ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΚΑΙ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ –
ΜΗΧΑΝΙΚΩΝ ΓΕΩΠΛΗΡΟΦΟΡΙΚΗΣ (ΣΑΤΜ –ΜΓ)
ΤΟΜΕΑΣ ΤΟΠΟΓΡΑΦΙΑΣ
ΕΡΓΑΣΤΗΡΙΟ ΤΗΛΕΠΙΣΚΟΠΗΣΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΥΤΟΜΑΤΗ ΕΞΑΓΩΓΗ ΑΝΤΙΚΕΙΜΕΝΩΝ ΑΠΟ ΤΗΛΕΠΙΣΚΟΠΙΚΕΣ ΑΠΕΙΚΟΝΙΣΕΙΣ ΜΕ ΤΕΧΝΙΚΕΣ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ

ΝΟΜΙΚΟΣ ΝΙΚΟΛΑΟΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΚΑΡΑΝΤΖΑΛΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ
ΑΘΗΝΑ, ΙΟΥΛΙΟΣ 2024

Τριμελής επιτροπή

Καράντζαλος Κωνσταντίνος	Παπουτσής Ιωάννης	Δουλάμης Νικόλαος
Καθηγητής	Επίκουρος Καθηγητής	Καθηγητής

ΕΥΧΑΡΙΣΤΙΕΣ

Θα επιθυμούσα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της παρούσας διπλωματικής εργασίας κ.Καράντζαλο Κωνσταντίνο για τις κατευθυντήριες γραμμές που έθεσε σε όλη τα στάδια της εργασίας και για την πολύτιμη συνεργασία και αμέριστη προσοχή του. Χάρη στην εμπειρία και τη καθοδήγηση του, κατόρθωσα να ανταπεξέλθω στις προκλήσεις και να ολοκληρώσω την εργασία με επιτυχία.

Επιπρόσθετα θα ήθελα να εκφράσω ειλικρινά τις ευχαριστίες μου προς τους υποψήφιους διδάκτορες Βασίλειο Τσιρώνη και Αθηνά Ψάλτα για τη συνεχή επιστημονική υποστήριξη και τις πολύτιμες συμβουλές τους σε τεχνικά θέματα ζωτικής σημασίας για την υλοποίηση της διπλωματικής εργασίας. Η αφοσίωση και η ειλικρινής προθυμία τους να βοηθήσουν σε κάθε στάδιο της εργασίας υπήρξαν ανεκτίμητες.

Τέλος, δε μπορώ να παραλείψω να ευχαριστήσω τη σύζυγο μου και την οικογένεια μου που ήταν συνεχώς δίπλα μου σε κάθε ακαδημαϊκό μου βήμα, δείχνοντας έμπρακτα την υποστήριξη τους. Η υπομονή και η κατανόησή τους κατά τη διάρκεια αυτής της απαιτητικής περιόδου ήταν καθοριστικές για την επιτυχή ολοκλήρωση της διπλωματικής μου εργασίας.

ΠΕΡΙΛΗΨΗ

Η αναγνώριση, κατηγοριοποίηση, ταξινόμηση και περιγραφή των αντικειμένων που αποτυπώνονται σε αεροφωτογραφίες και δορυφορικές εικόνες αποτελούν βασικό αντικείμενο μελέτης και δράσης της κλασσικής Φωτοερμηνείας. Ωστόσο, αυτές οι διαδικασίες απαιτούν σημαντικό χρόνο και εξειδικευμένο προσωπικό για να ολοκληρωθούν με ικανοποιητικά αποτελέσματα.

Η αλματώδης ανάπτυξη της τεχνολογίας και η εδραίωση της τεχνητής νοημοσύνης σε πληθώρα επιστημονικών πεδίων, συμβάλλουν σε σημαντικό βαθμό στην διαμόρφωση νέων τεχνικών αναγνώρισης και κατανόησης του περιεχομένου των δορυφορικών εικόνων. Ειδικά η επιστήμη της βαθιάς μάθησης (Deep learning), η οποία αποτελεί ένα υποσύνολο της τεχνητής νοημοσύνης (Artificial Intelligence or AI) οδήγησε στην ανάπτυξη της Υπολογιστικής Όρασης (Computer Vision) και τομέων της όπως η ταξινόμηση εικόνων (δυαδική ή πολλαπλών κατηγοριών), η ανίχνευση οντοτήτων (object detection) και η τμηματοποίηση εικόνας (segmentation).

Η παρούσα διπλωματική εργασία έχει εκπονηθεί με σκοπό την έρευνα, την εξοικείωση και την εφαρμογή τεχνικών βαθιάς μάθησης με σκοπό την εξαγωγή οντοτήτων από τηλεπισκοπικές απεικονίσεις. Συγκεκριμένα, μελετήθηκε και χρησιμοποιήθηκε το μοντέλο Segment Anything το οποίο αποτελεί θεμελιώδες μοντέλο για κατάτμηση ή τμηματοποίηση εικόνας. Στα πλαίσια της διπλωματικής εργασίας, έχει υλοποιηθεί εφαρμογή η οποία επιτρέπει σε κάθε χρήστη, χωρίς να διαθέτει κάποια πρότερη γνώση, να ανιχνεύει και να εξάγει οντότητες σε πραγματικό χρόνο από δορυφορικές απεικονίσεις. Επιπρόσθετα, πραγματοποιήθηκε περαιτέρω εκπαίδευση (fine tuning) του συγκεκριμένου μοντέλου για την ανίχνευση συγκεκριμένων οντοτήτων ενδιαφέροντος που θα αναλυθεί σε επόμενο κεφάλαιο.

Τα αποτελέσματα της έρευνας δείχνουν σημαντική βελτίωση στην ταχύτητα εξαγωγής οντοτήτων ενδιαφέροντος σε σχέση με τις παραδοσιακές μεθόδους ψηφιοποίησης. Η διαδικασία είναι πλήρως αυτοματοποιημένη, δεν απαιτεί ειδική εμπειρία από το χρήστη και δύναται να αποτελέσει το εφαλτήριο για περαιτέρω επεξεργασία των δεδομένων όπως στη διαδικασία της χαρτοσύνθεσης. Οι οντότητες ενδιαφέροντος που εξάγονται από τηλεπισκοπικές απεικονίσεις δύναται να χρησιμοποιηθούν σε τομείς με σημαντικές πρακτικές εφαρμογές όπως η παρακολούθηση της αστικής ανάπτυξης καθώς και ο έλεγχος των δασικών εκτάσεων για πιθανή τους καταπάτηση.

ΛΕΞΕΙΣ –ΚΛΕΙΔΙΑ

Βαθιά Μάθηση, Υπολογιστική όραση, Κατάτμηση, Νευρωνικά Δίκτυα, Segment Anything, συνάρτηση κόστους, συνάρτηση ενεργοποίησης, αλγόριθμος βελτιστοποίησης, σημασιολογική κατάτμηση, ανίχνευση αντικειμένων.

ABSTRACT

The recognition, categorization, classification, and description of objects captured in aerial photographs and satellite images are fundamental aspects of classical photointerpretation. However, these processes require significant time and specialized personnel to be completed satisfactorily.

The rapid advancement of technology and the establishment of artificial intelligence across various scientific fields significantly contribute to the development of new techniques for recognizing and understanding the content of satellite images. Specifically, the field of deep learning, a subset of artificial intelligence (AI), has led to the advancement of computer vision and its subfields, such as image classification (binary or multi-class), object detection, and image segmentation.

This thesis was conducted with the aim of researching, familiarizing with, and applying deep learning techniques for the extraction of entities from remote sensing imagery. Specifically, the Segment Anything model, a foundational model for image segmentation, was studied and utilized. As part of the thesis, an application was developed that allows any user, without prior knowledge, to detect and extract entities in real-time from satellite imagery. Additionally, further training of this model was conducted for the detection of specific entities of interest, which will be analyzed in the following chapter.

The research results show significant improvements in the speed of extracting entities of interest compared to traditional digitization methods. The process is fully automated, does not require special expertise from the user, and can serve as a springboard for further data processing, such as map composition. The data of interest can be used in fields with significant practical applications, such as monitoring urban development and controlling forest areas for potential encroachment.

KEYWORDS

Deep Learning, Computer Vision, Segmentation, Neural Networks, Segment Anything, Loss Function, Activation Function, Optimizer, semantic segmentation, object detection.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Ι. ΠΙΝΑΚΑΣ ΣΧΗΜΑΤΩΝ

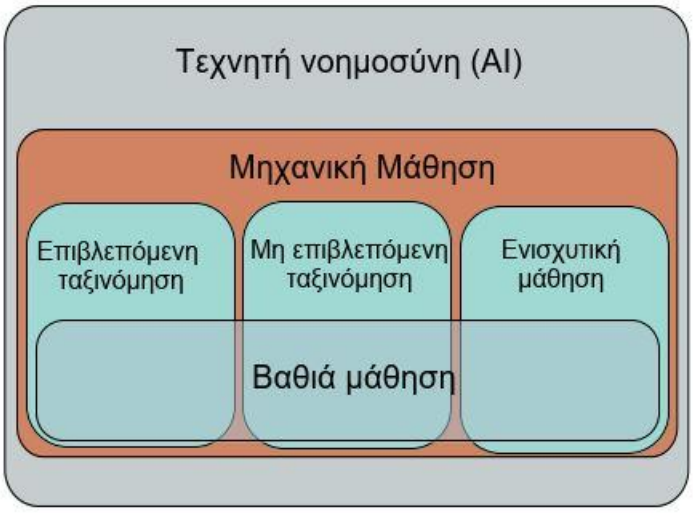
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΙΚΑ ΣΤΟΙΧΕΙΑ

Το παρόν κεφάλαιο εισάγει το ευρύτερο πλαίσιο στο οποίο θα κινηθεί η διπλωματική εργασία, καθορίζεται το αντικείμενο και ο σκοπός της και παρουσιάζονται οι διαδικασίες που ακολουθήθηκαν. Τέλος αναλύεται η δομή και το περιεχόμενο της.

Η τεχνητή νοημοσύνη (AI) έχει ως κύριο σκοπό τη δημιουργία συστημάτων που προσομοιώνουν την ανθρώπινη συμπεριφορά μέσω ενός ευρέως φάσματος προσεγγίσεων όπως η λογική και η πιθανολογική προσέγγιση. Η μηχανική μάθηση (machine learning) αποτελεί ένα υποσύνολο της τεχνητής νοημοσύνης το οποίο μαθαίνει να λαμβάνει αποφάσεις προσαρμόζοντας μαθηματικά μοντέλα σε δεδομένα. Ουσιαστικά, η μηχανική μάθηση μετατρέπει δεδομένα σε αριθμούς και αναζητά συσχετίσεις (patterns) μεταξύ αυτών των αριθμών. Έτσι δέχεται τα ορίσματα (inputs) και τα επιθυμητά αποτελέσματα (outputs) και καθορίζει τους κανόνες μεταξύ τους. Τα ορίσματα είναι γνωστά ως features ενώ τα αποτελέσματα ως labels. Η παραπάνω διαδικασία είναι γνωστή ως επιβλεπόμενη μάθηση (supervised learning). Η μηχανική μάθηση προτείνεται για δεδομένα σε δομημένη μορφή όπως πίνακες.

Η βαθιά μάθηση (deep learning) είναι ένας όρος που συνεχώς συγχέεται με τη μηχανική μάθηση αλλά στη πραγματικότητα αποτελεί ένα υποσύνολο της. Ένα βαθύ νευρωνικό δίκτυο αποτελεί ένα τύπο μοντέλου μηχανικής μάθησης και όταν αυτό τροφοδοτείται με δεδομένα αποκαλείται ως διαδικασία βαθιάς μάθησης. Η βαθιά μάθηση ασχολείται κυρίως με δεδομένα τα οποία δε βρίσκονται σε αυστηρή δομή όπως ο ήχος, τα κείμενα και οι εικόνες. Η βαθιά μάθηση, η μελέτη της οποίας αποτελεί βασικό συστατικό της παρούσας διπλωματικής έχει εισβάλλει στη καθημερινή μας ζωή σε πληθώρα τομέων. Οι σύγχρονες εφαρμογές της βαθιάς μάθησης αφορούν την υπολογιστική όραση (CV), την επεξεργασία φυσικής γλώσσας (NLP), την αναγνώριση μέσω ήχου και βίντεο (V/SP) καθώς και τομείς οικονομικών (F&B) . Η αναζήτηση εικόνων με βάση ένα συγκεκριμένο αντικείμενο στο διαδίκτυο αποτελεί αντικείμενο υπολογιστικής όρασης (computer vision). Η μετάφραση ενός κειμένου από μια γλώσσα σε άλλη γίνεται με βάση ένα αλγόριθμο επεξεργασίας φυσικής γλώσσας (Natural Language Processing). Αυτές οι εφαρμογές αλλά και πολλές άλλες είναι αποτέλεσμα της βαθιάς μάθησης. Στη παρούσα εργασία θα αναλυθούν κυρίως αντικείμενα υπολογιστικής όρασης όπως η ανίχνευση αντικειμένων (object detection) αλλά και η κατάτμηση εικόνας (image segmentation).

Η μηχανική μάθηση αλλά και η βαθιά μάθηση ως υποσύνολο της χωρίζεται σε τρεις κύριες κατηγορίες: την επιβλεπόμενη (supervised), τη μη επιβλεπόμενη (unsupervised) αλλά και την ενισχυτική μάθηση (reinforcement learning) όπως απεικονίζονται στο Σχήμα 1 παρακάτω.



Σχήμα 1 Συσχετισμοί μεταξύ Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης και Βαθιάς Μάθησης. (Πηγή : Understanding Deep Learning, 2023)

1.1 ΑΝΤΙΚΕΙΜΕΝΟ ΚΑΙ ΣΤΟΧΟΣ ΕΡΓΑΣΙΑΣ

Αντικείμενο της εργασίας αποτελεί η αποσαφήνιση βασικών εννοιών που σχετίζονται με τη βαθιά μάθηση και η εφαρμογή τεχνικών της με σκοπό την αυτοματοποιημένη εξαγωγή οντοτήτων από τηλεπισκοπικές απεικονίσεις. Συγκεκριμένα πραγματοποιείται εις βάθος μελέτη του προ-εκπαιδευμένου μοντέλου Segment Anything που σχετίζεται με τη κατάτμηση εικόνας , ενσωμάτωση του σε υλοποιημένη προγραμματιστική εφαρμογή , καθώς και περαιτέρω εκπαίδευση του μοντέλου σε δεδομένα του Εργαστηρίου Τηλεπισκόπησης με σκοπό την εξειδίκευση του σε συγκεκριμένες οντότητες ενδιαφέροντος. Τέλος πραγματοποιείται αξιολόγηση του εκπαιδευμένου μοντέλου σε δεδομένα ελέγχου με μια σειρά μέτρων αξιολόγησης (evaluation metrics) αλλά και μέσω οπτικών ελέγχων .

1.2 ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

Στο σημείο αυτό, κρίνεται σκόπιμο να παρουσιαστούν συνοπτικά τα κεφάλαια αλλά και τον απώτερο σκοπό τους για τη καλύτερη κατανόηση της παρούσας διπλωματικής εργασίας.

Στο πρώτο (1^0) κεφάλαιο, δηλαδή στα Εισαγωγικά Στοιχεία παρουσιάζεται το ευρύτερο πλαίσιο στο οποίο εντάσσεται η εργασία και καθορίζεται το αντικείμενο της αλλά και η δομή της.

Στο δεύτερο (2^0) κεφάλαιο, επιχειρείται θεωρητική αποσαφήνιση εννοιών που εισήχθησαν ονομαστικά στη πρώτη ενότητα. Επιπλέον, αναλύεται η έννοια της Υπολογιστικής Όρασης ως υποσύνολο της Βαθιάς Μάθησης αλλά και οι εφαρμογές της που θα χρησιμοποιηθούν για την επίτευξη του αντικείμενου της εργασίας, την εξαγωγή οντοτήτων από τηλεπισκοπικές απεικονίσεις. Τέλος, αποσαφηνίζονται έννοιες βασικές για τη κατανόηση των μηχανισμών της βαθιάς μάθησης.

Στο τρίτο (3^0) κεφάλαιο παρουσιάζεται αναλυτικά το μοντέλο Segment Anything, η αρχιτεκτονική του, η διαδικασία εκπαίδευσης του, οι μεθοδολογίες που ακολουθεί και η πληθώρα εφαρμογών και δυνατοτήτων που παρέχει στο χρήστη. Το Segment Anything μοντέλο (SAM) αποτελεί ένα προ εκπαιδευμένο θεμελιώδες μοντέλο για κατάτμηση εικόνας με χρήση προτροπών –οδηγιών (prompts).

Στο τέταρτο (4^0) κεφάλαιο παρουσιάζονται διαφορετικές μεθοδολογίες για κατάτμηση δορυφορικών εικόνων και εξαγωγή οντοτήτων μέσω του μοντέλου Segment Anything. Συγκεκριμένα, θα παρουσιαστεί η μεθοδολογία αυτόματης παραγωγής μασκών καθώς και η μεθοδολογία τμηματοποίησης μέσω γεωμετρικών προτροπών –οδηγιών (prompts). Επιπλέον, θα παρουσιαστεί μεθοδολογία που περιλαμβάνει συνεργασία ενός μοντέλου ανίχνευσης οντοτήτων (object detection), το Grounding Dino με το μοντέλο τμηματοποίησης εικόνας μέσω προτροπών, το Segment Anything. Επιπρόσθετα, στα πλαίσια της παρούσας διπλωματικής εργασίας έχει υλοποιηθεί προγραμματιστικά αυτόνομη εφαρμογή με σκοπό την ενσωμάτωση των πλήρων δυνατοτήτων που παρέχει το μοντέλο Segment Anything με σκοπό να αποτελέσει ένα αυτόνομο εργαλείο για κατάτμηση και εξαγωγή οντοτήτων από κάθε δορυφορική απεικόνιση.

Στο πέμπτο (5^0) κεφάλαιο παρουσιάζεται πλήρως η μεθοδολογία της περαιτέρω εκπαίδευσης του SAM (fine-tuning) για κάλυψη συγκεκριμένων αναγκών όπως εξαγωγή συγκεκριμένων οντοτήτων ενδιαφέροντος. Περιγράφεται αναλυτικά η μεθοδολογία εκπαίδευσης και αξιολόγησης του μοντέλου και διατυπώνονται συμπεράσματα για τη διαδικασία.

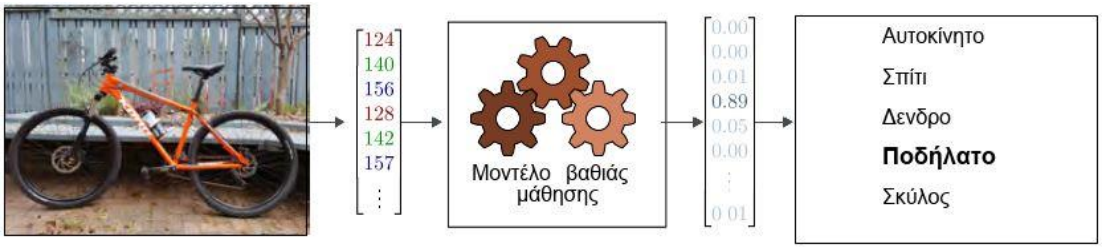
Στο έκτο (6^0) κεφάλαιο με γνώμονα τα ευρήματα των μεθοδολογιών που παρουσιάστηκαν σε προηγούμενες ενότητες, διατυπώνονται συμπεράσματα καθώς και προτείνονται μελλοντικές προεκτάσεις του μοντέλου οι οποίες θα είναι χρήσιμες σε μια πληθώρα τομέων.

ΚΕΦΑΛΑΙΟ 2: ΑΠΟΣΑΦΗΝΙΣΗ ΘΕΩΡΗΤΙΚΩΝ ΕΝΝΟΙΩΝ ΠΕΡΙ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ.

Στο κεφάλαιο αυτό αναλύονται βασικές θεωρητικές έννοιες που εισήχθηκαν σε προηγούμενο κεφάλαιο και αφορούν τρεις κύριες κατηγορίες της βαθιάς μάθησης. Επιπρόσθετα, εισάγεται το αντικείμενο της υπολογιστικής όρασης και οι σύγχρονες εφαρμογές της. Τέλος αποσαφηνίζονται έννοιες οι οποίες είναι απαραίτητες για τη κατανόηση των μηχανισμών βαθιάς μάθησης.

2.1 ΕΠΙΒΛΕΠΟΜΕΝΗ, ΜΗ-ΕΠΙΒΛΕΠΟΜΕΝΗ ΚΑΙ ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ

Η επιβλεπόμενη μάθηση (supervised learning) αναφέρεται σε εκπαίδευση μοντέλων με δεδομένα τα οποία αποτελούν ζευγάρια εισόδου-εξόδου. Αυτά τα δεδομένα (labeled data) έχουν το ρόλο του επιβλέποντα κατά της διάρκειας της εκπαίδευσης. Οι δύο κύριες κατηγορίες της επιβλεπόμενης μάθησης είναι η ταξινόμηση (classification) και η παλινδρόμηση (regression). Η παλινδρόμηση έχει ως αποτέλεσμα την επιστροφή ενός αριθμού ή πολλών. Η ταξινόμηση δύναται να είναι δυαδική (binary classification) ή πολλαπλών κατηγοριών (multi-class classification). Σε τέτοιου είδους προβλήματα το μοντέλο επιχειρεί να ταξινομήσει το όρισμα στη κατηγορία με τη μεγαλύτερη πιθανότητα. (Prince, n.d 2023.)



Σχήμα 2 Πρόβλημα ταξινόμησης πολλαπλών κατηγοριών. Το μοντέλο βαθιάς μάθησης ταξινομεί την εικόνα στη κατηγορία με τη μεγαλύτερη πιθανότητα (ποδήλατο) (Πηγή : Understanding Deep Learning, 2023)

Άλλο παράδειγμα επιβλεπόμενης μάθησης το οποίο θα μας απασχολήσει και στο 5^0 κεφάλαιο είναι ένα μοντέλο δυαδικής ταξινόμησης για σημασιολογική κατάτμηση εικόνας (semantic segmentation). Σε κάθε εικονοστοιχείο της εικόνας που εισάγεται στο μοντέλο ανατίθεται μια δυαδική ετικέτα (binary label) με βάση την οποία τα αντικείμενα ταξινομούνται σε υπόβαθρο ή σε αντικείμενο ενδιαφέροντος όπως φαίνεται στο σχήμα 3.

Σχήμα 3 Παράδειγμα επιβλεπόμενης ταξινόμησης. Σημασιολογική Κατάτμηση Εικόνας (semantic segmentation). Η τρικάναλη εικόνα (RGB) μετατρέπεται σε δυαδική εικόνα (αντικείμενο ενδιαφέροντος και υπόβαθρο εικόνας). (Πηγή : Understanding Deep Learning, 2023)

Η μη- επιβλεπόμενη μάθηση (unsupervised learning) αναφέρεται σε μοντέλα τα οποία εκπαιδεύονται σε δεδομένα (input data) χωρίς επισήμανση (not labels). Έτσι το μοντέλο δεν εξετάζει τις σχέσεις μεταξύ δεδομένων σε ζευγάρια εισόδου και εξόδου όπως στην επιβλεπόμενη αλλά περιγράφει και μελετά τη δομή των δεδομένων. Σημαντικό παράδειγμα μη επιβλεπόμενης μάθησης αποτελούν τα generative μη επιβλεπόμενα μοντέλα. Τα συγκεκριμένα μοντέλα μαθαίνουν να παράγουν νέα δεδομένα τα οποία μοιάζουν με τα δεδομένα εκπαίδευσης. Τα μοντέλα αυτά ακολουθούν τη κατανομή των δεδομένων εισόδου και με βάση αυτή την κατανομή παράγουν νέα δεδομένα. Τέτοια παραδείγματα μοντέλων ασχολούνται και με συμπλήρωση εικόνας (image inpainting) και συμπλήρωση κειμένου (text completion). (Prince, n.d 2023.) Παρακάτω παρατίθενται παραδείγματα generative μοντέλων στο Σχήμα 4 και 5.

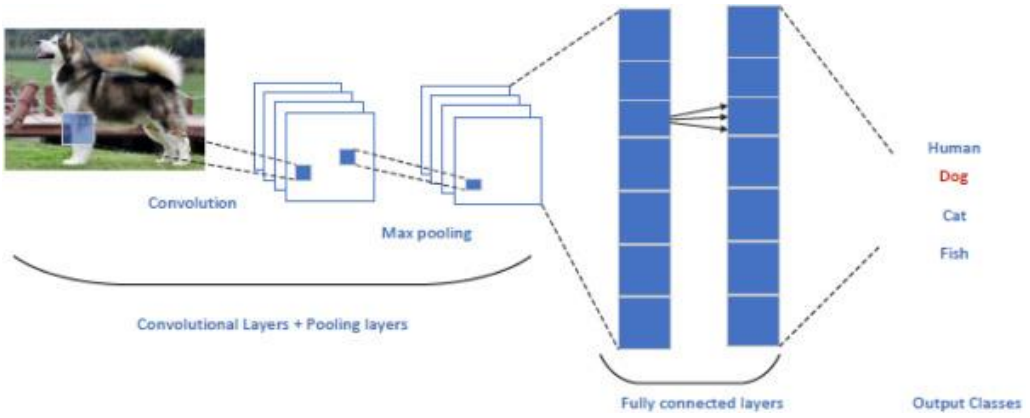
Σχήμα 4 Εικόνες που έχουν παραχθεί από generative μοντέλα. Αριστερά έχουν παραχθεί δύο εικόνες γάτας από μοντέλο που έχει εκπαιδευτεί σε παρόμοιο σετ δεδομένων. Ομοίως και για τα κτίρια δεξιά. (Πηγή : Understanding Deep Learning, 2023)

Σχήμα 5 Παράδειγμα συμπλήρωσης εικόνας (image inpainting) (Πηγή : Understanding Deep Learning, 2023)

Τέλος η ενισχυτική μάθηση (reinforcement learning) αναφέρεται σε μοντέλα τα οποία εκπαιδεύονται μέσω ανάδρασης. Αν το μοντέλο πετυχαίνει τους στόχους του, υπάρχει η ανάλογη επιβράβευση ενώ σε αντίθετη περίπτωση, επιβάλλεται ποινή στο μοντέλο. Έτσι το μοντέλο λαμβάνει αποφάσεις με βάση τις παρατηρήσεις του, τους στόχους του, την ανταμοιβή και τη ποινή που λαμβάνει για κάθε ενέργεια που εκτελεί.

2.2 ΥΠΟΛΟΓΙΣΤΙΚΗ ΟΡΑΣΗ

Η υπολογιστική όραση αποτελεί ένα τομέα της τεχνητής νοημοσύνης και της βαθιάς μάθησης που χρησιμοποιεί νευρωνικά δίκτυα με σκοπό την εξαγωγή πληροφοριών από ψηφιακές εικόνες, βίντεο και άλλα οπτικοακουστικά μέσα. Βασικός πυλώνας για την επιτυχία της υπολογιστικής όρασης αποτελούν τα συνελικτικά νευρωνικά δίκτυα (CNN : Convolutional Neural Networks) τα οποία θα αναλυθούν παρακάτω στο παρόν κεφάλαιο.



Σχήμα 6 Αρχιτεκτονική ενός Συνελικτικού Νευρωνικού Δικτύου για ταξινόμηση εικόνας (πηγή : Machine Learning with Applications, 2021)

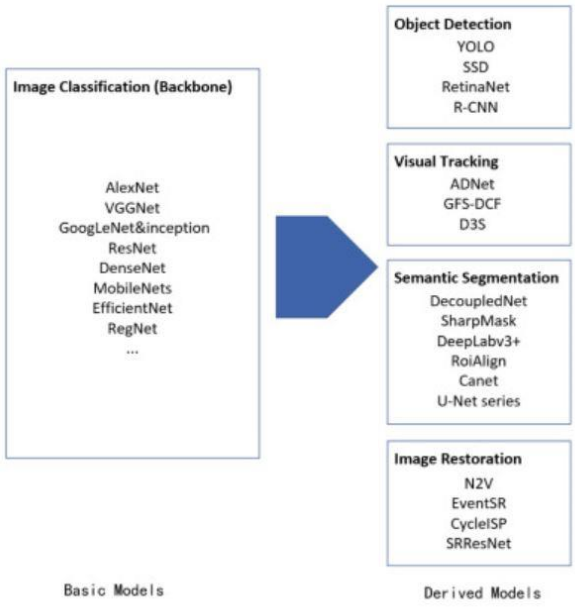
2.2.1 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ ΣΕ ΥΠΟΛΟΓΙΣΤΙΚΗ ΟΡΑΣΗ

Στα πρώιμα στάδια της εξέλιξης της υπολογιστικής όρασης (CV) , η προσέγγιση της βαθιάς μάθησης αντιμετώπισε δυσκολίες λόγω περιορισμών που απέρρεαν από την περιορισμένη μνήμη υπολογιστών, από τους κεντρικούς επεξεργαστές (CPU) και από τις κάρτες γραφικών (GPU) . Για αυτό το λόγο, πολλοί ερευνητές εξέταζαν τη προσέγγιση της μηχανικής μάθησης για την υπολογιστική όραση προτείνοντας μεθοδολογίες όπως ο K-means, ο ταξινομητής Naive-Bayes, την μέθοδο του πλησιέστερου γείτονα (K –Nearest Neighbor) καθώς και το Random Forest. Αξίζει να επισημανθεί η μεθοδολογία που ακολουθήθηκε από τους Viola και Jones το 2001 με σκοπό την ανίχνευση προσώπου (face detection) μέσω εκπαίδευσης ενός

ταξινομητή με βάση τον Adaboost αλγόριθμο. Ο ταξινομητής αυτός είναι γνωστός ως Viola-Jones ανιχνευτής (Viola and Jones, 2001).

Η ανάπτυξη της βαθιάς μάθησης τις τελευταίες δεκαετίες είναι ραγδαία και δύναται να ταξινομηθεί σε 10 κύριες κατηγορίες με βάση την αρχιτεκτονική των αντίστοιχων μοντέλων: συνελικτικά νευρωνικά δίκτυα (CNN), δίκτυα μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory Networks–LSTMs), αναδραστικά νευρωνικά δίκτυα (Recurrent Neural Networks – RNNs), τα Generative Adversalial Networks – GANs, τα Radial Basis Function Networks - RBFNs), οι Multilayer Perceptrons - MLPs), οι Self-Organizing Maps - SOMs), τα Δίκτυα Βαθιάς Πίστης (Deep Belief Networks - DBNs), οι Περιορισμένες Μηχανές Boltzmann (Restricted Boltzmann Machines - RBMs), και οι Αυτοκωδικοποιητές (Autoencoders). (Chai et al., 2021). Το 2016 από τον Guo και την ομάδα του διενεργήθηκε έρευνα που αφορούσε τη σύγκριση της απόδοσης των παραπάνω μοντέλων σε μια σειρά από αντικείμενα υπολογιστικής όρασης όπως η ταξινόμηση εικόνας, η ανίχνευση αντικειμένων, η ανάκτηση εικόνων και σημασιολογική κατάτμηση τους. Στο τέλος, αποφάνθηκαν ότι τα CNN αποτελούν τη βέλτιστη επιλογή για διαχείριση CV εργασιών (Guo et al., 2016).

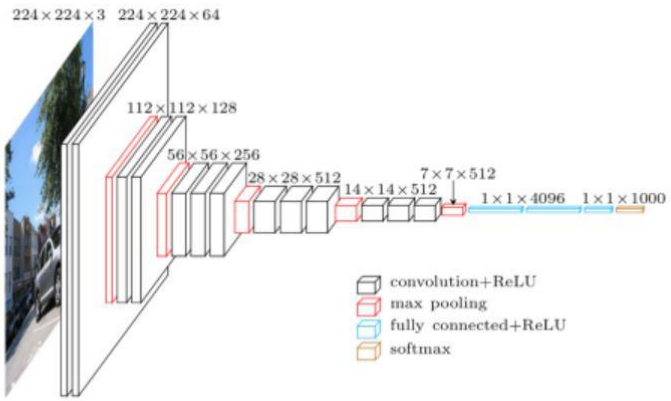
Βασικά CNN μοντέλα που έχουν χρήση σε πολλά πεδία εφαρμογών της Υπολογιστικής Όρασης (CV) και έχουν αποτελέσει προπομπούς μεταγενέστερων μοντέλων είναι το AlexNet, το VGGNet, το GoogleNet & Inception, το ResNet, το DenseNet, το MobileNets, το EfficientNet και το RegNet. Παρακάτω ακολουθεί μια συνοπτική περιγραφή κάποιων εκ των παραπάνω μοντέλων που σηματοδότησαν την ραγδαία εξέλιξη των αρχιτεκτονικών βαθιάς μάθησης.



Σχήμα 7 Βασικά CNN και παραγόμενα από αυτά μοντέλα (πηγή : Machine Learning with Applications, 2021)

1. **AlexNet**: Προτάθηκε από τους Krizhevsky και τους συνεργάτες του το 2012 στο άρθρο με όνομα «*ImageNet Classification with Deep Convolutional Neural Networks*». Αποτελείται από 5 συνελικτικά επίπεδα (Layers), ακολουθούμενα από τρία πλήρως συνδεδεμένα επίπεδα (Fully Connected Layers). Μετά από κάθε συνελικτικό επίπεδο υπάρχει μια activation function ReLU η οποία προσδίδει μη γραμμικότητα στα αποτελέσματα και ένα pooling layer (Max-pooling) το οποίο μειώνει τη διάσταση των δεδομένων που μαθαίνει το μοντέλο. (Krizhevsky, Sutskever and Hinton, 2012)

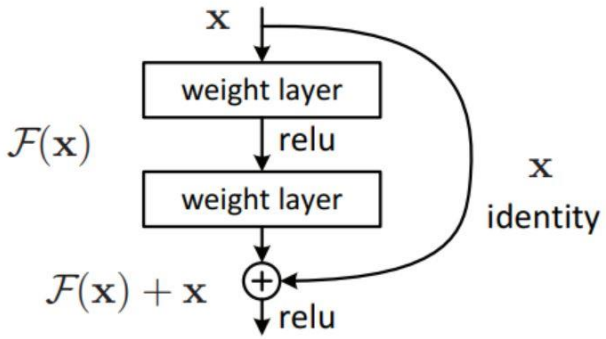
2. **VGGNet:** Προτάθηκε από τους K. Simonyan και A. Zisserman από το Πανεπιστήμιο της Οξφόρδης το 2015 στη δημοσίευση τους με όνομα «*Very Deep Convolutional Networks for Large Scale Image Recognition*». Το VGG (Visual Geometry Group) είναι ένα Συνελικτικό Νευρωνικό Δίκτυο το οποίο έχει μια τυπική βαθιά αρχιτεκτονική. Ο όρος «βαθιά» αναφέρεται στο πλήθος των συνελικτικών επιπέδων. Ανάλογα τον αριθμό των layers (16 και 19) υπάρχουν αντίστοιχα τα VGG-16 και VGG-19 μοντέλα. Η αρχιτεκτονική του VGG αποτελεί τη βάση για πρωτοποριακά μοντέλα αναγνώρισης αντικειμένων. Η επιτυχία του VGG δεν είναι μόνο λόγω του βάθους του μοντέλου (πλήθος επιπέδων) αλλά λόγω και του πλήθους των παραμέτρων που απαρτίζουν τα επίπεδα. Ενδεικτικά το μοντέλο VGG έχει 500M παραμέτρους ενώ το AlexNet έχει 200M παραμέτρους. (Simonyan and Zisserman, 2015)



Σχήμα 8 Αρχιτεκτονική VGG νευρωνικού δικτύου (πηγή: *Very Deep Convolutional Networks for Large Scale Image Recognition*,2015)

3. **ResNet:** Το ResNet (Residual Neural Network) είναι ένα μοντέλο βαθιάς μάθησης που χρησιμοποιείται για εφαρμογές υπολογιστικής όρασης. Η αρχιτεκτονική του παρουσιάστηκε από τον Kaiming He, Xiangyu Zhang,

Shaoqing Ren και Jian Sun το 2015 στο άρθρο με ονομασία «*Deep residual Learning for Image Recognition*» . Μέχρι την εμφάνιση της συγκεκριμένης αρχιτεκτονικής, η συνηθισμένη μεθοδολογία ήταν η αύξηση του βάθους των νευρωνικών δικτύων για μείωση του σφάλματος εκπαίδευσης (training error). Όμως μετά από ένα συγκεκριμένο αριθμό επιπέδων (layers), παρατηρήθηκε το φαινόμενο «Vanishing/Exploding Gradient » δηλαδή η κλίση είτε μηδενιζόταν είτε αυξανόταν σε υπερβολικό βαθμό. Έτσι, σε διάφορα πειράματα είχε παρατηρηθεί ότι συνελικτικά νευρωνικά δίκτυα που αποτελούταν από 56 επίπεδα είχαν μεγαλύτερο σφάλμα εκπαίδευσης σε σχέση με αντίστοιχα μοντέλα με 20 επίπεδα. Έτσι για να λυθεί το συγκεκριμένο πρόβλημα, στο μοντέλο ResNet εισήχθη η τεχνική «skip connections», δηλαδή οι νευρώνες από ένα επίπεδο συνδέονταν με περαιτέρω επίπεδα, παρακάμπτοντας κάποια ενδιάμεσα επίπεδα. Αυτό δημιουργεί ένα «residual block». Έτσι το ResNet αποτελείται από συσσώρευση πολλών τέτοιων μπλοκ. (He et al., 2015)



Σχήμα 9 Παράδειγμα skip connection (πηγή: *Deep residual Learning for Image Recognition*,2015)

2.2.2 ΕΦΑΡΜΟΓΕΣ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΟΡΑΣΗΣ

Τρεις ευρέως διαδεδομένες εφαρμογές της υπολογιστικής όρασης είναι η ταξινόμηση της εικόνας , η ανίχνευση αντικειμένων και η σημασιολογική κατάτμηση εικόνας. Οι συγκεκριμένες εφαρμογές θα παρουσιαστούν παρακάτω.

- **Ταξινόμηση Εικόνας (Image Classification)**

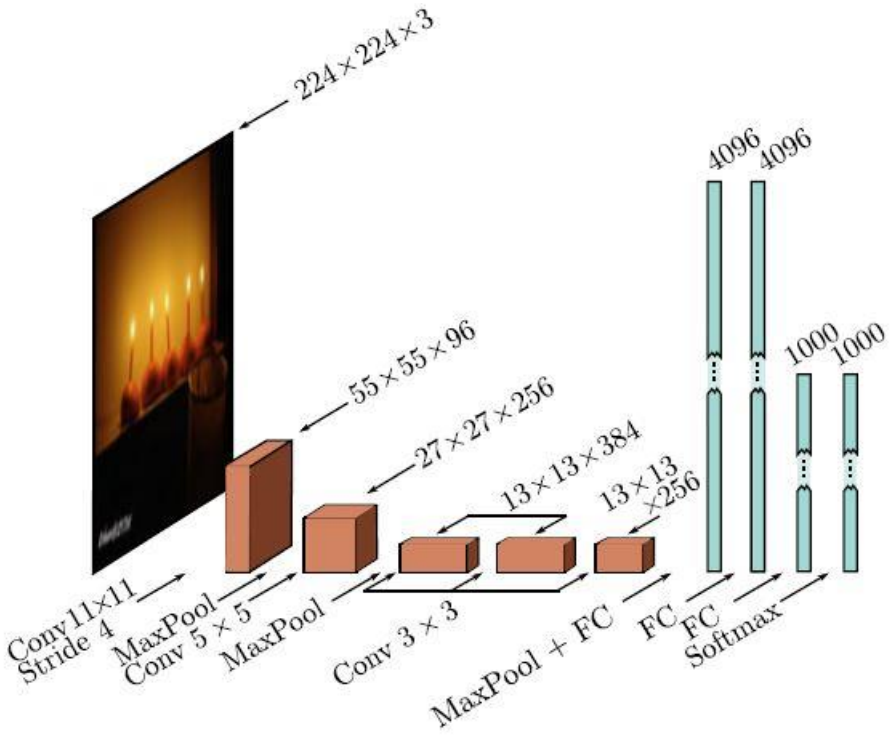
Τα συνελικτικά νευρωνικά δίκτυα χρησιμοποιούνται ευρέως στη ταξινόμηση εικόνας όπου ο σκοπός είναι η αντιστοίχηση της εικόνας σε μια από τις προκαθορισμένες κλάσεις. Ένα από τα πιο γνωστά σετ δεδομένων που με βάση αυτό έχουν εκπαιδευτεί νευρωνικά μοντέλα στην ταξινόμηση εικόνας είναι το ImageNet. Το συγκεκριμένο σετ δεδομένων αποτελείται από 1.281.167 εικόνες προς

εκπαίδευση (training images) , 50.000 εικόνες επικύρωσης (validation images) και 100.000 εικόνες προς έλεγχο (test images). Κάθε εικόνα έχει ταξινομηθεί σε μια από τις 1000 πιθανές κατηγορίες. (Deng et al., 2009).

Τα εκάστοτε μοντέλα που εκπαιδεύτηκαν πάνω στο ImageNet μετασχημάτιζαν την εικόνα που δεχόντουσαν ως όρισμα σε μέγεθος 224*224 με τρία κανάλια (RGB) και εξήγαγαν μια κατανομή πιθανοτήτων για 1000 κλάσεις. Το αντικείμενο εργασίας ήταν ιδιαίτερα απαιτητικό ειδικά για μοντέλα προ βαθιάς μάθησης. Συγκεκριμένα, τα μοντέλα αιχμής εκείνης της εποχής ταξινομούσαν τις εικόνες ελέγχου με ένα σφάλμα της τάξης του 25% για τις πέντε πιο πιθανές κλάσεις. . Με την εμφάνιση της βαθιάς μάθησης, τα μοντέλα υπερκέρασαν την ανθρώπινη απόδοση στην ταξινόμηση εικόνας.

Το 2012, το AlexNet που έχει παρουσιαστεί σε προηγούμενη ενότητα, ήταν το πρώτο συνελικτικό νευρωνικό δίκτυο το οποίο είχε αξιόπιστα αποτελέσματα στο συγκεκριμένο αντικείμενο. Συγκεκριμένα η μεθοδολογία που ακολουθήθηκε ήταν η παρακάτω. Το μοντέλο αρχικά ξεκινούσε με υποδειγματοληψία (downsampling) της εικόνας με ένα παράθυρο (kernel) διαστάσεων 11*11 και βήμα (stride) 4 για τη δημιουργία 96 καναλιών διαστάσεων 55*55. Έτσι η αρχική εικόνα από διαστάσεις 224*224*3 τροποποιούταν σε διαστάσεις 55*55*96 (height *width *channels). Το πρώτο βήμα εκτελούταν από ένα συνελικτικό επίπεδο 11*11 (convolutional layer). Στη συνέχεια ακολουθεί ένα max-pooling επίπεδο για τη μείωση των στοιχείων εκμάθησης (image features) και έπειτα ένα συνελικτικό επίπεδο (Conv layer) διαστάσεων 5*5 που άλλαζε τις διαστάσεις της αρχικής εικόνας σε 27*27*256. Τέλος ακολουθούν άλλα τρία συνελικτικά επίπεδα διαστάσεων 3*3 που μετατρέπουν την διάσταση της εικόνας σε 13*13*256. (height,width,channels). Επόμενο βήμα είναι η μετατροπή σε διάνυσμα διαστάσεων 43.264 *1 το οποίο τροφοδοτείται σε τρία πλήρως συνδεδεμένα επίπεδα (Fully Connected Layers) , το μήκος του καθενός (4096, 4096 και 1000 νευρώνες αντίστοιχα). Το τελικό επίπεδο εισάγεται σε μια συνάρτηση Softmax (*output activation function for multi-class classification*) και παράγεται μια κατανομή πιθανοτήτων για πάνω από 1000 κλάσεις. Το μοντέλο χρησιμοποιεί κατά την εκπαίδευση του μετασχηματισμούς των δεδομένων εκπαίδευσης (data augmentation) καθώς και dropout επίπεδα ανάμεσα στα πλήρως συνδεδεμένα (fully connected layers). Τα dropout επίπεδα αφαιρούν συνδέσεις μεταξύ των επιπέδων με σκοπό την αποτροπή της υπερπροσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης (over-fitting) . Το AlexNet είχε σφάλμα 16.4% για τις πέντε πιο πιθανές κλάσεις. Αυτό το αποτέλεσμα σηματοδότησε την απαρχή της εποχής της βαθιάς μάθησης και φανέρωσε τις πραγματικές τις δυνατότητες. (Krizhevsky, Sutskever and Hinton, 2012).

Μετά το AlexNet και το VGG μοντέλο εκπαιδεύτηκε ομοίως πάνω στο ImageNet σημειώνοντας καλύτερη απόδοση από το AlexNet με σφάλμα 6.8% για τις πέντε πιο πιθανές κλάσεις.



Σχήμα 10 Αρχιτεκτονική AlexNet μοντέλου (Πηγή: Understanding Deep Learning, 2023)

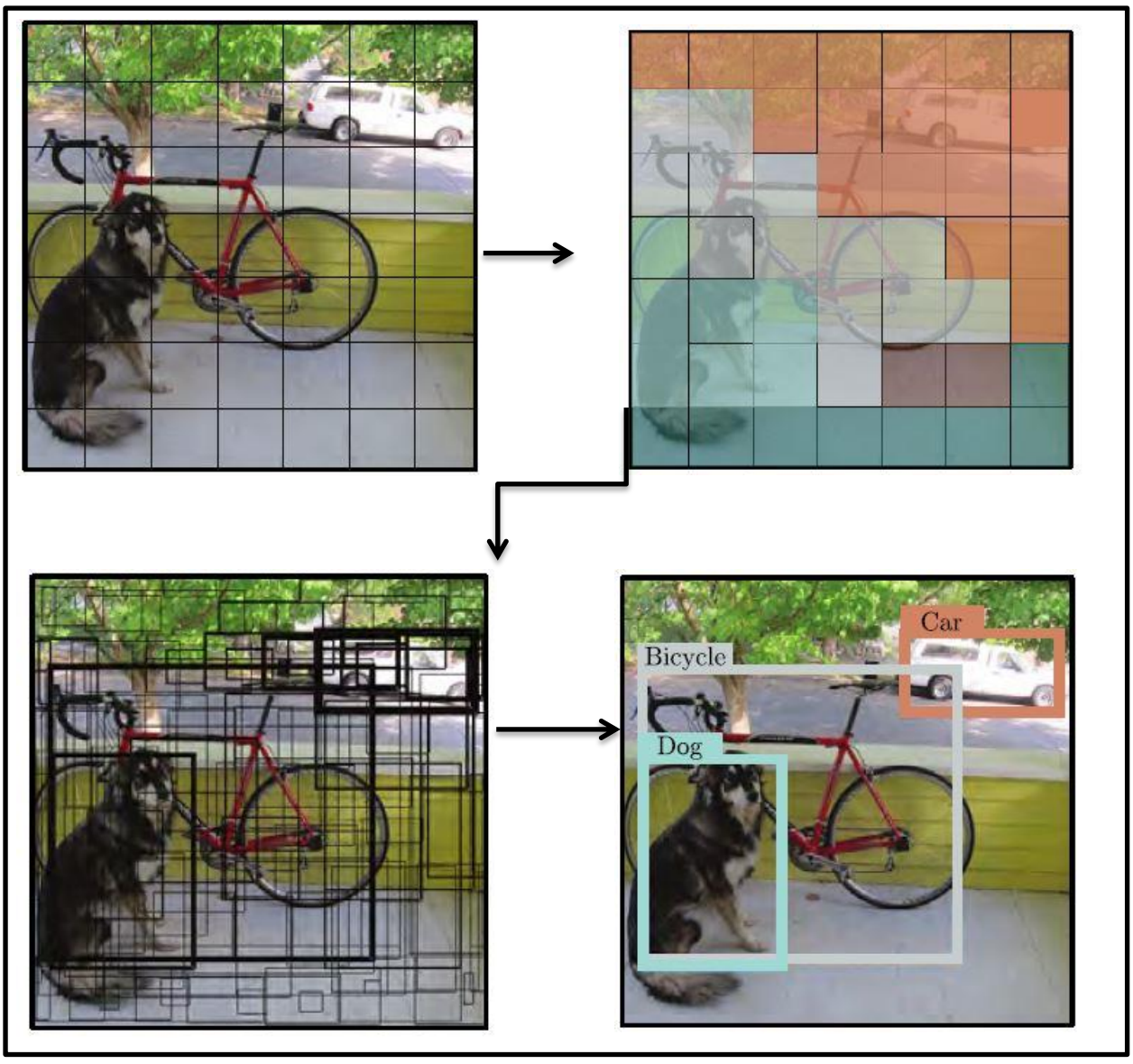
- **Ανίχνευση Αντικειμένων (Object Detection)**

Σκοπός της ανίχνευσης των αντικειμένων είναι η ανίχνευση και εντοπισμός πολλαπλών αντικειμένων μέσα σε μια εικόνα. Μια μέθοδος που βασίστηκε στα συνελικτικά νευρωνικά μοντέλα είναι το YOLO (You Only Look Once). Το YOLO δέχεται ως όρισμα μια εικόνα με τρία κανάλια (RGB) διαστάσεων 448*448. Η εικόνα περνάει από 24 συνελικτικά επίπεδα (Conv layers) τα οποία σταδιακά μειώνουν το μέγεθος της αναπαράστασης της εικόνας μέσω max pooling επιπέδων, ενώ παράλληλα αυξάνεται ο αριθμός των καναλιών. Το τελικό συνελικτικό επίπεδο έχει διαστάσεις 7*7 και 1024 κανάλια. Έπειτα αναδιαμορφώνεται σε ένα διάνυσμα το οποίο εισάγεται σε ένα πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer) το οποίο το αποδίδει σε 4096 τιμές. Ένα επιπλέον πλήρως συνδεδεμένο επίπεδο χαρτογραφεί αυτή την αναπαράσταση στην έξοδο.

Οι τιμές εξόδου δείχνουν ποια κλάση υπάρχει σε καθεμία από τις θέσεις ενός πλέγματος 7*7 στην εικόνα εισόδου. Σε κάθε θέση, οι τιμές εξόδου παρέχουν και ένα σταθερό αριθμό κουτιών (bounding boxes). Κάθε κουτί ορίζεται από πέντε παραμέτρους : συντεταγμένες x και y για το κέντρο του, ύψος και μήκος κουτιού και η εμπιστοσύνη της πρόβλεψης (confidence score). Η εμπιστοσύνη εκτιμά την επικάλυψη μεταξύ των κουτιών που έχουν προβλεφθεί και των πραγματικών

κουτιών (predicted and ground truth bounding boxes). Αξίζει να επισημανθεί το γεγονός ότι το μοντέλο έχει αρχικά εκπαιδευτεί στο ImageNet για ταξινόμηση και στη συνέχεια εκπαιδεύτηκε περαιτέρω (fine-tuning) για την ανίχνευση αντικειμένων. Αυτή η διαδικασία ονομάζεται μεταφερόμενη μάθηση (transfer-learning).

Στο τέλος, τα κουτιά με τη χαμηλότερη τιμή εμπιστοσύνης αφαιρούνται ενώ από τα κουτιά που αναφέρονται στο ίδιο αντικείμενο επιλέγεται το πιο αξιόπιστο. (Redmon et al., 2015)



Σχήμα 11 Μοντέλο YOLO (Πηγή: Understanding Deep Learning, 2023)

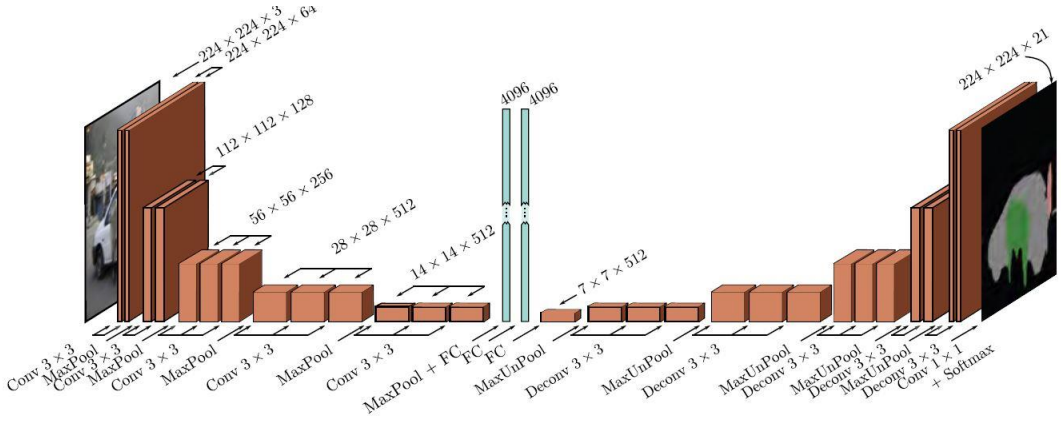
- **Σημασιολογική κατάτμηση (Semantic Segmentation)**

Ο στόχος της σημασιολογικής κατάτμησης είναι η απόδοση ή μη ετικέτας σε κάθε εικονοστοιχείο της εικόνας, ανάλογα αν αυτό αντιστοιχεί σε κάποιο αντικείμενο ή

όχι του σετ δεδομένων με βάση το οποίο έχει εκπαιδευτεί το μοντέλο. Η σημασιολογική κατάτμηση αναφέρεται στην σύνδεση κάθε εικονοστοιχείου μιας ψηφιακής εικόνας με μια κλάση (κατηγορία). Ουσιαστικά μια εικόνα διαμερίζεται σε κλάσεις αντικειμένων με τη καθεμιά να διακρίνεται χαρακτηριστικά από την άλλη. Χαρακτηριστικό παράδειγμα μοντέλου για σημασιολογική κατάτμηση είναι αυτό του *Hyeonwoo Noh* που δημοσιεύτηκε το 2015.

Αναλύοντας το συγκεκριμένο μοντέλο, παρατηρείται ότι χωρίζεται σε δύο βασικά μέρη. Η εικόνα εισόδου έχει διαστάσεις 224*224 και τρία κανάλια (RGB) ενώ το τελικό αποτέλεσμα έχει διαστάσεις 224*224*21 το οποίο περιέχει τη πιθανότητα από 21 πιθανές κλάσεις σε κάθε θέση της εικόνας. Το πρώτο μέρος του συνολικού μοντέλου είναι μια μικρή έκδοση του VGG που περιέχει 13 αντί για 15 συνελικτικά επίπεδα που μειώνουν την αναπαράσταση της εικόνας σε 14*14 αλλά με 512 επίπεδα. Μετά υπάρχει μια διεργασία με max pooling και δύο πλήρως συνδεδεμένα επίπεδα τα οποία μετατρέπουν την αναπαράσταση σε δυο μονοδιάστατες αναπαραστάσεις μήκους 4096. Τα δύο συγκεκριμένα επίπεδα δεν περιέχουν χωρική πληροφορία αλλά πληροφορία από όλη την εικόνα. Το δεύτερο μέρος του μοντέλου διαφοροποιείται από το VGG. Ένα πλήρως συνδεδεμένο επίπεδο μετασχηματίζει την αναπαράσταση σε 7*7 και 512 κανάλια. Αυτό ακολουθείται από μια σειρά αποσυνελικτικών επιπέδων και max unpooling επίπεδα. Το τελικό αποτέλεσμα είναι μια αναπαράσταση διαστάσεων 224*224*21. Το πρώτο μέρος του μοντέλου που αφορά μείωση της αναπαράστασης της εικόνας (downsampling) αναφέρεται ως encoder ενώ το δεύτερο μέρος του μοντέλου αφορά την αύξηση της αναπαράστασης της εικόνας (upsampling) και αναφέρεται ως decoder.Έτσι τα μοντέλα αυτά ονομάζονται συνήθως μοντέλα encoder-decoder.

Τέλος η τελική κατάτμηση της εικόνας παράγεται επιλέγοντας τη κλάση που εκπροσωπείται περισσότερο και προσδιορίζεται η περιοχή που αντιστοιχεί. Στη συνέχεια προστίθεται η αμέσως επόμενη κλάση που κυριαρχεί στα εναπομείναντα κενά εικονοστοιχεία. Η διαδικασία συνεχίζεται μέχρι να μην υπάρχουν επαρκή στοιχεία για να προστεθούν επιπλέον κλάσεις. (Noh, Hong and Han, 2015)

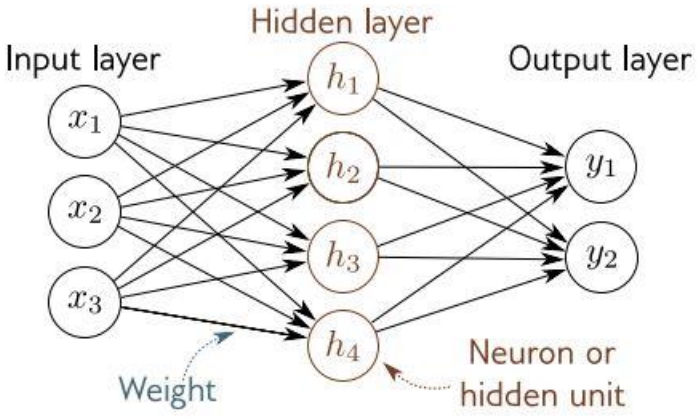


Σχήμα 12 Αρχιτεκτονική Μοντέλου Semantic Segmentation (Noh et al, 2015)

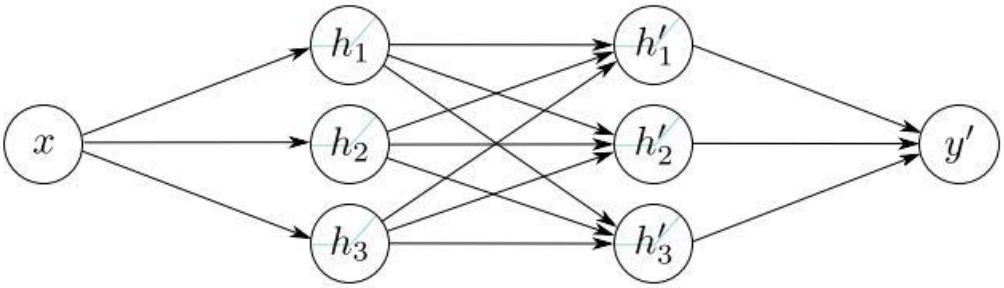
2.3 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΠΕΡΙ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ

2.3.1 ΡΗΧΑ ΚΑΙ ΒΑΘΙΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Η βασική διαφορά μεταξύ ρηχών και βαθιών νευρωνικών δικτύων (*shallow and deep neural networks*) είναι το βάθος (*depth*) του εκάστοτε δικτύου, δηλαδή ο αριθμός των ενδιάμεσων επιπέδων (*hidden layers*) που το απαρτίζουν. Στα ρηχά νευρωνικά δίκτυα, υπάρχει ένα ενδιάμεσο επίπεδο ενώ αντιθέτως στα βαθιά υπάρχουν αντίστοιχα περισσότερα από ένα. Και στις δύο περιπτώσεις πριν τα ενδιάμεσα επίπεδα υπάρχει το εισαγόμενο επίπεδο (*input layer*) ενώ στο τέλος το τελικό επίπεδο (*output layer*). Εκτός από το βάθος του εκάστοτε νευρωνικού δικτύου υπάρχει και το μήκος (*width*) του εκάστοτε επιπέδου του μοντέλου. Το μήκος αναφέρεται στον αριθμό των νευρώνων (*hidden units*) που απαρτίζουν το εκάστοτε επίπεδο του δικτύου. Ο συνολικός αριθμός των νευρώνων σε ένα δίκτυο ορίζεται ως η χωρητικότητα του δικτύου (*capacity*). Ο αριθμός των επιπέδων καθώς και των νευρώνων αποτελούν παραδείγματα υπερπαραμέτρων (*hyperparameters*), δηλαδή παραμέτρων που καθορίζονται από τον χρήστη ανάλογα το σκοπό που εξυπηρετεί το εκάστοτε μοντέλο. Επίσης αν κάθε στοιχείο σε ένα επίπεδο συνδέεται με όλα τα στοιχεία του επόμενου επιπέδου τότε το δίκτυο ονομάζεται πλήρως συνδεδεμένο (*fully-connected layer*). Αυτές οι συνδέσεις μεταξύ των επιπέδων αναφέρονται και ως βάρη (*weights*) και αποτελούν το κύριο στοιχείο το οποίο μαθαίνει ένα μοντέλο κατά την εκπαίδευση του. Εκτός από τα βάρη, σε κάθε επίπεδο προστίθεται και ένας επιπρόσθετος όρος που αναφέρεται ως απόκλιση (bias). Το σύνολο των βαρών και των αποκλίσεων (weights and biases) είναι οι παράμετροι που ένα μοντέλο μαθαίνει κατά την εκπαίδευση του. Ακολουθούν απεικονίσεις ενός ρηχού και ενός βαθύ νευρωνικού δικτύου.

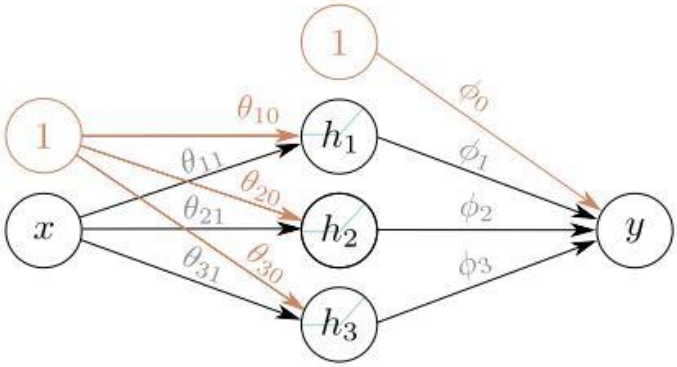


Σχήμα 13 Δομή ενός ρηχού νευρωνικού δικτύου (Πηγή: Understanding Deep Learning, 2023)



Σχήμα 14 Δομή ενός βαθύ νευρωνικού δικτύου (δύο ενδιάμεσα επίπεδα) (Πηγή: Understanding Deep Learning, 2023)

Προσεγγίζοντας τη λειτουργία ενός ρηχού νευρωνικού δικτύου από μαθηματικής απόψεως, ορίζεται το παρακάτω ρηχό νευρωνικό δίκτυο στο Σχήμα 15.



Σχήμα 15 Ρηχό νευρωνικό δίκτυο προς μελέτη (Πηγή: Understanding Deep Learning, 2023)

Στο Σχήμα 15 αριστερά είναι το επίπεδο x (*input layer*) ενώ δεξιά είναι το επίπεδο y (*output layer*). Ενδιάμεσα βρίσκεται ένα επίπεδο με 3 νευρώνες (*h1, h2, h3*). Ο υπολογισμός γίνεται από τα αριστερά προς τα δεξιά με τελικό αποτέλεσμα το y. Το πρώτο επίπεδο χρησιμοποιείται για να υπολογιστούν οι νευρώνες (*h1,h2,h3*), οι οποίοι στη συνέχεια συνδυάζονται για να υπολογιστεί το τελικό αποτέλεσμα. Κάθε βέλος στο σχήμα αποτελεί μια παράμετρο του μοντέλου (τα βάρη του μοντέλου είναι θ11, θ21, θ31,ϕ1,ϕ2,ϕ3 ενώ οι αποκλίσεις είναι θ10, θ20, θ30 και ϕ0. Η μαθηματική σχέση που περιγράφει το άνω μοντέλο είναι η παρακάτω:

$$y = \varphi 0 + \varphi 1 h1 + \varphi 2 h2 + \varphi 3 h3 \qquad (1)$$

Η σχέση 1 αποτελεί μια γραμμική συνάρτηση και μπορεί να αναλυθεί περαιτέρω όπως παρακάτω:

$$h1 = a[\theta 10 + \theta 11 x]$$

$$h2 = a[\theta 20 + \theta 21 x] \qquad (2)$$

$$h3 = a[\theta 30 + \theta 31 x]$$

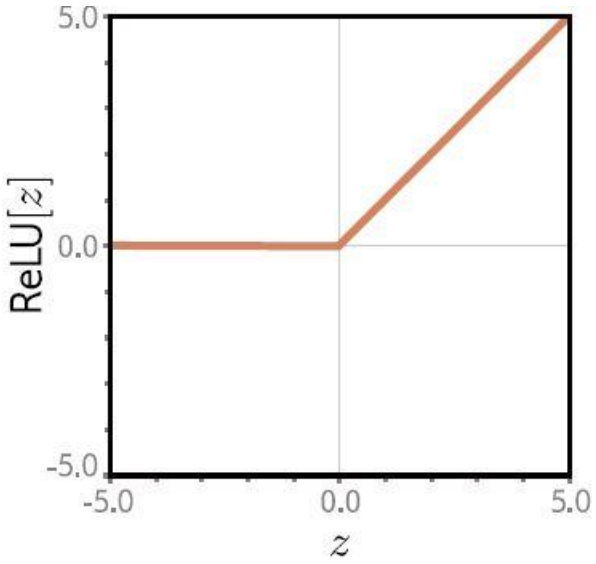
και στη πλήρη της μορφή:

$$\overset{1,2}{\Rightarrow} y = \varphi 0 + \varphi 1\alpha[\theta 10 + \theta 11x] + \varphi 2\alpha[\theta 20 + \theta 21x] + \varphi 3\alpha[\theta 30 + \theta 31x] \quad (3)$$

Το σύμβολο α[ ] συμβολίζει τη συνάρτηση ενεργοποίησης (*activation function*) που εισάγει την έννοια της μη-γραμμικότητας στο μοντέλο. Οι συναρτήσεις ενεργοποίησης έχουν καθοριστικό ρόλο στη λειτουργία του μοντέλου και θα αναλυθούν στην επόμενη υπο-ενότητα. Στη συγκεκριμένη περίπτωση ως συνάρτηση ενεργοποίησης χρησιμοποιείται η συνήθης ReLU ή αλλιώς *rectified linear unit.* (Abien Fred Agarap, 2018).

$$\alpha[z] = ReLU[z] = \begin{cases} 0, & z < 0 \\ z, & z \geq 0 \end{cases} \quad (4)$$

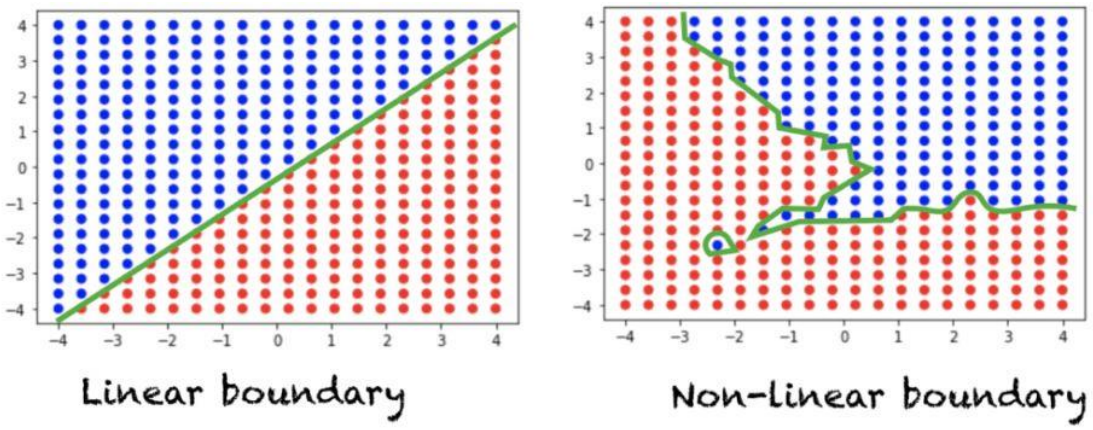
Ουσιαστικά η συγκεκριμένη συνάρτηση επιστρέφει το όρισμα όταν είναι θετικό και μηδέν όταν είναι αρνητικό.



Σχήμα 16 Συνάρτηση ενεργοποίησης ReLU (Πηγή: Understanding Deep Learning, 2023)

2.3.2 ΕΝΝΟΙΑ ΤΗΣ ΜΗ ΓΡΑΜΜΙΚΟΤΗΤΑΣ ΣΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Οι μη γραμμικές συναρτήσεις ενεργοποίησης αποτελούν τα μαθηματικά εργαλεία με τα οποία τα νευρωνικά δίκτυα εκπαιδεύονται από σύνθετα δεδομένα. Εισάγουν τη μη γραμμικότητα στο μοντέλο το οποίο το βοηθάει να μαθαίνει από σύνθετα δεδομένα και να κάνει ακριβείς προβλέψεις. Χωρίς τη μη γραμμικότητα , ένα νευρωνικό δίκτυο θα ήταν ένα απλό μοντέλο γραμμικής παλινδρόμησης (*linear regression model*), μη ικανό να διαχειριστεί σύνθετα δεδομένα.



Σχήμα 17 Η μη γραμμικότητας στη διαχείριση σύνθετων δεδομένων (πηγή: https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function)

Οι συναρτήσεις ενεργοποίησης εφαρμόζονται στα δεδομένα που εισάγονται σε κάθε νευρώνα ενός μοντέλου, μετασχηματίζοντας τα με μη γραμμικό τρόπο. Τα μετασχηματισμένα δεδομένα στη συνέχεια μεταβαίνουν στο επόμενο επίπεδο του μοντέλου. Ανάλογα το αντικείμενο που εξυπηρετεί το εκάστοτε μοντέλο εισάγεται και η αντίστοιχη συνάρτηση ενεργοποίησης. Παρακάτω θα παρατεθούν και θα αναλυθούν κάποιες από τις πιο γνωστές συναρτήσεις ενεργοποίησης.

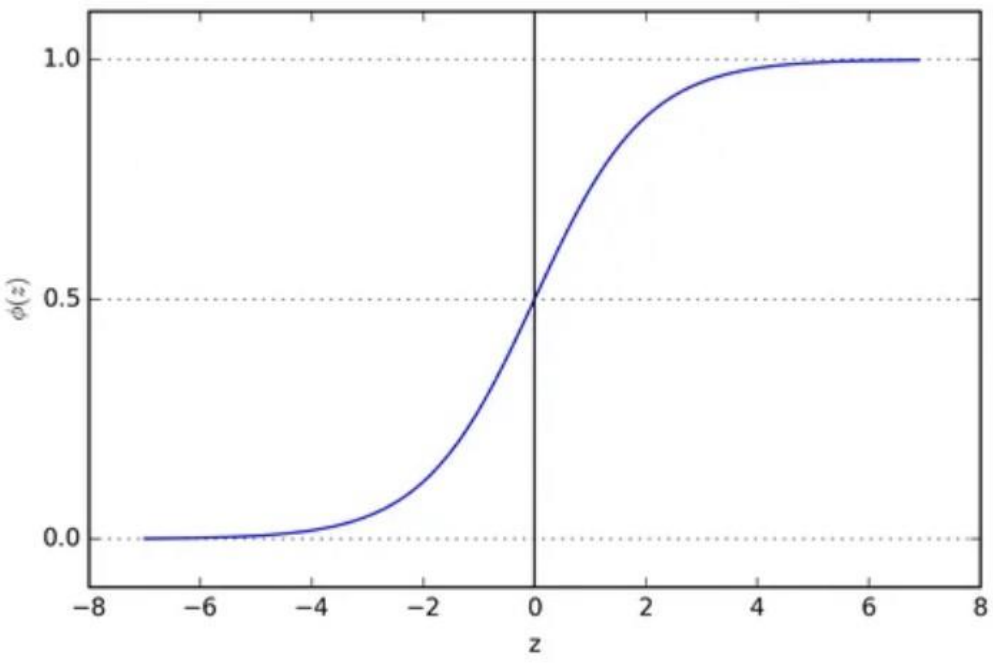
- **Συνάρτηση Sigmoid (Sigmoid function)**

Η σιγμοειδής συνάρτηση είναι μια από τις πιο γνωστές συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στα νευρωνικά δίκτυα. Η καμπύλη της έχει το σχήμα S και έχει ως πεδίο τιμών το διάστημα (0-1). (Narayan, 1997). Συνήθως χρησιμοποιείται για προβλήματα δυαδικής ταξινόμησης (*binary classification*).Η εξίσωση που τη περιγράφει είναι η παρακάτω:

$$\varphi\,(z) = \frac{1}{1 + e^{-z}}$$

Μειονεκτήματα της συγκεκριμένης συνάρτησης είναι ότι το αποτέλεσμα της δεν έχει ως κέντρο το μηδέν (*zero-centered*), γεγονός που επηρεάζει τη σύγκλιση του μοντέλου κατά την εκπαίδευση. Επιπλέον απαιτεί υψηλούς υπολογιστικούς πόρους λόγω του εκθέτη e^{-z} ειδικά σε βαθιά νευρωνικά δίκτυα. Τέλος παρατηρείται το πρόβλημα του κορεσμού (saturation problem) κατά την εκπαίδευση του μοντέλου. Τα άκρα της σιγμοειδής καμπύλης είναι σχεδόν ευθείες με αποτέλεσμα όταν τα δεδομένα που εισάγονται βρίσκονται στα άκρα του πεδίου ορισμού της (μεγάλες θετικές και μεγάλες αρνητικές τιμές) , η κλίση της τείνει στο 0. Αυτό έχει ως αποτέλεσμα την επιβράδυνση της εκπαίδευσης του μοντέλου. Επιπλέον, η παράγωγος της σιγμοειδής συνάρτησης κυμαίνεται στο διάστημα [0, 0.25]. Αν το νευρωνικό δίκτυο αποτελείται από πολλά επίπεδα, η μερική παράγωγος θα ισούται με το γινόμενο πολλών αριθμών μικρότερων του 0.25. Αυτό έχει ως αποτέλεσμα η

κλίση να τείνει στο 0 και να προκαλείται το πρόβλημα της «εξαφάνισης» της κλίσης (*vanishing gradient problem*). Όπως θα δούμε παρακάτω η κλίση (gradient) παίζει καθοριστικό ρόλο στο τρόπο που εκπαιδεύεται ένα μοντέλο, δηλαδή ουσιαστικά μαθαίνει τις παραμέτρους του μοντέλου (βάρη και αποκλίσεις).



Σχήμα 18 Σιγμοειδής Συνάρτηση (Πηγή: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6)

- **Συνάρτηση ReLU (Rectified Linear Unit Function)**

Η συνάρτηση ReLU είναι η πιο διάσημη συνάρτηση ενεργοποίησης στη βαθιά μάθηση. Ουσιαστικά πρόκειται για μια απλή συνάρτηση που αντιστοιχεί οποιαδήποτε αρνητική τιμή στο μηδέν ενώ οι θετικές τιμές παραμένουν αναλλοίωτες. Η απλότητα της επιτρέπει τον εύκολο υπολογισμό της και την εύκολη εφαρμογή της. Ορίζεται μαθηματικά όπως παρακάτω:

$$ReLU(x) = \max(0, x)$$

Τα κύρια της πλεονεκτήματα είναι η απλότητα υπολογισμού καθώς και η ικανότητα αντιμετώπισης του προβλήματος της εξαφάνισης των κλίσεων που αναφέρθηκε προηγουμένως (vanishing gradient problem) και αδυνατούσε να λύσει η σιγμοειδής συνάρτηση.

Χρησιμοποιήθηκε πρώτη φορά από τον Fukushima το 1969 στη δημοσίευση του με τίτλο «*Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements*» (Fukushima, 1969). Όμως αργότερα εγκαταλείφθηκε με αποτέλεσμα στα αρχικά στάδια της βαθιάς μάθησης να χρησιμοποιείται κυρίως η σιγμοειδής

συνάρτηση. Η ReLU ξαναέγινε γνωστή το 2009 με τη δημοσίευση «What is the Best Multi-Stage Architecture for Object Recognition?» (Jarrett et al., 2009), καθώς και με τις δημοσιεύσεις «*Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair*» (Nair and Hinton, 2010) και «Deep Sparce Rectifier Neural Networks» (Glorot, Bordes and Bengio, 2011) αναδείχθηκε η χρησιμότητα της και ο καθοριστικός της ρόλος για τη καθιέρωση των σύγχρονων νευρωνικών μοντέλων.

Όμως πρέπει να επισημανθεί και ένα σοβαρό μειονέκτημα που παρουσιάζει η χρήση της. Συγκεκριμένα, αν όλα τα δεδομένα που εισάγονται στη συνάρτηση είναι αρνητικά τότε το αποτέλεσμα που εξάγεται είναι μηδέν, καθιστώντας τον αντίστοιχο νευρώνα «απενεργοποιημένο- νεκρό» και να μη συμμετέχει στη περαιτέρω εκπαίδευση του μοντέλου και στην ανανέωση των βαρών κατά τη διάρκεια της οπισθοδιάδοσης (backpropagation). Η οπισθοδιάδοση είναι η τεχνική με την οποία ενημερώνονται οι παράμετροι του μοντέλου και πραγματοποιείται η εκπαίδευση του. Θα αναλυθεί σε επόμενη ενότητα. Αν πολλοί νευρώνες σε ένα μοντέλο απενεργοποιηθούν λόγω του παραπάνω προβλήματος, μπορεί να οδηγήσει σε σημαντική απώλεια χωρητικότητας του δικτύου με αποτέλεσμα να μην μπορεί να προσαρμοστεί ικανοποιητικά στα δεδομένα εκπαίδευσης. Αυτό παρατηρείται κυρίως στα μοντέλα με πολλά επίπεδα, δηλαδή βαθιά νευρωνικά δίκτυα. Το πρόβλημα που περιγράφεται παραπάνω λέγεται «*dying ReLU problem*».

- **Συνάρτηση Leaky ReLU**

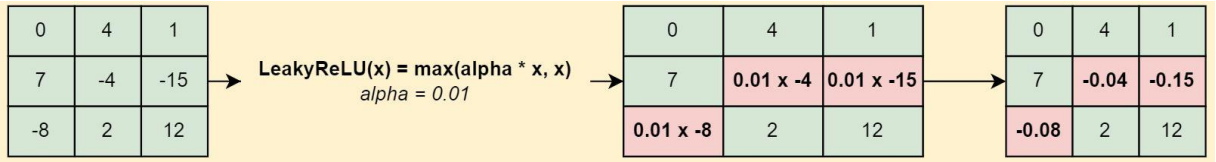
Η συνάρτηση Leaky ReLU αποτελεί μια βελτιωμένη εκδοχή της ReLU συνάρτησης. Όπως έχει προαναφερθεί, στη ReLU η κλίση για όλες τις τιμές μικρότερες του μηδενός ισούται με μηδέν, γεγονός που δύναται να απενεργοποιήσει τους νευρώνες και να προκαλέσει το «dying ReLU problem».

Αντιθέτως, η Leaky ReLU δημιουργήθηκε για να επιλύσει αυτό το ζήτημα. Στη συγκεκριμένη περίπτωση εισάγεται μια μικρή κλίση για τις αρνητικές τιμές (συνήθως 0.1 ή 0.01) αντί για 0 όπως συμβαίνει στη ReLU ενώ οι θετικές τιμές παραμένουν αναλλοίωτες. Η μαθηματική προσέγγιση είναι η παρακάτω:

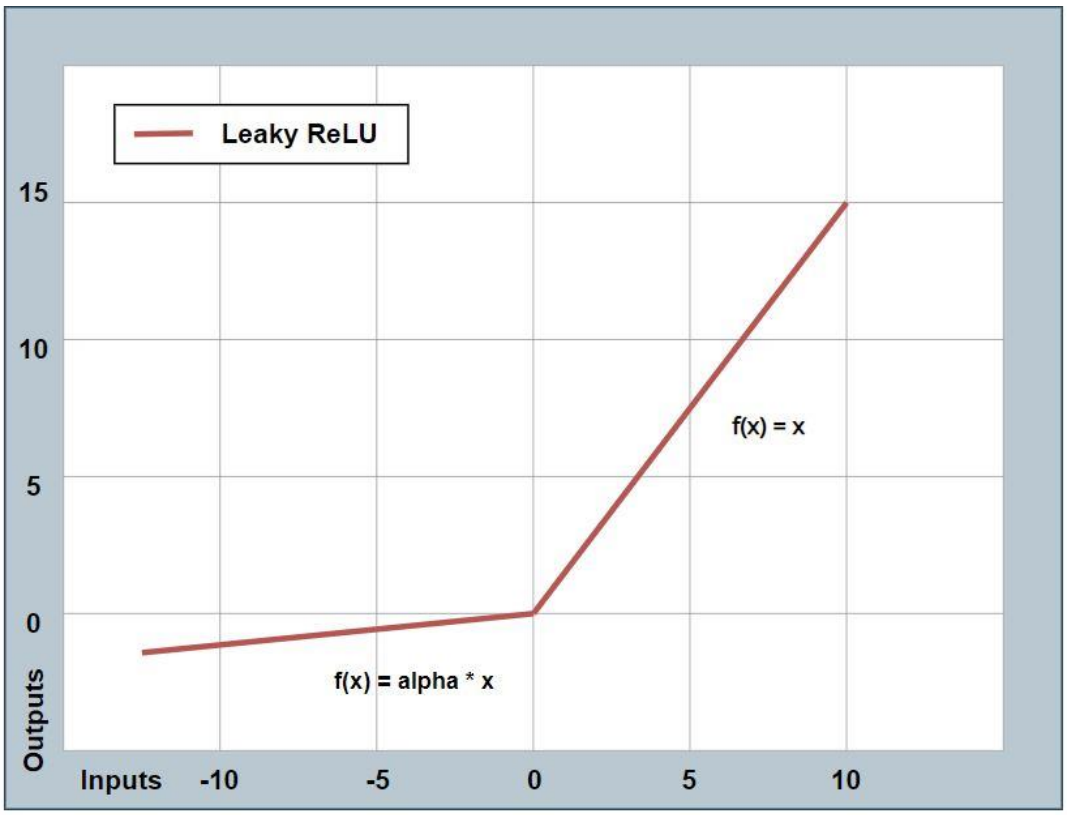
$$LeakyReLU(x) = \max(alpha * x, x)$$

$$\text{όπου } alpha \text{ συνήθως } 0.01 \text{ ή } 0.1$$

Έτσι οι αρνητικές τιμές προσεγγίζουν το μηδέν αλλά δεν ισούνται με αυτό όπως στη περίπτωση της ReLU, λύνοντας τα θέμα του «dying ReLU» (Maas, Hannun and Ng, 2013). Ακολουθεί ένα παράδειγμα στο Σχήμα 19.



Σχήμα 19 Παράδειγμα εφαρμογής συνάρτησης Leaky ReLU (πηγή: https://www.educative.io/answers/what-is-leaky-relu)



Σχήμα 20 Καμπύλης συνάρτησης Leaky ReLU (πηγή: https://www.educative.io/answers/what-is-leaky-relu)

2.3.3 Η ΕΝΝΟΙΑ ΤΗΣ ΣΥΝΑΡΤΗΣΗΣ LOSS (LOSS FUNCTION)

Η συνάρτηση loss γνωστή ως και συνάρτηση κόστους ή συνάρτηση σφάλματος αποτελεί μια από τις πιο σημαντικές έννοιες στα νευρωνικά δίκτυα αφού αποτελεί σημαντικό παράγοντα για την εκπαίδευση τους στα δεδομένα εκπαίδευσης.

Στην ενότητα 2.3.1 είχε παρουσιαστεί ο τρόπος που τα εισαγόμενα δεδομένα (input data) επεξεργάζονται από το νευρωνικό μοντέλο σε κάθε επίπεδο του και στο τέλος παράγεται ένα αποτέλεσμα.

Κατά την εκπαίδευση, το μοντέλο επιχειρεί να προσδιορίσει τη σχέση ανάμεσα στα δεδομένα εκπαίδευσης που αποτελούν τα εισαγόμενα δεδομένα και στο τελικό αποτέλεσμα.. Ο προσδιορισμός της σχέσης επιτυγχάνεται με την ανανέωση των τιμών των βαρών και των αποκλίσεων του μοντέλου. Μια γενική μορφής της εξίσωσης που ισχύει στο μοντέλο είναι η παρακάτω:

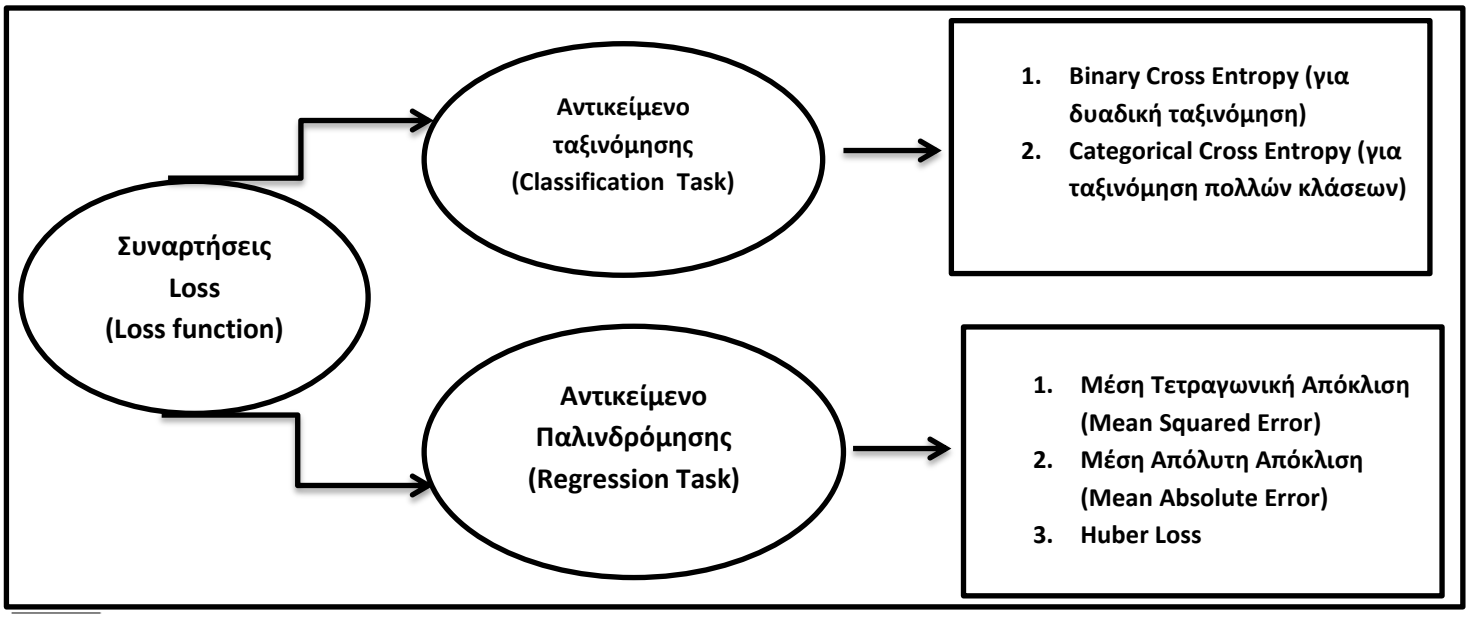
$$\hat{y} = \sigma(w^T x + b) \qquad (1)$$

Στη σχέση 1 το w αποτελεί το διάνυσμα των βαρών του μοντέλου. Το x συμβολίζει τα δεδομένα εκπαίδευσης που εισάγονται στο μοντέλο. Το b απεικονίζει το διάνυσμα της απόκλισης του μοντέλου. Η σχέση $w^T x + b$ αποτελεί το γραμμικό συνδυασμό των δεδομένων εκπαίδευσης, των βαρών και των αποκλίσεων (παράμετροι του μοντέλου). Το σ συμβολίζει τη συνάρτηση ενεργοποίησης (activation function) που εισάγει τη μη γραμμικότητα στο μοντέλο όπως έχει προαναφερθεί.

Η διαδικασία κατά την οποία τα δεδομένα εκπαίδευσης τροφοδοτούνται στο νευρωνικό δίκτυο και διέρχονται από τα επίπεδα του ονομάζεται «forward propagation» (προώθηση προς τα εμπρός). Όταν το μοντέλο έχει εξάγει το αποτέλεσμα (*predicted output*), αυτό συγκρίνεται με την αληθή τιμή (*target output*). Έπειτα μέσω της διαδικασίας που ονομάζεται «*backpropagation*» (αντίστροφη διάδοση) οι παράμετροι του μοντέλου προσαρμόζονται με σκοπό το τελικό αποτέλεσμα πρόβλεψης να προσεγγίζει την αληθή τιμή και να ελαχιστοποιείται η συνάρτηση loss.

Η συνάρτηση loss αποτελεί μια συνάρτηση που συγκρίνει την αληθή τιμή με τη πρόβλεψη του μοντέλου και αποτελεί μέτρο ελέγχου της απόδοσης του μοντέλου. Στόχος αποτελεί η ελαχιστοποίηση της κατά την εκπαίδευση του μοντέλου

Ανάλογα το αντικείμενο της Βαθιάς Μάθησης δηλαδή είτε ταξινόμηση (classification) είτε παλινδρόμηση (regression) , ο τύπος της loss συνάρτησης που χρησιμοποιείται είναι διαφορετικός.



Σχήμα 21 Συνηθισμένοι τύποι Loss συνάρτησης

Παρακάτω θα αναλυθεί η πιο διάσημη loss συνάρτηση η οποία χρησιμοποιείται σε αντικείμενα παλινδρόμησης (regression tasks), η μέση τετραγωνική απόκλιση (Mean Squared Error).

- **Mean Squared Error**

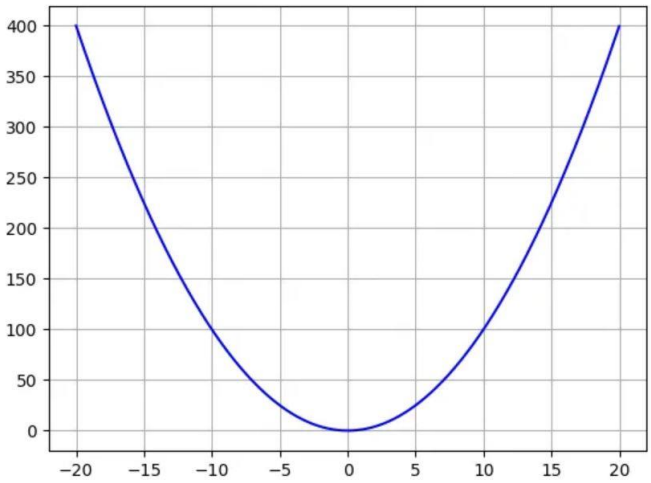
Η συγκεκριμένη συνάρτηση χρησιμοποιείται για τον υπολογισμό της διαφοράς μεταξύ της αληθής τιμής και της τιμής που προέβλεψε το μοντέλο. Συγκεκριμένα υπολογίζει τη μέση τιμή των τετραγώνων των διαφορών μεταξύ των αληθών τιμών και των τιμών που προέβλεψε το μοντέλο.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

όπου:

- n ο αριθμός παρατηρήσεων,
- y_i οι αληθείς τιμές του μοντέλου,
- $\hat{y}_i$ οι τιμές που προβλέπονται από το μοντέλο.

Τα κύρια πλεονεκτήματα της εν λόγω συνάρτησης είναι η ευκολία στον υπολογισμό της καθώς και ότι είναι συνεχής και παραγωγίσιμη σε όλο το πεδίο της, γεγονός που τη καθιστά κατάλληλη για χρήση σε αλγορίθμους βελτιστοποίησης. Έχει ένα ολικό ελάχιστο, γεγονός που επιτρέπει τη βελτιστοποίηση μέσω αλγορίθμων όπως ο «gradient descent». Ο όρος βελτιστοποίηση αναφέρεται στο τρόπο ελαχιστοποίησης της «loss» συνάρτησης. Ένα μειονέκτημα της συγκεκριμένης συνάρτησης είναι η ευαισθησία της στις ακραίες τιμές.



Σχήμα 22 Καμπύλη MSE Loss συνάρτησης (πηγή: https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression)

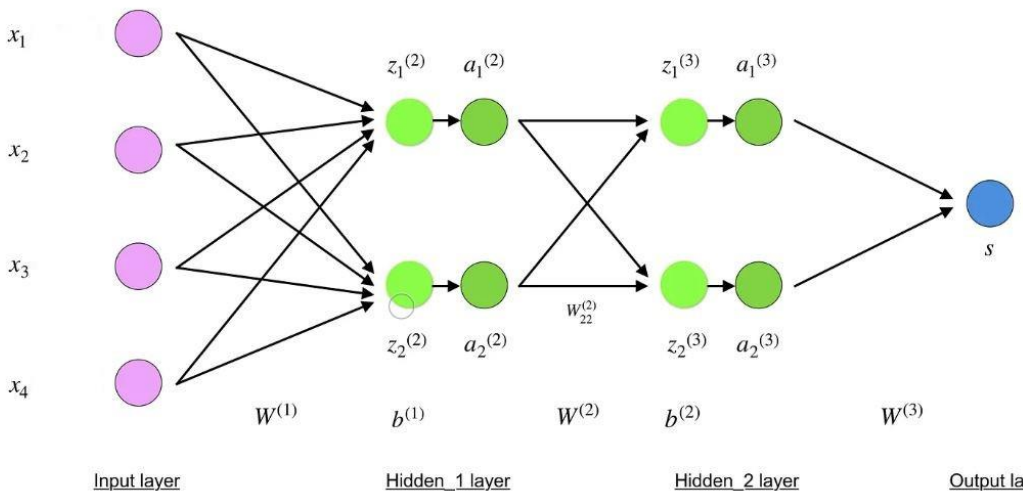
2.3.4 Ο ΑΛΓΟΡΙΘΜΟΣ BACKPROPAGATION ΣΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Ο αλγόριθμος του «*backpropagation*» αποτελεί έναν από τους πιο σημαντικούς μηχανισμούς στην εκπαίδευση των νευρωνικών δικτύων. Στη δεκαετία του 1960 διατυπώθηκε για πρώτη φορά ενώ με τη δημοσίευση «*Learning representations by back-propagating errors*» αναδείχθηκε η αξίας του αλγορίθμου για τη βαθιά μάθηση. (Rumelhart, Hinton and Williams, 1986).

Μετά από κάθε διαδικασία προώθησης των δεδομένων εκπαίδευσης μέσα από τα επίπεδα του νευρωνικού μοντέλου (forward pass), ακολουθεί μια διαδικασία αντίστροφης διάδοσης προς τα πίσω κατά την οποία υπολογίζονται οι παράμετροι του μοντέλου (βάρη και αποκλίσεις) οι οποίοι ελαχιστοποιούν την συνάρτηση loss.

Αφού έχει υπολογιστεί η τιμή πρόβλεψης του μοντέλου συγκρίνεται με τη αληθή τιμή του μέσω της συνάρτησης loss. Γνωρίζοντας τη τιμή της loss συνάρτησης, υπάρχει μέτρο εκτίμησης της απόδοσης του μοντέλου. Ύστερά, μέσω του αλγορίθμου του «backpropagation» επιδιώκεται ελαχιστοποίηση της συνάρτησης loss μέσω της ρύθμισης των βαρών και των παραμέτρων του μοντέλου. Το επίπεδο της ρύθμισης αποφασίζεται μέσω του υπολογισμού των μερικών παραγώγων της συνάρτησης loss ως προς κάθε παράμετρο του μοντέλου. Η μερική παράγωγος δείχνει το βαθμό που πρέπει να τροποποιηθεί κάθε παράμετρος του μοντέλου (σε θετική ή αρνητική διεύθυνση) με σκοπό την ελαχιστοποίηση της συνάρτησης loss. Ο υπολογισμός των μερικών παραγώγων γίνεται με μια τεχνική που ονομάζεται κανόνας της αλυσίδας.

Έστω ένα νευρωνικό δίκτυο με τέσσερα επίπεδα (1 εισαγόμενο επίπεδο (input layer), δύο ενδιάμεσα επίπεδα (hidden layers) και το τελικό επίπεδο (output layer)).



Σχήμα 23 Παράδειγμα νευρωνικού μοντέλου (πηγή: https://towardsdatascience.com/understanding-backpropagation-algorithm)

Στο εισαγόμενο επίπεδο υπάρχουν 4 νευρώνες και συμβολίζονται ως διάνυσμα $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$ με διάσταση (4,1).

Στο επίπεδο 2 (hidden layer 1) υπάρχουν 2 νευρώνες και συμβολίζονται ως διάνυσμα $z^{(2)} = \begin{bmatrix} z_1^{(2)} \\ z_2^{(2)} \end{bmatrix}$ όπου $z^{(2)} = W^{(1)}x + b^{(1)}$ (1) .

Το σύμβολο $W^{(1)}$ είναι ένας πίνακας βαρών σχήματος (2,4) όπου 2 είναι ο αριθμός των εξαγόμενων νευρώνων (νευρώνες στο hidden layer 1) και 4 ο αριθμός των εισαγόμενων νευρώνων (νευρώνες στο input layer). Ισούται με:

$$W^{(1)} = \begin{bmatrix} W_{11}^{(1)} & W_{12}^{(1)} & W_{13}^{(1)} & W_{14}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} & W_{23}^{(1)} & W_{24}^{(1)} \end{bmatrix}$$

Το σύμβολο $b^{(1)}$ είναι ένα διάνυσμα απόκλισης με σχήμα (2,1) όπου 2 ο αριθμός των νευρώνων στο hidden layer 1 και ισούται με :

$$b^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix}$$

Από σχέση 1 προκύπτει

$$z^{(2)} = W^{(1)}x + b^{(1)} = \begin{bmatrix} W_{11}^{(1)}x_1 & W_{12}^{(1)}x_2 & W_{13}^{(1)}x_3 & W_{14}^{(1)}x_4 \\ W_{21}^{(1)}x_1 & W_{22}^{(1)}x_2 & W_{23}^{(1)}x_3 & W_{24}^{(1)}x_4 \end{bmatrix} + \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix} \quad (2)$$

Έπειτα από το hidden layer 1 έπεται η συνάρτηση ενεργοποίησης η οποία εισάγει τη μη γραμμικότητα στο νευρωνικό μοντέλο και συμβολίζεται όπως παρακάτω:

$$a^{(2)} = f(z^{(2)})$$

Ομοίως στο hidden layer 2, οι νευρώνες συμβολίζονται όπως παρακάτω:

$$z^{(3)} = W^{(2)}a^{(2)} + b^{(2)}$$

και στη συνέχεια συνάρτηση ενεργοποίησης

$$a^{(3)} = f(z^{(3)})$$

Το τελικό επίπεδο (output layer) στο συγκεκριμένο παράδειγμα νευρωνικού μοντέλου παρουσιάζεται ως ένας νευρώνας και συμβολίζεται όπως παρακάτω.

$$s = W^{(3)}a^{(3)}$$

Αφού έχει υπολογιστεί και η τελική τιμή που προέβλεψε το μοντέλο, πραγματοποιείται η σύγκριση μεταξύ της τιμής πρόβλεψης s και της πραγματικής τιμής y μέσω της συνάρτησης loss C (MSE).

$$C = cost(s, y)$$

Όπως έχει προαναφερθεί, με βάση την τιμή της loss συνάρτησης , το μοντέλο γνωρίζει πόσο πρέπει να τροποποιήσει τις παραμέτρους του με σκοπό την ελαχιστοποίηση της. Η διαδικασία κατά την οποία πραγματοποιείται η ελαχιστοποίηση όπως έχει προαναφερθεί είναι η backpropagation.

Κατά τη διάρκεια της backpropagation οι παράμετροι του μοντέλου (βάρη και αποκλίσεις) ενημερώνονται όπως παρακάτω:

$$W = W - \varepsilon \frac{\theta C}{\theta W}$$

$$b = b - \varepsilon \frac{\theta C}{\theta b}$$

όπου ε είναι ο ρυθμός μάθησης (learning rate) δηλαδή το βήμα με το οποίο το νευρωνικό μοντέλο ενημερώνει τις παραμέτρους του για την ελαχιστοποίηση της συνάρτησης loss. Ο ρυθμός μάθησης αποτελεί υπερπαράμετρο του μοντέλου, δηλαδή καθορίζεται από τον χρήστη. Ο καθορισμός του ρυθμού μάθησης είναι κρίσιμος για την εκπαίδευση ενός νευρωνικού δικτύου επειδή ένας υψηλός ρυθμός μπορεί να οδηγήσει αποτυχία σύγκλισης υπερβαίνοντας το ελάχιστο σημείο της συνάρτησης loss αλλά και ένας χαμηλός ρυθμός δύναται να οδηγήσει σε μια αργή και δαπανηρή σε υπολογιστικούς πόρους διαδικασία εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 3: ΘΕΩΡΗΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΟΥ SEGMENT ANYTHING

Στο κεφάλαιο αυτό, επιχειρείται η θεωρητική προσέγγιση του προ-εκπαιδευόμενου μοντέλου Segment Anything, η συνολική του αρχιτεκτονική, η διαδικασία εκπαίδευσης του, καθώς και μεθοδολογίες που ακολουθεί με σκοπό την εξαγωγή οντοτήτων ενδιαφέροντος από τηλεπισκοπικές απεικονίσεις.

3.1 ΕΙΣΑΓΩΓΙΚΑ ΣΤΟΙΧΕΙΑ

Με τη δημοσίευση «Segment Anything» στις 05 Απριλίου 2023 από την επιστημονική ομάδα της *Meta's Ai*, εισάγεται η έννοια και το μοντέλο του Segment Anything. (Kirillov et al., 2023). Το συγκεκριμένο μοντέλο αποτελεί ένα πολλά υποσχόμενο μοντέλο θεμελίωσης (foundation model) για κατάτμηση εικόνας (image segmentation).

Ουσιαστικά αποτελείται από τρία αλληλένδετα στοιχεία, το σκοπό για το οποίο δημιουργήθηκε δηλαδή τμηματοποίηση με βάση κάποιο όρισμα (**promptable segmentation task**), το μοντέλο τμηματοποίησης SAM το οποίο χαρακτηρίζεται από δυνατότητα zero-shot μάθησης (**segment anything model**) και μια μηχανή δεδομένων η οποία συνέλεξε το σετ δεδομένων πάνω στο οποίο εκπαιδεύτηκε το συγκεκριμένο μοντέλο (**segment anything 1B, SA-1B**). Συγκεκριμένα το μοντέλο Segment Anything έχει εκπαιδευτεί σε ένα σετ δεδομένων το οποίο αποτελείται από 11 εκατομμύρια εικόνες και πάνω από 1.1 δισεκατομμύρια μάσκες. Το ίδιο το μοντέλο χρησιμοποιήθηκε με σκοπό τη συλλογή του συγκεκριμένου σετ δεδομένων. Επιπλέον το μοντέλο έχει σχεδιαστεί και εκπαιδευτεί με τέτοιο τρόπο ώστε με βάση τα ορίσματα που δέχεται (prompts) να τμηματοποιεί νέες κατανομές εικόνων. Αυτός είναι και ο απώτερος σκοπός της συγκεκριμένης ερευνητικής ομάδας, η δημιουργία ενός μοντέλου θεμελίωσης για κατάτμηση άγνωστων προς το μοντέλο εικόνων μέσω τη χρήσης ορισμάτων.

3.2 Η ΕΝΝΟΙΑ ΤΟΥ ZERO SHOT LEARNING (ZSL)

Όπως προαναφέρθηκε, την τεχνική του «zero-shot learning » τη χρησιμοποιεί το Segment Anything μοντέλο όπως και άλλοι σύγχρονοι αλγόριθμοι υπολογιστικής όρασης. Η συγκεκριμένη αρχή επιτρέπει σε ένα μοντέλο να ολοκληρώσει μια διεργασία χωρίς να έχει εκπαιδευτεί σε αντίστοιχα δεδομένα εκπαίδευσης,

χρησιμοποιώντας συμπληρωματικές πληροφορίες όπως περιγραφές κειμένου για να καταλάβει το περιεχόμενο μιας εικόνας.

Στην ZSL (Zero Shot Learning), ένα μοντέλο είναι προ-εκπαιδευμένο σε μια ομάδα κλάσεων (seen classes) και μετά επιχειρεί να γενικεύσει τη γνώση του σε μια διαφορετική ομάδα κλάσεων (unseen classes) χωρίς πρόσθετη εκπαίδευση. Ο σκοπός της συγκεκριμένης τεχνικής είναι η μεταφορά της υπάρχουσας γνώσης ενός μοντέλου που έχει προέλθει από δεδομένα εκπαίδευσης που αφορούν διαφορετικές κλάσεις από αυτές που καλείται να προβλέψει. Ουσιαστικά το ZSL αποτελεί ένα υποπεδίο της μεταφοράς μάθησης (transfer learning) και εντάσσεται στη κατηγορία της ετερογενούς μεταφοράς μάθησης (heterogeneous transfer learning) όπου οι χώροι χαρακτηριστικών (feature space) και ετικετών (label space) διαφέρουν. Παράδειγμα χρήσης του zero-shot learning στην υπολογιστική όραση είναι το CLIP, ένας ταξινομητής εικόνων που δημιουργήθηκε από την OpenAI. (Radford et al., 2021)

Πρώτη φορά, ο όρος zero-shot learning εμφανίστηκε το 2009 στη δημοσίευση «Zero-Shot Learning with Semantic Output Codes» (Palatucci et al., 2009). Για τη λειτουργία του, απαιτείται ο διαχωρισμός των δεδομένων σε τρεις κατηγορίες: τις κλάσεις με τις οποίες έχει εκπαιδευτεί το μοντέλο (seen classes), τις κλάσεις τις οποίες το μοντέλο καλείται να ταξινομήσει χωρίς περαιτέρω εκπαίδευση(unseen classes) και συμπληρωματικές πληροφορίες οι οποίες δύναται να είναι περιγραφές, σημασιολογικές πληροφορίες για τα δεδομένα στα οποία δεν έχει εκπαιδευτεί το μοντέλο. Οι συμπληρωματικές πληροφορίες βοηθούν το μοντέλο να κατανοήσει και να αναγνωρίσει τα δεδομένα από τις κλάσεις που δεν έχει εκπαιδευτεί βασιζόμενο σε περιγραφές ή σημασιολογικές σχέσεις με τις κλάσεις εκπαίδευσης.

Η διαδικασία του Zero-Shot Learning περιλαμβάνει δύο στάδια: την εκπαίδευση (training) και τη πρόβλεψη (inference). Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο μαθαίνει από ένα σετ δεδομένων που έχουν ετικέτα (label) δηλαδή έχει αναγνωριστεί και επισημανθεί η κατηγορία τους. Τα δεδομένα αυτά προέρχονται από τις «seen classes». Κατά τη διάρκεια της πρόβλεψης, το μοντέλο χρησιμοποιεί την υπάρχουσα γνώση και τη συμπληρωματική πληροφορία να ταξινομήσει νέες κλάσεις (unseen classes).

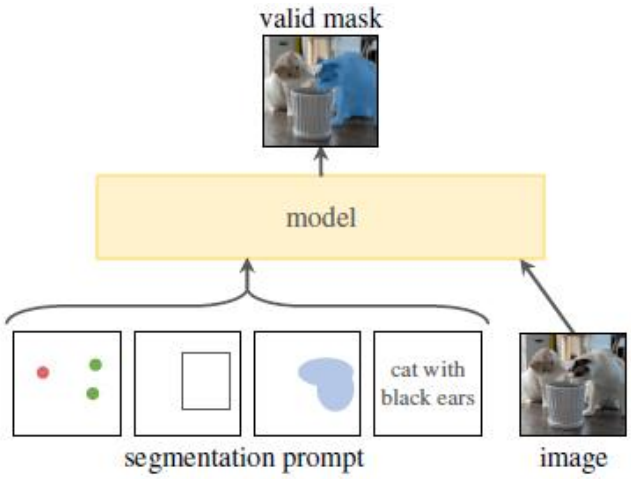
Ο άνθρωπος είναι ικανός να πραγματοποιήσει «Zero Shot Learning» εξαιτίας της υπάρχουσας γνώσης που έχουν όσο αφορά τη γλώσσα. Δύναται να πραγματοποιήσουν συνδέσεις ανάμεσα σε καινούργιες, άγνωστες κλάσεις δεδομένων και σε ήδη γνωστές, λόγω της υπάρχουσας γνώσης. Στην περίπτωση της υπολογιστικής όρασης, η γνώση μεταφράζεται ως ένα επισημασμένο (labeled) σετ δεδομένων από γνωστές και άγνωστες κλάσεις. Τα δεδομένα που παρέχονται πρέπει να συσχετίζονται σε ένα πολυδιάστατο διανυσματικό χώρο, γνωστό ως και σημασιολογικό χώρο (semantic space). Στο συγκεκριμένο χώρο πραγματοποιείται και η μεταφορά της γνώσης από τις γνωστές κλάσεις προς τις άγνωστες.

Μια τεχνική για την μεταφορά γνώσης από γνωστές σε άγνωστες κλάσεις στο σημασιολογικό χώρο είναι η σημασιολογική ενσωμάτωση χρησιμοποιώντας διάνυσμα χαρακτηριστικών. Έστω έχουμε δύο διαφορετικές κλάσεις που αποτελούνται από εικόνες από γάτες και πτηνά. Δημιουργείται ένα διάνυσμα χαρακτηριστικών (attribute vector) για κάθε κατηγορία που περιέχει τιμές ανάλογα τα χαρακτηριστικά των κλάσεων. Στη συγκεκριμένη περίπτωση, τα χαρακτηριστικά που εμπεριέχονται στο διάνυσμα είναι (ουρά, γούνα, ράμφος, φτερά, και μουστάκια). Στο Σχήμα 24, αποτυπώνονται οι τιμές των δυο διανυσμάτων των χαρακτηριστικών. Έστω ότι η γνωστή κλάση είναι η γάτα. Κατά τη διάρκεια τη φάσης της εκπαίδευσης, το μοντέλο εκπαιδεύεται με τη γνωστή κλάση και μαθαίνει να συσχετίζει τις εικόνες της με τα αντίστοιχα διανύσματα χαρακτηριστικών τους. Έπειτα κατά τη διάρκεια της πρόβλεψης (inference), το μοντέλο συναντά την άγνωστη κλάση και χρησιμοποιεί το διάνυσμα των χαρακτηριστικών της για να αναγνωρίσει τις άγνωστες εικόνες. Το μοντέλο συγκρίνει το διάνυσμα χαρακτηριστικών των πτηνών (άγνωστη κλάση) με το διάνυσμα της γάτας (γνωστή κλάση) για να αποφανθεί που θα το ταξινομήσει. Σε πραγματικά παραδείγματα ο αριθμός των κλάσεων είναι σαφώς μεγαλύτερος.

Σχήμα 24 Διανύσματα χαρακτηριστικών των διαφορετικών κλάσεων (Πηγή: https://blog.roboflow.com/zero-shot-learning-computer-vision)

3.3 Ο ΣΚΟΠΟΣ ΤΟΥ SEGMENT ANYTHING (SEGMENT ANYTHING TASK)

Το αντικείμενο που θα επιτελούσε το μοντέλο του Segment Anything έχει εμπνευστεί από τα συστήματα επεξεργασίας φυσικής γλώσσας (NLP) και τις νέες τεχνικές υπολογιστικής όρασης. Στα NLP και στο τομέα της σύγχρονης υπολογιστικής όρασης , υπάρχουν μοντέλα θεμελίωσης (foundation) που μπορούν να εκτελέσουν τη τεχνική zero-shot μάθησης σε νέα σετ δεδομένων μέσω τεχνικών που προϋποθέτουν συμπληρωματικές πληροφορίες ή οδηγίες (prompts). Έτσι μέσω του «Segment Anything» εισάγεται ο όρος της «promptable segmentation», όπου ο σκοπός είναι η επιστροφή μιας έγκυρης μάσκας τμηματοποίησης (segmentation mask) έχοντας κάποια οδηγία τμηματοποίησης (segmentation prompt).



Εικόνα 25 Αντικείμενο της τμηματοποίησης με οδηγίες (Task of promptable segmentation) (Πηγή: Kirillov et al., 2023)

Η οδηγία τμηματοποίησης ουσιαστικά καθοδηγεί το μοντέλο που θα πραγματοποιήσει τμηματοποίηση μέσα στην εικόνα. Η οδηγία τμηματοποίησης δύναται να περιλαμβάνει χωρική πληροφορία ή να αποτελεί κάποιο κείμενο με σκοπό τον εντοπισμό ενός αντικειμένου. Η έγκυρη μάσκα τμηματοποίησης αναφέρεται στο γεγονός ότι ακόμα και αν η οδηγία τμηματοποίησης (segmentation prompt) είναι αμφιλεγόμενη και αναφέρεται σε πολλαπλά αντικείμενα, το αποτέλεσμα θα είναι μια λογική μάσκα για τουλάχιστον ένα από αυτά τα αντικείμενα.

Τα είδη των οδηγιών τμηματοποίησης που χρησιμοποιούνται στο Segment Anything μοντέλο είναι τα παρακάτω:

- Ένα σετ σημειακών δεδομένων για καθορισμό υπόβαθρου ή προσκήνιου εικόνας (foreground or background points).
- Κάποιο πλαίσιο οριοθέτησης (bounding box)

- Περιγραφή αντικειμένου (text information)

Εικόνα 26 Διαφορετικά είδη οδηγιών για τμηματοποίηση (types of segmentation prompts) (Πηγή: https://medium.com/towards-data-science/segment-anything-promptable-segmentation-of-arbitrary-objects)

Το αντικείμενο της τμηματοποίησης της εικόνας με χρήση οδηγιών-προτροπών (promptable segmentation) προτείνει έναν αλγόριθμο για προ-εκπαίδευση ο οποίος προσομοιώνει μια σειρά από προτροπές (σημεία, κουτιά, μάσκες) για κάθε δείγμα εκπαίδευσης και συγκρίνει τα αποτελέσματα του μοντέλου (predictions) με τις αληθείς τιμές (ground truth). Η μέθοδος που χρησιμοποιείται για την προ-εκπαίδευση του μοντέλου προέρχεται από την διαδραστική τμηματοποίηση (interactive segmentation) (Xu et al., 2016), της οποίας ο σκοπός είναι η πρόβλεψη μιας έγκυρης μάσκας με είσοδο αρκετών πληροφοριών από τον χρήστη. Αντιθέτως στη περίπτωση του Segment Anything, ο στόχος είναι η πρόβλεψη μιας έγκυρης μάσκας για οποιαδήποτε προτροπή ακόμα και αν χαρακτηρίζεται ως ασαφής. Αυτό διασφαλίζει ότι το προ-εκπαιδευμένο μοντέλο θα είναι αποτελεσματικό σε περιπτώσεις που θα υπάρχει ασάφεια, όπως στη διαδικασία της αυτόματης εξαγωγής μασκών, διαδικασία που θα αναλυθεί παρακάτω.

Η τμηματοποίηση αποτελεί ένα ευρύ πεδίο που περιλαμβάνει διάφορες τεχνικές και εφαρμογές όπως η διαδραστική τμηματοποίηση (interactive segmentation), η ανίχνευση περιγράμματος (edge detection), η σημασιολογική τμηματοποίηση (semantic segmentation), η τμηματοποίηση αντικειμένων (instance segmentation) καθώς και άλλες. Ο στόχος της τμηματοποίησης με χρήση προτροπών είναι παραγωγή ενός μοντέλου με ευρείες δυνατότητες που να έχει τη δυνατότητα να προσαρμόζεται σχεδόν σε όλες τις εφαρμογές τμηματοποίησης μέσω της χρήσης προτροπών. Αυτή η δυνατότητα του μοντέλου αποτελεί μια μορφή γενίκευσης εργασίας. Το συγκεκριμένο μοντέλο δύναται να αποτελέσει ένα μέρος ενός μεγαλύτερου συστήματος εκτελώντας διαφορετικές εργασίες. Για παράδειγμα, το Segment Anything μπορεί να χρησιμοποιηθεί για διαδικασίες instance segmentation

(τμηματοποίηση αντικειμένων) αν συνδυαστεί με ένα υπάρχον ανιχνευτή αντικειμένων (object detection).

3.4 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΜΟΝΤΕΛΟΥ SEGMENT ANYTHING

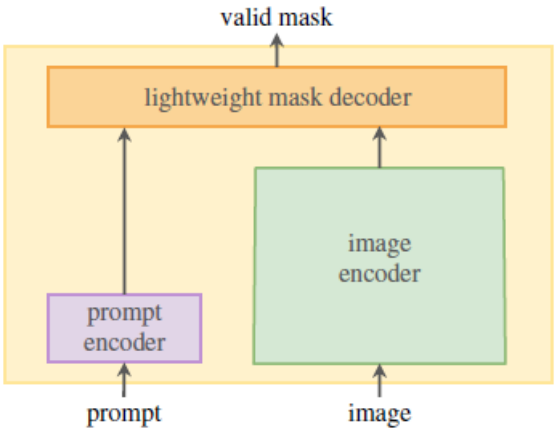
Λόγω των απαιτήσεων του σκοπού της τμηματοποίησης εικόνας βάσει οδηγιών καθώς και της χρήσης του μοντέλου σε πραγματικό χρόνο από χρήστες παρατηρούνται διάφοροι περιορισμοί στην αρχιτεκτονική του μοντέλου όπως παρακάτω:

- Το μοντέλο πρέπει να υποστηρίζει ευέλικτα είδη οδηγιών (prompts), δηλαδή να ανταποκρίνεται σε διαφορετικά είδη οδηγιών.
- Να υπολογίζει μάσκες σε πραγματικό χρόνο για να επιτρέπει τη διάδραση με το χρήστη.
- Το μοντέλο να είναι ικανό να διαχειρίζεται αμφιλεγόμενες οδηγίες που αφορούν πολλαπλά αντικείμενα.

Οι παραπάνω περιορισμοί ικανοποιούνται από μια απλή σχεδίαση αρχιτεκτονικής για το Segment Anything μοντέλο. Σύμφωνα με τη συγκεκριμένη αρχιτεκτονική το μοντέλο απαρτίζεται από τρία διαφορετικά μέρη όπως παρακάτω:

- Κωδικοποιητής Εικόνας (Image Encoder). Επεξεργάζεται την είσοδο της εικόνας και τη μετατρέπει σε ένα σύνολο χαρακτηριστικών (image embedding).
- Κωδικοποιητής Οδηγιών (Prompt Encoder). Μετατρέπει την οδηγία σε κάποια αναπαράσταση που θα χρησιμοποιηθεί σε μεταγενέστερο χρόνο από το μοντέλο.
- Αποκωδικοποιητής Μάσκας (Lightweight Mask Decoder). Το τμήμα του μοντέλου που συνδυάζει τα χαρακτηριστικά της εικόνας και των οδηγιών και προβλέπει τη τελική μάσκα τμηματοποίησης.

Συνολικά αυτά τα τρία μέρη συνθέτουν το Segment Anything μοντέλο όπως απεικονίζεται στο Σχήμα 27. Το μοντέλο έχει βασιστεί στα Transformer μοντέλα όρασης (transformer vision models) όπως αυτά έχουν προσδιοριστεί το 2020 στη δημοσίευση με όνομα «End-to End Object Detection with Transformers». Αξίζει να σημειωθεί ότι η ύπαρξη ενός ξεχωριστού αποκωδικοποιητή εικόνας (image encoder) στο μοντέλο, επιτρέπει την επαναχρησιμοποίηση των χαρακτηριστικών της εικόνας με διαφορετικές οδηγίες κάθε φορά. Ουσιαστικά η χρονοβόρα διεργασία στο συγκεκριμένο μοντέλο αποτελεί η αποκωδικοποίηση της εκάστοτε εικόνας. Αντιθέτως οι άλλες δύο διεργασίες απαιτούν ελάχιστο χρόνο και επιτρέπουν τη διάδραση με το χρήση σε πραγματικό χρόνο.



Σχήμα 27 Γενική δομή Segment Anything μοντέλου (πηγή: Kirillov et al., 2023)

3.4.1 ΚΩΔΙΚΟΠΟΙΗΤΗΣ ΕΙΚΟΝΑΣ (IMAGE ENCODER)

Ο κωδικοποιητής εικόνας που χρησιμοποιεί το Segment Anything μοντέλο είναι ένας προ-εκπαιδευμένος Vision Transformer (ViT) με MAE (Masked Autoencoder) για επεξεργασία υψηλής ανάλυσης εικόνας. Το συγκεκριμένο μέρος του μοντέλου λειτουργεί μια φορά για κάθε εικόνα και εφαρμόζεται πριν την υποβολή των οδηγιών (prompting) στο μοντέλο.

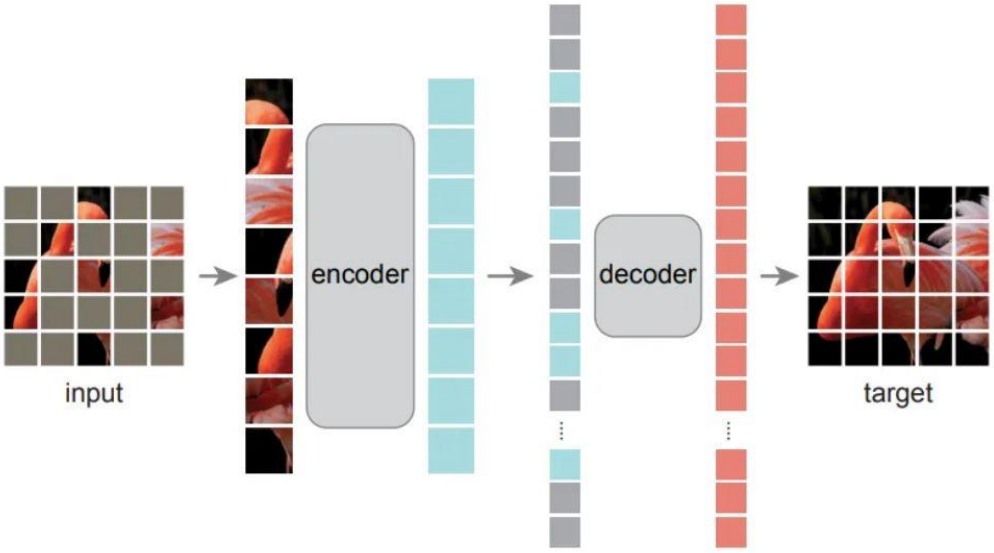
Όπως έχει προαναφερθεί, τα συνελικτικά δίκτυα ήταν κυρίαρχα στο τομέα της υπολογιστικής όρασης. Στη διαδικασία της συνέλιξης δεν είναι εύκολο να ενσωματωθούν δείκτες όπως ενσωματώσεις θέσεων (positional embeddings) και μάσκες (mask tokens). Αυτό το λειτουργικό κενό στην αρχιτεκτονική τους αντιμετωπίζεται με την έλευση των Vision Transformer (ViT). Οι Vision Transformer αποτελούν μια καινοτόμο προσέγγιση στην υπολογιστική όραση, φέρνουν μια νέα προσέγγιση στην επεξεργασία της εικόνας εφαρμόζοντας την αρχιτεκτονική των transformers η οποία ήταν ήδη επιτυχημένη στο τομέα της επεξεργασίας φυσικής γλώσσας. Σε αντίθεση με τα συνελικτικά δίκτυα, οι ViT μπορούν να ενσωματώσουν εύκολα δείκτες όπως ενσωματώσεις θέσεων και μάσκες. Συγκεκριμένα., ο ViT δέχεται ως όρισμα μια εικόνα και τη διαμερίζει σε επιμέρους κομμάτια των 16*16 εικονοστοιχείων. Ύστερα μέσω γραμμικού μετασχηματισμού (linear projection), τα επιμέρους κομμάτια της εικόνας (patches) μετατρέπονται σε μονοδιάστατα διανύσματα λόγω του γεγονότος ότι ο Transformer λειτουργεί με ακολουθίες διανυσμάτων. Η παραπάνω διαδικασία ορίζεται ως *flattening*. Η υπόλοιπη διαδικασία είναι παρόμοια με τους Transformers που χρησιμοποιούνται σε αντικείμενα φυσικής γλώσσας. Ύστερα στα διανύσματα που προκύπτουν ενσωματώνεται η έννοια της θέσης για να διατηρηθεί η χωρική πληροφορία των επιμέρους εικόνων. Η προσθήκη της χωρικής διάταξης στα διανύσματα

πραγματοποιείται μέσω των *positional embeddings*. Στη συνέχεια, ο Transformer προβλέπει τι απεικονίζει η εκάστοτε εικόνα. (Dosovitskiy et al., 2020).

Για την εφαρμογή του ViT, απαιτούνται μεγάλες ποσότητες δεδομένων με σκοπό τη μάθηση του τρόπου και του χώρου της εστίασης της προσοχής τους στην εικόνα με σκοπό την επίτευξη σωστών προβλέψεων. Αντιθέτως τα CNN μοντέλα περιορίζονται λόγω της συνέλιξης σε μια πιο τοπική εικόνα (local view), γεγονός όμως που μειώνει την ποσότητα των δεδομένων εκπαίδευσης για το μοντέλο αφού ξέρει ήδη τον τρόπο εστίασης αλλά όχι τον αντίστοιχο χώρο εστίασης μέσα στην εικόνα. Τα ViT ενδείκνυνται για αντικείμενα βαθιάς μάθησης που απαιτούν μεγάλες ποσότητες δεδομένων και υψηλούς υπολογιστικούς πόρους αφού δύναται να επιτύχουν καλύτερα αποτελέσματα από αντίστοιχα συνελικτικά μοντέλα. Αντιθέτως τα συνελικτικά νευρωνικά δίκτυα ενδείκνυνται για μικρότερου όγκου εργασίες με λιγότερους υπολογιστικούς πόρους..

Για την εισαγωγή του όρου Masked Autoencoder κρίνεται αναγκαία η κατανόηση των παρακάτω εννοιών. Μεταξύ γλώσσας και όρασης η πυκνότητα πληροφορίας διαφέρει σημαντικά. Η γλώσσα αποτελεί σήμα υψηλής σημασιολογίας και πυκνής πληροφορίας. Για αυτό το λόγο, η εκπαίδευση ενός μοντέλου για τη πρόβλεψη μερικών ελλειπόντων λέξεων σε μια πρόταση οδηγεί σε πολύπλοκη κατανόηση της γλώσσας. Αντιθέτως, οι εικόνες χαρακτηρίζονται από υψηλό χωρικό πλεονασμό, δηλαδή αν λείπει ένα μέρος της εικόνας αυτό μπορεί να ανακτηθεί από τα γειτονικά του κομμάτια χωρίς να απαιτείται κατανόηση υψηλού επιπέδου των αντικειμένων και των σκηνών της εικόνας. Για να ξεπεραστεί αυτή η διαφορά και να ενθαρρυνθεί η εκμάθηση χρήσιμων χαρακτηριστικών προτείνεται η στρατηγική των Masked Autoencoder η οποία ουσιαστικά αποκρύπτει ένα υψηλό ποσοστό τυχαίων κομματιών της εικόνας κατά την εκπαίδευση του μοντέλου. Η συγκεκριμένη στρατηγική δυσχεραίνει την εκπαίδευση του μοντέλου μειώνοντας το χωρικό πλεονασμό και προκαλώντας μια πιο πολύπλοκη κατανόηση της εκάστοτε εικόνας από το μοντέλο.

Έτσι ο Masked Autoencoder (MAE) αποκρύπτει (μασκάρει) τυχαία κομμάτια από την εικόνα εισόδου και ανακατασκευάζει τα ελλείποντα κομμάτια στο χώρο των εικονοστοιχείων. Η αρχιτεκτονική του MAE περιλαμβάνει δύο κύρια μέρη , έναν κωδικοποιητή (encoder) και έναν αποκωδικοποιητή (decoder). Ο κωδικοποιητής λειτουργεί μόνο στα ορατά μέρη της εικόνας που δεν έχουν μάσκες , ενώ ο αποκωδικοποιητής χρησιμοποιεί όλα τα κομμάτια της εικόνας (μάσκες και μη) για να ανακατασκευάσει την εικόνα εισόδου.



Σχήμα 28 Αρχιτεκτονική Masked Autoencoder (MAE) (πηγή: He et al., 2021)

Μελετώντας την αρχιτεκτονική ενός MAE, ακολουθώντας τη προσέγγιση ενός Vision Transformer, η εικόνα εισόδου διαμερίζεται σε επιμέρους μέρη τα οποία δεν επικαλύπτονται. Έπειτα ένα μέρος αυτών τυχαία επιλέγεται και τα υπόλοιπα αποκρύπτονται μέσω μάσκας. Στη συνέχεια ο κωδικοποιητής (MAE encoder) οποίος είναι ένας ViT και εφαρμόζεται μόνο στα ορατά κομμάτια (patches) της εικόνας. Τα ορατά μέρη της εικόνας αποτελούν περίπου το 25% του πλήρους συνόλου της εικόνας. Τα μασκαρισμένα κομμάτια αφαιρούνται και δε χρησιμοποιούνται από τον κωδικοποιητή. Αυτό επιτρέπει την εκπαίδευση μεγάλων κωδικοποιητών με ένα μικρό κλάσμα υπολογιστικής ισχύς και μνήμης. (He et al., 2021)

Σχετικά με τον MAE αποκωδικοποιητή, σε αυτόν εισάγεται το πλήρες σύνολο ενδείξεων που περιλαμβάνει τα κωδικοποιημένα ορατά μέρη αλλά και τις ενδείξεις μασκών (mask tokens). Η ένδειξη μάσκας αναφέρεται σε ένα διανυσματικό αντικείμενο το οποίο αντιπροσωπεύει την παρουσία μιας μάσκας σε ένα συγκεκριμένο τμήμα δεδομένων. Πρακτικά, οι μάσκες βοηθούν στην ανακατασκευή δεδομένων, βελτιώνοντας την κατανόηση των συστημάτων βαθιάς μάθησης για ελλειπή δεδομένα ή για ανάγκες πρόβλεψης. Έπειτα προστίθενται ενσωματώσεις θέσεων (positional embeddings) σε όλες τις ενδείξεις του πλήρες συνόλου. Χωρίς αυτές οι μάσκες δε θα είχαν πληροφορία για τη θέση τους στην εικόνα. Ο αποκωδικοποιητής του MAE χρησιμοποιείται μόνο κατά την προ-εκπαίδευση για την εκτέλεση του έργου ανακατασκευής εικόνας. Επομένως, η αρχιτεκτονική του αποκωδικοποιητή μπορεί να σχεδιαστεί με ευελιξία, ανεξάρτητα από τον σχεδιασμό του κωδικοποιητή. (He et al., 2021).

Στη περίπτωση του Segment Anything όπως έχει προαναφερθεί, ο κωδικοποιητής εικόνας είναι ένα προ-εκπαιδευόμενο ViT με MAE (Masked Autoencoder) που μετατρέπει την εικόνα σε ένα image embedding διαστάσεων C * H * W (κανάλια * ύψος * πλάτος) (Kirillov et al., 2023) . Το image embedding είναι μια πυκνή

πολυδιάστατη αναπαράσταση μιας εικόνας σε μορφή ενός διανύσματος. Συγκεκριμένα, χρησιμοποιείται ένας ViT –H/16 με παράθυρο προσοχής 14*14 και τέσσερα εξίσου κατανεμημένα επίπεδα παγκόσμιας προσοχής. Το ViT –H/16 αναλύεται όπως παρακάτω:

- Το H υποδηλώνει τη μεγάλη εκδοχή του Vision Transformer. (Dosovitskiy et al., 2020).
- Το 16*16 υποδηλώνει το μέγεθος των επιμέρους κομματιών. Στη προκειμένη περίπτωση, η εικόνα χωρίζεται σε επιμέρους εικόνες διαστάσεων 16*16 εικονοστοιχεία. (Dosovitskiy et al., 2020).

Το παράθυρο προσοχής (attention window) αναφέρεται στη περιοχή της εικόνας που το μοντέλο κοιτάζει ταυτόχρονα για να καθορίσει τις σχέσεις μεταξύ των εικονοστοιχείων. Έτσι το μοντέλο εξετάζει επιμέρους κομμάτια της εικόνας διαστάσεων 14*14 αντί ολόκληρη την εικόνα ταυτόχρονα.

Τα επίπεδα παγκόσμιας προσοχής (global attention blocks) αναφέρονται σε στρώσεις του μοντέλου που λαμβάνουν υπόψη ολόκληρη την εικόνα, επιτρέποντας τη κατανόηση ευρειών συσχετίσεων σε όλο το εύρος της εικόνας. Τα τέσσερα εξίσου κατανεμημένα μπλοκ αναφέρονται σε μπλοκ τα οποία είναι διανεμημένα σε διάφορα σημεία της αρχιτεκτονικής του μοντέλου. Αυτό επιτρέπει στο μοντέλο να συνδυάζει τοπική και παγκόσμια προσοχή σε διαφορετικά στάδια της επεξεργασίας.

Ο κωδικοποιητής της εικόνας παράγει το image embedding το οποίο έχει διαστάσεις κατά 16 φορές μικρότερες από την εικόνα εισόδου. Έτσι η εικόνα εισόδου έχει διαστάσεις 1024*1024*3, τυπική διάσταση για εικόνα υψηλής ανάλυσης, ενώ το image embedding που προκύπτει 64*64 εικονοστοιχεία. Για τη μείωση των καναλιών της εικόνας, χρησιμοποιείται συνέλιξη (1*1) με τελικό αριθμό καναλιών τα 256 και έπειτα συνέλιξη ξανά 3*3 διατηρώντας τον ίδιο αριθμό καναλιών. Το τελικό αποτέλεσμα είναι ένα image embedding διαστάσεων 64* 64 *256 .(Kirillov et al., 2023).

3.4.2 ΚΩΔΙΚΟΠΟΙΗΤΗΣ ΟΔΗΓΙΩΝ (PROMPT ENCODER)

Ο κωδικοποιητής οδηγιών-προτροπών (prompt encoder) είναι το τμήμα του μοντέλου που μετατρέπει τις οδηγίες (prompts) σε αναπαραστάσεις υψηλής διάστασης (embedding) ώστε να χρησιμοποιηθούν από το υπόλοιπο σύστημα.

Οι οδηγίες-προτροπές που παρέχονται στο μοντέλο για τη κατάτμηση της εικόνας (prompt segmentation) χωρίζονται σε δύο κύριες κατηγορίες, τις αραιές (sparse) και τις πυκνές (dense). Οι αραιές προτροπές που δέχεται το μοντέλο είναι της μορφής

σημείων, κουτιών και ελεύθερου κειμένου. Οι πυκνές προτροπές είναι μάσκες που απεικονίζουν αντικείμενα ή περιοχές ενδιαφέροντος.

Κάθε τύπος προτροπής ενσωματώνεται στην αναπαράσταση της εικόνας με διαφορετικό τρόπο ο οποίος θα παρουσιαστεί ανάλογα το είδος παρακάτω. Οι αραιές προτροπές κωδικοποιούνται σε διανυσματικές αναπαραστάσεις (vectorial embedding) διάστασης 256.

- **Κωδικοποίηση Προτροπής Σημείου (Point Prompt)**

Ένα σημείο αποτελεί μια προτροπή –οδηγία που αναφέρεται σε μια συγκεκριμένη θέση στην εικόνα. Η κωδικοποίηση θέσης (positional encoding) χρησιμοποιείται για την αναπαράσταση της θέσης του σημείου στην εικόνα και παρέχει πληροφορίες σχετικά με τις συντεταγμένες του σημείου (x,y). Επιπλέον υπάρχουν δύο αναπαραστάσεις (embeddings) οι οποίες υποδεικνύουν αν το σημείο είναι στο προσκήνιο (foreground) ή στο παρασκήνιο (background). Αυτές οι αναπαραστάσεις μαθαίνονται κατά τη διάρκεια εκπαίδευσης του μοντέλου. Η τελική αναπαράσταση της προτροπής σημείου προκύπτει από το άθροισμα της κωδικοποίησης θέσης με την αναπαράσταση που προκύπτει αν το σημείο ανήκει στο παρασκήνιο ή στο προσκήνιο.

- **Κωδικοποίηση Προτροπής Κουτιού (Box Prompt)**

Η αναπαράσταση μιας προτροπής κουτιού πραγματοποιείται από δυο βασικές αναπαραστάσεις. Η πρώτη αναπαράσταση (embedding) αφορά την κωδικοποίηση θέσης της αριστερής άνω γωνίας του κουτιού η οποία προστίθεται με την αναπαράσταση που συμβολίζει το χαρακτηρισμό «αριστερή άνω γωνία». Ομοίως για τη κάτω δεξιά γωνία του κουτιού εφαρμόζεται η ίδια διαδικασία με άθροισμα της κωδικοποίησης θέσης και της αναπαράστασης που συμβολίζει το χαρακτηρισμό «δεξιά κάτω γωνία». Αυτή η διαδικασία δημιουργεί μια πλήρη και ακριβή αναπαράσταση του κουτιού, η οποία περιλαμβάνει πληροφορίες τόσο για τις θέσεις των γωνιών όσο και για την έννοια της τοποθέτησης μέσα στην εικόνα.

- **Κωδικοποίηση Ελεύθερου Κειμένου (Text Prompt)**

Για την αναπαράσταση του ελεύθερου κειμένου χρησιμοποιείται ένας προ-εκπαιδευμένος κωδικοποιητής κειμένου (text encoder) από το μοντέλο CLIP. Το CLIP (Contrastive Language-Image Pretraining) είναι μοντέλο βαθιάς μάθησης που αναπτύχθηκε από την OpenAI το 2021 και συνδυάζει αναπαραστάσεις εικόνας και κειμένου σε κοινό χώρο, επιτρέποντας την κατανόηση και άμεση σύγκριση περιγραφών και εικόνας. (Radford et al., 2021). Το κείμενο μετατρέπεται σε αναπαράσταση (embedding) που περιέχει τις πληροφορίες του κειμένου μέσω του μοντέλου CLIP.

- **Κωδικοποίηση Πυκνών Προτροπών όπως Μάσκες**

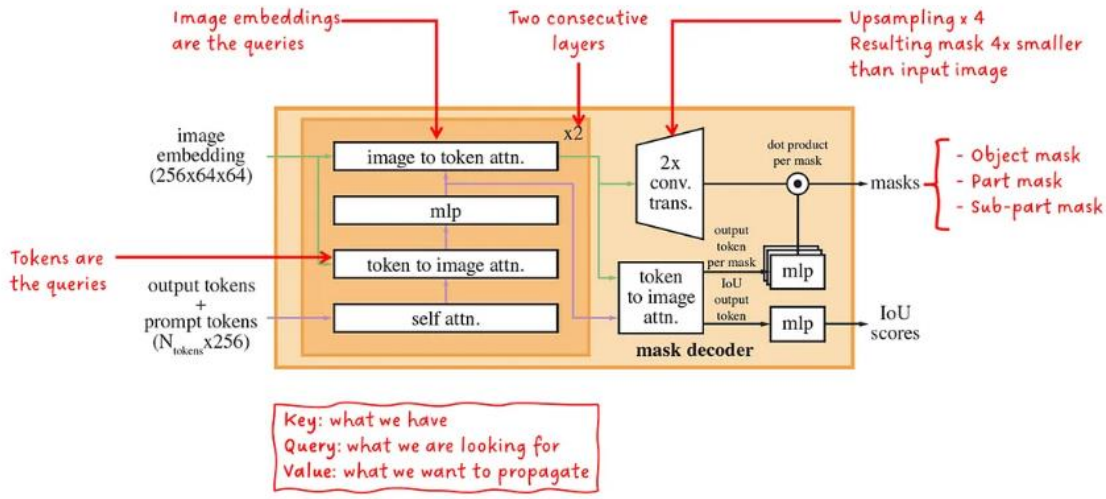
Οι μάσκες που χρησιμοποιούνται ως προτροπές έχουν χωρική αντιστοιχία με την εικόνα εισόδου, δηλαδή κάθε μέρος της μάσκας αντιστοιχεί σε ένα συγκεκριμένο μέρος της εικόνας. Οι μάσκες που εισάγονται έχουν 4 φορές χαμηλότερη ανάλυση από την ανάλυση της εικόνας. Στην συνέχεια, η μάσκα υποκλιμακώνεται περαιτέρω μέσω δύο συνελικτικών φίλτρων. Το πρώτο συνελιτκικό φίλτρο είναι διαστάσεων 2*2 με βήμα 2 (stride) και 4 κανάλια εξόδου. Το δεύτερο συνελικτικό φίλτρο έχει παρόμοιες διαστάσεις αλλά με 16 κανάλια εξόδου. Αυτά τα συνελικτικά επίπεδα έχουν ως στόχο τη μείωση των χωρικών διαστάσεων της μάσκας με ταυτόχρονη αύξηση του αριθμού των καναλιών. Έπειτα, ένα τελευταίο συνελικτικό επίπεδο διαστάσεων 1*1 χαρτογραφεί τη διάσταση των καναλιών από 16 σε 256. Κάθε συνελικτικό επίπεδο ακολουθείται από τη συνάρτηση ενεργοποίησης GELU (Gaussian Error Linear Unit) και από επίπεδα κανονικοποιήσης (layer normalization). Η GELU είναι μη γραμμική συνάρτηση ενεργοποίησης που χρησιμοποιείται στα νευρωνικά δίκτυα και σε αντίθεση με τις πιο παραδοσιακές συναρτήσεις ενεργοποίησης, παρέχει μια πιο ομαλή και συνεχή προσέγγιση. (Hendrycks and Gimpel, 2020). Τέλος η αναπαράσταση της μάσκας και της εικόνας προστίθενται. Αν δεν υπάρχει προτροπή μάσκας, προστίθεται μια αναπαράσταση (embedding) η οποία δηλώνει την απουσία της μάσκας σε κάθε θέση της αναπαράστασης της εικόνας.

3.4.3 ΑΠΟΚΩΔΙΚΟΠΟΙΗΤΗΣ ΜΑΣΚΩΝ (LIGHTWEIGHT MASK DECODER)

Η διαδικασία αποκωδικοποίησης μάσκας αναφέρεται στη μετατροπή των αναπαραστάσεων (embeddings) της εικόνας, των προτροπών καθώς και της αναπαράστασης ενός εξόδου token σε μια μάσκα. Ο αποκωδικοποιητής μάσκας (Mask Decoder) χρησιμοποιεί μια τροποποιημένη εκδοχή του μπλοκ αποκωδικοποίησης των Transformers (Vaswani et al., 2017) η οποία ακολουθείται από μια δυναμική κεφαλή πρόβλεψης μάσκας. Εισάγει τις ενσωματώσεις (embedding) εικόνας και τις ενσωματώσεις προτροπών, τις ενημερώνει μέσω των διαδικασιών της αυτο-προσοχής και της διασταυρούμενης προσοχής, και τελικά παράγει ένα σύνολο από μάσκες με τις αντίστοιχες βαθμολογίες τους, προσφέροντας μια ακριβή και αποτελεσματική πρόβλεψη της μάσκας. (Kirillov et al., 2023)

Ένα κρίσιμο στοιχείο αυτής της διαδικασίας είναι η εισαγωγή μιας αναπαράστασης που αποτελεί προϊόν μάθησης από το μοντέλο και αφορά το token εξόδου. (output token embedding). Η εισαγωγή της συγκεκριμένης αναπαράστασης (token εξόδου) στην αναπαράσταση της προτροπής πραγματοποιείται πριν την υποβολή της δεύτερης για επεξεργασία στον αποκωδικοποιητή. Η αναπαράσταση του token

εξόδου (output token εξόδου) παίζει καθοριστικό ρόλο στη λειτουργία του αποκωδικοποιητή αφού περιέχει αναγκαίες πληροφορίες που απαιτούνται για αντικείμενο της κατάτμησης της εικόνας. Αυτή η έννοια συναντάται και είναι παρόμοια με τα tokens κλάσης (class tokens) στους Vision Transformers για ταξινόμηση εικόνας. Στη ταξινόμηση εικόνας, τα tokens κλάσης είναι σημαντικά γιατί εμπεριέχουν πληροφορία για το σύνολο της εικόνας. Ομοίως στο συγκεκριμένο μοντέλο, η αναπαράσταση του token εξόδου (output token embedding) αποτελεί κρίσιμο παράγοντα που καθοδηγεί τη διαδικασία αποκωδικοποίησης προς την αποτελεσματική τμηματοποίηση της εικόνας.



Σχήμα 29 Αναλυτική Αρχιτεκτονική του Αποκωδικοποιητή Μάσκας (Πηγή : Kirillov et al., 2023)

Η αρχιτεκτονική του Αποκωδικοποιητή Μάσκας (Mask Decoder) απεικονίζεται στο Σχήμα 29. Κάθε επίπεδο του αποκωδικοποιητή εφαρμόζει 4 βήματα τα οποία είναι τα παρακάτω:

1. **Αυτό-προσοχή στα tokens** (*Self-attention on tokens*).

Στο πρώτο βήμα, εκτελείται αυτό-προσοχή στα tokens.Έτσι κάθε token ενημερώνεται λαμβάνοντας υπόψη όλα τα άλλα tokens, παρέχοντας μια ενιαία πληροφορία για την εικόνα.

2. **Διασταυρούμενη προσοχή από τα tokens στις αναπαραστάσεις εικόνας** (*Cross-attention from tokens to image embedding*)

Στο δεύτερο βήμα, τα tokens χρησιμοποιούνται ως ερωτήματα (queries) για την εκτέλεση διασταυρούμενης προσοχής (cross-attention) στην αναπαράσταση της εικόνας. Αυτό επιτρέπει στα tokens να ενσωματώσουν πληροφορίες από την αναπαράσταση της εικόνας, βελτιώνοντας την ακρίβεια των προβλέψεων.

3. **Ενημέρωση των tokens μέσω πολυεπίπεδου νευρωνικού δικτύου** (*MLP*)

Στο τρίτο βήμα, κάθε token ενημερώνεται με τη βοήθεια ενός πολυεπίπεδου νευρωνικού δικτύου (point-wise operation). Το MLP ενημερώνει τα tokens ατομικά, βελτιώνοντας τα ως προς την χωρητικότητα.

4. **Διασταυρούμενη προσοχή από τις αναπαραστάσεις της εικόνας προς τα tokens** (*Cross-attention from image embedding to tokens*)

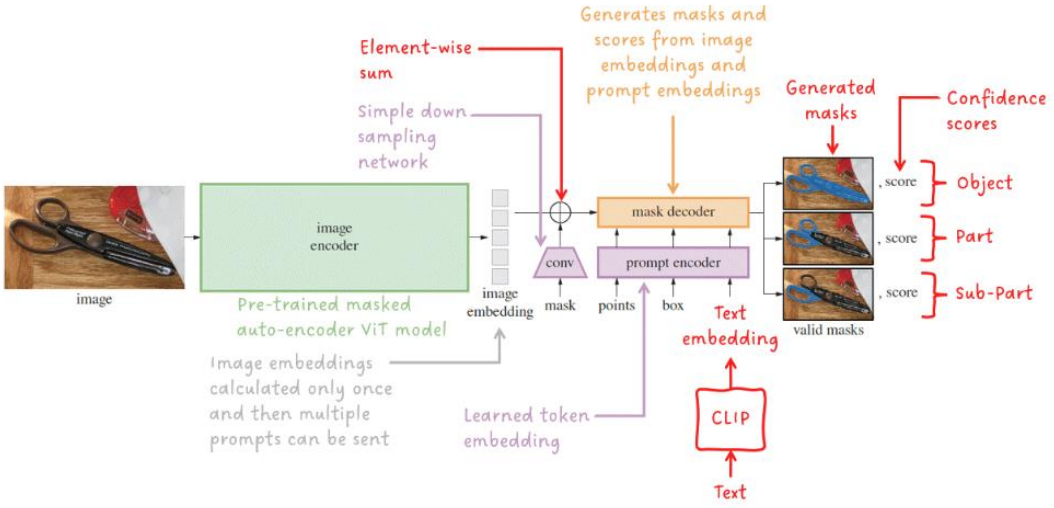
Στο τέταρτο και τελευταίο βήμα, η διασταυρούμενη προσοχή εκτελείται αντίστροφα σε σχέση με το δεύτερο βήμα, δηλαδή από τις αναπαραστάσεις της εικόνας που χρησιμοποιούνται ως ερωτήματα προς τα tokens. Το τελευταίο βήμα ενημερώνει την αναπαράσταση της εικόνας με πληροφορία από τις προτροπές-οδηγίες. (Kirillov et al., 2023).

Κατά τη διάρκεια της διασταυρούμενης προσοχής, η αναπαράσταση της εικόνας αντιμετωπίζεται ως σύνολο από 64^2 διανύσματα 256 διαστάσεων. Το επόμενο επίπεδο του αποκωδικοποιητή λαμβάνει τα ενημερωμένα tokens και τις ενημερωμένες αναπαραστάσεις της εικόνας από το προηγούμενο επίπεδο. Η διαδικασία χρησιμοποιεί δύο επίπεδα αποκωδικοποιητή.

Για να διασφαλιστεί ότι ο αποκωδικοποιητής έχει πρόσβαση σε σημαντικές γεωμετρικές πληροφορίες, στις αναπαραστάσεις της εικόνας (image embedding) προστίθενται κωδικοποιήσεις θέσης (positional encodings) κάθε φορά που συμμετέχουν σε επίπεδα προσοχής (attention layer). Αυτές οι κωδικοποιήσεις παρέχουν πληροφορίες σχετικά με τη γεωμετρική θέση των χαρακτηριστικών της εικόνας, επιτρέποντας στο μοντέλο να λαμβάνει υπόψη του τη χωρική διάταξη της εικόνας. Επιπλέον, οι αρχικές προτροπές (prompt tokens), συμπεριλαμβανομένων των θέσεων κωδικοποιήσεών τους, επαναπροστίθενται στα ενημερωμένα tokens κάθε φορά που αυτά συμμετέχουν σε ένα επίπεδο προσοχής. Όλες οι προαναφερόμενες διαδικασίες πραγματοποιούνται εντός του μπλοκ αποκωδικοποίησης του Transformer (Transformer Decoder Block).

Μετά την εκτέλεση του αποκωδικοποιητή, η διάσταση της ενημερωμένης αναπαράστασης της εικόνας αυξάνεται κατά 4 φορές (upsampling) με τη χρήση δύο αντίστροφων συνελικτικών επιπέδων. Έπειτα εφαρμόζεται ξανά διασταυρούμενη προσοχή από τα tokens στις αναπαραστάσεις της εικόνας και προκύπτουν ενημερωμένα tokens. Τα ενημερωμένα tokens εισάγονται σε ένα μικρό νευρωνικό δίκτυο με 3 επίπεδα (layers) το οποίο εξάγει ένα διάνυσμα που ταιριάζει με τη διάσταση των καναλιών της αναπαράστασης της εικόνας η οποία έχει αυξηθεί προηγουμένως. Τέλος, η πρόβλεψη της μάσκας πραγματοποιείται μέσω του γινόμενου των τιμών της αναπαράστασης της εικόνας και του αποτελέσματος του νευρωνικού δικτύου (3-layer MLP).

Στο Σχήμα 30 παρουσιάζεται η συνολική αρχιτεκτονική του Segment Anything μοντέλου όπως αυτή έχει αναλυθεί προηγουμένως.



Σχήμα 30 Συνολική Αρχιτεκτονική του Segment Anything μοντέλου (πηγή: Kirillov et al., 2023)

3.5 ΑΝΑΛΥΣΗ ΜΗΧΑΝΙΣΜΩΝ ΠΡΟΣΟΧΗΣ (SELF-ATTENTION, CROSS-ATTENTION)

Αφού αναλύθηκε η αρχιτεκτονική του μοντέλου Segment Anything, πρέπει να παρουσιαστούν οι μηχανισμοί προσοχής (attention mechanisms) οι οποίοι χρησιμοποιούνται εντός του μπλοκ αποκωδικοποίησης του Transformer (Transformer decoder block).

Η εισαγωγή των μηχανισμών προσοχής, ειδικά της αυτοπροσοχής (self-attention) και της διασταυρούμενης προσοχής (cross-attention) , επιτρέπει στους transformers να εστιάζουν επιλεκτικά στα σχετικά μέρη της εισερχόμενης ακολουθίας. Αυτή η προσέγγιση που βασίζεται στην προσοχή ενισχύει την ικανότητα του μοντέλου να ζυγίζει τη σημασία των διαφόρων στοιχείων, συμβάλλοντας σε πιο λεπτομερείς και πλούσιες σε πλαίσιο αναπαραστάσεις. (Vaswani et al., 2017)

3.5.1 ΑΝΑΛΥΣΗ ΜΗΧΑΝΙΣΜΟΥ ΑΥΤΟ- ΠΡΟΣΟΧΗΣ (SELF-ATTENTION)

Στον κωδικοποιητή (encoder) και στο αποκωδικοποιητή (decoder) των transformers, ο μηχανισμός της αυτό-προσοχής επιτρέπει στο μοντέλο να ζυγίζει τη σημασία κάθε στοιχείου της εισερχόμενης ακολουθίας σε σχέση με όλα τα άλλα στοιχεία. Αυτό διευκολύνει τη καταγραφή μακροχρόνιων εξαρτήσεων και τη κατανόηση εντός της ακολουθίας. (Vaswani et al., 2017).

Προσεγγίζοντας τον μηχανισμό της αυτό-μάθησης έστω η παρακάτω ακολουθία X.

$$X = \{x_1, x_2, \ldots, x_n\}$$

Ο μηχανισμός της αυτό- μάθησης υπολογίζει βαθμούς προσοχής (attention scores) για να αποφασίσει τη σημασία κάθε στοιχείου στην ακολουθία. Ο βαθμός προσοχής A_{ij} μεταξύ των στοιχείων i και j υπολογίζεται με το παρακάτω τύπο:

$$A_{ij} = \frac{\exp(Q_i * K_j)}{\sqrt{d}}$$

όπου Q_i, K_j, d συμβολίζουν τους όρους Ερώτημα (Query,Q) Κλειδί (Key, K) και Διάσταση (Dimension, D) των διανυσμάτων αντίστοιχα.

Ο όρος Ερώτημα (Query,Q) αναφέρεται στην αναπαράσταση του εισερχόμενου στοιχείου που αποτελεί προϊόν μάθησης του μοντέλου και υποδεικνύει τι πρέπει να αναζητήσει το μοντέλο στην ακολουθία. Αποτελεί τη βάση για την ανίχνευση σημαντικών πληροφοριών.

Ο όρος Κλειδί (Key, K) αναφέρεται στην αναπαράσταση του εισερχόμενου στοιχείου που αποτελεί προϊόν μάθησης του μοντέλου και υποδεικνύει τι πρέπει να συγκριθεί με το ερώτημα.

Ο όρος Αξία (Value, V) αντιπροσωπεύει τη πραγματική πληροφορία που σευδέεται με το εισερχόμενο στοιχείο.

Ο όρος Βαθμός Προσοχής (Attention Score – A) είναι το βάρος που αποδίδεται σε κάθε στοιχείο.

Τα πλεονεκτήματα που προσφέρει ο μηχανισμός της αυτό- μάθησης είναι η σύλληψη εξαρτήσεων μεγάλου εύρους εντός της ακολουθίας καθώς και η διαχείριση ακολουθιών μεταβλητού μήκους.

3.5.2 ΑΝΑΛΥΣΗ ΜΗΧΑΝΙΣΜΟΥ ΔΙΑΣΤΑΥΡΟΥΜΕΝΗΣ ΠΡΟΣΟΧΗΣ (CROSS-ATTENTION)

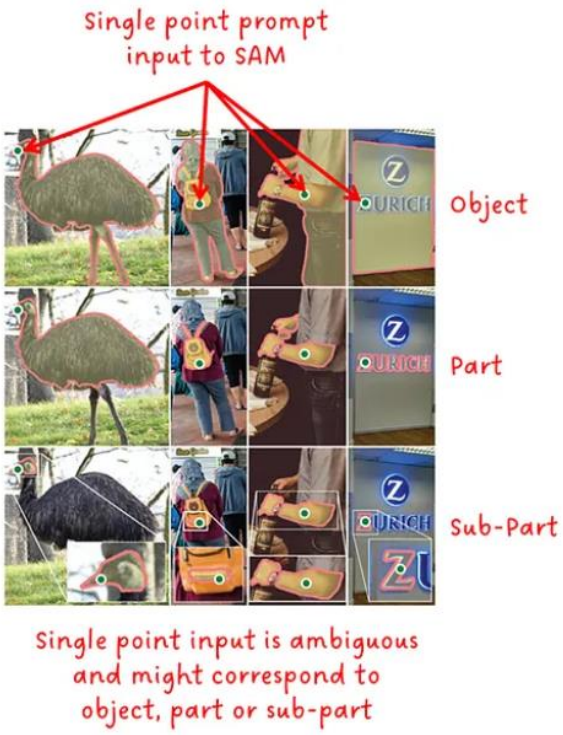
Ο μηχανισμός της διασταυρούμενης προσοχής εφαρμόζεται στους αποκωδικοποιητές των transformer. Επιτρέπει στο μοντέλο να συλλέξει πληροφορία από διαφορετικά σημεία της ακολουθίας εισόδου καθώς παράγει την ακολουθία εξόδου. Αντιθέτως με το μηχανισμό της αυτό- προσοχής ο οποίος επικεντρώνεται εντός της ίδιας ακολουθίας, ο μηχανισμός της διασταυρούμενης προσοχής επιτρέπει αλληλεπιδράσεις μεταξύ της εισερχόμενης και εξερχόμενης ακολουθίας.

Ο μηχανισμός της διασταυρούμενης προσοχής παίζει καθοριστικό ρόλο συνδέοντας τον αποκωδικοποιητή με την είσοδο του κωδικοποιητή. Επιτρέπει στο μοντέλο να

εστιάζει σε διαφορετικά μέρη της εισερχόμενης ακολουθίας ανάλογα με το τρέχον πλαίσιο στη διαδικασία αποκωδικοποίησης. Αυτό διευκολύνει την ενσωμάτωση σχετικών πληροφοριών από την εισερχόμενη ακολουθία στη δημιουργία κάθε στοιχείου της εξερχόμενης ακολουθίας.

3.6 ΔΙΑΧΕΙΡΙΣΗ ΤΗΣ ΑΣΑΦΕΙΑΣ (RESOLVING AMBIGUITY)

Σε μοντέλα που παράγουν ένα αποτέλεσμα, δηλαδή μια μάσκα, το μοντέλο θα επιλέξει μια μέση λύση από πολλαπλές έγκυρες μάσκες αν του δοθεί μια ασαφή προτροπή– οδηγία (prompt). Για την επίλυση του συγκεκριμένου προβλήματος, το μοντέλο παράγει πολλαπλές μάσκες από μια προτροπή. Συγκεκριμένα, για κάθε προτροπή παράγεται μια ομάδα τριών μασκών που αναφέρονται στο αντικείμενο, σε ένα μέρος του αντικείμενου και σε ένα υπό- μέρος του αντικειμένου. Μια ομάδα τριών μασκών είναι επαρκή για την αντιμετώπιση των περισσότερων συνθηκών. (Kirillov et al., 2023).



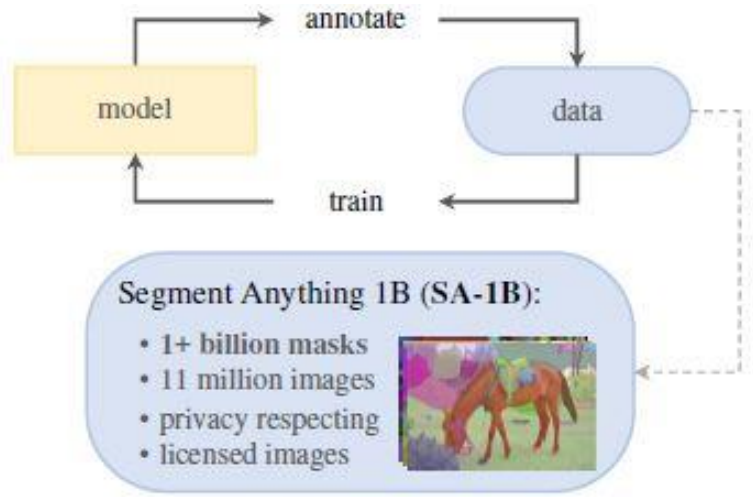
Σχήμα 31 Ομάδα τριών μασκών για επίλυση της ασάφειας από προτροπή ενός σημείου (πηγή: Kirillov et al., 2023)

Κατά την εκπαίδευση του μοντέλου, μόνο για την ελάχιστη τιμή loss μεταξύ των μασκών πραγματοποιείται η διαδικασία της backpropagation. Όπως φαίνεται στο Σχήμα 30, το μοντέλο μαζί με κάθε μάσκα προβλέπει και ένα σκορ εμπιστοσύνης (confidence score, IoU) (Kirillov et al., 2023)

3.7 ΜΗΧΑΝΗ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΟ SEGMENT ANYTHING (SEGMENT ANYTHING DATASET)

Για την επίτευξη μιας ισχυρής γενίκευσης σε κατανομές δεδομένων, διαπιστώθηκε η αναγκαιότητα της εκπαίδευσης του Segment Anything σε ένα ευρύ και ποικιλόμορφο σετ δεδομένων από μάσκες, μεγαλύτερο από οποιοδήποτε υπάρχον σύνολο δεδομένων τμηματοποίησης. Ενώ η τυπική προσέγγιση για τα θεμελιώδη μοντέλα (foundation models) είναι η απόκτηση δεδομένων από το διαδίκτυο, η ποσότητα διαθέσιμων μασκών είναι περιορισμένη με αποτέλεσμα την αναγκαιότητα ύπαρξης εναλλακτικής στρατηγικής.

Η λύση ήταν η δημιουργία μιας μηχανής δεδομένων (data engine) και η ανάπτυξη του μοντέλου παράλληλα με την χρήση του για την απόκτηση δεδομένων. (model in loop data annotation). (Kirillov et al., 2023). Η μηχανή δεδομένων χρησιμοποιήθηκε για τη συλλογή 1.1 δισεκατομμυρίων μασκών , το γνωστό σετ δεδομένων SA-1B.



Σχήμα 32 Χρήση του μοντέλου Segment Anything για δημιουργία δεδομένων (πηγή: Kirillov et al., 2023)

Η μηχανή δεδομένων έχει τρία στάδια για την απόκτηση των δεδομένων όπως παρακάτω:

1. **Υποβοηθούμενο – Χειροκίνητο Στάδιο (Assisted –Manual Stage).**

Στο πρώτο στάδιο, το μοντέλο βοηθά μια ομάδα επαγγελματιών στην επισήμανση μασκών, διαδικασία όμοια με μια κλασσική διαδραστική τμηματοποίηση. Οι επαγγελματίες επισημαίνουν μάσκες μέσω σημείων προσκηνίου και παρασκηνίου (foreground / background) χρησιμοποιώντας ένα εργαλείο διαδραστικής τμηματοποίησης το οποίο τροφοδοτείται από το Segment Anything. Μέσω διάδρασης, οι μάσκες επιδέχονται περαιτέρω βελτίωση από τους επαγγελματίες.

Επίσης η ομάδα που ασχολήθηκε με την επισήμανση των μασκών, όριζαν ετικέτα σε όσα αντικείμενα μπορούσαν να ονομάσουν ή να περιγράψουν.

Στην αρχή του πρώτου σταδίου, το SAM εκπαιδεύεται με κοινά δημόσια σετ δεδομένων τμηματοποίησης. Όσο πιο πολλές μάσκες συλλέγονται, ο κωδικοποιητής της εικόνας από Vision Transformer B αναβαθμίζεται σε Vision Transformer H και εξελίσσονται και άλλες αρχιτεκτονικές λεπτομέρειες. Συνολικά το μοντέλο επανακπαιδεύεται 6 φόρες. Όσο το μοντέλο βελτιώνεται, ο μέσος χρόνος επισήμανσης ανά μάσκα μειώνεται από 34 σε 14 δευτερόλεπτα. . Επιπλέον όσο το SAM βελτιώνεται ο μέσος αριθμός των μασκών ανά εικόνα από 20 αυξάνεται σε 44. Συνολικά κατά το πρώτο στάδιο συλλέχθηκαν 4.3 εκατομμύρια μάσκες από 120.000 εικόνες

2. **Ημι-αυτόματο Στάδιο (Semi-Automatic Stage)**.

Στο συγκεκριμένο στάδιο, στόχος είναι η αύξηση της ποικιλομορφίας των μασκών με σκοπό τη βελτίωση του μοντέλου στο να τμηματοποιεί τα πάντα μέσα στην εικόνα. Αρχικά, το μοντέλο παράγει αυτόματα μάσκες για ένα υποσύνολο αντικειμένων με υψηλό σκορ εμπιστοσύνης και με παροχή προτροπών-οδηγιών που αφορούν τοποθεσίες αντικειμένων. Οι επαγγελματίες εστιάζουν στην επισήμανση των υπόλοιπων μασκών , συμβάλλοντας στην αύξηση της ποικιλίας των μασκών. Κατά τη διάρκεια του συγκεκριμένου σταδίου συλλέγονται επιπλέον 5.9 εκατομμύρια μάσκες από 180.000 εικόνες. Όπως και στο πρώτο στάδιο, το μοντέλο επανακπαιδεύεται με τα νέα δεδομένα ακόμα 5 φορές. Ο μέσος αριθμός των μασκών ανά εικόνα από 44 αυξάνεται σε 72. Συνολικά και στα πρώτα δύο στάδια έχουν συλλεχθεί 10.2 εκατομμύρια μάσκες από 300.000 εικόνες.

3. **Πλήρως αυτόματο Στάδιο (Fully Automatic Stage)**.

Στο τελικό στάδιο, το μοντέλο δέχεται ως προτροπή –οδηγία ένα πλέγμα σημείων 32 * 32 που αναφέρονται ως σημεία προσκηνίου (foreground points). Για κάθε σημείο προβλέπεται μια ομάδα μασκών που αντιστοιχούν σε έγκυρα αντικείμενα. Το μοντέλο είναι ήδη σε ικανότητα να χειριστεί το θέμα της ασάφειας με αποτέλεσμα για σημεία που συμπίπτουν σε μέρη και υπό-μέρη αντικειμένων, να επιστρέφει μάσκες για το υπό-μέρος, μέρος και ολόκληρο το αντικείμενο. Οι διπλότυπες μάσκες αφαιρούνται από το σύνολο των μασκών. Για περαιτέρω βελτίωση της ποιότητας των μικρών μασκών, επεξεργάζονται πολλαπλές επικαλυπτόμενες περικοπές εικόνων σε μεγέθυνση. Έτσι δημιουργείται το τελικό σετ δεδομένων με 1.1 δισεκατομμύριο μάσκες από 11 εκατομμύρια εικόνες. Το τελικό σετ δεδομένων είναι γνωστό με το όνομα SA-1B.

Σχήμα 33 Αριθμός μασκών ανά εικόνα (πηγή: Kirillov et al., 2023)

3.8 SA -1B, ΤΟ ΣΕΤ ΔΕΔΟΜΕΝΩΝ ΤΟΥ SEGMENT ANYTHING

Το σύνολο δεδομένων του Segment Anything, το SA-1B, αποτελείται από 11 εκατομμύρια ποικίλες, υψηλής ανάλυσης, αδειοδοτημένες και προστατευμένες εικόνες καθώς και από 1.1 δισεκατομμύρια υψηλής ποιότητας μάσκες τμηματοποίησης, οι οποίες συλλέχθηκαν με τη χρήση της μηχανής δεδομένων που παρουσιάστηκε στη προηγούμενη ενότητα.

Οι εικόνες του SA -1B είναι υψηλής ανάλυσης (κατά μέσο όρο 3300x4950 εικονοστοιχεία), με αποτέλεσμα το μέγεθος των δεδομένων να παρουσιάσει προκλήσεις όσον αφορά την προσβασιμότητα και την αποθήκευση του. Για το λόγο αυτό, οι εικόνες που δημοσιοποιούνται έχουν υποστεί υποδειγματοληψία (downsampling) με τη μικρότερη πλευρά να ισούται με 1500 εικονοστοιχεία. Ακόμη και μετά την υποδειγματοληψία, οι εικόνες είναι σημαντικά υψηλότερης ανάλυσης από πολλά υπάρχοντα σύνολα δεδομένων υπολογιστικής όρασης (π.χ., οι εικόνες του COCO σετ δεδομένων είναι περίπου 480x640 εικονοστοιχεία).

Η μηχανή δεδομένων παρήγαγε 1,1 δισεκατομμύρια μάσκες, το 99,1% των οποίων δημιουργήθηκε πλήρως αυτόματα. Συμπεραίνεται ύστερα από σειρά πειραμάτων,

ότι οι αυτόματες μάσκες μας είναι υψηλής ποιότητας και αποτελεσματικές για την εκπαίδευση μοντέλων.

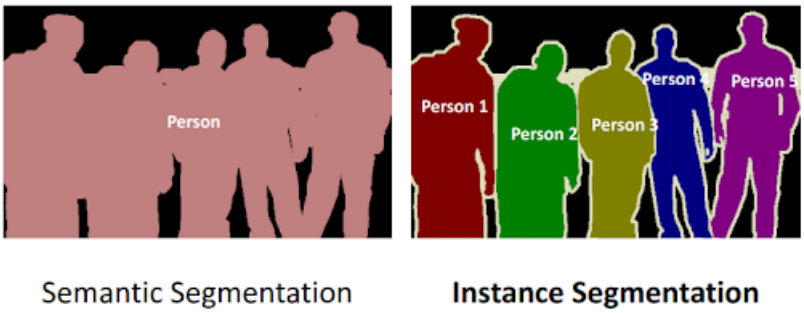
ΚΕΦΑΛΑΙΟ 4: ΠΑΡΟΥΣΙΑΣΗ ΜΕΘΟΔΟΛΟΓΙΩΝ ΕΦΑΡΜΟΓΗΣ SEGMENT ANYTHING ΜΟΝΤΕΛΟΥ

Στο κεφάλαιο αυτό, επιχειρείται η παρουσίαση διαφορετικών μεθοδολογιών τμηματοποίησης δορυφορικών απεικονίσεων και εξαγωγή οντοτήτων από αυτές μέσω του Segment Anything μοντέλου που παρουσιάστηκε στη προηγούμενη ενότητα. Συγκεκριμένα θα εφαρμοστεί η μεθοδολογία της αυτόματης παραγωγής μασκών στις εικόνες εισόδου (automatic mask generation) καθώς και μεθοδολογίες κατάτμησης απεικονίσεων μέσω γεωμετρικών προτροπών (prompts). Επιπλέον, επιχειρείται η ενσωμάτωση του Segment Anything σε ροή εργασίας (pipeline) σε συνδυασμό με την εφαρμογή ενός μοντέλου ανίχνευσης αντικειμένων όπως το Grounding Dino. Τέλος, θα παρουσιαστεί μια αυτόνομη εφαρμογή υλοποιημένη προγραμματιστικά στα πλαίσια της παρούσας διπλωματικής εργασίας που ενσωματώνει όλες τις δυνατότητες που παρέχει το Segment Anything μοντέλο. Η συγκεκριμένη εφαρμογή έχει υλοποιηθεί με σκοπό τη δημιουργία ενός αυτόνομου γεωχωρικού εργαλείου που δύναται να τμηματοποιεί και να εξάγει οντότητες από τηλεπισκοπικές απεικονίσεις.

4.1 ΜΕΘΟΔΟΛΟΓΙΑ ΑΥΤΟΜΑΤΗΣ ΠΑΡΑΓΩΓΗΣ ΜΑΣΚΩΝ ΜΕΣΩ SEGMENT ANYTHING

Στη παρούσα μεθοδολογία παρουσιάζεται η κατάτμηση δορυφορικών απεικονίσεων, μέσω μιας τεχνικής που χρησιμοποιήθηκε και στη παραγωγή του σετ δεδομένων SA-1B. Η τεχνική που περιγράφεται χρησιμοποιείται για την αυτόματη παραγωγή μασκών σε όλο το εύρος της εικόνας χωρίς κάποια προτροπή –οδηγία από τον εκάστοτε χρήστη. Στη πραγματικότητα, δημιουργείται ένα πλέγμα σημείων διαστάσεων 32 *32, τα οποία το μοντέλο Segment Anything χρησιμοποιεί ως σημειακές προτροπές. Τα σημεία χρησιμοποιούνται για κατάτμηση αντικειμένων στο προσκήνιο της εικόνας. Για κάθε σημείο του πλέγματος, το μοντέλο προβλέπει πολλαπλές μάσκες, οι οποίες φιλτράρονται για την ποιότητα τους και όσες είναι διπλότυπες αφαιρούνται. Το αποτέλεσμα είναι η εξαγωγή όλων των αντικειμένων εντός απεικόνισης ως ξεχωριστές οντότητες. Η συγκεκριμένη τεχνική είναι μια παραλλαγή της instance segmentation, δηλαδή τμηματοποίησης οντοτήτων

(instances). Στη φυσική της μορφή, η τμηματοποίηση οντοτήτων τμηματοποιεί την εικόνα σε ξεχωριστά αντικείμενα και όχι σε ξεχωριστές κλάσεις όπως πράττει η σημασιολογική κατάτμηση (semantic segmentation). Η διαφορά της όμως με τη παρούσα μεθοδολογία είναι το μοντέλο Segment Anything μέσω της αυτόματης παραγωγής μασκών τμηματοποιεί την εικόνα σε ξεχωριστές οντότητες αλλά χωρίς την απόδοση αντίστοιχων ετικετών (labels) όπως κάνει η τμηματοποίηση οντοτήτων (instance segmentation).



Σχήμα 34 Έννοια της σημασιολογικής κατάτμησης και κατάτμησης οντοτήτων (πηγή: https://blog.roboflow.com/difference-semantic-segmentation-instance-segmentation/)

4.1.1 ΠΡΟΕΤΟΙΜΑΣΙΑ ΠΕΡΙΒΑΛΛΟΝΤΟΣ ΕΡΓΑΣΙΑΣ ΚΑΙ ΔΙΑΘΕΣΙΜΟΤΗΤΑ ΔΕΔΟΜΕΝΩΝ

Για όλες τις μεθοδολογίες που θα ακολουθήσουν θα χρησιμοποιηθεί μια ομάδα δορυφορικών απεικονίσεων τριών καναλιών (Κόκκινο, Πράσινο, Μπλε, RGB) οι οποίες προέρχονται από ανοιχτά δεδομένα (open data) και συγκεκριμένα από το δορυφορικό υπόβαθρο της Google. Οι διαστάσεις τους ποικίλουν σε μήκος και πλάτος αλλά ο αριθμός καναλιών παραμένει κοινός. Επίσης η χωρική διακριτικότητα των απεικονίσεων δεν είναι ίδια και θα αναφέρεται. Αξίζει να επισημανθεί ότι οι δορυφορικές απεικονίσεις είναι γεωαναφερμένες και τα αποτελέσματα της κατάτμησης που θα προκύψουν δύναται να εισαχθούν σε Γεωγραφικά Συστήματα Πληροφοριών για περαιτέρω επεξεργασία.

Παρακάτω παρουσιάζονται οι δορυφορικές απεικονίσεις που θα χρησιμοποιηθούν.

Σχήμα 35 Δορυφορικές απεικονίσεις ως εικόνες εισόδου του Segment Anything (πηγή : Δορυφορικό Υπόβαθρο Goggle)

Όλες οι διαδικασίες που θα περιγραφούν έχουν υλοποιηθεί σε γλώσσα προγραμματισμού Python σε περιβάλλον Conda. Επισημαίνεται ότι έχει χρησιμοποιηθεί κώδικας από δύο αποθετήρια του Github, το επίσημο αποθετήριο της MetaAi για το μοντέλο segment anything (https://github.com/facebookresearch/segment-anything) καθώς και το αποθετήριο της βιβλιοθήκης samgeo που εφαρμόζει το μοντέλο Segment Anything σε γεωχωρικά δεδομένα όπως τηλεπισκοπικές απεικονίσεις. (https://github.com/opengeos/segment-geospatial). Το περιβάλλον στο οποίο έχουν υλοποιηθεί οι μεθοδολογίες εφαρμογής του Segment Anything περιλαμβάνουν και πολλές άλλες βιβλιοθήκες οι οποίες είναι απαραίτητες για τη διαχείριση , επεξεργασία και παρουσίαση των αποτελεσμάτων.

4.1.2 ΕΦΑΡΜΟΓΗ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ ΑΥΤΟΜΑΤΗΣ ΠΑΡΑΓΩΓΗΣ ΜΑΣΚΩΝ ΧΩΡΙΣ ΤΡΟΠΟΠΟΙΗΣΗ ΠΑΡΑΜΕΤΡΩΝ

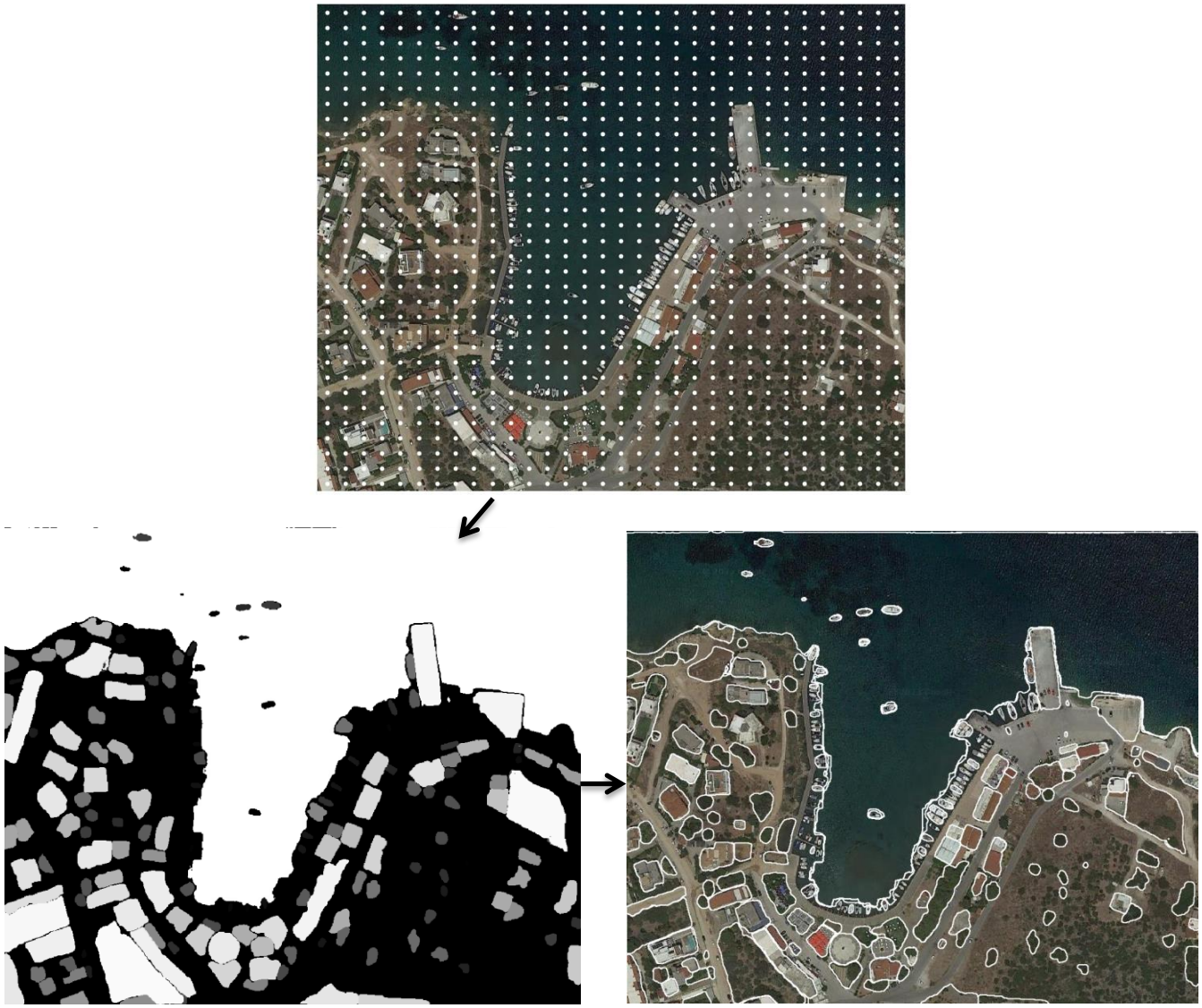
Το μοντέλο Segment Anything για τη παρούσα ενότητα θα χρησιμοποιήσει τα προ-εκπαιδευμένα βάρη, δηλαδή τις παραμέτρους του μοντέλου που υπολογίστηκαν κατά την εκπαίδευση του με το SA-1B σετ δεδομένων. Σε επόμενη ενότητα, θα παρουσιαστεί ο τρόπος περεταίρω εκπαίδευσης του μοντέλου για πιο ειδικευμένους σκοπούς. Τα βάρη του μοντέλου αφορούν και τα τρία κύρια μέρη του μοντέλου όπως αυτά έχουν παρουσιαστεί προηγουμένως, τον κωδικοποιητή εικόνας, τον κωδικοποιητή των προτροπών και τον αποκωδικοποιητή της μάσκας.

Επιπλέον ορίζεται ο Vision Transformer που θα χρησιμοποιήσει το μοντέλο, ο οποίος είναι ο ViT –H. Το πρώτο βήμα της διαδικασίας αφορούσε τον καθορισμό των βαρών και του Vision Transformer. Έπειτα ακολουθεί η εισαγωγή της εικόνα εισόδου στο μοντέλο μέσω του κωδικοποιητή εικόνας (image encoder) και ο υπολογισμός των αναπαραστάσεων της εικόνας. (image embeddings) .

Δεύτερο βήμα είναι η εισαγωγή του πλέγματος σημείων (32 *32) ως prompts μέσω του κωδικοποιητή προτροπών (prompt encoder) στο μοντέλο για να μετατραπούν σε αντίστοιχες αναπαραστάσεις . Στο τέλος για κάθε σημείο του πλέγματος προβλέπονται πολλαπλές μάσκες όπου αυτές με το μεγαλύτερο σκορ εμπιστοσύνης παρουσιάζονται ως τελικό αποτέλεσμα.

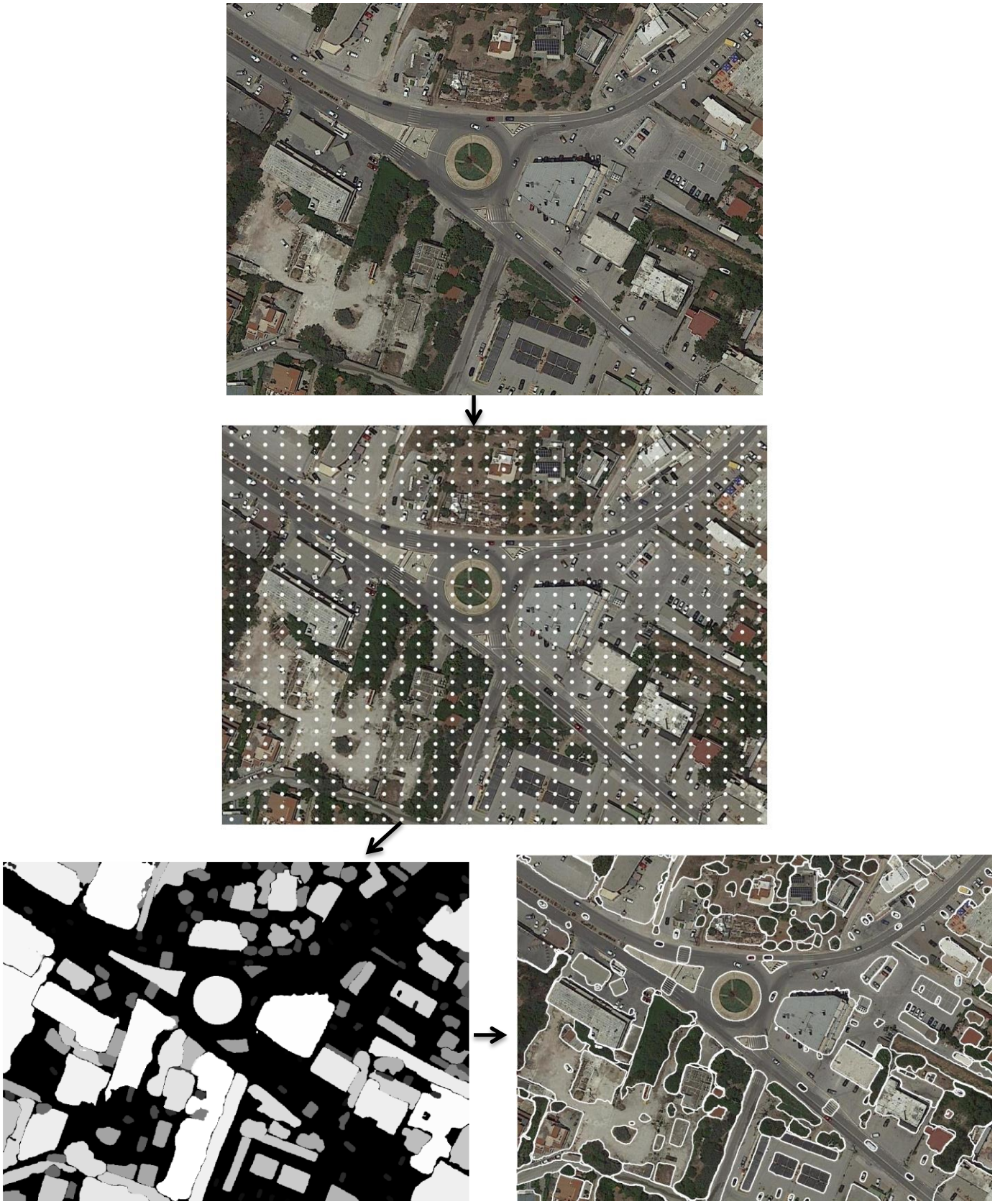
Σχήμα 36 Διαδικασία τμηματοποίησης εικόνας με αυτόματη παραγωγή μασκών. Άνω αριστερά η αρχική εικόνα που εισάγεται στον κωδικοποιητή εικόνας. Άνω δεξιά, πλέγμα σημείων (32*32) ως προτροπή –οδηγία το οποίο εισάγεται στον κωδικοποιητή οδηγιών. Κάτω αριστερά τελικό αποτέλεσμα τμηματοποίησης σε μορφή εικόνας. Κάτω δεξιά μετατροπή της εικόνας των μασκών σε διανυσματικό αρχείο

Στο Σχήμα 36 απεικονίζεται η διαδικασία της αυτόματης τμηματοποίησης μιας εκ των δορυφορικών απεικονίσεων του σετ δεδομένων. Η αρχική εικόνα έχει διαστάσεις (708, 743, 3) όπου 3 είναι ο αριθμός των καναλιών. Ο συνολικός αριθμός των μασκών που παράγεται ισούται με 496. Το τελικό αποτέλεσμα είναι ένα raster αρχείο με διαστάσεις (708,743). Όσο πιο ανοιχτόχρωμη είναι μια μάσκα, τόσο μεγαλύτερο το εμβαδόν της. Σε δεύτερο χρόνο, το συγκεκριμένο αρχείο μετατρέπεται προγραμματιστικά σε διανυσματικό αρχείο (πολύγωνα με λευκό περίγραμμα). Ομοίως και στις άλλες τηλεπισκοπικές απεικονίσεις εκτελέσθηκε η ίδια διαδικασία και παρακάτω παρουσιάζονται τα αποτελέσματα.



Σχήμα 37 Διαδικασία τμηματοποίησης με αυτόματη παραγωγή μασκών

Η εικόνα εισόδου έχει διαστάσεις (756, 913, 3) και ο συνολικός αριθμός μασκών που προβλέπεται ισούται με 216. Ομοίως στο τέλος, δημιουργήθηκε διανυσματικό αρχείο με τις μάσκες να εμφανίζονται ως πολύγωνα με λευκό περίγραμμα.



Εικόνα 38 Διαδικασία τμηματοποίησης με αυτόματη παραγωγή μασκών

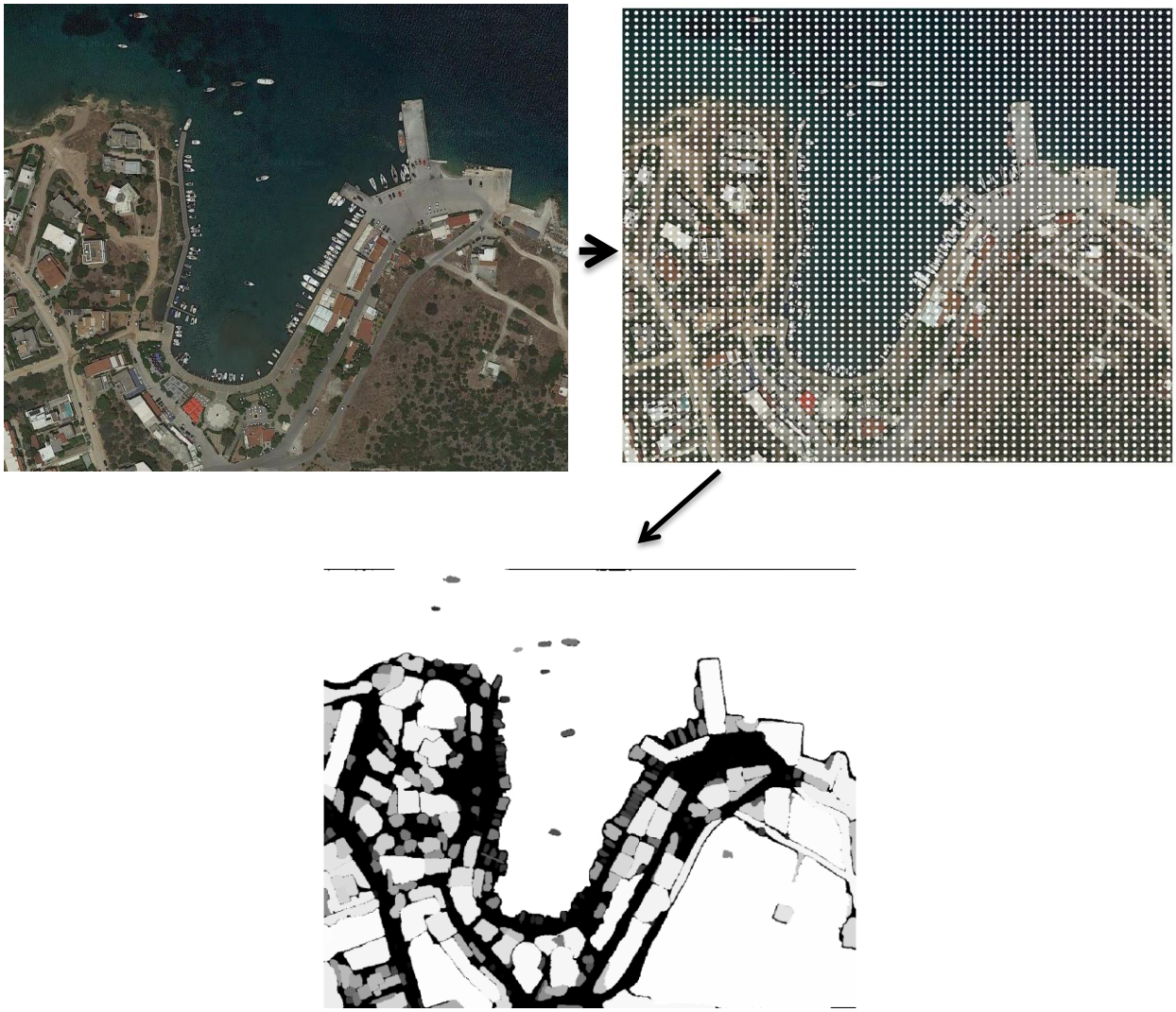
Η εικόνα στη τρίτη περίπτωση έχει διαστάσεις (709, 969,3) ενώ ο συνολικός αριθμός μασκών που προβλέπονται είναι 296. Ομοίως στο τέλος, δημιουργήθηκε

διανυσματικό αρχείο με τις μάσκες να εμφανίζονται ως πολύγωνα με λευκό περίγραμμα.

4.1.3 ΠΕΙΡΑΜΑΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΣΕ ΕΦΑΡΜΟΓΗ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ ΑΥΤΟΜΑΤΗΣ ΠΑΡΑΓΩΓΗΣ ΜΑΣΚΩΝ (ΤΡΟΠΟΠΟΙΗΣΗ ΠΑΡΑΜΕΤΡΩΝ)

Στη προηγούμενη υποενότητα, κατά την εκτέλεση της αυτόματης παραγωγής μασκών από τηλεπισκοπικές απεικονίσεις μέσω του μοντέλου Segment Anything, δε τροποποιήθηκε κάποια παράμετρος του μοντέλου. Επισημαίνεται ότι υπάρχουν αρκετές παράμετροι στην συγκεκριμένη μεθοδολογία που χρήζουν παραμετροποίηση και ελέγχουν την πυκνότητα του πλέγματος αλλά και την ποιότητα των μασκών που προβλέπονται μεταξύ άλλων.

Αρχικά η πυκνότητα του πλέγματος ρυθμίζεται από τον αριθμό των σημείων του πλέγματος στην εικόνα. Η προεπιλεγμένη τιμή που χρησιμοποιήθηκε και προηγουμένως είναι 32 (αριθμός σημείων ανά πλευρά εικόνας). Η τιμή που επιλέχθηκε για τη παρούσα πειραματική εκτέλεση της μεθοδολογίας είναι 64, άρα το πλέγμα θα έχει διαστάσεις 64*64. Επιπλέον η ποιότητα των μασκών που προβλέπονται από το μοντέλο ελέγχεται από ένα κατώφλι που λαμβάνει τιμές από 0 έως 1. Η προεπιλεγμένη τιμή για το συγκεκριμένο κατώφλι (pred_iou_thresh) ισούται με 0.88. Η τιμή που θα χρησιμοποιηθεί τώρα θα ισούται με 0.86. Αυτό σημαίνει ότι το κατώφλι είναι οριακά πιο ανεκτό με χαμηλότερης ποιότητας μάσκες. Η ροή εργασίας παραμένει ίδια με τη προηγούμενη υποενότητα. Παρακάτω παρατίθενται τα αποτελέσματα της αυτόματης τμηματοποίησης και πραγματοποιείται οπτικά σύγκριση με την τμηματοποίηση με τις προεπιλεγμένες τιμές.

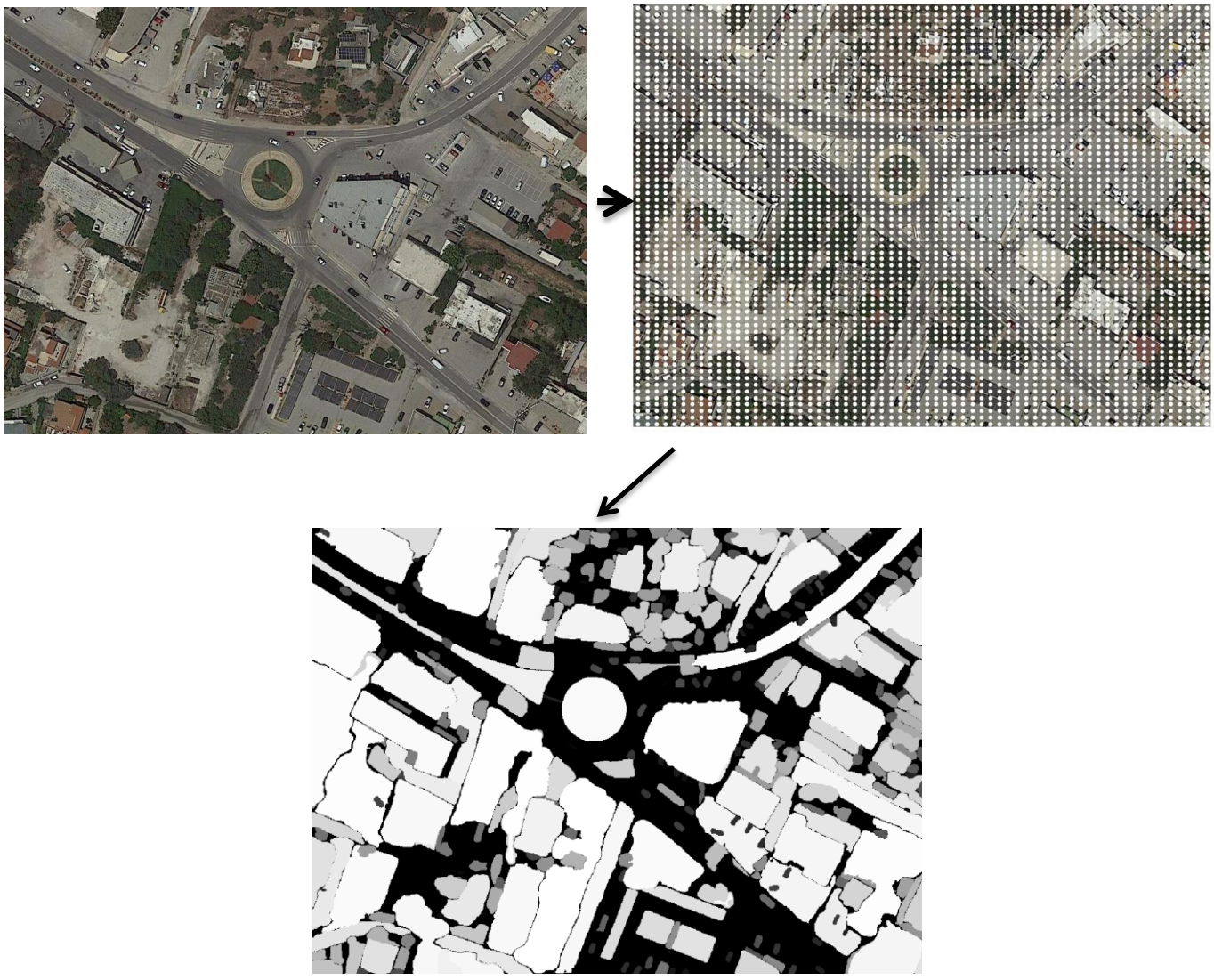


Σχήμα 39 Αριστερά η εικόνα εισόδου στο μοντέλο. Δεξιά το πυκνότερο πλέγμα (64 *64). Κάτω το αποτέλεσμα της αυτόματης κατάτμησης με τροποποιημένες παραμέτρους (πυκνότητα πλέγματος και κατώφλι ποιότητας μάσκας εξόδου)

Σχήμα 40 Αριστερά αποτέλεσμα τμηματοποίησης με τροποποιημένες παραμέτρους. Δεξιά αποτέλεσμα τμηματοποίησης με προεπιλεγμένες τιμές παραμέτρων

Ο αριθμός των μασκών με τις τροποποιημένες παραμέτρους ισούται με 728 ενώ με τις προεπιλεγμένες τιμές 216. Ο αυξημένος αριθμός προβλεπόμενων μασκών από το μοντέλο Segment Anything στη περίπτωση της παραμετροποίησης είναι λογικός για δύο λόγους. Αρχικά το μικρότερο κατώφλι σημαίνει ουσιαστικά και μεγαλύτερη ανεκτικότητα στην ποιότητα των μασκών. Αν το κατώφλι αυξανόταν, δηλαδή για τιμές μεγαλύτερες του 0.88 (προεπιλεγμένη τιμή) ο αριθμός των μασκών θα μειωνόταν. Ο δεύτερος λόγος είναι το πυκνότερο πλέγμα σημείων. Με περισσότερα σημεία κατανεμημένα ομοιόμορφα στην εικόνα , υπάρχουν περισσότερες σημειακές προτροπές για πρόβλεψη μασκών στην εικόνα.

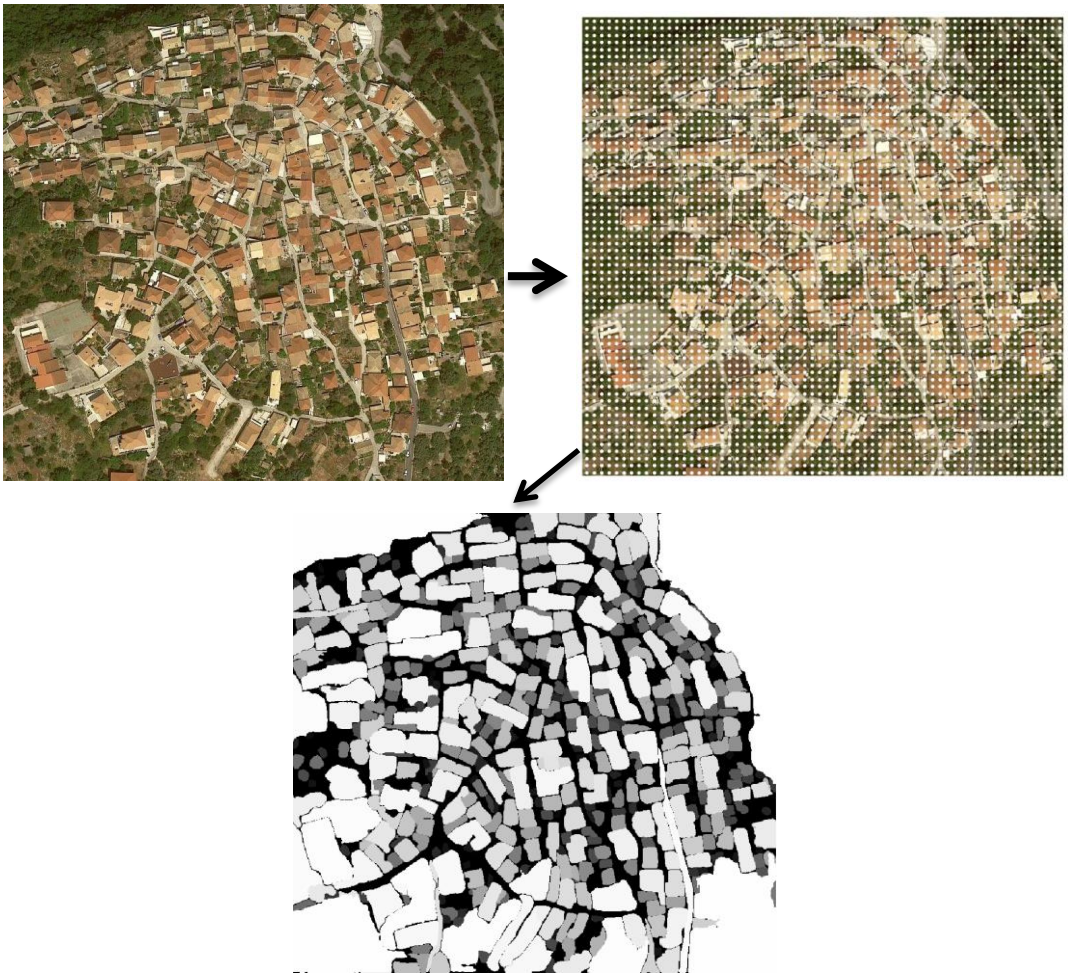
Ομοίως πραγματοποιείται η διαδικασία για τις άλλες δύο δορυφορικές απεικονίσεις.



Σχήμα 41 Αριστερά η εικόνα εισόδου στο μοντέλο. Δεξιά το πυκνότερο πλέγμα (64 *64). Κάτω το αποτέλεσμα της αυτόματης κατάτμησης με τροποποιημένες παραμέτρους (πυκνότητα πλέγματος και κατώφλι ποιότητας μάσκας εξόδου)

Σχήμα 42 Αριστερά αποτέλεσμα τμηματοποίησης με τροποποιημένες παραμέτρους. Δεξιά αποτέλεσμα τμηματοποίησης με προεπιλεγμένες τιμές παραμέτρων

Ο αριθμός των μασκών με τις τροποποιημένες παραμέτρους ισούται με 798 ενώ με τις προεπιλεγμένες τιμές 296. Ομοίως με το προηγούμενο παράδειγμα, ο αριθμός των μασκών είναι αυξημένος για τους ίδιους λόγους.



Σχήμα 43 Αριστερά η εικόνα εισόδου στο μοντέλο. Δεξιά το πυκνότερο πλέγμα (64 *64). Κάτω το αποτέλεσμα της αυτόματης κατάτμησης με τροποποιημένες παραμέτρους (πυκνότητα πλέγματος και κατώφλι ποιότητας μάσκας εξόδου)

Σχήμα 44 Αριστερά αποτέλεσμα τμηματοποίησης με τροποποιημένες παραμέτρους. Δεξιά αποτέλεσμα τμηματοποίησης με προεπιλεγμένες τιμές παραμέτρων

Ο αριθμός των μασκών με τις τροποποιημένες παραμέτρους ισούται με 1271 ενώ με τις προεπιλεγμένες τιμές 496. Ομοίως με το προηγούμενο παράδειγμα, ο αριθμός των μασκών είναι αυξημένος για τους ίδιους λόγους.

4.2 ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΤΑΤΜΗΣΗΣ ΔΟΡΥΦΟΡΙΚΩΝ ΕΙΚΟΝΩΝ ΚΑΙ ΕΞΑΓΩΓΗ ΟΝΤΟΤΗΤΩΝ ΜΕ ΧΡΗΣΗ ΓΕΩΜΕΤΡΙΚΩΝ ΠΡΟΤΡΟΠΩΝ

Στη παρούσα ενότητα, θα πραγματοποιηθεί τμηματοποίηση δορυφορικών εικόνων με χρήση γεωμετρικών προτροπών όπως σημειακά και πολυγωνικά δεδομένα. Η διαδικασία που ακολουθείται είναι παρόμοια, όμως αντί για πλέγμα σημείων που χρησιμοποιείται στην αυτόματη παραγωγή μασκών, θα εισαχθούν στο μοντέλο μέσω του κωδικοποιητή προτροπών συγκεκριμένα διανυσματικά δεδομένα. Οι απεικονίσεις που θα εισαχθούν στο μοντέλο είναι ίδιες με τη προηγούμενη ενότητα.

4.2.1 ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΤΑΤΜΗΣΗΣ ΔΟΡΥΦΟΡΙΚΩΝ ΕΙΚΟΝΩΝ ΚΑΙ ΕΞΑΓΩΓΗ ΚΤΙΡΙΑΚΩΝ ΟΝΤΟΤΗΤΩΝ ΜΕ ΧΡΗΣΗ ΠΟΛΥΓΩΝΩΝ ΩΣ ΠΡΟΤΡΟΠΩΝ

Η ροή εργασίας που θα ακολουθηθεί είναι σχεδόν παρόμοια με την αυτόματη παραγωγή μασκών. Αρχικά οι παράμετροι του μοντέλου που θα χρησιμοποιηθούν έχουν διαμορφωθεί από την ήδη υπάρχουσα εκπαίδευση του μοντέλου και ορίζονται κατά τον αρχικό ορισμό του μοντέλου. Ταυτόχρονα ορίζεται και ο Vision Transformer του κωδικοποιητή εικόνας (image encoder) ο οποίος είναι ίδιος με

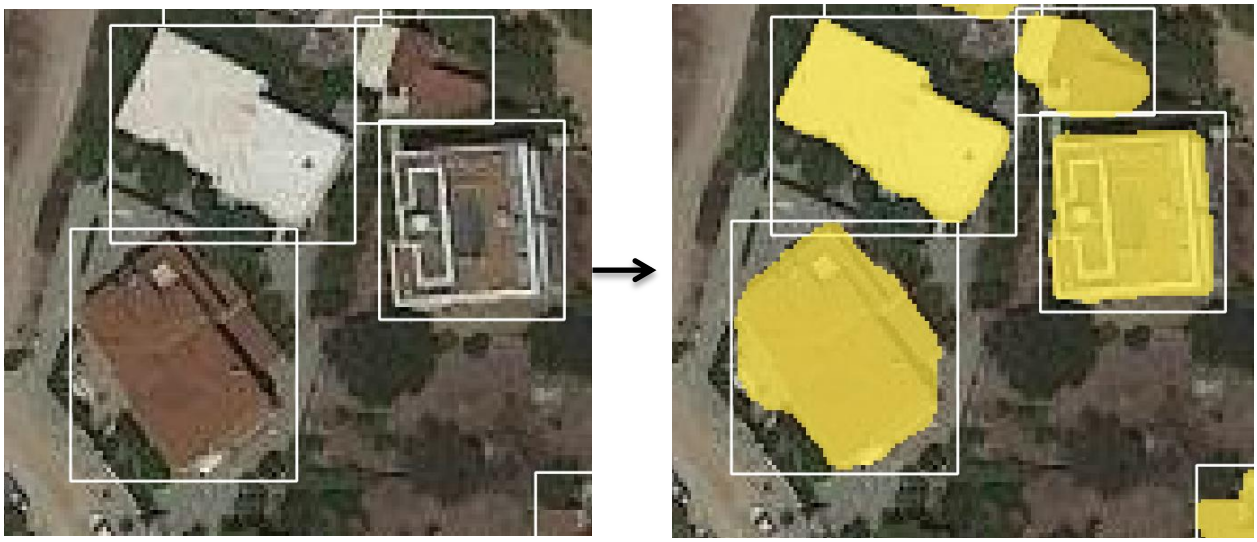
προηγουμένως (ViT –H) .Μια βασική διαφορά στο προγραμματιστικό κομμάτι, είναι η απενεργοποίηση της κλάσης που ευθύνεται για την αυτόματη παραγωγή μασκών «*SamAutomaticMaskGenerator*» και ενεργοποίηση της κλάσης «*SamPredictor*» που διαχειρίζεται την πρόβλεψη μασκών με εισαγωγή από τον χρήστη προτροπών.

Η εικόνα εισόδου εισάγεται στον κωδικοποιητή εικόνας και μετατρέπεται σε αναπαραστάσεις εικόνας (image embedding), ενώ τα πολύγωνα που θα χρησιμοποιηθούν ω προτροπές – οδηγίες στο μοντέλο, εισάγονται στο κωδικοποιητή προτροπών (prompt encoder) για να μετατραπούν και αυτά σε αντίστοιχες αναπαραστάσεις. Τέλος, μέσω του αποκωδικοποιητή μασκών (mask decoder) γίνεται η τελική πρόβλεψη των μασκών.

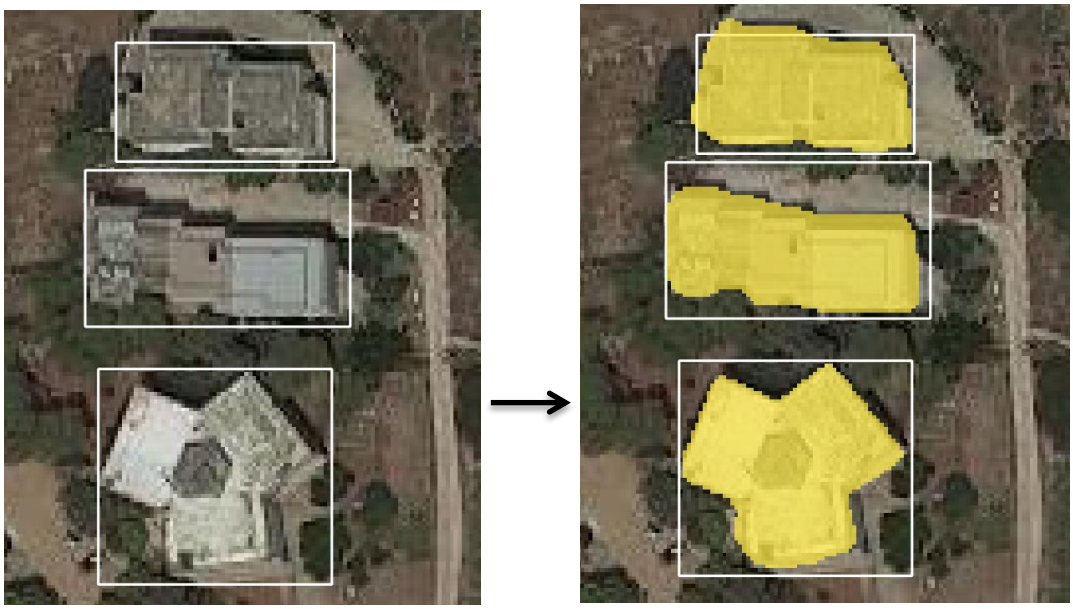
Τα πολύγωνα που θα χρησιμοποιηθούν είναι διανυσματικά δεδομένα μορφότυπου shapefile και αφορούν περιγράμματα κτιρίων στη περιοχή ενδιαφέροντος. Έχουν παραχθεί με διαδικασία ψηφιοποίησης σε περιβάλλον Γεωγραφικών Συστημάτων Πληροφοριών. Σκοπός είναι η πιστή εξαγωγή των μασκών των κτιριακών υποδομών της περιοχής ενδιαφέροντος. Συνολικά χρησιμοποιήθηκαν 33 πολύγωνα και εξάχθηκαν 33 αντίστοιχες μάσκες κτιρίων.

Σχήμα 45 Εικόνα εισόδου και πολυγωνικές προτροπές για τμηματοποίηση εικόνας

Σχήμα 46 Αποτέλεσμα τμηματοποίησης εικόνας με χρήση πολυγωνικών προτροπών.



Σχήμα 47 Παράδειγμα 1 εξαγωγής μάσκας κτιρίων με χρήση πολυγωνικών προτροπών



Σχήμα 48 Παράδειγμα 2 εξαγωγής μάσκας κτιρίων με χρήση πολυγωνικών προτροπών

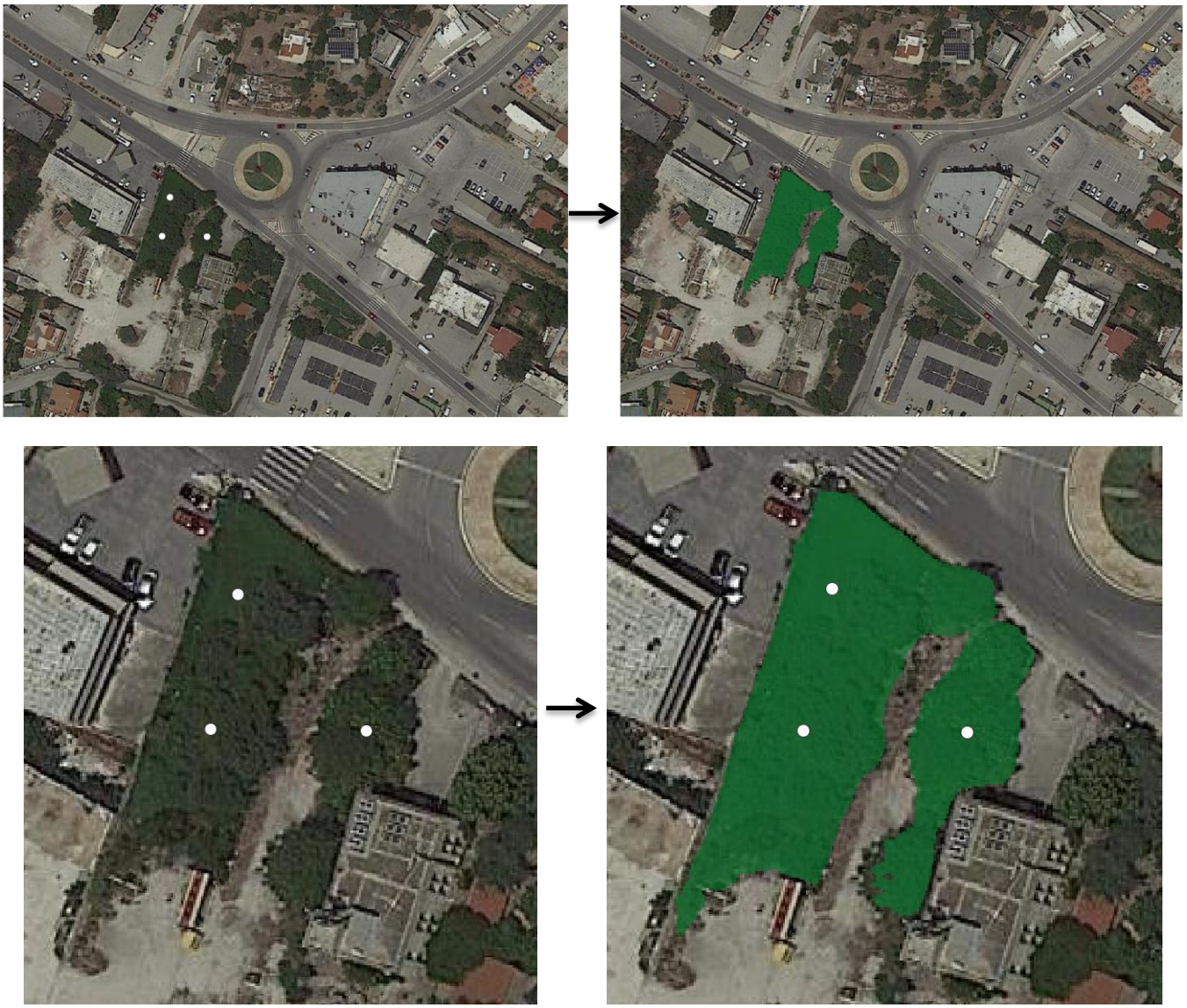
Σχήμα 49 Παράδειγμα 3 εξαγωγής μάσκας κτιρίων με χρήση πολυγωνικών προτροπών

4.2.2 ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΤΑΤΜΗΣΗΣ ΔΟΡΥΦΟΡΙΚΩΝ ΕΙΚΟΝΩΝ ΚΑΙ ΕΞΑΓΩΓΗ ΜΑΣΚΩΝ ΕΚΤΑΣΕΩΝ ΠΡΑΣΙΝΟΥ ΜΕ ΧΡΗΣΗ ΣΗΜΕΙΩΝ ΩΣ ΠΡΟΤΡΟΠΩΝ

Στη παρούσα υποενότητα, θα επαναληφθεί η ίδια μεθοδολογία με τα πολύγωνα – προτροπές με κάποιες μικρές διαφοροποιήσεις. Μια σημαντική διαφορά είναι η χρήση διανυσματικού αρχείου μορφότυπου shapefile που περιέχει εγγραφές σημείων ενδιαφέροντος που υποδεικνύουν στο μοντέλο που να τμηματοποιήσει την εικόνα εισόδου. Το διανυσματικό αρχείο εκτός από τη γεωχωρική πληροφορία, περιέχει πληροφορία για το είδος των σημείων του. Τα σημεία χωρίζονται σε δύο κλάσεις, τα σημεία στο προσκήνιο της εικόνας (foreground) και τα σημεία στο υπόβαθρο της εικόνας (background). Επιπλέον έχει υλοποιηθεί προγραμματιστικά συνάρτηση που λαμβάνει ως όρισμα το διανυσματικό αρχείο, το διαβάζει και δημιουργεί λίστες με τα ζεύγη συντεταγμένων των σημείων, με την κλάση που

ανήκουν (foreground, background) και το σύστημα αναφοράς τους . Στη συνέχεια, οι συγκεκριμένες πληροφορίες εισάγονται στον κωδικοποιητή προτροπών του μοντέλου. Ο μετασχηματισμός της εικόνας εισόδου σε αναπαραστάσεις εικόνας παραμένει ίδιος. Παρακάτω παρουσιάζονται τα αποτελέσματα και σχολιασμός των αποτελεσμάτων.

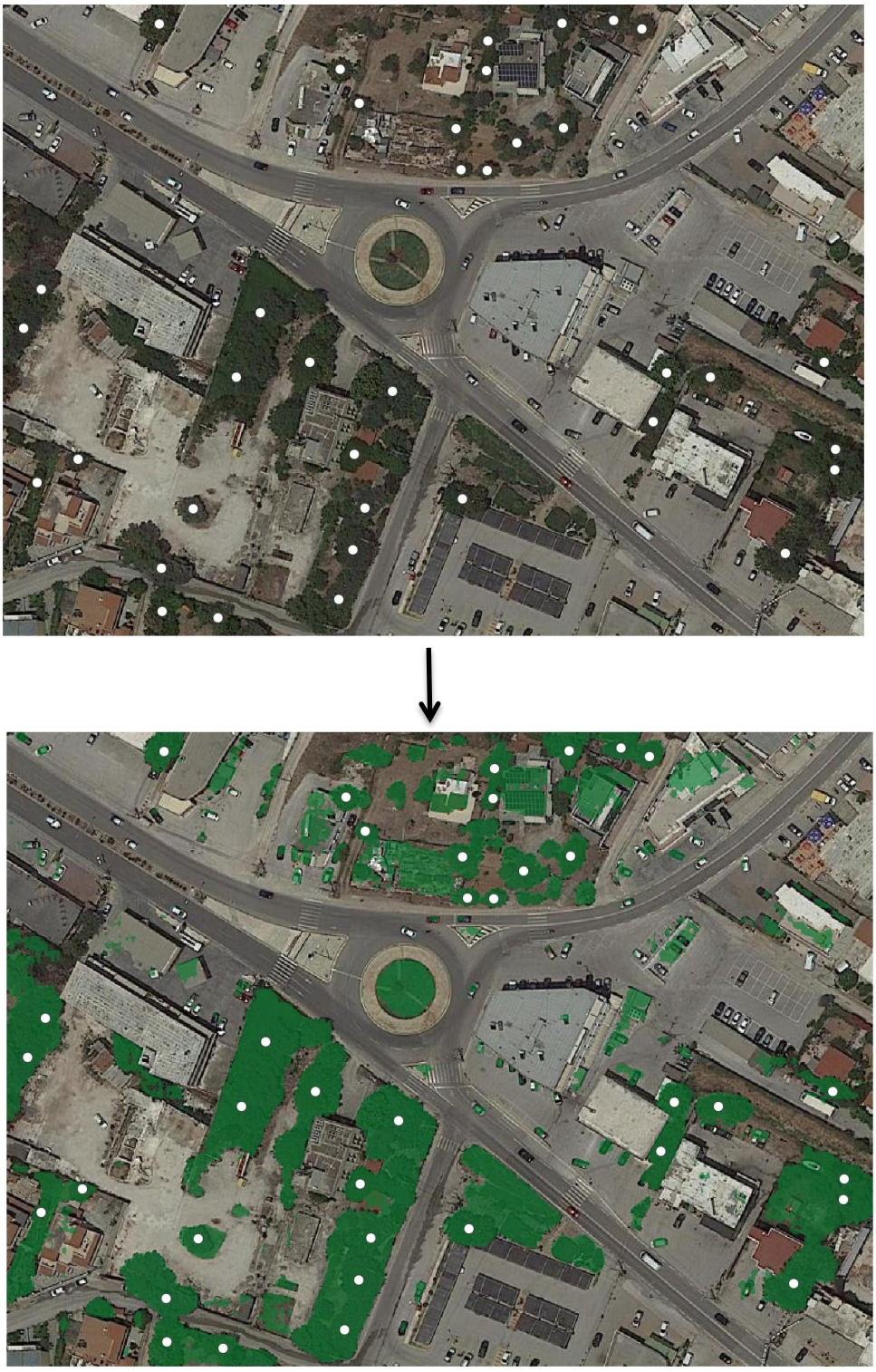
Στη πρώτη δοκιμή το διανυσματικό αρχείο που θα έχει το ρόλο της προτροπής του μοντέλου αποτελείται από τρία ζεύγη συντεταγμένων με σκοπό την εξαγωγή μιας μεμονωμένης μάσκας που αντιστοιχεί σε δασική έκταση.



Σχήμα 50 Χρησιμοποίηση τριών σημείων-προτροπών για εξαγωγή μεμονωμένης μάσκας

Όπως απεικονίζεται στο Σχήμα 49 το μοντέλο Segment Anything δέχεται ως προτροπή τρία σημεία που λειτουργούν ως σημεία προσκηνίου (foreground points) για τμηματοποίηση συγκεκριμένης περιοχής. Τα όρια της μάσκας συμπίπτουν με σχετική ακρίβεια με τα αληθή όρια της περιοχής όπως φαίνεται στο Σχήμα 49.

Επόμενη προσέγγιση της συγκεκριμένης μεθοδολογίας είναι η αύξηση του αριθμού των σημείων στην εικόνα με σκοπό την εξαγωγή όλων των εκτάσεων πρασίνου της εικόνας.



Σχήμα 51 Τμηματοποίηση εικόνας με χρήση σημείων -προτροπών

Σε γενικές γραμμές το μοντέλο έχει τμηματοποιήσει την εικόνα στις περιοχές ενδιαφέροντος όπως υποδεικνύουν τα σημεία – προτροπές (promptable segmentation) . Όμως με περαιτέρω παρατήρηση στο Σχήμα 50 διαπιστώνεται ότι

κάποια μέρη της εικόνας έχουν τμηματοποιηθεί και εξαχθεί λανθασμένα ως μάσκες εκτάσεων πρασίνου. Συμπεραίνεται ότι η κατάτμηση της εικόνας μέσω σημειακών προτροπών έχει καλύτερα αποτελέσματα για εξαγωγή μεμονωμένων μασκών παρά για γενικευμένη χρήση.

Σχήμα 52 Ορθή εξαγωγή εκτάσεων πρασίνου ως αποτέλεσμα τμηματοποίησης της εικόνας

Σχήμα 53 Λανθασμένη εξαγωγή εκτάσεων πρασίνου ως αποτέλεσμα τμηματοποίησης της εικόνας

4.3 ΜΕΘΟΔΟΛΟΓΙΑ ΣΥΝΕΡΓΑΣΙΑΣ ΜΟΝΤΕΛΟΥ ΑΝΙΧΝΕΥΣΗΣ ΟΝΤΟΤΗΤΩΝ ΚΑΙ ΜΟΝΤΕΛΟΥ ΚΑΤΑΤΜΗΣΗΣ ΔΟΡΥΦΟΡΙΚΩΝ ΕΙΚΟΝΩΝ

Στη συγκεκριμένη ενότητα θα υλοποιηθεί ροή εργασίας που θα περιλαμβάνει τη συνεργασία δύο μοντέλων, ενός μοντέλου ανίχνευσης αντικειμένων και ενός μοντέλου κατάτμησης εικόνας το οποίο είναι το Segment Anything. Το μοντέλο ανίχνευσης αντικειμένων που θα χρησιμοποιηθεί είναι το Grounding Dino.

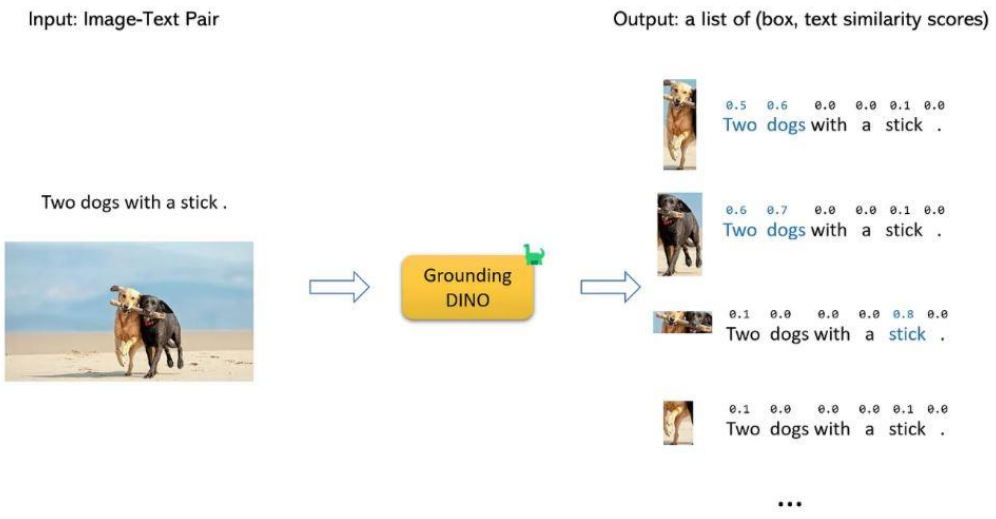
4.3.1 ΠΑΡΟΥΣΙΑΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ GROUNDING DINO, ΜΟΝΤΕΛΟΥ ΑΝΙΧΝΕΥΣΗΣ ΑΝΤΙΚΕΙΜΕΝΩΝ

Το μοντέλο Grounding Dino είναι ένα προ-εκπαιδευμένο zero-shot μοντέλο ανίχνευσης αντικειμένων. Κυκλοφόρησε τον Μάρτιο του 2023 και αναπτύχθηκε με σκοπό τη δημιουργία ενός ισχυρού συστήματος για την ανίχνευση αντικειμένων που ορίζονται από τον εκάστοτε χρήστη μέσω κειμένου (text prompt) χωρίς την ανάγκη επανεκπαίδευσης του. (Liu et al., 2023)

Τα περισσότερα από τα υπάρχοντα μοντέλα ανίχνευσης αντικειμένων έχουν εκπαιδευτεί για μια περιορισμένη ομάδα κλάσεων αντικειμένων που έχουν οριστεί κατά την εκπαίδευση τους. Η διαδικασία της προσθήκης νέων κλάσεων σε αυτά τα μοντέλα είναι μια επίπονη και χρονοβόρα διαδικασία. Έτσι το Grounding Dino που δεν απαιτεί περαιτέρω εκπαίδευση προσφέρει τεράστιο πλεονέκτημα σε εφαρμογές ανίχνευσης αντικειμένων.

Το Grounding Dino είναι ένας αλγόριθμος αυτό-επιτηρούμενης μάθησης (self-supervised) που συνδυάζει το μοντέλο DINO (ανιχνευτής αντικειμένων, **D**ETR with **I**mproved de**N**oising Anch**O**r boxes, 2022) και το μοντέλο GLIP (Grounded Language-Image Pre-training, 2022). Το DINO είναι μια μέθοδος ανίχνευσης βασισμένη σε Transformers, που πραγματοποιεί με σύγχρονες τεχνικές ανίχνευση αντικειμένων. Το GLIP έχει ως αντικείμενο τη συσχέτιση φράσεων ή λέξεων με αντίστοιχα οπτικά στοιχεία σε μια εικόνα ή ένα βίντεο, συνδέοντας αποτελεσματικά περιγραφές κειμένου με οπτικές αναπαραστάσεις.

Το Grounding Dino δουλεύει με ζευγάρια εικόνας- κειμένου (text prompt). Για τη κατανόηση του μηχανισμού λειτουργίας του μοντέλου παρατίθεται το παρακάτω παράδειγμα.



Σχήμα 54 Παράδειγμα λειτουργίας Grounding Dino μοντέλου (πηγή: Liu et al., 2023)

Πρώτο βήμα είναι το μοντέλο να χρησιμοποιήσει τους μηχανισμούς κατανόησης της γλώσσας που διαθέτει για να εντοπίσει τα αντικείμενα που αναφέρονται στη φράση εισόδου (text prompt). Άρα στη άνω φράση το μοντέλο θα προσπαθήσει να εντοπίσει τις λέξεις σκυλιά και ραβδί ως αντικείμενα.

Έπειτα, θα δημιουργήσει ένα σύνολο προτάσεων αντικειμένων για κάθε αντικείμενο που αναγνωρίστηκε στη περιγραφή σε φυσική γλώσσα (natural language). Οι προτάσεις αντικειμένων δημιουργούνται χρησιμοποιώντας μια ποικιλία χαρακτηριστικών όπως το χρώμα, το σχήμα και την υφή των αντικειμένων.

Στη συνέχεια το μοντέλο επιστρέφει την βαθμολογία για κάθε πρόταση αντικειμένου. Η βαθμολογία είναι ένα μέτρο της πιθανότητας η πρόταση του αντικειμένου να περιέχει όντως ένα αντικείμενο.

Τέλος, το μοντέλο θα επιλέξει τις προτάσεις αντικειμένων με την υψηλότερη βαθμολογία ως τις τελικές του επιλογές για ανίχνευση αντικειμένων. Οι τελικές του επιλογές είναι τα αντικείμενα για τα οποία το μοντέλο είναι βέβαιο ότι είναι παρόντα στην εικόνα.

4.3.2 ΠΑΡΟΥΣΙΑΣΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ ΣΥΝΔΥΑΣΜΟΥ ΤΩΝ ΔΥΟ ΜΟΝΤΕΛΩΝ (ΠΑΡΑΔΕΙΓΜΑ 1)

Έχει υλοποιηθεί προγραμματιστικά κλάση (LangSAM) η οποία ενσωματώνει και τα δύο μοντέλα, Grounding Dino και Segment Anything. Όπως και σε προηγούμενες μεθοδολογίες, ο κωδικοποιητής εικόνας (image encoder) λαμβάνει την εικόνα εισόδου και τη μετατρέπει σε αντίστοιχη αναπαράσταση (image embedding).

Επιπλέον ορίζεται ο Vision Transformer που θα χρησιμοποιήσει το μοντέλο, ο οποίος είναι ο ViT –H. Ως δεδομένα εισόδου θα χρησιμοποιηθούν δύο τρικάναλες δορυφορικές απεικονίσεις οι οποίες προέρχονται από ανοικτά δεδομένα και συγκεκριμένα από το δορυφορικό υπόβαθρο της Google.

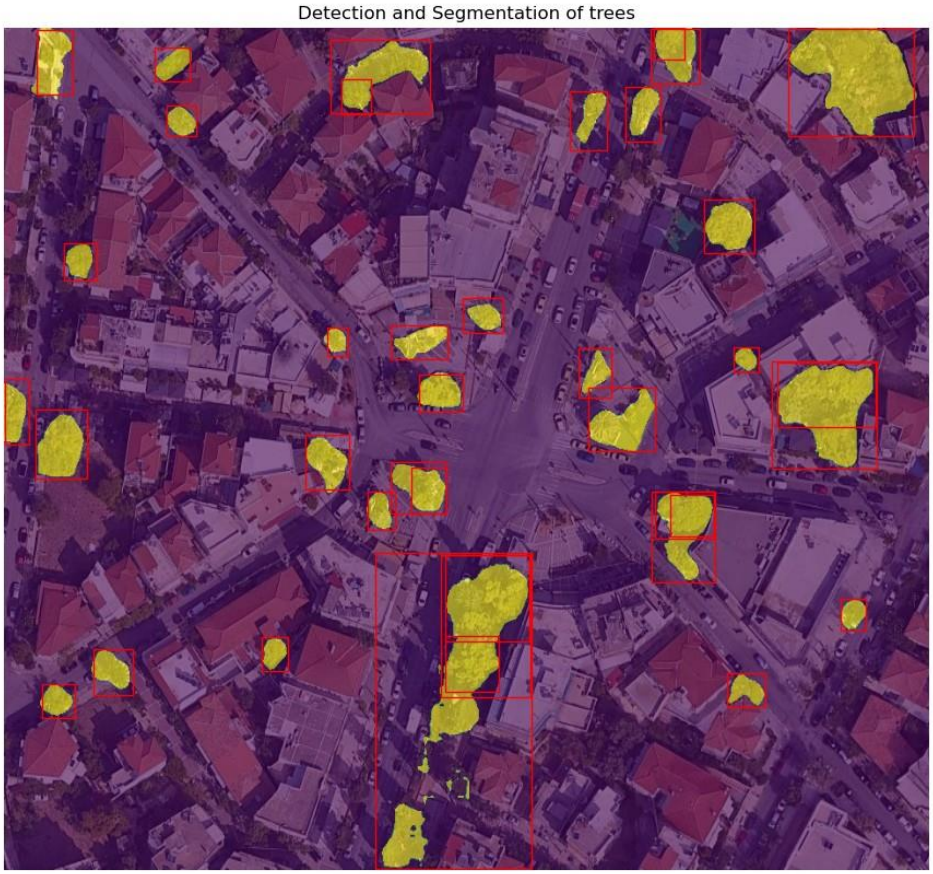
Στο πρώτο παράδειγμα, επιχειρείται η ανίχνευση δένδρων σε αστικό περιβάλλον. Έτσι το κείμενο εισόδου στο μοντέλο (text prompt) θα είναι η λέξη «tree». Με βάση αυτή τη λέξη το μοντέλο Grounding Dino θα δημιουργήσει πολύγωνα ανίχνευσης γύρω από τις τελικές του επιλογές (επιλογές με μεγαλύτερο σκορ εμπιστοσύνης). Στη συνέχεια το μοντέλο Segment Anything θα λάβει ως γεωμετρικές προτροπές τα πολύγωνα ανίχνευσης και με βάση αυτά θα επιχειρήσει να τμηματοποιήσει την εικόνα εισόδου.

Σημαντικό είναι να αναλυθούν οι δύο παράμετροι του μοντέλου με βάση τους οποίους γίνεται ανίχνευση των αντικειμένων. Οι δύο παράμετροι λαμβάνουν τιμές από 0 έως 1 και ουσιαστικά αποτελούν κατώφλια για την κλάση LangSAM.

1. Κατώφλι για πολύγωνα: Αυτή η τιμή χρησιμοποιείται για την ανίχνευση αντικειμένων στην εικόνα. Μια υψηλότερη τιμή κάνει το μοντέλο πιο επιλεκτικό, αναγνωρίζοντας μόνο τις πιο βέβαιες περιπτώσεις αντικειμένων, οδηγώντας σε λιγότερες συνολικές ανιχνεύσεις. Αντίθετα, μια χαμηλότερη τιμή κάνει το μοντέλο πιο ανεκτικό, οδηγώντας σε αυξημένες ανιχνεύσεις, συμπεριλαμβανομένων ενδεχομένως λιγότερο βέβαιων περιπτώσεων.

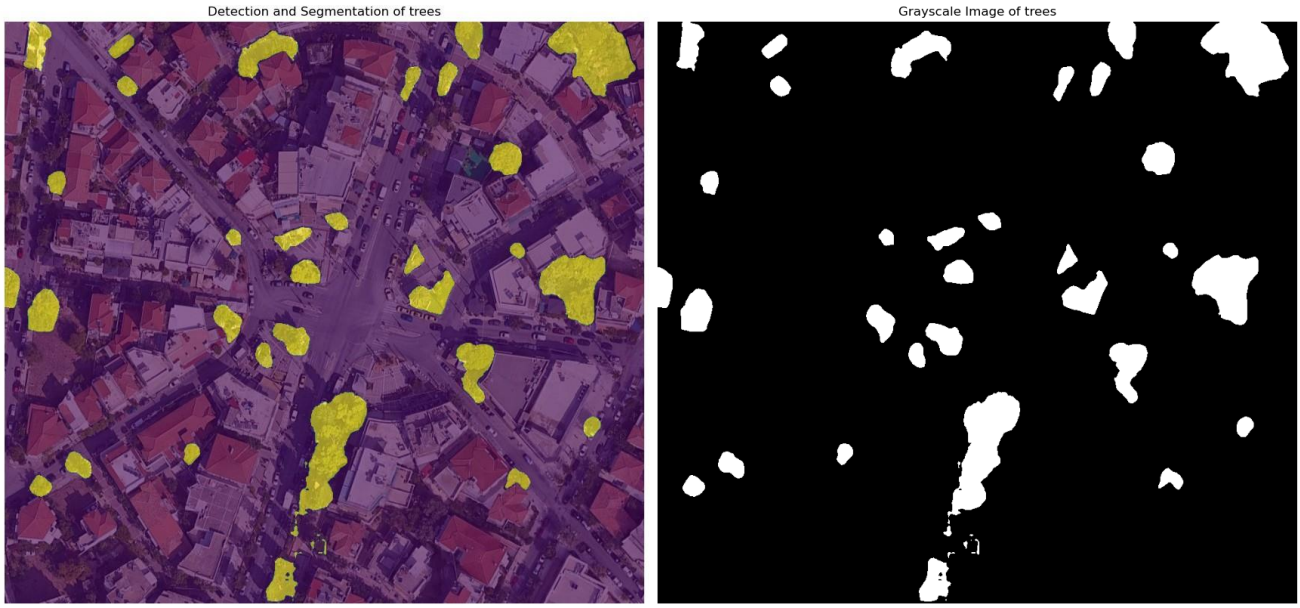
2. Κατώφλι κειμένου: Αυτή η τιμή χρησιμοποιείται για να συσχετίσει τα ανιχνευμένα αντικείμενα με την παρεχόμενη προτροπή κειμένου. Μια υψηλότερη τιμή απαιτεί ισχυρότερη συσχέτιση μεταξύ του αντικειμένου και της προτροπής, οδηγώντας σε πιο ακριβείς αλλά ενδεχομένως λιγότερες συσχετίσεις. Μια χαμηλότερη τιμή επιτρέπει χαλαρότερες συσχετίσεις, που θα μπορούσαν να αυξήσουν τον αριθμό των συσχετίσεων αλλά και να εισάγουν λιγότερο ακριβείς αντιστοιχίες.

Στη συγκεκριμένη περίπτωση λαμβάνεται μια κοινή τιμή και για τα δύο κατώφλια. Παρακάτω παρουσιάζονται τα αποτελέσματα της ανίχνευσης δένδρων σε αστικό ιστό μέσω της προτροπής κειμένου «tree» και τμηματοποίηση της εικόνας για εξαγωγή των μασκών ενδιαφέροντος.



Σχήμα 55 Αποτέλεσμα μεθοδολογίας ανίχνευσης αντικειμένων και εξαγωγή μασκών τους μέσω τμηματοποίησης εικόνας.

Το μοντέλο Grounding Dino έχει ανιχνεύσει 39 πολύγωνα οριοθέτησης (bounding boxes). Αυτά εισάγονται στο Segment Anything ως γεωμετρικές προτροπές και πραγματοποιείται η εξαγωγή μασκών. Παρακάτω παρατίθεται το τελικό αποτέλεσμα χωρίς τα πολύγωνα οριοθέτησης καθώς και οι μάσκες ως εικόνα ενός καναλιού (παγχρωματική εικόνα).



Σχήμα 56 Αποτέλεσμα της διαδικασίας (αριστερά). Μάσκες ως παγχρωματική εικόνα (δεξιά)

4.3.3 ΠΑΡΟΥΣΙΑΣΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ ΣΥΝΔΥΑΣΜΟΥ ΤΩΝ ΔΥΟ ΜΟΝΤΕΛΩΝ (ΠΑΡΑΔΕΙΓΜΑ 2)

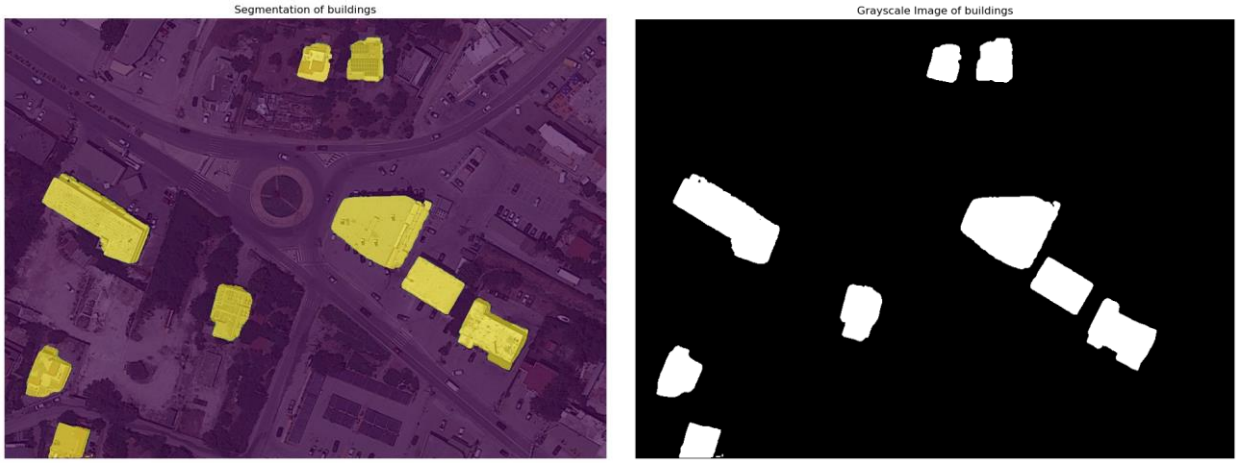
Στο Παράδειγμα 2 θα πραγματοποιηθεί παρόμοια διαδικασία με το Παράδειγμα 1 αλλά με τελικό αποτέλεσμα την ανίχνευση και την εξαγωγή κτιριακών εγκαταστάσεων από μια δορυφορική απεικόνιση. Η απεικόνιση αποτελείται από 3 κανάλια (RGB) και προέρχεται από ανοιχτά δεδομένα και συγκεκριμένα από το δορυφορικό υπόβαθρο της Google.

Η προτροπή κειμένου που θα χρησιμοποιηθεί είναι η λέξη «building». Η διαδικασία είναι ακριβώς ίδια με το προηγούμενο παράδειγμα. Τα πολύγωνα οριοθέτησης που υπολογίζονται είναι 9 στον αριθμό και αποτελούν τις γεωμετρικές προτροπές στο μοντέλο Segment Anything.

Αξίζει να σημειωθεί ότι οι τιμές των κατωφλιών δεν είναι ίδιες σε όλες τις περιπτώσεις. Εξαρτώνται από την απεικόνιση εισόδου αλλά και το αντικείμενο που πρέπει να ανιχνευτεί και να εξαχθεί η μάσκα του. Οι σωστές τιμές υπολογίζονται ύστερα από μια σειρά πειραματικών προσεγγίσεων του προβλήματος. Παρακάτω παρατίθενται τα αποτελέσματα της διαδικασίας για το Παράδειγμα 2.

Σχήμα 57 Αποτέλεσμα μεθοδολογίας ανίχνευσης αντικειμένων και εξαγωγή μασκών τους μέσω τμηματοποίησης εικόνας.

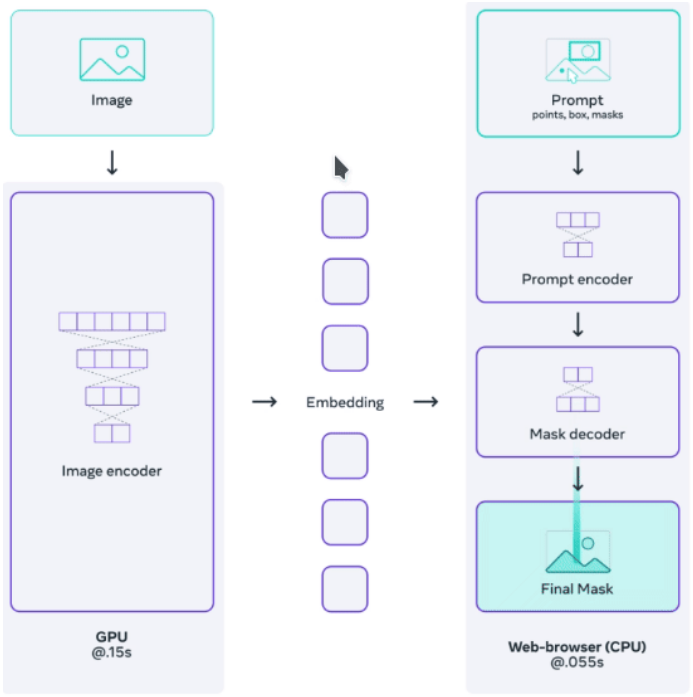
Παρακάτω παρατίθεται το τελικό αποτέλεσμα χωρίς τα πολύγωνα οριοθέτησης καθώς και οι μάσκες ως εικόνα ενός καναλιού (παγχρωματική εικόνα).



Σχήμα 58 Αποτέλεσμα της διαδικασίας (αριστερά). Μάσκες ως παγχρωματική εικόνα (δεξιά)

Παρατηρώντας και τα δύο παραδείγματα, οι μάσκες των κτιριακών εγκαταστάσεων καθώς και των δένδρων σε αστικό περιβάλλον έχουν αποδοθεί πιστά ως γεωμετρία. Παρατηρείται ότι το μοντέλο Grounding Dino χάνει κάποια αντικείμενα και δε τα περιβάλλει με πολύγωνα οριοθέτησης ώστε το Segment Anything να τα κατατμήσει. Αυτό οφείλεται στο γεγονός ότι δεν έχει εκπαιδευτεί με δορυφορικές απεικονίσεις με αποτέλεσμα να δυσχεραίνεται σε ορισμένες περιπτώσεις η λειτουργία του.

Τέλος ανεξάρτητα των παραλλαγών των μεθοδολογιών που ενσωματώνεται το μοντέλο κατάτμησης με χρήση προτροπών Segment Anything, η ροή εργασίας που πραγματοποιείται είναι παρόμοια. Η ροή εργασίας αποτυπώνεται ξεκάθαρα στο Σχήμα 59 και έχει περιγραφεί αναλυτικά σε πολλές διαδικασίες στις προηγούμενες ενότητες.



Σχήμα 59 Διάγραμμα ροής για το μοντέλο Segment Anything

4.4 ΥΛΟΠΟΙΗΣΗ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΑ ΑΥΤΟΝΟΜΗΣ ΕΦΑΡΜΟΓΗΣ ΓΙΑ ΑΝΙΧΝΕΥΣΗ ΚΑΙ ΕΞΑΓΩΓΗ ΟΝΤΟΤΗΤΩΝ ΑΠΟ ΔΟΡΥΦΟΡΙΚΕΣ ΑΠΕΙΚΟΝΙΣΕΙΣ

Στη παρούσα ενότητα θα παρουσιαστεί μια εφαρμογή που έχει υλοποιηθεί στα πλαίσια της διπλωματικής εργασίας, το «Segment it». Η συγκεκριμένη εφαρμογή είναι αυτόνομη και στοχεύει στην ενσωμάτωση όλων των δυνατοτήτων που παρέχει το μοντέλο «Segment Anything» και στη λειτουργία ως αυτόνομου εργαλείου για κατάτμηση και εξαγωγή οντοτήτων από κάθε δορυφορική απεικόνιση.

Η εφαρμογή έχει ως κύρια λειτουργία την κατάτμηση δορυφορικών εικόνων με βάση τις προτροπές (promptable segmentation). Οι προτροπές που χρησιμοποιεί το μοντέλο και η εφαρμογή για να κατατμήσει τις εικόνες είναι οι παρακάτω:

- Μια σειρά σημείων προσκηνίου/υποβάθρου (foreground, background points)
- Πλαίσια οριοθέτησης (bounding box)
- Ελεύθερο κείμενο (Free-form text)

Στην περίπτωση της χρήσης του κειμένου ως προτροπής, στην εφαρμογή ενσωματώνεται και το μοντέλο Grounding Dino που λειτουργεί ως μοντέλο ανίχνευσης αντικειμένων.

Η εφαρμογή «Segment it» είναι μια πολυσέλιδη εφαρμογή η οποία περιλαμβάνει 7 σελίδες, η καθεμία με ξεχωριστή λειτουργία οι οποίες είναι οι κάτωθι:

- Αρχική Σελίδα (Home)

Η πρώτη σελίδα της εφαρμογής περιλαμβάνει γενικές πληροφορίες για τις δυνατότητες της εφαρμογής και παρουσιάζει ένα παράδειγμα μασκών τμηματοποίησης μέσω εργαλείου κύλισης (slider)

- Σελίδα λήψης δορυφορικών απεικονίσεων (Download Imagery)
- Σελίδα ψηφιοποίησης γεωμετρικών προτροπών (Digitize Prompts)
- Σελίδα κατάτμησης εικόνας με χρήση προτροπών κειμένου (Segment it, Text Prompts)
- Σελίδα κατάτμησης με χρήση πλαισίων οριοθέτησης ως προτροπών (Segment it, Box Prompts)
- Σελίδα κατάτμησης με χρήση σημειακών προτροπών (Segment it, Point Prompts)
- Σελίδα αυτόματης παραγωγής μασκών (Segment it, Automated)

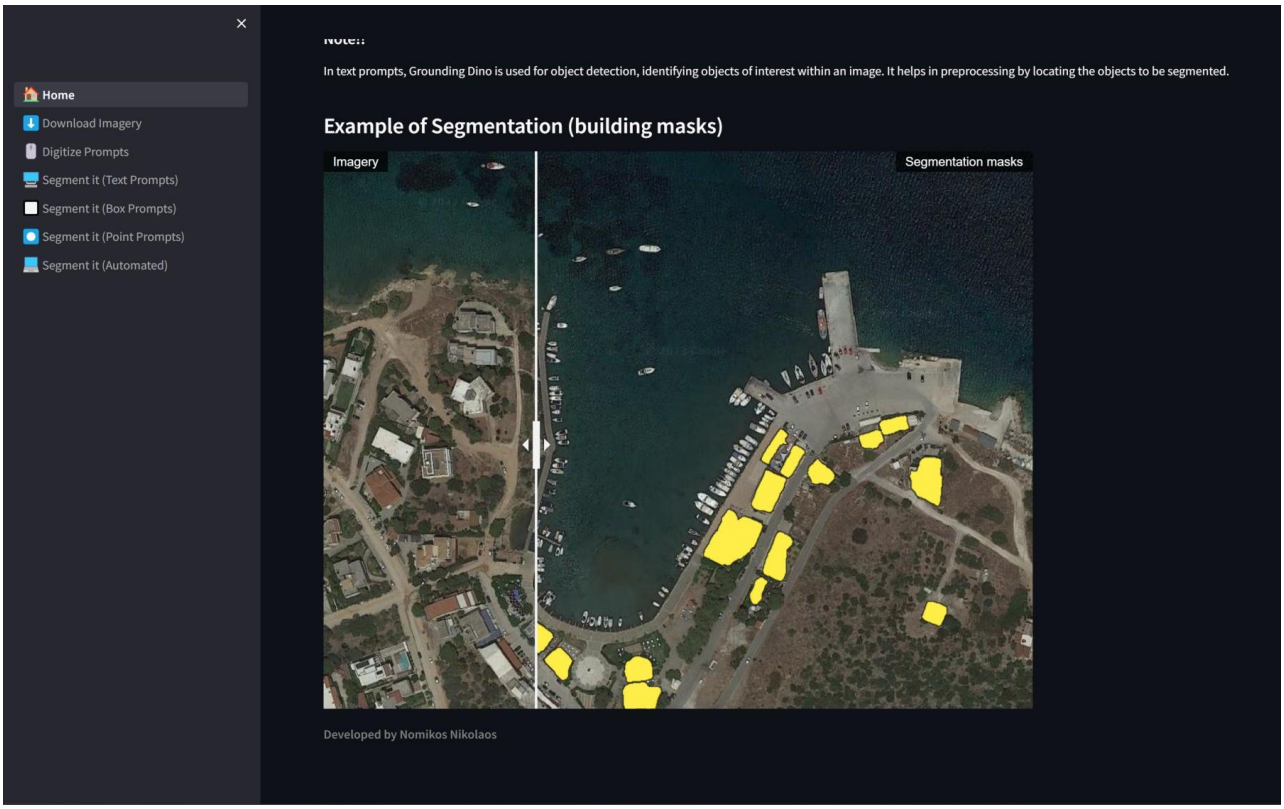
Η εφαρμογή αποτελεί μια ολοκληρωμένη προσέγγιση στη «promptable segmentation» των δορυφορικών εικόνων και συνδυάζει επιπλέον λειτουργικά στοιχεία από Γεωγραφικά Συστήματα Πληροφοριών όπως η δημιουργία

περιβάλλοντος χάρτη στο οποίο ενσωματώνονται οι εικόνες και τα διανυσματικά δεδομένα. Όλες οι εικόνες καθώς και τα προϊόντα τμηματοποίησης είναι γεωαναφερμένα.

Συνοπτικά οι κύριες τεχνολογίες που χρησιμοποιεί η εφαρμογή είναι οι παρακάτω:

- Streamlit: Το Streamlit είναι ένα ανοιχτού κώδικα πλαίσιο που χρησιμοποιείται για τη δημιουργία διαδικτυακών εφαρμογών με Python. Είναι ιδιαίτερα χρήσιμο για την ανάπτυξη εφαρμογών που απαιτούν γρήγορη και διαδραστική οπτικοποίηση δεδομένων, Στην εφαρμογή «Segment it», το Streamlit χρησιμοποιείται για την ανάπτυξη της διεπαφής χρήστη, προσφέροντας έναν φιλικό προς τον χρήστη τρόπο αλληλεπίδρασης με τις διάφορες λειτουργίες τμηματοποίησης δορυφορικών εικόνων.
- PyTorch: Το PyTorch είναι ένα ευρέως χρησιμοποιούμενο πλαίσιο μηχανικής μάθησης και βαθιάς μάθησης που αναπτύχθηκε από το Facebook's AI Research lab (FAIR). Παρέχει πλήρη υποστήριξη σε θέματα βαθιάς μάθησης με ενσωμάτωση μοντέλων και αντίστοιχων σετ δεδομένων. Χρησιμοποιείται ευρέως και σε εφαρμογές υπολογιστικής όρασης. Ο αρχικός κώδικας του μοντέλου Segment Anything έχει δομηθεί πάνω στη βιβλιοθήκη βαθιάς μάθησης, PyTorch.
- Segment-anything και segment-geospatial. Το πρώτο αποθετήριο κώδικα υλοποιεί προγραμματιστικά το μοντέλο Segment Anything και την αρχιτεκτονική του περιλαμβάνοντας όλες τις απαραίτητες κλάσεις και συναρτήσεις. Το δεύτερο αποθετήριο κώδικα, αξιοποιεί τις δυνατότητες του SAM σε γεωχωρικό επίπεδο με την τμηματοποίηση των δορυφορικών εικόνων.
- Geopandas και Rasterio. Το Geopandas είναι ένα εργαλείο ανοιχτού κώδικα που επιτρέπει την εύκολη διαχείριση και ανάλυση γεωχωρικών δεδομένων. Το Rasterio είναι μια βιβλιοθήκη Python για την ανάγνωση και τη συγγραφή δεδομένων εικόνας. Αυτές οι βιβλιοθήκες λειτουργούν συμπληρωματικά σε μια σειρά από διεργασίες που λαμβάνουν μέρος στην εφαρμογή.
- Matplotlib και PIL: Το Matplotlib είναι μια ευέλικτη βιβλιοθήκη για τη δημιουργία στατικών, κινούμενων και διαδραστικών γραφικών στην Python. Το PIL (Python Imaging Library) και η σύγχρονη εκδοχή του, Pillow, παρέχουν ισχυρά εργαλεία για την επεξεργασία εικόνων. Στην εφαρμογή «Segment it», αυτές οι βιβλιοθήκες χρησιμοποιούνται για την απεικόνιση των εικόνων και των μασκών τμηματοποίησης, επιτρέποντας στους χρήστες να βλέπουν και να επεξεργάζονται τα αποτελέσματα των τμηματοποιήσεων τους.
- Shapely: Το Shapely είναι μια βιβλιοθήκη για τον χειρισμό και την ανάλυση των γεωμετρικών σχημάτων στην Python. Υποστηρίζει διάφορους τύπους γεωμετριών, όπως σημεία, γραμμές και πολύγωνα, και παρέχει εργαλεία για την εκτέλεση χωρικών πράξεων, όπως η διασταύρωση, η συνένωση και η διαφορά γεωμετριών. Στην εφαρμογή «Segment It», το Shapely

χρησιμοποιείται για την διαχείριση και τον έλεγχο γεωμετρικών σχημάτων, εξασφαλίζοντας ότι οι γεωμετρίες των προτροπών των χρηστών ευθυγραμμίζονται σωστά με τις εικόνες που πρόκειται να τμηματοποιηθούν.



Σχήμα 60 Αρχική σελίδα της ψηφιακής αυτόνομης εφαρμογής κατάτμησης δορυφορικών εικόνων

Στο Σχήμα 60 απεικονίζεται η αρχική σελίδα της εφαρμογής. Αριστερά έχει δημιουργηθεί ένα μενού πλοήγησης που διευκολύνει τον εκάστοτε χρήστη στην μετάβαση των διάφορων υποσέλιδων της εφαρμογής.

4.4.1 ΣΕΛΙΔΑ ΛΗΨΗΣ ΔΟΡΥΦΟΡΙΚΩΝ ΑΠΕΙΚΟΝΙΣΕΩΝ

Η σελίδα αυτή, που αναπτύχθηκε με τη χρήση βιβλιοθηκών Python όπως οι Folium, Streamlit και segment-geospatial , επιτρέπει στους χρήστες να ορίσουν μια περιοχή ενδιαφέροντος (Area of Interest, AOI) σε έναν διαδραστικό χάρτη και να ληφθούν δορυφορικές απεικονίσεις της περιοχής αυτής σε μορφότυπο GeoTIFF. Οι συγκεκριμένες δορυφορικές απεικονίσεις αποτελούν τις μετέπειτα εικόνες εισόδου του μοντέλου Segment Anything για να τμηματοποιηθούν και να εξαχθούν οι οντότητες ενδιαφέροντος.

Σχήμα 61 Σελίδα λήψης δορυφορικών δεδομένων

Αναλυτικά, οι λειτουργίες της συγκεκριμένης σελίδας περιλαμβάνουν:

1. Δημιουργία Διαδραστικού Χάρτη:

Ο χρήστης έχει τη δυνατότητα περιήγησης σε δορυφορικό υπόβαθρο, εστίασης σε περιοχές ενδιαφέροντος (zoom in, zoom out) καθώς και δυνατότητα σχεδιασμού αυτών όπως απεικονίζεται στο Σχήμα 61.

2. Εξαγωγή Συντεταγμένων Περιοχής Ενδιαφέροντος:

Ο χρήστης έχει τη δυνατότητα προβολής και ανάκτησης των συντεταγμένων της περιοχής που σχεδιάζει στον διαδραστικό χάρτη, με βάση τις οποίες θα ληφθεί και η αντίστοιχη δορυφορική εικόνα.

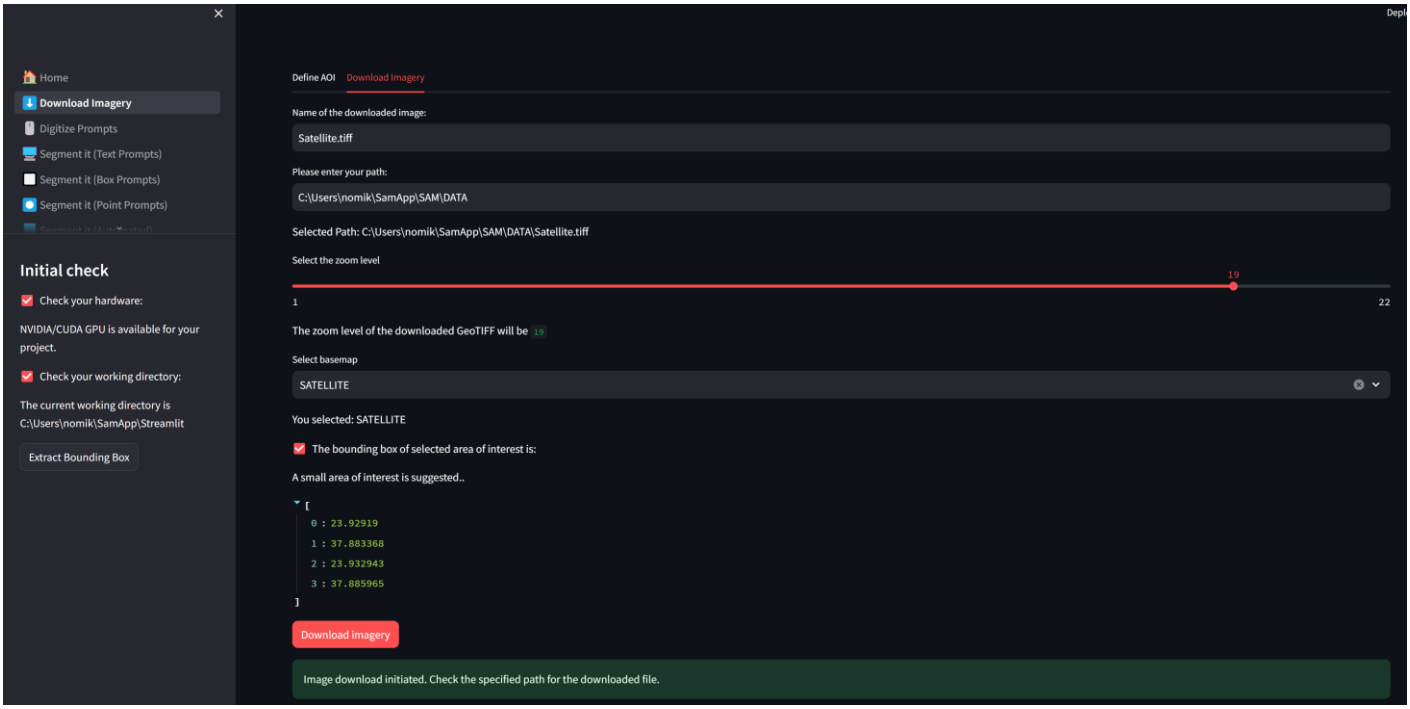
3. Λήψη Δορυφορικών Εικόνων:

Όπως απεικονίζεται στο Σχήμα 62, η εφαρμογή παρέχει φόρμα για τη λήψη δορυφορικών απεικονίσεων. Συγκεκριμένα, υπάρχει δυνατότητα καθορισμού ονόματος αρχείου και διαδρομής αποθήκευσης της εικόνας. Παρέχεται δυνατότητα επιλογής κλίμακας για την αποθήκευσης της αντίστοιχης απεικόνισης καθώς και επιλογής επιθυμητού υπόβαθρου. Μετά τη συμπλήρωση της φόρμας από τον εκάστοτε χρήστη, η απεικόνιση αποθηκεύεται

4. Δυνατότητες Ελέγχου και Πληροφοριών:

Στην εφαρμογή υπάρχουν ενσωματωμένοι έλεγχοι για τη διαθεσιμότητα του επεξεργαστή που χρησιμοποιεί η εφαρμογή (CPU ή GPU) καθώς και το τρέχον περιβάλλον εργασίας στον εκάστοτε υπολογιστή. Επίσης οποιαδήποτε περιοχή σχεδιάζεται στο χάρτη προβάλλεται δεξιά της σελίδας ως GeoJSON αρχείο για περαιτέρω έλεγχο από τον χρήστη.

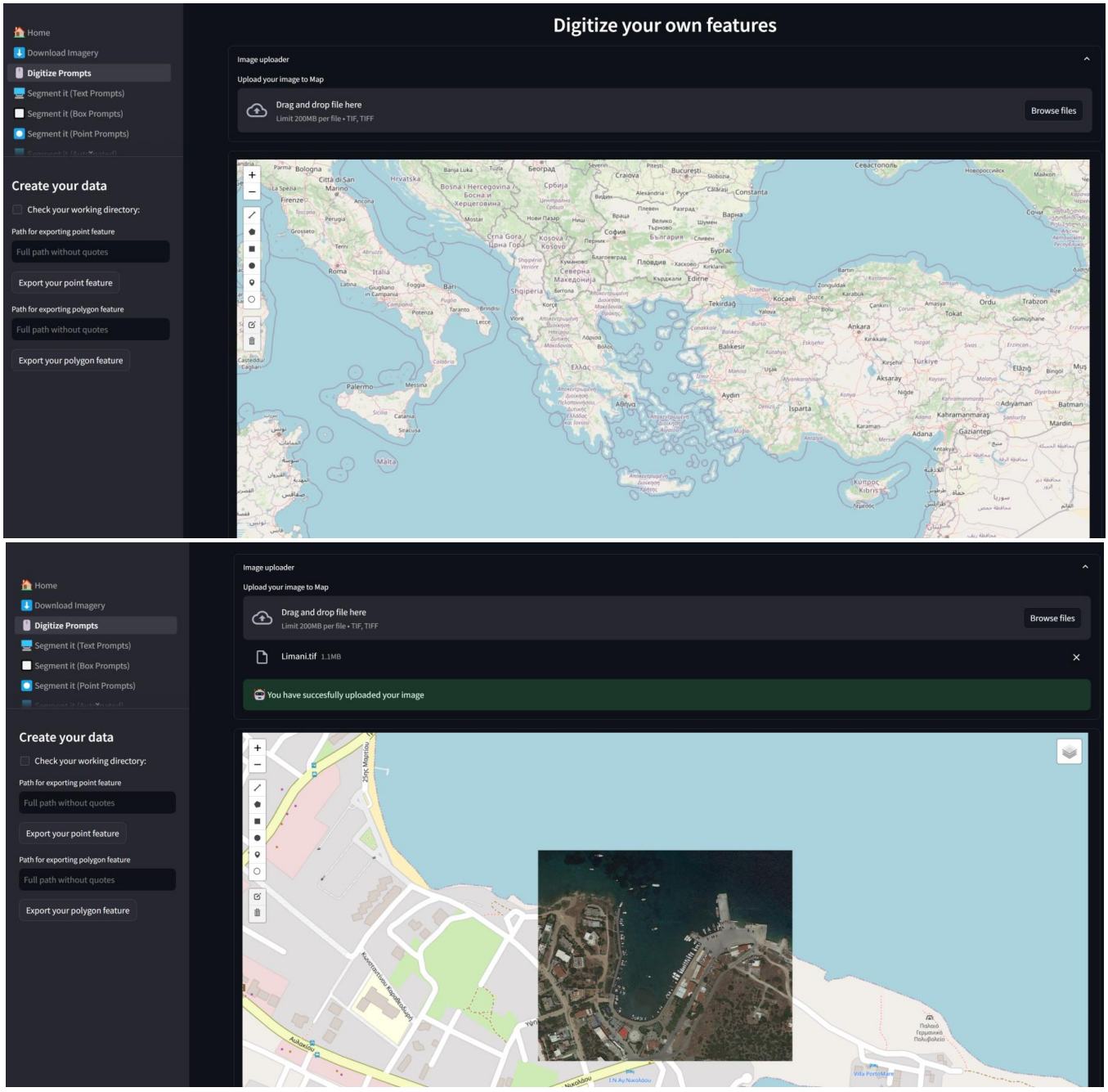
Η εφαρμογή παρέχει μια εύχρηστη διεπαφή που επιτρέπει στους χρήστες να επιλέγουν και να λαμβάνουν δορυφορικές εικόνες με απλότητα και ακρίβεια, καθιστώντας την ιδανική για χρήστες που χρειάζονται να εργαστούν με γεωχωρικά δεδομένα και δορυφορικές εικόνες.



Σχήμα 62 Φόρμα για λήψη δορυφορικής απεικόνισης

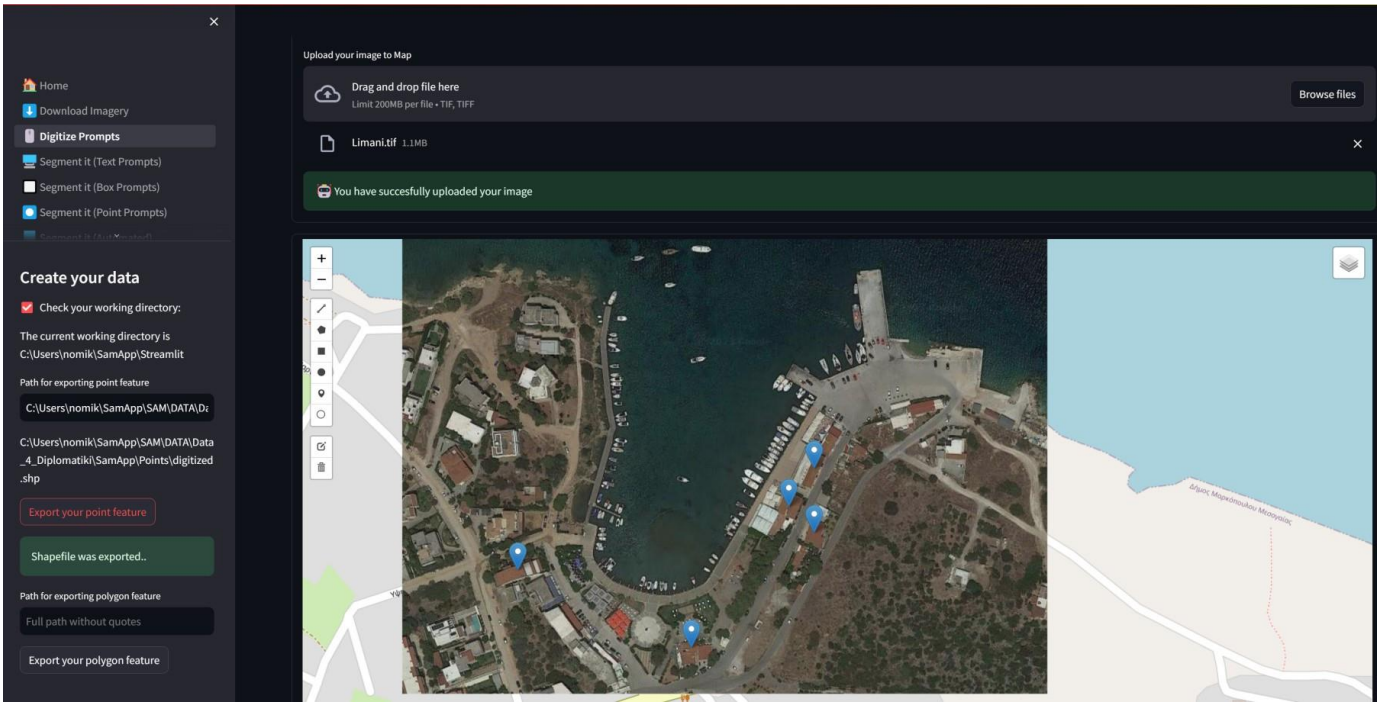
4.4.2 ΣΕΛΙΔΑ ΨΗΦΙΟΠΟΙΗΣΗΣ ΓΕΩΜΕΤΡΙΚΩΝ ΠΡΟΤΡΟΠΩΝ

Η τρίτη σελίδα της εφαρμογής αναπτύχθηκε προσφέροντας τη δυνατότητα ψηφιοποίησης διανυσματικών δεδομένων από δορυφορικές απεικονίσεις. Ο χρήστης έχει τη δυνατότητα να φορτώσει οποιαδήποτε δορυφορική εικόνα διαθέτει τοπικά στον υπολογιστή του, να τη προβάλλει σε διαδραστικό περιβάλλον και ψηφιοποιήσει πάνω σε αυτή διανυσματικά δεδομένα όπως πολύγωνα και σημεία. Τα διανυσματικά δεδομένα αποθηκεύονται τοπικά ως shapefile και δύναται να χρησιμοποιηθούν σε δεύτερο χρόνο ως γεωμετρικές προτροπές για το μοντέλο Segment Anything.

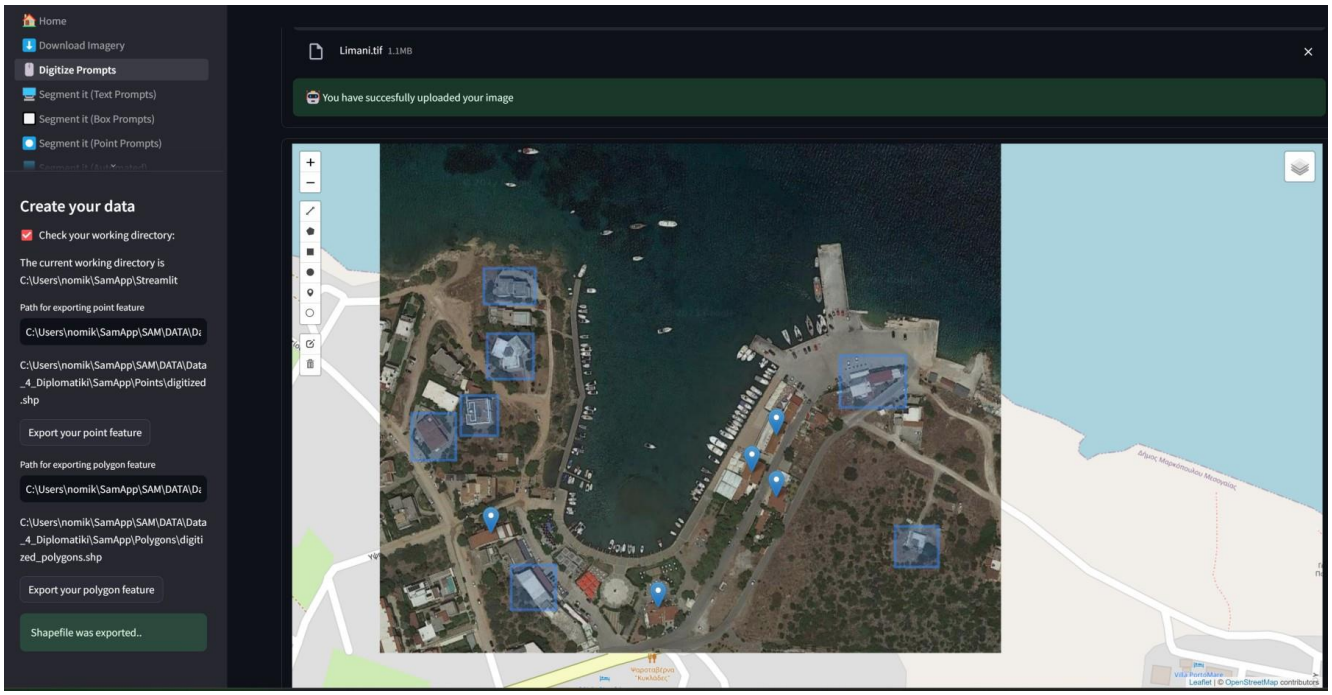


Σχήμα 63 Εισαγωγή Δορυφορικής Εικόνας στην εφαρμογή

Στο Σχήμα 63 απεικονίζεται η εισαγωγή δορυφορικής απεικόνισης στην εφαρμογή. Παρακάτω απεικονίζονται οι ψηφιοποιήσεις σημειακών και πολυγωνικών δεδομένων καθώς και η αποθήκευση τους τοπικά στο σύστημα.



Σχήμα 64 Ψηφιοποίηση σημειακού σετ δεδομένων



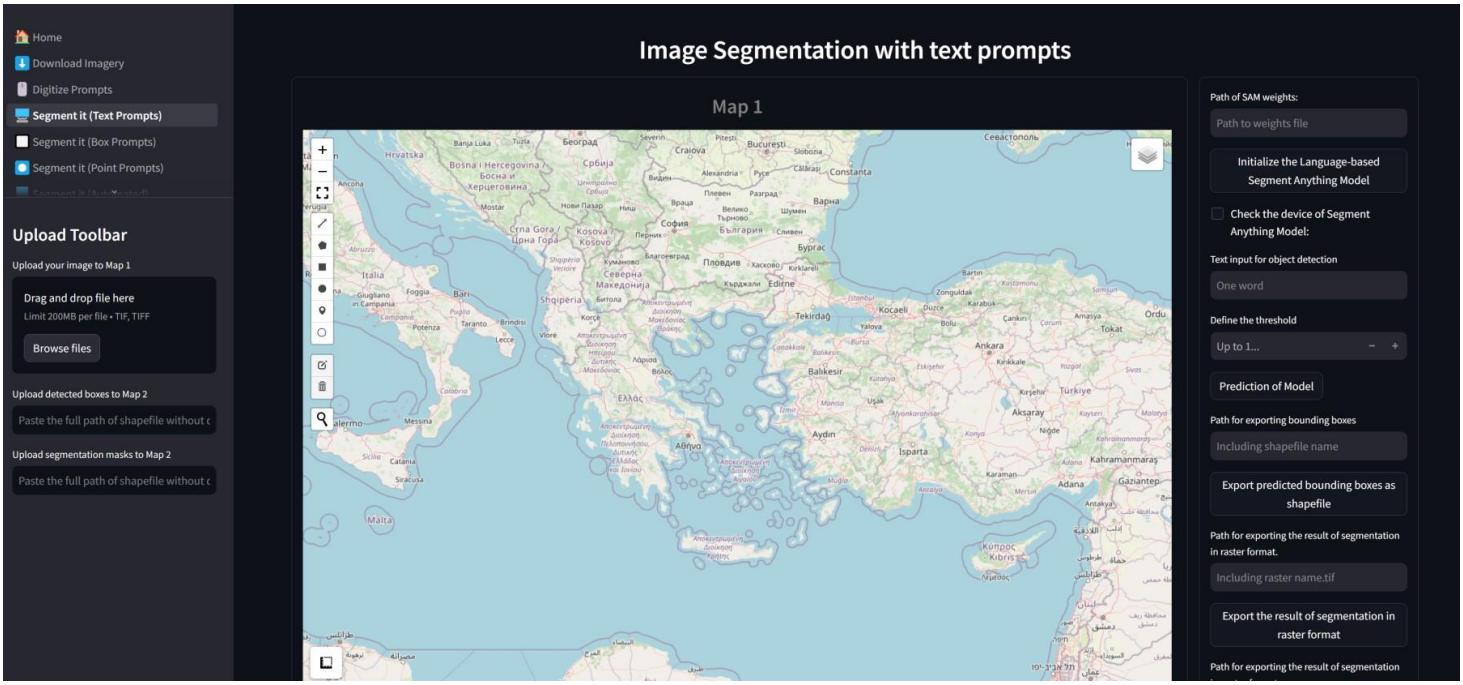
Σχήμα 65 Ψηφιοποίηση πολυγωνικού σετ δεδομένων

Η εφαρμογή παρέχει μια ολοκληρωμένη και φιλική προς τον χρήστη διεπαφή, επιτρέποντας την εύκολη ψηφιοποίηση και εξαγωγή γεωχωρικών δεδομένων από δορυφορικές εικόνες, ιδανική για τη μετέπειτα εισαγωγή τους ως γεωμετρικές προτροπές στο μοντέλο Segment Anything.

4.4.3 ΣΕΛΙΔΑ ΚΑΤΑΤΜΗΣΗΣ ΕΙΚΟΝΑΣ ΜΕ ΧΡΗΣΗ ΠΡΟΤΡΟΠΩΝ ΚΕΙΜΕΝΟΥ

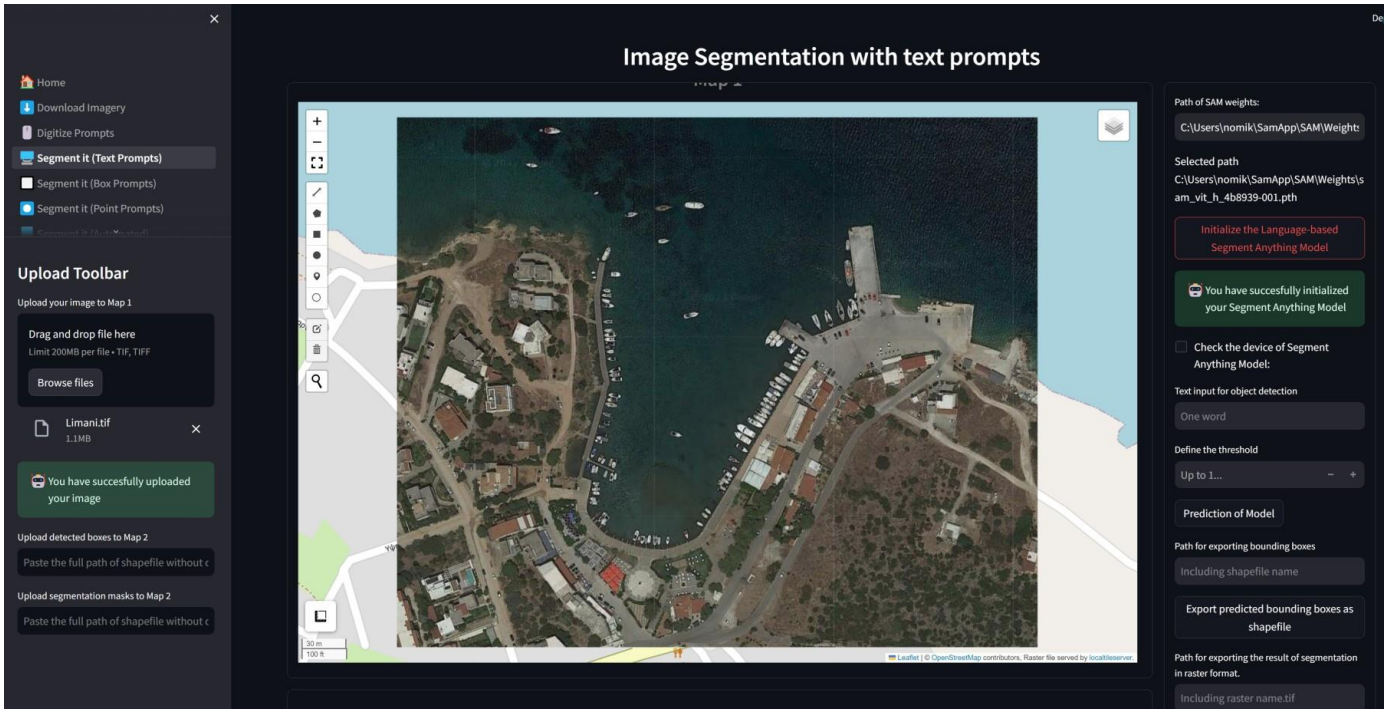
Η συγκεκριμένη σελίδα εφαρμόζει τη μεθοδολογία συνδυασμού των δύο μοντέλων Grounding Dino και Segment Anything, ενός μοντέλου ανίχνευσης αντικειμένων και του μοντέλου κατάτμησης εικόνων με προσθήκη προτροπών.

Η σελίδα αποτελείται από δύο διαδραστικούς χάρτες, ένα χάρτη για προβολή της αρχικής εικόνας εισόδου και ένα χάρτη για τη παρουσίαση των αποτελεσμάτων κατάτμησης του μοντέλου. Αριστερά υπάρχει ένα μενού πλοήγησης για μετάβαση σε διάφορες σελίδες της εφαρμογής καθώς και ένα δεύτερο μενού (upload toolbar) για μεταφόρτωση δεδομένων όπως η εικόνα εισόδου, τα πλαίσια οριοθέτησης που θα υπολογιστούν από το μοντέλο Grounding Dino και τις μάσκες τμηματοποίησης που υπολογίζονται από το μοντέλο SAM. Δεξιά έχει υλοποιηθεί μενού που ελέγχει τη μεθοδολογία της κατάτμησης εικόνας με χρήση προτροπών κειμένου και θα αναλυθούν παρακάτω. Αυτή είναι και η γενική δομή σε όλες τις σελίδες της εφαρμογής με μικρές διαφοροποιήσεις.



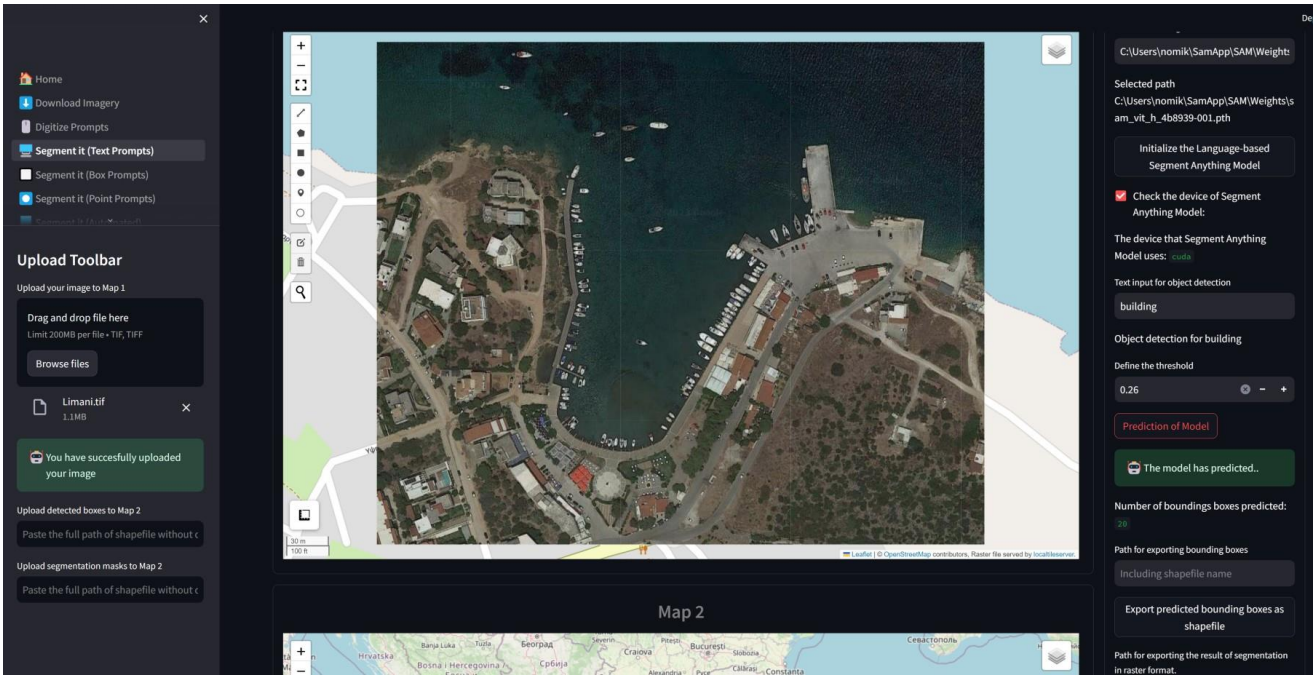
Σχήμα 66 Προεπισκόπηση σελίδας κατάτμησης εικόνας με προτροπές κειμένου

Ο χρήστης αρχικά εισάγει την δορυφορική απεικόνιση η οποία προβάλλεται στο πρώτο διαδραστικό χάρτη. Επιπλέον καθορίζει και τα βάρη του προ-εκπαιδευμένου μοντέλου SAM.



Σχήμα 67 Εισαγωγή δορυφορικής απεικόνισης και καθορισμός βαρών μοντέλου

Έπειτα ο χρήστης εισάγει τη προτροπή κειμένου με βάση την οποία το μοντέλο Grounding Dino θα πραγματοποιήσει ανίχνευση συναφών οντοτήτων. Στο Σχήμα 68 απεικονίζεται η λέξη building. Επιπλέον ορίζεται κοινή τιμή για τις δύο τιμές-κατώφλια που θα χρησιμοποιήσει το μοντέλο όπως έχει περιγραφεί σε προηγούμενη ενότητα. Ύστερα εκτελείται η πρόβλεψη του μοντέλου και εξάγονται τα πλαίσια ανίχνευσης σε μορφότυπο shapefile τοπικά σε διαδρομή αρχείου που έχει δοθεί από τον χρήστη. Ταυτόχρονα στην εφαρμογή αναγράφεται το πλήθος των πλαισίων ανίχνευσης όπως απεικονίζεται στο Σχήμα 68. (20 στο συγκεκριμένο παράδειγμα).



Σχήμα 68 Ορισμός προτροπής κειμένου, τιμής κατωφλιού και ανίχνευση οντοτήτων

Αφού εξαχθούν τα πλαίσια ανίχνευσης ως πολυγωνικό shapefile, εισάγονται στο δεύτερο διαδραστικό χάρτη. Αυτά τα πλαίσια θα χρησιμοποιηθούν ως γεωμετρικές προτροπές για το SAM για να εξαχθούν οι αντίστοιχες οντότητες που περικλείονται από αυτά.

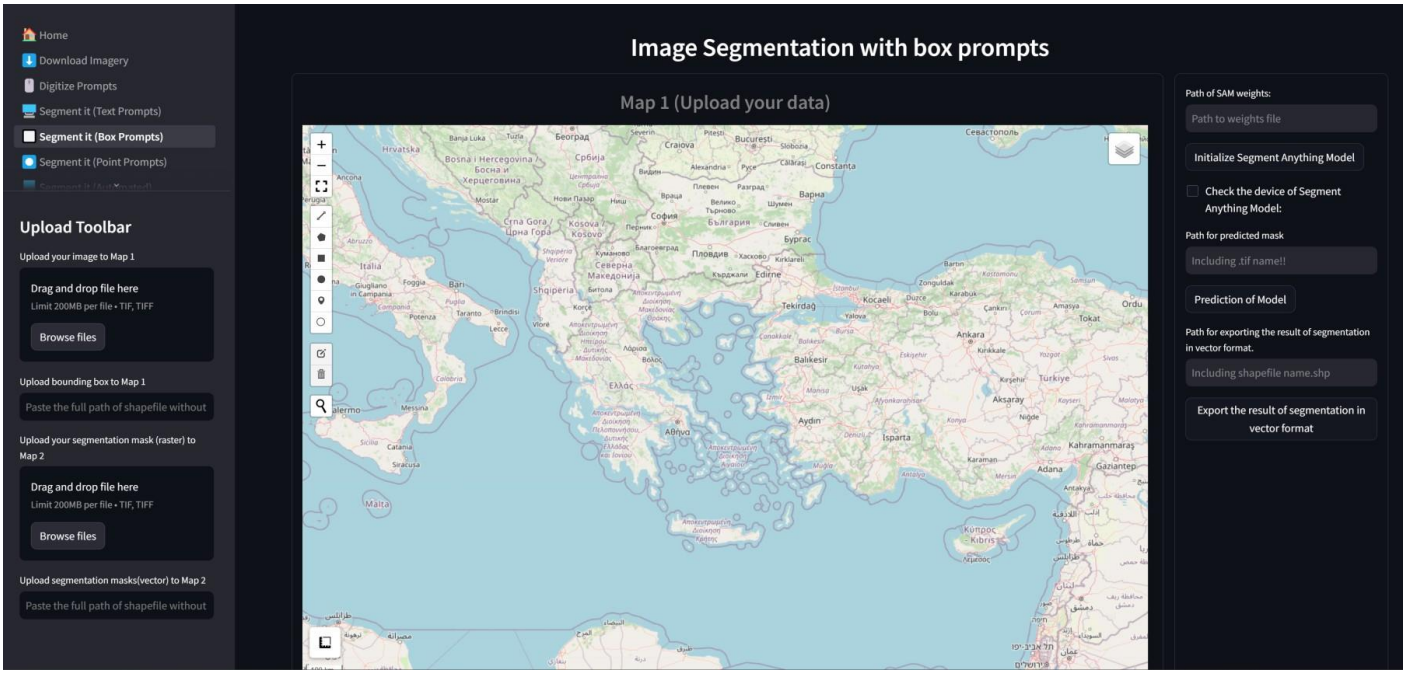
Σχήμα 69 Παρουσίαση αποτελεσμάτων ανίχνευσης οντοτήτων ενδιαφέροντος

Τέλος εκτελείται το SAM και εξάγονται οι μάσκες κατάτμησης σε μορφότυπο εικόνας (raster) και σε διανυσματική μορφή (vector) οι οποίες εισάγονται με παρόμοιο τρόπο στο δεύτερο διαδραστικό χάρτη (Σχήμα 70).

Σχήμα 70 Αποτέλεσμα κατάτμησης αντικειμένων ενδιαφέροντος

4.4.4 ΣΕΛΙΔΑ ΚΑΤΑΤΜΗΣΗΣ ΕΙΚΟΝΑΣ ΜΕ ΧΡΗΣΗ ΠΛΑΙΣΙΩΝ ΟΡΙΟΘΕΤΗΣΗΣ ΩΣ ΠΡΟΤΡΟΠΩΝ

Η συγκεκριμένη σελίδα εφαρμόζει τη μεθοδολογία κατάτμησης της εικόνας με χρήση πλαισίων οριοθέτησης (bounding box) ως προτροπών για το μοντέλο. Ο χρήστης έχει τη δυνατότητα εισαγωγής της απεικόνισης της επιλογής του καθώς και των πλαισίων οριοθέτησης (bounding boxes) εντός των οποίων θα πραγματοποιείται η κατάτμηση. Τα πλαίσια οριοθέτησης που εισάγονται είναι μορφότυπου shapefile.



Σχήμα 71 Αρχική εικόνα της σελίδας κατάτμησης εικόνας με πλαίσια οριοθέτησης

Σχήμα 72 Εισαγωγή δορυφορικής απεικόνισης και πλαισίων οριοθέτησης ως γεωμετρικές προτροπές μοντέλου

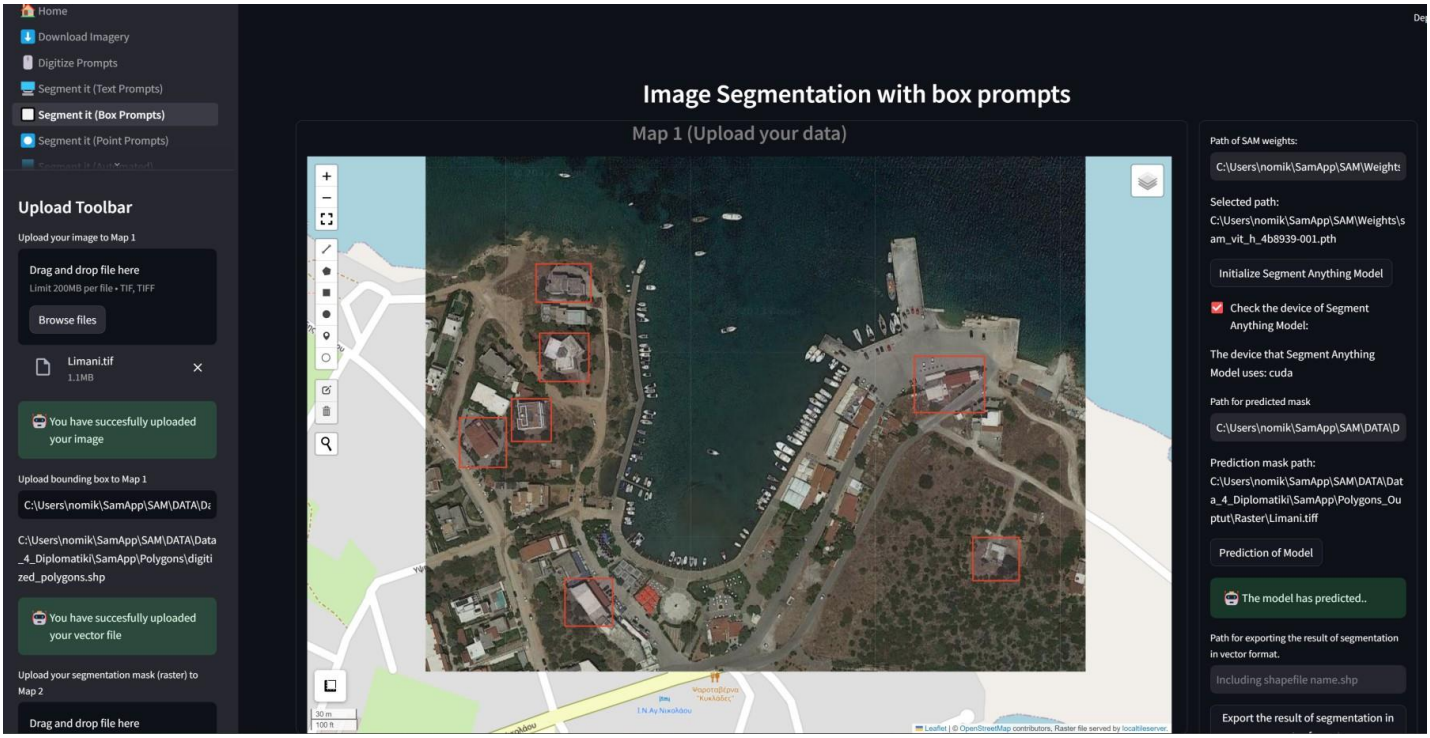
Η σελίδα αποτελείται από δύο διαδραστικούς χάρτες. Ο ένας έχει ως σκοπό την εισαγωγή των αρχικών δεδομένων (εικόνα εισόδου και πλαίσια οριοθέτησης) και προεπισκόπηση τους. Ο δεύτερος διαδραστικός χάρτης χρησιμοποιείται για τη προβολή της εικόνας εισόδου αλλά και την εισαγωγή των αποτελεσμάτων της κατάτμησης μέσω ενσωμάτωσης του μοντέλου Segment Anything. Συγκεκριμένα προβάλλονται οι μάσκες των οντοτήτων που εξάχθηκαν σε μορφή εικόνας (raster αρχείο) αλλά και σε διανυσματική μορφή (vector) για περαιτέρω επεξεργασία. Η

μετατροπή των μασκών σε διανυσματική μορφή αποτελεί λειτουργική προέκταση της συγκεκριμένης εφαρμογής και προσφέρει ευελιξία στην διαχείριση των δεδομένων και στην περαιτέρω επεξεργασία τους.

Ο χρήστης μέσω του μενού στο δεξί μέρος της σελίδας ορίζει τα βάρη του μοντέλου Segment Anything. Τα βάρη είναι τοπικά αποθηκευμένα και ανήκουν στο προ-εκπαιδευμένο μοντέλο SAM. Έπειτα ο χρήστης αρχικοποιεί το μοντέλο SAM. Ο Vision Transformer που χρησιμοποιεί το μοντέλο είναι ο ViT-H και είναι προεπιλεγμένος και καθορισμένος από τον κώδικα της εφαρμογής. Επιπλέον το μοντέλο αν υπάρχει δυνατότητα υλικού (hardware), λειτουργεί στη κάρτα γραφικών του εκάστοτε χρήστη για ταχύτερο υπολογισμό και εκτέλεση των διαδικασιών.

Σχήμα 73 Αρχικοποίηση μοντέλου Segment Anything και ελέγχου συσκευής του μοντέλου

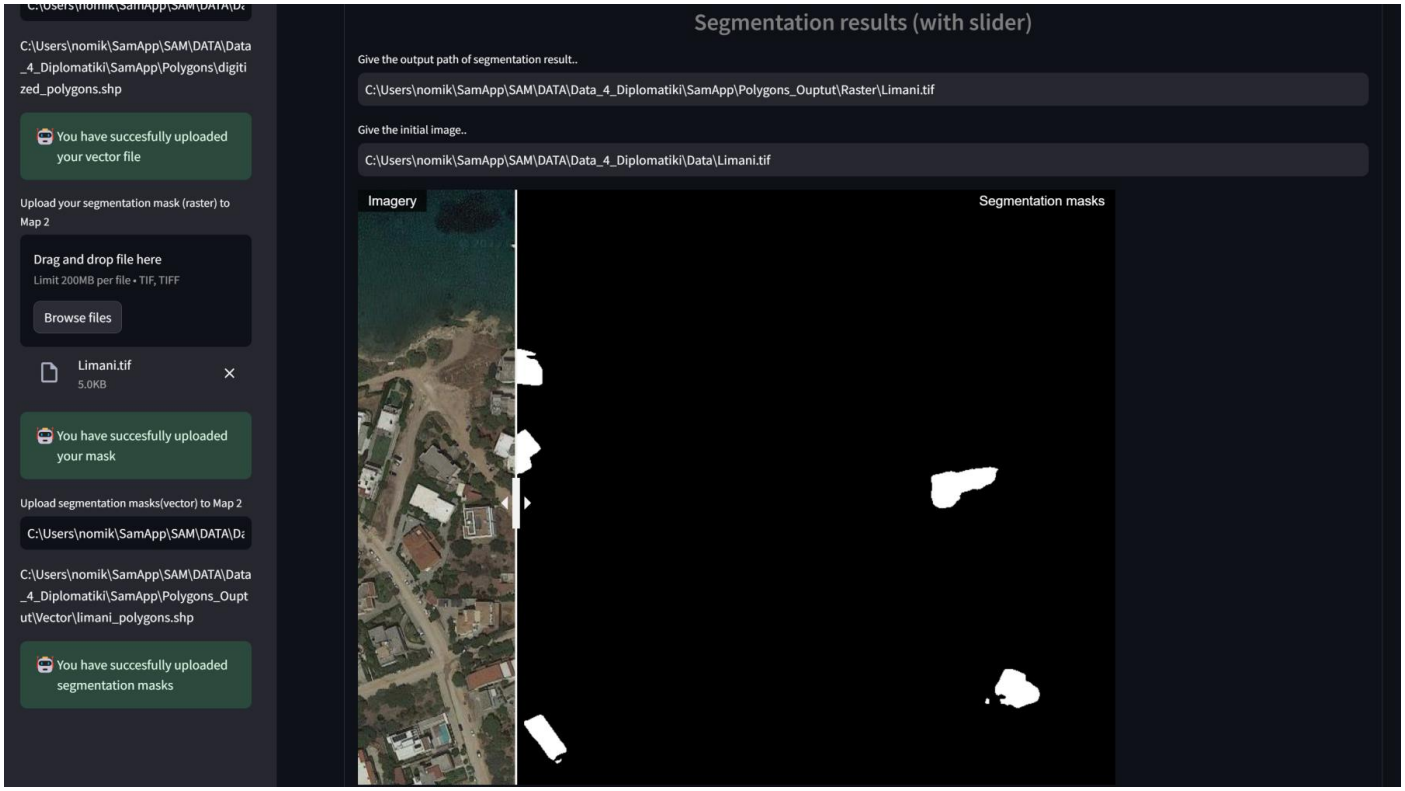
Μετά ο χρήστης εκτελεί τη διαδικασία της πρόβλεψης του μοντέλου δηλαδή κατάτμησης της εικόνας με βάση τα πλαίσια οριοθέτησης και ορίζει τη διαδρομή αρχείου για την αποθήκευση των αποτελεσμάτων τοπικά στον υπολογιστή. Τα αποτελέσματα (μάσκες τμηματοποίησης) αποθηκεύονται σε μορφή εικόνας αρχικά και δύναται να μετατραπούν και σε πολύγωνα.



Σχήμα 74 Πρόβλεψη μοντέλου Segment Anything

Τα αποτελέσματα της τμηματοποίησης της εικόνας εισάγονται στο δεύτερο διαδραστικό χάρτη. Επιπλέον έχει υλοποιηθεί ένα επιπρόσθετο εργαλείο οπτικοποίησης των αποτελεσμάτων τμηματοποίησης στο τέλος της σελίδας τύπου slider.

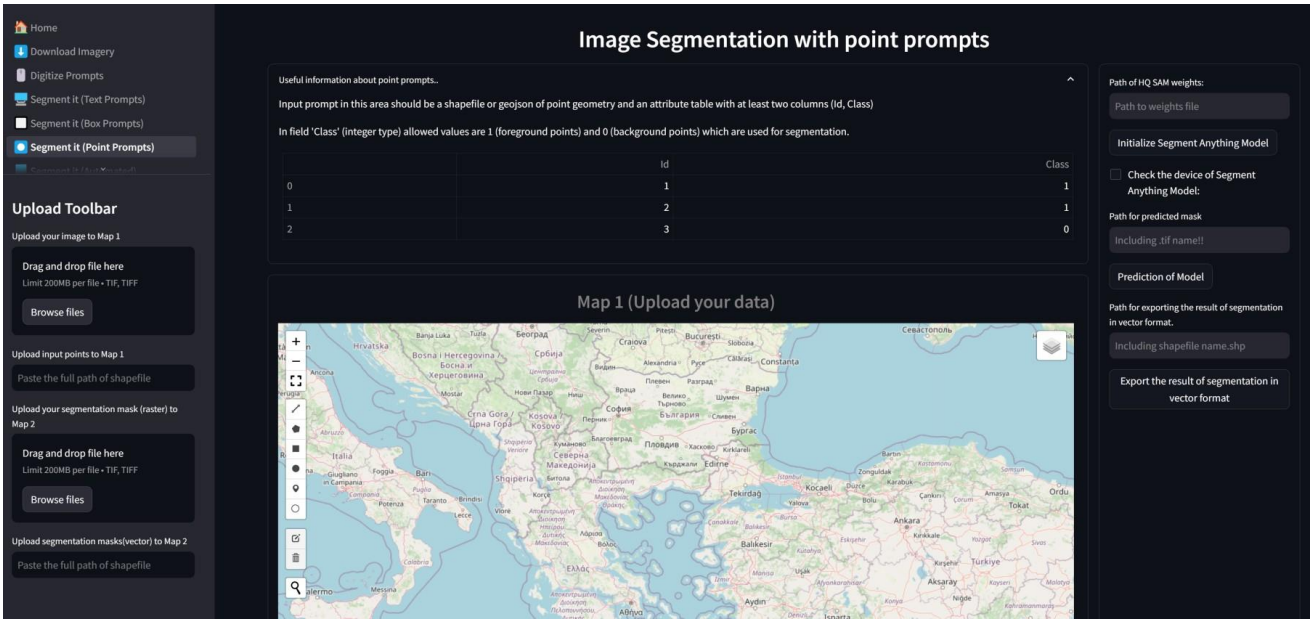
Σχήμα 75 Προβολή αποτελεσμάτων τμηματοποίησης στο δεύτερο διαδραστικό χάρτη



Σχήμα 76 Οπτικοποίηση αποτελεσμάτων μέσω εργαλείου slider

4.4.5 ΣΕΛΙΔΑ ΚΑΤΑΤΜΗΣΗΣ ΕΙΚΟΝΑΣ ΜΕ ΧΡΗΣΗ ΣΗΜΕΙΑΚΩΝ ΠΡΟΤΡΟΠΩΝ

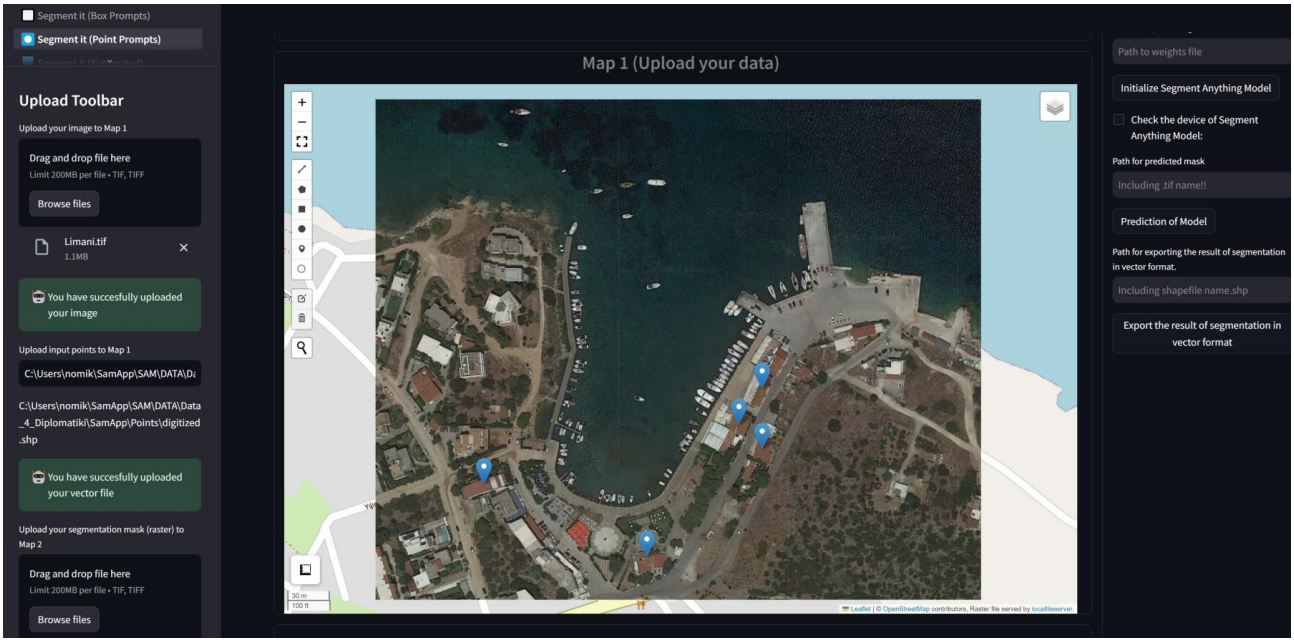
Η συγκεκριμένη σελίδα εφαρμόζει τη μεθοδολογία κατάτμησης της εικόνας με χρήση σημειακών προτροπών για το μοντέλο. Στην αρχή της σελίδας υπάρχει ένα ενημερωτικό πλαίσιο με τη δομή που πρέπει να έχει ο πίνακας χαρακτηριστικών (attribute table) των σημείων που θα χρησιμοποιηθούν ως προτροπή στο μοντέλο. Τα σημεία που εισέρχονται στην εφαρμογή είναι μορφότυπου shapefile. Ο πίνακας χαρακτηριστικών του σημειακού σετ δεδομένων πρέπει να περιέχει στήλη με όνομα «Class» που υποδηλώνει αν το σημείο ανήκει στο προσκήνιο της εικόνας (foreground) ή στο υπόβαθρο της εικόνας (background).



Σχήμα 77 Αρχική σελίδα τμηματοποίησης της εικόνας με σημειακές προτροπές

Ο χρήστης έχει τη δυνατότητα εισαγωγής της απεικόνισης της επιλογής του καθώς και των σημείων (points prompts), με βάση των οποίων πραγματοποιείται η κατάτμηση.

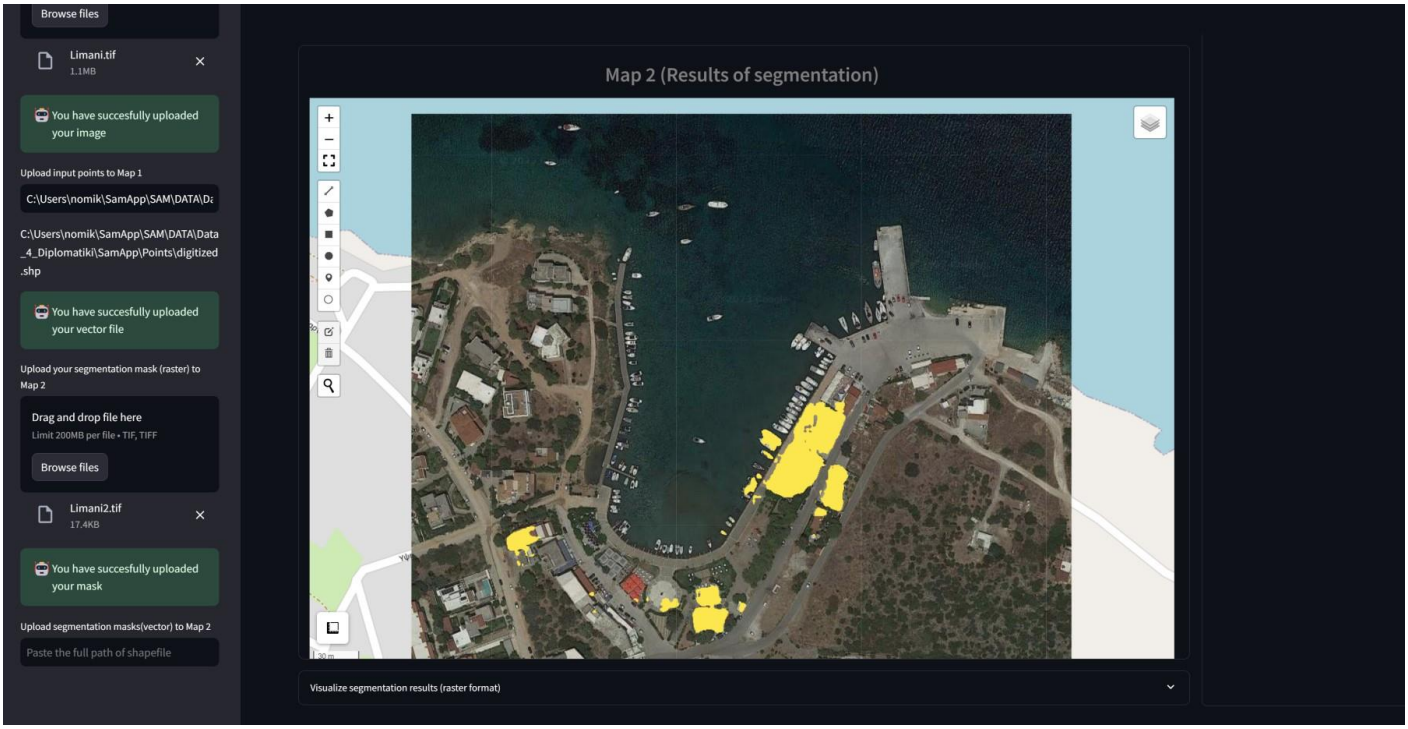
Η δομή της συγκεκριμένης σελίδας ακολουθεί την αντίστοιχη δομή της σελίδας της τμηματοποίησης με πλαίσια οριοθέτησης. Ο χρήστης εισάγει τα δεδομένα εισόδου και τα προβάλλει μέσω του πρώτου διαδραστικού χάρτη. Ύστερα αρχικοποιεί το μοντέλο Segment Anything και ορίζει τα προ-εκπαιδευμένα βάρη του. Το μοντέλο τμηματοποιεί την εικόνα εισόδου στις περιοχές που είναι τα σημεία εισόδου και εξάγει τις μάσκες σε επίπεδο εικόνας αλλά και σε διανυσματικό επίπεδο.



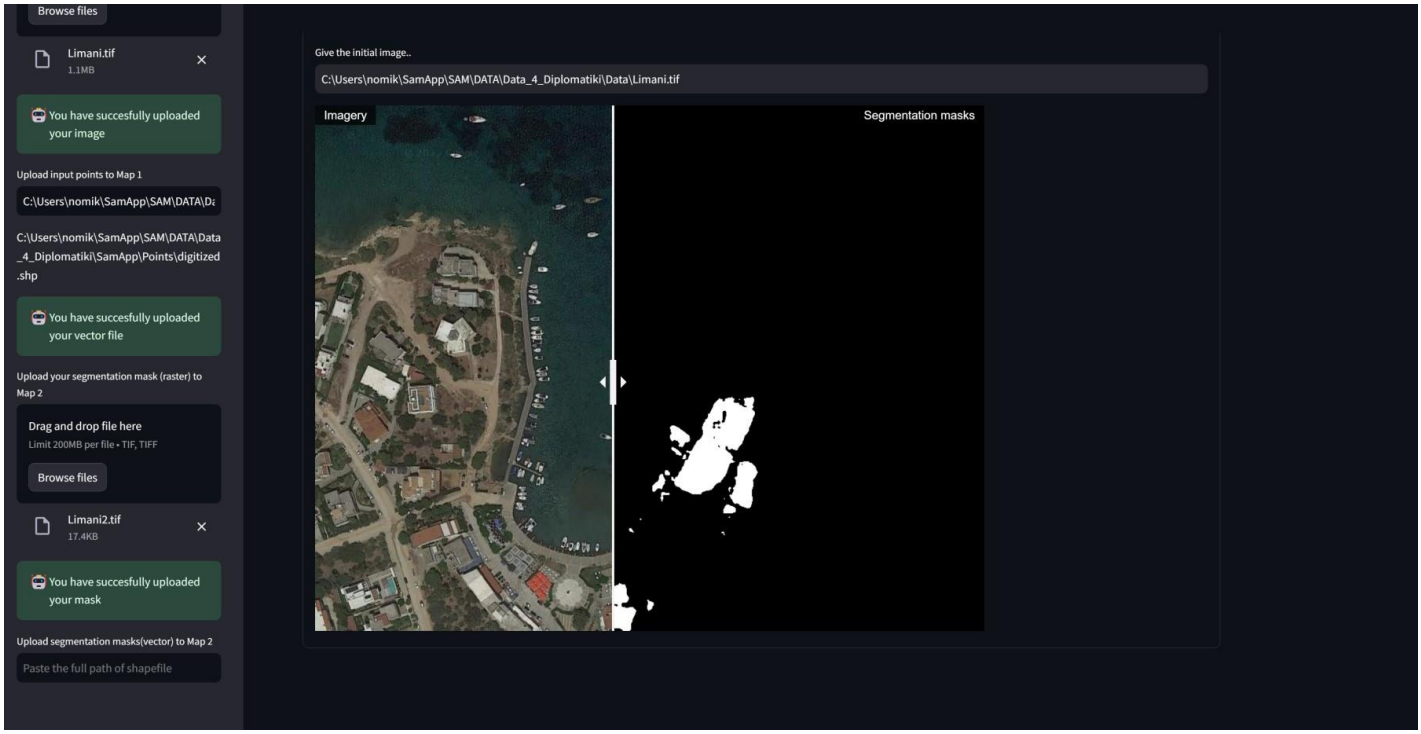
Σχήμα 78 Εισαγωγή δορυφορικής απεικόνισης και σημειακών προτροπών

Σχήμα 79 Εισαγωγή βαρών και αρχικοποίηση μοντέλου Segment Anything

Ο χρήστης προβάλλει τα αποτελέσματα της τμηματοποίησης στο δεύτερο διαδραστικό χάρτη της εφαρμογής για προεπισκόπηση των τελικών αποτελεσμάτων. Τέλος έχει υλοποιηθεί και πρόσθετο εργαλείο οπτικοποίησης των αποτελεσμάτων μορφής slider.



Σχήμα 80 Προβολή αποτελεσμάτων τμηματοποίησης με βάση σημειακές προτροπές στο μοντέλο



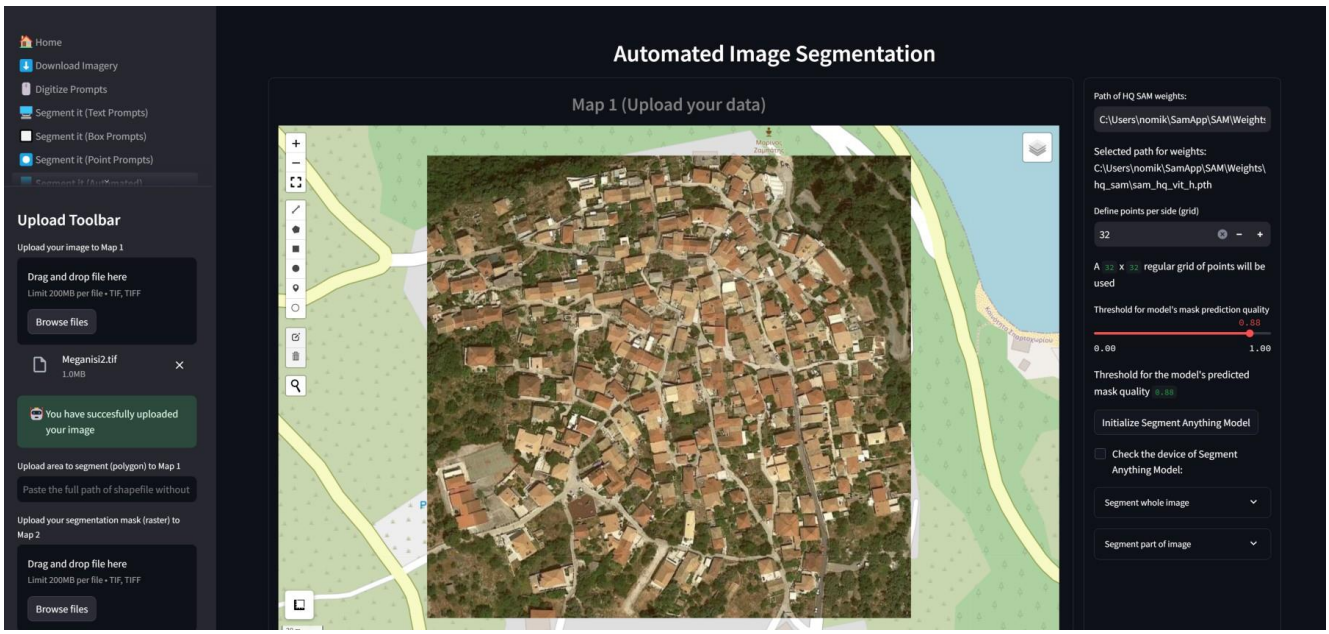
Σχήμα 81 Οπτικοποίηση αποτελεσμάτων τμηματοποίησης μέσω εργαλείου slider

4.4.6 ΣΕΛΙΔΑ ΑΥΤΟΜΑΤΗΣ ΠΑΡΑΓΩΓΗΣ ΜΑΣΚΩΝ

Στη τελευταία σελίδα της εφαρμογής έχει υλοποιηθεί η διαδικασία της αυτόματης παραγωγής μασκών από την εικόνα εισόδου, διαδικασία που αποτελεί παραλλαγή της instance segmentation. Όπως έχει αναλυθεί σε προηγούμενη ενότητα, στην συγκεκριμένη διαδικασία δημιουργείται ένα πλέγμα σημείων με βάση το οποίο προβλέπονται μάσκες για το σύνολο των οντοτήτων στην εικόνα εισόδου.

Με την ίδια λογική έχει υλοποιηθεί και η διαδικασία της αυτόματης παραγωγής μασκών στη συγκεκριμένη εφαρμογή με μια μικρή τροποποίηση. Έχει προστεθεί μια επιπρόσθετη λειτουργία στη παρούσα μεθοδολογία. Συγκεκριμένα, ο χρήστης δύναται εκτός από την εικόνα εισόδου να ορίσει και περιοχή ενδιαφέροντος εντός της εικόνας στο οποίο θα εκτελεστεί αποκλειστικά αυτόματη παραγωγή μασκών. Έτσι η διαδικασία της κατάτμησης θα εκτελεστεί σε ένα υποσύνολο της εικόνας και όχι σε όλο το εύρος της.

Ο χρήστης εκτός από την εικόνα εισόδου και το πολύγωνο της περιοχής ενδιαφέροντος (προαιρετική εισαγωγή) πρέπει να ορίσει τις διαστάσεις του πλέγματος σημείων με προκαθορισμένη τιμή το 32*32 (32 σημεία ομοιόμορφα κατανεμημένα σε κάθε πλευρά της εικόνας) αλλά και τη τιμή-κατώφλι για τη ποιότητα της προβλεπόμενης μάσκας. Αφού ορίσει όλες τις παραμέτρους του μοντέλου και εισάγει και τα βάρη του προ-εκπαιδευμένου μοντέλου SAM, εκτελείται η αυτοματοποιημένη παραγωγή των μασκών.

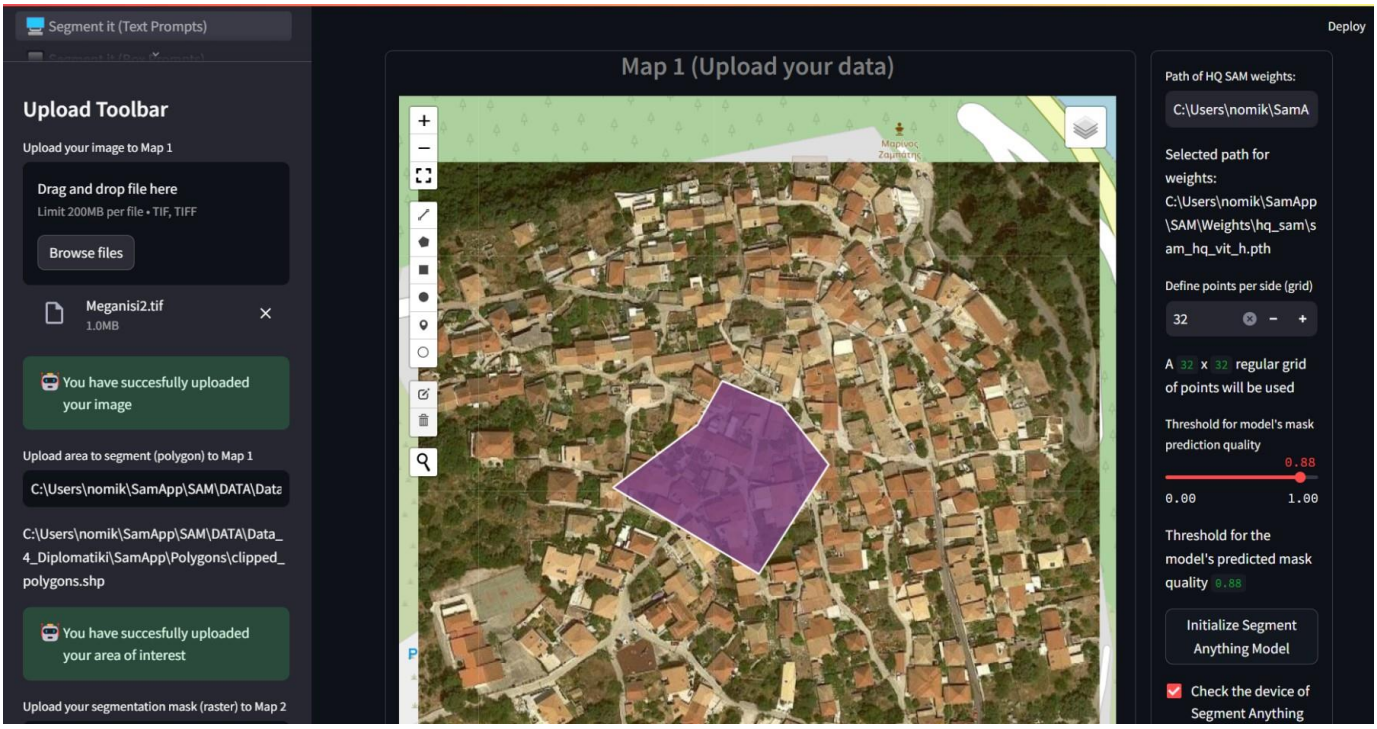


Σχήμα 82 Προβολή εικόνας εισόδου, ορισμός διαστάσεων πλέγματος και κατωφλιού για ποιότητα μάσκας

Αφού εκτελεστεί η πρόβλεψη του μοντέλου και παραχθούν οι μάσκες της εικόνας εισόδου σε μορφότυπο εικόνας αλλά και διανύσματος, εισάγονται στο δεύτερο διαδραστικό χάρτη για προεπισκόπηση των αποτελεσμάτων. (Σχήμα 83)

Σχήμα 83 Αποτέλεσμα κατάτμησης εικόνας εισόδου με πλέγμα σημείων (32*32)

Ακολουθεί το παράδειγμα εφαρμογής της ίδιας μεθοδολογίας αλλά με επιπρόσθετη είσοδο από τον χρήστη πολυγώνου ενδιαφέροντος για κατάτμηση εντός αυτού. Η διαδικασία της κατάτμησης παρουσιάζεται παρακάτω.



Σχήμα 84 Πολύγωνο ενδιαφέροντος για αυτόματη παραγωγή μασκών εντός αυτού

Σχήμα 85 Αποτέλεσμα κατάτμησης εντός πολυγώνου ενδιαφέροντος

4.4.7 ΣΥΜΠΕΡΑΣΜΑΤΑ ΥΛΟΠΟΙΗΣΗΣ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΑ ΑΥΤΟΝΟΜΗΣ ΕΦΑΡΜΟΓΗΣ ΓΙΑ ΑΝΙΧΝΕΥΣΗ ΚΑΙ ΕΞΑΓΩΓΗ ΟΝΤΟΤΗΤΩΝ ΑΠΟ ΔΟΡΥΦΟΡΙΚΕΣ ΑΠΕΙΚΟΝΙΣΕΙΣ

Η συγκεκριμένη εφαρμογή μέσω της ενσωμάτωσης δυο μοντέλων, ενός μοντέλου τμηματοποίησης με τη συμβολή προτροπών (promptable segmentation) και ενός μοντέλου ανίχνευσης οντοτήτων (object detection), έχει ως σκοπό την εξαγωγή οντοτήτων από τηλεπισκοπικές απεικονίσεις.

Μέσω πολλών δοκιμών και πειραματικών προσεγγίσεων, εξάγεται το συμπέρασμα ότι το Segment Anything μοντέλο παρουσιάζει κάποιες ιδιαιτερότητες και περιορισμούς στην διαχείριση δορυφορικών δεδομένων. Το χαρακτηριστικό του zero-shot segmentation, δηλαδή να τμηματοποιεί εικόνες και σετ δεδομένων που δεν έχει εκπαιδευτεί, του επιτρέπει εξαγωγή μασκών με σχετική ακρίβεια σε δορυφορικά δεδομένα. Αντίστοιχα, το ίδιο ισχύει και για την ανίχνευση οντοτήτων του Grounding Dino μοντέλου. Όμως σε οντότητες εξειδικευμένου ενδιαφέροντος τα αποτελέσματα χρήζουν βελτίωσης. Μέχρι στιγμής έχουν χρησιμοποιηθεί αποκλειστικά οι προ-εκπαιδευμένοι παράμετροι του SAM.

Στο επόμενο κεφάλαιο, επιχειρείται η διαδικασία της περαιτέρω εκπαίδευσης του SAM με σκοπό το μοντέλο να αποκτήσει τη δυνατότητα μέσω συγκεκριμένης μεθοδολογίας να τμηματοποιεί οντότητες συγκεκριμένου ενδιαφέροντος.

ΚΕΦΑΛΑΙΟ 5: ΜΕΘΟΔΟΛΟΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΚΠΑΙΔΕΥΣΗΣ ΤΟΥ ΠΡΟ-ΕΚΠΑΙΔΕΥΜΕΝΟΥ ΜΟΝΤΕΛΟΥ ΚΑΤΑΤΜΗΣΗΣ ΕΙΚΟΝΩΝ ΜΕΣΩ ΠΡΟΤΡΟΠΩΝ, SEGMENT ANYTHING

Στο κεφάλαιο αυτό, επιχειρείται η περαιτέρω εκπαίδευση του μοντέλου Segment Anything για κάλυψη εξειδικευμένων σκοπών σε αντικείμενα υπολογιστής όρασης. Μέχρι στιγμής στη παρούσα διπλωματική εργασία, όλες οι μεθοδολογίες που περιελάμβαναν το μοντέλο Segment Anything χρησιμοποιούσαν τις προ-εκπαιδευμένες παραμέτρους του μοντέλου αυτούσιες χωρίς περαιτέρω τροποποιήσεις. Όμως για την εξαγωγή οντοτήτων ειδικού ενδιαφέροντος από δορυφορικές απεικονίσεις απαιτείται περαιτέρω εκπαίδευση του μοντέλου βαθιάς μάθησης με χρήση συγκεκριμένων τεχνικών. Όλη η σχετική διαδικασία εκπαίδευσης καθώς και αξιολόγησης του νέου πλέον μοντέλου θα παρουσιαστεί στη παρούσα ενότητα.

5.1 ΠΑΡΟΥΣΙΑΣΗ ΑΝΤΙΚΕΙΜΕΝΟΥ ΜΕΘΟΔΟΛΟΓΙΑΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Το αντικείμενο της εργασίας που θα αναπτυχθεί στη παρούσα ενότητα είναι η περαιτέρω εκπαίδευση (fine tuning) του Segment Anything για ανίχνευση και εξαγωγή οντοτήτων ειδικού ενδιαφέροντος από δορυφορικές απεικονίσεις. Οι οντότητες που έχουν επιλεχθεί για ανίχνευση και εξαγωγή από δορυφορικές εικόνες είναι περιοχές με ομπρέλες σε παραλίες.

Το προ-εκπαιδευμένο μοντέλο SAM σε συνδυασμό με το μοντέλο ανίχνευσης Grounding Dino αδυνατεί να ανιχνεύσει τις συγκεκριμένες οντότητες, να δημιουργήσει πλαίσια οριοθέτησης και να εξάγει τις μάσκες τους. Η συγκεκριμένη διαδικασία εφαρμόστηκε σε άλλες οντότητες ενδιαφέροντος όπως μάσκες κτιρίων και δένδρων σε δορυφορικές εικόνες και παρουσιάστηκε σε προηγούμενη ενότητα.

Για την υλοποίηση της συγκεκριμένης εργασίας, παρουσιάζεται αναλυτικά η μεθοδολογία της περαιτέρω εκπαίδευσης των παραμέτρων του μοντέλου SAM. Για το συγκεκριμένο σκοπό, απαιτείται σετ δορυφορικών δεδομένων με βάση το οποίο θα πραγματοποιηθεί η επιπρόσθετη εκπαίδευση του μοντέλου.

Τα δορυφορικά δεδομένα που χρησιμοποιούνται προέρχονται από το Εργαστήριο Τηλεπισκόπησης του ΕΜΠ και αφορούν δορυφορικές απεικονίσεις παραλιών νησιών όπως η Πάρος, η Μύκονος, η Σαντορίνη αλλά και παραλίες της Χαλκιδικής. Όλες οι απεικονίσεις είναι υψηλής διακριτικής χωρικής ικανότητας (διαστάσεις εικονοστοιχείου 0.3m και 0.5m) και προέρχονται από λήψεις δορυφόρων Pleiades Neo και Maxar.

Εκτός από τις δορυφορικές απεικονίσεις για την εκπαίδευση ενός μοντέλου κατάτμησης όπως είναι το SAM απαιτούνται και δεδομένα μασκών που αντιστοιχούν σε αυτές. Τα δεδομένα μασκών ή αλλιώς groundtruth data αποτελούν δυαδικές εικόνες (binary raster data), οι τιμές των οποίων διαχωρίζονται σε 0 και 1. Τα εικονοστοιχεία με τιμή 1 αντιστοιχούν σε μάσκες οντοτήτων ενδιαφέροντος (ομπρέλες στη παρούσα φάση) και τα εικονοστοιχεία με τιμή 0 αποτελούν την υπόλοιπη εικόνα.

Συγκεκριμένα το σετ δεδομένων απαρτίζεται από 54 δορυφορικές εικόνες με ποικίλες διαστάσεις (3204 *3204, 2544 *2544, 3333*3333 και 2000*2000) υψηλής χωρικής διακριτικής ικανότητας (0.5m και 0.3m) οι οποίες αποτελούνται από τρία φασματικά κανάλια (RGB). Ομοίως στο σετ δεδομένων υπάρχουν 54 εικόνες αντίστοιχων διαστάσεων, ενός καναλιού και δυαδικής μορφής που εξυπηρετούν το ρόλο των μασκών (groundtruth data).

5.2 ΒΗΜΑ 1: ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Όπως αναφέρθηκε στη προηγούμενη υποενότητα οι δορυφορικές απεικονίσεις έχουν υψηλή χωρική διακριτική ικανότητα και μεγάλες διαστάσεις. Αυτό δυσχεραίνει τη διαδικασία εκπαίδευσης του μοντέλου.

Γι αυτό το σκοπό τα δεδομένα της εκπαίδευσης του μοντέλου (εικόνες και μάσκες) θα μετασχηματιστούν σε μικρότερες διαστάσεις και συγκεκριμένα σε διαστάσεις των 256*256 εικονοστοιχείων. Τα συγκεκριμένα υποσύνολα των αρχικών εικόνων ονομάζονται *patches.*

Σε κώδικα Python υλοποιούνται δύο συναρτήσεις οι οποίες διαβάζουν τις εικόνες και τις μάσκες αντίστοιχα και τις περικόπτουν σε υποσύνολα των 256*256 εικονοστοιχείων. Τα υποσύνολα (patches) αποθηκεύονται αντίστοιχα σε εικόνες και μάσκες. Οι εικόνες πλέον απαριθμούν στο σύνολο 5238 από 54 και το ίδιο συμβαίνει και με τις μάσκες.

Η επιλογή του μεγέθους 256x256 εικονοστοιχεία είναι σημαντική για διάφορους λόγους, ειδικά όταν υπάρχουν μεγάλες δορυφορικές εικόνες και προετοιμάζονται τα δεδομένα για εκπαίδευση του μοντέλου. Συγκεκριμένα, ο διαχωρισμός μεγάλων δορυφορικών εικόνων σε μικρότερες διευκολύνει τη διαχείριση τους και την επεξεργασία τους με τις υπάρχουσες δυνατότητες υλικού. Η επεξεργασία μεγάλων εικόνων απευθείας μπορεί να είναι υπολογιστικά δαπανηρή και χρονοβόρα. Επιπλέον, τα μικρότερα patches μπορούν να βοηθήσουν το μοντέλο να μάθει τοπικά χαρακτηριστικά πιο αποτελεσματικά. Για παράδειγμα, μπορεί να εστιάσει στην ανίχνευση συγκεκριμένων αντικειμένων μέσα σε κάθε patch, βελτιώνοντας τη συνολική ακρίβεια της τμηματοποίησης.

Επόμενο βήμα αποτελεί η εισαγωγή των κομμένων πλέον εικόνων και μασκών (patches) σε δύο πίνακες (arrays) της βιβλιοθήκης NumPy της γλώσσας Python. Η εισαγωγή των δεδομένων σε πίνακες διευκολύνει την μετέπειτα διαδικασία της εκπαίδευσης. Έτσι, δημιουργούνται δύο πίνακες, ένας για τις δορυφορικές εικόνες με διάσταση (5238, 256, 256, 3) και ένας για τις μάσκες με διάσταση (5238, 256, 256). Ο πίνακας των εικόνων έχει το 3 στο σχήμα του λόγω του αριθμού των καναλιών (RGB).

Επόμενο βήμα της επεξεργασίας των δεδομένων είναι η αφαίρεση των εικόνων από τον αντίστοιχο πίνακα που έχουν έστω ένα εικονοστοιχείο με τιμή 0. Η αφαίρεση αυτών των εικόνων κρίνεται απαραίτητη για αποφυγή σφαλμάτων κατά τη διάρκεια της εκπαίδευσης. Για τις εικόνες που αφαιρούνται, αφαιρούνται και οι αντίστοιχες μάσκες στο πίνακα των μασκών. Δημιουργούνται δύο νέοι πίνακες με φιλτραρισμένες πλέον εικόνες και μάσκες ως προς το χαρακτηριστικό που αναφέρθηκε προηγουμένως. Η διάσταση του νέου πίνακα εικόνων είναι (2868, 256, 256, 3) ενώ του πίνακα μασκών (2868, 256, 256).

Στη συνέχεια, τα δεδομένα φιλτράρονται επιπλέον αφαιρώντας τις κενές μάσκες, δηλαδή τις μάσκες που δεν έχουν έστω ένα εικονοστοιχείο ίσο με το 1. Για τις μάσκες που αφαιρούνται, αφαιρούνται και οι αντίστοιχες εικόνες. Η διαδικασία πραγματοποιείται για την αποφυγή σφαλμάτων κατά τη διάρκεια της εκπαίδευσης. Οι νέες διαστάσεις των δύο πινάκων είναι (219, 256, 256, 3) και (219, 256, 256) για εικόνες και μάσκες αντίστοιχα. Το επόμενο βήμα της διαδικασίας είναι ο διαχωρισμός των δεδομένων του μοντέλου σε δεδομένα εκπαίδευσης (train data) και σε δεδομένα ελέγχου (test data).

5.3 ΒΗΜΑ 2: ΔΙΑΧΩΡΙΣΜΟΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ

Επόμενο βήμα της διαδικασίας αποτελεί ο διαχωρισμός των δεδομένων (εικόνων και μασκών) σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Ο συγκεκριμένος διαχωρισμός είναι κρίσιμος για την αξιολόγηση της απόδοσης του μοντέλου. Τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση του μοντέλου, ενώ τα δεδομένα ελέγχου χρησιμοποιούνται για έλεγχο του βαθμού γενίκευσης του μοντέλου σε νέα, άγνωστα δεδομένα.

Επιπλέον ο διαχωρισμός των δεδομένων στις δύο κατηγορίες συμβάλλει στη αποφυγή της υπερπροσαρμογής (overfitting) του μοντέλου. Ο όρος overfitting είναι ένα σύνηθες πρόβλημα στη βαθιά μάθηση και παρατηρείται όταν ένα μοντέλο μαθαίνει πολύ καλά τα δεδομένα εκπαίδευσης αλλά αποτυγχάνει να αποδώσει καλά σε νέα δεδομένα που δεν έχει δει κατά τη διάρκεια της εκπαίδευσης.

Κατά το διαχωρισμό των δεδομένων ο οποίος έχει υλοποιηθεί προγραμματιστικά, έχει προστεθεί στη διαδικασία η λειτουργία της τυχαίας κατανομής (shuffling). Η συγκεκριμένη λειτουργία εξασφαλίζει τη τυχαία αναδιανομή των δεδομένων πριν το διαχωρισμό τους και τη βεβαιότητα ότι αποτελούν αντιπροσωπευτικά δείγματα, μειώνοντας τη πιθανότητα να υπάρξουν συστηματικές διαφορές μεταξύ των δύο συνόλων.

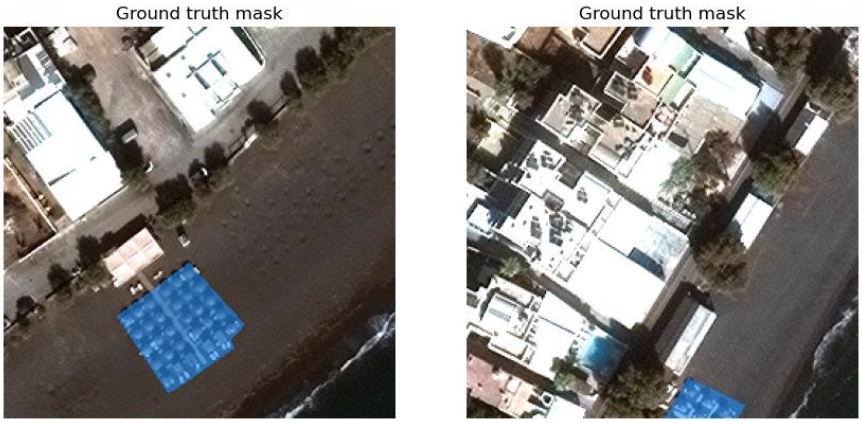
Από το σύνολο των δεδομένων, το 80% θα οριστεί ως δεδομένα εκπαίδευσης και το 20% ως δεδομένα ελέγχου, ποσοστά συνήθη στη εκπαίδευση μοντέλων. Τέλος δημιουργούνται 4 πίνακες, 2 για κάθε μια κατηγορία με τις παρακάτω διαστάσεις:

- Πίνακας Εικόνων Εκπαίδευσης με διάσταση (175, 256, 256, 3)
- Πίνακας Μασκών Εκπαίδευσης με διάσταση (175, 256, 256)
- Πίνακας Εικόνων Ελέγχου με διάσταση (44, 256, 256, 3)
- Πίνακας Μασκών Ελέγχου με διάσταση (44, 256, 256)

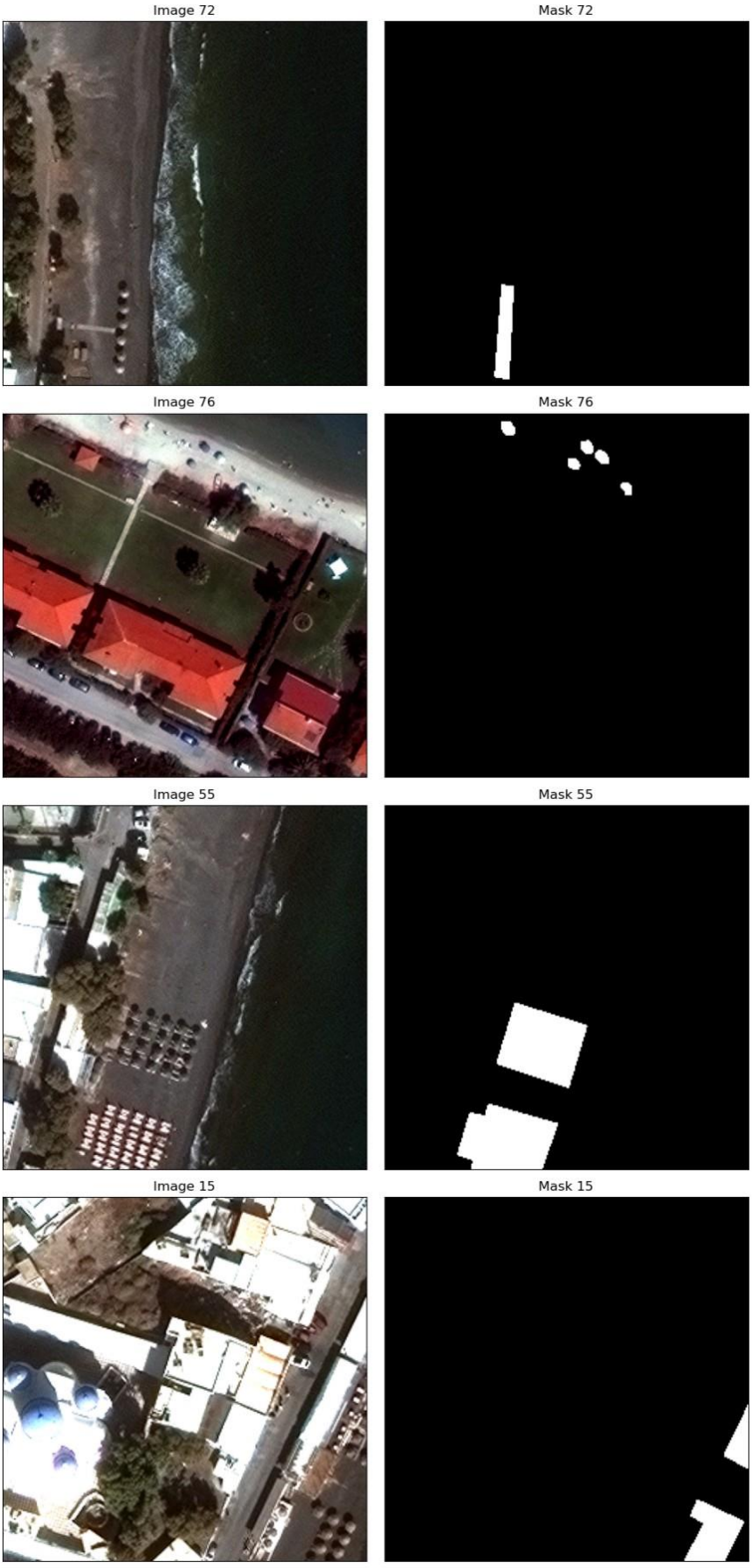
5.4 ΒΗΜΑ 3: ΟΠΤΙΚΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ ΕΚΠΑΙΔΕΥΣΗΣ ΚΑΙ ΕΛΕΓΧΟΥ

Επόμενο βήμα αποτελεί η φόρτωση των πινάκων της προηγούμενης υποενότητας σε ένα μορφότυπο στη γλώσσα Python που διευκολύνει τη διαχείριση τους για τα επόμενα βήματα της διαδικασίας, γνωστός ως «dictionary». Ο συγκεκριμένος τύπος δεδομένων περιέχει δύο στήλες, μια για τις εικόνες και την άλλη για τις μάσκες. Έτσι δημιουργούνται δύο «dictionaries», ένα για εκπαίδευση και ένα για έλεγχο.

Έπειτα, οπτικοποιούνται δειγματοληπτικά τα δεδομένα εκπαίδευσης και ελέγχου για να ελεγχθούν ότι έχουν εισαχθεί και μετασχηματιστεί ορθά. Παρακάτω ακολουθούν σχήματα με τις οπτικοποιήσεις των δεδομένων του μοντέλου.



Σχήμα 86 Δεδομένα εκπαίδευσης (αριστερά) Δεδομένα Ελέγχου (δεξιά)



Σχήμα 87 Οπτικοποίηση Δεδομένων Εκπαίδευσης (εικόνες και αντίστοιχες μάσκες)

Σχήμα 88 Οπτικοποίηση Δεδομένων Ελέγχου (εικόνες και αντίστοιχες μάσκες)

5.5 ΒΗΜΑ 4: ΑΡΧΙΚΟΠΟΙΗΣΗ SEGMENT ANYTHING ΜΟΝΤΕΛΟΥ ΚΑΙ ΤΕΛΙΚΗ ΔΙΑΜΟΡΦΩΣΗ ΔΕΔΟΜΕΝΩΝ

Για την εκπαίδευση του μοντέλου SAM, θα χρησιμοποιηθεί η ιδέα της αυτόματης παραγωγής μασκών σε μια εικόνα .

Αρχικά θα υλοποιηθεί προγραμματιστικά μια συνάρτηση που να δέχεται ως όρισμα την εκάστοτε εικόνα, να διαβάζει τις διαστάσεις της και να δημιουργεί ένα ομοιόμορφο πλέγμα στο εύρος της. Το πλέγμα των σημείων που θα δημιουργηθεί για την εκάστοτε εικόνα θα διαδραματίσει το ρόλο της γεωμετρικής προτροπής του μοντέλου κατά την εκπαίδευση του.

Επιπρόσθετα υλοποιείται μια προσαρμοσμένη «κλάση» *SAMDataset* (Python class) η οποία διαμορφώνει τη λειτουργικότητα για τη παροχή στο μοντέλο των εικόνων, των αντίστοιχων μασκών και των πλεγμάτων σημείων κατά την εκπαίδευση του. Μέσα στη συγκεκριμένη κλάση εδράζεται και η συνάρτηση που περιγράφεται προηγουμένως για τη δημιουργία του πλέγματος σημείων. Η κλάση δέχεται ως όρισμα το «dictionary» που περιέχει τις εικόνες και τις μάσκες καθώς και ένα επεξεργαστή ο οποίος είναι ο αντίστοιχος του Segment Anything. Επιστρέφει ένα «dictionary» που περιέχει τις εικόνες , τις αληθείς μάσκες και τις προτροπές του πλέγματος σημείων. Με αυτό τον τρόπο η κλάση *SAMDataset* προετοιμάζει τα δεδομένα με τέτοιο τρόπο ώστε να είναι κατάλληλα για την εκπαίδευση και αξιολόγηση του μοντέλου SAM.

Αφού χρησιμοποιηθεί η κλάση *SAMDataset* για τη δημιουργία δύο σετ δεδομένων (ένα για εκπαίδευση και ένα για έλεγχο) με τα απαραίτητα στοιχεία για την εκπαίδευση και αξιολόγηση του μοντέλου (εικόνες, μάσκες, πλέγματα σημείων), αυτά μεταφορτώνονται με τη σειρά τους σε μια κλάση της βιβλιοθήκης PyTorch, τη «DataLoader». Δημιουργούνται δυο οντότητες (instances) της κλάσης DataLoader, μια για εκπαίδευση και μια για έλεγχο. Αυτή είναι και η τελική διαμόρφωση των δεδομένων πριν τις διαδικασίες εκπαίδευσης και αξιολόγησης του μοντέλου.

Επόμενο βήμα είναι η αρχικοποίηση του μοντέλου Segment Anything με τις προ-εκπαιδευμένες παραμέτρους. Η αρχιτεκτονική του SAM όπως έχει ήδη παρουσιαστεί διακρίνεται σε τρία βασικά μέρη: τον κωδικοποιητή εικόνας (image encoder), τον κωδικοποιητή προτροπών (prompt encoder) και τον αποκωδικοποιητή μασκών (mask decoder). Κάθε μέρος του μοντέλου , έχει τις δικές του εκπαιδευμένες παραμέτρους. Στη παρούσα μεθοδολογία επιλέγεται η εκπαίδευση του μοντέλου μόνο στη περιοχή του αποκωδικοποιητή μάσκας. Η συγκεκριμένη επιλογή οφείλεται στο γεγονός ότι αποτελεί το μικρότερο μέρος του μοντέλου, με τις λιγότερες παραμέτρους προς εκπαίδευση με αποτέλεσμα να διευκολύνεται η διαδικασία της εκπαίδευσης και της αξιολόγησης του μοντέλου. Επιπλέον, οι απαιτήσεις σε λογισμικό είναι μικρότερες σε σχέση με την εκπαίδευση και των

τριών μέρων του μοντέλου SAM. Για αυτό το λόγο, αδρανοποιούνται τα δύο μέρη του μοντέλου ως προς την εκπαίδευση τους.

5.6 ΒΗΜΑ 5: ΕΠΙΛΟΓΗ ΑΛΓΟΡΙΘΜΟΥ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΚΑΙ ΣΥΝΑΡΤΗΣΗΣ LOSS

Τελευταίο βήμα πριν την εκτέλεση της εκπαίδευσης και της αξιολόγησης του μοντέλου είναι ο καθορισμός της συνάρτησης loss και του αλγόριθμου βελτιστοποίησης (optimizer) που θα χρησιμοποιηθούν κατά τη διάρκεια της εκπαίδευσης και της αξιολόγησης του μοντέλου.

Οι αλγόριθμοι βελτιστοποίησης ουσιαστικά οδηγούν τα μοντέλα βαθιάς μάθησης σε καλύτερη ακρίβεια και απόδοση. Καθοδηγούν το μοντέλο στον τρόπο που προσαρμόζουν τις παραμέτρους του κατά τη διάρκεια της εκπαίδευσης, με στόχο την ελαχιστοποίηση της συνάρτησης loss. Ανάμεσα στους αλγόριθμους βελτιστοποίησης, ο Adam (Adaptive Moment Estimation) προτιμάται σε πολλές περιπτώσεις λόγω της αποτελεσματικότητας και της προσαρμοστικότητας του. Βασίζεται στα πλεονεκτήματα δύο άλλων αλγόριθμων βελτιστοποίησης, τον AdaGrad και τον RMSProp. Κυρίαρχο χαρακτηριστικό του συγκεκριμένου αλγορίθμου είναι το γεγονός ότι ο ρυθμός μάθησης (learning rate) προσαρμόζεται δυναμικά για κάθε μεμονωμένη παράμετρο του μοντέλου αντί την χρησιμοποίηση κάποιου ενιαίου ρυθμού μάθησης για όλο το μοντέλο. (Kingma and Ba, 2014). Στην παρούσα περίπτωση επιλέγεται ο Adam ως αλγόριθμος βελτιστοποίησης.

Ως συνάρτησης loss χρησιμοποιείται η συνάρτηση DiceCELoss η οποία είναι συνδυαστική συνάρτηση της DiceLoss και της Cross Entropy Loss. Η συγκεκριμένη συνάρτηση χρησιμοποιείται ευρέως σε προβλήματα κατάτμησης εικόνων.

5.7 ΒΗΜΑ 6: ΕΚΤΕΛΕΣΗ ΒΡΟΓΧΩΝ ΕΚΠΑΙΔΕΥΣΗΣ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Επόμενο βήμα αποτελεί η υλοποίηση των βρόγχων εκπαίδευσης (training loop) και αξιολόγησης (test loop) του μοντέλου. Η εκπαίδευση και η αξιολόγηση κάθε μοντέλου πραγματοποιείται για ένα προκαθορισμένο αριθμό εποχών (epochs). Στη συγκεκριμένη μεθοδολογία το μοντέλο θα εκπαιδευτεί και αξιολογηθεί για 60 εποχές.

Σε κάθε εποχή υλοποιείται ένας βρόγχος εκπαίδευσης και ένας βρόγχος αξιολόγησης.

Στο βρόγχο εκπαίδευσης (training loop) , το μοντέλο τίθεται αρχικά σε κατάσταση εκπαίδευσης, δέχεται ως όρισμα τα δεδομένα εκπαίδευσης και πραγματοποιεί μια πρόβλεψη παράγοντας τις μάσκες κατάτμησης για αυτά. (forward pass). Έπειτα οι μάσκες που προβλέφθηκαν συγκρίνονται με τις αληθείς μάσκες (groundtruth masks) και υπολογίζεται η συνάρτηση loss. Μέσω της αντίστροφης μετάδοσης (backward pass) ενημερώνονται οι παράμετροι του μοντέλου με σκοπό της ελαχιστοποίηση της συνάρτησης loss. Στο τέλος του βρόγχου εκπαίδευσης, καταγράφεται η τιμής της συνάρτησης loss (training loss) . Εκτός από τη τιμή της συνάρτησης loss, καταγράφονται και δύο μέτρα αξιολόγησης τα οποία θα αναλυθούν παρακάτω. Η καταγραφή και των τριών τιμών πραγματοποιείται για τη παρακολούθηση της προόδου εκπαίδευσης.

Στο βρόγχο αξιολόγησης (testing loop), το μοντέλο αρχικά τίθεται σε κατάσταση αξιολόγησης με σκοπό την αποτροπή των ενημερώσεων των παραμέτρων του. Έπειτα δέχεται ως όρισμα τα δεδομένα ελέγχου και πραγματοποιεί μια πρόβλεψη παράγοντας μάσκες κατάτμησης για αυτά. (forward pass). Έπειτα οι μάσκες που προβλέφθηκαν συγκρίνονται με τις αληθείς μάσκες (groundtruth masks) και υπολογίζεται η συνάρτηση loss. Στην αξιολόγηση δεν υπάρχει αντίστροφη μετάδοση (backward pass) όπως στο βρόγχο εκπαίδευσης. Στο τέλος του βρόγχου αξιολόγησης καταγράφεται η τιμή της συνάρτησης loss (testing loss). Εκτός από τη τιμή της συνάρτησης loss, καταγράφονται και δύο μέτρα αξιολόγησης τα οποία θα αναλυθούν παρακάτω. Η καταγραφή και των τριών τιμών πραγματοποιείται για τη παρακολούθηση της προόδου αξιολόγησης .

Η διαδικασία επαναλαμβάνεται για 60 εποχές και αντίστοιχα καταγράφονται τιμές συνάρτησης loss και μέτρων αξιολόγησης για εκπαίδευση και αξιολόγηση αντίστοιχα.

Ακολουθούν Σχήματα που απεικονίζουν τα αποτελέσματα της παραπάνω διαδικασίας για τις 5 πρώτες εποχές.

```
Epoch 1/5: 100%|██████████| 88/88 [00:40<00:00,  2.17it/s]

EPOCH: 1
Mean training loss: 1.109500210393559
Mean training Dice score: 0.08675681784639716
Mean training IoU score: 0.048064612471992875
Evaluating on test set: 100%|██████████| 22/22 [00:09<00:00,  2.33it/s]
Mean test loss: 0.9617030756040053
Mean test Dice score: 0.12814155185962012
Mean test IoU score: 0.07164301718909719
Epoch 2/5: 100%|██████████| 88/88 [00:39<00:00,  2.26it/s]

EPOCH: 2
Mean training loss: 0.8690812933174047
Mean training Dice score: 0.16613202168881136
Mean training IoU score: 0.09779059821379382
Evaluating on test set: 100%|██████████| 22/22 [00:09<00:00,  2.28it/s]
Mean test loss: 0.8718561286276038
Mean test Dice score: 0.18467955136227168
Mean test IoU score: 0.11014091430115514
Epoch 3/5: 100%|██████████| 88/88 [00:39<00:00,  2.21it/s]

EPOCH: 3
Mean training loss: 0.7664930359883741
Mean training Dice score: 0.29098163469089755
Mean training IoU score: 0.188535329280141
Evaluating on test set: 100%|██████████| 22/22 [00:09<00:00,  2.25it/s]
Mean test loss: 0.7968964956023477
Mean test Dice score: 0.27509801233695313
Mean test IoU score: 0.166461567775431
Epoch 4/5: 100%|██████████| 88/88 [00:40<00:00,  2.18it/s]

EPOCH: 4
Mean training loss: 0.6595872723582116
Mean training Dice score: 0.40892088412195576
Mean training IoU score: 0.28580317611884964
Evaluating on test set: 100%|██████████| 22/22 [00:09<00:00,  2.26it/s]
Mean test loss: 0.7376768426461653
Mean test Dice score: 0.3261968098072843
Mean test IoU score: 0.21340216463431716
Epoch 5/5: 100%|██████████| 88/88 [00:40<00:00,  2.16it/s]

EPOCH: 5
Mean training loss: 0.5821354465389793
Mean training Dice score: 0.5093171753793616
Mean training IoU score: 0.37635013954819774
Evaluating on test set: 100%|██████████| 22/22 [00:10<00:00,  2.18it/s]
Mean test loss: 0.6679109944538637
Mean test Dice score: 0.4790465361015363
Mean test IoU score: 0.34108908305113966
```

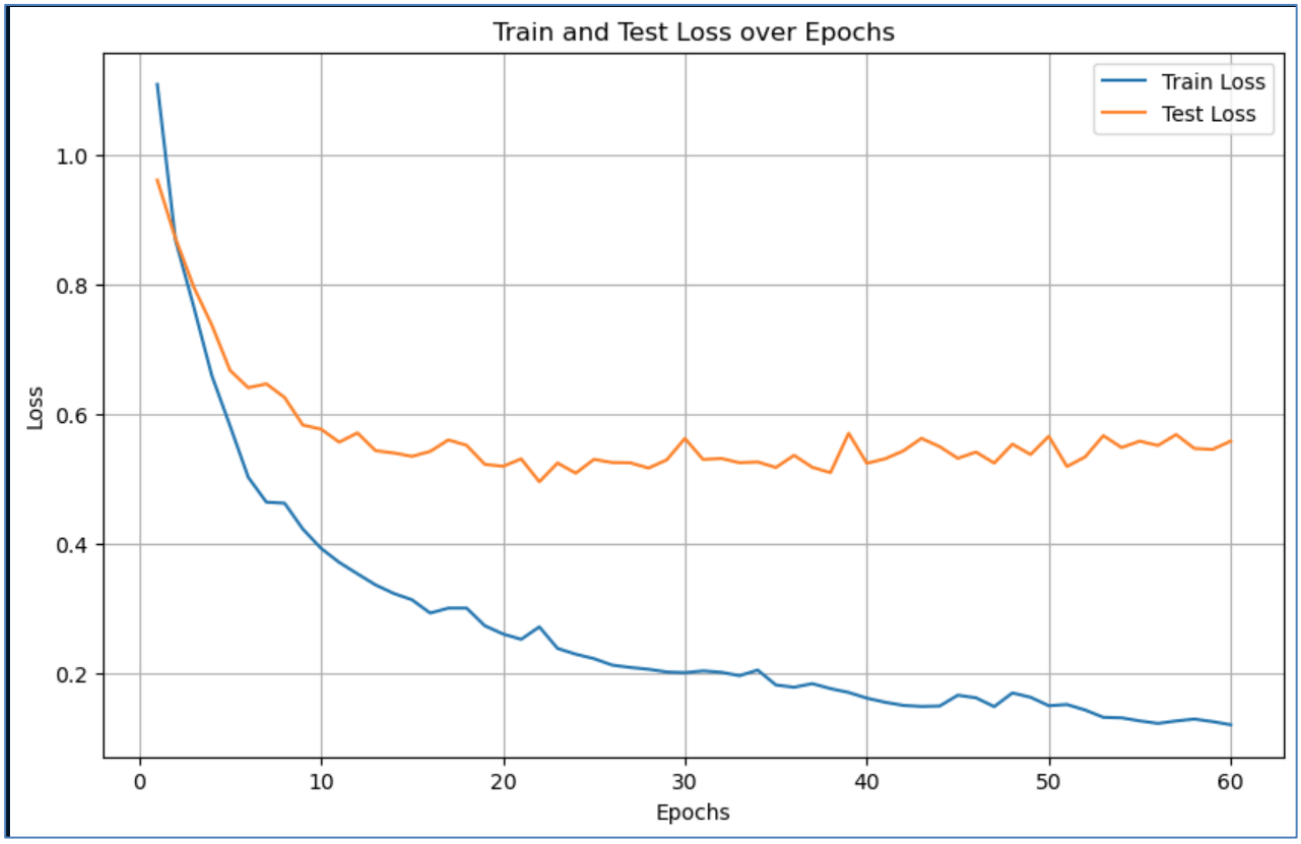
Σχήμα 89 Αποτελέσματα 5 πρώτων εποχών (βρόγχος εκπαίδευσης και βρόγχος αξιολόγησης)

```
Final Train Losses: [1.109500210393559, 0.8690812933174047, 0.7664930359883741, 0.6595872723582116, 0.5821354465389793]
Final Test Losses: [0.9617030756040053, 0.8718561286276038, 0.7968964956023477, 0.7376768426461653, 0.6679109944538637]
Final Train Dice Scores: [0.08675681784639716, 0.16613202168881136, 0.29098163469089755, 0.40892088412195576, 0.5093171753793616]
Final Test Dice Scores: [0.12814155185962012, 0.18467955136227168, 0.27509801233695313, 0.3261968098072843, 0.4790465361015363]
Final Train IoU Scores: [0.048064612471992875, 0.09779059821379382, 0.188535329280141, 0.28580317611884964, 0.37635013954819774]
Final Test IoU Scores: [0.07164301718909719, 0.11014091430115514, 0.166461567775431, 0.21340216463431716, 0.34108908305113966]
```

Σχήμα 90 Συγκεντρωτικά αποτελέσματα 5 πρώτων εποχών (Συναρτήσεις Loss και μέτρων αξιολόγησης)

5.8 ΒΗΜΑ 7: ΟΠΤΙΚΟΠΟΙΗΣΗ ΚΑΜΠΥΛΩΝ ΣΥΝΑΡΤΗΣΗΣ LOSS ΚΑΙ ΕΡΜΗΝΕΙΑ ΤΟΥΣ

Η διαδικασία της εκπαίδευσης και της αξιολόγησης του μοντέλου διήρκεσε για 60 εποχές. Για κάθε εποχή καταγράφονται δύο τιμές της συνάρτησης loss, μια τιμή που αφορά το βρόγχο εκπαίδευσης και μία που αφορά το βρόγχο αξιολόγησης. Παρακάτω παρατίθεται ένα διάγραμμα των δύο καμπύλων της συνάρτησης loss.



Σχήμα 91 Διάγραμμα Καμπύλων συνάρτησης loss για εκπαίδευση και αξιολόγηση του μοντέλου SAM

Ο οριζόντιος άξονας του διαγράμματος αντιπροσωπεύει το σύνολο των εποχών (60). Ο κατακόρυφος άξονας αντιπροσωπεύει την τιμή της συνάρτησης loss. Η μπλε καμπύλη συμβολίζει τις τιμές της loss για το βρόγχο εκπαίδευσης (training loss) ενώ η πορτοκαλί καμπύλη αντιπροσωπεύει τις τιμές της loss για το βρόγχο αξιολόγησης (testing loss).

Η καμπύλη της training loss μειώνεται σημαντικά κατά τη διάρκεια των εποχών, υποδεικνύοντας ότι το μοντέλο μαθαίνει και προσαρμόζεται καλά στα δεδομένα εκπαίδευσης. Η καμπύλη δείχνει μια απότομη πτώση αρχικά, κάτι που είναι συνηθισμένο καθώς το μοντέλο μαθαίνει γρήγορα τα πιο προφανή μοτίβα στα δεδομένα. Μετά την αρχική απότομη πτώση, η καμπύλη αρχίζει να εξομαλύνεται, υποδεικνύοντας ότι το μοντέλο πλησιάζει την ιδανική του απόδοση στα δεδομένα εκπαίδευσης.

Η καμπύλη της testing loss παρουσιάζει ένα πιο ασταθές μοτίβο σε σύγκριση με την καμπύλη της training loss. Παρόλο που παρατηρείται μια γενική πτωτική τάση, οι διακυμάνσεις υποδηλώνουν κάποια αστάθεια ή μεταβλητότητα στην απόδοση του μοντέλου στα δεδομένα ελέγχου. Η τιμής της ξεκινάει υψηλότερα από την αντίστοιχη της εκπαίδευσης και μειώνεται πιο αργά, κάτι που είναι αναμενόμενο καθώς το μοντέλο συνήθως αποδίδει καλύτερα στα δεδομένα εκπαίδευσης που έχει δει προηγουμένως. Αυτό υποδεικνύει ότι το μοντέλο έχει φτάσει σε ένα σημείο όπου δεν βελτιώνεται σημαντικά η απόδοσή του σε νέα δεδομένα (αξιολόγησης).

Στη τελευταία εποχή παρατηρείται ότι η τιμής της συνάρτησης loss για την εκπαίδευση είναι αρκετά χαμηλότερη από την αντίστοιχη της αξιολόγησης. Αυτό μπορεί να υποδεικνύει το γεγονός της υπερ-προσαρμογής (overfitting), όπου το μοντέλο έχει μάθει πολύ καλά τα δεδομένα εκπαίδευσης αλλά δεν γενικεύει καλά σε νέα δεδομένα με τα οποία δεν έχει εκπαιδευτεί.

Για την αντιμετώπιση του προβλήματος του overfitting, συστήνεται μεταξύ άλλων η διαδικασία του «data augmentation» όπου τα δεδομένα εκπαίδευσης υφίστανται μετασχηματισμούς ώστε να αποτρέπεται η υπερ-προσαρμογή του μοντέλου σε αυτά και να βελτιώνεται η γενίκευση του μοντέλου σε νέα άγνωστα δεδομένα. Εκτός από τη τεχνική του «data augmentation» το μοντέλο δύναται να εκπαιδευτεί σε μεγαλύτερα σετ δεδομένων για να αφομοιώσει μεγαλύτερη ποικιλία στα μοτίβα που παρακολουθεί. Τέλος, μια ακόμα τεχνική είναι η πρόωρη διακοπή της εκπαίδευσης του μοντέλου. Αυτό πραγματοποιείται όταν διαπιστωθεί κατά την εκπαίδευση και αξιολόγηση του μοντέλου ότι η καμπύλης της testing loss παύει να μειώνεται και το μοντέλο οδηγείται σε υπερ-προσαρμογή στα δεδομένα εκπαίδευσης.

5.9 ΒΗΜΑ 8: ΚΑΘΟΡΙΣΜΟΣ ΜΕΤΡΩΝ ΑΞΙΟΛΟΓΗΣΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ ΚΑΙ ΟΠΤΙΚΟΠΟΙΗΣΗ ΤΩΝ ΔΙΑΓΡΑΜΜΑΤΩΝ ΤΟΥΣ

Κατά τη διάρκεια της εκπαίδευσης και αξιολόγησης του μοντέλου, εκτός από τις τιμές της συνάρτησης loss καταγράφονταν και τιμές δύο ενδεικτικών μέτρων αξιολόγησης της διαδικασίας που υλοποιήθηκαν προγραμματιστικά.

5.9.1 Ο ΣΥΝΤΕΛΕΣΤΗΣ DICE ΩΣ ΜΕΤΡΟ ΑΞΙΟΛΟΓΗΣΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Το πρώτο μέτρο αξιολόγησης είναι ο συντελεστής DICE. Αποτελεί κοινό μέτρο αξιολόγησης που χρησιμοποιείται στην υπολογιστική όραση για σύγκριση της

ομοιότητας μεταξύ δύο συνόλων ή εικόνων. Χρησιμοποιείται ευρέως σε αντικείμενα κατάτμησης εικόνας και ανίχνευσης αντικειμένων.

Ισούται με

$$DICE\ Score = \frac{(2 * Τομή\ των\ δύο\ συνόλων)}{(Έκταση\ προβλεψης) + (Αληθή\ έκταση)}$$

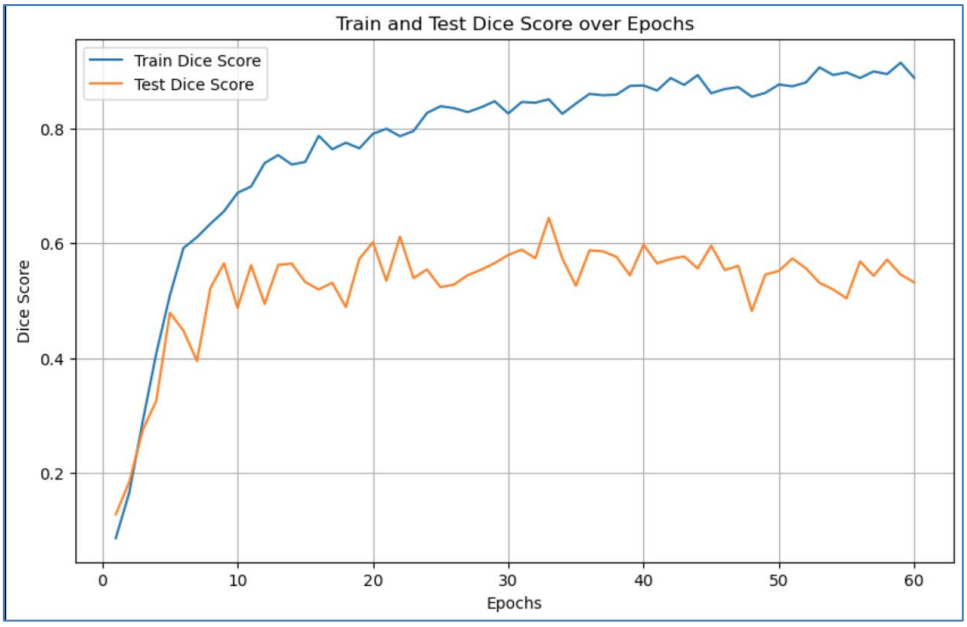
Όπου:

- Τομή των δύο συνόλων (intersection) αναφέρεται στις κοινές περιοχές μεταξύ των εικονοστοιχείων της πρόβλεψης του μοντέλου κατάτμησης (predicted segmentation output) και της αληθής μάσκας (groundtruth segmentation output).
- Έκταση πρόβλεψης είναι το σύνολο των εικονοστοιχείων στη πρόβλεψη του μοντέλου
- Αληθή έκταση είναι το σύνολο των εικονοστοιχείων στην αληθή μάσκα της εκάστοτε εικόνας.

Οι τιμές του συντελεστή Dice κυμαίνονται από 0 έως 1 όπου:

- Το 0 συμβολίζει καμία ομοιότητα μεταξύ της πρόβλεψης του μοντέλου και της αληθής μάσκας.
- Το 1 συμβολίζει ότι η πρόβλεψη του μοντέλου είναι ίδια ακριβώς με την αληθή μάσκα.

Ένας υψηλός συντελεστής DICE υποδηλώνει πιο ακριβές αποτέλεσμα κατάτμησης. Παρακάτω ακολουθεί το διάγραμμα της καμπύλης του συγκεκριμένου συντελεστή για τις διαδικασίες εκπαίδευσης και αξιολόγησης για 60 εποχές.



Σχήμα 92 Διάγραμμα καμπυλών Συντελεστή DICE για την εκπαίδευση και αξιολόγηση του SAM

Το διάγραμμα περιέχει δύο καμπύλες, τη μπλε καμπύλη που αντιπροσωπεύει το συντελεστή DICE κατά τη διάρκεια των εποχών στα δεδομένα εκπαίδευσης και την πορτοκαλί καμπύλη το συντελεστή DICE κατά τη διάρκεια των εποχών στα δεδομένα ελέγχου. Ο οριζόντιος άξονας συμβολίζει το σύνολο των εποχών (60) και ο κάθετος άξονας τις τιμές του συντελεστή DICE.

Σχετικά με την εκπαίδευση (train DICE Score), κατά τις πρώτες 15 εποχές, παρατηρείται μια σταθερή και σημαντική αύξηση που υποδεικνύει ότι το μοντέλο μαθαίνει από τα δεδομένα εκπαίδευσης και βελτιώνει την απόδοση του. Μετά τις πρώτες 15 εποχές, ο ρυθμός αύξησης του συντελεστή DICE μειώνεται και η καμπύλη αρχίζει να σταθεροποιείται μεταξύ των τιμών 0.75 και 0.85. Η συγκεκριμένη συμπεριφορά είναι ενδεικτική της διαδικασίας σύγκλισης του μοντέλου, όπου το μοντέλο συνεχίζει να βελτιώνει την απόδοση του αλλά με μικρότερους ρυθμούς.

Σχετικά με την αξιολόγηση (test DICE Score), η καμπύλη παρουσιάζει παρόμοια συμπεριφορά για τα δεδομένα ελέγχου προσεγγίζοντας μέγιστη τιμή περίπου στις 15 εποχές. Μετά τις πρώτες 15 εποχές, η καμπύλη παρουσιάζει αστάθεια και σημαντικές διακυμάνσεις μεταξύ των τιμών 0.5 και 0.7. Αυτές οι διακυμάνσεις υποδεικνύουν ότι το μοντέλο δε γενικεύει καλά στα δεδομένα ελέγχου με τα οποία δεν έχει εκπαιδευτεί.

Παρατηρείται όπως και στην ερμηνεία του διαγράμματος της συνάρτησης loss, το γεγονός του overfitting. Το μοντέλο δε μπορεί να γενικεύσει ικανοποιητικά σε νέα δεδομένα με τα οποία δεν έχει εκπαιδευτεί. Ως λύση προτείνεται όπως και στη προηγούμενη περίπτωση η εφαρμογή της τεχνικής «Data Augmentation» για τον εμπλουτισμό των δεδομένων με σκοπό την αύξηση της ποικιλίας των δεδομένων εκπαίδευσης.

5.9.2 Ο ΔΕΙΚΤΗΣ ΕΠΙΚΑΛΥΨΗΣ IoU ΩΣ ΜΕΤΡΟ ΑΞΙΟΛΟΓΗΣΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Ο Δείκτης Επικάλυψης (IoU, Intersection over Union) γνωστός ως και Jaccard index, αποτελεί μέτρο αξιολόγησης του μοντέλου κατάτμησης υπολογίζοντας την επικάλυψη μεταξύ δύο συνόλων, της πρόβλεψης του μοντέλου και της αληθής μάσκας (prediction output and groundtruth mask).

Ορίζεται ως :

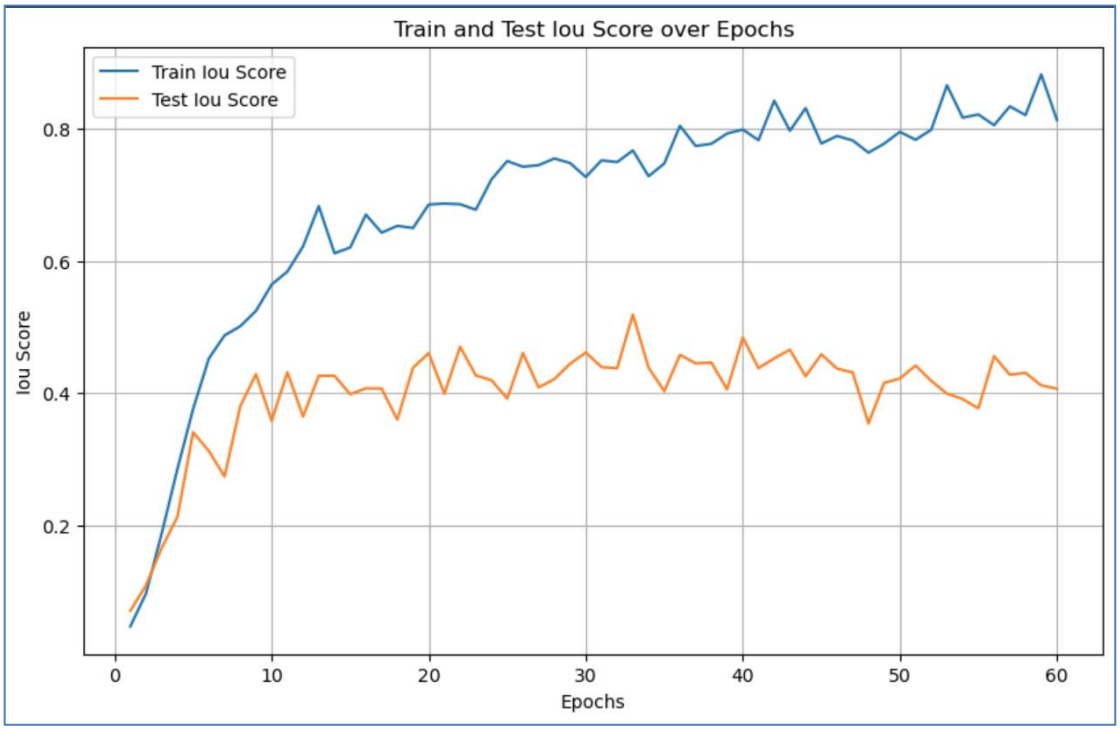
$$IoU = \frac{\alpha\lambda\eta\theta\eta\ \mu\acute{\alpha}\sigma\kappa\alpha\ \cap \mu\acute{\alpha}\sigma\kappa\alpha\ \pi\rho o\beta\lambda\epsilon\psi\eta\varsigma}{\alpha\lambda\eta\theta\acute{\eta}\ \mu\acute{\alpha}\sigma\kappa\alpha\ \cup \mu\acute{\alpha}\sigma\kappa\alpha\ \pi\acute{\rho}\beta\lambda\epsilon\psi\eta\varsigma}$$

όπου ο αριθμητής αναφέρεται στα κοινά εικονοστοιχεία της πρόβλεψης του μοντέλου κατάτμησης και της πραγματικής μάσκας (groundtruth mask) ενώ ο

παρανομαστής αναφέρεται στο πλήθος των εικονοστοιχείων που ανήκουν είτε στη πρόβλεψη είτε στη αληθή μάσκα είτε και στα δύο.

Οι τιμές του κυμαίνονται από 0 έως 1, όπου 0 σημαίνει ότι δεν υπάρχει καμία επικάλυψη μεταξύ των προβλέψεων του μοντέλου και των πραγματικών τιμών ενώ 1 σημαίνει τέλεια επικάλυψη. Ο δείκτης IoU είναι αυστηρότερο μέτρο αξιολόγησης σε σύγκριση με το συντελεστή DICE επειδή λαμβάνει υπόψη το συνδυασμένο μέγεθος των προβλέψεων και των πραγματικών τιμών

Παρακάτω ακολουθεί το διάγραμμα της καμπύλης του δείκτη επικάλυψης για τις διαδικασίες εκπαίδευσης και αξιολόγησης για 60 εποχές.



Σχήμα 93 Διάγραμμα καμπυλών Δείκτη Επικάλυψης IoU για την εκπαίδευση και αξιολόγηση του SAM

Το διάγραμμα του Δείκτη Επικάλυψης είναι παρόμοιο με του συντελεστή DICE. Το διάγραμμα περιέχει δύο καμπύλες, τη μπλε καμπύλη που αντιπροσωπεύει το συντελεστή IoU κατά τη διάρκεια των εποχών στα δεδομένα εκπαίδευσης και την πορτοκαλί καμπύλη το συντελεστή IoU κατά τη διάρκεια των εποχών στα δεδομένα ελέγχου. Ο οριζόντιος άξονας συμβολίζει το σύνολο των εποχών (60) και ο κάθετος άξονας τις τιμές του δείκτη επικάλυψης.

Σχετικά με την εκπαίδευση (train IoU Score), κατά τις πρώτες 15 εποχές, παρατηρείται μια σταθερή και σημαντική αύξηση που υποδεικνύει ότι το μοντέλο μαθαίνει από τα δεδομένα εκπαίδευσης και βελτιώνει την απόδοση του. Μετά τις πρώτες 15-20 εποχές, ο ρυθμός αύξησης μειώνεται και η καμπύλη αρχίζει να σταθεροποιείται μεταξύ των τιμών 0.65 και 0.80. Η συγκεκριμένη συμπεριφορά είναι ενδεικτική της διαδικασίας σύγκλισης του μοντέλου, όπου το μοντέλο συνεχίζει να βελτιώνει την απόδοση του αλλά με μικρότερους ρυθμούς.

Σχετικά με την αξιολόγηση (test IoU Score), η καμπύλη παρουσιάζει παρόμοια συμπεριφορά για τα δεδομένα ελέγχου προσεγγίζοντας μέγιστη τιμή περίπου στις 15 εποχές. Μετά τις πρώτες 15 εποχές, η καμπύλη παρουσιάζει αστάθεια και σημαντικές διακυμάνσεις μεταξύ των τιμών 0.4 και 0.55. Αυτές οι διακυμάνσεις υποδεικνύουν ότι το μοντέλο δε γενικεύει καλά στα δεδομένα ελέγχου με τα οποία δεν έχει εκπαιδευτεί.

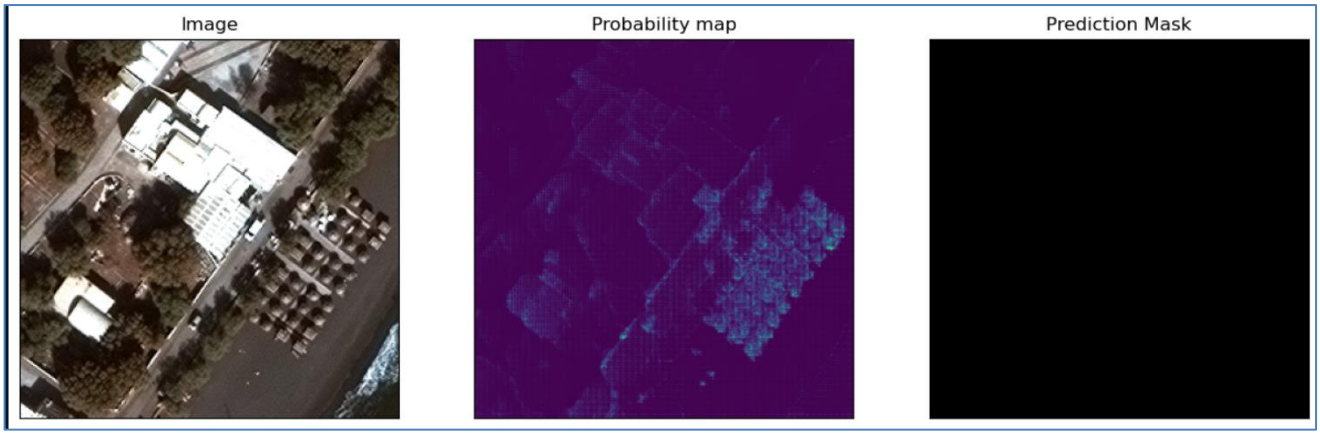
Όπως παρατηρείται ο δείκτης επικάλυψης είναι αυστηρότερος από το συντελεστή DICE ως προς την αξιολόγηση του μοντέλου. Όπως και προηγουμένως παρατηρείται το γεγονός του overfitting. Το μοντέλο δε μπορεί να γενικεύσει ικανοποιητικά σε νέα δεδομένα με τα οποία δεν έχει εκπαιδευτεί. Ως λύση προτείνεται όπως και στη προηγούμενη περίπτωση η εφαρμογή της τεχνικής «Data Augmentation» για τον εμπλουτισμό των δεδομένων με σκοπό την αύξηση της ποικιλίας των δεδομένων εκπαίδευσης.

5.10 ΒΗΜΑ 9: ΑΠΟΘΗΚΕΥΣΗ ΤΟΥ ΕΚΠΑΙΔΕΥΜΕΝΟΥ ΜΟΝΤΕΛΟΥ SAM ΚΑΙ ΟΠΤΙΚΟ ΕΛΕΓΧΟ ΤΗΣ ΑΠΟΔΟΣΗΣ ΤΟΥ ΣΕ ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ

Επόμενο βήμα αποτελεί η αποθήκευση του εκπαιδευμένου μοντέλου SAM. Συγκεκριμένα αποθηκεύονται τοπικά οι τροποποιημένοι παράμετροι του μοντέλου που έχουν εκπαιδευτεί κατά τη διάρκεια των 60 εποχών. Οι παράμετροι του μοντέλου έχουν τροποποιηθεί μόνο στην περιοχή του αποκωδικοποιητή μάσκας. (mask decoder). Έτσι σε μεταγενέστερο χρόνο δύναται να πραγματοποιηθεί αρχικοποίηση του SAM με τις νέες παραμέτρους.

Αφού αρχικοποιηθεί εκ νέου το μοντέλο SAM με τα νέα βάρη, πραγματοποιείται οπτικός έλεγχος της διαδικασίας κατάτμησης του μοντέλου σε δεδομένα ελέγχου, δηλαδή δεδομένα που δεν έχει εκπαιδευτεί. Θα πραγματοποιηθεί μια κατάτμηση μιας εικόνας ελέγχου με δύο διαφορετικές διαδικασίες.

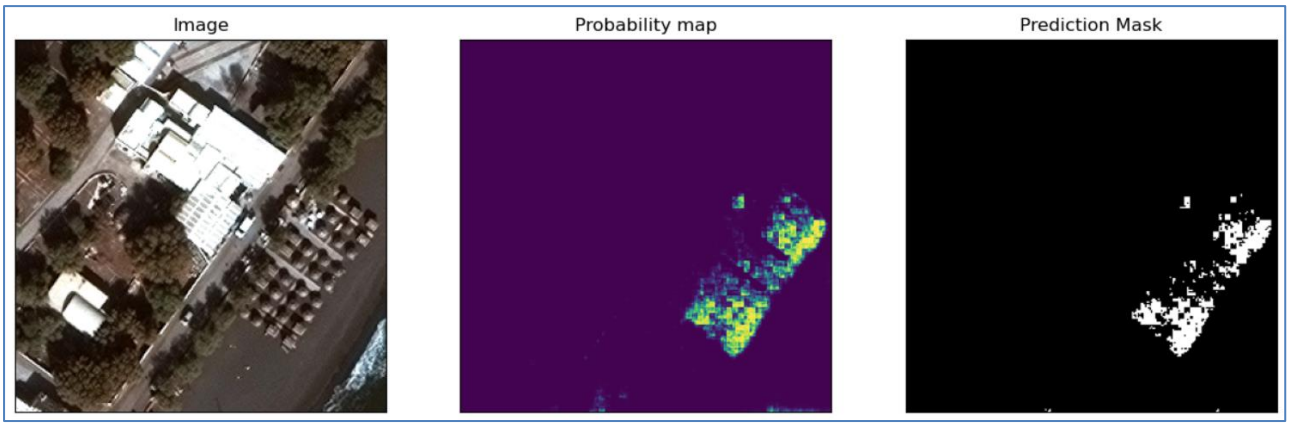
Στη πρώτη διαδικασία δε θα χρησιμοποιηθεί καθόλου γεωμετρική προτροπή στο μοντέλο. Δηλαδή δε θα δοθεί στο μοντέλο πλέγμα σημείων για να ανιχνεύσει και να κατατμήσει τις οντότητες ενδιαφέροντος όπως συνέβη και στη διαδικασία εκπαίδευσης του. Παρακάτω παρατίθεται το αποτέλεσμα της διαδικασίας χωρίς πλέγμα σημείων.



Σχήμα 94 Αποτέλεσμα κατάτμησης εικόνας χωρίς γεωμετρική προτροπή με εκπαιδευμένο μοντέλο SAM

Όπως απεικονίζεται και στο Σχήμα 94, από μια εικόνα προκύπτει ένας χάρτης πιθανοτήτων. Τα εικονοστοιχεία που χαρακτηρίζονται από πιθανότητα πάνω από μια ορισμένη τιμή εμφανίζονται ως μάσκα στο τελικό αποτέλεσμα της κατάτμησης. Το γεγονός μη παροχής πλέγματος σημείων ως γεωμετρική προτροπή στο μοντέλο το εμποδίζει από το να προβλέψει και να τμηματοποιήσει τις περιοχές με τις ομπρέλες στην εικόνα εισόδου.

Στη δεύτερη διαδικασία, μαζί με την εικόνα εισόδου θα δοθεί και ένα πλέγμα σημείων στο εκπαιδευμένο μοντέλο SAM για τμηματοποιήσει την εικόνα. Παρακάτω παρατίθεται το αποτέλεσμα της κατάτμησης.



Σχήμα 95 Αποτέλεσμα κατάτμησης εικόνας με γεωμετρική προτροπή με εκπαιδευμένο μοντέλο SAM

Στο Σχήμα 95 απεικονίζεται η τμηματοποίηση της εικόνας εισόδου με το εκπαιδευμένο μοντέλο SAM και με είσοδο του πλέγματος σημείων ως γεωμετρική προτροπή στο μοντέλο. Στο χάρτη πιθανοτήτων τα εικονοστοιχεία που αποτυπώνουν ομπρέλες εμφανίζονται με κίτρινο χρώμα και στο τελικό αποτέλεσμα εξάγονται ως μάσκα. Έτσι η οντότητα ενδιαφέροντος (ομπρέλες) ανιχνεύεται και εξάγεται μέσω του εκπαιδευμένου μοντέλου SAM. Σαφώς όπως διαπιστώθηκε και

στις συναρτήσεις loss και από την ερμηνεία των μέτρων αξιολόγησης, το αποτέλεσμα χρήζει περαιτέρω βελτίωσης.

ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ, ΔΥΝΑΤΟΤΗΤΕΣ ΒΕΛΤΙΩΣΗΣ ΚΑΙ ΠΡΟΕΚΤΑΣΕΙΣ ΤΩΝ ΥΠΑΡΧΟΥΣΩΝ ΜΕΘΟΔΟΛΟΓΙΩΝ

Στο κεφάλαιο αυτό, επιχειρείται μια σύνοψη των μεθοδολογιών που παρουσιάστηκαν στη παρούσα διπλωματική εργασία καθώς και σχολιασμός των αποτελεσμάτων και εξαγωγή χρήσιμων συμπερασμάτων. Τέλος, προτείνονται πιθανές προεκτάσεις του Segment Anything και της έννοιας της promptable segmentation σε άλλους τομείς αλλά και ενσωμάτωση του σε πιο σύνθετες και απαιτητικές διεργασίες και μεθοδολογίες.

Στη παρούσα διπλωματική εργασία επιχειρήθηκε η διαμόρφωση του πλαισίου μέσα στο οποίο παρουσιάζεται ο τρόπος με τον οποίο η επιστήμη της βαθιάς μάθησης και συγκεκριμένα ο τομέας της υπολογιστικής όρασης με χρήση διαφορετικών μεθοδολογιών προβαίνει στην ανάγνωση και επεξεργασία της εικόνας, στην ανίχνευση οντοτήτων εντός αυτής και στην εξαγωγή τους για περαιτέρω επεξεργασία. Παρουσιάστηκε το βασικό θεωρητικό υπόβαθρο που εδράζονται πολλές έννοιες της βαθιάς μάθησης, των νευρωνικών δικτύων, της αρχιτεκτονικής τους καθώς και πολλών τεχνικών που εφαρμόζουν.

Επιλέχθηκε το μοντέλο Segment Anything και κατ' επέκταση η έννοιας της κατάτμησης εικόνας με βάση την προτροπή-οδηγία (promptable segmentation) για να σχεδιαστούν και να εφαρμοστούν μεθοδολογίες για ανίχνευση και εξαγωγή οντοτήτων από τηλεπισκοπικές απεικονίσεις. Το χαρακτηριστικό του «zero-shot segmentation» δηλαδή η δυνατότητα της κατάτμησης εικόνων που δεν είχε συναντήσει ποτέ το μοντέλο κατά τη διάρκεια της εκπαίδευσης του, καθιστά το συγκεκριμένο μοντέλο ιδανικό για να σχεδιαστούν μεθοδολογίες που θα αφορούν τηλεπισκοπικές απεικονίσεις. Ένα μεγάλο πλεονέκτημα είναι ότι το συγκεκριμένο μοντέλο αποτελεί μια παραλλαγή διαδραστικής τμηματοποίησης (interactive segmentation) αλλά με ελάχιστη συμμετοχή από τον εκάστοτε χρήστη σε σχέση με τη κλασική της μορφή. Η αρχιτεκτονική του μοντέλου και η δυνατότητα της εισαγωγής προτροπών προσφέρουν ευελιξία στις εφαρμογές που μπορεί να χρησιμοποιηθεί το SAM. Επιπρόσθετα με την τεχνική της αυτόματης παραγωγής μασκών μέσω του πλέγματος σημείων δύναται να εξαχθούν σχεδόν όλα τα αντικείμενα που εμπεριέχονται στην εκάστοτε δορυφορική εικόνα. Η συγκεκριμένη τεχνική αποτελεί μια παραλλαγή της «instance segmentation», με τη διαφορά ότι δεν αποδίδονται ετικέτες για κάθε αντικείμενο αλλά μόνο εξάγονται ως διαφορετικές οντότητες.

Ο συνδυασμός του Grounding Dino και του Segment Anything σε μια κοινή μεθοδολογία επιτρέπει την ανίχνευση οντοτήτων μέσω προτροπών κειμένου σε δορυφορικές απεικονίσεις, δημιουργία πλαισίων οριοθέτησης και μεταφόρτωση αυτών των πλαισίων ως γεωμετρικές προτροπές στο SAM για εξαγωγή της μάσκας τους. Το μειονέκτημα της συγκεκριμένης μεθοδολογίας είναι η δυσκολία που αντιμετωπίζει το Grounding Dino για ανίχνευση κάποιων κλάσεων οντοτήτων που υπάρχουν στις δορυφορικές απεικονίσεις. Για παράδειγμα σε απεικονίσεις αντικειμένων (όχι δορυφορικές εικόνες), το Grounding Dino έχει τη δυνατότητα να αναγνωρίσει όλα τα αντικείμενα, να αποδώσει ετικέτα και να δημιουργήσει πλαίσια οριοθέτησης τα οποία με τη σειρά τους εισάγονται στο SAM για κατάτμηση εικόνας. Το αποτέλεσμα είναι η εξαγωγή όλων των αντικειμένων από την εικόνα, με διαφορετική μάσκα το καθένα και ετικέτα. Αυτός είναι ο ορισμός της instance segmentation που δύσκολα μπορεί να εφαρμοστεί σε δορυφορικές απεικονίσεις.

Για την αντιμετώπιση του παραπάνω μειονεκτήματος, δηλαδή την αυτόματη ανίχνευση κάποιων οντοτήτων ενδιαφέροντος και εξαγωγή τους από τις δορυφορικές απεικονίσεις χωρίς ο χρήστης να προσδώσει στο μοντέλο την ακριβή τοποθεσία της οντότητας όπως συμβαίνει με τις γεωμετρικές προτροπές του μοντέλου εφαρμόστηκε μεθοδολογία περαιτέρω εκπαίδευσης του SAM (fine-tuning SAM). Σε περίπτωση που το μοντέλο Grounding Dino θα μπορούσε να ανιχνεύσει οντότητες ειδικού ενδιαφέροντος σε μια δορυφορική απεικόνιση και να αποδώσει ορθά τα πλαίσια οριοθέτησης η διαδικασία του «fine-tuning» δε θα ήταν απαραίτητη. Η διαδικασία της περαιτέρω εκπαίδευσης εκτός από σημαντικούς υπολογιστικούς πόρους που απαιτούνται, απαιτεί και μεγάλα σετ δεδομένων εκπαίδευσης. Τα σετ δεδομένων πρέπει να χαρακτηρίζονται από ποικιλομορφία για να προσδώσουν στο μοντέλο τη δυνατότητα να γενικεύσει τις προβλέψεις του και σε δεδομένα με τα οποία δεν έχει εκπαιδευτεί. Η αρχιτεκτονική του SAM επιτρέπει την απομόνωση μέρων του μοντέλου όπως ο αποκωδικοποιητής μάσκας και την εκπαίδευση μόνο των σχετικών παραμέτρων του μοντέλου για ταχύτερα αποτελέσματα και οικονομία στους υπολογιστικούς πόρους. Το αποτέλεσμα της άνω διαδικασίας φανέρωσε την ικανότητα του Segment Anything με τη κατάλληλη εκπαίδευση να ανιχνεύει και να εξάγει οντότητες ειδικού ενδιαφέροντος από τηλεπισκοπικές απεικονίσεις.

Το Segment Anything εκτός από τις δορυφορικές απεικονίσεις χρησιμοποιείται ευρέως και στην αυτόματη τμηματοποίηση ιατρικών εικόνων συμβάλλοντας στην αναγνώριση και καταγραφή ανωμαλιών και πιθανών ασθενειών, αυξάνοντας την ακρίβεια της διάγνωσης. Επιπρόσθετα, όπως πραγματοποιήθηκε στη παρούσα διπλωματική η διαδικασία της περαιτέρω εκπαίδευσης του SAM για οντότητες ενδιαφέροντος, έτσι δύναται να πραγματοποιηθεί η ίδια διαδικασία με ιατρικές εικόνες ως δεδομένα εκπαίδευσης για ανίχνευση και εξαγωγή πιθανών καρκινικών κυττάρων. Εκτός από τις ιατρικές απεικονίσεις, και ο τομέας της γεωργίας ωφελείται

από την τμηματοποίηση εικόνων καλλιεργειών για την ανάλυση της υγείας των φυτών και την εκτίμηση της παραγωγής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Abien Fred Agarap (2018). Deep Learning using Rectified Linear Units (ReLU). arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.1803.08375.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. arXiv:2005.12872 [cs]. [online] Available at: https://arxiv.org/abs/2005.12872.

Chai, J., Zeng, H., Li, A. and Ngai, E.W.T. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications, 6, p.100-134. doi:https://doi.org/10.1016/j.mlwa.2021.100134.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. doi:https://doi.org/10.1109/cvpr.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs]. [online] Available at: https://arxiv.org/abs/2010.11929.

Fukushima, K. (1969). Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements. IEEE Transactions on Systems Science and Cybernetics, 5(4), pp.322–333. doi:https://doi.org/10.1109/tssc.1969.300225.

Glorot, X., Bordes, A. and Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. [online]proceedings.mlr.press. Available at : https://proceedings.mlr.press/v15/glorot11a.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S. (2016). Deep learning for visual understanding: A review. Neurocomputing, [online] 187, pp.27–48. doi:https://doi.org/10.1016/j.neucom.2015.09.116.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R. (2021). Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs]. [online] Available at: https://arxiv.org/abs/2111.06377.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep Residual Learning for Image Recognition. [online] arXiv.org. Available at: https://arxiv.org/abs/1512.03385.

Hendrycks, D. and Gimpel, K. (2020). Gaussian Error Linear Units (GELUs). arXiv:1606.08415 [cs]. [online] Available at: https://arxiv.org/abs/1606.08415.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? [online] nyuscholars.nyu.edu. doi:https://doi.org/10.1109/ICCV.2009.5459469.

Kingma, D.P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. [online] arXiv.org. Available at: https://arxiv.org/abs/1412.6980.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P. and Girshick, R. (2023). Segment Anything. arXiv:2304.02643 [cs]. [online] Available at: https://arxiv.org/abs/2304.02643.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, [online] 60(6), pp.84–90. doi:https://doi.org/10.1145/3065386.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J. and Zhang, L. (2023). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. [online] arXiv.org. Available at: https://arxiv.org/abs/2303.05499.

Maas, A., Hannun, A. and Ng, A. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. [online] Available at: https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

Nair, V. and Hinton, G. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines.[online] Available at: https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf.

Narayan, S. (1997). The generalized sigmoid activation function: Competitive supervised learning. Information Sciences, 99(1-2), pp.69–82. doi:https://doi.org/10.1016/s0020-0255(96)00200-9.

Noh, H., Hong, S. and Han, B. (2015). Learning Deconvolution Network for SemanticSegmentation. arXiv:1505.04366 [cs]. [online] Available at: https://arxiv.org/abs/1505.04366.

Palatucci, M., Pomerleau, D., Hinton, G. and Mitchell, T. (2009). Zero-Shot Learning with Semantic Output Codes. [online] Available at: https://www.cs.toronto.edu/~hinton/absps/palatucci.pdf.

Prince, S.J.D. (n.d.). Understanding Deep Learning. [online] MIT Press. Available at: https://mitpress.mit.edu/9780262048644/understanding-deep-learning/ [Accessed 5 Jun. 2024].

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs]. [online] Available at: https://arxiv.org/abs/2103.00020.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. [online] arXiv.org. Available at: https://arxiv.org/abs/1506.02640.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature, [online] 323(6088), pp.533–536. doi:https://doi.org/10.1038/323533a0.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. [online] arXiv.org. Available at: https://arxiv.org/abs/1409.1556.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017). Attention Is All You Need. [online] arXiv.org. Available at: https://arxiv.org/abs/1706.03762.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1. doi:https://doi.org/10.1109/cvpr.2001.990517.

Xu, N., Price, B., Cohen, S., Yang, J. and Huang, T. (2016). Deep Interactive Object Selection. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1603.04042.