ui, error, malformed hands, long neck, blurred, lowers, low res, bad anatomy, bad proportions, bad shadow, uncoordinated body, unnatural body, bad hands, fused hand, missing hand, disappearing arms, disappearing thigh, disappearing calf, disappearing legs, fused ears, bad ears, poorly drawn ears, extra ears, missing ears, old photo, low res, grayscale, colorless, (((objects in background, detailed background, harsh shadows, intense lighting, moody lighting, windows, studio lighting, professional lighting, illustration, art, professional photography, depth of field, sexy woman, person smiling, happy person, holding item in hands, holding an object, objects in foreground, beige clothes))), woman wearing skirt

Ο προσδιορισμός της στάσης σώματος γίνεται με την προσθήκη περιγραφής της αντίστοιχης στάσης στο κειμένου. Επιπλέον, ειδικοί όροι προστέθηκαν στις περιγραφές κειμένου για οπτικές γωνίες όπου ο άνθρωπος έχει γυρισμένη την πλάτη στην κάμερα για την αποφυγή παραγωγής εικόνων κακής ποιότητας. Στον παραπάνω πίνακα, οι όροι που αφορούν τη **στάση σώματος** αποτυπώνονται με έντονα γράμματα ενώ οι όροι που προσδιορίζουν την *οπτική γωνία* αποτυπώνονται με πλάγια γράμματα.

Οι παραγόμενες εικόνες, προτού χρησιμοποιηθούν για τη δημιουργία νέων συνθετικών σκηνών, δέχονται επιπλέον επεξεργασία για αφαίρεση του φόντου. Η αφαίρεση του φόντου γίνεται με τη χρήση του μοντέλου μηχανικής μάθησης Rembg [90] το οποίο αποτελείται από ένα συνελικτικό U-Net. Το Rembg δέχεται ως είσοδο έγχρωμες εικόνες και εξάγει εικόνες εντός των οποίων οι περιοχές που ανήκουν στο φόντο έχουν μαύρο χρώμα ενώ οι περιοχές που αντιστοιχούν στο αντικείμενο αποτυπώνονται με λευκό χρώμα. Οι εικόνες αυτές λέγονται μάσκες και χρησιμεύουν για την αφαίρεση του φόντου από τις συνθετικές εικόνες των ανθρώπων. Τέλος, με βάση τη διαφάνεια υπολογίζονται αυτόματα τα πλαίσια οριοθέτησης των ανθρώπων εντός της κάθε εικόνας. Η παραπάνω διαδικασία αποτυπώνεται στο Σχήμα Α.2.

Σχήμα Α.2. Αφαίρεση φόντου και υπολογισμός πλαισίου οριοθέτησης στις συνθετικές εικόνες ανθρώπων.

Στις παραγόμενες εικόνες για καθιστή στάση σώματος, διάφορα αντικείμενα, όπως γραφεία, απεικονίζονταν μπροστά από τους ανθρώπους, κρύβοντάς τους, τα οποία δεν αφαιρούνταν μαζί με το φόντο με αποτέλεσμα τον λάθος υπολογισμό των πλαισίων οριοθέτησης. Συνεπώς, δεν χρησιμοποιήθηκαν για τη δημιουργία του συνθετικού συνόλου δεδομένων.

Β. Αναλυτικές Παράμετροι Τυχαιοποίησης Σκηνής

Ο Πίνακας Β.1 περιλαμβάνει τις ακραίες τιμές των παραμέτρων οι οποίες μεταβάλλονται κατά την τυχαιοποίηση σκηνής της διαδικασίας δημιουργίας συνθετικών εικόνων μηχανολογικού εξοπλισμού. Οι τιμές αναγράφονται στη μορφή $[m, M]$ όπου m και M είναι η ελάχιστη και η μέγιστη τιμή της παραμέτρου, αντίστοιχα.

Πίνακας Β.1

Ακραίες τιμές παραμέτρων που μεταβάλλονται κατά την τυχαιοποίηση σκηνής για κάθε κατηγορία αντικειμένων.

	Παράμετροι				
Κατηγορία	Αζιμούθιο Κάμερας (°)	Απόσταση Κάμερας (m)	Ύψος Κάμερας (m)	Εστιακή Απόσταση (mm)	Κλίμακα Αντικειμένου
CNC Τόρνος	[-45, 45]	[2.8, 5]	[0.7, 1.7]	[30, 60]	[0.9, 1.2]
CNC Φρέζα	[-45, 45]	[3.3, 7]	[0.76, 1.36]	[30, 70]	[0.9, 1.2]
Τόρνος	[-45, 45]	[2.7, 5]	[0.76, 1.76]	[30, 60]	[1, 1.2]
Φρέζα	[-45, 45]	[2.2, 6]	[0.81, 2.06]	[30, 90]	[0.8, 1]
Περονοφόρο	[0, 360]	[4.2, 8]	[0.5, 1.5]	[30, 75]	[0.9, 1.2]
Παλέτα	[-90, 90]	[0, 5]	[1.5, 5]	[17.5, 116.5]	[1, 1.2]