ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μέθοδοι Επεξεργασίας και Ανάλυσης Εικόνων με τη Χρήση Τεχνητών Συνελικτικών Νευρωνικών Δικτύων

Παναγιώτα – Χρυσοβαλάντου Γατούλα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Δημήτριος Ιακωβίδης
Αναπληρωτής Καθηγητής

Λαμία, 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μέθοδοι Επεξεργασίας και Ανάλυσης Εικόνων με τη Χρήση Τεχνητών Συνελικτικών Νευρωνικών Δικτύων

Παναγιώτα – Χρυσοβαλάντου Γατούλα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Δημήτριος Ιακωβίδης
Αναπληρωτής Καθηγητής

Λαμία, 2021

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Μέθοδοι Επεξεργασίας και Ανάλυσης Εικόνων με τη χρήση Τεχνητών Συνελικτικών Νευρωνικών Δικτύων

Παναγιώτα – Χρυσοβαλάντου Γατούλα

Τριμελής Επιτροπή:

Δημήτριος Ιακωβίδης, Αναπληρωτής Καθηγητής (επιβλέπων)

Κωνσταντίνος Δελήμπασης, Αναπληρωτής Καθηγητής

Μιχαήλ Σαβελώνας, Επίκουρος Καθηγητής

Ευχαριστίες

Με την περάτωση της παρούσας πτυχιακής εργασίας, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ. Ιακωβίδη Δημήτριο, για την ανάθεση της εργασίας, την καθοδήγηση του και όλο το χρόνο που μου διέθεσε για την ολοκλήρωση της. Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Γεώργιο Δήμα για τις καίριες συμβουλές, τις γνώσεις που μου μετέδωσε, το χρόνο του, καθώς και για την υλική του υποστήριξη στην εκπαίδευση των νευρωνικών δικτύων.

ΠΕΡΙΛΗΨΗ

Η ανίχνευση σημαντικών αντικειμένων (Salient Object Detection - SOD), εμπνευσμένη από τον βιολογικό μηχανισμό της οπτικής προσοχής, αποσκοπεί στον εντοπισμό περιοχών των εικόνων όπου απαντώνται αντικείμενα στα οποία συγκεντρώνεται η οπτική προσοχή στην πρώτη παρατήρηση μιας οπτικής αναπαράστασης. Τα τελευταία χρόνια η ραγδαία ανάπτυξη των τεχνικών Βαθιάς Μάθησης έχει συμβάλλει καθοριστικά στην εξέλιξη του εν λόγω πεδίου. Βιβλιογραφικά τα συνελικτικά δίκτυα συνιστούν σήμερα την κυρίαρχη τάση αναφορικά με την ανίχνευση των σημαντικών αντικειμένων. Πρόσφατα, ερευνητικές εργασίες προσανατολισμένες στην ανίχνευση σημαντικών αντικειμένων έχουν αρχίσει να λαμβάνουν υπόψη την πληροφορία του βάθους μιας οπτικής αναπαράστασης προκειμένου να παρέχουν όσο το δυνατόν πιο συνεπή αποτελέσματα συναρτήσει του βιολογικού μηχανισμού της οπτικής προσοχής. Η πληροφορία του βάθους μιας οπτικής σκηνής, στις μεθοδολογίες που αναπτύσσονται (RGB-D Salient Object Detection methods) ανακτάται είτε μέσω συστημάτων στερεοσκοπικών καμερών είτε μέσω αισθητήρων βάθους. Ακόμη, τα συνελικτικά νευρωνικά δίκτυα έχουν αξιοποιηθεί για τη μονοφθαλμική εκτίμηση της πληροφορίας του βάθους οπτικών αναπαραστάσεων, επιτυγχάνοντας ακρίβεια που προσεγγίζει εκείνη των καμερών και των αισθητήρων βάθους. Ωστόσο, βιβλιογραφικά δε φαίνεται να επαληθεύεται η αξιολόγηση τους συναρτήσει της ανίχνευσης σημαντικών αντικειμένων. Η παρούσα εργασία αξιοποιεί τα συνελικτικά νευρωνικά δίκτυα και προτείνει μια μεθοδολογία για την υποβοηθούμενη από την πληροφορία του βάθους ανίχνευση σημαντικών αντικειμένων σε εικόνες. Η πρωτοτυπία της μεθοδολογίας έγκειται στο γεγονός ότι η πληροφορία του βάθους που αξιοποιείται σε αντίθεση με τις συμβατικές προσεγγίσεις, εκτιμάται από προεκπαιδευμένα συνελικτικά νευρωνικά δίκτυα. Η εφαρμογή της μεθοδολογίας σε γνωστά RGB-D σύνολα δεδομένων παράγει συγκρίσιμα αποτελέσματα έναντι των κλασσικών προσεγγίσεων.

ABSTRACT

Salient object detection (SOD) inspired by biological mechanism of visual attention, aims at detecting regions or objects more attentive at first glance. Rapid advances in deep learning techniques have critically contributed to the evolution of SOD over the last years. In the literature, Convolutional Neural Networks constitute a mainstream regarding SOD. Recently, research papers on SOD have started taking into account depth information so as to produce more consistent results concerning the biological mechanism of visual attention. In emerging methodologies (RGB-D Salient Object Detection methods), depth information is acquired either by stereo cameras or by depth sensors. Furthermore, Convolutional Neural Networks have been used for monocular depth estimation and they achieve a performance approximating that of RGB-D sensors. However, the aforementioned methods have yet to be assessed with respect to the task of SOD. This study uses Convolutional Neural Networks and proposes a methodology for RGB-D Salient Object Detection. The originality of the proposed method lies in the fact that the depth information used is estimated from pretrained Convolutional Neural Networks contrary to conventional methods. The evaluation of the proposed framework with reference to RGB-D datasets produces comparable results to conventional RGB-D methodologies.

Πίνακας περιεχομένων

1 Εισαγωγή

1.1. Επεξεργασία & Ανάλυση Εικόνας

Η υπερμεγέθης διαθέσιμη ποσότητα οπτικών πληροφορίων τις τελευταίες δεκαετίες διαδραμάτισε σημαίνοντα ρόλο στην ανάδειξη του πεδίου της Ψηφιακής Επεξεργασίας Εικόνων το οποίο αποτελεί προέκταση του πεδίου της Ψηφιακής Επεξεργασίας Σημάτων (Pitas, 2000). Ο όρος Ψηφιακή Επεξεργασία Εικόνων αφορά τη διαδικασία της επεξεργασίας ψηφιακών εικόνων με την αξιοποίηση αλγορίθμων και τη χρήση ηλεκτρονικών υπολογιστών (Gonzalez & Woods, 2011). Η ανάπτυξη του εν λόγω πεδίου υπήρξε ραγδαία εξαιτίας της συνεχούς βελτίωσης των υπολογιστικών συστημάτων, της εξέλιξης της επιστήμης της Πληροφορικής αλλά και συγγενών επιστημονικών κλάδων όπως εκείνος των Μαθηματικών, καθώς και των απαιτήσεων της σύγχρονης κοινωνίας για την ανάπτυξη εφαρμογών με απώτερο σκοπό την αντιμετώπιση και επίλυση προβλημάτων της καθημερινότητας. Οι παραπάνω παράγοντες έχουν καταστήσει την ψηφιακή επεξεργασία εικόνων πολύτιμο εργαλείο σε ένα ευρύ φάσμα, που κυμαίνεται από την τηλεπισκόπηση και τα συστήματα γεωγραφικών πληροφορίων μέχρι τη βιολογία, τη ρομποτική και την ιατρική διάγνωση (Sarfraz, 2020).

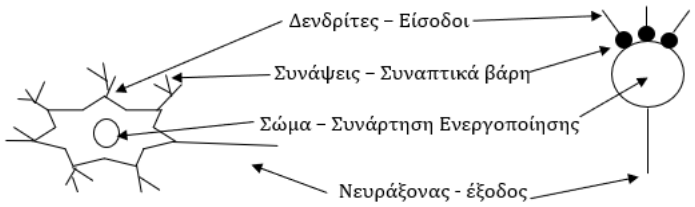
Αλληλένδετο με το πεδίο της Ψηφιακής Επεξεργασίας Εικόνων, το οποίο σήμερα συνιστά ολόκληρη επιστήμη (Παπαμάρκος, 2013), είναι το πεδίο της Ανάλυσης Ψηφιακών Εικόνων (Digital Image Analysis). Η ανάλυση ψηφιακών εικόνων αξιοποιεί τις τεχνικές που παρέχει η επεξεργασία εικόνων όπως η αναγνώριση αντικειμένων, η κατάτμηση εικόνων κα , με απώτερο σκοπό την απόδοση σημασιολογικής ερμηνείας στο περιεχόμενο των ψηφιακών εικόνων (Breckon, 2011). Σε αντίθεση με την επεξεργασία εικόνων η οποία αποσκοπεί κατά κύριο λόγο στην τροποποίηση του περιεχομένου των ψηφιακών εικόνων, η ανάλυση ψηφιακών εικόνων αναφέρεται στην προσπάθεια περιγραφής, αναγνώρισης και κατανόησης του περιεχομένου των ψηφιακών εικόνων (Sarfraz, 2020).

Οι τεχνικές ανάλυσης ψηφιακών εικόνων επιχειρούν να προσομοιάσουν τις περίπλοκες νευροφυσιολογικές λειτουργίες της ανθρώπινης όρασης οι οποίες ωστόσο έχουν αναλυθεί και κατανοηθεί προς το παρόν μόνο σε μερικό βαθμό (Sarfraz, 2020). Υπό αυτό το πρίσμα, τα όρια μεταξύ των ερευνητικών πεδίων της ανάλυσης εικόνων και της υπολογιστικής όρασης η οποία αποβλέπει στην εξομοίωση της ανθρώπινης όρασης με τη χρήση υπολογιστών, δεν είναι πλήρως σαφή και καθορισμένα (Gonzalez & Woods, 2011).

1.2. Τεχνητά Νευρωνικά Δίκτυα και Βαθιά Μάθηση

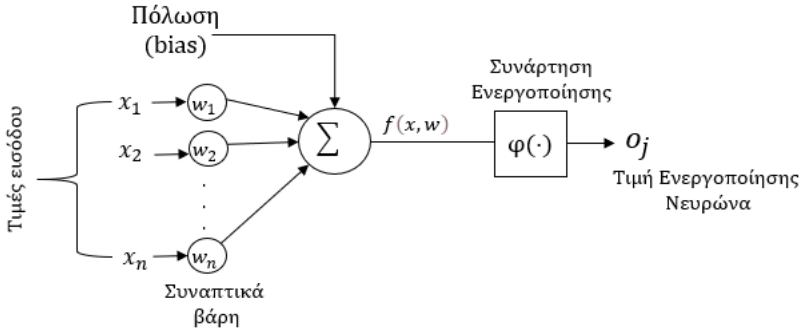
Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANNs) συνιστούν ένα από τα πιο δημοφιλή και ευρέως χρησιμοποιημένα συστήματα Μηχανικής Μάθησης (Machine Learning). Η δομή των τεχνητών νευρωνικών δικτύων είναι εμπνευσμένη από εκείνη των βιολογικών νευρωνικών δικτύων των έμβιων όντων και η λειτουργία τους επιτυγχάνεται κατ' αναλογία με τη μοντελοποίηση του μηχανισμού εκμάθησης που συναντάται στα βιολογικά νευρωνικά δίκτυα (Aggarwal, 2018).

Δομικές μονάδες των Τεχνητών Νευρωνικών Δικτύων είναι οι τεχνητοί νευρώνες. Ένας τεχνητός νευρώνας (perceptron) λειτουργεί κατά αναλογία του βιολογικού νευρώνα των βιολογικών νευρωνικών δικτύων (Aggarwal, 2018). Η αναλογία μεταξύ τεχνητών και βιολογικών νευρώνων μπορεί να παρατηρηθεί στην εικόνα (1.1) ενώ η δομή ενός τυπικού τεχνητού νευρώνα παρουσιάζεται στην εικόνα (1.2).



1. 1 Αναλογία βιολογικών και τεχνητών νευρώνων

Ένας τεχνητός νευρώνας δέχεται στην είσοδο του αριθμητικές τιμές που αντιστοιχούν σε χαρακτηριστικά (features) και είναι οι συντεταγμένες ενός διανύσματος $\vec{x} = (x_1, x_2, \dots, x_n)$. Ο τεχνητός νευρώνας πραγματοποιεί ταυτόχρονη επεξεργασία των πολλαπλών τιμών της εισόδου του και παράγει ως έξοδο μια νέα αριθμητική τιμή o_j, η οποία καλείται τιμή ενεργοποίησης του νευρώνα (Gonzalez & Woods, 2011).



1. 2 Δομή ενός τεχνητού νευρώνα

Αναλυτικότερα, καθεμία από τις τιμές εισόδου του τεχνητού νευρώνα $x_1, x_2, \dots, x_n$ πολλαπλασιάζεται με ένα συντελεστή βαρύτητας $w_1, w_2, \dots, w_n$. Οι τιμές των συντελεστών βαρύτητας $\vec{w} = (w_1, w_2, \dots, w_n)$ καλούνται συναπτικά βάρη του νευρώνα. Επιπλέον, κάθε τεχνητός νευρώνας διαθέτει μια τιμή εισόδου ίση με 1 η οποία πολλαπλασιάζεται με μια τιμή πόλωσης (bias) b_0. Το είδος της επεξεργασίας που μόλις περιγράφθηκε το οποίο διενεργείται σε κάθε νευρώνα αποτυπώνεται μαθηματικά στη σχέση (1.1) .

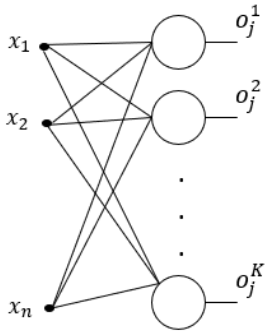
$$f(x, w) = b_0 + \sum_{i=1}^{n} x_i w_i \qquad \textbf{(1.1)}$$

Η τιμή ενεργοποίησης του νευρώνα από την εφαρμογή μιας συνάρτησης φ(·), η οποία καλείται συνάρτηση ενεργοποίησης (activation function), στο σταθμισμένο άθροισμα της εισόδου του νευρώνα, όπως υποδεικνύει η σχέση (1.2). Η επιλογή της συνάρτησης που χρησιμοποιείται ως συνάρτηση ενεργοποίησης πραγματοποιείται ώστε να ανταποκρίνεται στην επιθυμητή απόκριση του νευρώνα. Μεταξύ των συνηθέστερων επιλογών για τη συνάρτηση ενεργοποίησης είναι η σιγμοειδείς συνάρτηση, η συνάρτηση υπερβολικής εφαπτομένης και η συνάρτηση της ανορθωμένης γραμμικής μονάδας. (Gonzalez & Woods, 2011).

$$o_j = \varphi(f(x, w)) \qquad \textbf{(1.2)}$$

Τόσο οι τιμές των συναπτικών βαρών όσο και η τιμή της πόλωσης που διαθέτει ένας τεχνητός νευρώνας συνιστούν μεταβλητές παραμέτρους. Η διαδικασία της προσαρμογής των τιμών των παραμέτρων ενός νευρώνα πραγματοποιείται κατά αναλογία του μηχανισμού εκμάθησης των βιολογικών νευρώνων προκειμένου ο τεχνητός νευρώνας να είναι σε θέση να παράγει την επιθυμητή απόκριση.

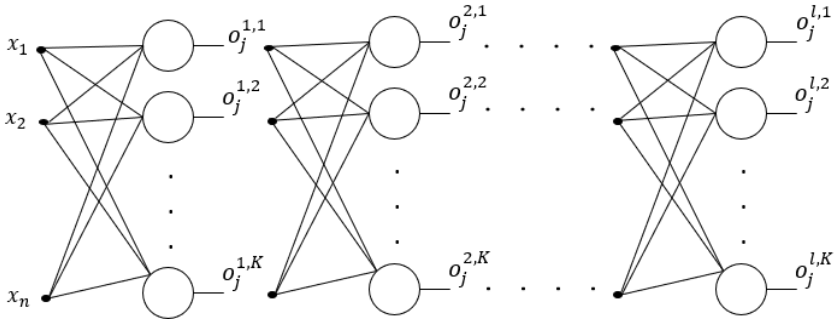
Ένα σύνολο K παράλληλα διατεταγμένων τεχνητών νευρώνων όπως αναπαρίσταται στην εικόνα (1.3) συνιστούν ένα τεχνητό νευρωνικό δίκτυο ενός επιπέδου (Single Layer Neural Network). Ένα πολυεπίπεδο τεχνητό νευρωνικό δίκτυο (Multilayer Neural Network) είναι ένα δίκτυο το οποίο περιλαμβάνει περισσότερα από ένα τεχνητά νευρωνικό δίκτυα ενός επιπέδου. Η διαδικασία εκπαίδευσης ενός τεχνητού νευρωνικού δικτύου αφορά την διαδικασία προσδιορισμού των τιμών των παραμέτρων των νευρώνων του.



1. 3 Τεχνητό Νευρωνικό Δίκτυο ενός επιπέδου

Σε ένα πολυεπίπεδο τεχνητό νευρωνικό δίκτυο – όπως αυτό της εικόνας (1.4) οι τιμές ενεργοποίησης των νευρώνων ενός επιπέδου αποτελούν τις εισόδους του αμέσως επόμενου. Το κάθε επίπεδο του δικτύου ενδέχεται να απαρτίζεται από διαφορετικό πλήθος νευρώνων, ωστόσο κάθε νευρώνας έχει πάντα μια τιμή ενεργοποίησης. Όταν τα συνεχόμενα επίπεδα ενός δικτύου τροφοδοτούν το ένα το άλλο στην με τέτοιο τρόπο ώστε η έξοδος του ενός να αποτελεί την είσοδο του αμέσως επόμενου και επιπλέον οι υπολογισμοί των τιμών ενεργοποίησης των νευρώνων διεξάγονται προς μια μόνο κατεύθυνση (από την είσοδο προς την έξοδο του δικτύου) απουσία ανατροφοδότησης τότε τα πολυεπίπεδα δίκτυα χαρακτηρίζονται δίκτυα εμπρόσθιας τροφοδότησης (Feed-Forward Neural Networks). Στην περίπτωση όπου όλοι οι νευρώνες ενός επιπέδου είναι συνδεδεμένοι με όλους τους νευρώνες του επομένου επιπέδου τότε το δίκτυο εμπρόσθιας διάδοσης ονομάζεται πλήρως συνδεδεμένο (fully connected) διαφορετικά καλείται μερικώς συνδεδεμένο (partially connected).

Σε ένα πολυεπίπεδο τεχνητό νευρωνικό δίκτυο οι νευρώνες του πρώτου επιπέδου συνιστούν το επίπεδο εισόδου (Input layer) του δικτύου, ενώ οι νευρώνες του τελευταίου επιπέδου – απ' όπου προκύπτει και η έξοδος ή οι έξοδοι του νευρωνικού δικτύου- συνιστούν το επίπεδο εξόδου (output layer) του δικτύου. Κάθε επίπεδο νευρώνων που εντοπίζεται ενδιάμεσα των επιπέδων εισόδου και εξόδου του δικτύου αποτελεί κρυφό επίπεδο του δικτύου. Σε γενικές γραμμές, τα νευρωνικά δίκτυα με ένα μόνο κρυφό στρώμα χαρακτηρίζονται ρηχά νευρωνικά δίκτυα, ενώ τα νευρωνικά δίκτυα που απαρτίζονται από δυο ή περισσότερα κρυφά στρώματα νευρώνων ονομάζονται βαθιά νευρωνικά δίκτυα (Deep Neural Networks – DNNs) (Gonzalez & Woods, 2011). Η διαδικασία εκπαίδευσης των Βαθέων Τεχνητών Νευρωνικών Δικτύων ονομάζεται Βαθιά Μάθηση (Deep Learning) (Aggarwal, 2018).



1. 4 Πολυεπίπεδο Τεχνητό Νευρωνικό Δίκτυο

Η σημερινή πληθώρα των διαθέσιμων επισημειωμένων δεδομένων που είναι απαραίτητα κατά τη διαδικασία της εκπαίδευσης των τεχνητών νευρωνικών δικτύων σε συνδυασμό με την ολοένα αυξανόμενη βελτίωση της υπολογιστικής ισχύος των μονάδων επεξεργασίας (CPUs και GPUs) έχουν συμβάλει με τρόπο καθοριστικό στην ραγδαία καθιέρωση των τεχνητών νευρωνικών δικτύων ως μια από τις πλέον αξιόπιστες και ακριβείς μεθόδους Βαθιάς Μάθησης (Gu et al., 2018).

Μεταξύ των ποικίλων τύπων Τεχνητών Νευρωνικών Δικτύων που έχουν προταθεί στη βιβλιογραφία, τα Συνελικτικά Νευρωνικά Δίκτυα (Convolution Neural Networks - CNNs) αποτελούν μέχρι στιγμής την πιο διαδεδομένη και μελετημένη κατηγορία Βαθέων Νευρωνικών Δικτύων (Deep Neural Networks – DNNs) (Gu et al., 2018).

Τα CNN επιτυγχάνουν γενικά υψηλότερα αποτελέσματα από ανάλογες μηχανές μάθησης σ’ ένα ευρύ φάσμα εφαρμογών που εκτείνεται μεταξύ των πεδίων της επεξεργασίας σημάτων, της υπολογιστικής όρασης και της ανάλυσης εικόνων. Η επεξεργασία ομιλίας και φυσικής γλώσσας, η ανίχνευση και αναγνώριση κειμένου, η αναγνώριση της ταυτότητας αντικειμένων και παρακολούθηση της τροχιάς τους, η εκτίμηση της θέσης αντικειμένων, η αναγνώριση κίνησης, η ανίχνευση οπτικά σημαντικών σημείων σε εικόνες και βίντεο, η ταξινόμηση εικόνων, ακόμη και η ιατρική ανάλυση εικόνων με απώτερο στόχο την υποβοήθηση της ιατρικής διάγνωσης είναι μόνο μερικά παραδείγματα επιτυχημένων εφαρμογών των συνελικτικών νευρωνικών δίκτυών (Gu et al., 2018).

1.3. Επεξεργασία & Ανάλυση Εικόνας με Συνελικτικά Νευρωνικά Δίκτυα

Ο τρόπος λειτουργίας των Συνελικτικών Νευρωνικών Δίκτυων, απορρέει από τον τρόπο με τον οποίο αντιλαμβάνονται και επεξεργάζονται την οπτική πληροφορία οι έμβιοι οργανισμοί (Gu et al., 2018).

Τα συνελικτικά δίκτυα είναι σχεδιασμένα ώστε να δέχονται και να επεξεργάζονται δεδομένα τα οποία είναι οργανωμένα σε μορφή πινάκων, όπως ακριβώς αναπαρίστανται και οι ψηφιακές εικόνες (Aggarwal, 2018).

Συγκεκριμένα μια ψηφιακή εικόνα παριστάνεται από ένα δισδιάστατο πίνακα ακέραιων αριθμών $I(i,j)$, διαστάσεων $N \times M$, όπου $i=1,\ldots,N$ και $j=1,\ldots,M$. Μια επιπλέον διάσταση προστίθεται στην παραπάνω αναπαράσταση προκειμένου να κωδικοποιήσει το βάθος χρώματος της εικόνας. Τελικά, μια εικόνα αναπαρίσταται ως ένας τρισδιάστατος πίνακας διαστάσεων $N \times M \times C$, όπου N, M αναφέρονται στο μήκος και στο πλάτος της εικόνας, ενώ η μεταβλητή C εκφράζει το πλήθος των χρωματικών καναλιών της (Παπαμάρκος, 2013).

Ένα συνελικτικό νευρωνικό δίκτυο είναι ένα νευρωνικό δίκτυο το όποιο συντίθεται από ένα ή περισσότερα συνελικτικά επίπεδα (Convolutional Layers) (Skansi, 2018). Συνήθως μεταξύ των συνελικτικών επιπέδων παρεμβάλλονται : επίπεδα απαλοιφής της γραμμικότητας (Non Linearity Layers), συγκεντρωτικά επίπεδα (Pooling Layers) ενώ στο πέρας αυτών συναντώνται κάποια επίπεδα πλήρως συνδεδεμένων νευρώνων (Fully Connected Layers) (Aggarwal, 2018). Επί της ουσίας, ο τρόπος λειτουργίας ενός συνελικτικού νευρωνικού δικτύου είναι συμβατός των δικτύων νευρωνικών εμπρόσθιας τροφοδότησης.

Κάθε συνελικτικό επίπεδο στο δίκτυο αποσκοπεί στην εξαγωγή χαρακτηριστικών από την εικόνα η οποία επιτυγχάνεται με την εφαρμογή μιας σειράς εκπαιδεύσιμων φίλτρων επί της εικόνας (διαδικασία συνέλιξης). Από την εφαρμογή της συνέλιξης μεταξύ κάθε εικόνας και των φίλτρων ενός συνελικτικού επιπέδου παράγεται ως έξοδος ένα σύνολο εικόνων, οι οποίες καλούνται χάρτες χαρακτηριστικών (Feature Maps). Οι χάρτες χαρακτηριστικών στην έξοδο κάθε επιπέδου, είναι ισάριθμοι του πλήθους των φίλτρων του. Το γεγονός ότι τα φίλτρα κάθε επιπέδου είναι εκπαιδεύσιμα αυτοματοποιεί την διαδικασία εξαγωγής χαρακτηριστικών και κατ' επέκταση της μάθησης.

Τα επίπεδα απαλοιφής της γραμμικότητας, χρησιμοποιούνται προκειμένου να αυξηθεί η μη-γραμμικότητα των αποφάσεων, τις οποίες εκπαιδεύεται να λαμβάνει το δίκτυο. Με άλλα λόγια, συμβάλλουν ούτως ώστε το εκπαιδευμένο μοντέλο να είναι σε θέση να προβλέπει με ακρίβεια την έξοδο του δικτύου ακόμη και όταν τα δεδομένα εισόδου είναι ιδιαίτερα σύνθετα.

Τα συγκεντρωντικά επίπεδα συνεισφέρουν στη μείωση της χωρικής διάστασης της εξόδου των συνελικτικών επιπέδων και κατά συνέπεια και των παραμέτρων του δικτύου. Σημαντική ιδιότητα αυτών των επίπεδων αποτελεί το γεγονός ότι παρά την οποία μείωση

πραγματοποιείται η σημαντική πληροφορία της εισόδου ενός συγκεντρωτικού επιπέδου τελικά διατηρείται.

Τέλος, τα πλήρως συνδεδεμένα επίπεδα επιτελούν ίδιες λειτουργίες με εκείνες των κλασσικών τεχνητών νευρωνικών δικτύων εμπρόσθιας τροφοδότησης (Feed-Forward Neural Networks).

1.4. Στόχοι της εργασίας

Η ανίχνευση σημαντικών αντικειμένων (Salient Object Detection - SOD), εμπνευσμένη από τον βιολογικό μηχανισμό της οπτικής προσοχής, αποσκοπεί στον εντοπισμό περιοχών των εικόνων όπου απαντώνται αντικείμενα στα οποία συγκεντρώνεται η οπτική προσοχή στην πρώτη παρατήρηση μιας οπτικής αναπαράστασης. Με άλλα λόγια πρόκειται για τα αντικείμενα εκείνα των εικόνων που θεωρείται ότι ξεχωρίζουν από το περιβάλλον τους. Το πεδίο της ανίχνευσης των σημαντικών αντικειμένων βρίσκει εφαρμογές σε ένα ευρύ φάσμα εργασιών επιπέδου αντικειμένων (object-level applications) τόσο στην υπολογιστική όραση όσο και στην ρομποτική (W. Wang et al., 2019).

Τα τελευταία χρόνια η ραγδαία ανάπτυξη των τεχνικών Βαθιάς Μάθησης έχει συμβάλλει καθοριστικά στην εξέλιξη του εν λόγω πεδίου. Ειδικότερα, τα συνελικτικά νευρωνικά δίκτυα με το τεχνολογικό υπόβαθρο που παρέχουν επιτυγχάνουν απαράμιλλη απόδοση σε εργασίες ανίχνευσης σημαντικών αντικειμένων σε σχέση με τους παραδοσιακούς ευριστικούς αλγορίθμους (Ullah et al., 2020). Για το λόγο αυτό τα συνελικτικά δίκτυα συνιστούν σήμερα την κυρίαρχη τάση αναφορικά με την ανίχνευση των σημαντικών αντικειμένων (Borji, Cheng, Hou, Jiang, & Li, 2019).

Πρόσφατα, ερευνητικές εργασίες προσανατολισμένες στην ανίχνευση σημαντικών αντικειμένων (Lang et al., 2012), (Niu, Geng, Li, & Liu, 2012) έχουν αρχίσει να λαμβάνουν υπόψη την πληροφορία του βάθους μιας οπτικής αναπαράστασης προκειμένου να παρέχουν όσο το δυνατόν πιο συνεπή αποτελέσματα ως προς τη βιολογικό μηχανισμό της οπτικής προσοχής. Η πληροφορία του βάθους μιας οπτικής σκηνής, στις μεθοδολογίες που αναπτύσσονται (RGB-D Salient Object Detection methods) ανακτάται είτε μέσω συστημάτων στερεοσκοπικών καμερών είτε μέσω αισθητήρων βάθους (Borji et al., 2019).

Η παρούσα εργασία αποσκοπεί:

i. Στην διερεύνηση υπολογιστικών αναπαραστάσεων της πληροφορίας του βάθους με απώτερο σκοπό την αξιοποίηση της σε εργασίες ανίχνευσης σημαντικών αντικειμένων. Για

την ανάκτηση της υπολογιστικής πληροφορίας του βάθους αξιοποιούνται προεκπαιδευμένες αρχιτεκτονικές συνελικτικών δικτύων. Επιπρόσθετα, σε μια προσπάθεια αναζήτησης εναλλακτικών υπολογιστικών προσεγγίσεων, η πληροφορία του βάθους που υπολογίζεται από τα προεκπαιδευμένα συνελικτικά δίκτυα συνδυάζεται με κανόνες ασαφούς λογικής.

ii. Στην αξιολόγηση της συμβολής των υπολογιστικών αναπαραστάσεων της πληροφορίας του βάθους αναφορικά με την υποβοήθηση της ανίχνευσης σημαντικών αντικειμένων.

iii. Στην αποτίμηση της συνεισφοράς των συνελικτικών νευρωνικών δικτύων συγκριτικά με τις συμβατικές μεθόδους ανάκτησης πληροφορίας του βάθους (κάμερες/αισθητήρες) στις εργασίες ανίχνευσης σημαντικών αντικειμένων.

1.5. Συνεισφορά της εργασίας

Πρόσφατα τα συνελικτικά νευρωνικά δίκτυα έχουν αξιοποιηθεί για τη μονοφθαλμική εκτίμηση της πληροφορίας του βάθους μιας οπτικής αναπαράστασης (Alhashim & Wonka, 2018), (Tateno, Tombari, Laina, & Navab, 2017). Ο όρος μονοφθαλμική εκτίμηση υποδηλώνει ότι η εκπαίδευση των συνελικτικών δικτύων διεξάγεται αποκλειστικά με μια σειρά εικόνων προερχόμενες από βίντεο χωρίς να απαιτείται κάποια επιπρόσθετη πληροφορία. Παρά το γεγονός ότι τα προαναφερόμενα μοντέλα είναι σε θέση να εκτιμούν την πληροφορία του βάθους μιας σκηνής με ακρίβεια που προσεγγίζει εκείνη των καμερών και των αισθητήρων βάθους, βιβλιογραφικά δε φαίνεται να επαληθεύεται η αξιολόγηση τους συναρτήσει της ανίχνευσης σημαντικών αντικειμένων.

Η παρούσα εργασία επιχειρεί να διερευνήσει αυτό το κενό και προτείνει ένα μοντέλο για την υποβοηθούμενη από την πληροφορία του βάθους ανίχνευση σημαντικών αντικειμένων σε εικόνες. Η πρωτοτυπία της εν λόγω μεθοδολογίας έγκειται στο γεγονός ότι η πληροφορία του βάθους που αξιοποιείται σε αντίθεση με τις συμβατικές προσεγγίσεις εκτιμάται από προεκπαιδευμένα συνελικτικά νευρωνικά δίκτυα. Ακόμα, σε χρόνο μεταγενέστερο της εκπαίδευσης του συνελικτικού δικτύου που είναι επιφορτισμένο με την ανίχνευση των σημαντικών αντικειμένων, ένας επιπρόσθετος επιδιορθωτικός μηχανισμός υιοθετείται προκειμένου να συνδράμει στην περαιτέρω βελτίωση των αποτελεσμάτων.

Αξίζει να σημειωθεί ότι μέρος της έρευνας που διεξήχθη στα πλαίσια της παρούσας εργασίας, υποβλήθηκε προς αξιολόγηση για δημοσίευση σε διεθνές επιστημονικό συνέδριο, και είναι ακόμα υπό κρίση:

G. Dimas, P. Gatoula, D. K. Iakovidis, *MonoSOD: Monocular Salient Object Detection based on Predicted Depth*, IEEE International Conference on Robotics and Automation, 2021, Submitted

1.6. Δομή της εργασίας

Η διάρθρωση του υπόλοιπου της εργασίας εκτείνεται σε έξι ενότητες.

Στην ενότητα 2, παρουσιάζεται η δομή που ακολουθούν τα συνελικτικά νευρωνικά δίκτυα και διασαφηνίζεται το πλαίσιο λειτουργίας τους. Στην ενότητα 3, αναλύεται το αντικείμενο που πραγματεύεται η ανίχνευση οπτικά σημαντικών αντικειμένων και διεξάγεται βιβλιογραφική ανασκόπηση των μεθόδων που εφαρμόζονται για την ανίχνευση οπτικά σημαντικών αντικειμένων τόσο σε RGB εικόνες, όσο και σε RGB-D εικόνες. Στην ενότητα 4, περιγράφεται λεπτομερώς η μεθοδολογία που συστήνει η παρούσα εργασία. Στην ενότητα 5, δίνονται διευκρινίσεις αναφορικά με τον τρόπο αξιολόγησης της μεθοδολογίας που υπαγορεύει η προηγούμενη ενότητα καθώς και εκτίθενται τα αποτελέσματα που προέκυψαν. Επιπλέον, πραγματοποιείται σύγκριση των αποτελεσμάτων που προέκυψαν, με αποτελέσματα που παρέχουν μεθοδολογίες αντιπροσωπευτικές για την υφιστάμενη κατάσταση αναφορικά με το πρόβλημα της υποβοηθούμενης από την πληροφορία του βάθους ανίχνευσης σημαντικών αντικειμένων σε εικόνες.

Στην τελευταία ενότητα, συνοψίζονται τα συμπεράσματα τα οποία διαμορφώθηκαν από τα αποτελέσματα της προτεινόμενης μεθοδολογίας. Επίσης, στην ίδια ενότητα παρέχονται περαιτέρω κατευθύνσεις για τη βελτίωση της υπό εξέταση μεθοδολογίας.

2. Συνελικτικά Νευρωνικά Δίκτυα

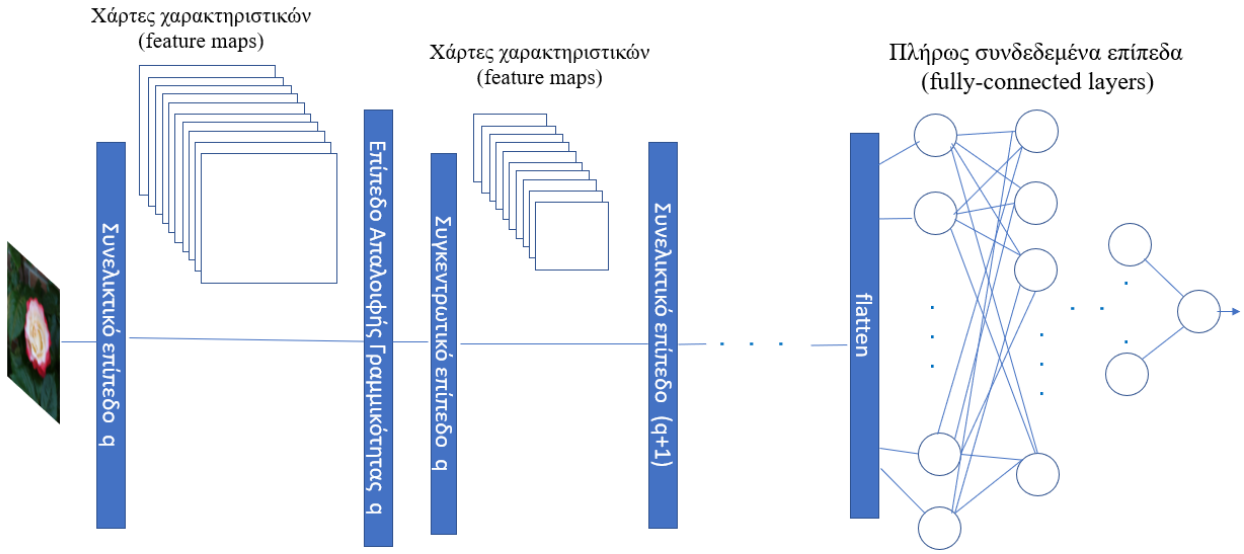
2.1. Βιολογικά και Τεχνητά Νευρωνικά Δίκτυα

Όπως αναφέρθηκε και στην εισαγωγή, ο τρόπος λειτουργίας των συνελικτικών νευρωνικών δικτύων είναι εμπνευσμένος από τον μηχανισμό με τον οποίο οι έμβιοι οργανισμοί επεξεργάζονται την οπτική πληροφορία. Η εργασία των Hubel & Wiesel (Hubel & Wiesel, 1959), που αφορούσε τη μελέτη του οπτικού συστήματος των θηλαστικών, συνέβαλε καθοριστικά στην αποκρυπτογράφηση του μηχανισμού λειτουργίας του οπτικού συστήματος και προσέφερε ευρήματά τα οποία αποτέλεσαν την έμπνευση για τον σχεδιασμό των συνελικτικών δικτύων. Οι συγγραφείς συμπέραναν ότι συγκεκριμένα οπτικά ερεθίσματα επάγουν τη λειτουργία καθορισμένων νευρώνων που εντοπίζονται στον οπτικό φλοιό των θηλαστικών. Επιπλέον, η εν λόγω εργασία αποκάλυψε, ότι τόσο ο προσανατολισμός όσο και το σχήμα των αντικειμένων μιας οπτικής αναπαράστασης προκαλούν τη διέγερση διαφορετικών νευρώνων, επηρεάζοντας με τρόπο καταλυτικό τη νευρική απόκριση. Ακόμη, η διαπίστωση ότι οι νευρώνες του οπτικού φλοιού συνδέονται ακολουθώντας μια διαστρωματωμένη αρχιτεκτονική, ενίσχυσε την εικασία πως τα θηλαστικά είναι σε θέση αντιλαμβάνονται σε διάφορα επίπεδα αφαιρετικότητας, το περιεχόμενο των οπτικών αναπαραστάσεων.

Θα μπορούσαμε να ισχυριστούμε ότι από την πλευρά της μηχανικής μάθησης, η ιεραρχική εξαγωγή χαρακτηριστικών επιχειρεί να προσομοιάσει τη τελευταία νευροφυσιολογική λειτουργία. Τα συνελικτικά δίκτυα, τα οποία επίσης ακολουθούν μια διαστρωματωμένη αρχιτεκτονική, αποπειρώνται να μιμηθούν τον προαναφερόμενο μηχανισμό, κωδικοποιώντας χαρακτηριστικά χαμηλού επιπέδου των εικόνων στα κατώτερα στρώματα του δικτύου, ενώ η υψηλού επιπέδου πληροφορία των εικόνων συλλαμβάνεται από τα ανώτερα στρώματα. Η αρχιτεκτονική των συνελικτικών δικτύων, η οποία είναι εμπνευσμένη από τον επιλεκτικό και εντοπισμένο τρόπο σύνδεσης των νευρώνων των θηλαστικών, εφαρμόζει αραιές τοπικές συνδέσεις και ταυτόχρονα επιτυγχάνει αρκετά υψηλό διαμοιρασμό των παραμέτρων που χρησιμοποιεί. Εκμεταλλεύεται -κατά τη διαδικασία ανάκτησης χαρακτηριστικών- το γεγονός ότι τα δεδομένα εισόδου, εφόσον αναπαρίσταται υπό τη μορφή πλέγματος, παρουσιάζουν υψηλή χωρική εξάρτηση σε τοπικό επίπεδο. Επιπλέον, ένα συνελικτικό δίκτυο, δεν εφαρμόζει άκριτη σύνδεση των χαρακτηριστικών που εξήχθησαν από τα διάφορα επίπεδα του. Αντί αυτού, συσχετίζει τα τοπικά εντοπισμένα χαρακτηριστικά κάθε επιπέδου, με μια συγκεκριμένη

περιοχή του αμέσως προηγούμενου επιπέδου. Με τον τρόπο αυτό οι χωρικές εξαρτήσεις κληροδοτούνται από τα ρηχότερα στρώματα του δικτύου προς τα βαθύτερα. Συγχρόνως, κατά την εξαγωγή των τοπικών χαρακτηριστικών, κάθε επίπεδο χρησιμοποιεί ένα σύνολο κοινών παραμέτρων ολόκληρη την εικόνα. Με τον τρόπο αυτό, επιτυγχάνεται τελικά σημαντική μείωση στο πλήθος των παραμέτρων που χρησιμοποιούνται στα συνελικτικά δίκτυα, έναντι των τυπικών νευρωνικών δικτύων (Aggarwal, 2018).

2.2. Επισκόπηση της Αρχιτεκτονικής των Συνελικτικών Νευρωνικών Δικτύων



Εικόνα 2. 1 Αρχιτεκτονική ενός τυπικού συνελικτικού νευρωνικού δικτύου

Η λειτουργία ενός συνελικτικού νευρωνικού δικτύου μοιάζει αρκετά με εκείνη των παραδοσιακών δικτύων εμπρόσθιας τροφοδότησης (Feed-Forward Networks). Η ειδοποιός διαφορά τους είναι ότι τα επίπεδα ενός συνελικτικού δικτύου, εξάγουν χαρακτηριστικά σε τοπικό επίπεδο και συνδέονται μεταξύ τους χρησιμοποιώντας αραιές, ωστόσο προσεκτικά σχεδιασμένες συζεύξεις (Aggarwal, 2018).

Ένα συνελικτικό δίκτυο λαμβάνει στην είσοδο του έναν πίνακα (τανυστή). Ειδικότερα, οι εικόνες αναπαρίστανται ως πίνακες τριών διαστάσεων $W \times H \times C$, όπου οι μεταβλητές W, H αναφέρονται στις διαστάσεις της εικόνας, ενώ η μεταβλητή C αφορά το πλήθος των χρωματικών καναλιών της. Στο επίπεδο εισόδου του δικτύου δε πραγματοποιείται κάποιου είδους επεξεργασία στα δεδομένα της εισόδου. Ακολούθως, η είσοδος του δικτύου προωθείται στα επόμενα επίπεδα που το απαρτίζουν, όπου σε καθένα από αυτά υφίσταται μια σειρά διαφορετικών και διαδοχικών ειδών επεξεργασίας.

Η θεμελιώδης δομική μονάδα των συνελικτικών δικτύων αποτελούν τα συνελικτικά επίπεδα. Ένα συνελικτικό επίπεδο εφαρμόζει την μαθηματική πράξη της συνέλιξης με απώτερο σκοπό την εξαγωγή τοπικά εντοπισμένων χαρακτηριστικών. Η συνέλιξη, ορίζεται ως πράξη εσωτερικού γινομένου (dot-product) μεταξύ ενός συνόλου παραμέτρων, που καλούνται βάρη του επιπέδου, και των όλων των δυνατών διαφορετικών τοπικών περιοχών μιας εικόνας. Η πράξη της συνέλιξης αποδεικνύεται ιδιαίτερα χρήσιμη για την απόκτηση χαρακτηριστικών στην περίπτωση που τα δεδομένα χαρακτηρίζονται από υψηλή χωρική εξάρτηση, όπως ακριβώς συμβαίνει με τις εικόνες. Αυτό οφείλεται στο γεγονός ότι επιτρέπει την εξαγωγή όμοιας πληροφορίας από περιοχές των εικόνων που φέρουν παρόμοια πρότυπα καθώς και εξαλείφει την εξάρτηση των δεδομένων εισόδου τόσο από την ύπαρξη θορύβου όσο και από τον μετασχηματισμό της μετατόπισης ως προς κάποια διεύθυνση.

Πέραν των συνελικτικών επιπέδων, ένα δίκτυο αυτής της κατηγορίας περιλαμβάνει επίπεδα απαλοιφής της γραμμικότητας και συγκεντρωτικά επίπεδα. Τα προαναφερθέντα επίπεδα παρεμβάλλονται μεταξύ δυο συνελικτικών επιπέδων του δικτύου. Την αρχιτεκτονική ενός συνελικτικού δικτύου συμπληρώνει ένα σύνολο πλήρως συνδεδεμένων επίπεδων, όπως εκείνα που συναντώνται στα τυπικά νευρωνικά δίκτυα. Η λειτουργικότητα των πλήρως συνδεδεμένων επιπέδων ορίζεται κάθε φορά συναρτήσει της εφαρμογής - προβλήματος το οποίο αναλαμβάνει να επιλύσει το συνελικτικό δίκτυο.

Εν συνεχεία θα παρουσιαστεί το είδος της επεξεργασίας που λαμβάνει χώρα από κάθε τύπο επιπέδου ενός συνελικτικού δικτύου.

2.2.1. Συνελικτικά Επίπεδα

Κάθε συνελικτικό επίπεδο q ενός ομώνυμου δικτύου φέρει ένα σύνολο τρισδιάστατων δομικών μονάδων, οι οποίες καλούνται φίλτρα ή αλλιώς πυρήνες. Κάθε φίλτρο έχει μήκος, πλάτος και βάθος. Η έννοια του βάθους ενός φίλτρου αναφέρεται στον όγκο της εισόδου που δέχεται το εκάστοτε συνελικτικό επίπεδο q. Είναι προφανές ότι ο όγκος της εισόδου που αντιστοιχεί στο πρώτο συνελικτικό επίπεδο του δικτύου, για το οποίο ισχύει ότι $q = 1$, ισούται με το πλήθος των χρωματικών καναλιών των εικόνων του συνόλου εκπαίδευσης. Τα φίλτρα ενός επιπέδου q, έχουν χωρικές διαστάσεις -μήκος και πλάτος- μικρότερες των αντίστοιχων χωρικών διαστάσεων του όγκου της εισόδου του. Επιπλέον, τα φίλτρα είναι συνήθως τετραγωνικά και οι διαστάσεις τους όσον αφορά το μήκος και το πλάτος, είναι περιττοί αριθμοί

(Aggarwal, 2018). Οι παραπάνω προϋποθέσεις μας επιτρέπουν να ορίσουμε τις διαστάσεις των φίλτρων για το q επίπεδο του δικτύου ως εξής:

$$F_q \times F_q \times d_q \quad \textbf{(2.1)}$$

όπου F_q οι χωρικές διαστάσεις των φίλτρων και d_q το βάθος τους.

Σε κάθε συνελικτικό επίπεδο του δικτύου q, κάθε φίλτρο του, συνελίσσεται με όλο τον όγκο των δεδομένων εισόδου του επιπέδου. Ο όγκος της εισόδου για το επίπεδο q ενός δικτύου έχει διαστάσεις που ορίζονται από την ακόλουθη σχέση:

$$L_q \times B_q \times d_q \quad \textbf{(2.2)}$$

όπου L_q το μήκος του όγκου της εισόδου, όπου B_q το πλάτος του όγκου της εισόδου, και d_q το βάθος του. Στο σημείο αυτό, να επισημάνουμε για ακόμη μια φορά ότι το βάθος του όγκου της εισόδου ενός συνελικτικού επιπέδου ισούται με το βάθος των φίλτρων του.

Αναλυτικότερα, καθένα από τα φίλτρα ενός επιπέδου q σαρώνει κατά πλάτος και κατά μήκος τον πίνακα εισόδου του επιπέδου σε όλο τον όγκο του και με τον τρόπο αυτό υπολογίζονται τα εσωτερικά γινόμενα μεταξύ των τιμών του φίλτρου και των τιμών της εισόδου στο επίπεδο q, για οποιαδήποτε θέση της. Από την απόκριση που αντιστοιχεί στην κάθε χωρική θέση του πίνακα εισόδου, από την οποία διήλθε ένα φίλτρο, παράγεται μια δισδιάστατη αναπαράσταση. Η αναπαράσταση αυτή καλείται χάρτης χαρακτηριστικών (Feature Map) ή εναλλακτικά χάρτης ενεργοποίησης (Activation Map). Γίνεται εύκολα αντιληπτό, ότι σε κάθε συνελικτκό επίπεδο παράγονται τόσοι χάρτες χαρακτηριστικών όσο και το πλήθος των φίλτρων που χρησιμοποιούνται σε αυτό. Το γεγονός αυτό μας επιτρέπει να ισχυριστούμε ότι ο αριθμός των φίλτρων που αξιοποιούνται στο κάθε συνελικτικό επίπεδο του δικτύου, καθορίζει το πλήθος των παραμέτρων για το επίπεδο αυτό και κατά συνέπεια ελέγχει τη χωρητικότητα του συνολικού μοντέλου, δηλαδή τον βαθμό στον οποίο αυτό μπορεί να μαθαίνει από οποιοδήποτε σύνολο εκπαίδευσης, χωρίς σφάλμα.

Οι δισδιάστατοι χάρτες χαρακτηριστικών που προκύπτουν από ένα συνελικτικό επίπεδο στοιβάζονται ως προς την τρίτη διάσταση, προκειμένου να παραχθεί ένας τρισδιάστατος όγκος εξόδου. Ο τρισδιάστατος όγκος εξόδου που παράγεται από ένα συνελικτικό επίπεδο αποτελεί τον τρισδιάστατο όγκο εισόδου του κατά σειρά αμέσως επόμενου συνελικτικού επιπέδου.

Έστω ότι ο τρισδιάστατος τανυστής W που ορίζεται από την ακόλουθη σχέση:

$$W^{(p,q)} = \left[w_{i,j,k}{}^{(p,q)}\right] \quad \textbf{(2.3)}$$

φέρει τις παραμέτρους (βάρη) που αντιστοιχούν στο υπ' αριθμό p φίλτρο του q συνελικτικού επιπέδου και οι δείκτες i, j, k, υποδεικνύουν τις θέσεις που αφορούν το μήκος, το πλάτος και το βάθος του φίλτρου.

Αν υποθέσουμε ότι ο τρισδιάστατος τανυστής H που ορίζεται από την παρακάτω σχέση:

$$H^{(q)} = \left[h_{i,j,k}{}^{(q)}\right] \tag{2.4}$$

περιλαμβάνει τους χάρτες χαρακτηριστικών που παρήχθησαν στο επίπεδο q και για $q=1$ ο τανυστής περιέχει τα τρισδιάστατα δεδομένα εισόδου του δικτύου, τότε οι συνελίξεις στο $(q+1)$ επίπεδο του δικτύου ορίζονται από τη σχέση:

$$h_{ijp}{}^{(q+1)} = \sum_{r=1}^{F_q}\sum_{s=1}^{F_q}\sum_{k=1}^{d_q} w_{rsk}{}^{(p,q)}\, h_{i+r-1,j+s-1,k}{}^{(q)}\,, \tag{2.5}$$

$$\forall\, i \in \{1 \ldots L_q - F_q + 1\},$$

$$\forall\, j \in \{1 \ldots B_q - F_q + 1\},$$

$$\forall\, p \in \{1 \ldots d_{q+1}\}$$

Σε αναλογία με ένα βιολογικό νευρωνικό δίκτυο, κάθε τιμή στον τρισδιάστατο τανυστή $H^{(q)}$, ο οποίος περιλαμβάνει τον τρισδιάστατο όγκου εξόδου που παράγεται από το συνελικτικό επίπεδο q του δικτύου, μπορεί να θεωρηθεί ως έξοδος της επεξεργασίας που λαμβάνει χώρα σε ένα νευρώνα. Κάθε νευρώνας ενός συνελικτικού επιπέδου q εστιάζει κάθε φορά σε μια συγκεκριμένη περιοχή του όγκου εισόδου που αντιστοιχεί στο επίπεδο q και συγχρόνως μοιράζεται τις ακριβώς ίδιες παραμέτρους με τους γειτονικούς νευρώνες του ίδιου επιπέδου. Με τον τρόπο αφενός αποφεύγεται η σύνδεση όλων των νευρώνων του επιπέδου (q+1) με όλους τους νευρώνες του αμέσως προηγούμενου επιπέδου q, αφετέρου επιτυγχάνεται μόνο η συσχέτιση κάθε νευρώνα στο επίπεδο (q+1) με μια συγκεκριμένη χωρική περιοχή του προηγούμενου επιπέδου q. Το μέγεθος της χωρικής περιοχής στην οποία εστιάζει κάθε φορά ένας νευρώνας, συνιστά μια υπερ-παράμετρο του δικτύου η οποία ονομάζεται δεκτικό πεδίο του νευρώνα (receptive field). Με την προσθήκη κάθε επιπλέον συνελικτικού επιπέδου σε ένα δίκτυο, αυξάνεται το δεκτικό πεδίο των νευρώνων του αναφορικά με τον αρχικό όγκο εισόδου του δικτύου. Το συμβάν αυτό, αποτελεί την εξήγηση στο γεγονός ότι ένα συνελικτικό δίκτυο απαθανατίζει χαρακτηριστικά χαμηλού επιπέδου στα κατώτερα στρώματα του και χαρακτηριστικά υψηλότερου επιπέδου στα βαθύτερα στρώματα.

Επιπλέον, κάθε νευρώνας ενός συνελικτικού επίπεδου πέραν του συνελικτικού πυρήνα p, διαθέτει και μια τιμή πόλωσης (bias). Η τιμή της πόλωσης προστίθεται κάθε φορά στο αποτέλεσμα του κουκκιδωτού προϊόντος που παράγεται από την πράξη της συνέλιξης.

Το πλήθος των παραμέτρων κάθε νευρώνα στο συνελικτικό επίπεδο q εκτιμάται ως εξής: από τη σχέση (2.2), γνωρίζουμε ότι ο όγκος της εισόδου στο επίπεδο q ενός δικτύου έχει διαστάσεις $L_q \times B_q \times d_q$. Από τη σχέση (2.1), $F_q \times F_q \times d_q$, η οποία υποδεικνύει τις διαστάσεις των φίλτρων του επιπέδου q, συνάγουμε ότι το δεκτικό πεδίο του νευρώνα στο επίπεδο q ισούται με $F_q \times F_q$. Κάθε νευρώνας στο συνελικτικό επίπεδο q θα έχει $F_q \times F_q \times d_q$ βάρη, δηλαδή παραμέτρους . Πέρα από τα βάρη κάθε νευρώνας, διαθέτει και μια τιμή πόλωσης. Συνεπώς, ο τελικός αριθμός των παραμέτρων που διαθέτει κάθε νευρώνας στο συνελικτικό επίπεδο q, ισούται με:

$$(F_q \times F_q \times d_q) + 1 \qquad \textbf{(2.6)}$$

Αν και ακόμη δεν έχουμε ορίσει τις διαστάσεις του όγκου εξόδου για το συνελικτικό επίπεδο q, έχουμε επισημάνει ήδη ότι το επίπεδο q παράγει τόσους χάρτες χαρακτηριστικών όσο και το πλήθος των φίλτρων/πυρήνων που χρησιμοποιούνται σε αυτό. Όπως σημειώσαμε νωρίτερα, όλοι οι νευρώνες ενός πυρήνα μοιράζονται ακριβώς τις ίδιες παραμέτρους. Δεδομένων των παραπάνω, αν υποθέσουμε ότι το συνολικό πλήθος των πυρήνων ενός επιπέδου q, ισούται με p, τότε το συνολικό πλήθος των παραμέτρων που χρησιμοποιεί το επίπεδο αυτό εκφράζεται από την παρακάτω σχέση :

$$p \times [(F_q \times F_q \times d_q) + 1] \qquad \textbf{(2.7)}$$

Προτού καθορίσουμε με σαφήνεια τις διαστάσεις του όγκου εξόδου ενός συνελικτικού επιπέδου, οφείλουμε να αναφερθούμε σε δυο υπερ-παραμέτρους των συνελικτικών επιπέδων. Πρόκειται για τον διασκελισμό ή αλλιώς βήμα (stride) και το γέμισμα (padding).

Η υπερ-παράμετρος που ονομάζεται βήμα υποδεικνύει τον αριθμό των χωρικών μονάδων, δηλαδή τον αριθμό των εικονοστοιχείων, που θα μετακινείται το δεκτικό πεδίο του πυρήνα p του συνελικτικού επιπέδου q, κατά τη διάρκεια συνέλιξης του φίλτρου p με τον όγκο εισόδου που αντιστοιχεί στο επίπεδο q. Αν υποθέσουμε ότι το βήμα του συνελικτικού επιπέδου q, ισούται με S_q , τότε ο πυρήνας p θα μετακινείται ούτως ώστε να υπολογιστούν οι αποκρίσεις για το σύνολο του όγκου της εισόδου στις οριζόντιες και κατακόρυφες θέσεις $1, S_q + 1, 2S_q + 1, \kappa o \kappa$. Αν θέσουμε τη τιμή της υπερ-παραμέτρου S_q ίση με 1 τότε το δεκτικό πεδίο του πυρήνα θα μετακινείται κατά 1 εικονοστοιχείο, καθιστώντας με αυτόν το τρόπο τη σάρωση

των δεδομένου του όγκου της εισόδου του συνελικτικού επιπέδου q, ιδιαίτερα πυκνή. Καθώς αυξάνουμε τη τιμή της υπερ-παραμέτρου S_q, αυξάνεται το δεκτικό πεδίο του πυρήνα. Ταυτόχρονα αραιώνει η σάρωση του όγκου της εισόδου του επιπέδου q, και συνεπώς μειώνονται οι χωρικές διαστάσεις (μήκος, πλάτος) του παραγόμενου όγκου εξόδου. Αν και με αυτό τον τρόπο δημιουργούνται πιο τραχιές αναπαραστάσεις για τον τρισδιάστατο όγκο εισόδου του επιπέδου q, η προσέγγιση αυτή ελαττώνει συνολικά την υπολογιστική και χρονική πολυπλοκότητα του μοντέλου.

Όταν ένας πυρήνας με δεκτικό πεδίο $F_q \times F_q$ και βήμα S_q, συνελίσσεται με μια αναπαράσταση διαστάσεων $L_q \times B_q$ τότε προκύπτει ένας χάρτης χαρακτηριστικών με διαστάσεις:

$$L_{q+1} \times B_{q+1} \tag{2.8}$$

όπου

$$L_{q+1} = \frac{L_q - F_q}{S_q} + 1 \tag{2.9}$$

και

$$B_{q+1} = \frac{B_q - F_q}{S_q} + 1 \tag{2.10}$$

Από τα παραπάνω γίνεται αντιληπτό ότι ένα συνελικτικό επίπεδο q, προκαλεί μείωση στις χωρικές διαστάσεις -μήκος και πλάτος- των δεδομένων της εισόδου του. Προκειμένου να αποφευχθεί αυτό το φαινόμενο χρησιμοποιείται η τεχνική του γεμίσματος με μηδενικά (zero padding). Η προσέγγιση αυτή προσθέτει εικονοστοιχεία στα άκρα της αναπαράστασης $L_q \times B_q$, τόσο κατά την κατεύθυνση του μήκους της όσο και κατά τη κατεύθυνση του πλάτους της. Η προσθήκη αυτή γίνεται έτσι ώστε ο χάρτης χαρακτηριστικών που θα παραχθεί να διατηρεί ίδιες διαστάσεις με την αναπαράσταση $L_q \times B_q$. Τα εικονοστοιχεία του προστίθενται έχουν είτε μηδενικές τιμές (zero padding) -στην περίπτωση αυτή δε θεωρείται ότι προστίθεται πληροφορία στο δισδιάστατο σήμα εισόδου του επιπέδου q, είτε σπανιότερα έχουν τιμές ίδιες με εκείνες των εικονοστοιχείων που βρίσκονται στα άκρα της αναπαράστασης $L_q \times B_q$ (reflection padding). Το πλήθος των εικονοστοιχείων που προστίθεται δίπλα σε κάθε εικονοστοιχείο το οποίο εντοπίζεται στα άκρα της αναπαράστασης $L_q \times B_q$, κατά τις κατευθύνσεις του μήκους και του πλάτους, δίνεται αντίστοιχα από τις σχέσεις (2.11) και (2.12):

$$P_q^{(l)} = \frac{F_q - S_q + (S_q - 1) \times L_q}{2} \quad \textbf{(2.11)}$$

$$P_q^{(b)} = \frac{F_q - S_q + (S_q - 1) \times B_q}{2} \quad \textbf{(2.12)}$$

Στο σημείο αυτό, αφού έχουν αποσαφηνιστεί και οι υπερ-παράμετροι του βήματος και του γεμίσματος είναι πλέον εφικτός ο καθορισμός των διαστάσεων του όγκου εξόδου ενός συνελικτικού επιπέδου q. Για όγκο εισόδου διαστάσεων $L_q \times B_q \times d_q$, υπερ-παραμέτρους S_q και $P_q^{(l)}, P_q^{(b)}$, και πλήθος πυρήνων p διαστάσεων $F_q \times F_q \times d_q$, έκαστος, το συνελικτικό επίπεδο q, παράγει όγκο εξόδου που είναι σύμφωνος με τις ακόλουθες σχέσεις:

$$L'_q \times B'_q \times d'_q \quad \textbf{(2.13)}$$

όπου,

$$L'_q = \frac{L_q - F_q + 2P_q^{(l)}}{S_q} + 1 \quad \textbf{(2.14)}$$

$$B'_q = \frac{B_q - F_q + 2P_q^{(d)}}{S_q} + 1 \quad \textbf{(2.15)}$$

$$d'_q = p \quad \textbf{(2.16)}$$

Συμπερασματικά, επεκτείνοντας τη σχέση (2.5), μπορούμε να ορίσουμε το είδος της επεξεργασίας που λαμβάνει χώρα σε ένα συνελικτικό επίπεδο του δικτύου ως εξής:

$$h_{ijp}^{(q+1)} = \left(\sum_{r=1}^{F_q}\sum_{s=1}^{F_q}\sum_{k=1}^{d_q} w_{rsk}^{(p,q)}\, h_{iS_q+r-1,\ jS_q+s-1,\ k}^{(q)}\right) + b^{(p,q)}, \quad \textbf{(2.17)}$$

$$\forall\, i \in \{1 \dots L_q - F_q + 1\},$$
$$\forall\, j \in \{1 \dots B_q - F_q + 1\},$$
$$\forall\, p \in \{1 \dots d_{q+1}\}$$

όπου $b^{(p,q)}$, η τιμή πόλωσης που σχετίζεται με τον συνελικτικό πυρήνα p του επιπέδου q και $h_{ijp}^{(q+1)}$ ο χάρτης χαρακτηριστικών που παράγεται από το συνελικτικό επίπεδο q.

2.2.2. Επίπεδα Απαλοιφής της Γραμμικότητας

Ένα επίπεδο απαλοιφής της γραμμικότητας q, συναντάται πάντα ακριβώς μετά από ένα συνελικτικό επίπεδο του δικτύου q. Ένα επίπεδο απαλοιφής της γραμμικότητας χρησιμοποιεί μια συνάρτηση ενεργοποίησης προκειμένου να μετασχηματίσει μη-γραμμικά την έξοδο του συνελικτικού επιπέδου που συνοδεύει. Πρόκειται για τις ίδιες μη-γραμμικές συναρτήσεις

ενεργοποίησης τις οποίες αξιοποιούν και τα τυπικά νευρωνικά δίκτυα. Οι πιο κοινές μη-γραμμικές συναρτήσεις ενεργοποίησης είναι: η σιγμοειδής συνάρτηση, η συνάρτηση υπερβολικής εφαπτομένης, η συνάρτηση ανορθωμένης γραμμικής μονάδας (ReLU), η συνάρτηση διαρρέουσας ανορθωμένης γραμμικής μονάδας (Leaky ReLU) και η συνάρτηση εκθετικής γραμμικής μονάδας (Exponential Linear Unit -ELU). Οι μαθηματικές εξισώσεις οι οποίες περιγράφουν τις παραπάνω εξισώσεις είναι αντίστοιχα οι (2.18)-(2.22). Οι γραφικές παραστάσεις των παραπάνω συναρτήσεων παρουσιάζονται στην εικόνα (2.2).

$$S(x) = \frac{e^x}{e^x + 1} \quad \textbf{(2.18)}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \textbf{(2.19)}$$

$$ReLU(x) = \max(0, x) \quad \textbf{(2.20)}$$

$$Leaky\_ReLU(x) = \max(0.1x, x) \quad \textbf{(2.21)}$$

$$ELU(x) = \begin{cases} x, & x \geq 0 \\ a(e^x + 1), & x < 0 \ με \ α \ παράμετρο \end{cases} \quad \textbf{(2.22)}$$

Αν συμβολίσουμε ως $a(\cdot)$, τη μη-γραμμική συνάρτηση ενεργοποίησης, τότε το είδος της επεξεργασίας που υλοποιείται από τη μη-γραμμική συνάρτηση σε ένα επίπεδο q, δίνεται από την ακόλουθη σχέση:

$$a_{ijk}{}^{(q)} = a(h_{i,j,k}{}^{(q)}) \quad \textbf{(2.23)}$$

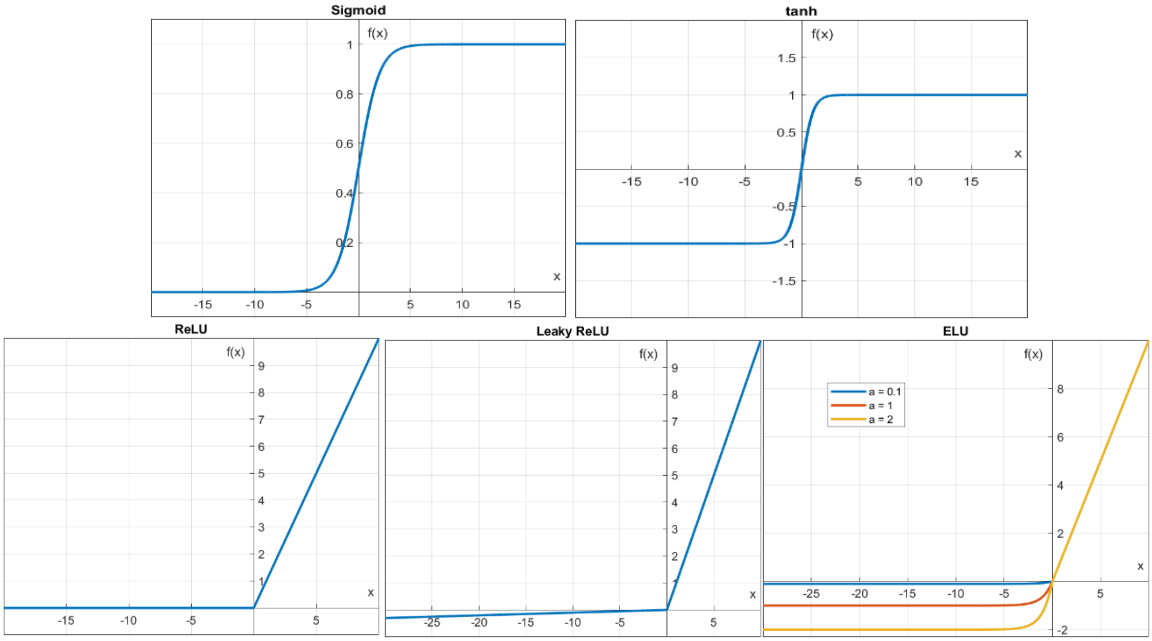
όπου $h_{i,j,k}{}^{(q)}$ ο χάρτης χαρακτηριστικών που παράχθηκε από το συνελικτικό επίπεδο q.

Τα επίπεδα απαλοιφής της γραμμικότητας δεν διαθέτουν παραμέτρους και συνεπώς δεν τίθεται θέμα εκπαίδευσης τους κατά τη διαδικασία της εκμάθησης των παραμέτρων του δικτύου, η οποία θα παρουσιαστεί αναλυτικά σε επόμενη υποενότητα. Επιπλέον, δεν επιδρούν στις διαστάσεις του όγκου της εισόδου που δέχονται. Με άλλα λόγια, οι διαστάσεις της εισόδου ενός επιπέδου απαλοιφής γραμμικότητας μένουν ανεπηρέαστες από την επεξεργασία που υφίστανται. Ως αποτέλεσμα, ο όγκος της εξόδου ενός επιπέδου απαλοιφής γραμμικότητας, φέρει ακριβώς τις ίδιες διαστάσεις με τον όγκο της εισόδου του.

Αν και όπως αναφέρθηκε παραπάνω, υπάρχουν διαφορετικές μη-γραμμικές συναρτήσεις οι οποίες μπορούν εν δυνάμει να χρησιμοποιηθούν από τα επίπεδα απαλοιφής της γραμμικότητας, στην πλειονότητα των περιπτώσεων τα συνελικτικά νευρωνικά δίκτυα χρησιμοποιούν συνάρτηση ανορθωμένης γραμμικής μονάδας ή αλλιώς ReLU. Η συνάρτηση ReLU σε ένα επίπεδο απαλοιφής της γραμμικότητας ορίζεται ως εξής:

$$h_{i,j,k}{}^{(q)\prime} = \max\{0, h_{i,j,k}{}^{(q)}\} \quad \textbf{(2.24)}$$

όπου $h_{i,j,k}{}^{(q)}$, ο χάρτης χαρακτηριστικών που παράχθηκε από το συνελικτικό επίπεδο q και ο οποίος αποτελεί την είσοδο του υπ' αριθμόν q επιπέδου απαλοιφής της γραμμικότητας.



Εικόνα 2.2 Γραφικές παραστάσεις συναρτήσεων ενεργοποίησης.

Ο λόγος που η συνάρτηση ReLU αποτελεί την κατεξοχήν αξιοποιήσιμη συνάρτηση στα επίπεδα απαλοιφής γραμμικότητας των συνελικτικών δικτύων είναι ότι έχει αποδειχθεί πειραματικά ότι η συγκεκριμένη συνάρτηση απλοποιεί και επιταχύνει τη διαδικασία εκπαίδευσης δια μέσου αλγορίθμου οπισθοδιάδοσης σφάλματος -ο οποίος θα παρουσιαστεί ακολούθως- συγκριτικά των υπολοίπων μη-γραμμικών συναρτήσεων ενεργοποίησης. Έτσι, η συνολική υπολογιστική πολυπλοκότητα του μοντέλου μειώνεται (Krizhevsky, Sutskever, & Hinton, 2012). Ωστόσο, οφείλουμε να αναφέρουμε ότι η συνάρτηση ReLU δύναται απενεργοποιήσει ορισμένους νευρώνες των επόμενων συνελικτικών επιπέδων, όταν ο ρυθμός με τον οποίο εκπαιδεύεται το μοντέλο έχει μεγάλη τιμή. Η εξαίρεση νευρώνων από τη διαδικασία της μάθησης, οδηγεί σε υποβάθμιση της απόδοσης του μοντέλου.

2.2.3. Επίπεδα Συγκέντρωσης

Από τη στιγμή που ένας χάρτης χαρακτηριστικών θα παραχθεί από το πρώτο συνελικτικό επίπεδο, θα διέλθει από ένα επίπεδο απαλοιφής της γραμμικότητας και, συνήθως, αμέσως μετά από ένα συγκεντρωτικό επίπεδο προτού προωθηθεί στο επόμενο συνελικτικό επίπεδο. Η επεξεργασία που πραγματοποιείται σε ένα συγκεντρωτικό επίπεδο q, αποσκοπεί στην προοδευτική μείωση των χωρικών διαστάσεων του χάρτη χαρακτηριστικών ${h_{i,j,k}}^{(q)}$, ο οποίος παράγεται από το συνελικτικό επίπεδο q.

Ένα συγκεντρωτικό επίπεδο προκαλεί ελάττωση στις διαστάσεις ενός χάρτη χαρακτηριστικών και άρα θα μπορούσε κανείς με έναν πρώτο συλλογισμό, να ισχυριστεί ότι αφαιρείται πληροφορία που έχει ανακτηθεί από προγενέστερα βήματα επεξεργασίας. Παρά το γεγονός αυτό, έχει αποδειχθεί ότι η επεξεργασία που υλοποιείται σε ένα συγκεντρωτικό επίπεδο κατορθώνει τελικά να διατηρεί τη σημαντική πληροφορία που φέρει ένας χάρτης χαρακτηριστικών.

Η λειτουργία των συγκεντρωτικών επιπέδων συμβάλλει καθοριστικά στον περιορισμό του πλήθους των παραμέτρων ενός συνελικτικού δικτύου. Η ύπαρξη των εν λόγω επιπέδων συνεισφέρει στον έλεγχο της υπερπροσαρμογής (overfitting) του μοντέλου. Με άλλα λόγια τα συγκεντρωντικά επίπεδα συνδράμουν στη δυνατότητα γενίκευσης του δικτύου. Δηλαδή στο να μπορεί να ανταποκριθεί εξίσου αποδοτικά όταν έρχεται αντιμέτωπο με άγνωστα σύνολα δεδομένων.

Ένα συγκεντρωτικό επίπεδο δέχεται στην είσοδο του ένα σύνολο από χάρτες χαρακτηριστικών και η επεξεργασία που πραγματοποιεί λαμβάνει χώρα μεμονωμένα σε καθέναν από αυτούς. Στο σημείο αυτό να διευκρινίσουμε ότι το βάθος του όγκου της εισόδου του συγκεντρωτικού επιπέδου διατηρείται ως έχει και μετά την επεξεργασία.

Τα επίπεδα αυτής της κατηγορίας, χρησιμοποιούν ένα δισδιάστατο πλέγμα $R_q \times R_q$, το οποίο μετακυλίει με βήμα S_q, κατά πλάτος και κατά μήκος του εκάστοτε χάρτη χαρακτηριστικών. Η επεξεργασία που υλοποιεί, συνίσταται στη διατήρηση μόνο μιας τιμής από την τοπική χωρική περιοχή $R_q \times R_q$ την οποία κάθε φορά υποδεικνύει το δισδιάστατο πλέγμα. όπως αντιλαμβάνεται κανείς πρόκειται για κάποιου είδους δειγματοληψία σε τοπικό επίπεδο. Η τιμή που εξάγεται από κάθε χωρική περιοχή του χάρτη χαρακτηριστικών, μπορεί να είναι είτε η μέγιστη, είτε η μέση. Όποτε κάνουμε αντίστοιχα λόγο για υποδειγματοληψία μέγιστης ή μέσης απόκρισης.

Αν συμβολίσουμε ως $pool(\cdot)$, τη συνάρτηση υποδειγματοληψίας, τότε το είδος της επεξεργασίας που υλοποιείται σε ένα συγκεντρωτικό επίπεδο q, δίνεται από την ακόλουθη σχέση:

$$a_{ijk}{}^{(q)'} = pool(\alpha_{i,j,k}{}^{(q)}) \quad \textbf{(2.25)}$$

όπου $\alpha_{i,j,k}{}^{(q)}$ ο χάρτης χαρακτηριστικών όπως αυτός διαμορφώθηκε από το επίπεδο απαλοιφής γραμμικότητας q.

Ειδικότερα, στην περίπτωση της υποδειγματοληψίας μέγιστης απόκρισης, οι τιμές του χάρτη χαρακτηριστικών $a_{ijk}{}^{(q)'}$ διαμορφώνονται σύμφωνα με τη σχέση (2.26)

$$a_{ijk}{}^{(q)'} = \max_{\substack{\forall\, r \in [1,R_q] \\ \forall\, s \in [1,R_q]}} (\alpha_{i,j,k}{}^{(q)}, \quad \alpha_{i\cdot S_q + r - 1, j\cdot S_q + s - 1, k}{}^{(q-1)}) \quad \textbf{(2.26)}$$

Στην περίπτωση της υποδειγματοληψίας μέσης απόκρισης, ο χάρτης χαρακτηριστικών $a_{ijk}{}^{(q)'}$ αποκτά τιμές σύμφωνα με την εξής σχέση:

$$a_{ijk}{}^{(q)'} = \frac{\sum_{r=1}^{R_q} \sum_{s=1}^{R_q} \alpha_{i\cdot S_q + r - 1, j\cdot S_q + s - 1, k}{}^{(q-1)}}{R_q \cdot R_q} \quad \textbf{(2.27)}$$

Ένα συγκεντρωτικό επίπεδο, δεν περιλαμβάνει εκπαιδεύσιμες παραμέτρους και συνεπώς δεν προσθέτει επιπλέον παραμέτρους στο μοντέλο. Ωστόσο, η διάσταση R_q, του τετραγωνικού πλέγματος, καθώς και το βήμα S_q, με το οποίο αυτό σαρώνει τον χάρτη χαρακτηριστικών αποτελούν υπερ-παραμέτρους για ένα συγκεντρωτικό επίπεδο q.

Ένα συγκεντρωτικό επίπεδο, δεν περιλαμβάνει εκπαιδεύσιμες παραμέτρους και συνεπώς δεν προσθέτει επιπλέον παραμέτρους στο μοντέλο. Ωστόσο, η διάσταση R_q, του τετραγωνικού πλέγματος, καθώς και το βήμα S_q, με το οποίο αυτό σαρώνει τον χάρτη χαρακτηριστικών αποτελούν υπερ-παραμέτρους για ένα συγκεντρωτικό επίπεδο q.

Δεδομένου ότι το επίπεδο μη γραμμικότητας, που προηγείται πάντα ενός συγκεντρωτικού επιπέδου δε προκαλεί μεταβολή των διαστάσεων του όγκου εισόδου που επεξεργάστηκε, οι διαστάσεις του όγκου εισόδου που επεξεργάζεται ένα συγκεντρωτικό επίπεδο q είναι σύμφωνες των διαστάσεων του όγκου εξόδου του συνελικτικού επιπέδου q. Οι σχέσεις (2.13)-(2.16) που ορίστηκαν νωρίτερα καθορίζουν τις διαστάσεις του όγκου εξόδου που αντιστοιχεί στο συνελικτικό επίπεδο q. Τώρα, αν σε συμφωνία με την (2.13), συμβολίσουμε τις διαστάσεις

του όγκου εισόδου στο συγκεντρωτικό επίπεδο q ως $L'_q \times B'_q \times d'_q$, τότε οι διαστάσεις του όγκου εξόδου αυτού του επιπέδου δίνονται από τις σχέσεις που έπονται:

$$L''_q \times B''_q \times d''_q \quad \textbf{(2.28)}$$

όπου,

$$L_q'' = \frac{L'_q - R_q}{S_q} + 1 \quad \textbf{(2.29)}$$

$$B_q'' = \frac{B'_q - R_q}{S_q} + 1 \quad \textbf{(2.30)}$$

$$d''_q = d'_q \quad \textbf{(2.31)}$$

και R_q , S_q , οι υπερ-παράμετροι του συγκεντρωτικού επιπέδου.

2.2.4. Πλήρως Συνδεδεμένα Επίπεδα

Όπως έχουμε επισημάνει νωρίτερα, η αρχιτεκτονική ενός συνελικτικού δικτύου περιλαμβάνει μια σειρά από διαδοχικά τοποθετημένα συνελικτικά επίπεδα ακολουθούμενα πάντα από επίπεδα απαλοιφής γραμμικότητας και συγκεντρωτικά επίπεδα. Το καταληκτικό στάδιο επεξεργασίας σε ένα συνελικτικό δίκτυο πραγματοποιείται από τα πλήρως συνδεδεμένα επίπεδα, τα οποία συναντώνται μετά το πέρας των παραπάνω επιπέδων (συνελικτικά, απαλοιφής γραμμικότητας, συγκεντρωτικά). Η ύπαρξη των πλήρως συνδεδεμένων επιπέδων σε ένα συνελικτικό δίκτυο, αποσκοπεί στην αξιοποίηση των χαρακτηριστικών που εξήχθησαν από τα συνελικτικά επίπεδα του δικτύου, με απώτερο σκοπό την πρόσδοση ενός είδους υψηλού επιπέδου συλλογιστικής στο συνολικό μοντέλο (Gu et al., 2018). Ο τρόπος λειτουργίας των πλήρως συνδεδεμένων επιπέδων ενός συνελικτικού δικτύου, είναι ο ίδιος ακριβώς με εκείνον που επιτελούν τα πλήρως συνδεδεμένα επίπεδα ενός κλασσικού νευρωνικού δικτύου εμπρόσθιας τροφοδότησης.

Ένα πλήρως συνδεδεμένο επίπεδο συντίθεται από ένα σύνολο μονάδων επεξεργασίας, δηλαδή νευρώνες, όπου -όπως υποδηλώνει και το όνομα των εν λόγω επιπέδων- καθένας από αυτούς διατηρεί συνδέσεις με κάθε νευρώνα που ανήκει στο αμέσως προηγούμενο επίπεδο του δικτύου. Η ισχύς της κάθε σύνδεσης που απαντάται σε ένα νευρώνα αποτιμάται από μια αριθμητική τιμή η οποία ονομάζεται βάρος του νευρώνα. Με τη βοήθεια των τιμών των βαρών, ένας νευρώνας υπολογίζει ένα σταθμισμένο άθροισμα για το σύνολο των εισόδων του - συνδέσεων με τους νευρώνες του προηγούμενου επιπέδου. Στο άθροισμα αυτό προστίθεται μια τιμή που διαθέτει ο κάθε νευρώνας και ονομάζεται βάρος πόλωσης. Τόσο οι τιμές των βαρών όσο και το βάρος της πόλωσης αποτελούν εκπαιδεύσιμες παραμέτρους για έναν

νευρώνα. Μια συνάρτηση ενεργοποίησης εφαρμόζεται στο αποτέλεσμα της παραπάνω επεξεργασίας και με τον τρόπο αυτό διαμορφώνεται η τελική τιμή της εξόδου ενός νευρώνα.

Αν θεωρήσουμε ως N το πλήθος των νευρώνων που διαθέτει ένα πλήρως συνδεδεμένο επίπεδο q, ως $L_q \times B_q \times d_q$, το μέγεθος της εισόδου (πλήθος συνδέσεων) κάθε νευρώνα στο επίπεδο q, και ως $a(\cdot)$, τη συνάρτηση ενεργοποίησης που χρησιμοποιούν οι νευρώνες του επίπεδου αυτού, τότε το είδος της επεξεργασίας που υλοποιείται στο πλήρως συνδεδεμένο επίπεδο q, υπαγορεύεται από την ακόλουθη σχέση:

$$y_n{}^{(q)} = a\left(\sum_{i=1}^{L_q}\sum_{j=1}^{B_q}\sum_{k=1}^{d_q}\left(w_{i,j,k,n}{}^{(q)} y_{i,j,k,n}{}^{(q-1)}\right) + b_n{}^{(q)}\right), \quad \forall\, n \in \{1 \ldots N\}, \tag{2.32}$$

όπου $y_{i,j,k}{}^{(q-1)}$, η έξοδος του αμέσως προηγούμενου επιπέδου.

Σε ένα πλήρως συνδεδεμένο επίπεδο q, κάθε νευρώνας έχει μέγεθος εισόδου $L_q \times B_q \times d_q$. Δηλαδή, διατηρεί $L_q \times B_q \times d_q$, συνδέσεις με τους νευρώνες του αμέσως προηγούμενου επιπέδου. Με άλλα λόγια διαθέτει $L_q \times B_q \times d_q$ βάρη ή αλλιώς παραμέτρους. Πέρα από τα βάρη, σε κάθε νευρώνα αποδίδεται και μια πόλωση. Συνεπώς, ο συνολικός αριθμός των παραμέτρων που διαθέτει κάθε νευρώνας στο επίπεδο q, ισούται με:

$$\left(L_q \times B_q \times d_q\right) + 1 \tag{2.33}$$

Αν υποθέσουμε ότι το συνολικό πλήθος νευρώνων ενός πλήρως συνδεδεμένου επιπέδου ισούται με n, τότε σύμφωνα με τη σχέση (2.33), ο συνολικός αριθμός των παραμέτρων που χρησιμοποιούνται στο επίπεδο εκφράζεται από τη σχέση που έπεται:

$$n \times [\left(L_q \times B_q \times d_q\right) + 1] \tag{2.34}$$

Είναι αντιληπτό το γεγονός ότι, ο υψηλός βαθμός συνδεσιμότητας που χαρακτηρίζει τους νευρώνες ενός πλήρως συνδεδεμένου επιπέδου, προσθέτει ένα διόλου ευκαταφρόνητο πλήθος παραμέτρων στο συνολικό μοντέλο. Εξαιτίας του παραπάνω γεγονότος, σε αρκετές αρχιτεκτονικές τα πλήρως συνδεδεμένα επίπεδα αντικαθίστανται από συνελικτικά επίπεδα με μέγεθος πυρήνα 1×1 (Gu et al., 2018). Με αυτόν τρόπο, τα πλήρως συνδεδεμένα επίπεδα εξαιρούνται εντελώς από την αρχιτεκτονική των συνελικτικών δικτύων.

Το τελευταίο πλήρως συνδεδεμένο επίπεδο ενός δικτύου, το οποίο καλείται και επίπεδο εξόδου, σχεδιάζεται ανάλογα με την εκάστοτε εφαρμογή στην οποία αξιοποιείται το δίκτυο. Το πλήθος των νευρώνων από τους οποίους θα συντίθεται καθώς και το είδος των

συναρτήσεων ενεργοποίησης που χρησιμοποιούν καθορίζουν μονοσήμαντα τη λειτουργικότητα του επιπέδου εξόδου και κατ' επέκταση του συνελικτικού δικτύου.

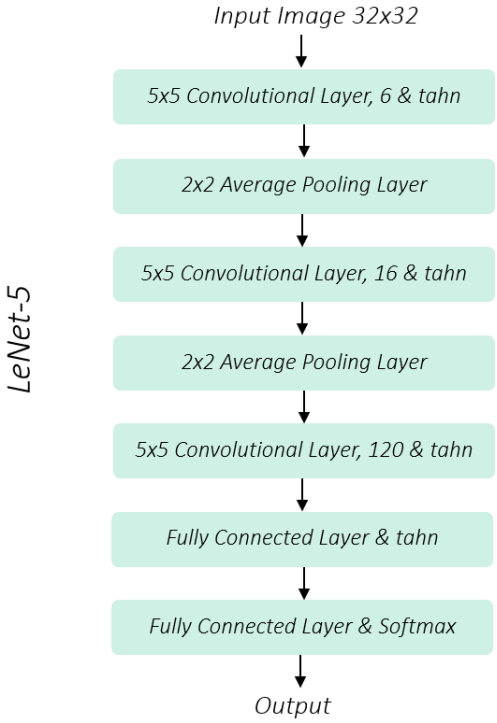
2.3. Γνωστές Αρχιτεκτονικές Συνελικτικών Νευρωνικών Δικτύων

Αν και ο πρόδρομος των συνελικτικών νευρωνικών δικτύων - το Neocognitron - παρουσιάστηκε για πρώτη φορά το 1980 από τους Fukushima & Miyake (Fukushima & Miyake, 1982), χρειάστηκε να περάσουν δέκα ολόκληρα χρόνια μέχρι ο LeCun και οι συνεργάτες του να θεμελιώσουν το πλαίσιο λειτουργίας των σύγχρονων συνελικτικών νευρωνικών δικτύων (LeCun et al., 1990) και συνολικά δεκαοχτώ χρόνια μέχρι να αναπτυχθεί από τον ίδιο και τους συνεργάτες του το πρώτο συνελικτικό νευρωνικό δίκτυο, το LeNet-5, το οποίο πραγματοποιούσε ταξινόμηση χειρόγραφων ψηφίων (LeCun, Bottou, Bengio, & Haffner, 1998).

Το δίκτυο LeNet-5 ακολουθούσε μια απλή αρχιτεκτονική συντιθέμενη από δυο συνελικτικά επίπεδα ακολουθούμενο έκαστο από ένα συγκεντρωτικό επίπεδο υποδειγματοληψίας μέσης απόκρισης καθώς και ένα συνελικτικό επίπεδο ακολουθούμενο από δυο πλήρως συνδεδεμένα επίπεδα. Κάθε συνελικτικό επίπεδο του δικτύου ακολουθείται από ένα επίπεδο απαλοιφής γραμμικότητας. Η αρχιτεκτονική του δικτύου συνοψίζεται στην εικόνα (2.3)

Οι προσπάθειες που έλαβαν χώρα τα επόμενα χρόνια αναφορικά με την χρήση των συνελικτικών νευρωνικών δικτύων σε περίπλοκες εργασίες όπως λχ η ταξινόμηση μεγάλης κλίμακας εικόνων, πέραν του γεγονότος ότι είναι περιορισμένες δεν σημείωσαν αξιοσημείωτη επίδοση κι ούτε έτυχαν ευρείας αναγνώρισης κυρίως λόγω της έλλειψης επαρκούς ποσότητας δεδομένων για την εκπαίδευση των δικτύων καθώς και των περιορισμένων υπολογιστικών δυνατοτήτων που προσέφεραν οι μονάδες επεξεργασίας δεδομένων της εποχής (Gu et al., 2018).

Από τα μέσα της δεκαετίας του 2000 καταβλήθηκαν αρκετές προσπάθειες που αποπειράθηκαν να αντιμετωπίσουν τις παραπάνω προκλήσεις που εγείρει η χρησιμοποίηση των συνελικτικών νευρωνικών δικτύων. Ωστόσο, η μεγάλη αλλαγή που συντέλεσε στην καθιέρωση των τελευταίων ως κατεξοχήν εργαλεία για την προσέγγιση και επίλυση προβλημάτων υπολογιστικής όρασης πραγματοποιήθηκε το 2012 με την παρουσίαση του AlexNet (Gu et al., 2018).



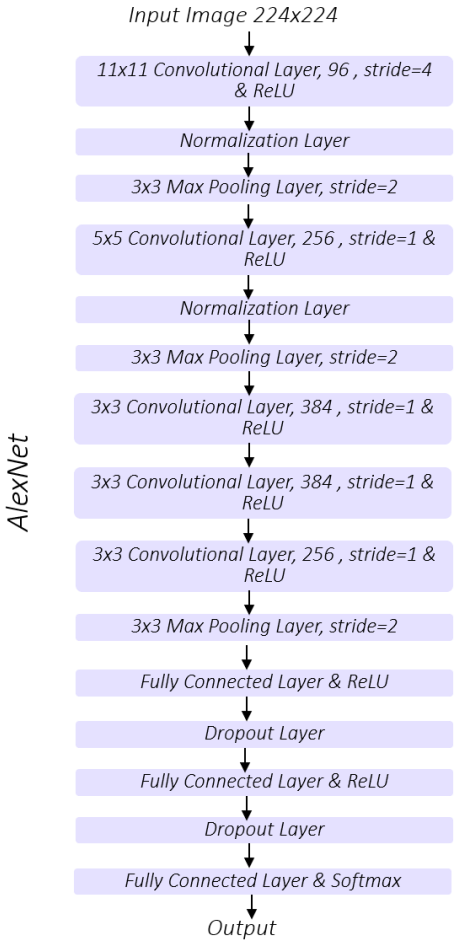
Εικόνα 2.3 Αρχιτεκτονική του δικτύου LeNet-5

Το εν λόγω δίκτυο που προτάθηκε από τον Krizhevsky και τους συνεργάτες του (Krizhevsky et al., 2012), σχεδιάστηκε με στόχο την ταξινόμηση ενός μεγάλου συνόλου εικόνων - του ImageNet. Στην πραγματικότητα το ImageNet αποτελεί μια βάση δεδομένων με περισσότερα από δεκατέσσερα εκατομμύρια εικόνων οργανωμένα σε χίλιες διαφορετικές κατηγορίες. Ο πρωταρχικός στόχος της σχεδίασης αυτής της βάσης αποτελούσε η αναγνώριση αντικειμένων και η ταξινόμηση των εικόνων. Για το σκοπό αυτό διεξάγεται ο ετήσιος διαγωνισμός ILSVRC (ImageNet Large Scale Visual Recognition Challenge) - ο οποίος συνιστά σημείο αναφοράς για την επιστημονική κοινότητα της Υπολογιστικής Όρασης. Στα πλαίσια του διαγωνισμού ILSVRC, έχουν παρουσιαστεί ορισμένες από τις πιο ρηξικέλευθες εξελίξεις όσον άφορα τα συνελικτικά νευρωνικά δίκτυα (Aggarwal, 2018).

Το δίκτυο AlexNet έφερε αρχιτεκτονική παρόμοια με αυτή του LeNet-5 αλλά βαθύτερη. Το δίκτυο των Krizhevsky et al. αποτελούταν από πέντε συνολικά συνελικτικά επίπεδα και τρία πλήρως συνδεδεμένα επίπεδα. Η ειδοποιός διαφορά του με το δίκτυο των LeCun et al ήταν το γεγονός ότι χρησιμοποιούσε τη συνάρτηση της ανορθωμένης γραμμικής μονάδας (ReLU) στα επίπεδα απαλοιφής γραμμικότητας. Επιπλέον, εισήγαγε τεχνικές οι οποίες χρησιμοποιούνται μέχρι σήμερα επικουρικά και συμπληρωματικά στα συνελικτικά δίκτυα όπως η επαύξηση δεδομένων (data augmentation), τα επίπεδα απόσυρσης νευρώνων (dropout

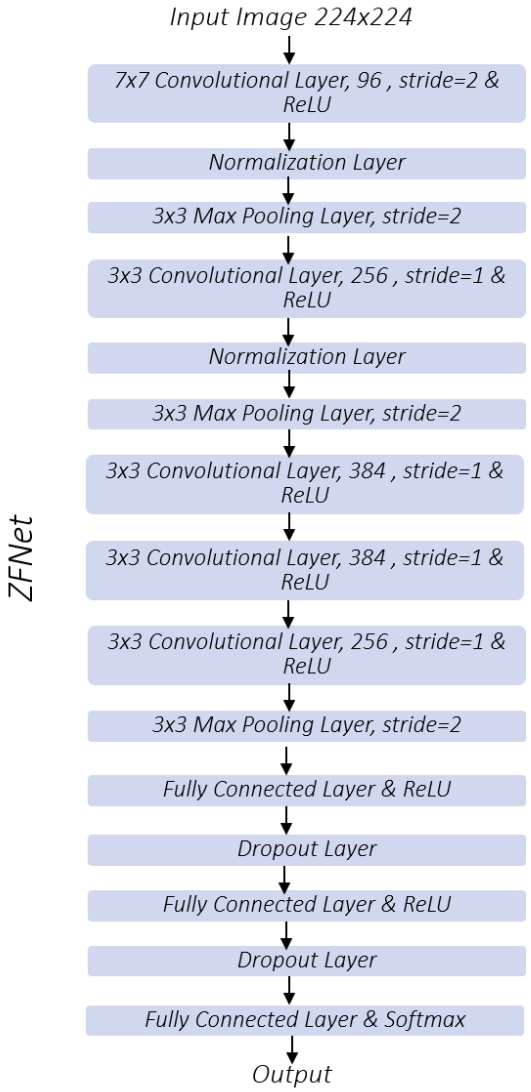
layers) και τα επίπεδα κανονικοποίησης τοπικής απόκρισης (local response normalization layers). Οι παραπάνω τεχνικές δύνανται να συμβάλλουν στην αποφυγή της υπερπροσαρμογής ενός δικτύου. Με άλλα λογία μειώνουν την ενδεχόμενη αδυναμία του τελικού μοντέλου να γενικεύσει τα αποτελέσματα του, δηλαδή να μπορεί να εφαρμοστεί σε διαφορετικά σύνολα δεδομένων. Η συνολική αρχιτεκτονική του δικτύου AlexNet μπορεί να παρατηρηθεί στην εικόνα (2.4).

Ακολούθως, της επιτυχίας - ορόσημο του δικτύου AlexNet η προσοχή της επιστημονικής κοινότητας στράφηκε για ακόμη μια φορά στη μελέτη των συνελικτικών νευρωνικών δικτύων. Έκτοτε, με έναυσμα τη περαιτέρω βελτίωση των αποτελεσμάτων της ταξινόμησης εικόνων που παρέχονται από την αρχιτεκτονική που υποδεικνύει το AlexNet, έχει παρατηρηθεί μια απότομη αύξηση των εργασιών που αξιοποιούν τα συνελικτικά δίκτυα. Ανάμεσα τους ξεχωρίζουν τα: ZFNet, VGGNet, GoogleNet, ResNet, DenseNet. Τα παραπάνω μοντέλα αναπτύχθηκαν και παρουσιάστηκαν στα πλαίσια του διαγωνισμού ILSVRC. (Gu et al., 2018).



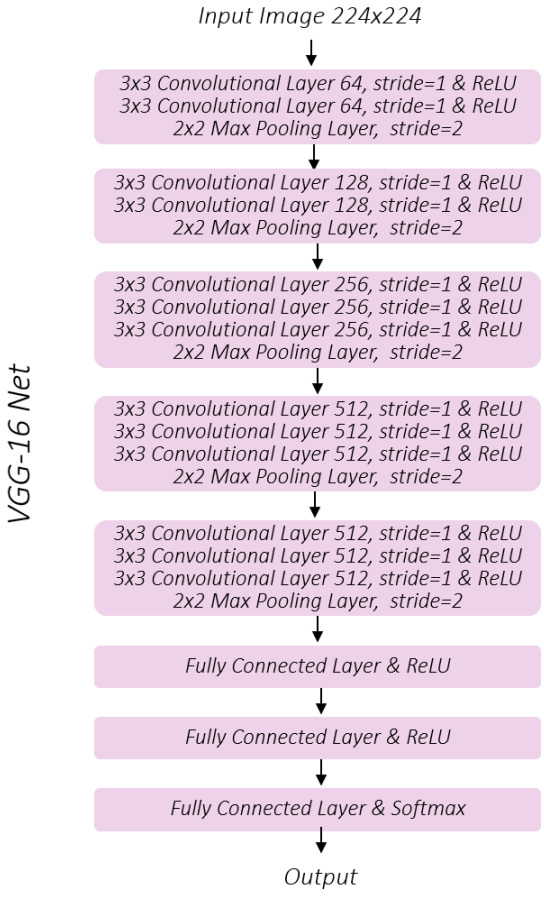
Εικόνα 2.4 Αρχιτεκτονική του δικτύου AlexNet

Το ZFNet προτάθηκε από τους Zeiler & Fergus (Zeiler & Fergus, 2014) προκειμένου να εξελίξει την αρχιτεκτονική του AlexNet. Για το λόγο αυτό η δική του αρχιτεκτονική επιδεικνύει μεγάλη ομοιότητα με εκείνη του AlexNet αλλά εισάγει βασικές αλλαγές με στόχο την βελτιστοποίηση των υπερ-παραμέτρων του τελευταίου. Μεταξύ αυτών συγκαταλέγεται η χρήση πυρήνων μικρότερου δεκτικού πεδίου συγκριτικά με το δίκτυο AlexNet ο αριθμός των οποίων αυξάνει στα βαθύτερα επίπεδα του δικτύου. Ωστόσο, διατηρεί όπως το AlexNet τη συνάρτηση της ανορθωμένης γραμμικής μονάδας (ReLU) στα επίπεδα απαλοιφής γραμμικότητας. Η συνολική αρχιτεκτονική του δικτύου παρουσιάζεται στην εικόνα (2.5). Επιπρόσθετα, στην εργασία αυτή γίνεται η πρώτη απόπειρα κατανόησης του τρόπου λειτουργίας των συνελικτικών νευρωνικών δικτύων με την οπτικοποίηση των αποτελεσμάτων που αντιστοιχούν στα ενδιάμεσα επίπεδα του δικτύου.

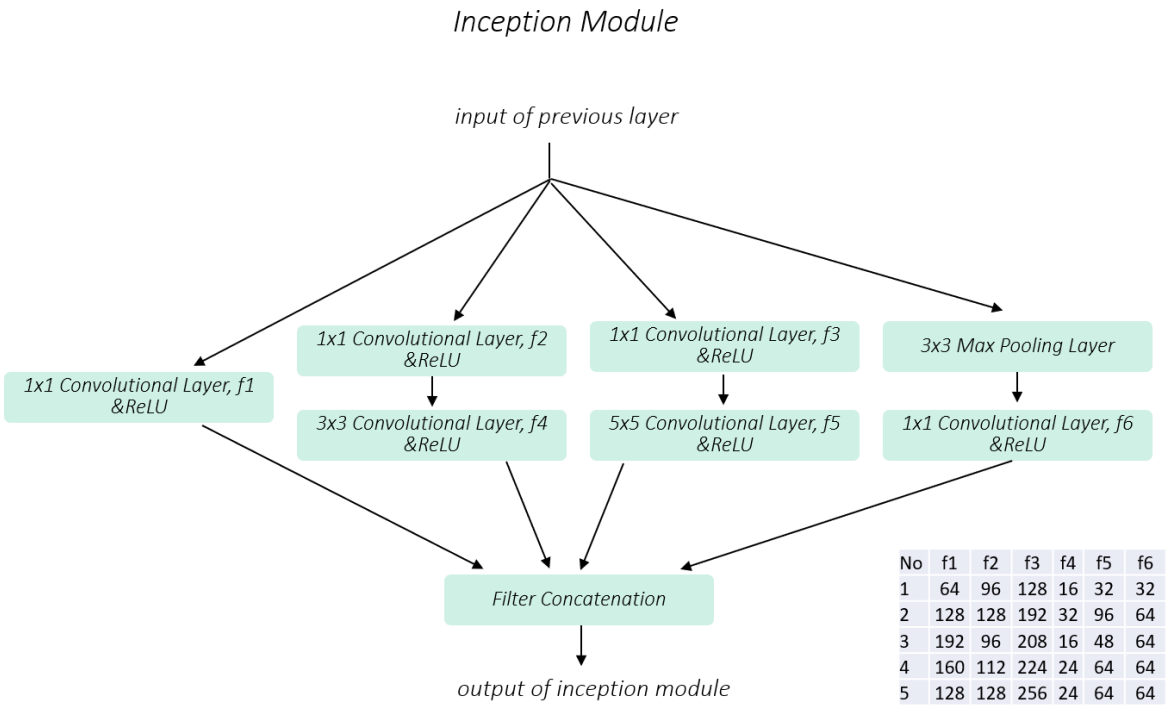


Εικόνα 2.5 Αρχιτεκτονική του δικτύου ZFNet

Οι Simonyan και Zisserman μας σύστησαν το VGGNet (Simonyan & Zisserman, 2014). Το VGGNet συνέδραμε με τρόπο ουσιαστικό αφενός στην ανάδειξη της άποψης ότι το βάθος ενός συνελικτικού δικτύου συνιστά κρίσιμο παράγοντα αναφορικά με την απόδοση του και αφετέρου στην αποκρυστάλλωση της αντίληψης ότι τα βαθιά μοντέλα έχουν την ικανότητα να γενικεύουν τα αποτελέσματα τους σε διάφορα σύνολα δεδομένων. Αν και το παραπάνω δίκτυο χρησιμοποιεί περισσότερες παραμέτρους από κάθε άλλη γνωστή αρχιτεκτονική, την παρούσα στιγμή φαίνεται ότι βιβλιογραφικά αποτελεί την συνηθέστερη επιλογή όταν είναι επιθυμητή η εξαγωγή χαρακτηριστικών από εικόνες. Το VGGNet χρησιμοποιεί αποκλειστικά πυρήνες δεκτικού μεγέθους 3×3. Οι συγγραφείς της εργασίας διαπιστώνουν ότι χρησιμοποιώντας δυο συνεχόμενα συνελικτικά επίπεδα πυρήνων μεγέθους 3×3 ισοδυναμεί με ένα συνελικτικό επίπεδο πυρήνων δεκτικού πεδίου 5×5. Ομοίως, τα τρία συνεχόμενα συνελικτικά επίπεδα πυρήνων μεγέθους 3×3 ισοδυναμούν με ένα συνελικτικό επίπεδο πυρήνων δεκτικού πεδίου 7×7. Με τον τρόπο αυτό οι οι Simonyan και Zisserman σχεδίασαν μια βαθιά αρχιτεκτονική δεκαέξι επιπέδων με λιγότερες παραμέτρους απ΄ ότι θα έφερε μια αντίστοιχου βάθους αρχιτεκτονική με πυρήνες μεγαλύτερου δεκτικού πεδίου. Η αρχιτεκτονική του δικτύου VGGNet συνοψίζεται στην εικόνα (2.6).



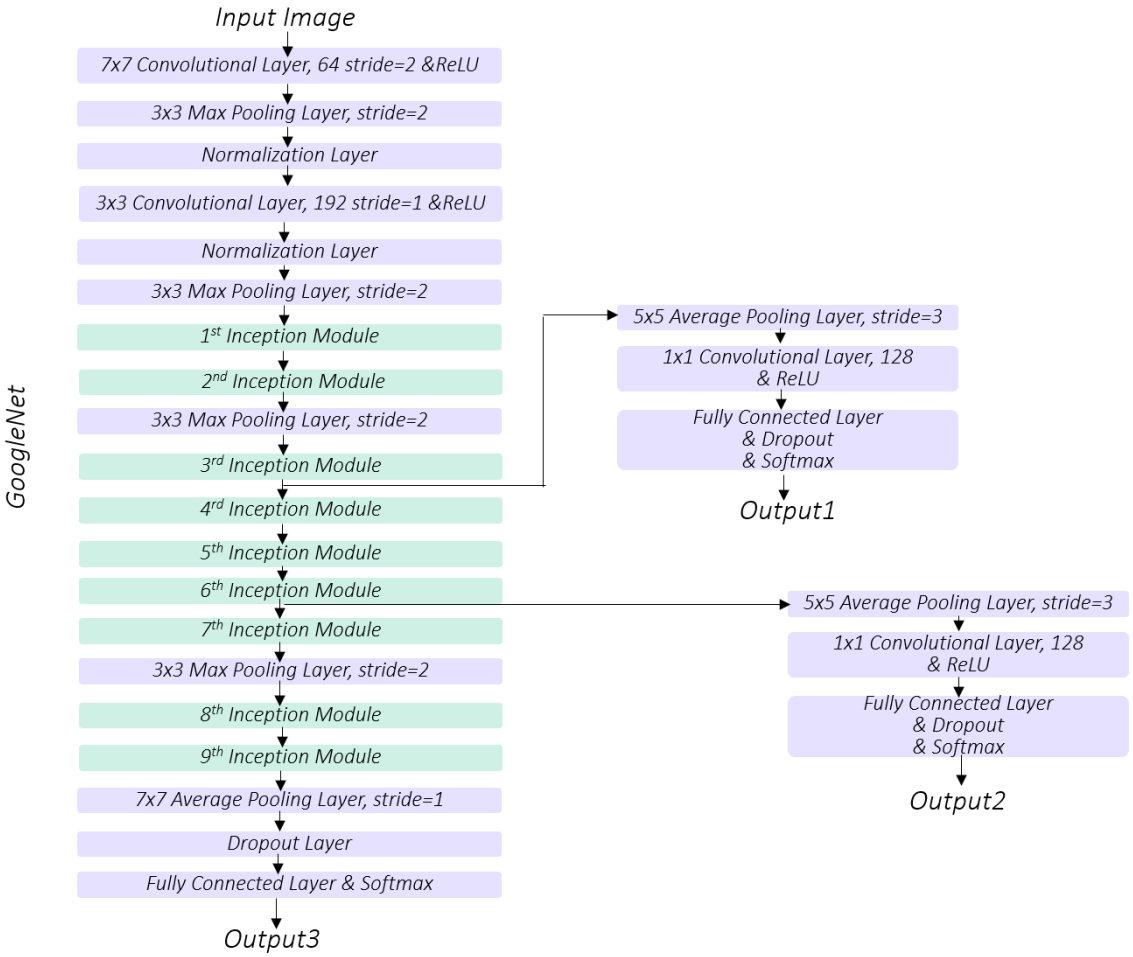
Εικόνα 2.6 Η αρχιτεκτονική του δικτύου VGGNet 16 επιπέδων



No	f1	f2	f3	f4	f5	f6
1	64	96	128	16	32	32
2	128	128	192	32	96	64
3	192	96	208	16	48	64
4	160	112	224	24	64	64
5	128	128	256	24	64	64
6	112	144	288	32	64	64
7	256	160	320	32	128	128
8	256	160	320	32	128	128
9	384	192	384	48	128	128

Εικόνα 2.7 Αρχιτεκτονική των δομικών μονάδων του GoogleNet

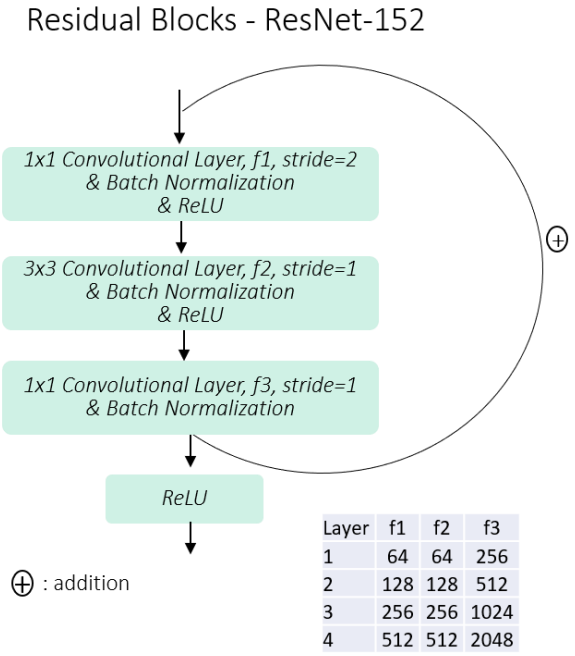
Την ίδια περίοδο, ο Szegedy και οι συνεργάτες του παρουσίασαν το GoogLeNet (Szegedy et al., 2015). Καινοτομία του GoogLeNet αποτελεί το γεγονός ότι η αρχιτεκτονική του δομείται πάνω σε μονάδες που ονομάζονται inception modules και οι οποίες συμβάλλουν με τρόπο δραστικό στην μείωση του συνολικού αριθμού παραμέτρων του δικτύου. Πράγματι, το προτεινόμενο δίκτυο φέρει πολύ λιγότερες παραμέτρους συγκριτικά με το AlexNet και το VGGNet. Ο αριθμός των εκπαιδεύσιμων παραμέτρων του GoogLeNet είναι περίπου 4 εκατομμύρια, ενώ τα AlexNet και το VGGNet χρησιμοποιούν αντίστοιχα περίπου 60 και 138 εκατομμύρια παραμέτρους. Η αρχιτεκτονική που υποδεικνύει το GoogLeNet αποτέλεσε αφετηρία για τον σχεδιασμό καινοτόμων δικτύων τα χρόνια που ακολούθησαν. Το εν λόγω δίκτυο συντίθεται από εννιά δομικές μονάδες inception. Η αρχιτεκτονική μιας δομικής μονάδας μπορεί να παρατηρηθεί στην εικόνα (2.7). Επιπλέον, σε αντίθεση με τις προαναφερόμενες αρχιτεκτονικές πέραν του ταξινομητή που συναντάται στο καταληκτικό επίπεδο του δικτύου, προτείνεται και η χρήση δυο επιπρόσθετων ταξινομητών στα ενδιάμεσα επίπεδα του δικτύου. Η τελική απόφαση για την ταξινόμηση λαμβάνει υπόψη όλους τους ταξινομητές. Η συνολική αρχιτεκτονική του δικτύου GoogLeNet παρουσιάζεται στην εικόνα (2.8).



Εικόνα 2.8 Αρχιτεκτονική του δικτύου GoogleNet

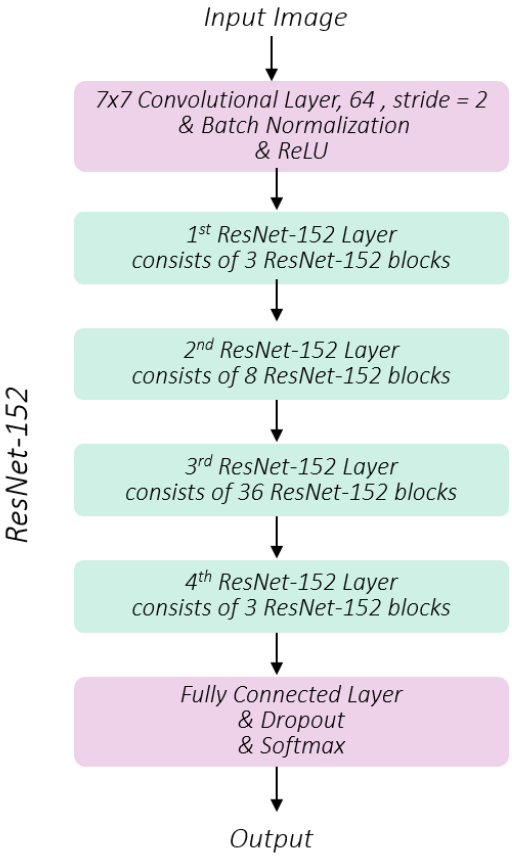
Το δίκτυο ResNet προτάθηκε από τον He και τους συνεργάτες του (K. He, Zhang, Ren, & Sun, 2016). Η πρωτοτυπία της εν λόγω μεθοδολογίας έγκειται στη διαπίστωση που πραγματοποιείται από τους συγγραφείς ότι η αύξηση του βάθους ενός συνελικτικού δικτύου δεν συνεπάγεται τη γραμμική βελτίωση της ακρίβειάς του. Με αφόρμηση την παραπάνω παρατήρηση, η βαθιά αρχιτεκτονική εκατό πενήντα δύο επιπέδων του ResNet μας εισάγει στη λεγόμενη «υπολειμματική» μάθηση (Resdual Learning). Η κεντρική ιδέα της υπολειμματικής μάθησης έγκειται στην αξιοποίηση της γνώσης που έχει αποκτηθεί από τα κατώτερα επίπεδα του δικτύου στα βαθύτερα μέσω παραλειπόμενων συνδέσεων (skip connections) μεταξύ των επιπέδων του. Με αυτόν τον τρόπο, παρέχεται η δυνατότητα δημιουργίας εξαιρετικά βαθέων συνελικτικών νευρωνικών δικτύων των οποίων η ακρίβεια αυξάνεται ανάλογα με το βάθος τους και ταυτόχρονα μειώνεται σημαντικά ο συνολικός αριθμός των απαιτούμενων παραμέτρων. Βασική δομική μονάδα του δικτύου ResNet-152 αποτελούν τα μπλοκ υπολειμματικής μάθησης (Residual Blocks), η αρχιτεκτονική των οποίων μπορεί να παρατηρηθεί στην εικόνα (2.9). Η συνολική αρχιτεκτονική του δικτύου συνοψίζεται στην

εικόνα (2.10). Τα υψηλά ποσοστά ακρίβειας του ResNet στην ταξινόμηση εικόνων, σε συνδυασμό με το γεγονός ότι πραγματοποιεί την ταξινόμηση με μικρότερο ποσοστό σφάλματος ακόμη και από τον άνθρωπο, συντέλεσε στην ευρεία αποδοχή αυτού του μοντέλου και στον χαρακτηρισμό της αρχιτεκτονικής του ως μία από τις καλύτερες που έχουν προταθεί μέχρι στιγμής. Για το λόγο αυτό στη βιβλιογραφία τα τελευταία χρόνια όχι μόνο χρησιμοποιείται ευρέως στις εργασίες που αφορούν συνελικτικά δίκτυα, άλλα και συναντώνται πολυάριθμες παραλλαγές του αναφορικά με το συνολικό πλήθος των επιπέδων από τα οποία απαρτίζεται.



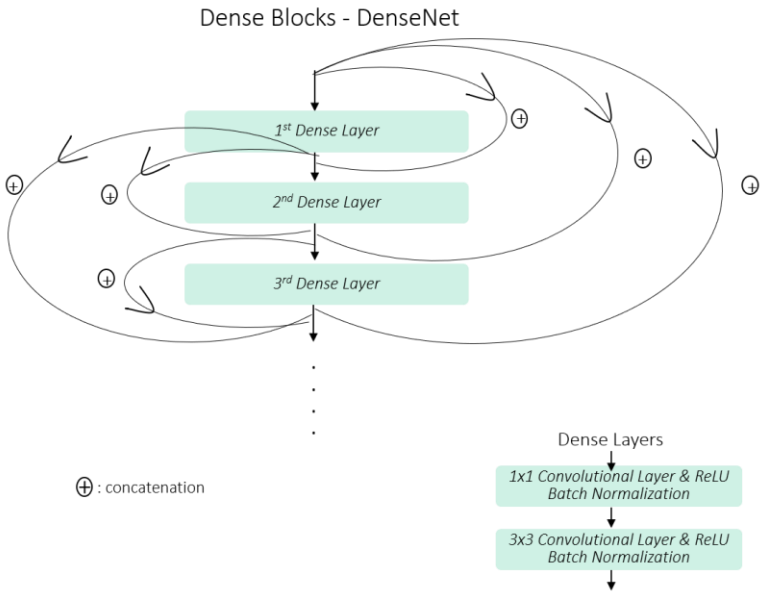
Εικόνα 2.9 Αρχιτεκτονική των μπλοκ υπολειμματικής μάθησης του δικτύου ResNet-152

Η αρχιτεκτονική των δικτύων DenseNet προτάθηκε από τους (Huang, Liu, Van Der Maaten, & Weinberger, 2017) προκειμένου να εξελίξει τα δίκτυα ResNet. Θεμελιώδεις δομικές μονάδες ενός δικτύου DenseNet είναι τα πυκνά μπλοκ (Dense Blocks). Η δομή ενός πυκνού μπλοκ μπορεί να παρατηρηθεί στην εικόνα (2.11). Ένα πυκνό μπλοκ συντίθενται από συνελικτικά επίπεδα τα οποία συνδέονται απευθείας μεταξύ τους. Με τον τρόπο αυτό κάθε επίπεδο δέχεται την είσοδο του την έξοδο όλων των προηγουμένων και καθιστά εφικτή την παραγωγή επιπλέον χαρακτηριστικών.

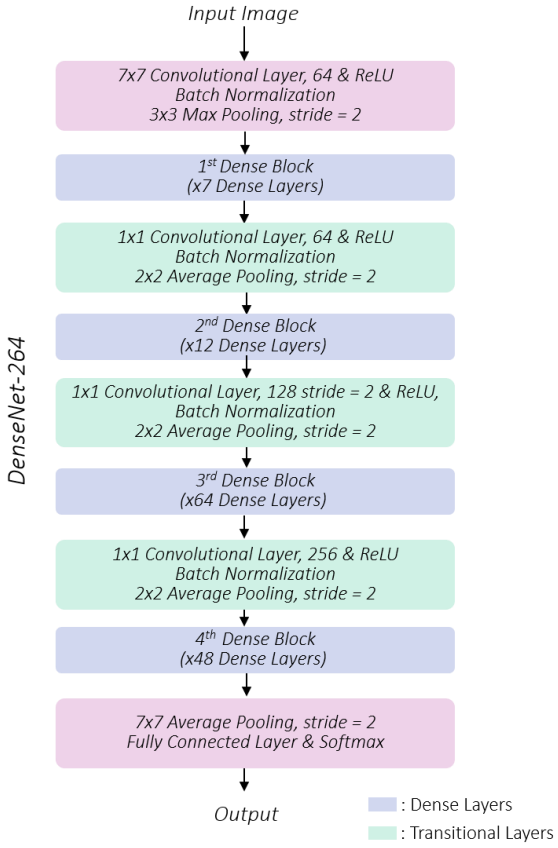


Εικόνα 2.10 Αρχιτεκτονική του δικτύου ResNet 152 επιπέδων

Σε αντίθεση με τα δίκτυα ResNet όπου πραγματοποιείται πρόσθεση στοιχείο προς στοιχείο της εισόδου με τις προηγούμενες εξόδους, σε ένα δίκτυο DenseNet πραγματοποιείται συνένωση (concatenation) των συνδεόμενων επιπέδων ως προς τη διάσταση του βάθους. Ο τρόπος σύνδεσης των επιπέδων σε ένα δίκτυο DenseNet επιτυγχάνει τη βελτιωμένη αξιοποίηση της πληροφορίας από τα ρηχότερα επίπεδα του δικτύου καθώς και την επαναχρησιμοποίηση της στα βαθύτερα επίπεδα του δικτύου, επιτρέποντας τελικά την έμμεση και βαθιά επίβλεψη της μάθησης. Ταυτόχρονα ελαττώνουν τόσο το πλήθος των συνολικών παραμέτρων όσο και το μέγεθος της μνήμης που χρησιμοποιούν αναφορικά με τα δίκτυα ResNet. Μεταξύ των πυκνών μπλοκ παρεμβάλλονται επίπεδα μετάβασης (Transition layers). Η συνολική αρχιτεκτονική ενός δικτύου DenseNet αποτελούμενο από 264 επίπεδα συνοψίζεται στην εικόνα (2.12).



Εικόνα 2.11 Αρχιτεκτονική ενός πυκνού μπλοκ



Εικόνα 2.12 Αρχιτεκτονική του δικτύου DenseNet 264 επιπέδων

2.4. Εκπαίδευση των Συνελικτικών Νευρωνικών Δικτύων

Η διαδικασία της εκπαίδευσης ενός νευρωνικού δικτύου αφορά την διαδικασία της εκμάθησης των τιμών των παραμέτρων του. Το σύνολο των παραμέτρων ενός συνελικτικού νευρωνικού δικτύου συγκροτούν τα βάρη και οι πολώσεις όλων των νευρώνων του. Ο προσδιορισμός των βέλτιστων τιμών των παραμέτρων ενός δικτύου το οποίο επιτελεί μια συγκεκριμένη διεργασία επιτυγχάνεται μέσω της ελαχιστοποίησης της κατάλληλης συνάρτησης κόστους ή αλλιώς συνάρτησης απωλειών (Loss Function).

2.4.1. Συνάρτηση Απωλειών

Η συνάρτηση απωλειών υπολογίζει για μια δεδομένη είσοδο του νευρωνικού δικτύου, το σφάλμα μεταξύ πραγματικής-επιθυμητής εξόδου και της πρόβλεψης που ανακτήθηκε από το επίπεδο εξόδου του δικτύου. Το σφάλμα συνιστά ένα δείκτη του βαθμού στον οποίο ανταποκρίνεται ένα νευρωνικό δίκτυο σε μια καθορισμένη διεργασία. Η όλη διαδικασία της εκπαίδευσης ενός νευρωνικού δικτύου, δηλαδή της εκμάθησης των παραμέτρων του αντιμετωπίζεται ως ένα πρόβλημα ολικής βελτιστοποίησης.

Ας ορίσουμε ως ϑ το σύνολο των παραμέτρων ενός συνελικτικού δικτύου. Ακόμη, ας υποθέσουμε ότι το σύνολο των δεδομένων εκπαίδευσης αποτελείται από N στο πλήθος δείγματα για τα οποία ισχύει η ακόλουθη σχέση:

$$\{x(n), y(n)\}, \qquad n \in [1, \dots, N] \tag{2.35}$$

όπου $x(n)$ είναι το n-οστό δείγμα των δεδομένων εισόδου του δικτύου και $y(n)$ η βάση αληθείας που αντιστοιχεί στο δείγμα $x(n)$.

Επιπλέον, ορίζοντας ως $o(n)$ την έξοδο του συνελικτικού δικτύου για το $x(n)$ δείγμα εισόδου τότε η συνάρτηση απώλειας του δικτύου δίνεται από την ακόλουθη σχέση:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} l(\vartheta; y^{(n)}, o^{(n)}) \tag{2.36}$$

2.4.2. Αλγόριθμος Καθοδικής Κλίσης

Όπως αναφέραμε νωρίτερα, η διαδικασία της εκπαίδευσης ενός νευρωνικού δικτύου συνιστά ένα πρόβλημα ολικής βελτιστοποίησης, το οποίο μπορεί να διατυπωθεί ως εξής:

$$\vartheta^* = \arg min_{\vartheta} \mathcal{L}(\vartheta) \quad \textbf{(2.37)}$$

Η προφανής λύση της τυχαίας επιλογής του συνδυασμού των παραμέτρων ϑ για τις οποίες ελαχιστοποιείται η συνάρτηση $\mathcal{L}(\vartheta)$, στην περίπτωση των νευρωνικών δικτύων θεωρείται υπολογιστικά ασύμφορη έως αδύνατη καθώς πρόκειται για ένα πολύ μεγάλο πλήθος παραμέτρων. Ωστόσο, υπό την παραδοχή ότι η συνάρτηση απώλειας που χρησιμοποιείται είναι διαφορίσιμη τότε ο αλγόριθμος της καθοδικής κλίσης (Gradient Descent) μπορεί να προσφέρει μια λύση στο πρόβλημα του προσδιορισμού των τιμών των παραμέτρων του δικτύου, για τις οποίες η συνάρτηση $\mathcal{L}(\vartheta)$ παρουσιάζει ολικό ελάχιστο. Η έννοια της κλίσης διαισθητικά αποτυπώνει το βαθμό στον οποίο μεταβάλλεται η έξοδος του δικτύου για μια δεδομένη μικρή διαφοροποίηση της εισόδου του δικτύου. Επί της ουσίας ο αλγόριθμος αυτός αναζητά τις βέλτιστες παραμέτρους ϑ, υπολογίζοντας την κλίση της συνάρτησης απώλειας $\nabla_{\vartheta}\mathcal{L}(\vartheta)$, δηλαδή τη μερική παράγωγο ως προς τη κάθε παράμετρο του δικτύου. Η μερική παράγωγος υποδεικνύει την κατεύθυνση προς την οποία θα πρέπει να κινηθούν οι παράμετροι ϑ , οι τιμές των οποίων ανανεώνονται με τη προσθήκη μιας μικρής ποσότητας προς την αντίθετη κατεύθυνση της κλίσης. Η σχέση (2.38) παρουσιάζει τον τρόπο με τον οποίο πραγματοποιείται η ανανέωση των παραμέτρων.

$$\vartheta_{t+1} = \vartheta_t - \eta \nabla_{\vartheta}\mathcal{L}(\vartheta_t), \quad \textbf{(2.38)}$$

όπου η είναι μια θετική σταθερά που δηλώνει το ρυθμό εκμάθησης (learning rate) δηλαδή τη ταχύτητα με την οποία το δίκτυο μαθαίνει να παράγει την επιθυμητή απόκριση για μια δεδομένη είσοδο.

Μεγάλη τιμή της σταθεράς του ρυθμού εκμάθησης συνεπάγεται γρήγορη εκπαίδευση του δικτύου αλλά αυξάνει σημαντικά το ενδεχόμενο αδυναμίας του τελικού μοντέλου να αντιμετωπίσει δεδομένα διαφορετικά από εκείνα στα οποία εκπαιδεύτηκε. Από την άλλη πλευρά, η μικρή τιμή της σταθεράς του ρυθμού εκμάθησης, αν και αποφεύγει το ενδεχόμενο της αστάθειας του δικτύου καθυστερεί την εκπαίδευση του. Επομένως, η εύρεση του κατάλληλου ισοζυγίου μεταξύ του ταχύτητας εκμάθησης και του βαθμού αξιοπιστίας του εκπαιδευμένου δικτύου αποτελεί ζητούμενο για τον προσδιορισμό της τιμής της σταθεράς του ρυθμού εκμάθησης. Για το λόγο αυτό έχουν προταθεί ποικίλες στρατηγικές (Loshchilov & Hutter, 2016), (Schaul, Zhang, & LeCun, 2013). Ωστόσο, μια αρκετά διαδεδομένη προσέγγιση είναι αρχικά να επιλέγεται μια μικρή τιμή για το ρυθμό εκμάθησης, ώστε ο τελευταίος να ξεκινήσει αργά και καθώς ο αλγόριθμος τείνει να συγκλίνει η τιμή αυτή να ελαττώνεται.

Ο αλγόριθμος της καθοδικής κλίσης εκτιμά κάθε φορά την κλίση της συνάρτησης απώλειας αξιοποιώντας όλο το σύνολο των δεδομένων εκπαίδευσης. Μια βελτιωμένη εκδοχή του συνιστά ο αλγόριθμος της στοχαστικής καθοδικής κλίσης (Stochastic Gradient Descent) ο οποίος υπολογίζει κάθε φορά την κλίση της συνάρτησης απώλειας βασιζόμενος μόνο σε ένα τυχαίο υποσύνολο των δειγμάτων εκπαίδευσης, το οποίο απαρτίζεται από ένα ή από μερικά δείγματα. Αν συμβολίσουμε το τυχαίο υποσύνολο των δειγμάτων εκπαίδευσης ως $(x^{(t)}, y^{(t)})$, τότε η ανανέωση των παραμέτρων ϑ με τη χρήση του αλγορίθμου της στοχαστικής καθοδικής κλίσης πραγματοποιείται ως εξής:

$$\vartheta_{t+1} = \vartheta_t - \eta_t \nabla_\vartheta \mathcal{L}\left(\vartheta_t; (x^{(t)}, y^{(t)})\right) \quad \textbf{(2.39)}$$

Στην πραγματικότητα ο αλγόριθμος της στοχαστικής καθοδικής κλίσης δε χρησιμοποιεί κάποιο μεμονωμένο δείγμα των δεδομένων εισόδου αλλά ένα μικρό σετ (minibatch) από αυτά. Η προσέγγιση που προτείνει, μειώνει την υπολογιστική πολυπλοκότητα και τη χρονική διάρκεια του ρυθμού εκπαίδευσης συγκριτικά τον κλασσικό αλγόριθμο της καθοδικής κλίσης εφόσον πραγματοποιεί συχνότερη ενημέρωση των τιμών των παραμέτρων του δικτύου. Ωστόσο, έχει αποδειχθεί ότι δεν οδηγεί πάντα σε καλή σύγκλιση γεγονός το οποίο καθιστά ασταθές το εκπαιδευμένο δίκτυο. Επιπλέον, η μέθοδος της στοχαστικής καθοδικής κλίσης εξαρτάται από το ρυθμό εκπαίδευσης (Ruder, 2016).

Για τους λόγους αυτούς, αρκετές παραλλαγές των αλγορίθμων καθοδικής κλίσης έχουν προταθεί προκειμένου, σε μια προσπάθεια διερεύνησης της περαιτέρω βελτίωσης της απόδοσης τους. Μεταξύ αυτών συγκαταλέγονται οι Momentum (Qian, 1999), RMSprop , Adagrad (Duchi, Hazan, & Singer, 2011), Adadelta (Zeiler, 2012), Adam (Kingma & Ba, 2014), Adamax (Kingma & Ba, 2014), Nadam (Dozat, 2016).

Στη μεθοδολογία που προτείνεται στην παρούσα εργασία, χρησιμοποιείται αποκλειστικά ο αλγόριθμος Adam. Για το σκοπό αυτό παρουσιάζεται ακολούθως το πλαίσιο λειτουργίας του αλγορίθμου Adam και της ορμής, καθώς η έννοια της ορμής αξιοποιείται από τον αλγόριθμο Adam.

2.4.3. Ορμή

Ειδικότερα, η έννοια της ορμής (momentum) συνυπολογίζει τις μεταβολές που προκάλεσαν στις παραμέτρους οι ανανεώσεις των πρόσφατων σετ δειγμάτων των δεδομένων της εισόδου. Ο παράγοντας της ορμής εισάγει τη χρήση ενός διανύσματος ταχύτητας u_t, που δείχνει την κατεύθυνση προς την οποία μεταβλήθηκε κάθε παράμετρος του δικτύου στην

τελευταία ανανέωση που πραγματοποιήθηκε. Το διάνυσμα της ταχύτητας συμβάλλει ούτως ώστε οι παράμετροι του δικτύου να μεταβάλλονται σε κάθε επόμενη ανανέωση προς την επιθυμητή κατεύθυνση, επισπεύδοντας με τον τρόπο αυτό τη σύγκλιση. Ταυτόχρονα, αποτρέπει την πιθανότητα εγκλωβισμού του αλγορίθμου καθοδικής κλίσης σε τοπικά ελάχιστα. Η ανανέωση των παραμέτρων ϑ του δικτύου γίνεται σύμφωνα με τις σχέσεις (2.40) και (2.41)

$$\vartheta_{t+1} = \vartheta_t - u_{t+1} \tag{2.40}$$

$$u_{t+1} = \gamma u_t - \eta_t \nabla_\vartheta \mathcal{L}\left(\vartheta_t; \left(x^{(t)}, y^{(t)}\right)\right) \tag{2.41}$$

όπου γ η παράμετρος της ορμής, συνήθως $\gamma = 0.9$

2.4.4. Αλγόριθμος Adam

Την έννοια της ορμής και της προσαρμοστικής ελάττωσης του ρυθμού εκμάθησης - τεχνική που υιοθετεί και ο αλγόριθμος Adagrad, συνδυάζει ο αλγόριθμος Adam (Adaptive Moment Estimation). Η ανανέωση των παραμέτρων δια μέσω του αλγορίθμου Adam πραγματοποιείται ως εξής:

$$\vartheta_{t+1} = \vartheta_t - \frac{\eta}{\sqrt{\hat{u}_t} + \varepsilon} \hat{m}_t \tag{2.42}$$

όπου $\hat{m}_t$ και $\hat{u}_t$ οι εκτιμήσεις για τον εκθετικό κινούμενο μέσο όρο και την εκθετική κινούμενη κεντροποιημένη διακύμανση των κλίσεων των παραμέτρων της συνάρτησης απώλειας.

Οι εκτιμήσεις για τις κεντρικές ροπές των κλίσεων δίνονται από τις σχέσεις (2.43) και (2.45):

$$\hat{m}_t = \frac{m_{t+1}}{1 - \beta_1^{t+1}} \tag{2.43}$$

όπου,

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1) \nabla_\theta (\mathcal{L}\left(\vartheta_t; \left(x^{(t)}, y^{(t)}\right)\right)) \tag{2.44}$$

και

$$\hat{u}_t = \frac{u_{t+1}}{1 - \beta_2^{t+1}} \tag{2.45}$$

όπου,

$$u_{t+1} = \beta_2 u_t + (1 - \beta_2) \nabla_\theta (\mathcal{L}\left(\vartheta_t; \left(x^{(t)}, y^{(t)}\right)\right)^2) \tag{2.46}$$

οι τιμές των υπερ-παραμέτρων β_1, β_2 , στις σχέσεις (2.43), (2.45) καθορίζουν τη τιμή της υπερ-παραμέτρου ε στη σχέση (2.42). Συνήθως οι παραπάνω υπερ-παράμετροι αρχικοποιούνται στις τιμές $\beta_1 = 0.9\,, \beta_2 = 0.999,\ \epsilon = 10^{-8}$

2.4.5. Αλγόριθμος Οπισθοδρόμησης Σφάλματος

Νωρίτερα αναφέραμε ότι ο υπολογισμός της κλίσης της συνάρτησης απωλειών καθιστά εφικτό για τον αλγόριθμο της καθοδικής κλίσης και των πολυάριθμων παραλλαγών του, τον προσδιορισμό ανανέωσης των παραμέτρων του δικτύου. Ωστόσο, δε διευκρινίστηκε ο τρόπος με τον οποίο εκτιμάται η κλίσης της συνάρτησης απώλειας. Όπως και στα κλασσικά νευρωνικά δίκτυα έτσι και στα συνελικτικά δίκτυα, η κλίση της συνάρτησης απώλειας υπολογίζεται μέσω του αλγορίθμου οπισθοδρόμησης σφάλματος (Error-Back Propagation). Ο αλγόριθμος οπισθοδρόμησης σφάλματος αποσκοπεί στην εύρεση των μερικών παραγώγων των ελεύθερων παραμέτρων του δικτύου, δηλαδή των βαρών και των πολώσεων του.

Τις μερικές παραγώγους των βαρών και των πολώσεων των δικτύου ως προς το επίπεδο q τις συμβολίζουμε ως εξής:

$$\nabla_{w^{(p,q)}} L = \frac{\partial L}{\partial w^{(p,q)}} \quad \textbf{(2.47)}$$

$$\nabla_{b^{(p,q)}} L = \frac{\partial L}{\partial b^{(p,q)}} \quad \textbf{(2.48)}$$

Προκειμένου να προσδιοριστούν οι ποσότητες των σχέσεων (2.47), (2.48) ορίζεται η συνάρτηση δέλτα σύμφωνα με τη σχέση που έπεται:

$$\delta_{i,j}{}^{(q)} = \frac{\partial L}{\partial y_{i,j}{}^{(q-1)}} \quad \textbf{(2.49)}$$

Η συνάρτηση δέλτα υπολογίζει σε κάθε επίπεδο q του δικτύου το σφάλμα της μερικής παραγώγου της συνάρτησης απώλειας ως προς την είσοδο του επιπέδου q (η οποία είναι η έξοδος του αμέσως προηγούμενου επιπέδου $q-1$)

Η έξοδος $y_{i,j}$ που αντιστοιχεί στο δισδιάστατο χάρτη χαρακτηριστικών που παράχθηκε από την επεξεργασία του πυρήνα και της πόλωσης p στο $(q-1)$ επίπεδο του δικτύου. Για λόγους απλούστευσης θα θεωρήσουμε ότι ο χάρτης χαρακτηριστικών $y_{i,j}$ περιλαμβάνει τις τιμές όπως αυτές διαμορφώθηκαν έπειτα από την επεξεργασία του επιπέδου απαλοιφής γραμμικότητας και του συγκεντρωτικού επιπέδου που ακολουθούν το συνελικτικό επίπεδο $(q-1)$.

Νωρίτερα, στη σχέση (2.17) ορίσαμε το είδος της επεξεργασίας για όλα τα φίλτρα και τον όγκο εισόδου σε ένα συνελικτικό επίπεδο του δικτύου. Ανάλογα, μειώνοντας την πολυπλοκότητα της προαναφερόμενης σχέσης θα ορίσουμε το είδος της επεξεργασίας μεταξύ της συνέλιξης ενός πυρήνα και μιας εκ των αναπαραστάσεων d_q του τρισδιάστατου όγκου εισόδου ως εξής:

$$z_{i,j}^{(q)} = \left(\sum_r \sum_s w_{r,s}^{(p,q)} \alpha_{i-r,\ j-s,}^{(q-1)}\right) + b^{(p,q)} \qquad \textbf{(2.50)}$$

$$= \qquad w^{(p,q)} * \alpha_{i,j,}^{(q-1)}) + b^{(p,q)}$$

Επιπλέον, μπορούμε να υποθέσουμε ότι από τη σχέση (2.51) λαμβάνουμε το χάρτη χαρακτηριστικών όπως διαμορφώνεται έπειτα από την συνάρτηση ενεργοποίησης $h(\cdot)$ του επιπέδου απαλοιφής γραμμικότητας.

$$\alpha_{i,j}^{(q)} = h(z_{i,j}^{(q)}) \qquad \textbf{(2.51)}$$

Πλέον, η σχέση (2.49) μπορεί να εκφραστεί ισοδύναμα σύμφωνα με την εξίσωση (2.52):

$$\delta_{i,j}^{(q)} = \frac{\partial L}{\partial z_{i,j}^{(q)}} \qquad \textbf{(2.52)}$$

Σύμφωνα με τον κανόνα σύνθετης παραγώγισης η σχέση (2.52) μπορεί να εκφραστεί ως εξής:

$$\delta_{i,j}^{(q)} = \frac{\partial L}{\partial z_{i,j}^{(q)}} = \sum_u \sum_v \frac{\partial L}{\partial z_{u,v}^{(q+1)}} \frac{\partial z_{u,v}^{(q+1)}}{\partial z_{i,j}^{(q)}} \qquad \textbf{(2.53)}$$

όπου u, v είναι οι δυο μεταβλητές άθροισης για την περιοχή των δυνατών τιμών του y.

Παρατηρώντας την εξίσωση (2.53) διαπιστώνουμε ότι πρώτος όρος της διπλής άθροισης είναι ο $\delta_{i,j}^{(q+1)}$. Επομένως, η εξίσωση (2.53) μπορεί να ξαναγραφεί ως εξής:

$$\delta_{i,j}^{(q)} = \frac{\partial L}{\partial z_{i,j}^{(q)}} = \sum_u \sum_v \delta_{i,j}^{(q+1)} \frac{\partial z_{u,v}^{(q+1)}}{\partial z_{i,j}^{(q)}} \qquad \textbf{(2.54)}$$

Αντικαθιστώντας τη σχέση (2.51) στη (2.50) και χρησιμοποιώντας την τιμή $z_{i,j}$ που προκύπτει, η προηγούμενη εξίσωση διαμορφώνεται σύμφωνα με την ακόλουθη σχέση:

$$\delta_{i,j}^{(q)} = \frac{\partial L}{\partial y_{i,j}^{(q-1)}} = \qquad \textbf{(2.55)}$$

$$= \sum_u \sum_v {\delta_{i,j}}^{(q+1)} \frac{\partial}{\partial {z_{i,j}}^{(q)}} [(\sum_r \sum_s {w_{r,s}}^{(p,q+1)} h({z_{u-r,v-s}}^{(q)}) + {b}^{(p,q+1)}]$$

Στην προηγούμενη σχέση η παράγωγος στο εσωτερικό των αγκυλών είναι ίση με το μηδέν εκτός εάν ισχύει ότι $u - r = i$ και $v - s = j$ και επειδή η παράγωγος ${b}^{(p,q+1)}$ ως προς το ${z_{i,j}}^{(q)}$ είναι μηδέν. Ωστόσο, αν ισχύει ότι $u - r = i$ και $v - s = j$ τότε θα ισχύει και ότι $u - i = r$ και $v - j = s$. Επομένως, η σχέση (2.55) μπορεί να εκφραστεί ως:

$$ {\delta_{i,j}}^{(q)} = \sum_u \sum_v {\delta_{i,j}}^{(q+1)} [\sum_{u-i} \sum_{v-j} {w_{u-i,v-j}}^{(p,q+1)} h'({z_{i,j}}^{(q)})] \quad \textbf{(2.56)}$$

Οι τιμές των i, j, u, v ορίζονται έξω από τους όρους στο εσωτερικό των αγκυλών. Από τη στιγμή που οι τιμές των μεταβλητών δεν αλλάζουν, οι ποσότητες $u - i$ και $v - j$ στο εσωτερικό των αγκυλών δεν είναι παρά δύο σταθερές. Άρα, η διπλή άθροιση αποτιμάται ίση με ${w_{u-i,v-j}}^{(p,q+1)} h'({z_{i,j}}^{(q)})$ και μπορούμε να ξαναγράψουμε την εξίσωση (2.56) ως:

$$ {\delta_{i,j}}^{(q)} = \sum_u \sum_v {\delta_{i,j}}^{(q+1)} {w_{u-i,v-j}}^{(p,q+1)} h'\left({z_{i,j}}^{(q)}\right) h'\left({z_{i,j}}^{(q)}\right)$$

$$= h'\left({z_{i,j}}^{(q)}\right) \sum_u \sum_v {\delta_{i,j}}^{(q+1)} {w_{u-i,v-j}}^{(p,q+1)} h'\left({z_{i,j}}^{(q)}\right) \quad \textbf{(2.57)}$$

$$= h'\left({z_{i,j}}^{(q)}\right) [{\delta_{i,j}}^{(q+1)} * {w_{-i,-j}}^{(p,q+1)}]$$

Οι αρνητικές τιμές στους παραπάνω δείκτες υποδηλώνουν ότι το w υφίσταται ανάκλαση ως προς αμφότερους τους χωρικούς άξονες. Η διαδικασία αυτή ισοδυναμεί με την περιστροφή του w κατά 180° δηλαδή:

$$ {\delta_{i,j}}^{(q)} = h'\left({z_{i,j}}^{(q)}\right) [{\delta_{i,j}}^{(q+1)} * rot180\{{w_{i,j}}^{(p,q+1)}\}] \quad \textbf{(2.58)}$$

επειδή οι πυρήνες είναι ανεξάρτητοι των i, j η παραπάνω σχέση αποκτά την τελική της μορφή στην εξίσωση (2.59).

$$ {\delta_{i,j}}^{(q)} = h'\left({z_{i,j}}^{(q)}\right) [{\delta_{i,j}}^{(q+1)} * rot180\{w^{(p,q+1)}\}] \quad \textbf{(2.59)}$$

Συνεπώς η μερική παράγωγος της σχέσης (2.47) μπορεί να εκφραστεί ισοδύναμα ως εξής:

$$\nabla_{w^{(p,q)}} L = \frac{\partial L}{\partial w^{(p,q)}} = \sum_i \sum_j \frac{\partial L}{\partial {z_{i,j}}^{(q)}} \frac{\partial {z_{i,j}}^{(q)}}{\partial {w_{r,s}}^{(p,q)}}$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} \frac{\partial {z_{i,j}}^{(q)}}{\partial {w_{r,s}}^{(p,q)}}$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} \left[\left(\sum_r \sum_s {w_{r,s}}^{(p,q)} h\left({z_{i-r,j-s}}^{(q-1)}\right) \right) + b^{(p,q)} \right]$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} \left[\left(\sum_r \sum_s {w_{r,s}}^{(p,q)} h\left({z_{i-r,j-s}}^{(q-1)}\right) \right) + b^{(p,q)} \right]$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} {z_{i-r,j-s}}^{(q-1)}$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} {a_{i-r,j-s}}^{(q-1)}$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} {a_{-(r-i),-(s-j)}}^{(q-1)}$$

$$= {\delta_{r,s}}^{(q)} * {a_{-r,-s}}^{(q-1)}$$

$$= {\delta_{r,s}}^{(q)} * rot180\{a^{(q-1)}\}$$

Συνεπώς,

$$\nabla_{w^{(p,q)}} L = \frac{\partial L}{\partial w^{(p,q)}} = {\delta_{r,s}}^{(q)} * rot180\{a^{(q-1)}\} \qquad \textbf{(2.60)}$$

Ομοίως η μερική παράγωγος της σχέσης (2.48) μπορεί να εκφραστεί ισοδύναμα ως εξής:

$$\nabla_{b^{(p,q)}} L = \frac{\partial L}{\partial b^{(p,q)}} = \sum_i \sum_j \frac{\partial L}{\partial {z_{i,j}}^{(q)}} \frac{\partial {z_{i,j}}^{(q)}}{\partial b^{(p,q)}}$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} \frac{\partial {z_{i,j}}^{(q+1)}}{\partial b^{(p,q)}}$$

$$= \sum_i \sum_j {\delta_{i,j}}^{(q)} \qquad \textbf{(2.61)}$$

αφού ισχύει ότι

$$\frac{\partial z_{i,j}{}^{(q)}}{\partial b^{(p,q)}} = 1 \; \forall \; i,j$$

Άρα,

$$\nabla_{b^{(p,q)}} L = \frac{\partial L}{\partial b^{(p,q)}} = = \sum_i \sum_j \delta_{i,j}{}^{(q)} \qquad \textbf{(2.62)}$$

Συμπερασματικά, για την εκπαίδευση ενός συνελικτικού δικτύου ακολουθείται η παρακάτω διαδικασία:

Σε πρώτο στάδιο, μικρών τυχαίων βαρών και τιμών πόλωσης αρχικοποιεί τις παραμέτρους του δικτύου. Ακολούθως, η είσοδος που λαμβάνεται στο επίπεδο εισόδου του δικτύου τροφοδοτείται σειριακά εξόδου του δικτύου. Κατά τη διαδικασία της εμπρόσθιας διάδοσης της εισόδου για κάθε νευρώνα που αντιστοιχεί στη θέση (i,j) σε κάθε χάρτη χαρακτηριστικών καθενός επιπέδου q, υπολογίζονται οι ποσότητες των σχέσεων (2.50) και (2.51). Έπειτα στο επίπεδο εξόδου του δικτύου εκτιμάται το σφάλμα της πρόβλεψης με τη βοήθεια της συνάρτησης απώλειας. Ακολούθως, ξεκινώντας από τα τελευταία επίπεδα του δικτύου με κατεύθυνση προς τα αρχικά υπολογίζονται για κάθε νευρώνα σε κάθε χάρτη χαρακτηριστικών οι μερικές παράγωγοι των παραμέτρων του δικτύου σύμφωνα με τις σχέσεις (2.60) και (2.62). Ωστόσο, οι συγκεντρωτικοί χάρτες χαρακτηριστικών ενός επιπέδου είναι μικρότεροι συγκριτικά με τους χάρτες χαρακτηριστικών που παρήχθησαν στο συνελικτικό επίπεδο. Προκειμένου, να επιτευχθεί συμφωνία διαστάσεων μεταξύ ενός συγκεντρωτικού χάρτη με τον αντίστοιχο χάρτη του συνελικτικού επιπέδου, ο κάθε συγκεντρωτικός χάρτης υφίσταται υπερδειγματοληψία (upsampling) πριν τον υπολογισμό των παραγώγων. Στη συνέχεια, οι τιμές των κλίσεων των ελεύθερων παραμέτρων του δικτύου τροφοδοτούνται στη συνάρτηση βελτιστοποίησης. Οι τιμές των βαρών και της πόλωσης που αντιστοιχούν σε κάθε χάρτη χαρακτηριστικών ενημερώνονται σύμφωνα με τον αλγόριθμο εκπαίδευσης που χρησιμοποιείται. Μετά την ανανέωση όλων των παραμέτρων του δικτύου τροφοδοτείται σε αυτό η επόμενη είσοδος.

3. Οπτική Σημαντικότητα

3.1. Οπτική Αντίληψη & Σημαντικότητα

Η ανθρώπινη οπτική προσοχή έχει αποτελέσει για δεκαετίες αντικείμενο εκτεταμένης μελέτης για τις επιστήμες της Νευρολογίας, της Γνωστικής Ψυχολογίας και τα τελευταία περίπου είκοσι χρόνια έχει κεντρίσει το ερευνητικό ενδιαφέρον στη κοινότητα της Υπολογιστικής Όρασης (Qiuxia et al., 2020).

Δεδομένου ότι ο μηχανισμός της οπτικής προσοχής διαδραματίζει ρόλο ζωτικής σημασίας για την ανθρώπινη αντίληψη (Qiuxia et al., 2020), η υπολογιστική προσομοίωση μοντέλων οπτικής προσοχής όχι μόνο συνεισφέρει στην αποκρυπτογράφηση του τρόπου λειτουργίας του συγκεκριμένου μηχανισμού (Itti & Koch, 2001), αλλά διευκολύνει και τον τομέα της υπολογιστικής όρασης προκειμένου να παραγάγει αποτελέσματα περισσότερο συνεπή ως προς τη βιολογική λειτουργία του ανθρώπινου οπτικού συστήματος (Qiuxia et al., 2020).

Το ανθρώπινο οπτικό σύστημα φαίνεται να παρουσιάζει την αξιοπερίεργη ιδιότητα της εξαιρετικά γρήγορης και ταυτόχρονης επεξεργασίας υπερμεγέθους ποσότητας οπτικής πληροφορίας. Η ποσότητα της οπτικής πληροφορίας που δύναται να επεξεργαστεί ένας άνθρωπος μπορεί να αγγίξει τα 108 -109 bits ανά δευτερόλεπτο (Koch et al., 2006).

Πίσω από αυτό το εντυπωσιακό φαινόμενο, κρύβεται ο επιλεκτικός μηχανισμός της οπτικής προσοχής ο οποίος επιτρέπει στο ανθρώπινο οπτικό σύστημα να στέφει την προσοχή του στα κατεξοχήν οπτικά σημαντικά (visual salient) μέρη μιας εικόνας αντί να αναλώνεται στην εξ' ολοκλήρου επεξεργασία μιας οπτικής αναπαράστασης (W. Wang & Shen, 2017).

Ομολογουμένως, η ανίχνευση περιεχομένου οπτικού ενδιαφέροντος (visual saliency detection) αποτελεί αντικείμενο προσεκτικής μελέτης για την υπολογιστική όραση, καθώς η αποδοτική αναγνώριση σημείων, αντικειμένων ή περιοχών αντιπροσωπευτικών για μια εικόνα συνδράμει στην αντιμετώπιση σύνθετων προβλημάτων αυτού του πεδίου όπως: η αναγνώριση σκηνής (scene understanding) (Bharath, Nicholas, & Cheng, 2013), η αναγνώριση αντικειμένων (object recognition) (Gao & Vasconcelos, 2005), η κατάτμηση αντικειμένων (object segmentation) (W. Wang, Shen, & Shao, 2015), η ιχνηλάτηση (visual tracking) (Mahadevan & Vasconcelos, 2009), η κατανόηση περιεχόμενου βίντεο (video understanding) (Ling Zhang, Zhang, & Xiao, 2015) κα.

Για τον εντοπισμό οπτικά σημαντικού περιεχομένου έχουν προταθεί δυο διαφορετικές υπολογιστικές προσεγγίσεις. Η ανίχνευση οπτικής προσήλωσης (Fixation Prediction) και η ανίχνευση οπτικά σημαντικών αντικειμένων (Salient Object Detection - SOD). Ειδικότερα, η ανίχνευση οπτικής προσήλωσης στοχεύει στον εντοπισμό σημείων στα οποία προσηλώνεται το βλέμμα (fixation points) στην πρώτη παρατήρηση μιας οπτικής αναπαράστασης (W. Wang et al., 2019), ενώ η ανίχνευση σημαντικών αντικειμένων αποσκοπεί στον εντοπισμό των αντικειμένων που ξεχωρίζουν σε μια δοθείσα οπτική σκηνή (Zhou, Fan, Cheng, Shen, & Shao, 2020).

Η ποσοτική αναπαράσταση της πρόβλεψης που υλοποιούν αντίστοιχα οι παραπάνω μέθοδοι αποτυπώνεται σε χάρτες προσήλωσης (Eye Fixation Maps) και σε χάρτες οπτικά σημαντικών χαρακτηριστικών (Saliency Maps). Ουσιαστικά, πρόκειται για εικόνες των οποίων η φωτεινότητα των εικονοστοιχείων τους, αντικατοπτρίζει την πιθανότητα το αντίστοιχο εικονοστοιχείο της προβληθείσας οπτικής σκηνής, να αποσπά την οπτική προσοχή του παρατηρητή της. Η σχέση μεταξύ φωτεινότητας και πιθανότητας είναι αναλογική (W. Wang & Shen, 2017).

Το γεγονός ότι η ανίχνευση οπτικά σημαντικών αντικειμένων δεν παρέχει απλώς τη δυνατότητα επισήμανσης σημαντικών σημείων αλλά υποδεικνύει ολόκληρες τις σημαντικές περιοχές μιας εικόνας, σε συνδυασμό με την ανάπτυξη των αλγορίθμων Βαθιάς Μηχανικής Μάθησης οι οποίοι βρίσκονται στο απόγειο τους τα τελευταία χρόνια, την καθιστούν εξαιρετικά χρήσιμο εργαλείο για την αντιμετώπιση σύνθετων προβλημάτων υπολογιστικής όρασης (W. Wang et al., 2019)

Πράγματι, η ανίχνευση οπτικά σημαντικών αντικειμένων μετρά ήδη πολυάριθμες εφαρμογές σε ένα ευρύ φάσμα προβλημάτων της υπολογιστικής όρασης καθώς και της επεξεργασίας και ανάλυσης εικόνων μεταξύ των οποίων συγκαταλέγονται: η κατανόηση εικόνας (image understanding) (J.-Y. Zhu, Wu, Xu, Chang, & Tu, 2014), η αυτόματη περιγραφή εικόνων (image captioning) (H. Fang et al., 2015), η ανάκτηση εικόνων ή βίντεο (image/video retrieval) (G. Liu & Fan, 2013), (Sun et al., 2016), η συμπίεση εικόνων (image compression) (S. Han & Vasconcelos, 2006), η βελτιστοποίηση εικόνων (image enhancement) (Lei et al., 2015), η σημασιολογική κατάτμηση εικόνων (semantic segmentation) (Yunchao Wei et al., 2016), η κατάτμηση ιατρικών εικόνων medical (image segmentation) (Fan, Ji, et al., 2020), η επισημείωση εικόνων (image annotation) (Lei, Duan, Wu, Ling, & Hou, 2016), η μη επιβλεπόμενη κατάτμηση αντικειμένων σε βίντεο (un-supervised video object segmentation) (W. Wang, Shen, Yang, & Porikli, 2017), η περίληψη βίντεο (video summarization) (Simakov,

Caspi, Shechtman, & Irani, 2008), η κωδικοποίηση βίντεο (video coding) (Lei et al., 2016), η ανίχνευση αντικειμένων (object detection) (D. Zhang, Meng, Zhao, & Han, 2017), η ιχνηλάτηση αντικειμένων (object tracking) (Mahadevan & Vasconcelos, 2009), η αναγνώριση κίνησης (action recognition) (Rapantzikos, Avrithis, & Kollias, 2009) και η επανα-ταύτιση ατόμων (person re-identification) (R. Zhao, Oyang, & Wang, 2017). Πέραν των προαναφερθέντων πεδίων η ανίχνευση σημαντικών αντικειμένων υιοθετείται σε εφαρμογές ρομποτικής όπως: η αλληλεπίδραση ανθρώπου – μηχανής (human-robot interaction) (Borji & Itti, 2014) ή η ανακάλυψη αντικειμένων (object discovery) (Karpathy, Miller, & Fei-Fei, 2013) με απώτερο σκοπό την καλύτερη κατανόηση της σκηνής και των αντικειμένων της.

3.2. Μέθοδοι ανίχνευσης οπτικά σημαντικών αντικειμένων σε RGB εικόνες

Αναφορικά με την ανίχνευση αντικειμένων οπτικού ενδιαφέροντος σε RGB εικόνες, δυο κύριες τεχνικές συναντώνται στη βιβλιογραφία. Οι ευριστικές (Heuristic-based) μεθοδολογίες και οι προσεγγίσεις που βασίζονται στη μάθηση (Learning-based) (Ullah et al., 2020).

Οι ευριστικές τεχνικές, στηρίζονται στην εξαγωγή χαρακτηριστικών χαμηλού επίπεδου και στην πλειονότητα των περιπτώσεων αξιοποιούν μη επιβλεπόμενους αλγόριθμους μάθησης (W. Wang et al., 2019). Η παρατήρηση ότι η αντίθεση συνιστά βαρυσήμαντο παράγοντα για την οπτική υπεροχή, αφού ο ανθρώπινος εγκέφαλος είναι εξαιρετικά επιρρεπής στην προσήλωση σε περιοχές μιας οπτικής σκηνής που εμφανίζουν μεγάλη αντίθεση, αποτελεί θεμέλιο λίθο των ευριστικών μεθοδολογιών (Jian et al., 2019). Αυτού του είδους οι προσεγγίσεις, χρησιμοποιούν ευριστικά χαρακτηριστικά όπως η αντίθεση, η υφή, το χρώμα και η θέση για να ανιχνεύσουν τα αντικείμενα εκείνα που ξεχωρίζουν σε μια εικόνα. Δηλαδή, μοντελοποιούν τον λεγόμενο από κάτω-προς-τα-πάνω (Bottom-up) ή αλλιώς εξωγενή μηχανισμό λειτουργίας της επιλεκτικής προσοχής (Qiuxia et al., 2020).

Όσον αφορά τις ευριστικές μεθοδολογίες, θα μπορούσε κανείς να τις ταξινομήσει σε έξι κύριες κατηγορίες (W. Wang et al., 2019).

Η πρώτη ομάδα, αναφέρεται σε εκείνες οι οποίες αξιοποιούν την τοπική αντίθεση (local-contrast-prior). Πρόκειται για μεθοδολογίες (Itti & Koch, 1999), (Perazzi, Krähenbühl, Pritch, & Hornung, 2012) που ποσοτικοποιούν τη διαφορά μεταξύ δυο ή περισσότερων εικονοστοιχείων ή περιοχών σε μια εικόνα για να εκτιμήσουν το βαθμό υπεροχής των αντικειμένων της. Τοπικά χαρακτηριστικά όπως το χρώμα, η φωτεινότητα, η υφή ή ο

προσανατολισμός ανακτώνται από περιοχές των εικόνων για να κατασκευαστούν οι χάρτες οπτικού ενδιαφέροντος μιας αναπαράστασης.

Μια δεύτερη κατηγορία αφορά τις μεθόδους εκείνες που συνεκτιμούν τη συνολική αντίθεση σε μια εικόνα (global-contrast-prior). Για να εκτιμήσουν το βαθμό στον οποίο μια περιοχή της μιας σκηνής θεωρείται σημαντική, οι μέθοδοι αυτές (M. Cheng, Zhang, Mitra, Huang, & ű, 2011), (Perazzi et al., 2012) διαχωρίζουν τα αντικείμενα μιας εικόνας και συγκρίνουν την εκάστοτε περιοχή με ολόκληρη την εικόνα.

Οι παρατηρήσεις ότι τα περισσότερο σημαντικά αντικείμενα έχουν την τάση να συγκεντρώνονται είτε στο κέντρο, είτε στο προσκήνιο μιας οπτικής αναπαράστασης, αποτέλεσαν αφορμή για την ανάπτυξη ευριστικών αλγορίθμων, που αξιοποιούν την αντίθεση της κεντρικής περιοχής (center-prior) σε σύγκριση με την υπόλοιπη εικόνα (Jian et al., 2018) ή την αντίθεση μεταξύ προσκηνίου και παρασκηνίου (backgroundness-prior) για να παραγάγουν χάρτες οπτικού ενδιαφέροντος (Yichen Wei, Wen, Zhu, & Sun, 2012).

Η πρωτότυπη ιδέα του Alexe και των συνεργατών του (Alexe, Deselaers, & Ferrari, 2010), να μετρήσουν την πιθανότητα να απαθανατιστεί ολόκληρο ένα αντικείμενο σε μια τυχαία περιοχή μιας εικόνας συντέλεσε στην δημιουργία μιας επιπλέον κατηγορίας μεθόδων, οι οποίες αξιοποιούν την έννοια της αντικειμενικότητας (objectness-prior). Η έννοια αυτή που διερευνά το κατά πόσο μια περιοχή σε μια εικόνα αποτελεί αντικείμενο της, έχει χρησιμοποιηθεί προκειμένου να ανιχνευθούν τα σημαντικά αντικείμενα μιας οπτικής αναπαράστασης (Chang, Liu, Chen, & Lai, 2011), (P. Jiang, Ling, Yu, & Peng, 2013) .

Η έκτη κατηγορία ευριστικών αλγορίθμων, προσεγγίζει το ζήτημα του εντοπισμού των αντικειμένων υψηλού ενδιαφέροντος σε μια εικόνα, αντιμετωπίζοντας το ως στατιστικό πρόβλημα υιοθετώντας Μπεϋζιανά μοντέλα (Xie, Lu, & Yang, 2012), (R. Liu, Cao, Lin, & Shan, 2014).

Πέραν, των παραπάνω μεθοδολογιών έχουν προταθεί ενδιαφέρουσες ευριστικές προσεγγίσεις που καταφεύγουν στην ανάλυση στο πεδίο της συχνότητας (Achanta, Hemami, Estrada, & Susstrunk, 2009), στη χρήση κυψελικών αυτομάτων (Qin, Lu, Xu, & Wang, 2015), σε μοντέλα αραιής αναπαράστασης (X. Li, Lu, Zhang, Ruan, & Yang, 2013) ή ακόμη και σε μοντέλα τυχαίων περιπάτων (C. Li, Yuan, Cai, Xia, & Dagan Feng, 2015) .

Στον αντίποδα των ευριστικών τεχνικών, βρίσκονται οι μεθοδολογίες που βασίζονται στη Μάθηση. Οι ευριστικές μεθοδολογίες, οι οποίες βασίζονται σε "χειροποίητα" (hand-crafted) χαρακτηριστικά, χαρακτηρίζονται συχνά από περιορισμένη δυνατότητα να ανιχνεύουν

αντικείμενα υψηλού ενδιαφέροντος σε εικόνες (Borji et al., 2019). Αντίθετα, οι μέθοδοι που βασίζονται στη μάθηση και οι οποίες είναι είτε επιβλεπόμενες είτε ημί-επιβλεπόμενες έχουν την δυνατότητα να ενσωματώνουν πέραν των χαρακτηριστικών χαμηλού επιπέδου και πληροφορία υψηλού επιπέδου, βελτιώνοντας σε αρκετές περιπτώσεις με καθοριστικό τρόπο την απόδοση τους στο ίδιο εγχείρημα (Ullah et al., 2020).

Μεταξύ των αλγορίθμων Βαθιάς Μάθησης που έχουν παρουσιαστεί σε αυτή την κατηγορία μεθοδολογιών συγκαταλέγονται: οι ταξινομητές τυχαίων δασών, οι διανυσματικές μηχανές υποστήριξης (Lihe Zhang, Zhang, Sun, Wei, & Bo, 2019), τα υπό συνθήκη τυχαία πεδία (Yang & Yang, 2016) κα. Πρόσφατα, δεδομένης της επιτυχίας των εφαρμογών των συνελικτικών νευρωνικών δικτύων στο πεδίο της υπολογιστικής όρασης, ο εν λόγω αλγόριθμος βαθιάς μάθησης, όχι απλώς έχει χρησιμοποιηθεί σε εφαρμογές εντοπισμού αντικειμένων οπτικής υπεροχής σε εικόνες, αλλά η συνδυαστική μελέτη τους συνιστά ανερχόμενη τάση όσον αφορά αυτό το ερευνητικό αντικείμενο (Ullah et al., 2020).

Τα συνελικτικά νευρωνικά δίκτυα, παρέχουν τη δυνατότητα επεξεργασίας χαρακτηριστικών προερχόμενων από ποικίλες κλίμακες σε πολλαπλά επίπεδα. Εξαιτίας αυτής της ικανότητας τους, είναι σε θέση να ανιχνεύουν τα αντικείμενα που ξεχωρίζουν σε μια εικόνα, χωρίς να απαιτείται να γνωρίζουν εκ των προτέρων κάποιου είδους πληροφορία όπως πχ γνώση κατάτμησης για την εικόνα. Επιπλέον, τα πολυδιάστατα χαρακτηριστικά που μαθαίνουν, τους επιτρέπουν να οριοθετούν καλύτερα τα αντικείμενα που εντόπισαν ως σημαντικά, ακόμη κι όταν αυτά εμφανίζουν σκιές ή αντανακλάσεις στις RGB εικόνες (Borji et al., 2019).

Οι μεθοδολογίες που αξιοποιούν τα συνελικτικά νευρωνικά δίκτυα διακρίνονται σε δυο κατηγορίες. Η πρώτη κατηγορία είναι εκείνη που μαζί με τα συνελικτικά δίκτυα αξιοποιεί και πολυστρωματικά νευρωνικά δίκτυα (Multi-Layer Perceptrons MLPs). Στο γενικό πλαίσιο λειτουργίας τους, τα μοντέλα που αναπτύσσονται υπό αυτό το πρίσμα, ανακτούν χαρακτηριστικά υψηλού επιπέδου με τη βοήθεια ενός συνελικτικού δικτύου από τις εικόνες εισόδου, οι οποίες είθισται σε προγενέστερο στάδιο να υφίστανται κατάτμηση και ακολούθως ένα πολυστρωματικό νευρωνικό δίκτυο χρησιμοποιείται για να προβλέψει το επίπεδο σημαντικότητας της κάθε περιοχής (Ullah et al., 2020). Οι προσεγγίσεις αυτής της κατηγορίας, οι οποίες καλούνται προσεγγίσεις βασισμένες σε συνελικτικά νευρωνικά δίκτυα (CNN-based approaches) ή αλλιώς μοντέλα που βασίζονται σε περιοχές (region-based models), αν και χρησιμοποιούν μεθόδους για να απαθανατίσουν την υψηλού επιπέδου σημασιολογική πληροφορία που ενυπάρχει στις εικόνες δεν καταφέρνουν να την αξιοποιήσουν με βέλτιστο

τρόπο καθώς η χωρική πληροφορία δε διατηρείται από τα τελευταία πλήρως συνδεδεμένα επίπεδα των πολυστρωματικών νευρωνικών δικτύων (Borji et al., 2019).

Στην προσπάθεια αντιμετώπισης των προαναφερόμενων αδυναμιών, τις οποίες παρουσιάζουν οι τεχνικές που βασίζονται σε συνελικτικά νευρωνικά δίκτυα, προέκυψαν οι προσεγγίσεις της δεύτερης κατηγορίας, οι επονομαζόμενες και μέθοδοι που βασίζονται πλήρως σε συνελικτικά δίκτυα (Fully Convolutional Networks, FCN-based approaches). Τα μοντέλα αυτής της κατηγορίας έχουν τη δυνατότητα να εκτιμούν χάρτες οπτικού ενδιαφέροντος σε επίπεδο εικονοστοιχείου (pixel-based models) (Ullah et al., 2020). Συγκριτικά, με τις μεθοδολογίες της πρώτης κατηγορίας, οι προσεγγίσεις αυτές προσφέρουν βελτιωμένα αποτελέσματα καθώς i) έχουν την ικανότητα να συλλαμβάνουν τόσο τις λεπτομερείς πληροφορίες σε τοπικό επίπεδο -από τα ρηχότερα επίπεδα του δικτύου- όσο και τη συνολική και σημασιολογικά σημαντική πληροφορία σε επίπεδο εικόνας - από τα βαθύτερα επίπεδα του δικτύου. ii) δεν είναι απαραίτητη η εκπαίδευση τους από τη αρχή, εφόσον μπορούν να αξιοποιήσουν μέσω της διαδικασίας προσαρμογής (fine-tuning) γνώσεις που έχουν αποκτηθεί από συνελικτικά δίκτυα σε διαφορετικά είδη εργασιών όπως για παράδειγμα η ταξινόμηση εικόνων iii) καθιστούν εφικτό τον σχεδιασμό ποικίλων και ευέλικτων αρχιτεκτονικών (Borji et al., 2019).

Ακολούθως θα παρουσιαστούν αναλυτικά μερικές από τις πιο αντιπροσωπευτικές εργασίες οι οποίες αξιοποιούν συνελικτικά νευρωνικά δίκτυα με απώτερο σκοπό τη διάκριση αντικειμένων που ξεχωρίζουν σε RGB εικόνες.

3.3. Συνελικτικά Δίκτυα & Ανίχνευση οπτικά σημαντικών αντικειμένων σε RGB εικόνες

3.3.1. Προσεγγίσεις που βασίζονται σε Συνελικτικά Νευρωνικά Δίκτυα

Μεταξύ των πιο χαρακτηριστικών εργασιών που έχουν παρουσιαστεί και εντάσσονται στην κατηγορία των μοντέλων που βασίζονται σε συνελικτικά δίκτυα (CNN-based approaches) είναι οι εξής:

Η πρώτη εργασία που χρησιμοποιεί τα συνελκτικά δίκτυα για τον εντοπισμό αντικειμένων που ξεχωρίζουν RGB εικόνες παρουσιάστηκε από τους Li & Yu (G. Li & Yu, 2015). Οι

συγγραφείς της εργασίας, κατασκευάζουν χάρτες οπτικού ενδιαφέροντος με ακρίβεια πρόβλεψης σε επίπεδο περιοχής, εκμεταλλευόμενοι την εξαγωγή χαρακτηριστικών σε πολλαπλές κλίμακες. Συγκεκριμένα, κάθε εικόνα διαχωρίζεται σε μη επικαλυπτόμενες περιοχές, κάθε μια από τις οποίες υφίσταται επεξεργασία από ένα συνελικτικό δίκτυο προ-εκπαιδευμένο στην ταξινόμηση εικόνων, με σκοπό την εξαγωγή χαρακτηριστικών σε τρεις διαφορετικές κλίμακες από την κάθε περιοχή. Τα χαρακτηριστικά που εξάγονται από όλες τις περιοχές για τις τρεις διαφορετικές κλίμακες συνενώνονται και διέρχονται από ένα πολυστρωματικό νευρωνικό δίκτυο μόλις δυο πλήρως συνδεδεμένων επιπέδων το οποίο παράγει τον χάρτη οπτικού ενδιαφέροντος. Επιπλέον, οι συγγραφείς προτείνουν την μετα-επεξεργασία των εκτιμώμενων χαρτών σχεδιάζοντας έναν αλγόριθμο που βασίζεται στην πληροφορία υπερ-εικονοστοιχείων.

Ο He και οι συνεργάτες του (S. He, Lau, Liu, Huang, & Yang, 2015) χρησιμοποιούν ένα συνελικτικό δίκτυο για να εκτιμήσουν χάρτες οπτικού ενδιαφέροντος σε επίπεδο υπερ-εικονοστοιχείου. Η προτεινόμενη προσέγγιση ενώ λαμβάνει υπόψη τη συνολική χωρική πληροφορία των εικόνων ελαχιστοποιεί σε σημαντικό βαθμό το υπολογιστικό κόστος συγκριτικά με τα μοντέλα που υλοποιούν την πρόβλεψη σε επίπεδο εικονοστοιχείου. Ωστόσο το γεγονός ότι τα υπερ-εικονοστοιχεία κωδικοποιούνται με βάση τη χρωματική πληροφορία σε συνδυασμό με το ότι το δίκτυο επεξεργάζεται τις εικόνες σε μια διάσταση και συνεπώς κάνει χρήση της πράξης της μονοδιάστατης συνέλιξης, περιορίζει σημαντικά την απόδοση του μοντέλου ιδιαίτερα όταν αυτό έρχεται αντιμέτωπο με περίπλοκες αναπαραστάσεις.

Σε μια άλλη εργασία ο Wang και οι συνεργάτες του (L. Wang, Lu, Ruan, & Yang, 2015) αξιοποιούν δυο διαφορετικά βαθιά νευρωνικά δίκτυα για εκμάθηση χαρακτηριστικών τόσο σε τοπικό όσο και σε συνολικό επίπεδο. Αρχικά, ένα δίκτυο υπολογίζει τη σημαντικότητα κάθε εικονοστοιχείου με τη βοήθεια τοπικών χαρακτηριστικών όπως το σχήμα, η υφή και η αντίθεση και έπειτα εφαρμόζεται ένας επιδιορθωτικός μηχανισμός που βασίζεται στην υψηλού επιπέδου έννοια της αντικειμενικότητας. Κατόπιν, τα αποτελέσματα αυτά αξιοποιούνται από ένα δεύτερο δίκτυο το οποίο χρησιμοποιεί εξάγει συνολικά χαρακτηριστικά όπως η γεωμετρική πληροφορία και η συνολική αντίθεση, για να προβλέψει σε επίπεδο περιοχής το βαθμό στον οποίο αυτή συγκεντρώνει ενδιαφέρον.

Μια ακόμη προσέγγιση που συνυπολογίζει τη τοπική και τη συνολική πληροφορία αναπτύχθηκε από το Zhao και τους συνεργάτες του (R. Zhao, Ouyang, Li, & Wang, 2015). Οι συγγραφείς σχεδίασαν ένα συνελικτικό δίκτυο που υποστηρίζει την παράλληλη επεξεργασία των εικόνων. Ειδικότερα, ο ένας κλάδος του δικτύου χρησιμοποιεί ολόκληρη την εικόνα για

να ανακτήσει πληροφορία του αφορά τη συνολική της αντίθεση, ενώ ένας δεύτερος κλάδος δέχεται στην είσοδο του ένα υπερ-εικονοστοιχείο για την εξαγωγή τοπικών χαρακτηριστικών. Ένας δυαδικός ταξινομητής αναλαμβάνει χαρακτηρίσει κάθε υπερ-εικονοστοιχείο ως σημαντικό ή μη, βελτιστοποιώντας ταυτόχρονα και τις δυο ροές δεδομένων. Επιπλέον, προτείνεται μια εξατομικευμένη στρατηγική αρχικοποίησης των παραμέτρων του δικτύου προκειμένου να παραγάγει όσο το δυνατόν πιο συνεπή ως προς τη βάση αλήθειας αποτελέσματα.

Δυο διαφορετικά δίκτυα για διαχωριστή εξαγωγή χαρακτηριστικών χαμηλού και υψηλού επιπέδου υιοθετούν στην εργασία τους ο Lee και οι συνεργάτες του (Lee, Tai, & Kim, 2016). Αρχικά, χαρακτηριστικά χαμηλού επιπέδου όπως το χρώμα, η υφή και χωρική πληροφορία ανακτώνται από κάθε υπερ-εικονοστοιχείο της εικόνας και αφού αυτή η πληροφορία κωδικοποιηθεί εισέρχεται σε ένα δίκτυο το οποίο συντίθεται από συνελικτικά επίπεδα που χρησιμοποιούν φίλτρα μεγέθους 1x1. Για την εξαγωγή πληροφορίας υψηλού επιπέδου αξιοποιείται το δίκτυο VGG-16, από την αρχιτεκτονική του οποίου νωρίτερα έχουν αφαιρεθεί τα πλήρως συνδεδεμένα επίπεδα. Τέλος, οι χαμηλού και υψηλού επιπέδου αναπαραστάσεις ενός υπερ-εικονοστοιχείου συνενώνονται και ένα πολυστρωματικό δίκτυο δύο επιπέδων χρησιμοποιείται προκειμένου να αποφανθεί αν η περιοχή που οριοθετείτε από το εκάστοτε εικονοστοιχείο θεωρείται υψηλού ενδιαφέροντος.

Ένα συνελικτικό δίκτυο που υποστηρίζει μια διπλή ροή δεδομένων στην είσοδο του, για την ταυτόχρονη επεξεργασία τόσο αδρών όσο και περισσότερο λεπτομερών περιοχών, συστήνεται από τον Kim και τους συνεργάτες του (Kim & Pavlovic, 2016). Η προτεινόμενη προσέγγιση δανείζεται την μέθοδο επιλεκτικής αναζήτησης αντικειμένων των (Uijlings, Van De Sande, Gevers, & Smeulders, 2013), η οποία αρχικά είχε προταθεί για την αναγνώριση αντικειμένων σε εικόνες, προκειμένου να επιλεγούν οι πιο αντιπροσωπευτικές περιοχές (αδρές και λεπτομερείς) που θα ληφθούν από τις εισόδους του δικτύου. Τελικά, οι αναπαραστάσεις που παράγονται από τους δυο κλάδους του δικτύου συνενώνονται και διέρχονται από ένα πλήρως συνδεδεμένο επίπεδο το οποίο προβλέπει ένα χάρτη οπτικού ενδιαφέροντος σε επίπεδο περιοχής. Ένας επιδιορθωτικός μηχανισμός που στηρίζεται σε μια μέθοδο ιεραρχικής κατάτμησης εφαρμόζεται για την περαιτέρω βελτίωση των αποτελεσμάτων.

Ο Wang και συνεργάτες του (X. Wang, Ma, & Chen, 2016) προσαρμόζουν το συνελικτικό δίκτυο Fast R-CNN του Girshick (Girshick, 2015), προκειμένου να εξάγουν πληροφορία υψηλού επιπέδου η οποία θα υποβοηθήσει στην ταυτοποίηση εκείνων των περιοχών μιας εικόνας που ξεχωρίζουν. Αρχικά, ένας αλγόριθμος κατάτμησης που διατηρεί τις ακμές

εφαρμόζεται στις εικόνες. Το μοντέλο Fast R-CNN χρησιμοποιείται για να απαθανατίσει την υψηλού επιπέδου πληροφορία που φέρουν οι εικόνες. Στο τέλος του, δίπλα στο λεγόμενο επίπεδο δειγματοληψίας της περιοχής ενδιαφέροντος (ROI pooling layer) εδράζεται ένα ρηχό δίκτυο το οποίο συντίθεται από πλήρως συνδεδεμένα επίπεδα και το οποίο είναι επιφορτισμένο ούτως ώστε να εκτιμά το βαθμό στον οποίο μια περιοχή της εικόνας θεωρείται σημαντική. Χαμηλού χαρακτηριστικά όπως η αντίθεση καθώς και πληροφορία που αφορά το προσκήνιο όπως και ένας αλγόριθμος ανάδειξης ακμών χρησιμοποιούνται για την περαιτέρω βελτίωση των αποτελεσμάτων.

3.3.2. Προσεγγίσεις που βασίζονται σε Πλήρως Συνελικτικά Νευρωνικά Δίκτυα

Σχετικά με τις μεθοδολογίες που θεμελιώνονται αποκλειστικά σε πλήρως συνελικτικά δίκτυα (FCN-based approaches), οι πιο χαρακτηριστικές που έχουν προταθεί είναι οι ακόλουθες:

Ο Li και οι συνεργάτες του (G. Li & Yu, 2016) παράγουν χάρτες οπτικού ενδιαφέροντος με τη βοήθεια ενός συνελικτικού δικτύου συντιθέμενο από δυο κλάδους, που επεξεργάζονται παράλληλα τα δεδομένα εισόδου. Ο πρώτος κλάδος του εν λόγω μοντέλου, χρησιμοποιεί το δίκτυο των Simonyan & Zisserman (Simonyan & Zisserman, 2014) για να εξάγει χαρακτηριστικά σε πολλαπλές κλίμακες, και να κατασκευάσει από τη σύντηξη τους, έναν αρχικό χάρτη οπτικά σημαντικών αντικειμένων με ακρίβεια εικονοστοιχείου. Ο δεύτερος κλάδος του δικτύου, δανείζεται και αυτός την αρχιτεκτονική του δικτύου VGG-16 για να εκτιμήσει ένα δεύτερο χάρτη σε επίπεδο υπερ-εικονοστοιχείων αυτή φορά, με σκοπό να αξιοποιήσει την πληροφορία του ώστε να αντιμετωπίσει τις ενδεχόμενες ασυνέχειες του πρώτου χάρτη, στις περιοχές γύρω από τα όρια των αντικειμένων. Οι δυο προβλέψεις συνενώνεται και ένα συνελικτικό επίπεδο εκπαιδεύεται προκειμένου να επιτελεί τη βέλτιστη σταθμισμένη σύντηξη των εκτιμώμενων χαρτών.

Παρόμοια προσέγγιση ακολουθούν και οι Tang & Wu (Tang & Wu, 2016). Το μοντέλο που προτείνουν οι συγγραφείς, τροποποιεί το δίκτυο VGG-16 (Simonyan & Zisserman, 2014), ούτως ώστε να είναι σε θέση να ανακτά χαρακτηριστικά σε πολλαπλές κλίμακες, και από τη σύντηξη τους να υπολογίζει σε επίπεδο εικονοστοιχείου, έναν αρχικό χάρτη για τα αντικείμενα της σκηνής που ξεχωρίζουν. Ταυτόχρονα, το δίκτυο των Zeiler & Fergus (Zeiler & Fergus, 2014) -ZFNet- χρησιμοποιείται για να εκτιμήσει σε επίπεδο υπερ-εικονοστοιχείων ένα δεύτερο χάρτη οπτικού ενδιαφέροντος. Οι δυο παραγόμενες εκτιμήσεις μαζί με την αρχική εικόνα

λαμβάνονται από ένα ρηχό συνελικτικό δίκτυο, το οποίο υπολογίζει τον τελικό χάρτη οπτικού ενδιαφέροντος. Στην προσέγγιση αυτή είναι αξιοσημείωτο το γεγονός, ότι η εκπαίδευση όλου του μοντέλου λαμβάνει χώρα ταυτόχρονα.

Τα πλεονεκτήματα των ανατροφοδοτούμενων συνελικτικών δίκτυων και της βαθιάς επιβλεπόμενης μάθησης συνδιάζουν στην προσέγγισή τους ο Tang και οι συνεργάτες του (Tang, Wu, & Bu, 2016). Το μοντέλο που προτείνεται από τους συγγραφείς, τροποποιεί την αρχιτεκτονική του δικτύου VGG-16 (Simonyan & Zisserman, 2014), αντικαθιστώντας τα συνελικτικά του επίπεδα με ανατροφοδοτούμενα συνελικτικά επίπεδα και την επεκτείνει προσθέτοντας πλευρικά συνελικτικά επίπεδα, προκειμένου να επαυξήσει την ικανότητα του να μαθαίνει με τρόπο αποδοτικό και ανθεκτικό υψηλού επιπέδου χαρακτηριστικά, τόσο τοπικά όσο και για το σύνολο της εκάστοτε εικόνας. Τελικά, ο χάρτης οπτικού ενδιαφέροντος παράγεται από τη σύντηξη των αναπαραστάσεων όλων των πλευρικών επιπέδων.

Ένα μοντέλο αποτελούμενο από δυο συνελικτικά υποδίκτυα παρουσιάστηκε από τους Liu & Han (N. Liu & Han, 2016) . Οι συγγραφείς της εργασίας, χρησιμοποιούν το δίκτυο VGG-16 (Simonyan & Zisserman, 2014), για να εκτιμήσουν σε γενικές γραμμές ένα χάρτη οπτικού ενδιαφέροντος και ακολούθως ένα ιεραρχικό δίκτυο συντιθέμενο από συνελικτικά επίπεδα που λειτουργούν με ανατροφοδότηση αναλαμβάνει να ραφινάρει σε ποικίλες κλίμακες την αρχική πρόβλεψη.

Ο Kuen και οι συνεργάτες του προτείνουν ένα μοντέλο για τον υπολογισμό χαρτών εξεχόντων αντικειμένων σε δυο διακριτά στάδια (Kuen, Wang, & Wang, 2016). Αρχικά, ένα δίκτυο του τύπου αυτό-κωδικοποιητή βασισμένο στο δίκτυο VGG-16 (Simonyan & Zisserman, 2014), χρησιμοποιείται για να παραχθεί μια αρχική εκτίμηση του χάρτη οπτικού ενδιαφέροντος. Σε μεταγενέστερο στάδιο ένα δεύτερο δίκτυο, επίσης της μορφής αυτό-κωδικοποιητή το οποίο συντίθεται από ανατροφοδοτούμενα συνελικτικά επίπεδα χρησιμοποιείται για να προσδώσει μεγαλύτερη ακρίβεια στην αρχική πρόβλεψη, βελτιώνοντας σταδιακά την κάθε περιοχή της.

Μια ενδιαφέρουσα μεθοδολογία είναι αυτή του Kruthiventi και των συνεργατών του (Kruthiventi, Gudisa, Dholakiya, & Venkatesh Babu, 2016). Οι συγγραφείς, δανείζονται τα πέντε πρώτα συνελικτικά μπλοκ του δικτύου VGG-16 (Simonyan & Zisserman, 2014), τα οποία επεκτείνουν ένα επιπλέον συνελικτικό μπλοκ. Δομικά στοιχεία, που συντίθενται από πολλαπλά συνδεδεμένα συνελικτικά επίπεδα (inception modules) - παρόμοια των δομικών στοιχείων που χρησιμοποιεί το GoogleNet (Szegedy et al., 2015) - εισάγονται πλευρικά του

πρώτου και των τριών τελευταίων μπλοκ για να εξάγουν σημασιολογική πληροφορία σε πολλαπλά επίπεδα. Οι πληροφορία που ανακτάται από τα δομικά στοιχεία συνενώνεται και διέρχεται από ένα συνελικτικό δίκτυο μόλις δυο επιπέδων, προκειμένου να υπολογιστεί ο τελικός χάρτης οπτικού ενδιαφέροντος.

Σε μια εναλλακτική προσέγγιση, ο Hou και οι συνεργάτες του (Hou et al., 2017), προτείνουν ένα συνελικτικό μοντέλο, το οποίο λαμβάνει υπόψη χαρακτηριστικά προερχόμενα από πολλαπλά επίπεδα και κλίμακες. Ειδικότερα, η προτεινόμενη μέθοδος συνδυάζει την έξοδο των συνελικτικών μπλοκ του δικτύου VGG-16 (Simonyan & Zisserman, 2014), ώστε να παραγάγει χάρτες οπτικού ενδιαφέροντος. Σε αντίθεση με τη συμβατική σύντηξη στο καταληκτικό συνελικτικό επίπεδο του δικτύου, προσέγγιση που κατά κανόνα ακολουθούν οι περισσότερες μεθοδολογίες, η παρούσα μεθοδολογία εισάγει τη χρήση σύντομων συνδέσεων μεταξύ κάθε ζεύγους των ενδιάμεσων αναπαραστάσεων του δικτύου. Η πρωτοτυπία αυτή, επιτρέπει στην υψηλού επιπέδου σημασιολογική πληροφορία από τα ανώτερα επίπεδα του δικτύου, να εμπλουτίσει τη χαμηλού επιπέδου -πλούσια ωστόσο σε λεπτομέρειες- πληροφορία που έχει εξαχθεί, η οποία ταυτόχρονα δρα και ως επιδιορθωτικός μηχανισμός στη τελική πρόβλεψη.

Σε αντίθεση με τις προαναφερθείσες εργασίες, οι οποίες παραλλάσουν την αρχιτεκτονική που υποδεικνύει το δίκτυο VGG-16, ο Zhang και οι συνεργάτες του (J. Zhang, Dai, & Porikli, 2017), υιοθετούν την αρχιτεκτονική δικτύου ResNet-101 (K. He et al., 2016), για να εκτιμήσουν με ακρίβεια εικονοστοιχείου χάρτες για τα σημαντικά αντικείμενα μιας σκηνής. Επιπλέον, κατά τη διαδικασία του ελέγχου, ένας επιδιορθωτικός μηχανισμός που συνδυάζει πληροφορία σε επίπεδο υπερ-εικονοστοιχείων για τρεις διαφορετικές κλίμακες, εφαρμόζεται για την περαιτέρω βελτίωση των αποτελεσμάτων.

3.4. Ανίχνευση οπτικά σημαντικών αντικειμένων υποβοηθούμενη από την πληροφορία του βάθους

3.4.1. Η πληροφορία του βάθους & η ανίχνευση οπτικά σημαντικών αντικειμένων

Παρόλο που φαίνεται να έχει σημειωθεί σημαντική πρόοδος στον τομέα της ανίχνευσης οπτικά σημαντικών αντικειμένων σε RGB εικόνες (RGB-based SOD), οι εγγενείς αδυναμίες

αυτού του τύπου εικόνων, δυσχεραίνουν τη περαιτέρω βελτίωση της απόδοσης των προαναφερθέντων μοντέλων όταν αυτά έρχονται αντιμέτωπα με απαιτητικά σενάρια εικόνων.

Σε αυτό το σημείο αξίζει να επισημάνουμε ότι ακόμη και οι προσεγγίσεις που θεμελιώνονται σε πλήρη συνελικτικά δίκτυα (FCN – based approaches) και οι οποίες αποτελούν την καλύτερη δυνατή προσέγγιση στο πεδίο της ανίχνευσης αντικειμένων υψηλού ενδιαφέροντος σε RGB εικόνες, τουλάχιστον μέχρι τη παρούσα χρονική στιγμή, συχνά αποτυγχάνουν ή εμφανίζουν ανεπαρκή απόδοση όταν έρχονται αντιμέτωπες με ιδιαίτερα δύσκολες οπτικές αναπαραστάσεις (Borji et al., 2019).

Έχει παρατηρηθεί ότι σε φυσικές σκηνές τα αντικείμενα που θεωρούνται σημαντικά και βρίσκονται στο προσκήνιο και συνεπώς πιο κοντά στο παρατηρητή της σκηνής, έχουν την τάση να μοιράζονται συχνά παρόμοια εμφάνιση και χαρακτηριστικά με αντικείμενα που βρίσκονται στο παρασκήνιο. Κι ενώ οι άνθρωποι είναι σε θέση να διακρίνουν εύκολα τα σημαντικά αντικείμενα του προσκηνίου, από τα αντικείμενα που παρασκηνίου λόγω των διαφορετικών γεωμετρικών και δομικών χαρακτηριστικών τους όπως η απόσταση τους, το μέγεθος τους και η διαφορετική γωνία παρατήρησης τους, τα υπολογιστικά συστήματα δεν είναι σε θέση να πραγματοποιήσουν αυτή τη διάκριση χωρίς να τους παρέχεται κάποια επιπλέον πληροφορία. Είναι εύκολα αντιληπτό, ότι η διάκριση καθίσταται ακόμη πιο δύσκολη όταν η σκηνή στο προσκήνιο είναι ιδιαίτερα περίπλοκη, όταν η εικόνα παρουσιάζει διακυμάνσεις στη φωτεινότητα της, όταν οι συνθήκες φωτισμού υπό τις οποίες έχει ληφθεί η εικόνα είναι ανεπαρκείς ή τα αντικείμενα της είναι διαφανή (Borji et al., 2019).

Προκειμένου να ανταπεξέλθουν τα RGB μοντέλα ανίχνευσης σημαντικών αντικειμένων σε απαιτητικά σενάρια εικόνων, όπως αυτά που περιγράφηκαν παραπάνω, αναπτύχθηκαν νέες μεθοδολογίες οι οποίες λαμβάνουν υπόψη την έννοια του βάθους σε μια απεικονιζόμενη σκηνή.

Η πρώτη εργασία η οποία αξιοποιούσε την πληροφορία του βάθους για να παραγάγει βελτιωμένα αποτελέσματα αναφορικά με την ανίχνευση περιοχών υψηλής σημαντικότητας σε εικόνες, παρουσιάστηκε μόλις το 2012 από τον Lang και τους συνεργάτες του (Lang et al., 2012). Η προαναφερόμενη μεθοδολογία χρησιμοποιούσε Μεικτά Γκαουσιανά Μοντέλα (Gaussian Mixture Models), για να αποσαφηνίσει τη σχέση μεταξύ των σημαντικών περιοχών μιας εικόνας και του βάθους της τελευταίας. Η αποτίμηση της σχέσης αυτής πραγματοποιούταν μέσω της εκτίμησης της από κοινού συνάρτησης πυκνότητας. Η διαπίστωση στη οποία κατέληξαν οι συγγραφείς, ότι η πληροφορία του βάθους μπορεί να

συνεισφέρει ουσιαστικά στη ενίσχυση των αποτελεσμάτων της ανίχνευσης των σημαντικών αντικειμένων σε μια εικόνα, έγινε η απαρχή για την ανάπτυξη πολυάριθμων και καινοτόμων ερευνητικών μεθοδολογιών σ' αυτό το πεδίο.

Θα μπορούσε κανείς να ισχυριστεί ότι η ανίχνευση σημαντικών αντικείμενων σε μια σκηνή υποβοηθούμενη από την πληροφορία του βάθους (RGB-D Salient Object Detection) παρουσιάζει μεγαλύτερη συνέπεια ως προς τη βιολογική λειτουργία του ανθρώπινου οπτικού συστήματος, αφού το τελευταίο στην θέαση μιας σκηνής, προκειμένου να προσδιορίσει τη θέση των αντικειμένων της και να πραγματοποιήσει τη διάκριση τους, δε περιορίζεται απλώς στο σχήμα τους, αλλά εκμεταλλεύεται και την πληροφορία του βάθους που χαρακτηρίζει τη σκηνή. Η αντίληψη του βάθους (στερεοσκοπική όραση) είναι αυτή που επιτρέπει στο ανθρώπινο οπτικό σύστημα, την τρισδιάστατη κατανόηση της εκάστοτε προβαλλόμενης οπτικής αναπαράστασης (Cong et al., 2018).

3.4.2. Χάρτες βάθους

Στην υπολογιστική όραση, η πληροφορία του βάθους για την ανίχνευση των σημαντικών αντικείμενων μιας σκηνής, είναι αξιοποιήσιμη μέσω των εικόνων βάθους. Οι τιμές της φωτεινότητας των εικονοστοιχείων σε μια εικόνα βάθους ή αλλιώς χάρτη βάθους (Depth Map), αναπαριστούν την απόσταση αυτού του σημείου από τη θέση παρατήρησης της οπτικής σκηνής.

Ένας χάρτης βάθους παρέχει επιπρόσθετη χωρική πληροφορία, η οποία συμπληρώνει τη χρωματική πληροφορία που φέρει μια RGB εικόνα (Zhou et al., 2020).

Πράγματι, οι χάρτες βάθους δύνανται να αποκαλύψουν χρήσιμες ιδιότητες των αντικειμένων που βρίσκονται στο προσκήνιο, όπως το σχήμα ή το περίγραμμα τους, επιτρέποντας με αυτό τον τρόπο τη διάκριση των τελευταίων από τα αντικείμενα του παρασκήνιου μιας σκηνής όσο πολύπλοκη κι αν είναι η αναπαράσταση του παρασκήνιου (Cong et al., 2018).

Πλέον, εξαιτίας της μεγάλης διαθεσιμότητας αισθητήρων βάθους αλλά και της ανάπτυξης των απεικονιστικών συσκευών, είναι σχετικά εύκολη η απόκτηση της πληροφορίας του βάθους για μια οπτική σκηνή. Ποικίλες τεχνικές που προσπαθούν να προσομοιάσουν τη διοπτρική ανθρώπινη όραση έχουν αναπτυχθεί για αποτύπωση του βάθους μιας σκηνής. Αναλυτικότερα, οι τεχνικές προβολής δομημένου φωτός (Structured Light Pattern) που υιοθετεί και η Lytro Light Field κάμερα, ανακτούν τη πληροφορία του βάθους από την ποσοτικοποίηση της

παραμόρφωσης που εμφανίζει ένα προκαθορισμένο μοτίβο - πρότυπο όταν αυτό προβληθεί πάνω σε μια σκηνή. Από την άλλη πλευρά στα συστήματα Time-Of-Flight τα οποία ενσωματώνουν και οι πιο πρόσφατες εκδόσεις του συστήματος Kinect, η αποτύπωση του βάθους πραγματοποιείται από τον υπολογισμό του χρόνου ανάκλασης μιας δέσμης φωτός πίσω στον αισθητήρα που φέρουν, από τη στιγμή που αυτή θα προσπέσει σε κάποιο αντικείμενο. Τέλος, οι στερεοσκοπικές κάμερες συλλαμβάνουν μια αναπαράσταση από δύο διαφορετικές οπτικές γωνίες και πραγματοποιούν έμμεση εκτίμηση του χάρτη βάθους, από το χάρτη της στερεοσκοπικής μετατόπισης (Disparity Map), αφού πρώτα ληφθούν υπόψη οι παράμετροι της στερεοσκοπικής εγκατάστασης. Ο χάρτης της στερεοσκοπικής μετατόπισης, περιέχει μια ποσοτικοποίηση της διαφοράς εικονοστοιχείο προς εικονοστοιχείο των δυο διαφορετικών στιγμιότυπων που ελήφθησαν από την κάμερα (Cong et al., 2018).

3.4.3. Μεθοδολογίες ανίχνευσης οπτικά σημαντικών αντικειμένων υποβοηθούμενες από την πληροφορία του βάθους

Η πληροφορία που φέρουν οι χάρτες βάθους, είναι αξιοποιήσιμη από τις μεθοδολογίες ανίχνευσης σημαντικών αντικειμένων είτε άμεσα (Depth Feature-Based Methods), είτε έμμεσα (Depth – Measure Based Methods). Στην πρώτη περίπτωση, η πληροφορία του βάθους χρησιμοποιείται απευθείας και επικουρικά για να ενισχύσει και να συμπληρώσει τη χρωματική πληροφορία, ενώ στη δεύτερη περίπτωση χρησιμοποιούνται ειδικά σχεδιασμένες μετρικές, οι οποίες εξάγουν από τους χάρτες βάθους με όσο το δυνατόν πιο αντιπροσωπευτικό τρόπο, τα χαρακτηριστικά που ενυπάρχουν στην πληροφορία που αυτοί φέρουν όπως το σχήμα ή η δομή των αντικειμένων της απεικονιζόμενης σκηνής (Ullah et al., 2020).

3.4.3.1. Έμμεση αξιοποίηση της πληροφορίας του βάθους

Μεταξύ των διάφορων Depth Measure - Based μεθοδολογιών που έχουν προταθεί, πιο χαρακτηριστικές είναι αυτές των (Ju, Ge, Geng, Ren, & Wu, 2014) και (Feng, Barnes, You, & McCarthy, 2016).

Οι συγγραφείς της πρώτης εργασίας (Ju et al., 2014), αφορμόμενοι από τις παρατηρήσεις ότι συνήθως οι περιοχές μιας εικόνας που βρίσκονται εγγύτερα στον παρατηρητή της, περιέχουν την περισσότερο σημαντική πληροφορία, καθώς επίσης και ότι τα σημαντικά αντικείμενα τείνουν να βρίσκονται στο κέντρο μιας εικόνας, προτείνουν τη χρήση του μέτρου της ανισοτροπικής κεντρο-περιφερειακής διαφοράς (Anisotropic Center-Surround Difference, ACSD) για την εύρεση σημαντικών περιοχών. Στη μεθοδολογία αυτή, η σημαντικότητα ενός

σημείου εκτιμάται λαμβάνοντας υπόψη το βαθμό στον οποίο αυτό διαφέρει από τα υπόλοιπα συνυπολογίζοντας τη συνολική δομική πληροφορία του χάρτη βάθους.

Δεδομένου ότι αρκετά συχνά οι τιμές των περιοχών που βρίσκονται στο παρασκήνιο ενός χάρτη βάθους παρουσιάζουν μεγάλη διακύμανση, οι περιοχές του παρασκηνίου καταλήγουν πολλές φορές να θεωρούνται εσφαλμένα σημαντικές, ιδίως όταν εμφανίζουν μεγάλη αντίθεση συγκριτικά με τις περιοχές του προσκηνίου. Για την αποφυγή αυτού του φαινομένου ο Feng και οι συνεργάτες του (Feng et al., 2016), σχεδίασαν ένα μέτρο (Local Background Enclosure, LBE) το οποίο βασίζεται την αντίθεση μεταξύ προσκηνίου και παρασκηνίου, που παρουσιάζει ο χάρτης βάθους και συνεκτιμά το ποσοστό του περιγράμματος των αντικειμένων που δε βρίσκονται στο προσκήνιο.

3.4.3.2. Άμεση αξιοποίηση της πληροφορίας του βάθους

Όσον αφορά τις Depth Feature – Based μεθοδολογίες, εκεί συναντάμε τα λεγόμενα κλασσικά μοντέλα, τα οποία στηρίζονται σε "χειροποίητα" (handcrafted) χαρακτηριστικά που εξάγονται τόσο από τις χρωματικές εικόνες όσο και από τους χάρτες βάθους, καθώς και πολυάριθμες έξυπνες προσεγγίσεις οι οποίες υιοθετούν μεθόδους Βαθιάς Μηχανικής Μάθησης με σκοπό την εκμάθηση χαρακτηριστικών από τις RGB-D εικόνες, για τον εντοπισμό των σημαντικών αντικειμένων σε μια σκηνή. Τα Συνελικτικά Νευρωνικά Δίκτυα φαίνεται ότι αποτελούν την πιο διαδεδομένη βιβλιογραφικά μέθοδο μηχανικής μάθησης που χρησιμοποιείται για το σκοπό αυτό. Ωστόσο, πρόσφατα έχουν παρουσιαστεί ενδιαφέρουσες εργασίες οι οποίες αξιοποιούν Παραγωγικά Ανταγωνιστικά Δίκτυα (Generative Adversarial Networks - GANs), (Z. Liu, Zhang, & Zhao, 2020), (B. Jiang, Zhou, Wang, Tang, & Luo, 2020), (Z. Liu, Tang, Xiang, & Zhao, 2020) ή ακόμη και Conditional Variational Autoencoders (CVAEs) (J. Zhang et al., 2020).

Αναλυτικότερα, ο Niu και οι συνεργάτες του (Niu et al., 2012), προσαρμόζουν στους χάρτες της στερεοσκοπικής μετατόπισης - που φέρουν την πληροφορία βάθους μιας σκηνής - τη μέθοδο των (M. Cheng et al., 2011), η οποία χρησιμοποιεί την έννοια της συνολικής αντίθεσης για να ανιχνεύσει σημαντικές περιοχές σε RGB εικόνες. Από την εφαρμογή αυτή υπολογίζεται ένας αρχικός χάρτης οπτικού ενδιαφέροντος. Ένας δεύτερος χάρτης προκύπτει και πάλι από τους χάρτες της στερεοσκοπικής μετατόπισης, μέσω της μοντελοποίησης εμπειρικών και εξειδικευμένων παρατηρήσεων για τα σημαντικά αντικείμενα μιας σκηνής και τους χάρτες της στερεοσκοπικής μετατόπισης. Η τελική εκτίμηση προκύπτει από το πολλαπλασιασμό των δυο παραγόμενων εκτιμήσεων.

Ο Fang και οι συνεργάτες του (Y. Fang, Wang, Narwaria, Le Callet, & Lin, 2014), εκμεταλλεύονται τον διακριτό μετασχηματισμό συνημίτονου για να εξάγουν από τις RGB εικόνες χαρακτηριστικά χρώματος, φωτεινότητας και υφής καθώς και χαρακτηριστικά βάθους από τις ομώνυμες εικόνες. Ακολούθως, ένα Γκαουσιανό μοντέλο παράγει ένα χάρτη αντίθεσης για καθένα από τα τέσσερα εξαγόμενα χαρακτηριστικά. Ο τελικός χάρτης σημαντικότητας, υπολογίζεται με τη βοήθεια ενός πρωτότυπου αλγορίθμου, ο οποίος συμπτύσσει τους χάρτες αντίθεσης με σταθμισμένο τρόπο, δίνοντας κάθε φορά μεγαλύτερο βάρος στους χάρτες εκείνους που εμφανίζουν τη μικρότερη χωρική διακύμανση τιμών, δηλαδή χαρακτηρίζονται από μεγαλύτερη συνεκτικότητα. Για τη περαιτέρω βελτίωση των αποτελεσμάτων, αναπτύσσονται τεχνικές εμπνευσμένες από λειτουργίες του ανθρώπινου οπτικού συστήματος όπως η οπτική οξύτητα και η τάση συγκέντρωσης της οπτικής προσοχής στην κεντρική περιοχή μιας εικόνας.

Λίγο αργότερα, ο Peng και οι συνεργάτες του (Peng, Li, Xiong, Hu, & Ji, 2014) ανέπτυξαν μια μεθοδολογία για την εκτίμηση χαρτών οπτικού ενδιαφέροντος προερχόμενους από τις εικόνες βάθους, η οποία μπορεί να λειτουργήσει συνδυαστικά με οποιοδήποτε RGB μοντέλο ανίχνευσης σημαντικών αντικειμένων. Συγκεκριμένα, συστήνουν έναν πρωτότυπο αλγόριθμο ο οποίος χρησιμοποιεί έναν Γκαουσιανό εκτιμητή πυρήνα και πραγματοποιεί εξαγωγή χαρακτηριστικών χαμηλού επιπέδου, τα οποία κωδικοποιούν την αντίθεση του χάρτη βάθους σε τρία διαφορετικά επίπεδα – τοπικά, συνολικά και μεταξύ προσκηνίου και παρασκηνίου. Επιπλέον, επεκτείνουν την παραπάνω μεθοδολογία τροποποιώντας την ώστε να μπορεί να παραγάγει χάρτες σημαντικών χαρακτηριστικών χαμηλού επιπέδου από RGB-D εικόνες. Ένας αλγόριθμος που παραλλάσσει αυτόν του Prim (Prim, 1957), χρησιμοποιείται με στόχο να ομαδοποιήσει τους χάρτες που παρήχθησαν και να δημιουργήσει χάρτες σημαντικών χαρακτηριστικών μεσαίου επιπέδου για τις RGB-D εικόνες. Ένα Μπεϋζιανό πλαίσιο συνδυάζει τους χάρτες χαρακτηριστικών χαμηλού και μεσαίου επιπέδου για να παράγει τον τελικό χάρτη σημαντικότητας ο οποίος υποβάλλεται σε έναν επιδιορθωτικό μηχανισμό για την περαιτέρω βελτίωση των αποτελεσμάτων. Ο μηχανισμός αυτός αξιοποιεί πληροφορία υψηλού επιπέδου η οποία ανακτάται από τις RGB-D εικόνες με τη βοήθεια ενός Γκαουσιανού μοντέλου.

Σε μια άλλη εργασία, ο Song και οι συνεργάτες του (Song et al., 2017) βασιζόμενοι στο γεγονός ότι η κλίμακα των προς ανίχνευση σημαντικών αντικειμένων είναι άγνωστη, προτείνουν τη σύντηξη χαρακτηριστικών τα οποία έχουν εξαχθεί από πολλαπλές κλίμακες τόσο από τις χρωματικές όσο και από τις εικόνες βάθους. Αυτό επιτυγχάνεται ως εξής: αρχικά,

εφαρμόζεται και στα δυο είδη εικόνων μια μέθοδος κατάτμησης σε πολλαπλές κλίμακες. Στη συνέχεια, πραγματοποιείται εξαγωγή χαρακτηριστικών χαμηλού, μεσαίου και υψηλού επιπέδου για όλες τις κλίμακες. Στα χαμηλού επιπέδου χαρακτηριστικά περιλαμβάνονται χαρακτηριστικά χρώματος και βάθους από τις αντίστοιχες εικόνες, ενώ και από τα δύο είδη εικόνων εξάγονται χαρακτηριστικά υφής καθώς και χαρακτηριστικά που βασίζονται στη γεωδαισιακή απόσταση. Ένας σταθμισμένος συνδυασμός των χαμηλού επιπέδου χαρακτηριστικών χρώματος και βάθους συνιστά τα χαρακτηριστικά του μεσαίου επιπέδου. Η εξαγωγή χαρακτηριστικών υψηλού επιπέδου είναι προσανατολισμένη στις περιοχές που βρίσκονται κοντά στο κέντρο της εικόνας και στις τμηματοποιημένες περιοχές όπως αυτές διαμορφώθηκαν νωρίτερα. Προκειμένου να αποφευχθεί η γραμμική σύντηξη των εξαγόμενων χαρτών των παραπάνω χαρακτηριστικών για τις διάφορες κλίμακες, χρησιμοποιείται ένας αλγόριθμος παλινδρόμησης τυχαίων δασών για να παραχθεί ο τελικός χάρτης. Επιπροσθέτως, στην προκειμένη εργασία προτείνεται και μια μεθοδολογία κατάτμησης σημαντικών αντικειμένων αξιοποιώντας τους εκτιμώμενους χάρτες σημαντικότητας, τις διανυσματικές μηχανές υποστήριξης και τη στατιστική τεχνική της δειγματοθέτησης (bootstrap sampling). Τα πειραματικά αποτελέσματα της εργασίας αυτής, επιβεβαίωσαν για ακόμη μια φορά τον ισχυρισμό ότι η πληροφορία που εξάγεται από τους χάρτες βάθους, συνδράμει με τρόπο καθοριστικό στην ανίχνευση των σημαντικών αντικειμένων μιας σκηνής.

3.4.3.3. Μεθοδολογίες ανίχνευσης οπτικά σημαντικών αντικειμένων σε RGB-D εικόνες με τη χρήση Συνελικτικών Νευρωνικών Δικτύων

Παρόλο που διαφαίνεται από τις παραπάνω εργασίες, ότι η προσθήκη της πληροφορίας του βάθους επιδρά θετικά στην απόδοση των μοντέλων που υλοποιούν εργασίες SOD, το πως μπορεί να επιτευχθεί η αξιοποίηση της πληροφορίας του βάθους στο μέγιστο δυνατό βαθμό και ο συνδυασμός των δύο διαφορετικών ειδών πληροφορίας με βέλτιστο τρόπο, από τις handcrafted feature based προσεγγίσεις παραμένει ένα ζήτημα υπό διερεύνηση (J. Han, Chen, Liu, Yan, & Li, 2017) .

Σε μια προσπάθεια αντιμετώπισης των προαναφερθέντων προκλήσεων, προτείνονται συνεχώς στη βιβλιογραφία τα τελευταία έτη, μοντέλα Βαθιάς Μηχανικής Μάθησης (Deep Models) – τα οποία στην πλειοψηφία τους χρησιμοποιούν τα Συνελικτικά Νευρωνικά Δίκτυα.

Τα Συνελικτικά Νευρωνικά Δίκτυα εκπαιδεύονται για να συνδυάζουν τις δυο διαφορετικές αναπαραστάσεις μιας οπτικής σκηνής και να εξάγουν αυτόματα χαρακτηριστικά

υψηλού επιπέδου, αποκαλύπτοντας δυσδιάκριτες μεν, ουσιώδες δε συσχετίσεις μεταξύ των RGB εικόνων και των χαρτών βάθους (Zhou et al., 2020).

Η πρώτη μεθοδολογία η οποία καταφεύγει στη χρήση Βαθιάς Μηχανικής Μάθησης για να επεξεργαστεί την πληροφορία βάθους συνδυαστικά με αυτήν του χρώματος, παρουσιάστηκε μόλις το 2017 (Qu et al., 2017). Οι συγγραφείς της εν λόγω εργασίας, εκπαιδεύουν ένα σχετικά απλό - από άποψη αρχιτεκτονικής - συνελικτικό νευρωνικό δίκτυο, ώστε να μπορεί να αποφασίσει για κάθε περιοχή μιας οπτικής σκηνής αν αυτή είναι σημαντική, γνωρίζοντας μόνο χαρακτηριστικά χαμηλού επιπέδου για αυτή την περιοχή. Σύμφωνα με τη προτεινόμενη προσέγγιση, τα χαρακτηριστικά ανακτώνται από τις RGB-D εικόνες μέσω μιας παραδοσιακής (hand-crafted) μεθοδολογίας. Συγκεκριμένα, από μια περιοχή που ορίζεται από ένα υπέρ-εικονοστοιχείο, και διαμορφώνεται με τη βοήθεια του αλγορίθμου SLIC (Achanta et al., 2012), χαρακτηριστικά χαμηλού επιπέδου όπως η αντίθεση στις εικόνες χρώματος και βάθους - τοπικά και συνολικά - καθώς και ο βαθμός συνεκτικότητας χρώματος και παρασκηνίου, κωδικοποιούνται σε διανύσματα. Ακολούθως, ένα συνελικτικό δίκτυο λαμβάνει ως είσοδο τα διανύσματα χαρακτηριστικών ενός υπέρ-εικονοστοιχείου και μαθαίνει να προβλέπει αν η είσοδος του αντιστοιχεί σε σημαντική περιοχή, αντιμετωπίζοντας το εγχείρημα ως ένα πρόβλημα δυαδικής λογιστικής παλινδρόμησης. Αναφορικά με την αρχιτεκτονική του, αποτελείται από τρία συνελικτικά επίπεδα, ακολουθούμενα από ένα πλήρως συνδεδεμένο επίπεδο και ένα καταληκτικό επίπεδο λογιστικής παλινδρόμησης που χρησιμοποιείται για τη τελική πρόβλεψη. Επιπλέον, για τη περαιτέρω βελτίωση των αποτελεσμάτων ενσωματώνουν στο δίκτυο ένα μηχανισμό Λαπλασιανής Οπισδρόμησης. Η μεθοδολογία αυτή επαληθεύεται σε γνωστά RGB-D σύνολα δεδομένων (NLPR (Peng et al., 2014), NJUDS2000 (Ju et al., 2014), LFSD (N. Li, Ye, Ji, Ling, & Yu, 2014)). Τα αποτελέσματα καταδεικνύουν τόσο την υπεροχή της προσέγγισης αυτής συγκριτικά με τις συμβατικές προγενέστερες μεθοδολογίες, όσο και τη συνεισφορά του μηχανισμού της Λαπλασιανής Οπισδρόμησης ο οποίος δύναται να λειτουργήσει ως επιδιορθωτικός μηχανισμός του τελικού αποτελέσματος.

Λίγο, αργότερα ο Han και οι συνεργάτες του (J. Han et al., 2017) χρησιμοποίησαν ένα βαθύ μοντέλο το οποίο συντίθεται από δυο διαφορετικά συνελικτικά δίκτυα για να προβλέψουν με ακρίβεια εικονοστοιχείου τις τιμές του χάρτη οπτικής υπεροχής. Αναλυτικότερα, δυο διαφορετικά δίκτυα που ωστόσο έχουν την ίδια αρχιτεκτονική χρησιμοποιούνται για να παραγάγουν μια υψηλού επιπέδου αναπαράσταση για τα δυο διαφορετικά είδη εικόνων (χρώματος και βάθους). Η αρχιτεκτονική των δικτύων είναι εμπνευσμένη από αυτή που υποδεικνύει το δίκτυο VGG-16 (Simonyan & Zisserman, 2014),

καθώς περιλαμβάνει, τα πέντε πρώτα επίπεδα του (δηλαδή δεκατρία συνολικά συνελικτικά επίπεδα) και επιπλέον ένα πλήρως συνδεδεμένο επίπεδο. Για να προκύψει ο τελικός χάρτης οπτικού ενδιαφέροντος, ακολουθείται μια τεχνική συνδυασμού των δυο τροπικοτήτων στο τελικό στάδιο (late modality fusion), όπου οι δυο υψηλού επιπέδου αναπαραστάσεις που προκύπτουν από τα αντίστοιχα δίκτυα συνενώνονται και διέρχονται από ένα πλήρως συνδεδεμένο επίπεδο, που υλοποιεί την πρόβλεψη. Μια συνάρτηση απώλειας που λαμβάνει υπόψη τη συνολική δομική πληροφορία της εικόνας, χρησιμοποιείται κατά τη διάρκεια της εκπαίδευσης από τα δυο δίκτυα, καθώς και από το μηχανισμό σύντηξης. Ωστόσο, πραγματική πρωτοτυπία αυτής της μεθοδολογίας αποτελεί το γεγονός ότι οι εικόνες βάθους δε χρησιμοποιούνται αυτούσιες, αλλά κωδικοποιούνται σε μορφή HHA (Gupta, Girshick, Arbeláez, & Malik, 2014) πριν ληφθούν στην είσοδο του δικτύου που είναι υπεύθυνο για την επεξεργασία τους. Η αναπαράσταση αυτή, χρησιμοποιεί τρία κανάλια, ένα για την οριζόντια ανομοιότητα των εικονοστοιχείων, ένα για το ύψος κάθε εικονοστοιχείου από την επιφάνεια του εδάφους και ένα για τη γωνία του διανύσματος επιφάνειας θέσης που σχηματίζει το εκάστοτε εικονοστοιχείο με τοπική κατεύθυνση της βαρύτητας. Τέλος, ακόμη μια καινοτομία αυτού του μοντέλου, θεωρείται το γεγονός ότι η γνώση που εξάγεται από το δίκτυο που είναι υπεύθυνο για την επεξεργασία της χρωματικής πληροφορίας, μεταφέρεται στο δίκτυο που επεξεργάζεται την πληροφορία του βάθους μέσω της διαδικασίας της αρχικοποίησης των βαρών. Τα πειράματα που διεξάχθηκαν σε γνωστά RGB-D σύνολα δεδομένων (DES (Y. Cheng, Fu, Wei, Xiao, & Cao, 2014), NLPR (Peng et al., 2014), STEREO (Niu et al., 2012), NJUDS2000 (Ju et al., 2014)), επιβεβαιώνουν την παροχή συνεπών αποτελεσμάτων της μεθοδολογίας αυτής ως προς τη βάση αλήθεια, έναντι σύγχρονών της προσεγγίσεων.

Ένα συνελικτικό δίκτυο με διπλή ροή εισόδου που υποστηρίζει την ταυτόχρονη επεξεργασία RGB εικόνας και χάρτη βάθους παρουσίασαν οι Chen & Li (Chen & Li, 2018). Το προτεινόμενο δίκτυο χρησιμοποιεί και αυτό την αρχιτεκτονική του δικτύου VGG-16 (Simonyan & Zisserman, 2014) - κρατώντας απαράλλακτα τα πέντε πρώτα μπλοκ του - και την επαυξάνει προσθέτοντας ένα έκτο μπλοκ, αποτελούμενο από ένα μόνο συνελικτικό επίπεδο. Η καινοτομία που εισάγει η εργασία αυτή, είναι τα δομικά στοιχεία (Complementarity - Aware Fusion modules) που παρεμβάλλονται στο τέλος κάθε μπλοκ του δικτύου και τα οποία εκπαιδεύονται ώστε να πραγματοποιούν προοδευτικά σύντηξη των δυο διαφορετικών ειδών πληροφορίας για πέντε διαφορετικά επίπεδα, λαμβάνοντας κάθε φορά υπόψη τη συμπληρωματικότητα των δυο διαφορετικών τροπικοτήτων που επεξεργάζεται το μοντέλο. Όπως και στην προηγούμενη εργασία, η πληροφορία του βάθους κωδικοποιείται στην HHA

αναπαράσταση (Gupta et al., 2014), πριν χρησιμοποιηθεί από το προτεινόμενο μοντέλο. Από τα πειράματα που πραγματοποιήθηκαν σε RGB-D σύνολα δεδομένων (NLPR (Peng et al., 2014), NJUDS2000 (Ju et al., 2014), STEREO (Niu et al., 2012)) διαπιστώθηκε ότι ενώ το συγκεκριμένο μοντέλο παράγει αρκετά καλή ποιότητας χάρτες σημαντικότητας, αδυνατεί να διατηρήσει τις χαρακτηριστικές λεπτομέρειες των σημαντικών αντικειμένων.

Σε μια άλλη εργασία, ο Chen και οι συνεργάτες του (Chen, Li, & Su, 2019) συστήνουν ένα βαθύ μοντέλο με διπλή ροή εισόδου που υλοποιεί σύντηξη των δυο διαφορετικών ειδών πληροφορίας σε πολλαπλά επίπεδα. Αρχικά, εκπαιδεύεται μόνο η ροή που επεξεργάζεται τη πληροφορία του χρώματος και η οποία αποτελείται από δυο κλάδους - ένα για την εξαγωγή τοπικών και ένα για την εκμάθηση συνολικών χαρακτηριστικών. Κάθε κλάδος του δικτύου υπακούει στην αρχιτεκτονική των 13 συνελικτικών επιπέδων του δικτύου VGG-16 (Simonyan & Zisserman, 2014). Ακολούθως, η γνώση που αποκτήθηκε από αυτή τη διαδικασία, δηλαδή οι τιμές των βαρών, μεταφέρονται στη ροή που επεξεργάζεται τους χάρτες βάθους και η οποία έχει την ίδια δομή και αρχιτεκτονική με την πρώτη ροή. Μετά το πέρας της διαδικασίας εκπαίδευσης και της δεύτερης ροής, οι δυο κλάδοι που εξάγουν ομότιμα χαρακτηριστικά και από τις δυο τροπικότητες συνενώνονται μεταξύ τους, όπως επίσης και οι δυο τροπικότητες σε πολλαπλά επίπεδα τους και κατόπιν το δίκτυο επανεκπαιδεύεται, αθροίζοντας τις προβλέψεις των συνενωμένων πλέον τοπικών και συνολικών κλάδων για να υπολογιστεί ο τελικός χάρτης οπτικού ενδιαφέροντος. Η πληροφορία του φέρουν οι χάρτες βάθους αντιστοιχίζεται στην HHA αναπαράσταση (Gupta et al., 2014), για να αξιοποιηθεί από το μοντέλο αυτό. Η αποτελεσματικότητα της μεθοδολογίας αυτής εξετάζεται σε RGB-D σύνολα δεδομένων (NLPR (Peng et al., 2014), NJUDS2000 (Ju et al., 2014), STEREO (Niu et al., 2012)) και η απόδοση της φαίνεται ικανοποιητική ακόμη και όταν έρχεται αντιμέτωπη με ιδιαίτερα απαιτητικά σενάρια εικόνων.

Σε μια μεταγενέστερη εργασία οι Chen & Li (Chen & Li, 2019) προτείνουν ένα μοντέλο που συντίθεται από τρείς διαφορετικές ροές δεδομένων. Πέραν, των δυο ροών για την επεξεργασία της πληροφορίας του βάθους και του χρώματος που χρησιμοποιούν οι προαναφερθείσες προσεγγίσεις, συστήνουν μια επιπρόσθετη παράλληλη ροή για την σύγχρονη - συνεργατική επεξεργασία και των δυο τροπικοτήτων σε πολλαπλά επίπεδα. Οι δυο συνηθισμένες ροές, οι οποίες επεξεργάζονται ταυτόχρονα τις δυο τροπικότητες αξιοποιούν την αρχιτεκτονική του δικτύου VGG-16 (Simonyan & Zisserman, 2014), από την οποία δανείζονται τα πέντε πρώτα μπλοκ της και την επεκτείνουν προσθέτοντας ένα ακόμη συνελικτικό επίπεδο. Η τρίτη ροή, φέρει παρόμοια αρχιτεκτονική με τις προηγούμενες και

εκπαιδεύεται για να συνδυάσει με βέλτιστο τρόπο σε πολλαπλές κλίμακες τις εξόδους των άλλων δυο ροών όπως αυτές διαμορφώνονται έπειτα από κάθε συνελικτικό μπλοκ. Για το σκοπό αυτό, η λειτουργία της επιβλέπεται από εκπαιδευόμενα δομικά στοιχεία (Attention-Aware Cross-Modal Cross-Level modules). Η εκπαίδευση και των τριών ροών διεξάγεται ταυτόχρονα, ενώ η πληροφορία του βάθους εισάγεται στο δίκτυο αυτό αφού πρώτα κωδικοποιηθεί στη μορφή HHA (Gupta et al., 2014). Από την επαλήθευση της προτεινόμενης μεθόδου σε RGB-D σύνολα δεδομένων (NLPR (Peng et al., 2014), NJUDS2000 (Ju et al., 2014), STEREO (Niu et al., 2012)), παρατηρείται ότι όχι απλώς αυτή παράγει αποτελέσματα συνεπή ως προς τη βάση αλήθειας αλλά έχει και τη δυνατότητα να διατηρεί λεπτομέρειες των σημαντικών αντικειμένων.

Οι Wang & Gong (N. Wang & Gong, 2019) προτείνουν επίσης ένα συνελικτικό δίκτυο δυο ροών για την ταυτόχρονη επεξεργασία των εικόνων RGB και των χαρτών βάθους, καθώς και ένα εκπαιδευόμενο μηχανισμό που υλοποιεί με βέλτιστο τρόπο το συνταίριασμα της πληροφορίας των δυο τροπικοτήτων. Κάθε ροή δεδομένων χαρακτηρίζεται από αρχιτεκτονική του τύπου κωδικοποιητή – αποκωδικοποιητή. Ως κωδικοποιητής χρησιμοποιείται το δίκτυο VGG-16 (Simonyan & Zisserman, 2014), πλην των πλήρως συνδεδεμένων επιπέδων του (δηλαδή πέντε μπλοκ με δεκατρία συνολικά συνελικτικά επίπεδα), ενώ ο αποκωδικοποιητής συγκροτείται συνελικτικά επίπεδα και επίπεδα υπερδειγματοληψίας (Upsampling layers) οργανωμένα σε μπλοκ ισάριθμα του κωδικοποιητή. Για κάθε ροή δεδομένων, η γνώση που αποκτάται στον κωδικοποιητή μεταφέρεται στον αποκωδικοποιητή μέσω κατάλληλων συνδέσεων μεταξύ ομότιμων μπλοκ. Οι υψηλού επιπέδου αναπαραστάσεις που προκύπτουν από τις δυο ροές για τις δυο τροπικότητες συνενώνονται στο τελικό στάδιο (late modality fusion), και κατόπιν διέρχονται από ένα καταληκτικό συνελικτικό επίπεδο, για να διαμορφώσουν ένα χάρτη επιλογής (Switch Map). Ο χάρτης επιλογής υποβοηθάει στην εκτίμηση του τελικού χάρτη σημαντικότητας, καθώς παρέχει κατά περίπτωση τα κατάλληλα βάρη για τη σταθμισμένη σύντηξη των δυο υψηλού επιπέδου αναπαραστάσεων. Όλα τα δομικά στοιχεία του προτεινόμενου μοντέλου εκπαιδεύονται μαζί και μια ειδικά σχεδιασμένη συνάρτηση απώλειας που συνυπολογίζει τη συνεισφορά κάθε δομικού στοιχείου, χρησιμοποιείται κατά τη διαδικασία αυτή. Αν και από τα πειράματα που αξιολογούν την απόδοση του μοντέλου αυτού σε RGB-D σύνολα δεδομένων (NLPR (Peng et al., 2014), NJUDS2000 (Ju et al., 2014) , STEREO (Niu et al., 2012)) συνάγεται η υπεροχή της μεθόδου αυτής, διαπιστώνεται ότι το προτεινόμενο μοντέλο, όπως και τα συγκρίσιμα με αυτό, δε

δύνανται να παρέχουν ικανοποιητικά αποτελέσματα σε απαιτητικά σενάρια εικόνων, όπου τα σημαντικά αντικείμενα είναι δυσδιάκριτα στις RGB-D εικόνες.

Μια εναλλακτική προσέγγιση για τη σύντηξη χαρακτηριστικών από πολλαπλές κλίμακες συστήνεται από το Zhao και τους συνεργατές του (J.-X. Zhao et al., 2019). Οι συγγραφείς της εργασίας, εισάγουν μέσα στο μοντέλο τους δομικά στοιχεία ενίσχυσης χαρακτηριστικών (Feature-Enhanced Modules) καθώς και μια αρχιτεκτονική πυραμίδας για τη συγχώνευση των δυο διαφορετικών ειδών πληροφορίας με ιεραρχικό τρόπο. Το προτεινόμενο μοντέλο, φέρει ίδια δομή με τα πέντε πρώτα μπλοκ του δικτύου VGG-16 (Simonyan & Zisserman, 2014), ενώ στο πέρας κάθε μπλοκ συναντάται μια μονάδα ενίσχυσης χαρακτηριστικών. Το στοιχείο αυτό, ενσωματώνει την πληροφορία του βάθους μέσα στο μοντέλο, αφού πρώτα τη βελτιώσει, αξιοποιώντας την αντίθεση μεταξύ προσκηνίου και παρασκηνίου. Με τον τρόπο αυτό, κατασκευάζονται ενισχυμένα χαρακτηριστικά που ενέχουν πληροφορία και από τις δυο τροπικότητες. Τα χαρακτηριστικά αυτά, που αποκτώνται από τις πέντε διαφορετικές κλίμακες του μοντέλου, συνδυάζονται ιεραρχικά σε μια αρχιτεκτονική πυραμίδας για να παραχθεί ο τελικός χάρτης οπτικού ενδιαφέροντος. Τα πειράματα που διεξάγονται σε RGB-D σύνολα δεδομένων (NJU2000 (Ju et al., 2014), NLPR (Peng et al., 2014), STEREO (Niu et al., 2012), LFSD (N. Li et al., 2014), DES (Y. Cheng et al., 2014)), φανερώνουν την αποτελεσματικότητα της προτεινόμενης προσέγγισης ακόμη και όταν αυτή αφορά απαιτητικά σενάρια εικόνων (περίπλοκες σκηνές, χαμηλή αντίθεση, μικρά ή πολλαπλά αντικείμενα).

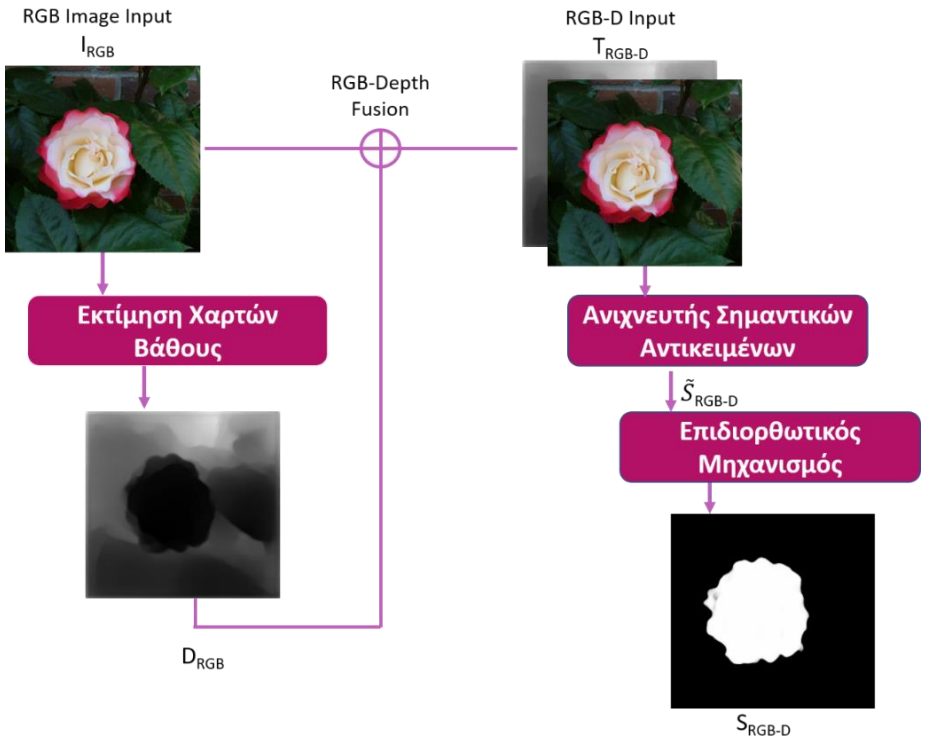
Μια πιο σύγχρονη εργασία για την ανίχνευση σημαντικών αντικειμένων σε RGB-D εικόνες είναι αυτή των Fan, Lin και των συνεργατών τους (Fan, Lin, Zhang, Zhu, & Cheng, 2020). Το προτεινόμενο μοντέλο (*D^3Net*), χρησιμοποιεί μια τριπλή ροή για εκμάθηση χαρακτηριστικών καθώς και μια μονάδα εντοπισμού σφαλμάτων βάθους (Depth Depurator Unit). Όλες οι ροές δεδομένων, μοιράζονται την ίδια αρχιτεκτονική αλλά επεξεργάζονται διαφορετικού είδους πληροφορία. Η πρώτη ροή επεξεργάζεται τις εικόνες χρώματος, η δεύτερη τους χάρτες βάθους και η τρίτη δέχεται στην είσοδο της μια RGB-D εικόνα. Κάθε μια από τις τρεις ροές δεδομένων χαρακτηρίζεται από αρχιτεκτονική του τύπου κωδικοποιητή – αποκωδικοποιητή. Ο κωδικοποιητής επεκτείνει την αρχιτεκτονική του δικτύου VGG-16 (Simonyan & Zisserman, 2014), από την οποία δανείζεται μόνο τα πέντε πρώτα μπλοκ του και της προσθέτει ένα επιπλέον μπλοκ αποτελούμενο από ένα μόνο συνελικτικό επίπεδο. Ο αποκωδικοποιητής συγκροτείται από επίπεδα υπερδειγματοληψίας (Upsampling layers) και συνελικτικά επίπεδα οργανωμένα σε ισάριθμα μπλοκ. Για κάθε ροή δεδομένων, η γνώση που αποκτάται στον κωδικοποιητή μεταφέρεται στον αποκωδικοποιητή μέσω κατάλληλων

συνδέσεων. Οι τρεις ροές δεδομένων εκπαιδεύονται ανεξάρτητα η μια από την άλλη και κάθε μια εκτιμά ένα χάρτη σημαντικότητας. Πρωτοτυπία αυτής της προσέγγισης, αποτελεί η μονάδα εντοπισμού σφαλμάτων βάθους, η οποία κατά τη διάρκεια της διαδικασίας του ελέγχου, απομακρύνει τους χάρτες οπτικά σημαντικών αντικειμένων που υπολογίστηκαν από τη ροή η οποία χειρίζεται αποκλειστικά την πληροφορία του βάθους, όταν αυτοί είναι χαμηλής ποιότητας. Επιπλέον, αποφασίζει ποια από τις εναπομείνασες ροές δεδομένων έχει προβλέψει με μεγαλύτερη συνέπεια ως προς τη βάση αλήθειας, το χάρτη οπτικού ενδιαφέροντος. Οι συγγραφείς επαληθεύουν την εν λόγω μεθοδολογία σε αρκετά RGB-D σύνολα δεδομένων (NJU2K (Ju et al., 2014), NLPR (Peng et al., 2014), STEREO (Niu et al., 2012), DES (Y. Cheng et al., 2014), SSD (C. Zhu & Li, 2017), LFSD (N. Li et al., 2014), SIP (Fan, Lin, et al., 2020)). Τα αποτελέσματα των πειραμάτων τους κατατάσσουν το μοντέλο αυτό ως καλύτερο έναντι όλων όσων παρουσιάστηκαν νωρίτερα.

4. Μεθοδολογία

Η προτεινόμενη μεθοδολογία αναπτύχθηκε προκειμένου να αντιμετωπίσει το ζήτημα της ανίχνευσης σημαντικών αντικειμένων σε RGB εικόνες όταν αυτή διεξάγεται με την υποβοήθηση της πληροφορία του βάθους, που χαρακτηρίζει τη δοθείσα οπτική αναπαράσταση. Η πρωτοτυπία της εν λόγω μεθοδολογίας, έγκειται στο γεγονός ότι αξιοποιεί αρχιτεκτονικές συνελικτικών νευρωνικών δικτύων με απώτερο σκοπό την παραγωγή χαρτών βάθους για την εκάστοτε απεικονιζόμενη οπτική σκηνή. Η προσέγγιση αυτή έρχεται σε αντίθεση με τις υπάρχουσες μεθοδολογίες, οι οποίες στα πλαίσια της ανίχνευσης σημαντικών αντικειμένων σε εικόνες, αξιοποιούν πληροφορία βάθους προερχόμενη από κάμερες ή αισθητήρες. Εν συνεχεία, παρουσιάζεται αναλυτικά η μεθοδολογία που προτείνεται, η οποία πέραν της πρωτοτυπίας που εισάγει ακολουθεί τη λογική των προσεγγίσεων που χρησιμοποιούν εξ ολοκλήρου συνελικτικά δίκτυα (Fully Convolutional Network approaches), για να εντοπίσουν τα οπτικά σημαντικά αντικείμενα σε μια οπτική αναπαράσταση.

4.1. Αρχιτεκτονική του προτεινόμενου μοντέλου



Εικόνα 4.1 Συνολική αρχιτεκτονική της προτεινόμενης μεθοδολογίας

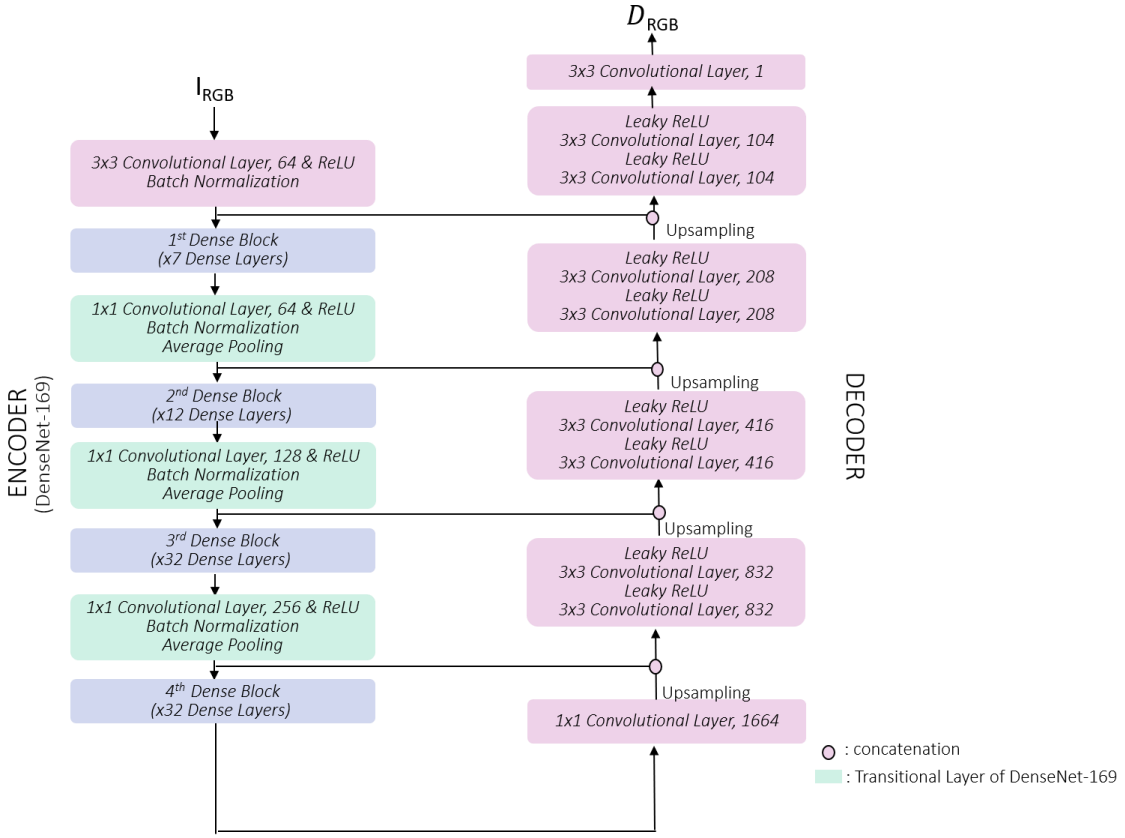
Το γενικό πλαίσιο λειτουργίας της μεθοδολογίας συνοψίζεται στην εικόνα (4.1). Δοθείσας μιας RGB εικόνας I_{RGB}, ένα προεκπαιδευμένο συνελικτικό δίκτυο χρησιμοποιείται για να παραγάγει το χάρτη βάθους D_{RGB} , ο οποίος αντικατοπτρίζει την πληροφορία του βάθους που αντιστοιχεί στην αναπαράσταση I_{RGB}. Ο εκτιμώμενος χάρτης βάθους D_{RGB} και η RGB εικόνα I_{RGB} συνδυάζονται και συνθέτουν ένα τρισδιάστατο τανυστή $T_{RGB\text{-}D}$ βάθους τεσσάρων χρωματικών καναλιών, ο οποίος περιλαμβάνει την RGB-D πληροφορία της εκάστοτε οπτικής αναπαράστασης. Ο τανυστής $T_{RGB\text{-}D}$ προωθείται σε ένα συνελικτικό νευρωνικό δίκτυο που φέρει αρχιτεκτονική αυτοκωδικοποιητή και έχει επιφορτιστεί με την ανίχνευση των σημαντικών αντικείμενων που φέρει η εικόνα I_{RGB}. Από το δίκτυο αυτό προκύπτει μια αρχική εκτίμηση $S_{RGB\text{-}D}$, για το χάρτη οπτικού ενδιαφέροντος της RGB εικόνας. Σε χρόνο μεταγενέστερο της εκπαίδευσης του συνελικτικού αυτοκωδικοποιητή, ο χάρτης οπτικού ενδιαφέροντος $\tilde{S}_{RGB\text{-}D}$, ο οποίος υπολογίστηκε από την αρχιτεκτονική συνελικτικού αυτοκωδικοποιητή, υφίσταται την επεξεργασία ενός επιδιορθωτικού μηχανισμού. Ο επιδιορθωτικός μηχανισμός χρησιμοποιεί ένα πυρήνα και εκπαιδεύει ένα κατώφλι προκειμένου να βελτιώσει μέσω της πράξης της συνέλιξης την αρχική εκτίμηση του χάρτη οπτικού ενδιαφέροντος $\tilde{S}_{RGB\text{-}D}$. Το αποτέλεσμα που προκύπτει από την εφαρμογή του επιδιορθωτικού μηχανισμού στον χάρτη οπτικού ενδιαφέροντος $\tilde{S}_{RGB\text{-}D}$, και το οποίο συμβολίζεται ως $S_{RGB\text{-}D}$, συνιστά την καταληκτική πρόβλεψη του προτεινόμενου μοντέλου.

4.2. Εκτίμηση των χαρτών βάθους

Για την παραγωγή των χαρτών βάθους μελετήθηκαν δυο γνωστές αρχιτεκτονικές συνελικτικών δικτύων, οι οποίες έχουν προταθεί πρόσφατα. Πρόκειται για τα δίκτυα *DenseDepth* (Alhashim & Wonka, 2018) και *MonoDepth2* (Godard, Mac Aodha, Firman, & Brostow, 2019), τα οποία αντίστοιχα θεωρούνται μοντέλα επιβλεπόμενης (supervised) και αυτο-επιβλεπόμενης μάθησης (self-supervised). Σε μια απόπειρα διερεύνησης εναλλακτικών υπολογιστικών αναπαραστάσεων όσον αφορά την πληροφορία του βάθους, εξετάστηκε μια επιπλέον προσέγγιση. Η προσέγγιση αυτή, διερευνά τη δυνητική συνεισφορά της ασαφούς συλλογιστικής στην πρόβλεψη που πραγματοποιείται από το μοντέλο *DenseDepth*. Οι υποενότητες (4.2.1), (4.2.2), (4.2.3) είναι αντίστοιχα αφιερωμένες στα προεκπαιδευμένα συνελικτικά δίκτυα *DenseDepth*, *MonoDepth2*, και στην υβριδική προσέγγιση των συνελικτικών δικτύων και της ασαφούς λογικής.

4.2.1. DenseDepth

Το μοντέλο που προτείνει το συνελικτικό δίκτυο *DenseDepth*, είναι σε θέση να εκτιμά τον χάρτη βάθους μιας δοθείσης οπτικής αναπαράστασης, έχοντας εκπαιδευτεί μονοφθαλμικά δηλαδή χρησιμοποιώντας μόνο μια εικόνα. Ειδικότερα, το δίκτυο λαμβάνει στην είσοδο του μια RGB εικόνα I_{RGB} και εξάγει τον χάρτη βάθους D_{RGB} που αντιστοιχεί σε αυτήν. Οι διαστάσεις του χάρτη βάθους D_{RGB} , που παράγεται από το μοντέλο *DenseDepth* είναι σύμφωνες με τις διαστάσεις της εικόνας I_{RGB}. Το δίκτυο *DenseDepth* φέρει αρχιτεκτονική συνελικτικού αυτοκωδικοποιητή στην οποία ενσωματώνει παραλειπόμενες συνδέσεις, προσομοιάζοντας με τον τρόπο αυτό τη χαρακτηριστική αρχιτεκτονική του συνελικτικού δικτύου U-Net (Ronneberger, Fischer, & Brox, 2015). Ο κωδικοποιητής του μοντέλου ακολουθεί την αρχιτεκτονική του συνελικτικού δικτύου DenseNet-169 (Huang et al., 2017), το οποίο αποτελεί ένα εξαιρετικά βαθύ νευρωνικό δίκτυο. Αντίθετα, ο αποκωδικοποιητής του δικτύου *DenseDepth* χαρακτηρίζεται από μια ρηχότερη αρχιτεκτονική η οποία περιλαμβάνει παραλειπόμενες συνδέσεις, με τα επίπεδα του κωδικοποιητή.



Εικόνα 4.2 Αρχιτεκτονική του δικτύου DenseDepth

Αναλυτικότερα, η αρχιτεκτονική του κωδικοποιητή συνοψίζεται στην εικόνα (4.2). Το πρώτο συνελικτικό επίπεδο του δικτύου DenseNet-169 συντίθεται από 64 πυρήνες μεγέθους

3×3 με βήμα σάρωσης 2 και μέγεθος γεμίσματος 3. Το πρώτο συνελικτικό επίπεδο ακολουθεί ένα συγκεντρωτικό επίπεδο υποδειγματοληψίας μέγιστης απόκρισης, το οποίο έχει παράθυρο χωρικών διαστάσεων 3×3 και βήμα σάρωσης ίσο με 2. Ακολούθως, το δίκτυο DenseNet169 περιλαμβάνει τέσσερα πυκνά μπλοκ (Dense Blocks), μεταξύ των οποίων παρεμβάλλονται τρία επίπεδα μετάβασης (Transition Layers). Καθένα από τα πυκνά μπλοκ αποτελείται από ένα πλήθος σειριακά τοποθετημένων πυκνών επιπέδων (Dense Layers). Το πρώτο πυκνό μπλοκ αποτελείται από επτά πυκνά επίπεδα, το δεύτερο πυκνό μπλοκ αποτελείται από δώδεκα πυκνά επίπεδα, ενώ τα δυο τελευταία συγκροτούνται από τριάντα δύο πυκνά επίπεδα έκαστο. Μέσα σε ένα πυκνό μπλοκ το κάθε πυκνό επίπεδο συνδέεται με όλα τα επόμενα του. Η δομή ενός πυκνού επιπέδου περιλαμβάνει ένα συνελικτικό επίπεδο 192 πυρήνων μεγέθους 1×1 με βήμα σάρωσης 1 και ένα συνελικτικό επίπεδο 128 πυρήνων δεκτικού πεδίου 3×3 με βήμα σάρωσης και μέγεθος γεμίσματος ίσα με 1. Καθένα από τα τρία επίπεδα μετάβασης συντίθεται από ένα συνελικτικό επίπεδο 64, 128 και 256 πυρήνων αντίστοιχα μεγέθους 1×1 με βήμα σάρωσης 1 ακολουθούμενο από ένα συγκεντρωτικό επίπεδο υποδειγματοληψίας μέγιστης απόκρισης που χρησιμοποιεί παράθυρα χωρικών διαστάσεων 2×2 και βήμα σάρωσης ίσο με 2. Όλα τα συνελικτικά επίπεδα που απαντώνται στο δίκτυο DenseNet-169 ακολουθούνται πάντα από ένα επίπεδο που εφαρμόζει κανονικοποίηση παρτίδων (batch normalization), καθώς και από ένα επίπεδο απαλοιφής της γραμμικότητας, το οποίο υιοθετεί ως συνάρτηση ενεργοποίησης τη συνάρτηση της διορθωμένης γραμμικής μονάδας (ReLU).

Ο αποκωδικοποιητής του μοντέλου *DenseDepth* λαμβάνει στην είσοδο του την αναπαράσταση που εξάγεται από το τέταρτο πυκνό επίπεδο του κωδικοποιητή. Ο αποκωδικοποιητής περιλαμβάνει ένα συνελικτικό επίπεδο αποτελούμενο από 1664 πυρήνες μεγέθους 1×1 με βήμα σάρωσης ίσο με 1. Τέσσερα συνελικτικά μπλοκ έπονται του πρώτου επιπέδου. Κάθε μπλοκ αποτελείται από δυο συνελικτικά επίπεδα. Το πλήθος των πυρήνων που χρησιμοποιούν τα συνελικτικά επίπεδα του πρώτου μπλοκ ισούται με 832 και ο αριθμός τους υποδιπλασιάζεται για κάθε επόμενο μπλοκ. Τα συνελικτικά επίπεδα σε όλα τα μπλοκ έχουν μέγεθος πυρήνων 3×3 με βήμα σάρωσης και μέγεθος γεμίσματος ίσα με 1. Επιπλέον, κάθε συνελικτικό επίπεδο ακολουθείται από ένα επίπεδο μη γραμμικότητας το οποίο χρησιμοποιεί ως συνάρτηση ενεργοποίησης τη συνάρτηση της διακεκομμένης ανορθωμένης γραμμικής μονάδας (Leaky ReLU). Η είσοδος που λαμβάνει ένα συνελικτικό μπλοκ του αποκωδικοποιητή είναι το αποτέλεσμα της συνένωσης της εξόδου του αμέσως προηγούμενου μπλοκ (στην περίπτωση του πρώτου μπλοκ είναι η έξοδος του πρώτου συνελικτικού επιπέδου του κωδικοποιητή), η οποία προηγουμένως έχει υπερδειγματοληπτηθεί (Upsampling)

προκειμένου να διπλασιαστούν οι χωρικές διαστάσεις της (μήκος και πλάτος) και της εξόδου των επιπέδων μετάβασης του κωδικοποιητή. Έτσι, η αναπαράσταση της εξόδου που προέρχεται από το τρίτο, δεύτερο και πρώτο επίπεδο μετάβασης του κωδικοποιητή χρησιμοποιείται αντιστοίχως στο πρώτο, δεύτερο και τρίτο μπλοκ του αποκωδικοποιητή. Το τελευταίο επίπεδο του κωδικοποιητή συνενώνεται με την έξοδο του συγκεντρωτικού επιπέδου που ακολουθεί το πρώτο συνελικτικό επίπεδο του κωδικοποιητή (πρόκειται για την ίδια αναπαράσταση που εισέρχεται στο πρώτο πυκνό μπλοκ του κωδικοποιητή). Μετά το πέρας των μπλοκ συναντάμε το επίπεδο εξόδου του δικτύου, το οποίο αποτελείται από έναν πυρήνα μεγέθους 3x3 με βήμα σάρωσης και μέγεθος γεμίσματος ίσα με 1.

Μια ειδικά σχεδιασμένη συνάρτηση απώλειας χρησιμοποιείται κατά τη διάρκεια της εκπαίδευσης η οποία πραγματοποιείται με τον αλγόριθμο Adam. Η συνάρτηση απώλειας που χρησιμοποιείται είναι το σταθμισμένο άθροισμα της L1 νόρμας των τιμών του βάθους, της L1 νόρμας των τιμών της παραγώγου του βάθους και του μέτρου της δομικής ομοιότητας (Structural Similarity Index Measure - SSIM) (Z. Wang, Simoncelli, & Bovik, 2003). Υποθέτοντας ότι $\hat{y}$ είναι ο χάρτης βάθους όπως εκτιμάται από την αρχιτεκτονική του δικτύου και y είναι η βάση αλήθειας του εκτιμώμενου χάρτη βάθους τότε η συνάρτηση απώλειας ορίζεται ως εξής:

$$\mathcal{L}(y, \hat{y}) = \lambda_1 \mathcal{L}_{depth}(y, \hat{y}) + \lambda_2 \mathcal{L}_{grad}(y, \hat{y}) + \lambda_3 \mathcal{L}_{SSIM}(y, \hat{y}) \quad \textbf{(4.1)}$$

όπου,

$$\mathcal{L}_{depth}(y, \hat{y}) = \frac{1}{n} \sum_{p}^{n} \left| y_p - \hat{y}_p \right| \quad \textbf{(4.2)}$$

$$\mathcal{L}_{grad}(y, \hat{y}) = \frac{1}{n} \sum_{p}^{n} \left| g_x(y_p, \hat{y}_p) \right| + \left| g_y(y_p, \hat{y}_p) \right| \quad \textbf{(4.3)}$$

$$\mathcal{L}_{SSIM}(y, \hat{y}) = \frac{1 - SSIM(y, \hat{y})}{2} \quad \textbf{(4.4)}$$

με $\lambda_1 = 0{,}1$ και $\lambda_2 = \lambda_3 = 1$.

Το δίκτυο εκπαιδεύτηκε στο σύνολο δεδομένων NYU Depth v2 (Nathan Silberman & Fergus, 2012), το οποίο περιλαμβάνει RGB εικόνες και τους αντίστοιχους χάρτες βάθους για ένα πλήθος ποικίλων σεναρίων εσωτερικών χώρων. Ο κωδικοποιητής του δικτύου αρχικοποιήθηκε με τις τιμές των βαρών όπως αυτές έχουν διαμορφωθεί από την ταξινόμηση

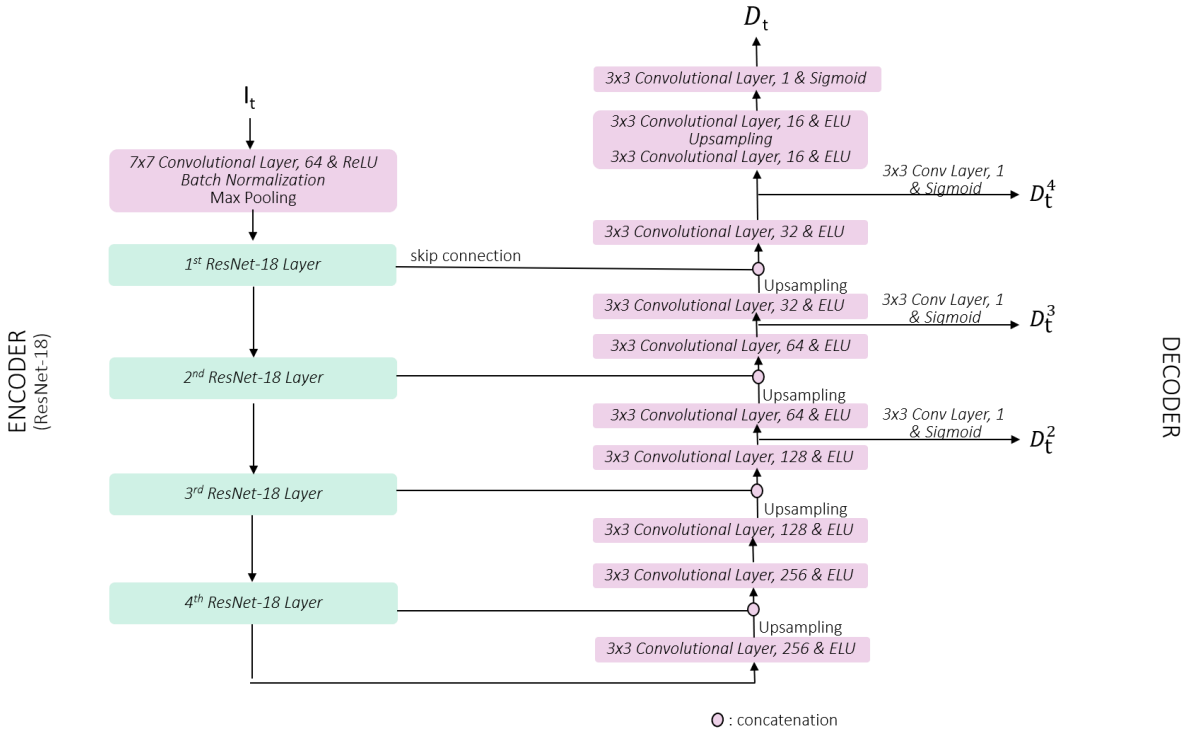
του συνόλου δεδομένων ImageNet, στην οποία είχε εκπαιδευτεί σε πρότερο χρόνο το μοντέλο DenseNet-169.

4.2.2. MonoDepth2

Ένα αυτο-επιβλεπόμενο μοντέλο για μονοφθαλμική εκτίμηση του βάθους μιας δεδομένης οπτικής αναπαράστασης μας συστήνει το συνελικτικό δίκτυο *MonoDepth2*. Η μονοφθαλμική εκτίμηση του βάθους μιας οπτικής αναπαράστασης επαναπροσδιορίζεται από την κατηγορία των μοντέλων της αυτό-επιβλεπόμενης μάθησης, η οποία αντιμετωπίζει την εκτίμηση ενός χάρτη βάθους ως ένα πρόβλημα σύνθεσης εικόνας. Μια αυτό-επιβλεπόμενη μεθοδολογία που πραγματοποιεί μονοφθαλμική εκτίμηση του βάθους, εκπαιδεύει ένα δίκτυο προκειμένου να παράγει χάρτες βάθους εξαιρώντας από τη διαδικασία της μάθησης τη γνώση του συνόλου δεδομένων που συνιστά τη βάση αλήθειας (Godard et al., 2019). Για την εκπαίδευση ενός μοντέλου αυτό-επιβλεπόμενης μάθησης χρησιμοποιούνται είτε σε ζεύγη στερεοσκοπικών εικόνων (στερεοσκοπική εκπαίδευση) είτε σε εικόνες βίντεο που έχουν ληφθεί από μια κάμερα (μονοφθαλμική εκπαίδευση). Πάντως, σε κάθε περίπτωση το εκπαιδευμένο μοντέλο είναι σε θέση να εκτιμά χάρτες βάθους χρησιμοποιώντας μόνο μια εικόνα για την εκάστοτε οπτική αναπαράσταση (C. Zhao, Sun, Zhang, Tang, & Qian, 2020).

Το μοντέλο *MonoDepth2* αξιοποιεί ταυτόχρονα κατά τη διαδικασία της εκπαίδευσης τόσο ζεύγη στερεοσκοπικών εικόνων όσο και εικόνες μονοφθαλμικού βίντεο, ενώ ο υπολογισμός των χαρτών βάθους πραγματοποιείται έμμεσα μέσω των χαρτών της στερεοσκοπικής μετατόπισης. Αναλυτικότερα, δοθείσης μιας RGB εικόνας I_t, μια αρχιτεκτονική συνελικτικού αυτοκωδικοποιητή (depth network) εκπαιδεύεται ώστε να εκτιμά τον χάρτη της στερεοσκοπικής μετατόπισης, μέσω του οποίου υπολογίζεται ο χάρτης βάθους D_t . Η ανάλυση του εκτιμώμενου χάρτη βάθους είναι ίδια με εκείνη της εικόνας I_t. Ένα δεύτερο δίκτυο (pose network) , συνεπιτηρεί τη διαδικασίας της εκπαίδευσης του συνελικτικού αποκωδικοποιητή. Το δίκτυο αυτό λαμβάνει στην είσοδο του την εικόνα I_t καθώς και μια δεύτερη εικόνα $I_{t'}$, και παράγει ένα μετασχηματισμό $T_{t\to t'}$ με έξι βαθμούς ελευθερίας. Από τον μετασχηματισμό αυτό συνάγεται η σχετική μετατόπιση -ιδιοκίνηση των αντικειμένων- της εικόνα I_t ως προς την εικόνα $I_{t'}$. Μεταξύ των εικόνων $I_{t'}$ συγκαταλέγονται: είτε η αντίθετη όψη της εικόνας I_t, είτε κάποια χρονικά παρακείμενη I_{t-1} , I_{t+1}στην I_t εικόνα. Στην πρώτη περίπτωση το δίκτυο *MonoDepth2* εκπαιδεύεται με ζεύγη στερεοσκοπικών εικόνων ενώ στις άλλες δυο με εικόνες μονοφθαλμικού βίντεο.

Αναλυτικότερα, η αρχιτεκτονική του συνελικτικού αυτοκωδικοποιητή (depth network), συνοψίζεται στην εικόνα (4.3). Όπως φαίνεται από την εικόνα πρόκειται για την χαρακτηριστική αρχιτεκτονική του δικτύου U-Net (Ronneberger et al., 2015) η οποία περιλαμβάνει κωδικοποιητή, αποκωδικοποιητή και παραλειπόμενες συνδέσεις μεταξύ των δύο.



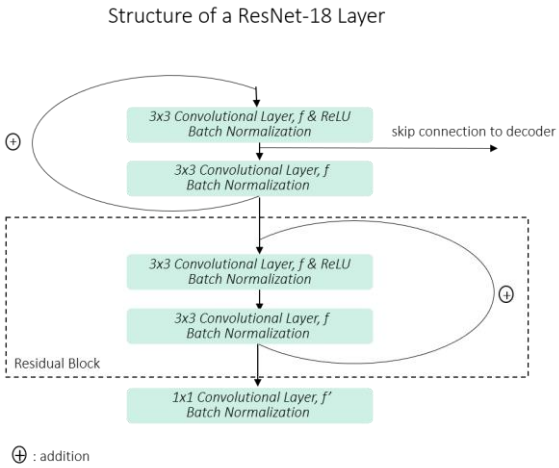
Εικόνα 4.3 Η αρχιτεκτονική του δικτύου depth του μοντέλου MonoDepth2

Ο κωδικοποιητής δανείζεται την αρχιτεκτονική του δικτύου ResNet-18. Το δίκτυο ResNet-18 περιλαμβάνει ένα συνελικτικό επίπεδο που φέρει από 64 πυρήνες μεγέθους 7×7 με βήμα σάρωσης 2 και μέγεθος γεμίσματος (padding) 3. Το συνελικτικό επίπεδο ακολουθούν ένα επίπεδο κανονικοποίησης παρτίδων και ένα συγκεντρωτικό επίπεδο υποδειγματολειψίας μέγιστης απόκρισης, το οποίο έχει παράθυρο χωρικών διαστάσεων 3×3 και βήμα σάρωσης ίσο με 2. Ακολούθως, το δίκτυο ResNet-18 περιλαμβάνει τέσσερα επίπεδα καθένα από τα οποία συντίθεται από δυο ίδια υπολειμματικά μπλοκ (Residual Blocks). Η δομή ενός επιπέδου ResNet-18 μπορεί να παρατηρηθεί στην εικόνα (4.4).

Όπως μπορεί να παρατηρηθεί, το κάθε υπολειμματικό μπλοκ αποτελείται από δυο συνελικτικά επίπεδα εκ των οποίων μόνο το πρώτο ακολουθείται από ένα επίπεδο απαλοιφής της γραμμικότητας που χρησιμοποιεί τη συνάρτηση της ανορθωμένης γραμμικής μονάδας (ReLU). Ακόμη, ένα επίπεδο κανονικοποίησης παρτίδων ακολουθεί καθένα από τα συνελικτικά επίπεδα των υπολειμματικών μπλοκ. Κάθε φορά στην έξοδο του δεύτερου

συνελικτικού επιπέδου ενός υπολειμματικού μπλοκ προστίθεται και η είσοδος του υπολειμματικού μπλοκ (μηχανισμός της υπολειμματικής μάθησης). Τα συνελικτικά επίπεδα των υπολειμματικών μπλοκ φέρουν πυρήνες διαστάσεων 3x3 και τιμές γεμίσματος (padding) και βήματος σάρωσης ίσες με 1. Το πλήθος των πυρήνων - στην εικόνα 4.4 δηλώνεται ως f - που χρησιμοποιούνται από τα συνελικτικά επίπεδα των υπολειμματικών μπλοκ για τα υπ΄αριθμόν επίπεδα 1, 2, 3,4 ισούνται αντίστοιχα με 64, 128, 256, 512. Στο σημείο αυτό γίνεται αντιληπτό το γεγονός, ότι το δεύτερο υπολειμματικό μπλοκ του πρώτου επιπέδου παρουσιάζει ασυμφωνία ως προς τη διάσταση του βάθους με το πρώτο υπολειμματικό μπλοκ του δεύτερου επιπέδου (64 έναντι 128). Επιπλέον, ούτε οι χωρικές διαστάσεις του μήκους και του πλάτους βρίσκονται σε συμφωνία, καθώς υποδιπλασιάζονται σε κάθε επίπεδο του κωδικοποιητή. Προκειμένου να αντιμετωπιστούν οι παραπάνω ανακολουθίες -που προκαλούνται από τις μεταβάσεις μεταξύ του πρώτου-δεύτερου, δεύτερου-τρίτου και τρίτου-τέταρτου επιπέδου- προτού πραγματοποιηθεί η πρόσθεση η αναπαράσταση του ρηχότερου επιπέδου διέρχεται από ένα επιπλέον συνελικτικό επίπεδο. Αυτό το επιπλέον συνελικτικό επίπεδο αποτελείται από πλήθος πυρήνων ίδιο με εκείνο των συνελικτικών επιπέδων του βαθύτερου επιπέδου για να δημιουργηθεί συμφωνία ως προς την διάσταση του βάθους. Το πλήθος των πυρήνων του προαναφερόμενου συνελικτικού επιπέδου των τριών πρώτων υπολειμματικών μπλοκ - στην εικόνα 4.4 δηλώνεται ως f’ - ισούται αντίστοιχα με 128, 256, 512. Γίνεται εύκολα αντιληπτό ότι το τέταρτο υπολειμματικό μπλοκ δεν διαθέτει επιπλέον συνελικτικό επίπεδο. Η τιμή του βήματος σάρωσης του πρόσθετου συνελικτικού επιπέδου, το οποίο συναντάται μόνο στα τρία πρώτα υπολειμματικά μπλοκ, ισούται με 2, ούτως ώστε να επιτευχθεί ο υποδιπλασιασμός των διαστάσεων του μήκους και του πλάτους. Το δεκτικό πεδίο των πυρήνων είναι ισούται με 1×1.

Ο δε αποκωδικοποιητής χαρακτηρίζεται από μια απλούστερη αρχιτεκτονική την οποία συνθέτουν πέντε συνελικτικά μπλοκ. Κάθε συνελικτικό μπλοκ αποτελείται από δυο συνελικτικά επίπεδα αποτελούμενα από ίδιο αριθμό πυρήνων μεγέθους 3x3 με βήμα σάρωσης και μέγεθος γεμίσματος (padding) 1. Ο αριθμός αυτός είναι 256 για τα συνελικτικά επίπεδα του πρώτου μπλοκ και υποδιπλασιάζεται σε κάθε επόμενο.



Εικόνα 4.4. Η δομή καθενός από τα τέσσερα επίπεδα του δικτύου ResNet-18.

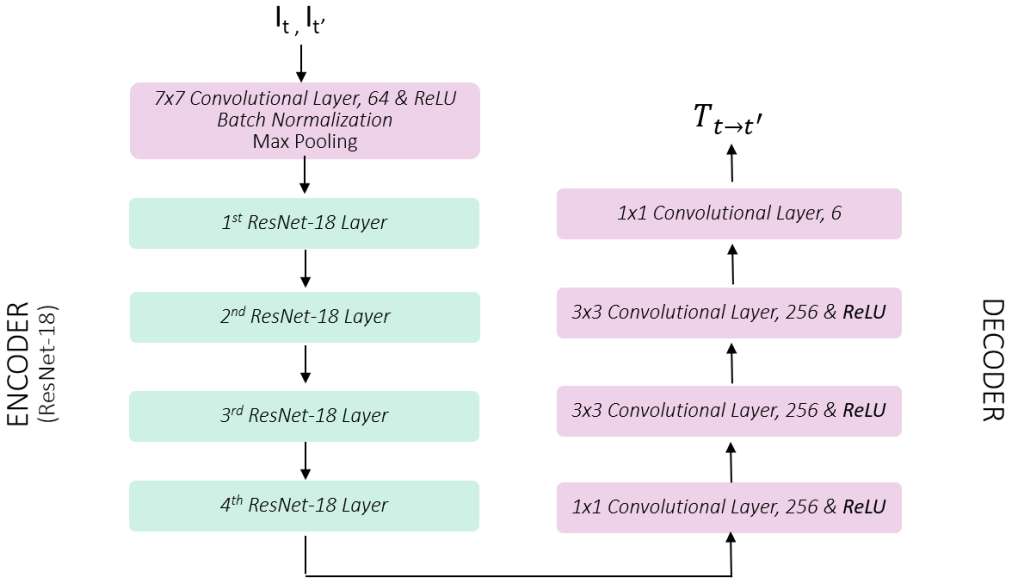
Προτού η έξοδος του πρώτου συνελικτικού επιπέδου ενός μπλοκ του αποκωδικοποιητή προωθηθεί στο δεύτερο επίπεδο του, υπερδειγματοληπτείται (upsampling) έτσι ώστε να διπλασιαστούν οι διαστάσεις του μήκους και του πλάτους της. Επιπλέον, στα τέσσερα πρώτα μπλοκ, το δεύτερο συνελικτικό επίπεδο επεξεργάζεται την υπερδειγματοληπτημένη (upsampled) έξοδο του πρώτου συνελικτικού επιπέδου αφού προηγουμένως αυτή συνενωθεί (skip connection) με την αναπαράσταση της εξόδου του υπολειμματικού μπλοκ του κωδικοποιητή, με την οποία παρουσιάζει συμφωνία ως προς τις χωρικές διαστάσεις του μήκους και του πλάτους. Έτσι, η έξοδος του τέταρτου, τρίτου, δεύτερου, πρώτου υπολειμματικού μπλοκ συνενώνεται αντίστοιχα με την αναπαράσταση της εξόδου του πρώτου συνελικτικού επιπέδου που ανήκει στο πρώτο, δεύτερο, τρίτο, τέταρτο μπλοκ του αποκωδικοποιητή. Κάθε συνελικτικό επίπεδο ακολουθείται από ένα επίπεδο απαλοιφής της γραμμικότητας το οποίο χρησιμοποιεί ως συνάρτηση ενεργοποίησης τη συνάρτηση της εκθετικής γραμμικής μονάδας (Exponential Linear Unit -ELU). Η τεχνική γεμίσματος της αντανάκλασης (reflection padding) υιοθετείται από όλα τα συνελικτικά επίπεδα προκειμένου να περιοριστεί ο αριθμός των ορατών σφαλμάτων στα όρια του χάρτη βάθους. Χάρτες της στερεοσκοπικής μετατόπισης παράγονται σε τέσσερις διαφορετικές κλίμακες από τα μπλοκ 2-5. Η επινόηση αυτή δηλαδή η πρόβλεψη σε πολλαπλές κλίμακες, συμβάλλει στη βελτίωση της εκπαίδευσης του συνολικού μοντέλου όπως θα εξηγηθεί και παρακάτω. Ο υπολογισμός ενός χάρτη στερεοσκοπικής μετατόπισης γίνεται από ένα συνελικτικό επίπεδο ενός πυρήνα μεγέθους 3×3 με βήμα σάρωσης και μέγεθος γεμίσματος (padding) 1. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται στο επίπεδο απαλοιφής της γραμμικότητας που συνοδεύει το προαναφερόμενο συνελικτικό επίπεδο είναι η σιγμοειδής.

Οι χάρτες της στερεοσκοπικής μετατόπισης που εξάγονται από τον κωδικοποιητή του δικτύου depth ανάγονται σε χάρτες βάθους D_t έπειτα από την εφαρμογή της σχέσης (4.5)

$$D_t = \frac{1}{a\sigma + \beta} \tag{4.5}$$

όπου σ οι τιμές του χάρτη στερεοσκοπικής μετατόπισης ενώ οι τιμές των a, β επιλέγονται έτσι ώστε να περιορίσουν το εύρος των τιμών του χάρτη βάθους D_t στο διάστημα [0.1, 100].

Η αρχιτεκτονική του δεύτερου δικτύου (pose network) συνοψίζεται στην εικόνα (4.5). Πρόκειται επίσης για έναν συνελικτικό αποκωδικοποιητή χωρίς ωστόσο παραλειπόμενες συνδέσεις. Ο κωδικοποιητής και αυτού του δικτύου υιοθετεί την αρχιτεκτονική του συνελικτικού δικτύου ResNet-18. Ο αποκωδικοποιητής του δικτύου συντίθεται από μόλις τέσσερα συνελικτικά επίπεδα. Τα τρία πρώτα επίπεδα αποτελούνται από 256 πυρήνες μεγέθους 1×1, 3×3, 3×3 αντίστοιχα και έχουν βήμα μετατόπισης ίσο με 1. Τα επίπεδα αυτά ακολουθούνται από επίπεδα απαλοιφής γραμμικότητας που χρησιμοποιούν τη συνάρτηση της ανορθωμένης γραμμικής μονάδας (ReLU). Το επίπεδο εξόδου του αποκωδικοποιητή συνιστά ένα συνελικτικό επίπεδο 6 πυρήνων μεγέθους 1x1 και βήματος σάρωσης 1. Το επίπεδο εξόδου δεν ακολουθείται από επίπεδο απαλοιφής της γραμμικότητας.



Εικόνα 4.5 Η αρχιτεκτονική του δικτύου pose του μοντέλου MonoDepth2

Η συνάρτηση απώλειας που χρησιμοποιείται κατά τη διαδικασία της εκπαίδευσης των δυο δικτύων που συνθέτουν το μοντέλο *MonoDepth2* δίνεται από την ακόλουθη σχέση:

$$\mathcal{L} = \mu\mathcal{L}_p + \lambda\mathcal{L}_s \tag{4.6}$$

όπου $\mathcal{L}_p, \mathcal{L}_s, \mu$ δίνονται αντίστοιχα από τις σχέσεις (4.7), (4.10), (4.11) και η τιμή της παραμέτρου λ ισούται με 0,001

$$\mathcal{L}_p = \min_{t\prime} pe\ (I_t\ , I_{t\prime\to t}) \tag{4.7}$$

όπου pe το φωτομετρικό σφάλμα επαναπροβολής (photometric reprojection error) των σημείων των εικόνων I_t $και$ $I_{t\prime\to t}$.

Το φωτομετρικό σφάλμα pe ανακατασκευής των σημείων δυο εικόνων I_a, I_β ορίζεται σύμφωνα με τη σχέση (4.8)

$$pe\left(I_a, I_\beta\right) = \frac{\alpha}{2}\Big(1 - SSIM\big(I_a, I_\beta\big)\Big) + (1-a)\left\|I_a - I_\beta\right\|_1 \tag{4.8}$$

στην συγκεκριμένη περίπτωση η παράμετρος a λαμβάνει την τιμή 0,85.

Οι τιμές της εικόνας $I_{t\prime\to t}$ υπολογίζονται από τη σχέση (4.9)

$$I_{t\prime\to t} = I_{t\prime}\langle proj(D_t, T_{t\to t'}, K)\rangle \tag{4.9}$$

όπου K οι εσωτερικές παράμετροι της κάμερας για τις οποίες έχει γίνει η σύμβαση ότι παραμένουν αναλλοίωτες.

Η ύπαρξη του δεύτερου όρου της σχέσης (4.6) αποσκοπεί στη βελτιστοποίηση της διαδικασίας της εκπαίδευσης όταν αυτή πραγματοποιείται με τη χρήση των εικόνων μονοφθαλμικού βίντεο. Η εκτίμηση του πραγματοποιείται σύμφωνα με τη σχέση (4.10).

$$\mathcal{L}_s = |\partial \mathrm{x}\ d_t^*|\ e^{-|\partial \mathrm{x}\ \mathrm{I}_t|} + \ |\partial \mathrm{y}\ d_t^*|\ e^{-|\partial \mathrm{y}\ \mathrm{I}_t|} \tag{4.10}$$

όπου $d_t^* = d_t/\overline{d_t}$ η μέση κανονικοποιημένη και ανάστροφη αναπαράσταση του εκτιμώμενου χάρτη βάθους. Ο όρος αυτός ποσοτικοποιεί την ομαλότητα των ακμών.

Τέλος, η παράμετρος μ την οποία χρησιμοποιεί η εξίσωση (4.6) λαμβάνει τιμές σύμφωνα με την ακόλουθη συνάρτηση (4.11)

$$\mu = \begin{cases} 1, & \alpha\nu\ \min_{t\prime} pe\ (I_t\ , I_{t\prime\to t}) < \min_{t\prime} pe\ (I_t\ , I_{t\prime}) \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{4.11}$$

Η παράμετρος μ αυτή επιτρέπει στο δίκτυο να αγνοεί τα εικονοστοιχεία που παραμένουν ίδια σε δυο χωρικά παρακείμενα πλαίσια της ίδιας οπτικής αναπαράστασης καθώς επίσης και ολόκληρα πλαίσια εικόνων όταν η κάμερα παραμένει στατική

Η συνάρτηση απώλειας υπολογίζεται για κάθε μια από τις τέσσερις κλίμακες από τις οποίες εκτιμάται ένας χάρτης βάθους και η τελική τιμή της διαμορφώνεται από τον μέσο των επιμέρους τιμών. Πριν τον υπολογισμό οι χάρτες βάθους των ενδιάμεσων επιπέδων (2-4)

υπερδειγματοληπτούνται (upsampling) ούτως ώστε ο υπολογισμός της να διεξαχθεί σε ανάλυση σύμφωνη με την εκείνη της εικόνας I_t.

Το δίκτυο εκπαιδεύτηκε στο σύνολο δεδομένων KITTI 2015 stereo (Menze & Geiger, 2015), το οποίο περιλαμβάνει ζεύγη στερεοσκοπικών εικόνων τα οποία στην πλειοψηφία τους είναι χωρικά παρακείμενα. Πρόκειται αποκλειστικά για σενάρια εξωτερικών χώρων με εικόνες φυσικού περιβάλλοντος (αστικά και υπαίθρια τοπία, αυτοκινητόδρομοι). Η εκπαίδευση του μοντέλου πραγματοποιήθηκε με τη χρήση του αλγορίθμου Adam. Οι κωδικοποιητές των δυο δικτύων αρχικοποιήθηκαν με βάρη οι τιμές των οποίων προέκυψαν από την ταξινόμηση του συνόλου δεδομένων ImageNet, στην οποία είχε εκπαιδευτεί σε πρότερο χρόνο το δίκτυο ResNet-48 .

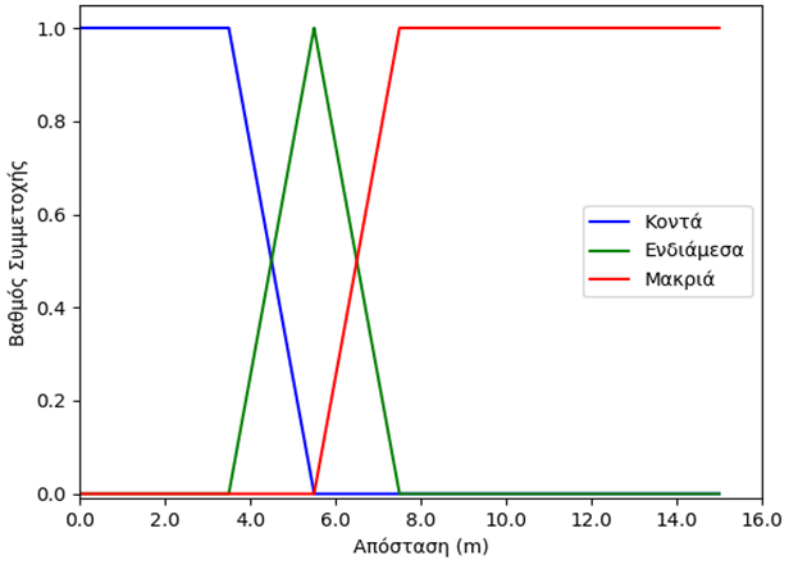
4.2.3. Υβριδική προσέγγιση του μοντέλου DenseDepth και της Ασαφούς Λογικής

Η προσέγγιση αυτή παρέχει μια ασαφή αναπαράσταση για τους χάρτες βάθους D_{RGB} που παρήχθησαν από το συνελικτικό δίκτυο *DenseDepth*. Για το σκοπό αυτό, θεωρούμε τρία διαφορετικά ασαφή σύνολα $D_1,\ D_2,\ D_3$ καθένα από τα οποία περιγράφει την απόσταση στην οποία βρίσκονται τα εικονιζόμενα αντικείμενα της εκάστοτε οπτικής σκηνής από το σημείο θέασης τους. Τα ασαφή σύνολα $D_1,\ D_2,\ D_3$, αντιπροσωπεύουν αντίστοιχα τις ασαφείς/λεκτικές τιμές: κοντινή, ενδιάμεση, μακρινή, οι οποίες αναφέρονται στην απόσταση. Καθένα από τα παραπάνω τρία ασαφή σύνολα, που αναπαριστά ένα διαφορετικό επίπεδο απόστασης ορίζεται σε πεδίο με εύρος τις τιμές βάθους των ομώνυμων χαρτών που έχουν παραχθεί από το συνελικτικό μοντέλο *DenseDepth*. Στο σημείο αυτό οφείλουμε να διευκρινίσουμε ότι μεταξύ των συνόλων κοντινής-ενδιάμεσης και ενδιάμεσης-μακρινής απόστασης υπάρχει επικάλυψη που αντιστοιχεί στις μεταβατικές καταστάσεις. Τα ασαφή σύνολα $D_1,\ D_2,\ D_3$, περιγράφονται αντίστοιχα από τις συναρτήσεις συμμετοχής $d_1(z), d_2(z), d_3(z),$ με $z\ \epsilon\ [0, \infty)$, οι οποίες εικονίζονται στην εικόνα (4.6).

Από την απόκριση της κάθε συνάρτησης συμμετοχής $d_i(z)\ με\ i\ \epsilon\ \{1, 2, 3\}$ παράγεται μια ασαφής αναπαράσταση των χαρτών βάθους D_{RGB} όπως αυτοί διαμορφώθηκαν από το συνελικτικό δίκτυο *DenseDepth*. Η προαναφερθείσα διαδικασία συνοψίζεται στη σχέση (4.12).

$$D_M^i(\mathrm{D_{RGB}}) = d_i(\mathrm{D_{RGB}})\ \forall\ i\ \epsilon\ \{1, 2, 3\} \qquad \textbf{(4.12)}$$

Από την εφαρμογή της προηγούμενης σχέσης σε κάθε χάρτη βάθους D_{RGB} που παράχθηκε από το δίκτυο *DenseDepth* και ο οποίος αντιστοιχεί στην RGB αναπαράσταση I_{RGB} προκύπτουν τρεις διαφορετικοί ασαφείς χάρτες βάθους D_M^1, D_M^2, D_M^3. Ο κάθε χάρτης βάθους D_M^i απεικονίζει τις περιοχές της εικόνας I_{RGB} οι οποίες βρίσκονται εντός των ορίων της απόστασης που ορίζονται από την ασαφή συνάρτηση d_i, κωδικοποιώντας όλη την πληροφορία σε ένα μόνο χρωματικό κανάλι. Οι αποστάσεις των εικονοστοιχείων μετριούνται από τη θέση παρατήρησης της αναπαράστασης I_{RGB}. Συγκεκριμένα, στο χάρτη βάθους D_M^1, αντικατοπτρίζονται οι περιοχές που βρίσκονται σε κοντινή απόσταση από το σημείο θέασης της εικόνας I_{RGB} ενώ στους χάρτες βάθους D_M^2 *και* D_M^3 εικονίζονται αντίστοιχα οι περιοχές που βρίσκονται σε ενδιάμεση και μακρινή απόσταση από τη θέση παρατήρησης της οπτικής αναπαράστασης I_{RGB}.



Εικόνα 4.6 Συναρτήσεις συμμετοχής

Στα πειράματα που διεξάγονται στην παρούσα εργασία και τα οποία παρουσιάζονται στο επόμενο κεφάλαιο οι ασαφείς αναπαραστάσεις της πληροφορίας του βάθους χρησιμοποιούνται έπειτα από την εφαρμογή των σχέσεων (4.13) ή (4.14).

$$\mathrm{D_{RGB}} = con(D_M^1, D_M^2, D_M^3) \quad \textbf{(4.13)}$$

$$\mathrm{D_{RGB}} = con(D_M^1, D_M^2) \quad \textbf{(4.14)}$$

όπου η συνάρτηση $con(\cdot)$ αναφέρεται στη συνένωση των αναπαραστάσεων D_M^i ως προς την τρίτη διάσταση (του βάθους).

4.3. Ανιχνευτής Σημαντικών Αντικειμένων

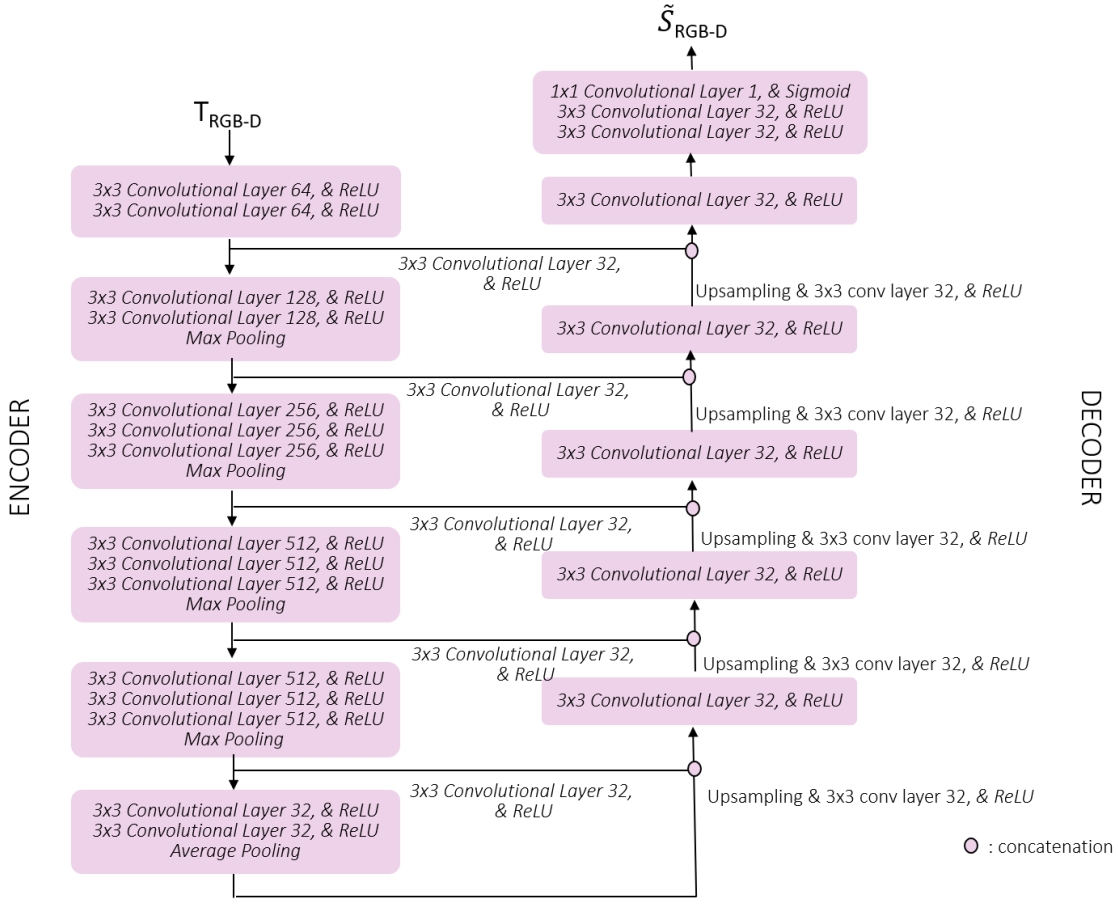
Η αρχιτεκτονική που υιοθετεί το δίκτυο από το οποίο πραγματοποιείται η ανίχνευση των σημαντικών αντικειμένων είναι εκείνη που υποδεικνύει το *RGB-D* υποδίκτυο της τριπλής αρχιτεκτονικής του μοντέλου D^3Net (Fan, Lin, et al., 2020). Το μοντέλο D^3Net αξιοποιεί την πληροφορία του βάθους για να επιτελέσει ανίχνευση των σημαντικών αντικειμένων σε μια σκηνή. Πρόκειται για ένα επίσης πρόσφατα προτεινόμενο μοντέλο, που επιτυγχάνει αξιοσημείωτη απόδοση. Το πλαίσιο λειτουργίας του μοντέλου D^3Net παρουσιάστηκε στην υποενότητα (3.4.3.3).

Το *RGB-D* υποδίκτυο του μοντέλου D^3Net λαμβάνει στην είσοδο του έναν τανυστή $T_{RGB\text{-}D}$ ο οποίος περιλαμβάνει την εικόνα I_{RGB} και τον αντίστοιχο χάρτη βάθους D_{RGB} ο οποίος βάσει των παραπάνω μεθοδολογιών. Οι δυο είδη της πληροφορίας που αντιστοιχούν στην αναπαράσταση I_{RGB} συνενώνονται ως προς τη τρίτη τους διάσταση για να συγκροτήσουν τον τανυστή $T_{RGB\text{-}D}$ Ο όγκος της εισόδου που λαμβάνει το εν λόγω δίκτυο στο επίπεδο εισόδου του διαφέρει κατά περίπτωση. Σε αντίθεση με τη χρωματική πληροφορία I_{RGB} η οποία κωδικοποιείται με τρία κανάλια η πληροφορία του βάθους αναπαρίσταται κάθε φορά με ένα διαφορετικό πλήθος καναλιών. Συγκεκριμένα, όταν αυτή ανακτάται από τα συνελικτικά δίκτυα *DenseDepth* και *MonoDepth2* -μεθοδολογίες (4.2.1) και (4.2.2)- συντίθεται αρχικά από τρία κανάλια πληροφορίας. Ωστόσο, όλη η πληροφορία του βάθους κωδικοποιείται τελικά με ένα μόνο κανάλι πληροφορίας και η συνολική διάσταση του βάθους του τανυστή $T_{RGB\text{-}D}$ ισούται με τέσσερα. Στην περίπτωση όπου ο χάρτης βάθους D_{RGB} που αντιστοιχεί στην αναπαράσταση I_{RGB} προέκυψε σύμφωνα με την μεθοδολογία που περιγράφθηκε στην υποενότητα (4.2.3) , τότε αυτός αποτελείται είτε από τρία είτε από δύο κανάλια πληροφορίας ανάλογα με τον αν κατασκευάστηκε σύμφωνα με την εξίσωση (4.13) ή (4.14). Άρα, για αυτές τις περιπτώσεις η τρίτη διάσταση του τανυστή $T_{RGB\text{-}D}$ θα έχει αντίστοιχα τιμή ίση είτε με έξι είτε με πέντε.

Εικόνα 4. 7 Χάρτες βάθους. Από αριστερά προς τα δεξιά εικονίζονται οι χάρτες βάθους όπως απαθανατίζονται από την κάμερα βάθους, το συνελικτικό δίκτυο DenseDepth, το συνελικτικό δίκτυο MonoDepth2, την εφαρμογή των κανόνων Ασαφούς Λογικής στους χάρτες του δικτύου DenseDepth.

Το εν λόγω δίκτυο φέρει τη χαρακτηριστική αρχιτεκτονική του δικτύου U-Net (Ronneberger et al., 2015). Πρόκειται για ένα συνελικτικό αυτοκωδικοποιητή, με παραλειπόμενες συνδέσεις μεταξύ των δυο κύριων δομικών δικτύων του (συνελικτικός κωδικοποιητής και συνελικτικός αποκωδικοποιητής). Η αρχιτεκτονική του εν λόγω δικτύου συνοψίζεται στην εικόνα (4.8).

Ο κωδικοποιητής του δικτύου επεκτείνει την αρχιτεκτονική του συνελικτικού δικτύου VGG-16.Το δίκτυο VGG-16 συντίθεται από πέντε συνελικτικά μπλοκ ακολουθούμενα από τρία πλήρως συνδεδεμένα επίπεδα. Όπως παρατηρείτε και στην εικόνα (4.8), καθένα από τα δυο πρώτα συνελικτικά μπλοκ του δικτύου αποτελείται από δυο συνελικτικά επίπεδα ακολουθούμενα από ένα συγκεντρωτικό επίπεδο υποδειγματοληψίας μέγιστης απόκρισης, ενώ καθένα από τα επόμενα τρία συνελικτικά μπλοκ απαρτίζεται από τρία συνελικτικά επίπεδα ακολουθούμενα από ένα συγκεντρωτικό επίπεδο υποδειγματοληψίας μέγιστης απόκρισης. Ο κωδικοποιητής του δικτύου εξαιρεί από την αρχιτεκτονική του τα πλήρως συνδεδεμένα επίπεδα και χρησιμοποιεί μόνο τα πέντε συνελικτικά μπλοκ του δικτύου VGG-16.



Εικόνα 2.8 Αρχιτεκτονική του συνελικτικού αυτοκωδικοποιητή που πραγματοποιεί την ανίχνευση των σημαντικών αντικειμένων

Προκειμένου ο κωδικοποιητής να είναι σε θέση να διατηρεί όσο το δυνατόν περισσότερη από τη σημασιολογική πληροφορία των RGB-D δεδομένων που επεξεργάζεται, η αρχιτεκτονική του δικτύου VGG-16 εμπλουτίζεται με τη προσθήκη ενός έκτου συνελικτικού μπλοκ. Το τελευταίο συνελικτικό μπλοκ του κωδικοποιητή περιλαμβάνει δυο συνελικτικά επίπεδα 32 πυρήνων, ακολουθούμενα από ένα συγκεντρωτικό επίπεδο υποδειγματοληψίας μέσης απόκρισης. Κάθε συνελικτικό επίπεδο του κωδικοποιητή χρησιμοποιεί πυρήνες μεγέθους 3×3 και τις ίδιες υπερ-παραμέτρους. Συγκεκριμένα, όλα έχουν βήμα σάρωσης ίσο με 1 και μέγεθος γεμίσματος επίσης ίσο με 1. Κάθε συνελικτικό επίπεδο ακολουθείται από ένα επίπεδο απαλοιφής της γραμμικότητας το οποίο υιοθετεί ως συνάρτηση ενεργοποίησης τη συνάρτηση της ανορθωμένης γραμμικής μονάδας (ReLU). Τέλος, τα συγκεντρωτικά επίπεδα χρησιμοποιούν παράθυρα χωρικών διαστάσεων 2×2 και η υπερ-παράμετρος του βήματος σάρωσης των παραθύρων είναι ορισμένη στη τιμή 2.

Ο αποκωδικοποιητής του δικτύου συντίθεται και αυτός από έξι συνελικτικά μπλοκ. Καθένα από τα πέντε πρώτα μπλοκ του αποτελείται από ένα συνελικτικό επίπεδο 32 πυρήνων και μεγέθους 3×3 και έχει υπερ-παραμέτρους (βήμα σάρωσης και μέγεθος γεμίσματος) ίδιες με εκείνες των συνελικτικών επιπέδων του κωδικοποιητή. Το κάθε μπλοκ λαμβάνει στην είσοδο του το αποτέλεσμα της συνένωσης του αμέσως προηγούμενου μπλοκ και της εξόδου του ανάστροφα ομόλογου μπλοκ του κωδικοποιητή. Δηλαδή, η έξοδος του πέμπτου κατά σειρά μπλοκ του κωδικοποιητή προωθείται στην είσοδο του πρώτου κατά σειρά μπλοκ του αποκωδικοποιητή. Η έξοδος του τέταρτου κατά σειρά μπλοκ του κωδικοποιητή προωθείται στην είσοδο του δεύτερου κατά σειρά μπλοκ του αποκωδικοποιητή κοκ. Οι αναπαραστάσεις που προέρχονται από τον κωδικοποιητή του δικτύου πριν συνενωθούν με εκείνες του αποκωδικοποιητή διέρχονται από ένα συνελικτικό επίπεδο 32 πυρήνων μεγέθους 1×1, με βήμα σάρωσης ίσο με 1. Ο λόγος ύπαρξης αυτών των συνελικτικών επιπέδων είναι για να διατηρηθεί συμφωνία ως προς τη τιμή του βάθους (32) του όγκου της εισόδου των συνελικτικών επιπέδων του αποκωδικοποιητή. Επίσης, πριν την συνένωση των αναπαραστάσεων που προέρχονται από τον κωδικοποιητή με τις αναπαραστάσεις που αντιστοιχούν στο αμέσως προηγούμενο μπλοκ του αποκωδικοποιητή, οι τελευταίες υπερδειγματοληπτούνται (upsampled) προκειμένου να διπλασιαστούν οι χωρικές διαστάσεις τους (μήκος και πλάτος) και κατόπιν διέρχονται από ένα συνελικτικό επίπεδο 32 πυρήνων μεγέθους 3×3, με βήμα σάρωσης ίσο με 1. Τέλος, το έκτο και καταληκτικό μπλοκ του αποκωδικοποιητή, το οποίο θεωρείται και επίπεδο εξόδου του συνελικτικού αυτοκωδικοποιητή, περιλαμβάνει τρία συνελικτικά επίπεδα. Τα δυο πρώτα εκ των τριών συνελικτικών επιπέδων φέρουν 32 πυρήνες μεγέθους 3×3, με τιμές βήματος

σάρωσης και μέγεθος γεμίσματος (padding) ίσες με 1, ενώ το τελευταίο διαθέτει έναν πυρήνα μεγέθους 1×1, και βήμα σάρωσης ίσο με 1. Κάθε συνελικτικό επίπεδο που επεξεργάζεται δεδομένα για λογαριασμό του αποκωδικοποιητή ακολουθείται από ένα επίπεδο απαλοιφής της γραμμικότητας. Όλα τα υπόλοιπα επίπεδα με απαλοιφής της γραμμικότητας, με εξαίρεση το καταληκτικό, χρησιμοποιούν τη συνάρτηση της ανορθωμένης γραμμικής μονάδας (ReLU). Το τελικό επίπεδο απαλοιφής της γραμμικότητας, το οποίο έπεται του καταληκτικού συνελικτικού επίπεδου, όπου υλοποιείται και τη πρόβλεψη των τιμών του χάρτη οπτικού ενδιαφέροντος $\tilde{S}$RGB-D, χρησιμοποιεί ως συνάρτηση ενεργοποίησης τη σιγμοειδή.

Η συνάρτηση απώλειας που χρησιμοποιήθηκε κατά τη διαδικασία της εκπαίδευσης του δικτύου είναι εκείνη της διασταυρωμένης εντροπίας, η οποία εκτιμάται σύμφωνα με τη σχέση (4.15)

$$\mathcal{L}(S,G) = -\frac{1}{N}\sum_{i=1}^{N}(g_i \log s_i + (1-g_i)\log(1-s_i)), \quad s_i \in S \text{ και } g_i \in G \tag{4.15}$$

όπου S ο χάρτης σημαντικών αντικειμένων όπως αυτός εκτιμάται από τον συνελικτικό αυτοκωδικοποιητή και G η βάση αλήθειας που φέρει την πληροφορία των σημαντικών αντικειμένων της εκάστοτε οπτικής αναπαράστασης.

Το δίκτυο εκπαιδεύτηκε με τη χρήση του αλγορίθμου Adam και ο βαθμός εκμάθησης αρχικοποιήθηκε στη τιμή 1e-4. Για την αρχικοποίηση των παραμέτρων του δικτύου ακολουθήθηκε η μέθοδος των K. He et al (K. He, Zhang, Ren, & Sun, 2015) την οποία άλλωστε εφαρμόζουν και οι συγγραφείς της εργασίας του *D³Net* (Fan, Lin, et al., 2020).

4.4. Επιδιορθωτικός Μηχανισμός

Ο επιδιορθωτικός μηχανισμός του μοντέλου αξιοποιεί την παραμετρική συνάρτηση μοναδιαίου βήματος, η οποία ορίζεται ως εξής:

$$H_p(x) = \begin{cases} 0, & x < a \\ x, & x \geq a \end{cases} \tag{4.16}$$

όπου, η μεταβλητή a υποδηλώνει την εκπαιδεύσιμη παράμετρο με πεδίο ορισμού το διάστημα [0, 1]. Στη συγκεκριμένη περίπτωση οι τιμές της μεταβλητής x αντιστοιχούν στις τιμές των εικονοστοιχείων του χάρτη οπτικού ενδιαφέροντος $\tilde{S}$RGB-D, οι οποίες εκτιμήθηκαν από την αρχιτεκτονική του συνελικτικού αυτοκωδικοποιητή.

Ο επιδιορθωτικός μηχανισμός χρησιμοποιεί τη συνάρτηση H_p, με συνελικτικό τρόπο λαμβάνοντας υπόψη τη τοπική χωρική περιοχή διαστάσεων $m \times n$, που περιβάλλει το εκάστοτε εικονοστοιχείο (i, j) του χάρτη οπτικού ενδιαφέροντος $\tilde{S}$RGB-D. Η σχέση (4.17) αποτυπώνει τον προαναφερόμενο τρόπο λειτουργίας του επιδιορθωτικού μηχανισμού.

$$S_{RGB-D}(i,j) = H_p\left(\frac{1}{m*n}\sum_{k=-a}^{a}\sum_{l=-b}^{b}\tilde{S}_{RGB-D}(i-k, j-l)\right) \quad \textbf{(4.17)}$$

όπου S_{RGB-D} ο τελικός χάρτης οπτικού ενδιαφέροντος όπως διαμορφώθηκε από τον επιδιορθωτικό μηχανισμό.

5. Αποτελέσματα

5.1. Σύνολα Δεδομένων

Η αξιολόγηση της προτεινόμενης μεθοδολογίας πραγματοποιήθηκε βάσει γνωστών RGB-D συνόλων δεδομένων ειδικά σχεδιασμένων για την ανίχνευση σημαντικών αντικειμένων. Πρόκειται για τα σύνολα δεδομένων: NLPR (Peng et al., 2014), NJU2K (Ju et al., 2014)και STEREO (Niu et al., 2012).

Το σύνολο δεδομένων NLPR περιλαμβάνει 1000 ζεύγη RGB-D εικόνων σε ποικίλα σενάρια εσωτερικών και εξωτερικών χώρων. Οι χάρτες βάθους έχουν διαμορφωθεί από τον αισθητήρα βάθους Microsoft Kinect.

Το σύνολο δεδομένων NJU2K συντίθεται από περίπου 2000 εικόνες. Οι εικόνες του συνόλου δεδομένων προέρχονται από το τρισδιάστατες ταινίες, από το Internet ή έχουν ληφθεί από φωτογράφους. Οι χάρτες βάθους του εν λόγω συνόλου δεδομένων έχουν ανακτηθεί από τη στερεοσκοπική κάμερα Fuji W3 stereo.

Το σύνολο δεδομένων STEREO αποτελείται από 1000 RGB εικόνες και τους αντίστοιχους χάρτες βάθους τους. Οι χάρτες βάθους του εν λόγω συνόλου δεδομένων έχουν ανακτηθεί από στερεοσκοπικές κάμερες. Όλες οι εικόνες (χρωματικές και βάθους) έχουν συλλεχθεί από διαδικτυακούς ιστότοπους (Flickr, NVIDIA 3D Vision Live, Stereoscopic Image Gallery).

Χάρτες βάθους για τις χρωματικές εικόνες και των τριών συνόλων δεδομένων εκτιμήθηκαν σύμφωνα με τις προσεγγίσεις της ενότητας (4.2). Προκειμένου τα αποτελέσματα της προτεινόμενης μεθοδολογίας να είναι συγκρίσιμα με εκείνα των RGB-D μεθοδολογιών ανίχνευσης σημαντικών αντικειμένων που παρουσιάστηκαν αναλυτικά στην ενότητα (3.4.3.3) υιοθετήθηκε ο διαμοιρασμός δεδομένων που προτείνει ο Fan (Fan, Lin, et al., 2020). Συγκεκριμένα, 700 εικόνες από το σύνολο δεδομένων NLPR και 1485 εικόνες από το σύνολο δεδομένων NJU2K επιλέχθηκαν για να κατασκευάσουν το σύνολα εκπαίδευσης και επικύρωσης (training and validation datasets) του δικτύου που πραγματοποιεί την ανίχνευση των σημαντικών αντικειμένων. Οι εναπομείνασες 300 και 500 εικόνες από τα παραπάνω σύνολα δεδομένων καθώς και το σύνολο των εικόνων STEREO συνέθεσαν το σύνολο ελέγχου (test dataset).

Οι εικόνες που χρησιμοποιήθηκαν για την εκπαίδευση του ανιχνευτή σημαντικών αντικείμενων είχαν ανάλυση 224x224. Επιπλέον, πραγματοποιήθηκε επαύξηση του συνόλου δεδομένων με την εφαρμογή του μετασχηματισμού της οριζόντιας περιστροφής.

5.2. Μετρικές Αξιολόγησης

Για την ποσοτική αξιολόγηση των αποτελεσμάτων της προτεινόμενης μεθοδολογίας χρησιμοποιήθηκαν τέσσερις μετρικές. Πρόκειται για το μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE) , και τα μεγέθη F-measure (Fβ), E-measure (Eξ) (Fan et al., 2018) και S-measure (Sa) (Fan, Cheng, Liu, Li, & Borji, 2017). Ακολούθως, παρατίθεται ο τρόπος με τον οποίο υπολογίζονται οι παραπάνω μετρικές.

Το μέσο απόλυτο σφάλμα (MAE), παρέχει μια ποσοτικοποίηση της απόκλισης μεταξύ του εκτιμώμενου χάρτη σημαντικότητας και του χάρτη σημαντικότητας που αντιστοιχεί στη βάση αλήθειας μιας δοθείσης οπτικής αναπαράστασης. Αν συμβολίσουμε ως G το δυαδικό χάρτη σημαντικότητας που συνιστά τη βάση αλήθειας και ως S τον χάρτη σημαντικότητας ο οποίος περιέχει εκτιμώμενες τιμές σημαντικότητας για κάθε εικονοστοιχείο τότε, το μέσο απόλυτο σφάλμα υπολογίζεται ως εξής:

$$MAE(S,G) = \frac{1}{W \cdot H} \sum_{i=1}^{W} \sum_{j=1}^{H} \left| S_{ij} - G_{ij} \right| \qquad \textbf{(5.1)}$$

όπου W, H οι διαστάσεις του χάρτη σημαντικότητας.

Το μέσο απόλυτο σφάλμα λαμβάνει τιμές στο διάστημα [0,1]. Γενικά μπορούμε να παρατηρήσουμε ότι η τιμή του μέσου απόλυτου σφάλματος αυξάνει αναλογικά με το μέγεθος των αντικειμένων μιας οπτικής αναπαράστασης. Για το λόγο αυτό περιοχές με μικρότερες τιμές σφαλμάτων μπορούν να θεωρηθούν περιοχές που αντιστοιχούν σε μικρά αντικείμενα, ενώ συνήθως μεγαλύτερες τιμές σφαλμάτων αντιστοιχίζονται σε μεγαλύτερα αντικείμενα. Παρά το γεγονός ότι το μέσο απόλυτο σφάλμα παρέχει μια άμεση εκτίμηση αναφορικά με το σφάλμα της εκτίμησης, αγνοεί πλήρως τις χωρικές συσχετίσεις των εσφαλμένων προβλέψεων.

Από την άλλη πλευρά, η μετρική F-measure (F_β) μπορεί να θεωρηθεί ένα μέτρο αξιολόγησης της ομοιότητας μεταξύ του εκτιμώμενου χάρτη σημαντικότητας και της βάσης αλήθειας, το οποίο λαμβάνει υπόψη το σχήμα των εντοπισμένων περιοχών. Το μέτρο F_β, αποτελεί τον αρμονικό μέσο όρο της ακρίβειας (Precision) και της ανάκλησης (Recall) και εκτιμάται σύμφωνα με τη σχέση (5.2).

$$F_\beta = \frac{(1+\beta^2) \cdot Presicion \cdot Recall}{\beta^2 \cdot Presicion + Recall} \qquad \textbf{(5.2)}$$

Στις συγκρίσεις που πραγματοποιήθηκαν στη παρούσα εργασία η τιμή της παραμέτρου β^2 τέθηκε ίση με 0.3 δίνοντας μεγαλύτερο βάρος στην ακρίβεια έναντι της ανάκλησης. Η τιμή

του μέτρου F_β, υπολογίστηκε για ένα πλήθος διαφορετικών τιμών κατωφλιών στο διάστημα [0,255] και η μέγιστη τιμή που έλαβε το μέτρο αποτέλεσε και τη τελική τιμή F_β.

Πλήθος μελετών συμπεριφορικής όρασης (behavioral vision) καταδεικνύουν το γεγονός ότι το ανθρώπινο οπτικό σύστημα παρουσιάζει υψηλή ευαισθησία στη δομική πληροφορία των οπτικών ερεθισμάτων που δέχεται. Η μετρική S-measure (Sa) , σχεδιάστηκε αποκλειστικά για την αξιολόγηση των προβλεπόμενων χαρτών σημαντικότητας, προκειμένου να λαμβάνει υπόψη τη δομική ομοιότητα μεταξύ των εκτιμώμενων χαρτών και των χαρτών που αντιστοιχούν στη βάση αλήθειας. Το μέτρο Sa, αξιολογεί ταυτόχρονα τόσο τη δομική ομοιότητα των αντικειμένων (Object-aware Structural Similarity) – S_r, όσο και ολόκληρων περιοχών (Region-aware Structural Similarity)- S_o σύμφωνα με την ακόλουθη σχέση:

$$S_a = b \cdot S_o + (1 - b) \cdot S_r \quad \textbf{(5.3)}$$

όπου b η παράμετρος ισορροπίας η οποία λαμβάνει τιμές στο διάστημα [0,1]. Για τους σκοπούς της επερχόμενης αξιολόγησης των αποτελεσμάτων η τιμή της παραμέτρου b τέθηκε ίση με 0.5.

Οι σχέσεις (5.4) και (5.5) παρουσιάζουν αντίστοιχα τον τρόπο υπολογισμού της δομικής ομοιότητας των αντικειμένων (S_r), και των περιοχών (S_o).

$$S_r = \sum_{k=1}^{K} w_k \cdot SSIM\,(k) \quad \textbf{(5.4)}$$

με K το πλήθος των περιοχών στις οποίες υποδιαιρούνται οι προς σύγκριση χάρτες σημαντικότητας, w_k ένας παράγοντας στάθμισης που ανατίθεται στη κάθε περιοχή ευθέως ανάλογος του ποσοστού των εικοστοιχείων που καλύπτει η περιοχή k στο προσκήνιο του χάρτη της βάσης αλήθειας. Ο όρος $SSIM(k)$ αφορά το μέτρο της δομικής ομοιότητας των υπό σύγκριση περιοχών και υπολογίζεται όπως έχει οριστεί από τον Wang και του συνεργάτες του (Z. Wang et al., 2003) .

$$S_o = \mu \cdot O_{FG} + (1 - \mu) \cdot O_{BG} \quad \textbf{(5.5)}$$

όπου μ ο λόγος της περιοχής του προσκηνίου στο χάρτη της βάσης αλήθειας, ενώ οι τιμές των O_{BG}, O_{FG} εκτιμώνται από τις σχέσεις (5.6) και (5.7).

$$O_{FG} = \frac{2\bar{x}_{FG}}{(\bar{x}_{FG})^2 + 1 + 2\lambda \cdot \sigma_{x_{FG}}} \quad \textbf{(5.6)}$$

με $\bar{x}_{FG}$ $και$ $\sigma_{x_{FG}}$ να αντιστοιχούν στη μέση τιμή και την τυπική απόκλιση των εικονοστοιχείων που εντοπίζονται στο προσκήνιο του εκτιμώμενου χάρτη σημαντικότητας.

Ανάλογα,

$$O_{BG} = \frac{2\bar{x}_{BG}}{(\bar{x}_{BG})^2 + 1 + 2\lambda \cdot \sigma_{x_{BG}}} \quad \textbf{(5.7)}$$

με, $\bar{x}_{BG}$ *και* $\sigma_{x_{BG}}$ να υποδεικνύουν τη μέση τιμή και την τυπική απόκλιση των εικονοστοιχείων που ανήκουν στο παρασκήνιο του εκτιμώμενου χάρτη σημαντικότητας.

Η τιμή της παραμέτρου λ στις σχέσεις (5.6), (5.7) ορίστηκε 0.5 σύμφωνα με τις υποδείξεις των συγγραφέων.

Οι παραπάνω μετρικές εντοπίζουν σφάλματα πρόβλεψης σε χάρτες σημαντικότητας είτε σε επίπεδο εικονοστοιχείου (MAE, Fβ) είτε με επίπεδο δομικών χαρακτηριστικών (Sα). Το μέτρο αξιολόγησης E_ξ (E-measure), αφορμώμενο από την παρατήρηση μελετών γνωστικής όρασης ότι το ανθρώπινο οπτικό σύστημα είναι εξίσου ευαίσθητο τόσο στη τοπική όσο και στη συνολική πληροφορία μιας παρατηρούμενης οπτικής σκηνής, έχει σχεδιαστεί προκειμένου να εντοπίζει σφάλματα τόσο σε τοπικό επίπεδο όσο και σε επίπεδο εικόνας. Η μετρική αξιολόγησης E_ξ η οποία έχει προταθεί για την σύγκριση δυαδικών χαρτών σημαντικότητας συστήνει τη χρήση του πίνακα ενισχυμένης ευθυγράμμισης (Enhanced Alignment Matrix) ο οποίος καταγράφει στατιστικά σε επίπεδο εικόνας και πληροφορία ταιριάσματος σε επίπεδο εικονοστοιχείων. Ο υπολογισμός του μέτρου αξιολόγησης E_ξ, πραγματοποιείται σύμφωνα με την παρακάτω σχέση:

$$E_\xi = \frac{1}{w \cdot h} \sum_{x=1}^{w} \sum_{y=1}^{h} \varphi_{F_M}(x, y) \quad \textbf{(5.8)}$$

όπου φ_{F_M} ο πίνακας ενισχυμένης ευθυγράμμισης, ο οποίος εκτιμάται σύμφωνα με τις σχέσεις (5.9), (5.10), (5.11).

$$\varphi_{F_M} = \mathrm{f}(\xi_{\mathrm{FM}}) \quad \textbf{(5.9)}$$

όπου f(·) μια κυρτή συνάρτηση - στη συγκεκριμένη περίπτωση έγινε χρήση της τετραγωνικής συνάρτησης που υποδεικνύει η σχέση (5.10) και ξ_{FM} ο πίνακας ενισχυμένης ευθυγράμμισης, οποίος ορίζεται στη σχέση (5.11).

$$\mathrm{f}(x) = \frac{1}{4}(1 + x)^2 \quad \textbf{(5.10)}$$

$$\xi_{\mathrm{FM}} = \frac{2\varphi_{GT} \odot \varphi_{FM}}{\varphi_{GT} \odot \varphi_{GT} + \varphi_{FM} \odot \varphi_{FM}} \quad \textbf{(5.11)}$$

όπου Θ το γινόμενο Hadamard. Ο πίνακας φ_I περιλαμβάνει την απόσταση κάθε εικονοστοιχείου της αναπαράστασης I από τη μέση τιμή της. Ο όρος φ_I υπολογίζεται σύμφωνα με την επόμενη σχέση.

$$\varphi_I = I - \mu_I \mathbb{A}, \qquad \text{με I} \in \{\text{GT}, \text{FM}\} \tag{5.12}$$

με μ_Iτη μέση τιμή των εικονοστοιχείων της αναπαράστασης I και $\mathbb{A}$ ένας πίνακας ίδιων διαστάσεων με την αναπαράσταση I του οποίου όλες οι τιμές ισούνται με 1. Όταν $I = \text{GT}$ τότε ο υπολογισμός αφορά τον δυαδικό χάρτη σημαντικότητας που εικονίζει τη βάση αλήθειας και όταν $I = \text{FM}$ τότε υπολογισμός αφορά τον προβλεπόμενο χάρτη σημαντικότητας ο οποίος προηγουμένως έχει υποστεί προσαρμοστική κατωφλίωση προκειμένου να ανακτηθεί η δυαδική μορφή του.

Η αξιολόγηση της απόδοσης του προτεινόμενου μοντέλου ανίχνευσης σημαντικών αντικειμένων πραγματοποιείται ως εξής: Συμβολίζοντας ως I_j^k την j εικόνα που ανήκει στο σύνολο δεδομένων ελέγχου D_k με $k \in \{NLPR, NJU2K, STEREO\}$ και ως $\bar{\zeta}(I_j^k)$ το αποτέλεσμα από εφαρμογή του μέτρου αξιολόγησης $\zeta \in \{MAE, F_\beta, S_\alpha, E_\xi\}$ στην εικόνα I_j^k υπολογίζουμε τις παρακάτω ποσότητες για κάθε σύνολο δεδομένων ελέγχου D_k :

$$M_\zeta(D_k) = \frac{1}{|D_k|}\sum_j \bar{\zeta}(I_j^k) \tag{5.13}$$

με $M_\zeta(D_k)$ να δηλώνει τη μέση τιμή του μέτρου ζ στις εικόνες του συνόλου δεδομένων ελέγχου D_k ενώ ο όρος $|D_k|$ αναφέρεται στο συνολικό πλήθος των εικόνων που περιλαμβάνει το σύνολο δεδομένων D_k.

$$\sigma_\zeta(D_k) = \sqrt{\frac{1}{|D_k|}\sum_j (\bar{\zeta}(I_j^k) - M_\zeta(D_k))^2} \tag{5.14}$$

με $\sigma_\zeta(D_k)$ να δηλώνει τη τυπική απόκλιση του μέτρου ζ στις εικόνες του συνόλου δεδομένων ελέγχου D_k . Ο όρος $|D_k|$ αφορά το συνολικό πλήθος των εικόνων που περιέχονται στο σύνολο δεδομένων D_k. Ο όρος $M_\zeta(D_k)$ είναι η μέση τιμή του μέτρου ζ για τις εικόνες του συνόλου δεδομένων D_k, η οποία εκτιμήθηκε σύμφωνα με τη σχέση (5.13)

Η μέση τιμή και η τυπική απόκλιση των στατιστικών μέτρων MAE, F_β, S_α και E_ξ ανά σύνολο δεδομένων ελέγχου για καθεμία από τις πτυχές της προτεινόμενης προσέγγισης παρουσιάζονται στην ακόλουθη υποενότητα.

5.3. Αποτελέσματα

5.3.1. Ποσοτική σύγκριση των αποτελεσμάτων

Στους πίνακες (5.1) – (5.4) εικονίζονται τα αποτελέσματα της προτεινόμενης μεθοδολογίας για την ανίχνευση σημαντικών αντικειμένων σε RGB εικόνες όταν αυτή υποβοηθείται από την πληροφορία του βάθους η οποία έχει ανακτηθεί προσεγγιστικά από συνελικτικά νευρωνικά δίκτυα σύμφωνα με τις μεθόδους των υποενοτήτων (4.2.1)-(4.2.3). Τα προηγούμενα αποτελέσματα αντιπαραβάλλονται με αποτελέσματα που έχουν προκύψει όταν χρησιμοποιούνται οι χάρτες βάθους των RGB-D συνόλων δεδομένων, οι οποίοι έχουν ανακτηθεί από συστήματα στερεοσκοπικών καμερών ή αισθητήρες βάθους.

Ειδικότερα, ο πίνακας (5.1) συγκεντρώνει τα αποτελέσματα όπως διαμορφώθηκαν από την εφαρμογή της προτεινόμενης μεθοδολογίας ανίχνευσης σημαντικών αντικειμένων, με τη χρήση των χαρτών βάθους που ανακτήθηκαν από τα δίκτυα *DenseDepth* (προσέγγιση που παρουσιάστηκε στην υποενότητα 4.2.1) και *MonoDepth2* (προσέγγιση που παρουσιάστηκε στην υποενότητα 4.2.2). Τα παραπάνω αποτελέσματα αντιπαραβάλλονται με τα αποτελέσματα που ανακτήθηκαν από την εφαρμογή της μεθοδολογίας ανίχνευσης σημαντικών αντικειμένων της υποενότητας (4.3) με τη χρήση των χαρτών βάθους που φέρουν τα RGB-D σύνολα δεδομένων και οι οποίοι έχουν υπολογιστεί με την βοήθεια στερεοσκοπικών καμερών ή αισθητήρων βάθους. Σε κάθε περίπτωση, οι τιμές των αποτελεσμάτων αναφέρονται στην αξιολόγηση που πραγματοποιήθηκε χωρίς την εφαρμογή του επιδιορθωτικού μηχανισμού.

Σύμφωνα με τα αποτελέσματα του πίνακα (5.1) για τους χάρτες βάθους που έχουν ανακτηθεί από το δίκτυο *DenseDepth* (υποενότητα 4.2.1) οι μετρικές Μέσο Απόλυτο Σφάλμα και E-measure επιτυγχάνουν απόκλιση της τάξης του ~10-3, σε αντίθεση με τις μετρικές F-measure και S-measure που παρουσιάζουν αντίστοιχα διαφορά της τάξης του ~10-2 και ~10-1. Ανάλογα, για τους χάρτες βάθους που προέρχονται από το δίκτυο *MonoDepth2* (υποενότητα 4.2.2) με εξαίρεση τη τάξη διαφοράς του E-measure η οποία είναι εκείνη του ~10-2 , η τάξη απόκλισης που σημειώνουν οι υπόλοιπες μετρικές παραμένει η ίδια με τους χάρτες βάθους του δικτύου *DenseDepth*.

Πίνακας 5.1 Αποτελέσματα της προτεινόμενης μεθοδολογίας ανίχνευσης σημαντικών αντικειμένων, χωρίς την εφαρμογή του επιδιορθωτικού μηχανισμού με χάρτες βάθους από τα συνελικτικά δίκτυα DenseDepth και MonoDepth2 καθώς και από τους χάρτες βάθους των RGB-D συνόλων δεδομένων.

Σύνολο Δεδομένων	**Μέτρο Αξιολόγησης**	**Πληροφορία του βάθους**					
		DenseDepth		***Monodepth2***		***Κάμερα/Αισθητήρας***	
		Μέση τιμή	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*
NJU2K	$MAE\downarrow$[1]	0.051	± 0.060	0.051	± 0.058	0.048	± 0.063
	$F_\beta\uparrow$[2]	0.861	± 0.186	0.860	± 0.178	0.863	± 0.195
	$S_a\uparrow$	0.892	± 0.111	0.891	± 0.111	0.898	± 0.117
	$E_\xi\uparrow$	0.908	± 0.154	0.908	± 0.151	0.913	± 0.153
NLPR	$MAE\downarrow$	0.032	± 0.045	0.034	± 0.049	0.031	± 0.046
	$F_\beta\uparrow$	0.855	± 0.157	0.844	± 0.172	0.858	± 0.156
	$S_a\uparrow$	0.909	± 0.106	0.904	± 0.117	0.910	± 0.104
	$E_\xi\uparrow$	0.939	± 0.112	0.936	± 0.113	0.941	± 0.108
STEREO	$MAE\downarrow$	0.050	± 0.056	0.049	± 0.053	0.046	± 0.049
	$F_\beta\uparrow$	0.857	± 0.170	0.855	± 0.173	0.859	± 0.169
	$S_a\uparrow$	0.894	± 0.104	0.893	± 0.104	0.898	± 0.098
	$E_\xi\uparrow$	0.915	± 0.140	0.914	± 0.142	0.919	± 0.135
Συνολικά	$\boldsymbol{MAE\downarrow}$	**0.044**	**± 0.053**	**0.044**	**± 0.053**	**0.042**	**± 0.052**
	$\boldsymbol{F_\beta\uparrow}$	**0.857**	**± 0.171**	**0.853**	**± 0.174**	**0.860**	**± 0.173**
	$\boldsymbol{S_a\uparrow}$	**0.898**	**± 0.107**	**0.896**	**± 0.110**	**0.902**	**± 0.106**
	$\boldsymbol{E_\xi\uparrow}$	**0.920**	**± 0.135**	**0.919**	**± 0.135**	**0.924**	**± 0.132**

Πριν από την αξιολόγηση της συνεισφοράς του επιδιορθωτικού μηχανισμού στην προτεινόμενη μεθοδολογία θα εξετάσουμε τα αποτελέσματα που προκύπτουν έπειτα από τη εφαρμογή του επιδιορθωτικού μηχανισμού. Ο πίνακας (5.2) περιλαμβάνει τα αποτελέσματα που αφορούν τους χάρτες σημαντικών αντικειμένων έπειτα από την διαδικασία της επιδιόρθωσης, η οποία εφαρμόζεται όπως υποδεικνύει η υποενότητα (4.4). Η τιμή της παραμέτρου α που χρησιμοποιεί ο επιδιορθωτικός μηχανισμός όπως υποδεικνύει η σχέση (4.16) προσδιορίστηκε πειραματικά και ορίστηκε ίση με 0,376.

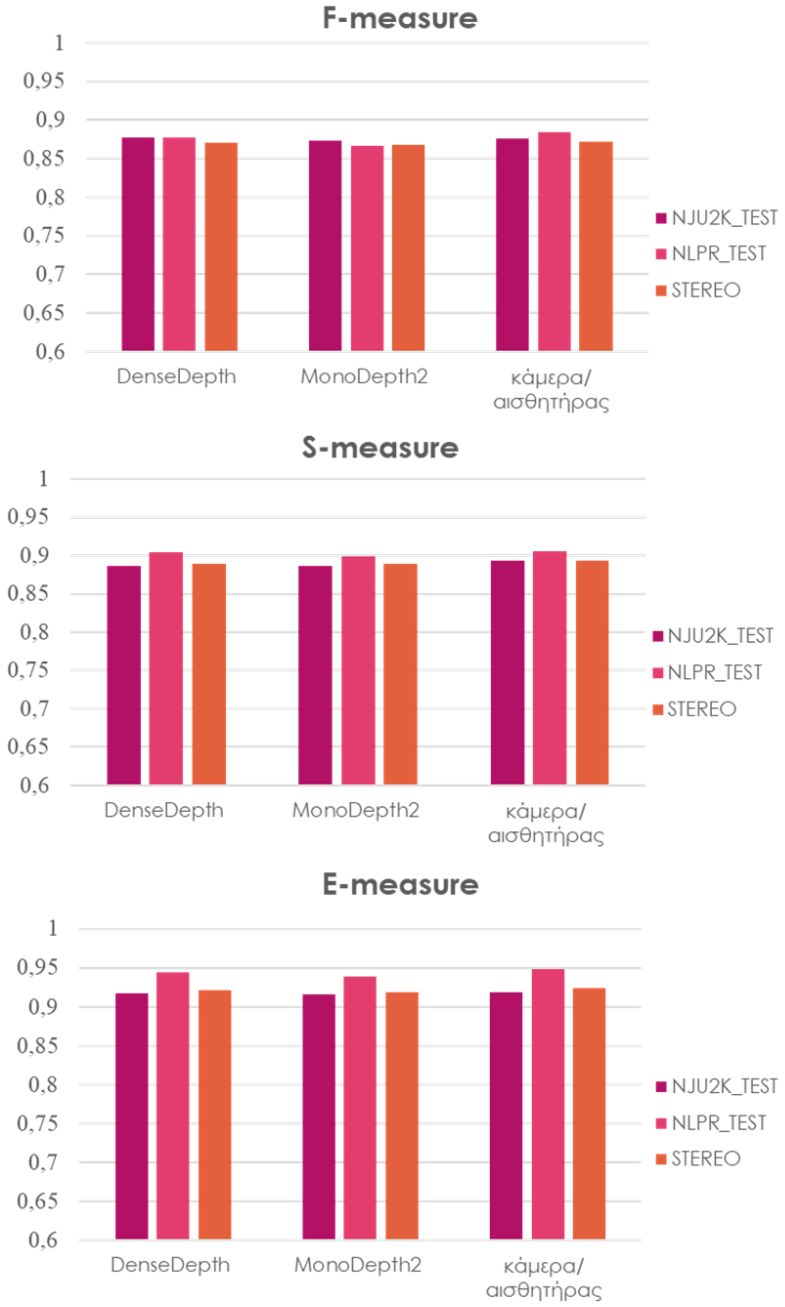
[1] Το σύμβολο ↓υποδηλώνει ότι όσο χαμηλότερη είναι η τιμή τόσο το αποτέλεσμα προσεγγίζει τη βάση αλήθειας

[2] Το σύμβολο ↑υποδηλώνει ότι όσο υψηλότερη είναι η τιμή τόσο το αποτέλεσμα προσεγγίζει τη βάση αλήθειας

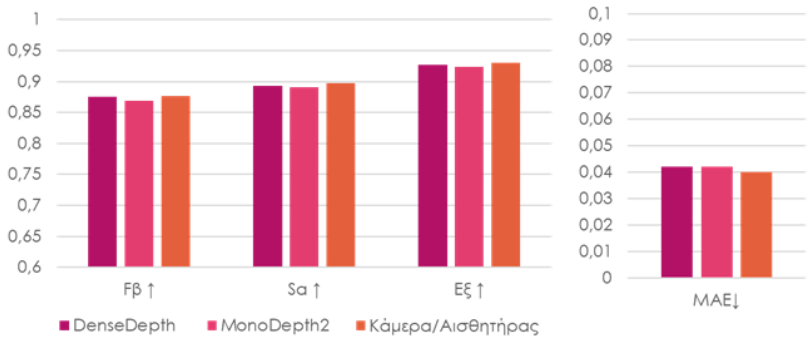
Πίνακας 5.2 Αποτελέσματα της προτεινόμενης μεθοδολογίας ανίχνευσης σημαντικών αντικειμένων, με την εφαρμογή του επιδιορθωτικού μηχανισμού. Εξετάζεται η ανίχνευση σημαντικών αντικειμένων με χάρτες βάθους από τα συνελικτικά δίκτυα DenseDepth και MonoDepth2 καθώς και από τους χάρτες βάθους των RGB-D συνόλων δεδομένων.

Σύνολο Δεδομένων	**Μέτρο Αξιολόγησης**	**Πληροφορία του βάθους**					
		DenseDepth		***MonoDepth2***		***Κάμερα/Αισθητήρας***	
		Μέση τιμή	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*
NJU2K	*MAE*↓	0.049 ±	0.06	0.048 ±	0.059	0.046 ±	0.064
	F_β ↑	0.877 ±	0.174	0.873 ±	0.172	0.876 ±	0.188
	S_a ↑	0.886 ±	0.114	0.886 ±	0.113	0.893 ±	0.112
	E_ξ ↑	0.917 ±	0.146	0.916 ±	0.145	0.918 ±	0.153
NLPR	*MAE*↓	0.030 ±	0.045	0.032 ±	0.048	0.030 ±	0.046
	F_β ↑	0.878 ±	0.16	0.867 ±	0.181	0.884 ±	0.149
	S_a ↑	0.904 ±	0.111	0.899 ±	0.121	0.905 ±	0.107
	E_ξ ↑	0.944 ±	0.117	0.939 ±	0.127	0.948 ±	0.105
STEREO	*MAE*↓	0.047 ±	0.056	0.046 ±	0.053	0.044 ±	0.049
	F_β ↑	0.871 ±	0.163	0.868 ±	0.167	0.872 ±	0.163
	S_a ↑	0.889 ±	0.107	0.889 ±	0.106	0.893 ±	0.101
	E_ξ ↑	0.921 ±	0.137	0.919 ±	0.141	0.924 ±	0.134
Συνολικά	***MAE*↓**	**0.042 ±**	**0.053**	**0.042 ±**	**0.053**	**0.040 ±**	**0.053**
	F_β ↑	**0.875 ±**	**0.165**	**0.869 ±**	**0.173**	**0.877 ±**	**0.166**
	S_a ↑	**0.893 ±**	**0.110**	**0.891 ±**	**0.113**	**0.897 ±**	**0.109**
	E_ξ ↑	**0.927 ±**	**0.133**	**0.924 ±**	**0.137**	**0.930 ±**	**0.130**

Ειδικότερα, από την παρατήρηση των αποτελεσμάτων του πίνακα (5.2) συνάγονται οι ακόλουθες παρατηρήσεις: για τους χάρτες βάθους που έχουν εκτιμηθεί από το δίκτυο *DenseDepth* οι μετρικές Μέσο Απόλυτο Σφάλμα, F-measure και S-measure σημειώνουν απόκλιση της τάξης του ~10-3 ενώ η διαφορά του E-measure είναι της τάξης του ~10-2. Ομοίως, όταν η ανίχνευση των σημαντικών αντικειμένων διεξάγεται με τη χρήση χαρτών βάθους υπολογισμένων από το συνελικτικό δίκτυο *MonoDepth2* τα μέτρα Μέσο Απόλυτο Σφάλμα και S-measure επιτυγχάνουν διαφορά της τάξης του ~10-3, ενώ οι εναπομείνασες μετρικές παρουσιάζουν απόκλιση της τάξης του ~10-2.

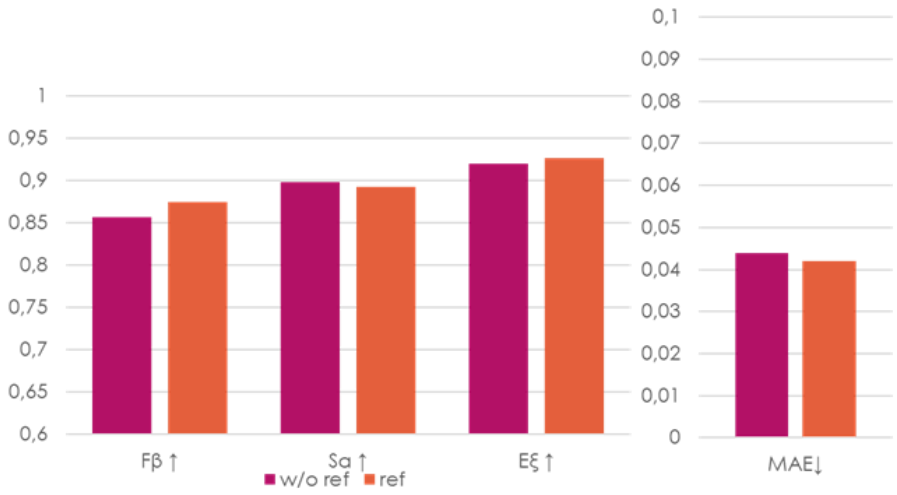


Εικόνα 5.1 Γραφική αναπαράσταση των αποτελεσμάτων του πίνακα 5.2

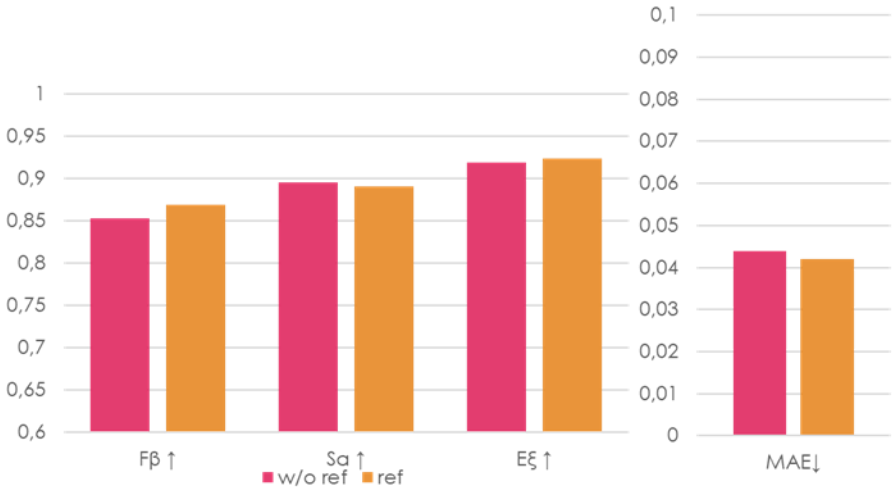


Εικόνα 5.2 Γραφήματα αντίστοιχα του πίνακα (5.2) – Οι τιμές αφορούν τα αποτελέσματα για το σύνολο των δεδομένων εκπαίδευσης (NLPR, NJU2K, STEREO)

Πράγματι, η συνεισφορά του επιδιορθωτικού μηχανισμού στην παρούσα μεθοδολογία επιβεβαιώνεται από τα αποτελέσματα του πίνακα (5.2). Διαπιστώνεται, ότι ο επιδιορθωτικός μηχανισμός παρέχει βελτιωμένα αποτελέσματα για τις περισσότερες μετρικές (Μέσο Απόλυτο Σφάλμα, F-measure, E-measure), γεγονός το οποίο αποδεικνύεται και από τα γραφήματα των εικόνων (5.3) και (5.4).



Εικόνα 5.3 Συνεισφορά του επιδιορθωτικού μηχανισμού στην ανίχνευση σημαντικών αντικειμένων που πραγματοποιήθηκε με χάρτες βάθους από το συνελικτικό δίκτυο DenseDepth



Εικόνα 5.4 Συνεισφορά του επιδιορθωτικού μηχανισμού στην ανίχνευση σημαντικών αντικειμένων που πραγματοποιήθηκε με χάρτες βάθους από το συνελικτικό δίκτυο MonoDepth2

Γενικότερα, από τους πίνακες (5.1) και (5.2) παρατηρούμε ότι οι όλες οι τιμές που προκύπτουν από την αξιολόγηση των αποτελεσμάτων των χαρτών βάθους που ανακτώνται από το δίκτυο *MonoDepth2* είναι λίγο χαμηλότερες από τις τιμές που αντιστοιχούν στην αξιολόγηση με τους χάρτες βάθους που έχουν παραχθεί από το δίκτυο *DenseDepth*. Μια προφανής εξήγηση αυτού του λόγου αποτελεί το γεγονός ότι το δίκτυο *MonoDepth2* έχει εκπαιδευτεί ακολουθώντας έναν αυτό-επιβλεπόμενο μηχανισμό επιτήρησης, ο οποίος

αναλύθηκε σε βάθος στην υποενότητα (4.2.2), ενώ το δίκτυο *DenseDepth* εκπαιδεύτηκε με επιβλεπόμενο τρόπο (υποενότητα 4.2.1).

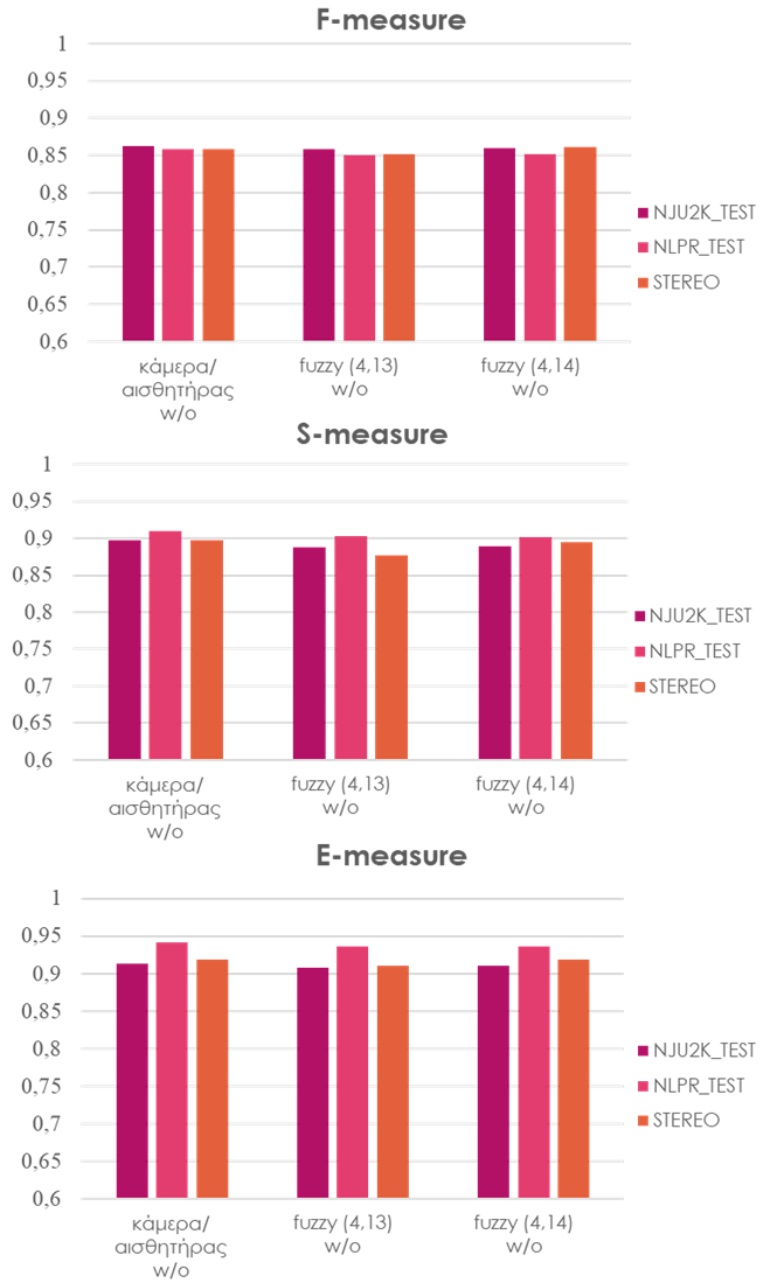
Όσον αφορά τις ασαφείς αναπαραστάσεις της πληροφορίας του βάθους όπως παρατηρούμε και από τα συγκεντρωτικά αποτελέσματα του πίνακα (5.3) επιτυγχάνουν ελάχιστα χαμηλότερες τιμές συγκριτικά με τους χάρτες βάθους του δικτύου *MonoDepth2*. Παρ' όλα αυτά, οι τιμές αυτές επιδέχονται σύγκριση με εκείνες που προκύπτουν όταν οι χάρτες βάθους που αξιοποιούνται προέρχονται από κάμερες ή αισθητήρες. Λεπτομερέστερα, το Μέσο Απόλυτο Σφάλμα το F-measure και το S-measure επιτυγχάνουν αντίστοιχα διαφορά της τάξης του ~10-3, ~10-2 και ~10-1, ενώ η απόκλιση του E-measure κυμαίνεται μεταξύ ~10-2 και ~10-1 αναλόγως της προσέγγισης.

Εξετάζοντας προσεκτικά τα αποτελέσματα που συνοψίζονται στον πίνακα (5.3) συνάγεται ότι όταν οι ασαφείς χάρτες βάθους φέρουν λιγότερη πληροφορία, τότε η ανίχνευση των σημαντικών αντικειμένων διεξάγεται με μεγαλύτερη επιτυχία. Να υπενθυμίσουμε στο σημείο αυτό ότι η προσέγγιση fuzzy(4.13) χρησιμοποιεί και τα τρία κανάλια της ασαφούς πληροφορίας του βάθους, σε αντίθεση με την προσέγγιση fuzzy(4.14) η οποία αξιοποιεί μόνο δύο εκ των τριών καναλιών από την πληροφορία του βάθους. Η τελευταία προσέγγιση αγνοεί τη μακρινή πληροφορία του βάθους της εκάστοτε οπτικής σκηνής καθώς διατηρεί μόνο εκείνη που αναφέρεται σε κοντινές και ενδιάμεσες αποστάσεις από το σημείο παρατήρησης.

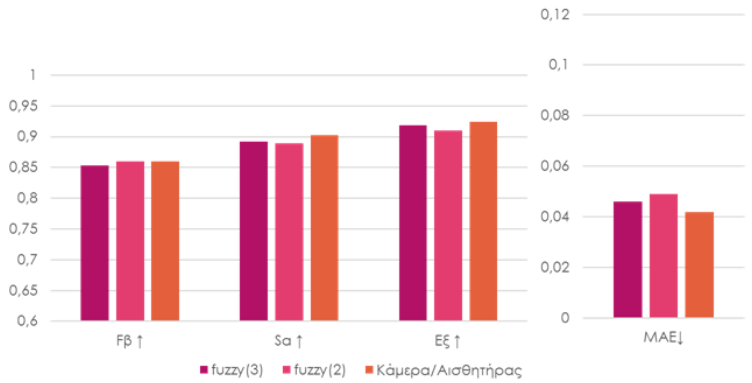
Ακόμα, οφείλουμε να επισημάνουμε ότι στην περίπτωση όπου η ανίχνευση των σημαντικών αντικειμένων πραγματοποιείται με τη χρήση χαρτών βάθους που ανακτήθηκαν από την εφαρμογή κανόνων ασαφούς λογικής στην πληροφορία του βάθους που παράχθηκε από το συνελικτικό δίκτυο *DenseDepth*, δεν έχει εφαρμοστεί ο επιδιορθωτικός μηχανισμός για την περαιτέρω βελτίωση των αποτελεσμάτων. Συνεπώς, και οι τιμές των αποτελεσμάτων που αναφέρονται στους χάρτες βάθους που φέρουν τα RGB-D σύνολα δεδομένων είναι αυτές που διαμορφώνονται χωρίς τη χρήση του επιδιορθωτικού μηχανισμού.

Πίνακας 5.3 Αποτελέσματα της προτεινόμενης μεθοδολογίας όπως διαμορφώθηκαν όταν οι χάρτες βάθους που χρησιμοποιήθηκαν ανακτήθηκαν σύμφωνα με τη προσέγγιση που προτείνεται στην υποενότητα (4.2.3).

Σύνολο Δεδομένων	**Μέτρο Αξιολόγησης**	**Πληροφορία του βάθους**					
		fuzzy (4.13)		***fuzzy (4.14)***		***Κάμερα/Αισθητήρας***	
		Μέση τιμή	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*
NJU2K	$MAE\downarrow$	0.053	0.060	0.049	0.061	0.048	0.063
	$F_\beta\uparrow$	0.859	0.187	0.86	0.186	0.863	0.195
	$S_a\uparrow$	0.888	0.114	0.889	0.118	0.898	0.117
	$E_\xi\uparrow$	0.908	0.148	0.910	0.156	0.913	0.153
NLPR	$MAE\downarrow$	0.033	0.048	0.032	0.044	0.031	0.046
	$F_\beta\uparrow$	0.850	0.167	0.851	0.164	0.858	0.156
	$S_a\uparrow$	0.903	0.110	0.902	0.117	0.910	0.104
	$E_\xi\uparrow$	0.936	0.114	0.936	0.117	0.941	0.108
STEREO	$MAE\downarrow$	0.052	0.054	0.046	0.051	0.046	0.049
	$F_\beta\uparrow$	0.852	0.178	0.861	0.169	0.859	0.169
	$S_a\uparrow$	0.877	0.108	0.895	0.104	0.898	0.098
	$E_\xi\uparrow$	0.911	0.145	0.919	0.141	0.919	0.135
Συνολικά	$\mathbf{MAE\downarrow}$	**0.046**	**0.054**	**0.049**	**0.061**	**0.042**	**0.052**
	$\mathbf{F_\beta\uparrow}$	**0.853**	**0.177**	**0.860**	**0.186**	**0.860**	**0.173**
	$\mathbf{S_a\uparrow}$	**0.892**	**0.110**	**0.889**	**0.118**	**0.902**	**0.106**
	$\mathbf{E_\xi\uparrow}$	**0.918**	**0.135**	**0.910**	**0.156**	**0.924**	**0.132**



Εικόνα 5. 5 Γραφική αναπαράσταση των αποτελεσμάτων του πίνακα (5.3)



Εικόνα 5.6 Γραφήματα αντίστοιχα του πίνακα (5.3) – Οι τιμές αφορούν τα αποτελέσματα για το σύνολο των δεδομένων εκπαίδευσης (NLPR, NJU2K, STEREO)

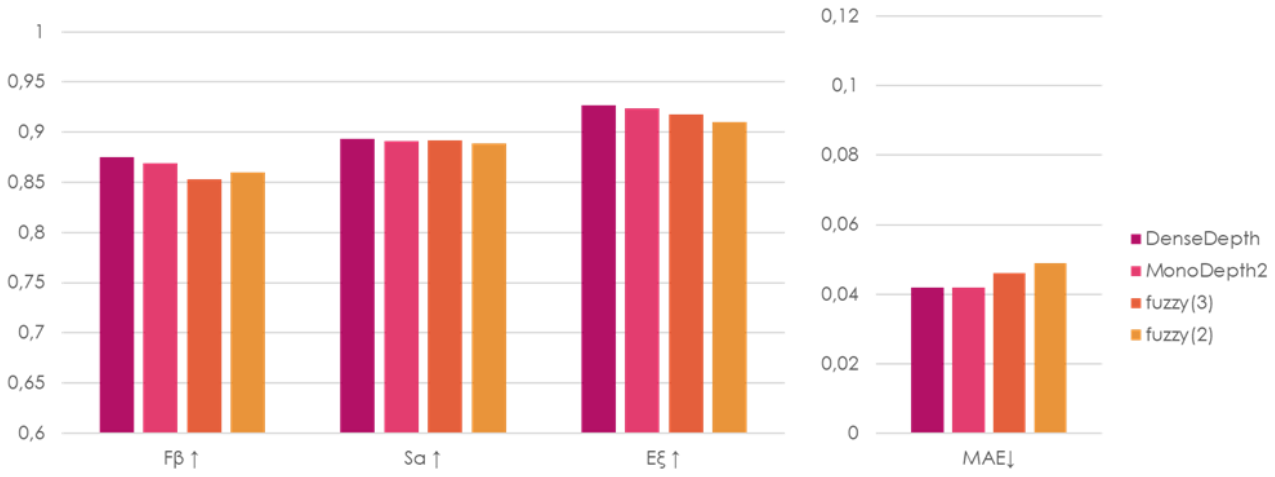
Στο σημείο αυτό , μπορούμε να ισχυριστούμε ότι οι εκτιμώμενες από συνελικτικά νευρωνικά δίκτυα αναπαραστάσεις της πληροφορίας του βάθους μιας οπτικής σκηνής, δύνανται να προσφέρουν συγκρίσιμα αποτελέσματα με αντίστοιχες αναπαραστάσεις οι οποίες έχουν εκτιμηθεί με τη βοήθεια στερεοσκοπικών καμερών ή αισθητήρων βάθους.

Ο παραπάνω ισχυρισμός συνάγεται από την παρατήρηση των αποτελεσμάτων που συγκεντρώνουν οι πίνακες (5.1) και (5.3). Συγκεκριμένα, παρατηρούμε ότι συνολικά για όλα τα σύνολα δεδομένων στα οποία επαληθεύτηκε η προτεινόμενη μεθοδολογία, όλες οι μετρικές (Μέσο Απόλυτο Σφάλμα, F-measure, S-measure, E-measure) παρουσιάζουν ελαφρώς υψηλότερες τιμές όταν χρησιμοποιείται η πληροφορία του βάθους από αισθητήρες ή κάμερες. Ωστόσο, οι αντίστοιχες τιμές για τις υπολογιστικές αναπαραστάσεις του βάθους παρουσιάζουν μικρή απόκλιση και σαφώς μπορούν να θεωρηθούν συγκρίσιμες.

Ο πίνακας (5.4) συγκεντρώνει τα αποτελέσματα της ανίχνευσης των σημαντικών αντικειμένων μιας δοθείσας οπτικής αναπαράστασης με την αξιοποίηση όλων των υπολογιστικών προσεγγίσεων χαρτών βάθους που διερευνήθηκαν στην παρούσα εργασία. Από την αντιπαραβολή των αποτελεσμάτων του πίνακα (5.4) διαπιστώνουμε ότι η διεξαγωγή της ανίχνευσης σημαντικών αντικειμένων σε συνδυασμό με τους χάρτες βάθους που υπολογίζει το συνελικτικό δίκτυο *DenseDepth* παρέχει τα καλύτερα αποτελέσματα έναντι των υπολοίπων προσεγγίσεων. Έπονται τα αποτελέσματα που αφορούν τους χάρτες βάθους του αυτό-επιβλεπόμενου συνελικτικού δικτύου *MonoDepth2* , τα οποία ακολουθούν οι προσεγγίσεις που αξιοποιούν τη ασαφή λογική συνδυαστικά με το συνελικτικό δίκτυο *DenseDepth*.

Πίνακας 5.4 Συγκεντρωτικά αποτελέσματα όλων των προσεγγίσεων που προτείνονται στην παρούσα εργασία.

Σύνολο Δεδομένων	**Μέτρο Αξιολόγησης**	**Πληροφορία του βάθους**							
		DenseDepth		***Monodepth2***		***fuzzy (4.13)***		***fuzzy (4.14)***	
		Μέση τιμή	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*
NJU2K	$MAE\downarrow$	0.049	± 0.06	0.048	± 0.059	0.053	± 0.060	0.049	± 0.061
	$F_\beta\uparrow$	0.877	± 0.174	0.873	± 0.172	0.859	± 0.187	0.86	± 0.186
	$S_a\uparrow$	0.886	± 0.114	0.886	± 0.113	0.888	± 0.114	0.889	± 0.118
	$E_\xi\uparrow$	0.917	± 0.146	0.916	± 0.145	0.908	± 0.148	0.910	± 0.156
NLPR	$MAE\downarrow$	0.030	± 0.045	0.032	± 0.048	0.033	± 0.048	0.032	± 0.044
	$F_\beta\uparrow$	0.878	± 0.16	0.867	± 0.181	0.850	± 0.167	0.851	± 0.164
	$S_a\uparrow$	0.904	± 0.111	0.899	± 0.121	0.903	± 0.110	0.902	± 0.117
	$E_\xi\uparrow$	0.944	± 0.117	0.939	± 0.127	0.936	± 0.114	0.936	± 0.117
STEREO	$MAE\downarrow$	0.047	± 0.056	0.046	± 0.053	0.052	± 0.054	0.046	± 0.051
	$F_\beta\uparrow$	0.871	± 0.163	0.868	± 0.167	0.852	± 0.178	0.861	± 0.169
	$S_a\uparrow$	0.889	± 0.107	0.889	± 0.106	0.877	± 0.108	0.895	± 0.104
	$E_\xi\uparrow$	0.921	± 0.137	0.919	± 0.141	0.911	± 0.145	0.919	± 0.141
Συνολικά	$\mathbf{MAE\downarrow}$	**0.042**	**± 0.053**	**0.042**	**± 0.053**	**0.046**	**± 0.054**	**0.049**	**± 0.061**
	$\mathbf{F_\beta\uparrow}$	**0.875**	**± 0.165**	**0.869**	**± 0.173**	**0.853**	**± 0.177**	**0.86**	**± 0.186**
	$\mathbf{S_a\uparrow}$	**0.893**	**± 0.110**	**0.891**	**± 0.113**	**0.892**	**± 0.110**	**0.889**	**± 0.118**
	$\mathbf{E_\xi\uparrow}$	**0.927**	**± 0.133**	**0.924**	**± 0.137**	**0.918**	**± 0.135**	**0.910**	**± 0.156**



Εικόνα 5.7 Συγκεντρωτικά αποτελέσματα όλων των υπολογιστικών αναπαραστάσεων του βάθους που προτείνονται για την ανίχνευση σημαντικών αντικειμένων σε εικόνες. Οι τιμές αφορούν τα αποτελέσματα για το σύνολο των δεδομένων εκπαίδευσης (NLPR, NJU2K, STEREO)

Όπως αναφέραμε και νωρίτερα οι χάρτες βάθους από το δίκτυο *DenseDepth* παράγουν ελαφρώς χαμηλότερα αποτελέσματα όταν χρησιμοποιούνται στην προτεινόμενη μεθοδολογία για την ανίχνευση των σημαντικών αντικειμένων μιας οπτικής σκηνής, συγκριτικά με τους αντίστοιχους χάρτες που έχουν ανακτηθεί από κάμερες ή αισθητήρες.

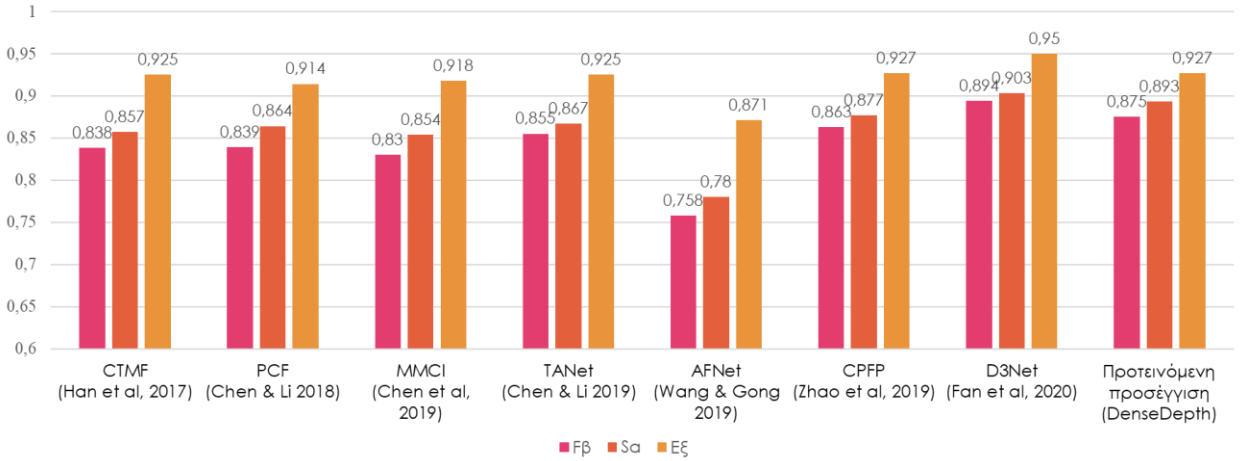
Προκειμένου να αξιολογηθεί η συνεισφορά της πρωτοτυπίας της προτεινόμενης μεθοδολογίας η οποία δεν είναι άλλη από την αντικατάσταση της πληροφορίας του βάθους προερχόμενης από κάμερες ή αισθητήρες με πληροφορία βάθους που εκτιμάται από συνελικτικά νευρωνικά δίκτυα, διεξάγεται σύγκριση των συναγόμενων αποτελεσμάτων με εκείνα της υφιστάμενης κατάστασης (state-of-the-art). Συγκεκριμένα, η παρούσα μεθοδολογία ανίχνευσης σημαντικών αντικειμένων αντιπαραβάλλεται με τις πρόσφατες μεθοδολογίες ανίχνευσης σημαντικών αντικειμένων, οι οποίες αξιοποιούν πληροφορία του βάθους που έχει ανακτηθεί με τη βοήθεια αισθητήρων ή καμερών. Πρόκειται για τις μεθοδολογίες που παρουσιάστηκαν αναλυτικά στην υποενότητα (3.4.3.3). Τα αποτελέσματα της προαναφερόμενης σύγκρισης παρουσιάζονται στον πίνακα (5.5). Με έντονο κόκκινο χρώμα για την εκάστοτε μετρική αξιολόγησης επισημαίνεται το μοντέλο με την βέλτιστη συνολικά επίδοση, ενώ με έντονο μπλε χρώμα υπογραμμίζεται η μεθοδολογία με τη δεύτερη καλύτερη επίδοση.

Από τον πίνακα (5.5) διαπιστώνουμε ότι για έξι από τις επτά μεθοδολογίες υφιστάμενης κατάστασης που χρησιμοποιούν πληροφορία βάθους απευθείας από κάμερα ή αισθητήρα, η συστηνόμενη προσέγγιση επιτυγχάνει συνολικά καλύτερα αποτελέσματα. Τα αποτελέσματα που αντιστοιχούν στην ανίχνευση των σημαντικών αντικειμένων που πραγματοποιήθηκε σύμφωνα με τη προτεινόμενη μεθοδολογία - η οποία αξιοποιεί πληροφορία βάθους που εκτιμάται από το δίκτυο *DenseDepth*- την κατατάσσουν ως το μοντέλο με τη δεύτερη συνολικά καλύτερη επίδοση. Το μοντέλο το οποίο σημειώνει την καλύτερη επίδοση είναι το *D^3Net* (Fan, Lin, et al., 2020). Πρόκειται για ένα επίσης συνελικτικό μοντέλο με όμοια τριπλή ροή δεδομένων (RGB, RGB-D, Depth) το οποίο αποτέλεσε και την έμπνευση για το σχεδιασμό του συνελικτικού αποκωδικοποιητή που χρησιμοποιείται στην παρούσα εργασία προκειμένου να εντοπιστούν τα σημαντικά αντικείμενα μιας οπτικής σκηνής. Όπως ήδη αναφέρθηκε και στο τέταρτο κεφάλαιο, η παρούσα εργασία δανείζεται την αρχιτεκτονική του υποδικτύου RGB-D. Πράγματι, δίκτυο *D^3Net* επιτυγχάνει συνολικά την καλύτερη επίδοση. Ωστόσο, οφείλουμε να επισημάνουμε ότι το μοντέλο D3Net φέρει υψηλότερη πολυπλοκότητα εν αντιθέσει με την προτεινόμενη μεθοδολογία, καθώς το πρώτο αξιοποιεί μια τριπλή ροή δεδομένων ενώ η παρούσα προσέγγιση μόνο μια ροή δεδομένων. Να διευκρινίσουμε ότι στη

σύγκριση των αποτελεσμάτων του πίνακα (5.5) παραλείπονται οι τιμές τις τυπικής απόκλισης καθώς καμία από τις state-of-the-art μεθοδολογίες δεν παρέχει αυτού του είδους την πληροφορία.

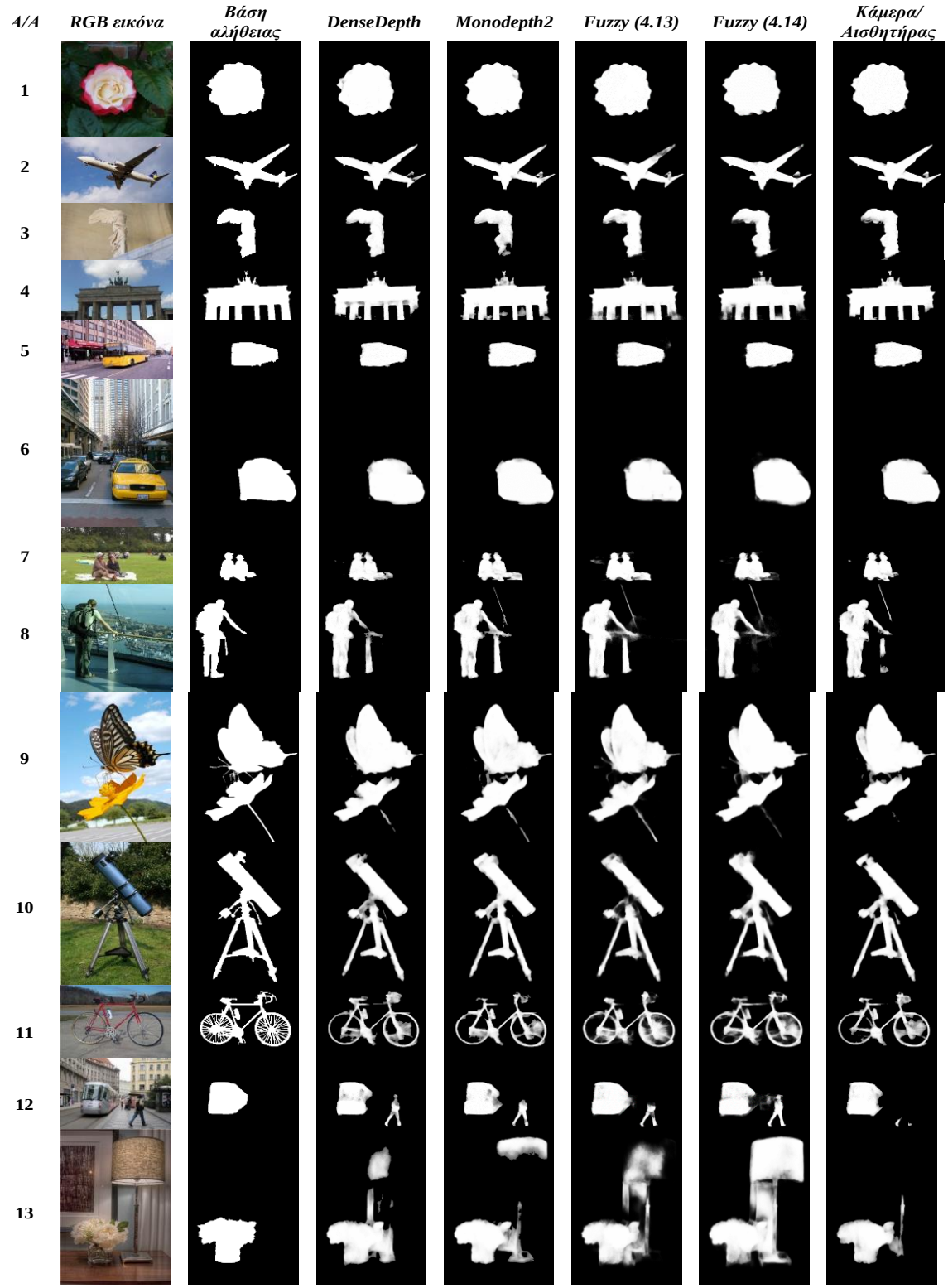
5.5 Αποτελέσματα της σύγκρισης της προτεινόμενης προσέγγισης με state-of-the-art μεθοδολογίες ανίχνευσης σημαντικών αντικειμένων υποβοηθούμενες από την πληροφορία του βάθους.

Σύνολο Δεδο-μένων	**Μέτρο Αξιολό-γησης**	**CTMF (Han et al, 2017)**	**PCF (Chen & Li 2018)**	**MMCI (Chen et al, 2019)**	**TANet (Chen & Li 2019)**	**AFNet (Wang & Gong 2019)**	**CPFP (Zhao et al, 2019)**	**D^{3}Net (Fan et al, 2020)**	**Προτεινόμενη προσέγγιση (DenseDepth)**
NJU2K	*MAE*↓	0.085	0.059	0.079	0.060	0.100	0.053	0.041	0.049
	F_β ↑	0.845	0.872	0.852	0.874	0.775	0.877	0.900	0.877
	S_a ↑	0.849	0.877	0.858	0.878	0.772	0.879	0.900	0.886
	E_ξ ↑	0.913	0.924	0.915	0.925	0.853	0.926	0.950	0.917
NLPR	*MAE*↓	0.056	0.044	0.059	0.041	0.058	0.036	0.030	0.030
	F_β ↑	0.825	0.841	0.815	0.863	0.771	0.867	0.897	0.878
	S_a ↑	0.860	0.874	0.856	0.866	0.799	0.88	0.912	0.904
	E_ξ ↑	0.929	0.925	0.913	0.941	0.879	0.932	0.953	0.944
STEREO	*MAE*↓	0.055	0.049	0.065	0.046	0.068	0.038	0.031	0.047
	F_β ↑	0.844	0.804	0.822	0.827	0.728	0.846	0.885	0.871
	S_a ↑	0.863	0.842	0.848	0.858	0.770	0.872	0.898	0.889
	E_ξ ↑	0.932	0.893	0.928	0.910	0.881	0.923	0.946	0.921
Συνολικά	***MAE*↓**	**0.065**	**0.051**	**0.068**	**0.049**	**0.075**	**0.042**	**0.034**	**0.042**
	F_β ↑	**0.838**	**0.839**	**0.830**	**0.855**	**0.758**	**0.863**	**0.894**	**0.875**
	S_a ↑	**0.857**	**0.864**	**0.854**	**0.867**	**0.78**	**0.877**	**0.903**	**0.893**
	E_ξ ↑	**0.925**	**0.914**	**0.918**	**0.925**	**0.871**	**0.927**	**0.950**	**0.927**



Εικόνα 5.8 Σύγκριση με μοντέλα υφιστάμενης κατάστασης

5.3.2. Ποιοτική σύγκριση των αποτελεσμάτων



Εικόνα 5. 9 Χάρτες σημαντικών αντικειμένων όπως εκτιμώνται από το σύνολο των προσεγγίσεων που εξετάζονται στην παρούσα εργασία

Σε συνέχεια της αξιολόγησης των παραγόμενων αποτελεσμάτων διεξάγεται και μια ποιοτική σύγκριση. Από την ποιοτική σύγκριση των αποτελεσμάτων της προτεινόμενης μεθοδολογίας συνάγονται οι ακόλουθες παρατηρήσεις:

Η αντικατάσταση της πληροφορίας του βάθους η οποία έχει ανακτηθεί από κάμερες/ αισθητήρες με αντίστοιχη πληροφορία η οποία προέρχεται από προ-εκπαιδευμένα συνελικτικά νευρωνικά δίκτυα ή/και σε συνδυασμό με κανόνες ασαφούς λογικής δύναται να προσφέρει εξίσου καλά αποτελέσματα σε εργασίες ανίχνευσης σημαντικών αντικειμένων (σειρές 1, 2).

Ακόμη και στην περίπτωση όπου τα σημαντικά αντικείμενα μιας σκηνής έχουν περίπλοκα σχήματα (σειρά 3η) ή φέρουν ιδιαίτερες λεπτομέρειες (σειρά 4η). Μάλιστα σε ορισμένες περιπτώσεις (*MonoDepth2* – σειρά 2η , *DenseDepth* – σειρά 3η) οι χάρτες σημαντικότητας που προκύπτουν με την αξιοποίηση της πληροφορίας του βάθους από τα συνελικτικά δίκτυα, φαίνεται να προσεγγίζουν ελαφρώς καλύτερα τη βάση αλήθειας συγκριτικά με τους χάρτες σημαντικότητας που προέκυψαν με χάρτες βάθους ανακτώμενους από κάμερες/αισθητήρες.

Οι προτεινόμενες προσεγγίσεις ανταποκρίνονται το ίδιο καλά σε σχέση με τη συμβατική μεθοδολογία -πληροφορία βάθους από κάμερα/αισθητήρα-σε περισσότερο απαιτητικά σενάρια εικόνων. Μεταξύ των σεναρίων αυτών συγκαταλέγονται περιπτώσεις όπου η λήψη της παρατηρούμενης σκηνής -δηλαδή της εικόνας RGB - έχει πραγματοποιηθεί υπό γωνία (σειρά 5), όπου υπάρχουν απαθανατίζονται πολλαπλά αντικείμενα στη σκηνή (σειρά 6η) ή όπου τα πολλαπλά αντικείμενα βρίσκονται σε διαφορετικό βάθος (σειρά 7η).

Ακόμα, διαπιστώνεται ότι το προτεινόμενο πλαίσιο μεθοδολογίας είναι σε θέση να εντοπίζουν τις σημαντικές αναπαραστάσεις μιας σκηνής είτε πρόκειται για αντικείμενα (σειρές 1-6) είτε πρόκειται για ανθρώπους (σειρές 7-8).

Στην ιδιαίτερη περίπτωση της 8ης σειράς παρατηρούμε ότι όλες οι προσεγγίσεις αποτυγχάνουν να ανιχνεύσουν ακριβώς το σημαντικό αντικείμενο της σκηνής καθώς όλες θεωρούν σε μεγαλύτερο ή μικρότερο βαθμό ως σημαντικά αντικείμενα που η βάση αλήθειας δε συγκαταλέγει στα σημαντικά. Ωστόσο, να εξετάσουμε προσεκτικότερα τα αποτελέσματα συμπεραίνουμε ότι η ασαφής αναπαράσταση των χαρτών βάθους του δικτύου *DenseDepth* που χρησιμοποιεί μόνο τα δυο κανάλια πληροφορίας παράγει καλύτερα αποτελέσματα ακόμη και από τη συμβατική προσέγγιση.

Ένα ακόμη σενάριο όπου η ασαφής αναπαράσταση των χαρτών βάθους του δικτύου *DenseDepth* (fuzzy-4.13) φαίνεται να παράγει αποτελέσματα περισσότερο συνεπή ως προς τη βάση αλήθειας από κάθε άλλη προσέγγιση είναι αυτό της 9η σειράς. Στην προκειμένη

περίπτωση ο χάρτης που παράγεται απαθανατίζει το σημαντικό αντικείμενο της σκηνής και τις λεπτομέρειες του με μεγάλη ακρίβεια. Επιπλέον, παρατηρούμε ότι στο σενάριο αυτό η ασαφής αναπαράσταση των προσεγγιστικών χαρτών βάθους εντοπίζει λεπτομέρειες που η συμβατική προσέγγιση αγνοεί.

Επιπρόσθετα, οι προσεγγίσεις που παρουσιάζονται σε αυτή την εργασία δίνουν οπτικά εξίσου καλά αποτελέσματα με τη συμβατική προσέγγιση ακόμη και όταν τα σημαντικά αντικείμενα φέρουν δυσδιάκριτες λεπτομέρειες (10η σειρά).

Τέλος, οφείλουμε να επισημάνουμε ότι σε ορισμένα σενάρια εικόνων όπως αυτά των σειρών 11-13 τόσο η συμβατική όσο και οι προτεινόμενες προσεγγίσεις αποτυγχάνουν να απαθανατίσουν επακριβώς τη βάση αλήθειας. Καθώς είτε δεν εντοπίζουν ευδιάκριτα λεπτομέρειες που φαίνεται να βρίσκονται στο ίδιο βάθος (ακτίνες του ποδηλάτου στην 11η σειρά) είτε θεωρούν πως τα σημαντικά αντικείμενα της σκηνής είναι περισσότερα από ένα (12η και 13η σειρά).

6. Συμπεράσματα

Η παρούσα εργασία παρούσα εργασία διερευνά την συνεισφορά υπολογιστικών προσεγγίσεων της πληροφορίας του βάθους σε εργασίες ανίχνευσης σημαντικών αντικειμένων σε RGB εικόνες. Ειδικότερα, προτείνεται η αντικατάσταση της πληροφορίας του βάθους που αντιστοιχεί σε μια οπτική σκηνή και η οποία έχει ανακτηθεί από στερεοσκοπικές κάμερες ή αισθητήρες βάθους με ομόλογη πληροφορία η οποία υπολογίζεται από προεκπαιδευμένες αρχιτέκτονες συνελικτικών νευρωνικών δικτύων ή/και σε συνδυασμό με κανόνες ασαφούς λογικής.

Η ανίχνευση των σημαντικών αντικειμένων πραγματοποιείται από ένα συνελικτικό νευρωνικό δίκτυο που φέρει αρχιτεκτονική αυτοκωδικοποιητή, ενώ σε χρόνο μεταγενέστερο της εκπαίδευσης ένας πρόσθετος επιδιορθωτικός μηχανισμός υιοθετείται προκειμένου να βελτιώσει τα αποτελέσματα της πρόβλεψης. Η μεθοδολογία που συστήνεται επαληθεύεται έναντι γνωστών RGB-D συνόλων δεδομένων καθιερωμένων στην ανίχνευση σημαντικών αντικειμένων.

Σύμφωνα με τα αποτελέσματα που συνάγονται, οι διαφορετικές προσεγγιστικές αναπαραστάσεις του βάθους οι οποίες έχουν εκτιμηθεί από συνελικτικά νευρωνικά δίκτυα, δύνανται να προσφέρουν συγκριτικά αποτελέσματα με την πληροφορία του βάθους ανακτώμενη από συσκευές αναφορικά με την προτεινόμενη μεθοδολογία. Επιπλέον, έναντι των μεθοδολογιών της υφιστάμενης κατάστασης για την ανίχνευση σημαντικών αντικειμένων σε RGB-D εικόνες, τα αποτελέσματα της συγκεκριμένης εργασίας επιτυγχάνουν υψηλότερη επίδοση για την πλειοψηφία των μοντέλων που έχουν προταθεί μέχρι στιγμής. Στο σημείο αυτό να επισημάνουμε ότι σε όλα τα μοντέλα της υφιστάμενης κατάστασης η πληροφορία του βάθους που αξιοποιείται έχει ανακτηθεί από στερεοσκοπικές κάμερες ή αισθητήρες βάθους.

Το γεγονός ότι η προσέγγιση που εισάγεται στην παρούσα εργασία παρέχει συγκρίσιμα αποτελέσματα, μας επιτρέπει να αποφανθούμε ότι η αντικατάσταση της πληροφορίας του βάθους από υπολογιστικές προσεγγίσεις που αξιοποιούν συνελικτικά νευρωνικά δίκτυα, είναι εφικτή για την ανίχνευση των σημαντικών αντικειμένων μιας οπτικής αναπαράστασης. Η δυνατότητα αυτή καταργεί την εξάρτηση του εν λόγω πεδίου της υπολογιστικής όρασης από τα υπάρχοντα RGB-D σύνολα δεδομένων τα οποία είναι σχετικά περιορισμένα έναντι των RGB συνόλων δεδομένων που προορίζονται για την ανίχνευση σημαντικών αντικειμένων. Επιπλέον, ελαττώνει το κόστος του υλικού καθώς αναιρεί την αναγκαιότητα ύπαρξης στερεοσκοπικών καμερών ή αισθητήρων βάθους για την ανάκτηση της πληροφορίας του

βάθους. Ο παράγοντας αυτός θεωρείται σημαντικός όταν γίνεται λόγος για εφαρμογές ρομποτικής όπου όπως έχει ήδη ειπωθεί εφαρμόζεται η ανίχνευση των σημαντικών αντικειμένων.

Σε μια προσπάθεια βελτίωσης των αποτελεσμάτων της παρούσας εργασίας προκειμένου η προτεινόμενη προσέγγιση να παρέχει υψηλότερα αποτελέσματα ακόμη και από το ενδεχόμενο αξιοποίησης της συμβατικής πληροφορίας του βάθους θα παρουσίαζε ιδιαίτερο ενδιαφέρον:

i. η διερεύνηση διαφορετικών αρχιτεκτονικών ή ακόμη και συναρτήσεων απώλειας όσον αφορά το δίκτυο που διεξάγει την ανίχνευση των σημαντικών αντικειμένων,

ii. η ενσωμάτωση του επιδιορθωτικού μηχανισμού στη διαδικασία της εκπαίδευσης του δικτύου που πραγματοποιεί την ανίχνευση των σημαντικών αντικειμένων,

iii. καθώς και η ταυτόχρονη εκπαίδευση των συνελικτικών αρχιτεκτονικών της ανίχνευσης των σημαντικών αντικειμένων και της εκτίμησης των χαρτών του βάθους.

Βιβλιογραφία

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. 2009 IEEE conference on computer vision and pattern recognition (pp. 1597–1604). IEEE.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence, 34(11), 2274–2282.

Aggarwal, C. C. (2018). Neural Networks and Deep Learning: A Textbook. Springer International Publishing. Retrieved from https://books.google.gr/books?id=achqDwAAQBAJ

Alexe, B., Deselaers, T., & Ferrari, V. (2010). What is an object? 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 73–80). IEEE.

Alhashim, I., & Wonka, P. (2018). High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941.

Bharath, R., Nicholas, L. Z. J., & Cheng, X. (2013). Scalable scene understanding using saliency-guided object localization. 2013 10th IEEE International Conference on Control and Automation (ICCA) (pp. 1503–1508). IEEE.

Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., & Li, J. (2019). Salient object detection: A survey. Computational visual media, 1–34.

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. Journal of vision, 14(3), 29–29.

Breckon, C. Toby; Solomon. (2011). Fundamentals of digital image processing : a practical approach with examples in Matlab (1st ed.). Wiley-Blackwell. Retrieved from http://gen.lib.rus.ec/book/index.php?md5=d097a6b40bfe517e69c73dbf8bcd5260

Chang, K.-Y., Liu, T.-L., Chen, H.-T., & Lai, S.-H. (2011). Fusing generic objectness and visual saliency for salient object detection. 2011 International Conference on Computer Vision (pp. 914–921). IEEE.

Chen, H., & Li, Y. (2018). Progressively complementarity-aware fusion network for RGB-D salient object detection. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3051–3060).

Chen, H., & Li, Y. (2019). Three-stream attention-aware network for RGB-D salient object detection. IEEE Transactions on Image Processing, 28(6), 2825–2835.

Chen, H., Li, Y., & Su, D. (2019). Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognition, 86, 376–385.

Cheng, M., Zhang, G., Mitra, N. J., Huang, X., & ű, S. (2011). Global contrast based salient region detection. CVPR 2011 (pp. 409–416). doi:10.1109/CVPR.2011.5995344

Cheng, Y., Fu, H., Wei, X., Xiao, J., & Cao, X. (2014). Depth enhanced saliency detection method. Proceedings of international conference on internet multimedia computing and service (pp. 23–27).

Cong, R., Lei, J., Fu, H., Cheng, M.-M., Lin, W., & Huang, Q. (2018). Review of visual saliency detection with comprehensive information. IEEE Transactions on circuits and Systems for Video Technology, 29(10), 2941–2959.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(7).

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., & Borji, A. (2017). Structure-measure: A new way to evaluate foreground maps. Proceedings of the IEEE international conference on computer vision (pp. 4548–4557).

Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M., & Borji, A. (2018). Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421.

Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020). Pranet: Parallel reverse attention network for polyp segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 263–273). Springer.

Fan, D.-P., Lin, Z., Zhang, Z., Zhu, M., & Cheng, M.-M. (2020). Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks. IEEE Transactions on Neural Networks and Learning Systems.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., et al. (2015). From captions to visual concepts and back. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1473–1482).

Fang, Y., Wang, J., Narwaria, M., Le Callet, P., & Lin, W. (2014). Saliency detection for stereoscopic images. IEEE Transactions on Image Processing, 23(6), 2625–2636.

Feng, D., Barnes, N., You, S., & McCarthy, C. (2016). Local background enclosure for RGB-D salient object detection. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2343–2350).

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. Competition and cooperation in neural nets (pp. 267–285). Springer.

Gao, D., & Vasconcelos, N. (2005). Discriminant saliency for visual recognition from cluttered scenes. Advances in neural information processing systems (pp. 481–488).

Girshick, R. (2015). Fast r-cnn. Proceedings of the IEEE international conference on computer vision (pp. 1440–1448).

Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. Proceedings of the IEEE international conference on computer vision (pp. 3828–3838).

Gonzalez, R. C., & Woods, R. E. (2011). Digital Image Processing. Pearson Education. Retrieved from https://books.google.gr/books?id=MaYuAAAAQBAJ

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., et al. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354–377.

Gupta, S., Girshick, R., Arbeláez, P., & Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation. European conference on computer vision (pp. 345–360). Springer.

Han, J., Chen, H., Liu, N., Yan, C., & Li, X. (2017). CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE transactions on cybernetics, 48(11), 3171–3183.

Han, S., & Vasconcelos, N. (2006). Image compression using object-based regions of interest. 2006 International Conference on Image Processing (pp. 3097–3100). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision (pp. 1026–1034).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

He, S., Lau, R. W., Liu, W., Huang, Z., & Yang, Q. (2015). Supercnn: A superpixelwise convolutional neural network for salient object detection. International journal of computer vision, 115(3), 330–344.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. (2017). Deeply supervised salient object detection with short connections. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3203–3212).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700–4708).

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology, 148(3), 574.

Itti, L., & Koch, C. (1999). Comparison of feature combination strategies for saliency-based visual attention systems. Human vision and electronic imaging IV (Vol. 3644, pp. 473–482). International Society for Optics and Photonics.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. Nature reviews neuroscience, 2(3), 194–203.

Jian, M., Zhang, W., Yu, H., Cui, C., Nie, X., Zhang, H., & Yin, Y. (2018). Saliency detection based on directional patches extraction and principal local color contrast. Journal of Visual Communication and Image Representation, 57, 1–11.

Jian, M., Zhou, Q., Cui, C., Nie, X., Luo, H., Zhao, J., & Yin, Y. (2019). Assessment of feature fusion strategies in visual attention mechanism for saliency detection. Pattern Recognition Letters, 127, 37–47.

Jiang, B., Zhou, Z., Wang, X., Tang, J., & Luo, B. (2020). cmSalGAN: RGB-D Salient Object Detection with Cross-View Generative Adversarial Networks. IEEE Transactions on Multimedia.

Jiang, P., Ling, H., Yu, J., & Peng, J. (2013). Salient region detection by ufo: Uniqueness, focusness and objectness. Proceedings of the IEEE international conference on computer vision (pp. 1976–1983).

Ju, R., Ge, L., Geng, W., Ren, T., & Wu, G. (2014). Depth saliency based on anisotropic center-surround difference. 2014 IEEE international conference on image processing (ICIP) (pp. 1115–1119). IEEE.

Karpathy, A., Miller, S., & Fei-Fei, L. (2013). Object discovery in 3d scenes via shape analysis. 2013 IEEE International Conference on Robotics and Automation (pp. 2088–2095). IEEE.

Kim, J., & Pavlovic, V. (2016). A shape-based approach for salient object detection using deep learning. European Conference on Computer Vision (pp. 455–470). Springer.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry II, M. J., Balasubramanian, V., & Sterling, P. (2006). How much the eye tells the brain. Current Biology, 16(14), 1428–1434.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems (pp. 1097–1105).

Kruthiventi, S. S., Gudisa, V., Dholakiya, J. H., & Venkatesh Babu, R. (2016). Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5781–5790).

Kuen, J., Wang, Z., & Wang, G. (2016). Recurrent attentional networks for saliency detection. Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (pp. 3668–3677).

Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., & Yan, S. (2012). Depth matters: Influence of depth cues on visual saliency. European conference on computer vision (pp. 101–115). Springer.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems (pp. 396–404).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

Lee, G., Tai, Y.-W., & Kim, J. (2016). Deep saliency with encoded low level distance map and high level features. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 660–668).

Lei, J., Duan, J., Wu, F., Ling, N., & Hou, C. (2016). Fast mode decision based on grayscale similarity and inter-view correlation for depth map coding in 3D-HEVC. IEEE Transactions on Circuits and Systems for Video Technology, 28(3), 706–718.

Lei, J., Zhang, C., Fang, Y., Gu, Z., Ling, N., & Hou, C. (2015). Depth sensation enhancement for multiple virtual view rendering. IEEE Transactions on Multimedia, 17(4), 457–469.

Li, C., Yuan, Y., Cai, W., Xia, Y., & Dagan Feng, D. (2015). Robust saliency detection via regularized random walks ranking. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2710–2717).

Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5455–5463).

Li, G., & Yu, Y. (2016). Deep contrast learning for salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 478–487).

Li, N., Ye, J., Ji, Y., Ling, H., & Yu, J. (2014). Saliency detection on light field. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2806–2813).

Li, X., Lu, H., Zhang, L., Ruan, X., & Yang, M.-H. (2013). Saliency detection via dense and sparse reconstruction. Proceedings of the IEEE international conference on computer vision (pp. 2976–2983).

Liu, G., & Fan, D. (2013). A model of visual attention for natural image retrieval. 2013 International Conference on Information Science and Cloud Computing Companion (pp. 728–733). IEEE.

Liu, N., & Han, J. (2016). Dhsnet: Deep hierarchical saliency network for salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 678–686).

Liu, R., Cao, J., Lin, Z., & Shan, S. (2014). Adaptive partial differential equation learning for visual saliency detection. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3866–3873).

Liu, Z., Tang, J., Xiang, Q., & Zhao, P. (2020). Salient object detection for RGB-D images by generative adversarial network. Multimedia Tools and Applications, 79(35), 25403–25425.

Liu, Z., Zhang, W., & Zhao, P. (2020). A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection. Neurocomputing, 387, 210–220.

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

Mahadevan, V., & Vasconcelos, N. (2009). Saliency-based discriminant tracking. 2009 IEEE conference on computer vision and pattern recognition (pp. 1007–1013). IEEE.

Menze, M., & Geiger, A. (2015). Object scene flow for autonomous vehicles. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3061–3070).

Nathan Silberman, P. K. Derek Hoiem, & Fergus, R. (2012). Indoor Segmentation and Support Inference from RGBD Images. ECCV.

Niu, Y., Geng, Y., Li, X., & Liu, F. (2012). Leveraging stereopsis for saliency analysis. 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 454–461). IEEE.

Peng, H., Li, B., Xiong, W., Hu, W., & Ji, R. (2014). Rgbd salient object detection: a benchmark and algorithms. European conference on computer vision (pp. 92–109). Springer.

Perazzi, F., Krähenbühl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. 2012 IEEE conference on computer vision and pattern recognition (pp. 733–740). IEEE.

Pitas, I. (2000). Digital Image Processing Algorithms and Applications (1st ed.). USA: John Wiley & Sons, Inc.

Prim, R. C. (1957). Shortest connection networks and some generalizations. The Bell System Technical Journal, 36(6), 1389–1401.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. Neural networks, 12(1), 145–151.

Qin, Y., Lu, H., Xu, Y., & Wang, H. (2015). Saliency detection via cellular automata. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 110–119).

Qiuxia, L., Khan, S., Nie, Y., Hanqiu, S., Shen, J., & Shao, L. (2020). Understanding More about Human and Machine Attention in Deep Neural Networks. IEEE Transactions on Multimedia.

Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., & Yang, Q. (2017). RGBD salient object detection via deep fusion. IEEE Transactions on Image Processing, 26(5), 2274–2285.

Rapantzikos, K., Avrithis, Y., & Kollias, S. (2009). Dense saliency-based spatiotemporal feature points for action recognition. 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1454–1461). Ieee.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention (pp. 234–241). Springer.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Sarfraz, M. (2020). Introductory Chapter: On Digital Image Processing. In M. Sarfraz (Ed.), Digital Imaging. Rijeka: IntechOpen. doi:10.5772/intechopen.92060

Schaul, T., Zhang, S., & LeCun, Y. (2013). No more pesky learning rates. International Conference on Machine Learning (pp. 343–351).

Simakov, D., Caspi, Y., Shechtman, E., & Irani, M. (2008). Summarizing visual data using bidirectional similarity. 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–8). IEEE.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Skansi, S. (2018). Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence. Undergraduate Topics in Computer Science. Springer International Publishing. Retrieved from https://books.google.gr/books?id=5cNKDwAAQBAJ

Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., & Ren, T. (2017). Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE Transactions on Image Processing, 26(9), 4204–4216.

Sun, J., Liu, X., Wan, W., Li, J., Zhao, D., & Zhang, H. (2016). Video hashing based on appearance and attention features fusion via DBN. Neurocomputing, 213, 84–94.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., et al. (2015). Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1–9).

Tang, Y., & Wu, X. (2016). Saliency detection via combining region-level and pixel-level predictions with CNNs. European Conference on Computer Vision (pp. 809–825). Springer.

Tang, Y., Wu, X., & Bu, W. (2016). Deeply-supervised recurrent convolutional neural network for saliency detection. Proceedings of the 24th ACM international conference on Multimedia (pp. 397–401).

Tateno, K., Tombari, F., Laina, I., & Navab, N. (2017). Cnn-slam: Real-time dense monocular slam with learned depth prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6243–6252).

Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. International journal of computer vision, 104(2), 154–171.

Ullah, I., Jian, M., Hussain, S., Guo, J., Yu, H., Wang, X., & Yin, Y. (2020). A brief survey of visual saliency detection. Multimedia Tools and Applications, 1–41.

Wang, L., Lu, H., Ruan, X., & Yang, M.-H. (2015). Deep networks for saliency detection via local estimation and global search. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3183–3192).

Wang, N., & Gong, X. (2019). Adaptive fusion for RGB-D salient object detection. IEEE Access, 7, 55277–55284.

Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., & Yang, R. (2019). Salient object detection in the deep learning era: An in-depth survey. arXiv preprint arXiv:1904.09146.

Wang, W., & Shen, J. (2017). Deep visual attention prediction. IEEE Transactions on Image Processing, 27(5), 2368–2378.

Wang, W., Shen, J., & Shao, L. (2015). Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing, 24(11), 4185–4196.

Wang, W., Shen, J., Yang, R., & Porikli, F. (2017). Saliency-aware video object segmentation. IEEE transactions on pattern analysis and machine intelligence, 40(1), 20–33.

Wang, X., Ma, H., & Chen, X. (2016). Salient object detection via fast R-CNN and low-level cues. 2016 IEEE International Conference on Image Processing (ICIP) (pp. 1042–1046). IEEE.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003 (Vol. 2, pp. 1398–1402). Ieee.

Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.-M., Feng, J., Zhao, Y., et al. (2016). Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(11), 2314–2320.

Wei, Y., Wen, F., Zhu, W., & Sun, J. (2012). Geodesic saliency using background priors. European conference on computer vision (pp. 29–42). Springer.

Xie, Y., Lu, H., & Yang, M.-H. (2012). Bayesian saliency via low and mid level cues. IEEE Transactions on Image Processing, 22(5), 1689–1698.

Yang, J., & Yang, M.-H. (2016). Top-down visual saliency via joint CRF and dictionary learning. IEEE transactions on pattern analysis and machine intelligence, 39(3), 576–588.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. European conference on computer vision (pp. 818–833). Springer.

Zhang, D., Meng, D., Zhao, L., & Han, J. (2017). Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. arXiv preprint arXiv:1703.01290.

Zhang, J., Dai, Y., & Porikli, F. (2017). Deep salient object detection by integrating multi-level cues. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1–10). IEEE.

Zhang, J., Fan, D.-P., Dai, Y., Anwar, S., Saleh, F. S., Zhang, T., & Barnes, N. (2020). UC-Net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8582–8591).

Zhang, L., Zhang, D., Sun, J., Wei, G., & Bo, H. (2019). Salient object detection by local and global manifold regularized SVM model. Neurocomputing, 340, 42–54.

Zhang, L., Zhang, Q., & Xiao, C. (2015). Shadow remover: Image shadow removal based on illumination recovering optimization. IEEE Transactions on Image Processing, 24(11), 4623–4636.

Zhao, C., Sun, Q., Zhang, C., Tang, Y., & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences, 1–16.

Zhao, J.-X., Cao, Y., Fan, D.-P., Cheng, M.-M., Li, X.-Y., & Zhang, L. (2019). Contrast prior and fluid pyramid integration for RGBD salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3927–3936).

Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1265–1274).

Zhao, R., Oyang, W., & Wang, X. (2017). Person re-identification by saliency learning. IEEE transactions on pattern analysis and machine intelligence, 39(2), 356–370.

Zhou, T., Fan, D.-P., Cheng, M.-M., Shen, J., & Shao, L. (2020). RGB-D Salient Object Detection: A Survey. arXiv preprint arXiv:2008.00230.

Zhu, C., & Li, G. (2017). A three-pathway psychobiological framework of salient object detection using stereoscopic technology. Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 3008–3014).

Zhu, J.-Y., Wu, J., Xu, Y., Chang, E., & Tu, Z. (2014). Unsupervised object class discovery via saliency-guided multiple class learning. IEEE transactions on pattern analysis and machine intelligence, 37(4), 862–875.

Παπαμάρκος. (2013). Ψηφιακή επεξεργασία & ανάλυση εικόνας.