ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Μοντελοποίηση Οπτικής Προσοχής σε Δεδομένα Βίντεο με Ενσωμάτωση του Βάθους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ιωάννας Π. Διαμάντη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Μοντελοποίηση Οπτικής Προσοχής σε Δεδομένα Βίντεο με Ενσωμάτωση του Βάθους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ιωάννας Π. Διαμάντη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 21/07/2021.

.....	.....	.....
Πέτρος Μαραγκός Καθηγητής Ε.Μ.Π.	Κωνσταντίνος Τζαφέστας Αναπληρωτής Καθηγητής Ε.Μ.Π.	Γεράσιμος Ποταμιάνος Αναπληρωτής Καθηγητής Παν/μιο Θεσσαλίας

Αθήνα, Ιούλιος 2021

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

Περίληψη

Το θέμα της παρούσας Διπλωματικής Εργασίας είναι η αντιμετώπιση του προβλήματος της μοντελοποίησης της προσοχής στην φύση μέσω της πρόβλεψης της Εμφάνειας σε βίντεο. Αντίθετα με τις υπάρχουσες μεθόδους οπτικής Εμφάνειας, οι οποίες χρησιμοποιούν μόνο τις RGB ακολουθίες εικόνων των βίντεο ως είσοδο, η προτεινόμενη μέθοδος χρησιμοποιεί και το βάθος ως μία επιπλέον πληροφορία. Το υπό εξέταση πρόβλημα διαφέρει από το πρόβλημα της Αναγνώρισης Σημαντικών Αντικειμένων (Salient Object Detection), καθώς ο σκοπός είναι η πρόβλεψη της ανθρώπινης προσοχής σε βίντεο σε μία γενικότερη σκοπιά και όχι περιορισμένα σε συγκεκριμένα αντικείμενα.

Το προτεινόμενο μοντέλο αποτελείται από δύο οπτικές ροές, μία για τις RGB εικόνες και μία για τις αντίστοιχες εικόνες βάθους. Και οι δύο ροές ακολουθούν μία αρχιτεκτονική Κωδικοποιητή-Αποκωδικοποιητή και συγχωνεύονται προκειμένου να προκύψει ένας ενιαίος τελικός χάρτης Εμφάνειας. Το δίκτυο εκπαιδεύεται από άκρο σε άκρο και αξιολογείται πάνω σε 9 διαφορετικά σύνολα δεδομένων παρακολούθησης ματιού, τα οποία αποτελούνται από μεγάλο εύρος περιεχομένου βίντεο. Διεξάχθηκαν εκτενή πειράματα τόσο όσον αφορά τις διαφορετικές μεθόδους που εφαρμόστηκαν για τον υπολογισμό του βάθους από τα αρχικά δεδομένα παρακολούθησης ματιού καθώς αυτά δεν περιέχουν αυτή την πληροφορία, όσο και την αλληλεπίδραση και συγχώνευση των δύο πληροφοριών (RGB και βάθους) κατά τη διαδικασία της εκπαίδευσης, προκειμένου να εξεταστεί η συνεισφορά του βάθους στο πρόβλημα της οπτικής εμφάνειας.

Η προτεινόμενη μέθοδος στις περισσότερες περιπτώσεις αποδίδει καλύτερα από πολλές άλλες state-of-the-art μεθόδους όπως και από την RGB-μόνο εκδοχή του μοντέλου, κάτι το οποίο υποδεικνύει την συνεισφορά της πληροφορίας του βάθους στην αποτελεσματική εκτίμηση της Εμφάνειας σε βίντεο τα οποία έχουν προβληθεί σε μία δισδιάστατη οθόνη. Από όσο γνωρίζουμε, αυτή είναι η πρώτη ανταγωνιστική προσέγγιση βαθιάς μάθησης του προβλήματος της εκτίμησης της Εμφάνειας σε βίντεο που συνδυάζει τόσο τις RGB εικόνες όπως και το Βάθος προκειμένου να αντιμετωπίσει το γενικότερο πρόβλημα της εκτίμησης της Εμφάνειας στη φύση.

Λέξεις Κλειδιά

Τρισδιάστατα Συνελικτικά Δίκτυα, Συνελικτικά Νευρωνικά Δίκτυα Διπλής Ροής, Όραση Υπολογιστών, Εμφάνεια, Οπτική Προσοχή, Δεδομένα Βίντεο, Βάθος

Abstract

The scope of the following thesis is to address the problem of attention modeling "in-the-wild", via saliency prediction in videos. Contrary to existing visual saliency approaches using only RGB frames as input, the proposed method also employs depth as an additional modality. The addressed problem differs from salient object detection, because its goal is to predict human attention in videos in a more general aspect and not restricted to specific objects.

The proposed model consists of two visual streams, one for the RGB frames, and one for the corresponding depth frames. Both streams follow an encoder-decoder approach and are fused to obtain a final saliency map. The network is trained end-to-end and is evaluated on 9 different databases that contain eye-tracking data and consist of a wide range of video content. Extensive experiments were carried out, regarding the different methods applied to estimate depth from the initial eye-tracking datasets, since they do not contain the information of depth, as well as regarding the interaction and fusion of the two pieces of information (RGB and depth) during the model's training procedure, in order to investigate the contribution of depth in the problem of visual saliency.

The proposed method outperforms in most cases other state-of-the-art models and the RGB-only variant of the model, which indicates the contribution of depth in accurately estimating saliency in videos displayed on a 2D screen. To the best of our knowledge, this is the first competitive deep learning video saliency estimation approach that combines both RGB and Depth features to address the general problem of saliency estimation "in-the-wild".

Keywords

3D Convolutional Neural Networks, Two-Stream Convolutional Neural Networks, Computer Vision, Saliency, Visual Attention, Video Data, Depth

Ευχαριστίες

Κατ΄ αρχάς θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Πέτρο Μαραγκό, για την ανάθεση της διπλωματικής εργασίας και την εμπιστοσύνη που μου έδειξε για την διεκπεραίωσή της. Οι διαλέξεις του στα μαθήματα της Όρασης Υπολογιστών και της Αναγνώρισης Προτύπων ήταν η αιτία να γνωρίσω αυτά τα ερευνητικά πεδία καθώς και τη μηχανική μάθηση. Είμαι ιδιαίτερα ευγνώμων που μου δόθηκε η ευκαιρία να εξερευνήσω αυτούς τους τομείς της επιστήμης μέσα από τα μαθήματά του, αλλά και την εκπόνηση αυτής της διπλωματικής εργασίας. Επίσης θα ήθελα να ευχαριστήσω θερμά την Δρ. Αντιγόνη Τσιάμη για την πολύ καλή συνεργασία που είχαμε στο πλαίσιο της διπλωματικής, την καθοδήγηση και τις συμβουλές που μου έδωσε, συμβάλλοντας σημαντικά στην πρόοδο της έρευνας.

Θα ήθελα επίσης να ευχαριστήσω το φιλικό μου περιβάλλον για την στήριξη που είχα και ιδιαίτερα τις φίλες και συμφοιτήτριές μου Σοφία, Εύη, Μπιάνκα και Νίκη καθώς και τον φίλο μου Γιάννη με τους οποίους εκτός από τις πολλές υπέροχες στιγμές που μοιραστήκαμε κατά τη διάρκεια των φοιτητικών μας χρόνων, είχαμε επίσης πολύ καλή και παραγωγική συνεργασία σε όλα τα χρόνια της φοίτησής μας στο Εθνικό Μετσόβιο Πολυτεχνείο.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για όλη την πολύτιμη στήριξη που είχα, όχι μόνο στα φοιτητικά μου χρόνια αλλά από πολύ νωρίτερα, κάτι που αποτελεί βασική αιτία της μέχρι τώρα πορείας μου. Ιδιαίτερα θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς τους γονείς μου Πέτρο και Βασιλική αλλά και προς την αδερφή μου Γεωργία για την υπομονή που έδειξαν και την συνεχή καθοδήγηση που μου παρείχαν όλα αυτά τα χρόνια.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Όραση Υπολογιστών και Μηχανική Μάθηση

Η Όραση Υπολογιστών αποτελεί ένα πεδίο της γενικότερης επιστήμης της Επεξεργασίας Σημάτων, το οποίο ασχολείται με την επεξεργασία δισδιάστατων (εικόνες) ή και τρισδιάστατων ψηφιακών σημάτων (βίντεο). Συγκεκριμένα σκοπός είναι η ανάπτυξη αλγορίθμων για την εξαγωγή πληροφορίας υψηλού επιπέδου, προκειμένου να μοντελοποιηθούν οι αυτοματοποιημένες διαδικασίες που πραγματοποιεί το ανθρώπινο οπτικό σύστημα σε μία μηχανή, όπως ένας ηλεκτρονικός υπολογιστής ή ένα ρομπότ. Η πληροφορία του σήματος λαμβάνεται από κατάλληλες κάμερες ή αισθητήρες, επεξεργάζεται από τους αντίστοιχους αλγόριθμους, οι οποίοι αναπτύσσονται με τη βοήθεια επιστημών όπως η φυσική, η γεωμετρία, η στατιστική κ.ά., και τελικά μετασχηματίζεται σε μία μορφή χρήσιμη για την μηχανή. Χαρακτηριστικά υποπροβλήματα της Όρασης Υπολογιστών αποτελούν η ταξινόμηση εικόνων, η κατηγοριοποίηση δράσεων σε βίντεο, η κατάτμηση εικόνας, η εκτίμηση του βάθους σε δισδιάστατες εικόνες, η ανίχνευση και αναγνώριση αντικειμένων, η εκτίμηση πόζας και κίνησης κ.ά.

Η έρευνα στο πεδίο της Όρασης Υπολογιστών πραγματοποιείται εδώ και πολλά χρόνια, με την αρχή να γίνεται κατά τα τέλη του 1960. Τα πρώτα χρόνια και μέχρι σχετικά πρόσφατα, τα προβλήματα της Όρασης Υπολογιστών αντιμετωπίζονταν όπως αναφέρθηκε με μαθηματικούς αλγόριθμους βασισμένους σε διάφορες επιστήμες, αποκαλούμενους πλέον και ως Κλασικοί Αλγόριθμοι Επεξεργασίας Σημάτων, οι οποίοι απέδωσαν ενθαρρυντικά αποτέλεσματα και προώθησαν την περαιτέρω έρευνα αυτής της επιστήμης. Τα τελευταία χρόνια, με την ανάπτυξη της Τεχνητής Νοημοσύνης, της Μηχανικής Μάθησης και των Νευρωνικών Δικτύων η προσέγγιση αυτού του επιστημονικού πεδίου έχει αλλάξει, με την έρευνα να στρέφεται επίσης στην χρήση αυτών των μεθόδων και την διασύνδεσή τους με την Όραση Υπολογιστών και τους υπάρχοντες αλγόριθμούς, κάτι το οποίο φαίνεται να επιφέρει σημαντικά καλύτερα αποτελέσματα σε πληθώρα προβλημάτων.

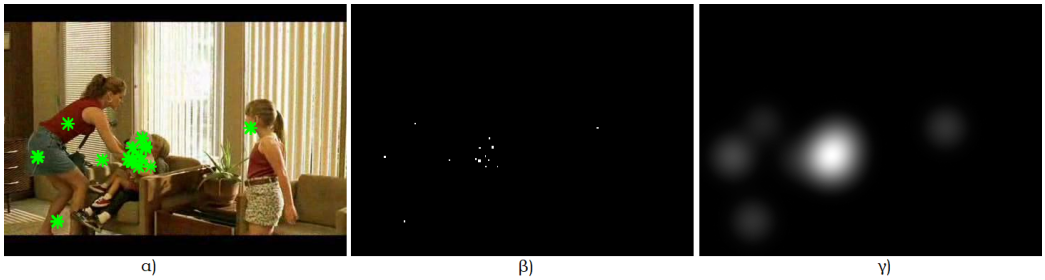
1.2 Περιγραφή του Προβλήματος

Η προσοχή μπορεί να οριστεί ως ένας γνωσιακός μηχανισμός, ο οποίος χρησιμοποείται από διάφορους οργανισμούς (τον άνθρωπο, ζώα κ.λ.π) αλλά και από τεχνητά συστήματα προκειμένου να διαχωριστεί η πιο σημαντική πληροφορία από ένα σύνολο ερεθισμάτων και να διεξαχθεί στη συνέχεια ένα σύνολο σύνθετων διεργασιών. Συγκεκριμένα η ανθρώπινη οπτική προσοχή αποτελεί ένα υποσύνολο του μηχανισμού αυτού, όπου ο άνθρωπος μπορεί να εστιάζει επιλεκτικά σε ένα συγκεκριμένο κομμάτι πληροφορίας, αγνοώντας όλα τα υπόλοιπα ερεθίσματα που αντιλαμβάνεται μέσω του οπτικού του συστήματος. Αυτή η περιοχή αποτέλεσε για χρόνια ένα ενεργό ερευνητικό αντικείμενο για ψυχοφυσικούς και ερευνητές της γνωσιακής επιστήμης, καθώς οι μηχανισμοί προσοχής παίζουν έναν ρόλο ζωτικής σημασίας για τους ανθρώπους.

Η οπτική προσοχή αποτελεί μια ευρεία έννοια, η οποία συχνά περιλαμβάνει πολλά θέματα, όπως η καθοδική επεξεργασία της γνωσιακής πληροφορίας, που εξαρτάται από του στόχους και τις προσδοκίες μας, η αναζήτηση αντικειμένων ή οι απαιτήσεις της εκάστοτε εργασίας. Από την άλλη, η οπτική εμφάνεια είναι μια ανοδική διεργασία και βασίζεται στα αισθητηριακά στοιχεία του κάθε ερεθίσματος τα οποία καθιστούν συγκεκριμένες περιοχές της εικόνας περισσότερο ευδιάκριτες.

Για την μοντελοποίηση του συγκεκριμένου προβλήματος, χρησιμοποιούμε σύνολα δεδομένων, στατικά (εικόνες) ή δυναμικά (βίντεο), αποκαλούμενα ως Δεδομένα Παρακολούθησης Ματιού (Eye-Tracking Data), τα οποία περιέχουν την πληροφορία σχετικά με τις περιοχές εστίασης της οπτικής προσοχής του ανθρώπου πάνω στην εικόνα ή το βίντεο κατά την θέαση. Για την συλλογή αυτής της πληροφορίας και την δημιουργία των συνόλων δεδομένων, το περιεχόμενο τους προβάλλεται σε ορισμένους θεατές - εθελοντές και κατα τη διάρκεια της διαδικασίας, η εστίαση της προσοχής τους χρησιμοποιώντας καταγράφεται από μία συσκευή παρακολούθησης της θέσης της κόρης του ματιού του ανθρώπου (Eye Tracker). Στο Σχήμα 1.1 απεικονίζεται η μορφή των δεδομένων αυτών. Στα 1.1α, 1.1β φαίνονται τα δεδομένα στη μορφή που αυτά καταγράφονται από τον Eye Tracker. Με την καταγραφή της συσκευής αυτής μπορούμε πρακτικά να υπολογίσουμε τις περιοχές (διακριτά pixel) στις οποίες εστίασε ο άνθρωπος. Η αναπαράσταση αυτής της πληροφορίας γίνεται με έναν διακριτό δισδιάστατο χάρτη, ο οποίος ονομάζεται δυαδικός χάρτης εστίασης ή προσοχής (Binary Fixation Map)(1.1β). Στον διακριτό αυτό χάρτη εφαρμόζεται ένα Γκασουσιανό φιλτράρισμα προκειμένου να προκύψει ο συνεχής χάρτης εμφάνειας ή προσοχής (Saliency Map), ο οποίος φαίνεται στο Σχήμα 1.1γ. Στον χάρτη εμφάνειας αναπαριστάται η ίδια πληροφορία με τον χάρτη εστίασης αλλά σε συνεχή μορφή και αποτελεί πρακτικά το ζητούμενο του προβλήματος της εμφάνειας. Ο χάρτης αυτός ονομάζεται επίσης και χάρτης πυκνότητας (Density Map), καθώς αναπαριστά την πυκνότητα ή συγκέντρωση των σημείων που αποτελούν σημεία εστίασης της προσοχής σε μία περιοχή της εικόνας.

Το πρόβλημα της Εμφάνειας (Saliency) σε βίντεο είναι το πρόβλημα της εκτίμησης της εστίασης του ανθρώπινου ματιού όταν αυτό αντιλαμβάνεται δυναμικές σκηνές. Το πρόβλημα αυτό έχει κερδίσει όλο και περισσότερο ενδιαφέρον τα τελευταία χρόνια



Σχήμα 1.1: Η μορφή των Δεδομένων Παρακολούθησης Ματιού (Eye-Tracking Data). (α) RGB εικόνα μαζί με τα διακριτά σημεία (pixel με πράσινο χρώμα) στα οποία εστίασαν οι θεατές. (β) Δυαδικός χάρτης εστίασης (Fixation Map) με τα διακριτά σημεία εστίασης (άσπρα pixel). (γ) Συνεχής χάρτης εμφάνειας με τις άσπρες περιοχές να αναπαριστούν τις περιοχές με τη μεγαλύτερη εμφάνεια.

εξαιτίας και της σημαντικής συνεισφοράς του σε πληθώρα εφαρμογών όπως η σύνοψη και συμπίεση βίντεο, η εικονική πραγματικότητα, η ρομποτική κ.ά. Η παράλληλη ανάπτυξη των τεχνικών βαθιάς μάθησης και συγκεκριμένα των Συνελικτικών Νευρωνικών Δικτύων (Convolutional Neural Networks - CNNs) έχει βοηθήσει ιδιαίτερα στην επίτευξη αξιόλογων αποτελεσμάτων σε διάφορα προβλήματα της Όρασης Υπολογιστών όπως η εκτίμηση της Εμφάνειας, η κατάτμηση της εικόνας, η κατηγοριοποίηση εικόνων κ.ά.

Η εκτίμηση της εμφάνειας σε βίντεο είναι ένα πιο απαιτητικό πρόβλημα σε σχέση με την εκτίμηση της εμφάνειας σε στατικές εικόνες. Αυτό συμβαίνει, διότι στα βίντεο χρειάζεται η ακριβής εξαγωγή τόσο χωρικών όσο και χρονικών χαρακτηριστικών από τις ακολουθίες εικόνων και η αποτελεσματική συγχώνευσή τους προκειμένου να προκύψει ένας εννιαίος χάρτης Εμφάνειας. Προηγούμενες μέθοδοι έχουν προσπαθήσει να αντιμετωπίσουν αυτή την πρόκληση με χρήση διάφορων τεχνικών ενσωμάτωσης της χρονικής πληροφορίας όπως η οπτική ροή, τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNNs), οι τρισδιάστατες συνελίξεις κ.ά.

Η ανθρώπινη προσοχή επηρεάζεται ιδιαίτερα από πολλά διαφορετικά ερεθίσματα, τα οποία μπορεί να είναι παρόντα σε κάποια δυναμική σκηνή βίντεο και να ξυπνήσουν διάφορες ανθρώπινες αισθήσεις. Πρόσφατες έρευνες και μελέτες έχουν δείξει πως τέτοια ερεθίσματα μπορεί να είναι η πληροφορία του βάθους όπως και ο ήχος. Όπως φαίνεται από αυτές, η ενσωμάτωση τέτοιας πληροφορίας στην μοντελοποίηση του προβλήματος μπορεί να βοηθήσει σημαντικά στην ανίχνευση των περιοχών της εικόνας όπου ο άνθρωπος εστιάζει την προσοχή του.

Η πληροφορία του βάθους είναι άρρηκτα συνδεδεμένη με το οπτικό ερέθισμα καθώς ο ανθρώπινος εγκέφαλος έχει την δυνατότητα να αντιληφθεί σημασιολογικά το περιεχόμενο της σκηνής που παρατηρεί και συνεπώς μπορεί να διαχωρίσει τα διαφορετικά αντικείμενα, τα οποία απεικονίζονται σε αυτήν και να εκτιμήσει την σχετική τους απόσταση ακόμα και όταν αυτά απεικονίζονται στις δύο διαστάσεις. Το γεγονός πως

αυτή η πληροφορία γίνεται φυσικά αντιληπτή και επεξεργάζεται από τον ανθρώπινο εγκέφαλο καθιστά το βάθος ένα ενδιαφέρον επιπλέον στοιχείο προς εξέταση σχετικά με την συνεισφορά του την εκτίμηση της Ανθρώπινης Προσοχής. Επιπλέον το βάθος είναι σήμερα εύκολα υπολογίσιμο, τόσο από αισθητήρες βάθους οι οποίοι εμφανίζονται όλο και περισσότερο σε απλές καθημερινές συσκευές, όπως τα κινητά τηλέφωνα, όσο και με την εφαρμογή τεχνικών βαθιάς μάθησης, οι οποίες πετυχαίνουν ιδιαίτερα ακριβή αποτελέσματα.

1.3 Εφαρμογές

Η μοντελοποίηση της Ανθρώπινης Προσοχής είναι ένα πρόβλημα γύρω από το οποίο έχει γίνει εκτεταμμένη έρευνα και ιδιαίτερα τα τελευταία χρόνια το πρόβλημα έχει επεκταθεί σε δυναμικές σκηνές με την ολοένα και μεγαλύτερη παρουσία των βίντεο στην καθημερινότητα του ανθρώπου. Οι μελέτες γύρω από το πρόβλημα αυτό αυξήθηκαν επίσης εξαιτίας της πληθώρας εφαρμογών στις οποίες η μοντελοποίηση της ανθρώπινης προσοχής μπορεί να συνεισφέρει σημαντικά.

Ένα πρόβλημα στο οποίο η εκτίμηση της ανθρώπινης προσοχής έχει χρησιμοποιηθεί αρκετά είναι το πρόβλημα της σύνοψης βίντεο (Video Summarization). Η ανάλυση των βίντεο με αυτοματοποιημένη σύνοψη είναι ένα από τα προβλήματα που απασχολούν ιδιαίτερα την κοινότητα της Όρασης Υπολογιστών, καθώς μας παρέχει μία πιο σύντομη εκδοχή του συνολικού βίντεο, με βάση την οποία μπορούμε να αποφανθούμε πολύ πιο γρήγορα σε σχέση με το περιεχόμενο του. Η εκτίμηση της ανθρώπινης προσοχής σε βίντεο αποτελεί μία πολύ σημαντική συνεισφορά στο συγκεκριμένο πρόβλημα, καθώς δίνει τη δυνατότητα επιλογής των καρέ στα οποία ο άνθρωπος πράγματι εστίασε την προσοχή του και συνεπώς περιέχουν την σημαντικότερη πληροφορία [57, 18, 52, 19, 34].

Ένα ακόμα πρόβλημα στο οποίο η μοντελοποίηση της ανθρώπινης προσοχής μπορεί να συνεισφέρει σημαντικά είναι το πρόβλημα της συμπίεσης βίντεο (Video Compression). Με την ολοένα και μεγαλύτερη ανάπτυξη της τεχνολογίας και του υλικού της κάμερας ακόμα και σε καθημερινές συσκευές όπως τα κινητά τηλέφωνα, προκύπτει η ανάγκη αποτελεσματικής συμπίεσης και αποθήκευσης του καταγεγραμμένου υλικού. Έρευνες έχουν δείξει πως η εκτίμηση των περιοχών στις οποίες ο άνθρωπος εστιάζει περισσότερο την προσοχή του μπορεί να συνεισφέρει σημαντικά στην συμπίεση των βίντεο, εφαρμόζοντας τους αλγόριθμους συμπίεσης προσαρμοσμένους σε κάθε περιοχή της εικόνας ανάλογα με το ποσοστό εμφάνειας που συγκεντρώνεται σε αυτήν και διατηρώντας περισσότερη πληροφορία στις περιοχές που το ποσοστό αυτό είναι συγκριτικά υψηλότερο [21, 42, 41].

Η μοντελοποίηση της Ανθρώπινης Προσοχής και η εκτίμηση της Εμφάνειας έχει επίσης χρησιμοποιηθεί αρκετα στο επιστημονικό πεδίο της ρομποτικής. Ήδη πολλά ρομπότ διαθέτουν κάμερες και αισθητήρες βάθους προκειμένου να μπορούν να κατευθυνθούν στο χώρο. Σε ένα πραγματικό περιβάλλον, οι αισθητήρες αυτοί συλλαμβάνουν πλούσια και πολύπλοκη πληροφορία τόσο από εσωτερικές όσο και από εξωτερικές σκηνές. Προκύπτει συνεπώς η ανάγκη επεξεργασίας αυτής της πληροφορίας και η

εστίαση σε συγκεκριμένους στόχους [85]. Επίσης η ενσωμάτωση της ανθρώπινης αντίληψης από ένα ρομπότ επιτρέπει την αυτόνομη αναζήτηση και ανακάλυψη νέων περιβάλλοντων και τον σχεδιασμό μονοπατιών [13].

Η εικονική πραγματικότητα (Virtual Reality-VR) είναι ακόμα ένα πεδίο εφαρμογής της μοντελοποίησης της ανθρώπινης προσοχής. Η προβολή του οπτικού υλικού υψηλής ποιότητας σε πραγματικό χρόνο είναι μία πρόκληση, καθώς απαιτείται υψηλός αριθμός καρέ ανά δευτερόλεπτο (Frame Rate) και η ανάλυση των εικόνων αυξάνεται συνεχώς. Η ενσωμάτωση της ανθρώπινης προσοχής σε εικονικά περιβάλλοντα μπορεί να συνεισφέρει σημαντικά στην αντιμετώπιση του προβλήματος αυτού, μειώνοντας την ποιότητα και την ανάλυση σε σημεία όπου η εστίαση της ανθρώπινης προσοχής είναι μειωμένη. [67]. Στα εικονικά περιβάλλοντα υπάρχει επίσης η πρόκληση του σχεδιασμού του χώρου και ο τρόπος με τον οποίο ο άνθρωπος τον εξερευνά μέσα σε μία εικονική πραγματικότητα. Η πρόβλεψη της ανθρώπινης προσοχής στην περίπτωση αυτή μπορεί να συνεισφέρει στον αποτελεσματικότερο σχεδιασμό του περιβάλλοντος προσφέροντας καλύτερη και πιο στοχευμένη εμπειρία [68].

1.4 Στόχοι και Συνεισφορά της Διπλωματικής Εργασίας

Στόχος της διπλωματικής εργασίας είναι η υλοποίηση ενός μοντέλου, το οποίο θα μπορεί να παράξει ακριβείς εκτιμήσεις σχετικά με τις περιοχές των καρέ κάποιου βίντεο, στις οποίες ο θεατής εστίασε περισσότερο την προσοχή του όταν το παρακολουθούσε. Για το σκοπό αυτό, η διπλωματική αυτή στοχεύει να ενσωματώσει στα RGB βίντεο των eye tracking δεδομένων την πληροφορία του βάθους, όπως αυτή έχει υπολογιστεί από τρεις διαφορετικές μεθόδους και να εξετάσει την συνεισφορά της πληροφορίας αυτής στην μοντελοποίηση της ανθρώπινης προσοχής. Επίσης η παρούσα Διπλωματική Εργασία σκοπεύει να ανοίξει το δρόμο προς την έρευνα σχετικά με την συνεισφορά του βάθους στο γενικότερο πρόβλημα της Εμφάνειας, κάτι το οποίο φαίνεται να έχει παραμεληθεί παρά την εκτεταμμένη έρευνα που έχει γίνει αντίστοιχα για το πρόβλημα του Salient Object Detection (SOD). Η συνεισφορά της Διπλωματικής Εργασίας μπορεί να συνοψιστεί στα παρακάτω σημεία:

- Σχεδιασμός και υλοποίηση ενός μοντέλου πρόβλεψης της ανθρώπινης προσοχής σε δισδιάστατα βίντεο.
- Σχεδιασμός και υλοποίηση ενός μοντέλου πρόβλεψης της ανθρώπινης προσοχής σε τρισδιάστατα βίντεο.
- Σύγκριση των δύο μοντέλων μεταξύ τους και εξαγωγή συμπερασμάτων σχετικά με την συνεισφορά του βάθους.
- Σύγκριση των μοντέλων με άλλα state-of-the-art μοντέλα σε 9 διαφορετικά σύνολα δεδομένων και εξαγωγή συμπερασμάτων.

- Σύγκριση 3 διαφορετικών μεθόδων υπολογισμού βάθους από δισδιάστατες εικόνες.

- Εμπλουτισμός 9 υπάρχοντων συνόλων δεδομένων παρακολούθησης ματιού (Eye-tracking datasets) με τους αντίστοιχους χάρτες βάθους και από τις 3 μεθόδους.

- Επισημείωση μεγάλου μέρους των δεδομένων του NYU Depth Dataset V2 με τους αντίστοιχους χάρτες βάθους.

1.5 Διάρθρωση της Διπλωματικής Εργασίας

Όσον αφορά το περιεχόμενο και την δομή της Διπλωματικής Εργασίας:

- Στο Κεφάλαιο 2 γίνεται μια βιβλιογραφική ανασκόπηση των μεθόδων που έχουν αναπτυχθεί σχετικά με την μοντελοποίηση της ανθρώπινης προσοχής σε βίντεο καθώς, των προβλημάτων που χρήζουν αντιμετώπισης και της εξέλιξης αυτών των μεθόδων από τις πιο αρχικές μορφές τους μέχρι τις πιο πρόσφατες προσεγγίσεις. Επίσης στο ίδιο κεφάλιο γίνεται αναφορά στα θεωρητικά εργαλεία που χρησιμοποιήθηκαν στην Διπλωματική Εργασία για την επίλυση του συγκεκριμένου προβλήματος. Αναλύονται οι βασικές έννοιες γύρω από την Μηχανική Μάθηση και τα Τεχνητά Νευρωνικά Δίκτυα, οι πρακτικές σχεδιασμού και εκπαίδευσης αυτών καθώς και οι μετρικές αξιολόγησης του προβλήματος της Εμφάνειας.

- Στο Κεφάλαιο 3 παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην Διπλωματική Εργασία τόσο για την μοντελοποίηση της ανθρώπινης προσοχής όσο και τα δεδομένα που χρησιμοποιήθηκαν για την εκτίμηση του βάθους καθώς επίσης και η προεργασία που χρειάστηκε να γίνει σε μερικά σύνολα δεδομένων προκειμένου να μπορέσουν να χρησιμοποιηθούν στην εκπαίδευση.

- Στο Κεφάλαιο 4 εξετάζονται οι τρεις διαφορετικές μέθοδοι εξαγωγής βάθους από δισδιάστατες εικόνες και συγκρίνονται τα αποτελέσματα της κάθε μεθόδου ως προς την ακρίβεια και την ικανότητα γενίκευσης σε νέα δεδομένα.

- Στο Κεφάλαιο 5 παρουσιάζεται η βασική συνεισφορά της Διπλωματικής Εργασίας, όπου αναλύεται η προτεινόμενη μέθοδος και εξηγείται η αρχιτεκτονική και ο τρόπος εκπαίδευσης του προτεινόμενου μοντέλου. Επίσης παρουσιάζονται τα διαφορετικά πειράματα που έγιναν μέχρι να φτάσουμε στην τελική μέθοδο, καθώς επίσης και τα τελικά αποτελέσματα. Με βάση τα αποτελέσματα αυτά συγκρίνονται τόσο η δισδιάστατη με την τρισδιάστατη εκδοχή του μοντέλου μεταξύ τους, καθώς και η προτεινόμενη μέθοδος συνολικά με άλλες state-of-the-art προσεγγίσεις.

- Στο Κεφάλαιο 6 συνοψίζονται η συνεισφορά της εργασίας, τα συμπεράσματα που προέκυψαν από την συγκεκριμένη έρευνα καθώς και οι μελλοντικές επεκτάσεις αυτής για την περαιτέρω βελτίωση της μοντελοποίησης του προβλήματος.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Σχετική Βιβλιογραφία

Η μοντελοποίηση της ανθρώπινης προσοχής είναι ένα πρόβλημα της όρασης υπολογιστών, το οποίο έχει μελετηθεί πάνω σε στατικές εικόνες αρκετά ήδη από την δεκαετία του 1990 και στο οποίο έχουν επιτευχθεί καλά αποτελέσματα τόσο από πιο παραδοσιακές μεθόδους επεξεργασίας σημάτων, όσο από πιο σύγχρονες ποσεγγίσεις όπως αυτή των βαθιών νευρωνικών δικτύων. Το πρόβλημα έχει μελετηθεί ακόμα πιο έντονα τα τελευταία χρόνια και έχει επεκταθεί στην επεξεργασία βίντεο λόγω της σημαντικής του συνεισφοράς σε πληθώρα προβλημάτων και εφαρμογών όπως η σύνοψη βίντεο, η συμπίεση βίντεο, η εικονική πραγματικότητα, η ρομποτική κ.ά. Συνολικά το πρόβλημα της οπτικής προσοχής μπορεί να χωριστεί σε δύο διαφορετικά προβλήματα, αυτό της μοντελοποίησης της ανθρώπινης προσοχής, κατά το οποιο προσπαθούμε να προβλέψουμε την περιοχή εστίασης του ματιού του ανθρώπου και το πρόβλημα της ανίχνευσης σημαντικών αντικειμένων (Salient Object Detection - SOD), κατά το οποίο προσπαθούμε να αναγνωρίσουμε σημαντικές περιοχές και αντικείμενα.

2.1.1 Μοντελοποίηση Προσοχής με χρήση RGB δεδομένων βίντεο

Σχετικά με την μοντελοποίηση του προβλήματος πάνω σε RGB δεδομένα, τόσο στατικών εικόνων όσο και δυναμικού περιεχομένου βίντεο, υπάρχει ιδιαίτερα πλούσια βιβλιογραφία με πληθώρα μεθόδων και αποτελεσματικών προσεγγίσεων του προβλήματος. Συγκεκριμένα για την επίλυση του προβλήματος και την παραγωγή ακριβέστερων προβλέψεων σε δεδομένα βίντεο, απαιτείται η εξαγωγή τόσο χωρικών όσο και των χρονικών χαρακτηριστικών που περιέχονται στα δεδομένα αυτά αλλά και η αποτελεσματική συγχώνευση αυτών των χαρακτηριστικών ώστε να προκύψει μία εννιαία πρόβλεψη βασισμένη σε όλες τις διαστάσεις. Προκειμένου λοιπόν να εκμεταλλευτούν και την χρονική πληροφορία των βίντεο, παλιότερες μέθοδοι συμπεριέλαβαν την πληροφορία της οπτικής ροής μεταξύ διαδοχικών καρέ. Μία δημοφιλής προσέγγιση είναι

η χρήση δικτύων δύο ροών, μίας ροής για την επεξεργασία των καρέ των βίντεο και μίας ροής για την εξαγωγή χαρακτηριστικών από τις αντίστοιχες προυπολογισμένες εικόνες οπτικής ροής των καρέ αυτών. Αυτή η μέθοδος εφαρμόστηκε χαρακτηριστικά στο STSConvNet [1], εμφανίζει όμως κάποιους σημαντικούς περιορισμούς. Αρχικά προστίθεται υπολογιστικό κόστος εφόσον απαιτείται ο υπολογισμός της οπτικής ροής για τα σύνολα δεδομένων καθώς επίσης και πως υπάρχει περιορισμένη αντίληψη των χρονικών χαρακτηριστικών πέρα από τον αριθμό των καρέ που συνήθως χρησιμοποιούνται για τον υπολογισμό της οπτικής ροής. Μία προσπάθεια αντιμετώπισης του τελευταίου προβλήματος έγινε στο STRA-Net [36] όπου για κάθε εικόνα εισόδου, το δίκτυο παίρνει αντίστοιχα πέντε διαδοχικές εικόνες οπτικής ροής, οι οποίες συνδέονται σειριακά και δίνουν μεγαλύτερο εύρος χρονικής πληροφορίας.

Ως μία εναλλακτική της οπτικής ροής για την μοντελοποίηση της χρονικής πληροφορίας από τα βίντεο εισόδου, χρησιμοποιήθηκε αργότερα και με την περαιτέρω ανάπτυξη των νευρωνικών δίκτυων, μία συγκεκριμένη κλάση των νευρωνικών δικτύων: τα Αναδρομικά Νευρωνικά Δίκτυα (RNNs). Τα δίκτυα αυτά είναι ένα είδος νευρωνικών δικτύων, στα οποία οι συνδέσεις μεταξύ των νευρώνων διαμορφώνουν έναν κατευθυνόμενο γράφο στη διεύθυνση μίας χρονικής ακολουθίας. Σε αντίθεση με τα παραδοσιακά δίκτυα εμπρόσθιας τροφοδότησης, τα αναδρομικά νευρωνικά δίκτυα (RNNs) διαθέτουν μία εσωτερική κατάσταση (μνήμη), η οποία τους επιτρέπει να επεξεργαστούν αποτελεσματικά χρονικές ακολουθίες δεδομένων και να εξάγουν τις μεταξύ τους χρονικές σχέσεις. Αυτή η κλάση δικτύων έχει χρησιμοποιηθεί σε πολλές προτεινόμενες μεθόδους για την καλύτερη μοντελοποίηση της οπτικής προσοχής σε βίντεο και σε πολλές περιπτώσεις οι μονάδες αυτές έχουν τροποποιηθεί έτσι ώστε να ενσωματώνουν την πράξη της δισδιάστατης συνέλιξης κάνοντάς τες ακόμα πιο κατάλληλες για προβλήματα όρασης υπολογιστών (ConvLSTM) [78, 80, 40, 28, 81].

Επίσης για την κωδικοποίηση των χωροχρονικών χαρακτηριστικών των ακολουθιών εισόδου, πολλές μέθοδοι χρησιμοποιούν τρισδιάστατα συνελικτικά νευρωικά δίκτυα (3D Convolutional Networks), τα οποία επεκτείνουν την παραδοσιακή αρχιτεκτονική των συνελικτικών δικτύων (CNNs) στις τρεις διαστάσεις, προκειμένου οι συνελικτικοί πυρήνες να εξάγουν παράλληλα χρονικά και χωρικά χαρακτηριστικά. Μία τέτοια προσέγγιση έχει χρησιμοποιηθεί από πολλές σύγχρονες μεθόδους όπως οι [74, 46, 26, 5].

Ταυτόχρονα με τις παραπάνω πρακτικές για την εκμετάλλευση των χρονικών χαρακτηριστικών σε βίντεο, κάποιες μέθοδοι προτείνουν την χρήση κάποιας ακόμα πρότερης γνώσης σχετικά με την είσοδο εκτός από την οπτική ροή, τόσο για την μοντελοποίηση της ανθρώπινης προσοχής σε βίντεο όσο και σε στατικές εικόνες. Τέτοια πρότερη γνώση είναι η προκατάληψη υπέρ του κέντρου της εικόνας (Center Bias), η αναγώριση προσώπων στην εικόνα (Faces Map) [10], η ανίχνευση ακμών (Edge Detection), οι σημασιολογικές σχέσεις των αντικειμένων (Semantic relationships) [87] κ.ά. Η λογική πίσω από την ενσωμάτωση τέτοιας πρόωρης γνώσης στην διαδικασία μοντελοποίησης του συγκεκριμένου προβλήματος έγκειται στο γεγονός πως η ανθρώπινη προσοχή πολλές φορές εστιάζεται σε περιοχές της εικόνας για λόγους που δεν μπορεί εύκολα να ανιχνευθούν από ένα συνελικτικό δίκτυο, όπως για παράδειγμα σε κάποιο αντικείμενο που βρίσκεται στο προσκήνιο, κάτι πιο είναι πιο κεντραρισμένο

στην εικόνα, σε ανθρώπινα πρόσωπα κ.ά.

2.1.2 Μοντελοποίηση Προσοχής με χρήση RGB-D δεδομένων βίντεο

Σχετικά με την πρότερη γνώση της σχετικής θέσης των αντικειμένων που απεικονίζονται (προσκήνιο/παρασκήνιο) έχει γίνει τα προηγούμενα χρόνια αρκετή έρευνα και μάλιστα κατά κύριο λόγο στο πρόβλημα της Ανίχνευσης Σημαντικών Αντικειμένων στην εικόνα (SOD). Σύμφωνα με μία πρόσφατη έρευνα [88], παραπάνω από 100 μοντέλα έχουν χρησιμοποιήσει την πληροφορία του βάθους μαζί με τις RGB εικόνες για το πρόβλημα του SOD ήδη από το 2012 και συνεχίζοντας μέχρι σήμερα με την χρήση βαθιών νευρωνικών δικτύων [7, 55, 89, 86]. Στις περισσότερες μεθόδους, το βάθος συμπεριλαμβάνεται στην εκπαίδευση του δικτύου σε μία διαφορετική επιπλέον ροή του δικτύου και επεξεργάζεται από αυτό είτε ανεξάρτητα από τις RGB εικόνες, είτε με κάποια αλληλεπίδραση μεταξύ του στα ενδιάμεσα στρώματα των μοντέλων. Σε όλες τις περιπτώσεις οι δύο ροές πληροφορίας του δικτύου εν τέλει συνδυάζονται προκειμένου να προκύψει ένας εννιαίος χάρτης εμφάνειας με τα πιο σημαντικά αντικείμενα της εικόνας. Η ενσωμάτωση του βάθους σε τέτοια προβλήματα έχει φέρει πολύ καλά αποτελέσματα ιδιαίτερα σε πολύπλοκες σκηνές με πολλά ή διάφανα αντικείμενα αλλά και σε σκηνές με δύσκολο διαχωρισμό του προσκηνίου από το παρασκήνιο.

Παρά την συνεισφορά που αποδεδειγμένα έχει η πληροφορία του βάθους στο πρόβλημα της Ανίχνευσης Σημαντικών Αντικειμένων, παρατηρείται πως δεν έχει δωθεί αντίστοιχη σημασία στην έρευνα της συνεισφοράς του βάθους στο γενικότερο πρόβλημα της μοντελοποίησης της ανθρώπινης προσοχής και ακόμα περισσότερο στην μοντελοποίηση με χρήση βαθιάς μάθησης. Σε μία αρχική και πιο απλοϊκή εφαρμογή με χρήση βαθιών νευρωνικών δικτύων εξετάστηκε η ενσωμάτωση του χάρτη βάθους στα δεδομένα εισόδου από τους Leifman et. al. [37], όπου η εικόνα του βάθους απλώς συνδέεται σειριακά με τα υπόλοιπα δεδομένα και αυτά επεξεργάζονται όλα μαζί από το δίκτυο προκειμένου να παραχθεί η εκτίμηση ενός εννιαίου χάρτη Εμφάνειας. Εκτός από αυτήν την προσέγγιση, δεν υπάρχουν ιδιαίτερα πολλές άλλες προτεινόμενες μέθοδοι που να εξετάζουν την συνεισφορά του βάθους καθώς και τους διάφορους τρόπους που η πληροφορία αυτή μπορεί να χρησιμοποιηθεί αποτελεσματικότερα στην μοντελοποίηση του προβλήματος, όπως έχει γίνει εκτεταμμένα στο πρόβλημα της Ανίχνευσης Σημαντικών Αντικειμένων.

Σκοπός της Διπλωματικής Εργασίας είναι τόσο να εξετάσει την αποτελεσματικότητα του βάθους ως μία επιπλέον πληροφορία στο γενικότερο πρόβλημα της Εκτίμησης της Ανθρώπινης Προσοχής όσο και να ανοίξει τον δρόμο της έρευνας προς αυτήν την κατεύθυνση, η οποία φαίνεται να έχει παραμεληθεί στο συγκεκριμένο πρόβλημα παρά τα ενθαρρυντικά αποτελέσματα που πετυχαίνει σε άλλα πολύ παρόμοια προβλήματα της όρασης υπολογιστών.

2.2 Θεωρητικά Εργαλεία

2.2.1 Τεχνικές βαθιάς μάθησης (Deep Learning)

Στο κεφάλαιο αυτό παρουσιάζεται η έννοια της βαθιάς μάθησης. Η έννοια αυτή είναι άρρηκτα συνδεδεμένη με το πλέον δημοφιλέστερο πεδίο της μηχανικής μάθησης, τα νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα είναι μία κλάση μοντέλων τα οποία έχουν πρόσφατα δημιουργήσει μια τεχνολογική επανάσταση μέσα στο γενικότερο πεδίο της μηχανικής μάθησης, φέρνοντας πολύ καλά αποτελέσματα και λύνοντας σημαντικά προβλήματα, μεταξύ άλλων, και στο πεδίο της Όρασης Υπολογιστών. Συγκεκριμένα για την επίλυση προβλημάτων Όρασης Υπολογιστών χρησιμοποιούνται σε μεγάλο βαθμό τα συνελικτικά νευρωνικά δίκτυα (CNNs), τα οποία θα παρουσιαστούν αργότερα στο κεφάλαιο.

2.2.2 Τεχνητά Νευρωνικά Δίκτυα

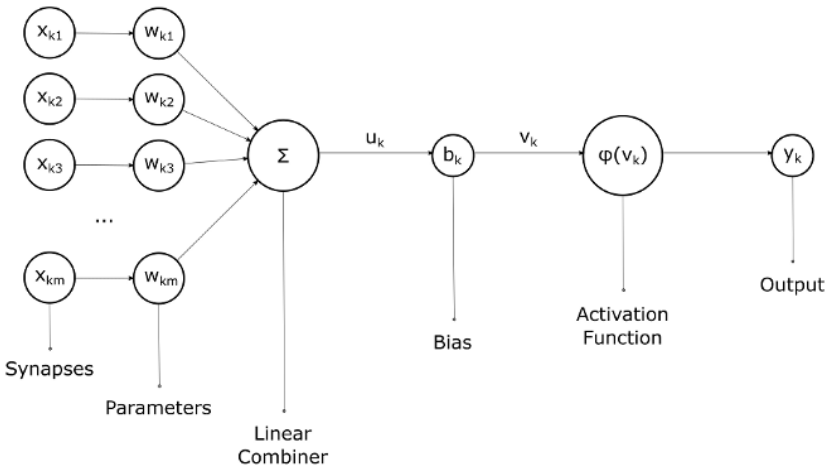
Ένα τεχνητό νευρωνικό δίκτυο είναι μία σειρά αλγορίθμων, που προσπαθεί να κατανοήσει τις υποκείμενες συσχετίσεις μέσα σε ένα ή περισσότερα σύνολα δεδομένων μέσω μίας διαδικασίας που προσομοιώνει τον ανθρώπινο εγκέφαλο. Το έργο στο επιστημονικό πεδίο των τεχνητών νευρωνικών δικτύων βασίστηκε στην βιολογία και συγκεκριμένα στον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Ο εγκέφαλος είναι ένας εξαιρετικά πολύπλοκος, μη γραμμικός, παράλληλος υπολογιστής. Έχει τη δυνατότητα να οργανώνει τα δομικά του στοιχεία, γνωστά ως νευρώνες, με τρόπο ώστε να εκτελούν συγκεκριμένους υπολογισμούς με ταχύτητα πολλαπλάσια από αυτή του γρηγορότερου ψηφιακού υπολογιστή που υπάρχει σήμερα. Εκτιμάται ότι υπάρχουν περίπου 86 δισεκατομμύρια νευρώνες στο νευρικό σύστημα, οι οποίοι συνδέονται μεταξύ τους με περίπου 10^{15} συνάψεις. Οι συνάψεις, ή νευρικές απολήξεις, είναι οι στοιχειώδεις δομικές και λειτουργικές μονάδες που παίζουν διαμεσολαβητικό ρόλο κατά τις αλληλεπιδράσεις μεταξύ των νευρώνων.

Μοντελοποίηση Νευρώνων

Όπως και ο ανθρώπινος εγκέφαλος, τα νευρωνικά δίκτυα αποτελούνται από νευρώνες και τις μεταξύ τους διασυνδέσεις ή συνάψεις (synapses). Ένας νευρώνας, συχνά αποκαλούμενος και ως κόμβος ή μονάδα, είναι μία μονάδα επεξεργασίας πληροφορίας η οποία είναι θεμελιώδης για την λειτουργία ενός νευρωνικού δικτύου. Στο σχηματικό διάγραμμα του Σχήματος 2.1 παρουσιάζεται το μαθηματικό μοντέλο ενός νευρώνα που αποτελεί τη βάση για τη σχεδίαση πολλών διαφορετικών αρχιτεκτονικών νευρωνικών δικτύων, κάποια από τα οποία θα μελετηθούν στη συνέχεια. Τα βασικότερα στοιχεία του μαθηματικού μοντέλου ενός νευρώνα είναι:

- Ένα σύνολο συνάψεων (Synapses) ή διασυνδέσεων. Η σύναψη είναι κατά βάση ένα σήμα εισόδου, το οποίο προέρχεται από κάποιο άλλο νευρώνα (γι αυτό ονομάζεται και διασύνδεση). Συμβολίζεται ως x_{kj}, όπου το k συμβολίζει

τον νευρώνα και το j αντιπορσωπεύει το της σύναψης. Ο συνολικός αριθμός συνάψεων (εισόδων) του νευρώνα συμβολίζεται με m.



Σχήμα 2.1: Μη γραμμικό μαθηματικό μοντέλο ενός νευρώνα.
Σχήμα από [60]

- Οι παράμετροι (Parameters) ή συναπτικά βάρη W_{kj}. Η τιμή των παραμέτρων αρχικοποιείται σε τυχαίες ή συγκεκριμένες τιμές και επαναπροσδιορίζεται κατά την εκπαίδευση του δικτύου. Κάθε σύναψη X_{kj} πολλαπλασιάζεται με το αντίστοιχο βάρος W_{kj}, ενώ ο συνολικός αριθμός των παραμέτρων του δικτύου είναι ίσος με τον συνολικό αριθμό συνάψεων m.

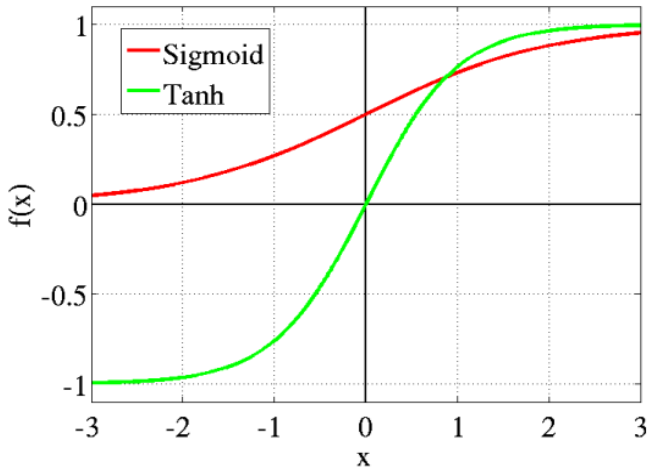
- Ένας γραμμικός συνδυαστής (Linear combiner) ή αθροιστής. Ο συνδυαστής αυτός αθροίζει τα σταθμισμένα με τα αντίστοιχα συναπτικά βάρη του νευρώνα σήματα εισόδου.

- Μία εξωτερικά εφαρμοζόμενη πόλωση (bias) b_k. Η πόλωση έχει ως αποτέλεσμα την μείωση (αν είναι αρνητική) ή αύξηση (αν είναι θετική) της εξόδου του αθροιστή. Η πόλωση του νευρώνα μπορεί να απενεργοποιηθεί (μηδενική πόλωση) και να μην συμπεριλαμβάνεται στην εκπαίδευση του δικτύου.

- Μία συνάρτηση ενεργοποίησης (Activation Function), η οποία χρησιμεύει στον περιορισμό του πλάτους του σήματος εξόδου y_k του νευρώνα. Ανάλογα με την έξοδο της συνάρτησης ενεργοποίησης, ένας νευρώνας μπορεί να ενεργοποιηθεί (activated) ή να μείνει αδρανής. Παρακάτω αναφέρονται κάποιες από τις σημαντικότερες συναρτήσεις ενεργοποίησης, οι οποίες χρησιμοποιούνται κατά κύριο λόγο στα περισσότερα νευρωνικά δίκτυα:

Sigmoid (Σιγμοειδής): Όπως φαίνεται στο Σχήμα 2.2, η σιγμοειδής συνάρτηση έχει το σχήμα του αγγλικού γράμματος S, υπολογίζεται από τον μαθηματικό τύπο: $\sigma(x) = \frac{1}{1+e^{-x}}$ και συνεπώς το εύρος τιμών της είναι [0, 1]. Λόγω του εύρους τιμών της, χρησιμοποιείται πολύ σε δίκτυα τα οποία θέλουν να προβλέψουν πιθανότητες ως την έξοδό τους, οι οποίες ως γνωστον έχουν εύρος τιμών [0, 1]. Η συνάρτηση είναι μονότονη, διαφορίσιμη καθώς και υπολογιστικά ακριβή λόγω των εκθετικών υπολογισμών που απαιτούνται.

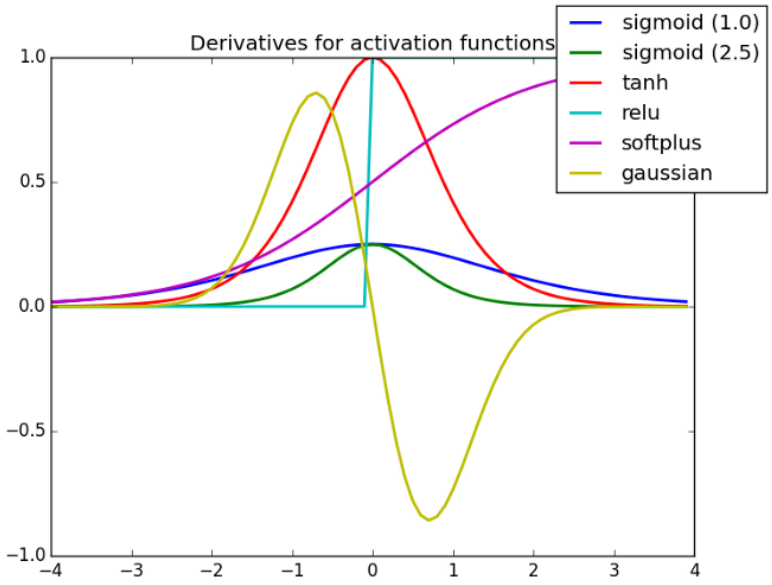
Tanh (Υπερβολική Εφαπτομένη): Η συνάρτηση υπερβολικής εφαπτομένης μοιάζει με την σιγμοειδή. Έχει και αυτή το σχήμα του αγγλικού γράμματος s όπως φαίνεται στο Σχήμα 2.2 και υπολογίζεται από τον μαθηματικό τύπο: $\tanh(x) = 2\sigma(2x) - 1$, όπου $\sigma(x)$ η σιγμοειδής συνάρτηση που ορίστηκε παραπάνω. Το εύρος τιμών της είναι το [-1, 1], είναι μονότονη και διαφορίσιμη συνάρτηση.



Σχήμα 2.2: Συναρτήσεις Ενεργοποίησης: Σιγμοειδής, Υπερβολική Εφαπτομένη. Σχήμα από [66]

Το κυρίως μειονέκτημα τόσο της σιγμοειδούς όσο και της υπερβολικής εφαπτομένης είναι το λεγόμενο Πρόβλημα των Κορεσμένων Παραγώγων (Saturated Gradients Problem), το οποίο εμφανίζεται σε βαθύτερα νευρωνικά δίκτυα. Το πρόβλημα αυτό προκύπτει από το γεγονός ότι από κάποια στιγμή και μετά κατά την εκπαίδευση, το γραμμικό κομμάτι του νευρώνα (δηλαδή η έξοδος του συνδυαστή), θα έχει τιμές είτε πολύ μεγάλες είτε αρκετά μικρές. Συνεπώς η είσοδος της συνάρτησης ενεργοποίησης (στην προκειμένη της tanh ή της sigmoid συνάρτησης) θα έχει μεγάλη κατά απόλυτο τιμή, η οποία αντιστοιχεί στο κορεσμένο πλέον κομμάτι των συναρτήσεων, στο οποίο όπως φαίνεται στο

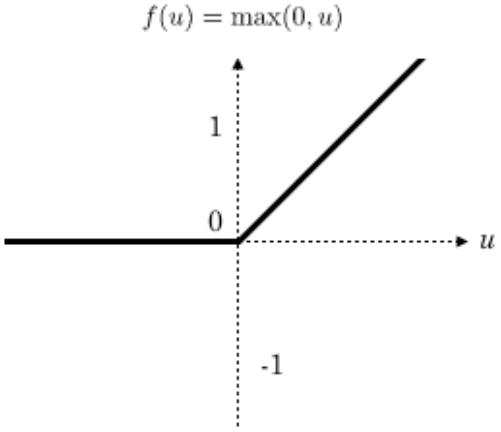
Σχήμα 2.3 οι τιμές των παραγώγων είναι αρκετά μικρές. Αυτό έχει ως αποτέλεσμα μετά από κάποιες επαναλήψεις της διαδικασίας εκπαίδευσης, το δίκτυο να μην μπορεί να εκπαιδευτεί περαιτέρω εξαιτίας της τόσο μικρής τιμής της παραγώγου των συναρτήσεων ενεργοποίησης των νευρώνων του δικτύου.



Σχήμα 2.3: Παράγωγοι Συναρτήσεων Ενεργοποίησης
Σχήμα από [66]

ReLU (Ανορθωμένη Γραμμική Μονάδα): Η ReLU (Rectified Linear Unit) είναι η πλέον συχνότερα χρησιμοποιούμενη συνάρτηση ενεργοποίησης σήμερα στα τεχνητά νευρωνικά δίκτυα. Χρησιμοποιείται σχεδόν σε όλα τα βαθιά και τα συνελικτικά νευρωνικά δίκτυα (CNNs). Περιγράφεται από τον μαθηματικό τύπο: $f(x) = \max(0, x)$ και συνεπώς το εύρος τιμών της είναι $[0, \infty]$, όπως φαίνεται και στο Σχήμα 2.4. Έτσι η έξοδος της συνάρτησης είναι ίση με την είσοδο αν αυτή έχει μη αρνητική τιμή, ενώ σε αντίθετη περίπτωση η έξοδος περιορίζεται στο μηδέν. Η συγκεκριμένη συνάρτηση ενεργοποίησης είναι τόσο δημοφιλής εξαιτίας του χαμηλού της υπολογιστικού κόστους, καθώς και λόγω του ότι αντιμετωπίζει το Πρόβλημα των Κορεσμένων Παραγώγων, όπως αυτό εξηγήθηκε προηγουμένως. Αυτό επιβεβαιώνεται από το Σχήμα 2.3, στο οποίο βλέπουμε πως σε αντίθεση με τις δύο προηγούμενες συναρτήσεις ενερ-

γοποίησης, η παράγωγος της ReLU δεν εμφανίζει τόσο μικρές τιμές για είσοδο μακριά από το μηδέν. Η ReLU παρουσιάζει επίσης ένα μειονέκτημα, γνωστό ως Dying ReLU Problem. Το πρόβλημα αυτό προκύπτει όταν η έξοδος του συνδυαστή του νευρώνα είναι αρνητική και συνεπώς με βάση τον ορισμό της ReLU, η έξοδος του νευρώνα θα είναι ίση με μηδέν. Όταν ένας νευρώνας βρεθεί σε μία τέτοια κατάσταση, δεν μπορεί να ανακάμψει με κάποιο τρόπο, καθώς η παράγωγος είναι επίσης μηδενική και συνεπώς δεν μπορεί να ανανεωθεί η τιμή των συναπτικών βαρών. Το πρόβλημα αυτό έρχεται να αντιμετωπίσει μία παραλλαγή της συνάρτησης αυτής, η Leaky ReLU, η οποία αντί να μηδενίζει τις τιμές $x < 0$ όπως η απλή ReLU, τις αντιστοιχίζει στις τιμές $y = \alpha x$ με $\alpha < 1$.



Σχήμα 2.4: Συνάρτηση Ενεργοποίησης ReLU
Σχήμα από [54]

Πλήρως Συνδεδεμένα Δίκτυα Πρόσθιας Τροφοδότησης

Στην ενότητα αυτή παρουσιάζεται η βασική αρχιτεκτονική των πλήρως συνδεδεμένων (fully connected) τεχνητών νευρωνικών δικτύων πρόσθιας τροφοδότης (feed-forward), δηλαδή ο τρόπος με τον οποίο δομούνται και συνδέονται οι νευρώνες, όπως αυτοί αναλύθηκαν στην προηγούμενη ενότητα, προκειμένου να δημιουργηθεί και να εκπαιδευτεί ολόκληρο το δίκτυο.

Η βασική δομή ενός τέτοιου δικτύου παρουσιάζεται γραφικά στο σχήμα 2.5. Όπως παρατηρούμε, το δίκτυο αποτελείται αρχικά από τρία διαφορετικά στρώματα (layers). Το πρώτο στρώμα ονομάζεται στρώμα εισόδου (input layer), και είναι το στρώμα το οποίο δέχεται πρώτο την πληροφορία εισόδου και την προωθεί βαθύτερα στο δίκτυο. Στη συνέχεια υπάρχουν τα λεγόμενα κρυφά στρώματα (hidden layers) του δικτύου. Στα στρώματα αυτά γίνεται πρακτικά η εξαγωγή όλων των απαραίτητων χαρακτηριστικών από το σήμα εισόδου προκειμένου να επιστευθχεί η κωδικοποίησή του. Ο αριθμός

των κρυφών στρωμάτων μπορεί να είναι από πολύ μικρός (ένα στρώμα) μέχρι οσοδήποτε μεγάλος απαιτεί το εκάστωτε πρόβλημα. Συνήθως τα πιο απαιτητικά προβλήματα απαιτούν τον υπολογισμό πιο πολύπλοκων χαρακτηριστικών και συνεπώς μεγαλύτερο αριθμό από κρυφά στρώματα στο εσωτερικό του δικτύου. Ένα δίκτυο θεωρείται βαθύ όταν αποτελείται από τουλάχιστον 2 κρυφά στρώματα. Τέλος έχουμε το στρώμα εξόδου (output layer) το οποίο είναι αυτό που πρακτικά δίνει την τελική απόφαση του δικτύου για την λύση του προβλήματος με τα συγκεκριμένα δεδομένα εισόδου.

Το δίκτυο του Σχήματος 2.5 λοιπόν αποτελείται από ένα στρώμα εισόδου, τρία κρυφά στρώματα και ένα στρώμα εξόδου. Φαίνεται πως κάθε νευρώνας ενός στρώματος συνδέεται με κάθε νευρώνα του αμέσως επόμενου στρώματος μέσω των παραμέτρων (ή βαρών) του δικτύου. Για το λόγο αυτό τα δίκτυα αυτά ονομάζονται πλήρως συνδεδεμένα δίκτυα πρόσθιας τροφοδότησης. Η πληροφορία ταξιδεύει από το στρώμα εισόδου προς το στρώμα εξόδου και κάθε νευρώνας συνδέεται με κάθε νευρώνα του επόμενου στρώματος. Όπως φάνηκε και στο Σχήμα 2.1 η έξοδος κάθε νευρώνα του δικτύου μπορεί να αποτυπωθεί μαθηματικά ως:

$$u_k = \sum_{j=1}^{m} x_{kj} w_{kj} \tag{2.1}$$

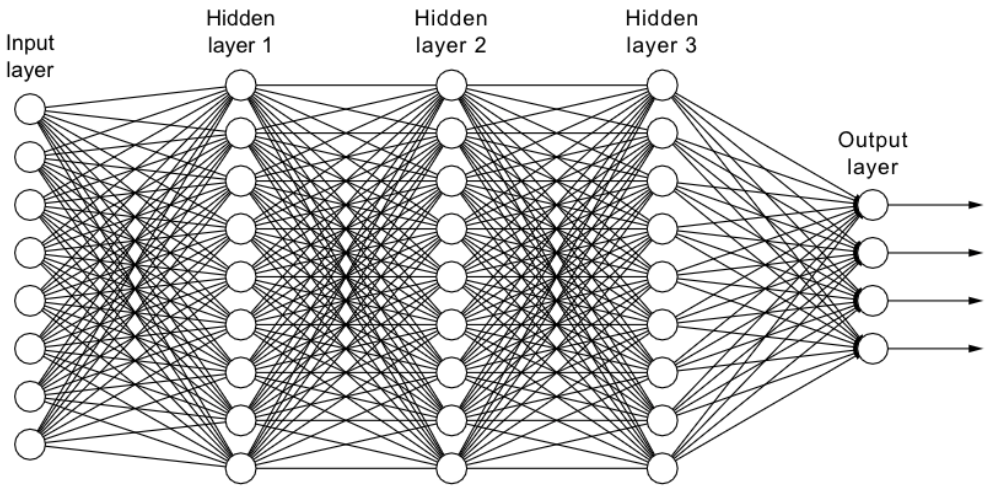
$$v_k = b_k + u_k \tag{2.2}$$

$$y_k = \phi(v_k) \tag{2.3}$$

2.2.3 Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Στο σημείο αυτό θα αναφερθούμε σε ένα διαφορετικό τύπο νευρωνικού δικτύου εμπρόσθιας τροφοδότησης, το οποίο αποτελεί ένα από τα αποτελεσματικότερα εργαλεία επίλυσης προβλημάτων στο πεδίο της Όρασης Υπολογιστών. Ο τύπος αυτός δικτύου ονομάζεται Συνελικτικό Νευρωνικό Δίκτυο (CNN). Όπως γίνεται φανερό και από το όνομα του, το δίκτυο αυτό χρησιμοποιεί την μαθηματική πράξη της συνέλιξης ως τον βασικό υπολογισμό για την επεξεργασία των δεδομένων και την εξαγωγή συμπερασμάτων για αυτά. Η δομή ενός τέτοιου δικτύου απεικονίζεται στο Σχήμα 2.6.

Όπως φαίνεται και στο σχήμα, το δίκτυο αυτό αποτελείται από ένα στρώμα εισόδου (input), μερικά ενδιάμεσα στρώματα και ένα στρώμα εξόδου (output), κατά αντιστοιχία με τα πλήρως συνδεδεμένα δίκτυα που παρουσιάστηκαν. Η μεγαλύτερη διαφορά των Συνελικτικών Νευρωνικών δικτύων έγγειται στο γεγονός πως η είσοδος του δικτύου είναι πλέον δισδιάστατη ή και τρισδιάστατη, στη μορφή των παραμέτρων του δικτύου αλλά και στον τρόπο διασύνδεσης των νευρώνων διαφορετικών στρωμάτων. Το δίκτυο παίρνει σαν είσοδο μία εικόνα (αν είναι δισδιάστατο) ή μία αλληλουχία εικόνων (αν είναι τρισδιάστατο), υπολογίζει από αυτήν όλα τα χρήσιμα χαρακτηριστικά για την εξαγωγή ενός τελικού συμπεράσματος και δίνει την αντίστοιχη έξοδο. Για παράδειγμα στο Σχήμα 2.6 βλέπουμε ένα δισδιάστατο Συνελικτικό Νευρωνικό Δίκτυο (CNN), το οποίο δέχεται σαν είσοδο μία εικόνα η οποία απεικονίζει κάποιο ψηφίο,



Σχήμα 2.5: Αρχιτεκτονική πλήρως συνδεδεμένου δικτύου εμπρόσθιας τροφοδότησης
Σχήμα από [16]

υπολογίζει κάποια χαρακτηριστικά της εικόνας με τη βοήθεια δισδιάστατων συνελίξεων και τελικά μπορεί να αποφανθεί για το ποιο ψηφίο (0-9) απεικονίζεται στην εικόνα εισόδου.

Τα διάφορα χαρακτηριστικά της εικόνας εισόδου που χρησιμοποιεί το δίκτυο προκειμένου να κάνει κάποια πρόβλεψη προκύπτουν από την συνέλιξη της εικόνας εισόδου με διάφορους πυρήνες. Η μαθηματική πράξη της δισδιάστατης συνέλιξης διακριτού χρόνου μίας εικόνας εισόδου x με ένα πυρήνα h στο σημείο (m, n) ορίζεται ως:

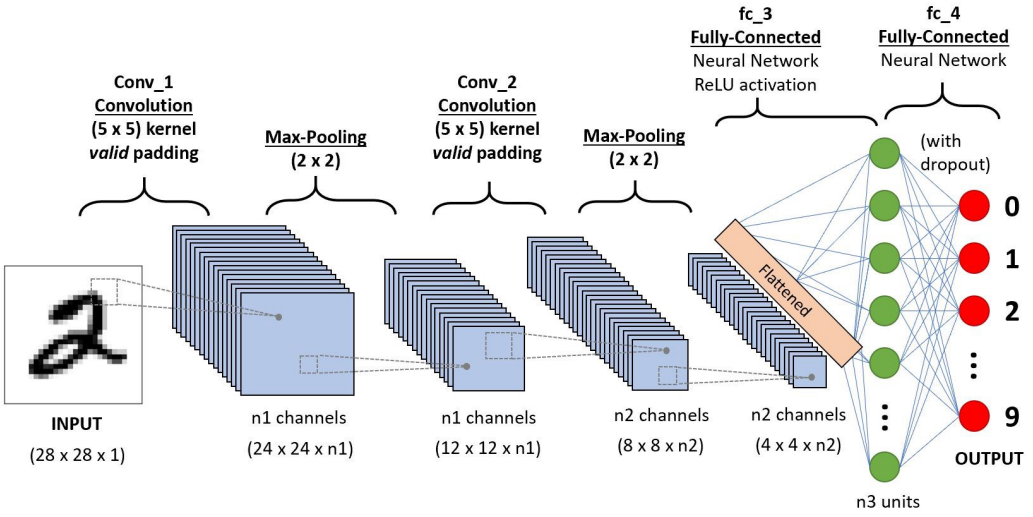
$$y[m,n] = h[m,n] * x[m,n] = \sum_j \sum_i h[i,j] * x[m-i, n-j] \tag{2.4}$$

όπου το εύρος των i, j εξαρτάται από το μέγεθος του κάθε πυρήνα h. Συνεπώς, το μεγαλύτερο ποσοστό των παραμέτρων (ή βαρών), των οποίων οι τιμές χρειάζεται να επαναπροσδιοριστούν κατά την διαδικασία της εκπαίδευσης είναι οι παράμετροι των πυρήνων με τους οποίους το δίκτυο πραγματοποιεί τις διάφορες συνελίξεις.

Ανάλογα με το πρόβλημα που προσπαθούμε κάθε φορά να λύσουμε και την μορφή που επιθυμούμε να έχει η έξοδος του δικτύου, ένα Συνελικτικό Νευωρνικό Δίκτυο μπορεί να είναι πλήρως συνελικτικό (fully convolutional) δηλαδή να αποτελείται μόνο από συνελικτικά στρώματα (convolutional layers) διατηρώντας τις διαστάσεις της εισόδου σε όλο το δίκτυο και συνεπώς παράγοντας πολυδιάστατη έξοδο. Διαφορετικά μπορεί από ένα σημείο και μετά να μειώνει τις διαστάσεις της πληροφορίας σε μία και

να έχει μερικά πλήρως συνδεδεμένα στρώματα (fully connected layers) στο τέλος του δικτύου, παράγοντας μονοδίασταση έξοδο (όπως το δίκτυο του Σχήματος 2.6).

Τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) έφεραν την επανάσταση σε πολλά κλασικά προβλήματα της Όρασης Υπολογιστών, στα οποία τα πλήρως συνδεδεμένα δίκτυα πετυχαίνουν κακή επίδοση, όπως η ανίχνευση αντικειμένων (object detection), η κατάτμηση εικόνας (image segmentation), η ταξινόμηση αντικειμένων (object classification) κ.ά. Η επιτυχία του συγκεκριμένου τύπου δικτύου σε τέτοια προβήματα, οφείλεται στη δυνατότητα υπολογισμού χωρο-χρονικών (spatio-temporal) χαρακτηριστικών που παρέχει η πράξη της συνέλιξης. Η δυνατότητα αυτή υπολογισμού χωρικών χαρακτηριστικών και συσχετίσεων ανάμεσα σε pixel τόσο κοντινότερων όσο και πιο απομακρυσμένων περιοχών της εικόνας, καθώς και η δυνατότητα εξαγωγής χρονικών χαρακτηριστικών ανάμεσα σε διαδοχικά αλλά και χρονικά πιο απομακρυσμένα καρέ δεν παρέχονται από τα κλασικά νευρωνικά δίκτυα και είναι ιδιαίτερα σημαντικές σε προβλήματα Όρασης Υπολογιστών, όπου οι χωρο-χρονικές εξαρτήσεις μεταξύ των pixel παίζουν σημαντικό ρόλο στην μοντελοποίηση του προβλήματος.



Σχήμα 2.6: Αρχιτεκτονική Συνελικτικών Νευρωνικών Δικτύων (CNNs)
Σχήμα από [61]

2.2.4 Εκπαίδευση Νευρωνικών Δικτύων

Στην ενότητα αυτή θα παρουσιαστεί ο τρόπος με τον οποίο ένα νευρωνικό δίκτυο προσδιορίζει αποτελεσματικά τις υπό εκπαίδευση παραμέτρους του προκειμένου να μοντελοποιήσει κάποιο πρόβλημα. Υπάρχουν αρχικά δύο είδη εκπαίδευσης ενός νευρωνικού δικτύου, η επιβλεπόμενη (supervised) και η μη-επιβλεπόμενη (unsupervised)

μάθηση. Στα πλαίσια της παρούσας εργασίας, χρησιμοποιήθηκε μόνο η επιβλεπόμενη μάθηση για τον σχεδιασμό και την εκπαίδευση των μοντέλων, η οποία χρησιμοποιείται μάλιστα και στην πλειοψηφία των προβλημάτων μηχανικής μάθησης.

Προκειμένου ένα νευρωνικό δίκτυο να εκπαιδευτεί με επιβλεπόμενη μάθηση, πρέπει να χρησιμοποιεί επισημειωμένα δεδομένα (annotated data) ως το σύνολο εκπαίδευσης του. Επισημειωμένα ονομάζονται τα δεδομένα στα οποία μαζί με την πληροφορία εισόδου, παρέχεται και η αντίστοιχη ετικέτα (label ή ground truth), η οποία πρακτικά είναι η επιθυμητή έξοδος του δικτύου. Έτσι το δίκτυο, αφού παράξει κάποια πρόβλεψη για συγκεκριμένα δεδομένα εισόδου, είναι σε θέση να αποφανθεί για την ορθότητα της πρόβλεψης του, αφού του παρέχονται τα πραγματικά δεδομένα και να προσαρμόσει τις παραμέτρους του με βάση της απόκλιση της πρόβλεψής του και της πραγματικής τιμής ώστε και να την ελαχιστοποιήσει.

Η προσαρμογή των παραμέτρων γίνεται με αλγόριθμους οι οποίοι χρησιμοποιούνται και εκτός του πεδίου των νευρωνικών δικτύων για την ανάπτυξη μοντέλων μηχανικής μάθησης. Ένας βασικός τέτοιος αλγόριθμος βελτιστοποίησης είναι ο αλγόριθμος Καθόδου Κλίσης (Gradient Descent). Σκοπός του αλγορίθμου είναι να ελαχιστοποιήσει την συνάρτηση σφάλματος (Loss function), την οποία χρησιμοποιεί το δίκτυο για να εκπαιδευτεί.

Συνάρτηση Σφάλματος

Κάθε νευρωνικό δίκτυο που εκπαιδεύεται χρησιμοποιώντας επιβλεπόμενη μάθηση, χρησιμοποιεί μία ή περισσότερες συναρτήσεις σφάλματος προκειμένου να προσδιορίζει την απόκλιση των προβλέψεών του από την πραγματική τιμή των δεδομένων και να βελτιστοποιήσει τις παραμέτρους του ώστε τελικά να προσεγγίσει το πρόβλημα με όσο το δυνατόν ακριβέστερο τρόπο. Η τιμή που υπολογίζεται από την συνάρτηση σφάλματος καθοδηγεί το δίκτυο σχετικά με το πως και πόσο πρέπει να μεταβάλει την κάθε εκπαιδεύσιμη παράμετρό του στην εκάστωτε επανάληψη. Οι συναρτήσεις σφάλματος μπορεί να είναι ειδικά σχεδιασμένες προκειμένου να συμβάλλουν στην ταχύτητα εκπαίδευσης και την ακρίβεια του δικτύου σε κάποιο συγκεκριμένο πρόβλημα (custom). Υπάρχουν επίσης κάποιες ευρέως χρησιμοποιούμενες συναρτήσεις σφάλματος με γενική εφαρμογή σε διάφορα προβλήματα μηχανικής μάθησης. Τέτοιες συναρτήσεις είναι:

- $L_1 = \sum_{i=1}^{n} |y_i - \hat{y}_i|$
- $L_2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- $MeanAbsoluteError(MAE) = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$
- $MeanSquareError(MSE) = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$

όπου y_i είναι η πραγματική τιμή και $\hat{y}_i$ η πρόβλεψη του δικτύου για την παρατήρηση i.

Κάθοδος Κλίσης (Gradient Descent)

Η συνάρτηση σφάλματος λοιπόν ενός δικτύου χρησιμοποιεί τα πραγματικά δεδομένα μαζί με τις αντίστοιχες προβλέψεις προκειμένου να υπολογιστεί και να ελαχιστοποιηθεί η απόκλιση του δικτύου από την πραγματικότητα. Η υπάρξη της πρόβλεψης του δικτύου $\hat{y}$ ως συνιστώσα στην συνάρτηση σφάλματος, κάνει την τελευταία να είναι μία συνάρτηση όλων των παραμέτρων του δικτύου οι οποίες χρησιμοποιήθηκαν προκειμένου να υπολογιστεί η συγκεκριμένη πρόβλεψη. Αυτό μας δίνει την δυνατότητα, υπολογίζοντας αναλυτικά τις εξισώσεις από τις οποίες προκύπτει η τελική πρόβλεψη, να υπολογίσουμε την παράγωγο $\frac{\partial \mathbf{J}}{\partial \mathbf{w_k}}$ της συνάρτησης σφάλματος $\mathbf{J}$ για όλες τις παραμέτρους του δικτύου $\mathbf{w_k}$, για κάθε επανάληψη k.

Συνεπώς για κάθε παράμετρο $\mathbf{w_k}$ μπορούμε να υπολογίσουμε την παράγωγο της συνάρτησης σφάλματος ως προς την παράμετρο αυτή, η οποία θα έχει μία μορφή παρόμοια με αυτή του Σχήματος 2.7. Σύμφωνα με το γνωστό θεώρημα του Fermat, αν μία συνάρτηση $\mathbf{J}$ παρουσιάζει τοπικό ακρότατο σε ένα σημείο $\mathbf{w_0}$ και είναι παραγωγίσιμη στο σημείο αυτό, τότε η τιμή της παραγώγου της στο σημείο $\mathbf{w_0}$ έχει τιμή μηδέν: $\frac{\partial \mathbf{J}}{\partial \mathbf{w_0}} = 0$. Προκειμένου λοιπόν να ελαχιστοποιήσουμε την συνάρτηση σφάλματος και συνεπώς να βελτιστοποιήσουμε τις προβλέψεις του δικτύου, αρκεί να κατευθύνουμε την παράγωγο της συνάρτησης σφάλματος ως προς τις διάφορες παραμέτρους του δικτύου σε κάποιο ελάχιστο (ιδανικά στο ολικό ελάχιστο).

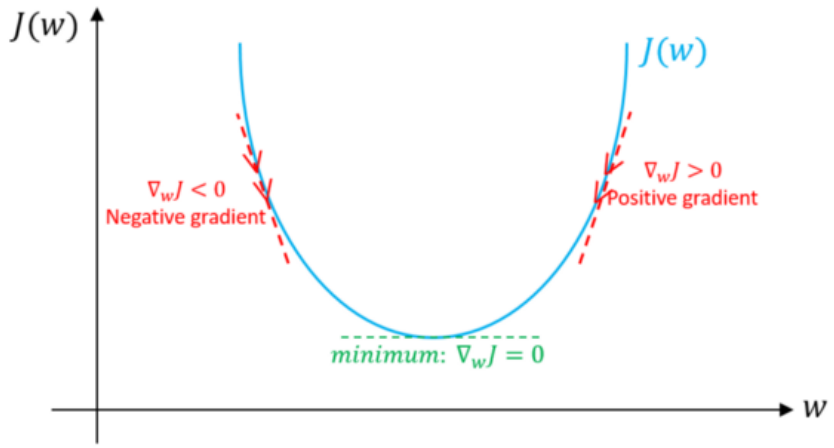
Για να το πετύχουμε αυτό, πρέπει να μεταβάλουμε τις τιμές των διάφορων παραμέτρων κατά κάποιο παράγοντα σε κάθε επανάληψη. Στο σημείο αυτό χρησιμοποιείται η υπερ-παράμετρος α αποκαλούμενη ως Ρυθμός Μάθησης (Learning Rate). Ο Ρυθμός Μάθησης α συνήθως κυμαίνεται από 10^{-1} μέχρι 10^{-5} και ρυθμίζει το πόσο γρήγορα θα μεταβάλλονται οι παράμετροι προκειμένου να επιτευχθεί το ελάχιστο της συνάρτησης σφάλματος. Με βάση τον αλγόριθμο Gradient Descent κάθε παράμετρος μεταβάλλεται με τον εξής τρόπο:

$$\mathbf{w_{k+1}} = \mathbf{w_k} - \alpha \frac{\partial \mathbf{J}}{\partial \mathbf{w_k}} \tag{2.5}$$

Από την εξίσωση καταλαβαίνουμε πως όταν η παράγωγος του σφάλματος στο σημείο $\mathbf{w_k}$ είναι αύξουσα, έχει δηλαδή θετική τιμή, τότε η παράμετρος θα μειωθεί κατά έναν παράγοντα $\alpha \frac{\partial \mathbf{J}}{\partial \mathbf{w_k}}$. Αντίστοιχα όταν η παράγωγος είναι φθίνουσα, τότε η παράμετρος θα αυξηθεί κατά έναν παράγοντα $\alpha \frac{\partial \mathbf{J}}{\partial \mathbf{w_k}}$, φέρνοντας σε κάθε περίπτωση τη συνάρτηση σφάλματος πιο κοντά στο ελάχιστο. Αυτή είναι και η βασική λογική του αλγορίθμου Gradient Descent, ο οποίος είναι ένας αλγόριθμος βελτιστοποίησης πρώτης τάξης, λαμβάνει υπόψιν του δηλαδή μόνο την πρώτη παράγωγο όταν ανανεώνει τις παραμέτρους του δικτύου.

Υπάρχουν τρεις διαφορετικές παραλλαγές του συγκεκριμένου αλγορίθμου, των οποίων η κύρια διαφορά έγγειται στο πόσα δεδομένα λαμβλανουμε υπόψιν σε κάθε επανάληψη του αλγορίθμου προκειμένου να επαναπροσδιορίσουμε τις παραμέτρους:

- **Batch Gradient Descent:** Ο αλγόριθμος εκτελείται και ανανεώνει τις παραμέτρους του δικτύου κάθε φορά που ολόκληρο το σύνολο των δεδομένων εκπα-



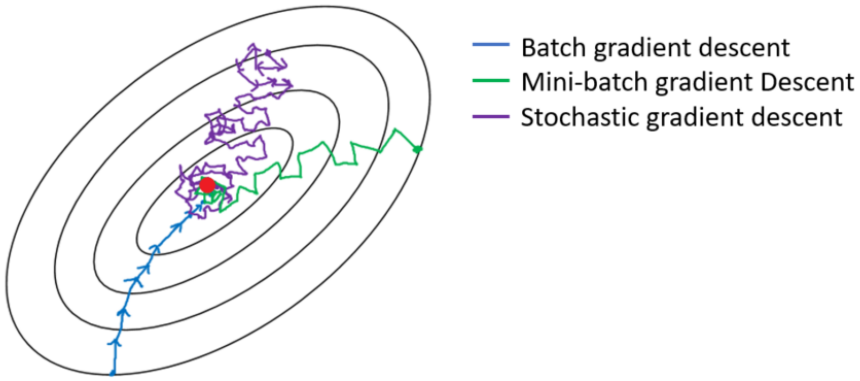
Σχήμα 2.7: Παράγωγος Συνάρτησης Σφάλματος
Σχήμα από [12]

ίδευσης έχει περάσει από το δίκτυο και συνεπώς έχει παραχθεί μία πρόβλεψη για κάθε δείγμα, χρησιμοποιώντας τις ίδιες τιμές παραμέτρων για κάθε πρόβλεψη. Έτσι στο τέλος κάθε επανάληψης όπου όλα τα δεδομένα έχουν επεξεργαστεί, κάθε παράμετρος ανανεώνεται με βάση τη Σχέση 2.5 για κάθε δεδομένο εισόδου.

- **Mini-batch Gradient Descent:** Σε αυτήν την παραλλαγή του αλγορίθμου, οι παράμετροι του δικτύου δεν περιμένουν να παραχθούν προβλέψεις για ολά τα διαθέσιμα δεδομένα, παρά μόνο για ένα μικρότερο υποσύνολο αυτών σταθερού μεγέθους (batch size) και ανανεώνουν την τιμή τους κάθε φορά με βάση τις προβλέψεις που παράχθηκαν για το υποσύνολο αυτό των δεδομένων. Έτσι για μία εποχή εκπαίδευσης του δικτύου, μόνο οι προβλέψεις δεδομένων που ανήκαν στο ίδιο batch έχουν παραχθεί χρησιμοποιώντας τις ίδιες τιμές παραμέτρων.

- **Stochastic Gradient Descent:** Τέλος, στην μορφή αυτή του αλγορίθμου, οι διάφορες παράμετροι του δικτύου ανανεώνονται κάθε φορά που παράγεται μία πρόβλεψη για κάποιο από τα δεδομένα εισόδου, το οποίο επιλέγεται τυχαία από το σύνολο των δεδομένων. Συνεπώς με βάση αυτή τη μορφή του αλγορίθμου, κάθε τυχαίο δείγμα των δεδομένων εισόδου περνάει από το δίκτυο χρησιμοποιώντας διαφορετικές τιμές παραμέτρων από τα προηγούμενα, καθώς αυτές ανανεώνονται με κάθε πέρασμα κάποιου δείγματος.

Και οι τρεις μορφές του αλγορίθμου χρησιμοποιούνται ευρέως για την επίλυση πολλών και διαφορετικών προβλημάτων σε αλγόριθμους μηχανικής αλλά και βαθιάς μάθησης. Κάθε μορφή έχει τόσο πλεονεκτήματα, όσο και μειονεκτήματα σε σχέση με το υπολογιστικό κόστος, την ανάγκη παρακολούθησης άλλων υπερπαραμέτρων (όπως ο ρυθμός μάθησης του δικτύου), την ταχύτητα καθώς και την δυνατότητα σύγκλισης στο ολικό ελάχιστο, με τον Mini-batch Gradient Descent να χρησιμοποιείται στην πλειοψηφία των προβλημάτων, εξαιτίας της μεγαλύτερης ταχύτητας του, την καλύτερη διαχείριση μεγάλων συνόλων δεδομένων καθώς και την καλύτερη αποφυγή των τοπικών ελαχίστων λόγω του θορύβου που προσθέτει στην διαδικασία της εκπαίδευσης. Στο Σχήμα 2.8 βλέπουμε τις τρεις διαφορετικές μορφές του αλγορίθμου καθώς κατευθύνονται προς το ελάχιστο.



Σχήμα 2.8: Οι τρεις παραλλαγές του αλγορίθμου Gradient Descent κατευθύνονται προς το ελάχιστο.
Σχήμα από [12]

Βλέπουμε πως καθώς ο αριθμός των δειγμάτων που λαμβάνονται υπόψιν στην ανανέωση των παραμέτρων μειώνεται, προστίθεται όλο και περισσότερος θόρυβος στην διαδικασία της εκπαίδευσης του δικτύου, με την στοχαστική μορφή του αλγορίθμου να εισάγει τόσο θόρυβο, που το δίκτυο γυρίζει γύρω από το επιθυμητό σημείου του ελάχιστου χωρίς να μπορεί να το πετύχει. Επίσης βλέπουμε πως ο αλγόριθμος Batch Gradient Descent φαίνεται να έχει την καλύτερη σύγκλιση σε λιγότερα βήματα στο ελάχιστο λόγω της έλλειψης θορύβου. Παρόλαυτα δεν χρησιμοποιείται τόσο συχνά λόγω των απαιτήσεων του σε χώρο μνήμης, σε υπολογιστικό χρόνο αλλά και εξαιτίας του ότι είναι επιρρεπής στο να «κολλήσει» σε κάποιο τοπικό ελάχιστο χωρίς να έχει τη δυνατότητα να ξεφύγει από αυτό λόγω της έλλειψης θορύβου. Έτσι χρησιμοποιείται κυρίως για την επίλυση σχετικά απλούστερων προβλημάτων όπου η συνάρτηση σφάλματος είναι κυρτή και συνεπώς το τοπικό ελάχιστο είναι και ολικό ελάχιστο και το σύνολο των δεδομένων εκπαίδευσης είναι μικρότερο.

Προβλήματα κατά την Εκπαίδευση ενός Δικτύου

Στην προηγούμενη ενότητα αναφέρθηκε ο βασικός αλγόριθμος που εκτελείται για την διαδικασία της εκπαίδευσης ενός βαθιού νευρωνικού δικτύου. Στο κομμάτι αυτό θα αναφερθούμε στα πιο βασικά προβλήματα που μπορεί να αποτελέσουν εμπόδιο στην εκπαίδευση του δικτύου και να αποτρέψουν την βέλτιστη μοντελοποίηση του προβλήματος, καθώς και στον τρόπο με τον οποίο προσπαθούμε να τα ξεπεράσουμε:

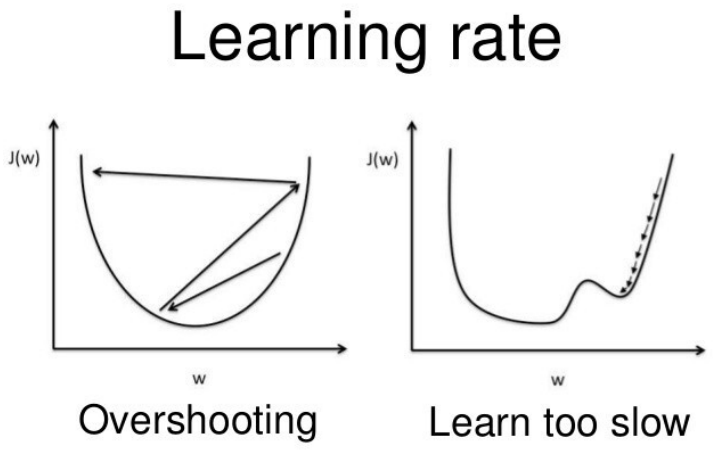
- **Εγκλωβισμός σε τοπικά ελάχιστα:** Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος καθόδου με βάση την κλίση, μπορεί να εγκλωβιστεί σε κάποιο τοπικό ελάχιστο της συνάρτησης σφάλματος, ανάλογα με την μορφή που αυτή έχει αλλά και την παραλλαγή του αλγορίθμου που χρησιμοποιείται για την εκπαίδευση του δικτύου. Ένας τρόπος να αποφευχθεί ο εγκλωβισμός σε κάποιο τοπικό ελάχιστο, προκειμένου να μπορέσει ο αλγόριθμος να εντοπίσει το ολικό ελάχιστο της συνάρτησης είναι να χρησιμοποιηθεί ένας ακόμα παράγοντας στην εκπαίδευση, τον οποίο ονομάζουμε Momentum. Ο αλγόριθμος Gradient Descent με Momentum μετατρέπει την Σχέση 2.5 σε:

$$\mathbf{w_{k+1}} = \mathbf{w_k} - \alpha \frac{\partial \mathbf{J}}{\partial \mathbf{w_k}}^{t} - \gamma \frac{\partial \mathbf{J}}{\partial \mathbf{w_k}}^{t-1} \tag{2.6}$$

  με τον τελευταίο παράγοντα της εξίσωσης να αντιπροσωπεύει το Momentum. Με τον τρόπο αυτό σε κάθε ανανέωση των συναπτικών βαρών $\mathbf{w}$ στο βήμα k, λαμβάνεται υπόψιν και η κλίση που είχε η συνάρτηση στο προηγούμενο βήμα του αλγόριθμου, η οποία μάλιστα μπορεί να είναι και αντίθετη από την κλίση του αλγόριθμου στο τρέχον βήμα, στην περίπτωση που έχουμε ξεπεράσει το σημείο του τοπικού ελαχίστου. Αυτό μπορεί να βοηθήσει στην αποφυγή κυρίως των μικρότερων τοπικών ελαχίστων αν θεωρήσουμε πως στις περισσότερες περιπτώσεις, ένα τοπικό ελάχιστο είναι σχετικά μικρού εύρους, κάνοντας τον παράγοντα $\gamma \frac{\partial \mathbf{J}}{\partial \mathbf{w}}^{t-1}$ αρκετό ώστε ο αλγόριθμος να το προσπεράσει. Αντίθετα σε ένα μεγαλύτερου εύρους τοπικό ελάχιστο ο αλγόριθμος δεν θα έχει τοσο αποτελεσματική συμπεριφορά, καθώς ο παράγοντας δεν είναι αρκετά μεγάλος ώστε ο αλγόριθμος να το προσπεράσει. Παρόλαυτά σε ένα μεγαλύτερο τοπικό ελάχιστο είμαστε αρκετές φορές ικανοποιητικά κοντά και στο ολικό ελάχιστο.

- **Προσαρμογή ρυθμού μάθησης:** Ένα ακόμη συνηθισμένο πρόβλημα είναι η αρχική επιλογή αλλά και η προσαρμογή του ρυθμού μάθησης α κατά την διάρκεια της εκπαίδευσης. Αρκετά μικρός ρυθμός μάθησης σημαίνει μικρότερη μεταβολή των παραμέτρων σε κάθε επανάληψη και συνεπώς περισσότερος απαιτούμενος χρόνος εκπαίδευσης του μοντέλου. Μεγάλος ρυθμός μάθησης α οδηγεί σε πολύ μεγάλη μεταβολή των παραμέτρων του δικτύου με αποτέλεσμα το δίκτυο να αναπηδά γύρω από το σημείο του ελαχίστου χωρίς να μπορεί ποτέ να το προσεγγίσει αποτελεσματικά όπως φαίνεται στο Σχήμα 2.9. Η πιο συνηθισμένη πρακτική που οδηγεί τόσο σε γρήγορη αλλά και βέλτιστη εκπαίδευση του μοντέλου είναι η χρήση διαφορετικού ρυθμού μάθησης ανάλογα με

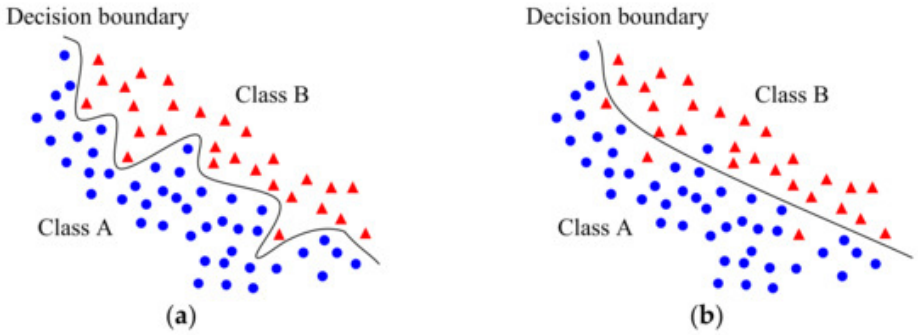
την φάση που βρίσκεται η εκπαίδευση του μοντέλου. Συγκεκριμένα στην αρχή που το δίκτυο είναι συνήθως μακριά από το επιθυμητό ελάχιστο χρησιμοποιείται μεγαλύτερος ρυθμός μάθησης ώστε να μπορούν να αποφευχθούν και τα σημεία τοπικών ελαχίστων, ενώ μετά από κάποιες επαναλήψεις ο ρυθμός μάθησης μειώνεται ώστε να γίνονται μικρότερα βήματα και να ποσεγγιστεί ακριβέστερα το σημείο ελαχίστου. Η τεχνική αυτή ονομάζεται Multi-step Learning Rate.



Σχήμα 2.9: Επίδραση του ρυθμού μάθησης στην εκπαίδευση του δικτύου. Σχήμα από [35]

- **Υπερ-προσαρμογή (Overfitting):** Ένα ακόμη συνιθησμένο πρόβλημα που καλείται ένα νευρωνικό δίκτυο να αντιμετωπίσει κατά την εκπαίδευση του είναι το πρόβλημα της υπερ-προσαρμογής. Το πρόβλημα αυτό σχετίζεται με την δυνατότητα που έχει το δίκτυο μετά την εκπαίδευση του να γενικεύσει και να λειτουργεί αποτελεσματικά για πολλά διαφορετικά δείγματα του προβλήματος που προσπαθεί να μοντελοποιήσει, τα οποία πιθανότατα να μην έχει συναντήσει προηγουμένως στα σύνολα δεδομένων με τα οποία εκπαιδεύτηκε. Όταν ένα μοντέλο εκπαιδεύεται για πολλές επαναλήψεις πάνω στα ίδια ή πολύ παρόμοια δεδομένα ενός συνόλου, τότε υπερ-προσαρμόζεται σε αυτά με αποτέλεσμα να έχει αλλάξει τα συναπτικά του βάρη λαμβάνοντας υπόψιν ένα πολύ συγκεκριμένο τύπο δειγμάτων, με αποτέλεσμα να πάρα πολύ καλή επίδοση σε αυτά τα δεδομένα αλλά να μην μπορεί να διαχειριστεί αποτελεσματικά δεδομένα που δεν έχει συναντήσει στο παρελθόν. Το πρόβλημα αυτό φαίνεται και γραφικά στο Σχήμα 2.10. Στο Σχήμα (a) βλέπουμε πως το περιθώριο απόφασης έχει υπερ-προσαρμοστεί στα συγκεκριμένα δεδομένα, ενώ στο Σχήμα (b) έχει γενικεύσει καλύτερα, μοντελοποιώντας το γενικότερο πρόβλημα και όχι τόσο τα συγκεκριμένα δεδομένα. Υπάρχουν κάποιες τεχνικές που βοηθούν στην

εντιμετώπιση του συγκεκριμένου προβλήματος. Τέτοιες τεχνικές είναι η εκπαίδευση του δικτύου για λιγότερες εποχές, η χρήση του Dropout Layer και η εφαρμογή Επαύξησης των Δεδομένων Εκπαίδευσης (Data Augmentation). Το Dropout είναι πρακτικά η τυχαία απενεργοποίηση κάποιων νευρώνων του δικτύου σε κάθε επανάληψη. Αυτό έχει ως αποτέλεσμα να αλλάζει με τυχαίο τρόπο η ακριβής αρχιτεκτονική και η διασύνδεση των στρωμάτων του δικτύου και συνεπώς ο τρόπος ανανέωσης των βαρών σε κάθε επανάληψη. Επίσης κάθε νευρώνας αναγκάζεται να αναλάβει περισσότερη ευθύνη για την μοντελοποίηση του προβλήματος και το δίκτυο μαθαίνει εν τέλει μία πιο αραιή και γενικευμένη αναπαράσταση των δεδομένων βοηθώντας σε καλύτερη γενίκευση. Η τεχνική της Επαύξησης Δεδομένων είναι πρακτικά ο εμπλουτισμός των διαθέσιμων δεδομένων εκπαίδευσης με την εφαρμογή κάποιων γραμμικών ή μη μετασχηματισμών, όπως η αντίθεση της εικόνας, η περιστροφή, η περικοπή κ.ά. Αυτό έχει ως αποτέλεσμα την συνεχή τροφοδότηση του δικτύου με όσο το δυνατόν νέα και διαφορετικά δεδομένα, κάτι που αποτρέπει το δίκτυο από την υπερπροσαρμογή σε συγκεκριμένα δεδομένα και την αδυναμία γενίκευσης.



Σχήμα 2.10: Υπερ-προσαρμογή δικτύου σε πρόβλημα δυαδικής ταξινόμησης. Σχήμα από [53]

- **Εξαφάνιση/Εκτόξευση της κλίσης (Vanishing/Exploding Gradient):** Όπως αναφέρθηκε προηγουμένως, προκειμένου να επαναπροσδιοριστούν σε κάθε βήμα οι παράμετροι του δικτύου και να οδηγηθούμε στην βέλτιστη λύση, απαιτείται ο υπολογισμός της παραγώγου της συνάρτησης σφάλματος ως προς την κάθε παράμετρο. Για να το πετύχουμε αυτό χρησιμοποιούμε τον κανόνα της αλυσίδας, κάτι το οποίο οδηγεί στον πολλαπλασιαμό πολλών επιμέρους παραγώγων, με τον αριθμό των όρων του πολλαπλασιαμού (μήκος αλυσίδας) να αυξάνεται καθώς μετακινούμασε από το τέλος του δικτύου προς τα πιο αρχικά επίπεδα. Αυτό έχει ως αποτέλεσμα σε περίπτωση που οι τιμές των επιμέρους παραγώγων είναι ακραίες (πολύ μικρές ή πολύ μεγάλες) να καταλήγουμε αντίστοιχα σε ένα αποτέλεσμα όπου η παράγωγος της συνάρτησης σφάλματος ως προς κάποια παράμετρο είναι πάρα πολύ μικρή (εξαφάνιση κλίσης) ή πάρα

πολύ μεγάλη (εκτόξευση κλίσης). Αυτό με τη σειρά του οδηγεί στην αδυναμία του μοντέλου να αλλάξει κάποιες παραμέτρους στην περίπτωση της εξαφάνισης της κλίσης ή να στην απροσδιόριστη και ανεξέλεγκτη αλλαγή των παραμέτρων στην περίπτωση της εκτόξευσης της κλίσης. Για την επίλυση αυτού του προβλήματος χρησιμοποιούμε τα στρώματα Κανονικοποίησης της Παρτίδας (Batch Normalization), τα οποία χρησιμοποιούνται σε διάφορα σημεία του δικτύου, προκειμένου να κανονικοποιήσουν και να μεταβάλλουν την κατανομή των δειγμάτων κάθε παρτίδας.

2.3 Μετρικές Εμφάνειας

Όπως σε όλα τα προβλήματα, έτσι και στο πρόβλημα της Εμφάνειας έχουν αναπτυχθεί ορισμένες μετρικές, οι οποίες υπολογίζονται και χρησιμοποιούνται ώστε να αξιολογηθούν οι διάφορες μέθοδοι ως προς την αποτελεσματικότητα και την ακρίβειά τους. Στην ενότητα αυτή θα παρουσιαστούν οι πέντε σημαντικότερες και πιο συχνά χρησιμοποιούμενες τέτοιες μετρικές, όπως ορίζονται και εξηγούνται αναλυτικά στο [3] και οι οποίες είναι και οι μετρικές που χρησιμοποιήθηκαν στην Διπλωματική Εργασία για την αξιολόγηση του προτεινόμενου μοντέλου και την σύγκρισή του με άλλες μεθόδους. Ορίζουμε ως P τον συνεχή χάρτη εμφάνειας, όπως αυτός έχει προβλεφθεί από κάποια μέθοδο, ενώ ως Q τον πραγματικό χάρτη εμφάνειας. Συγκεκριμένα με Q^D ορίζουμε τον πυκνό (συνεχή) πραγματικό χάρτη, ενώ με Q^B τον δυαδικό (διακριτό) πραγματικό χάρτη εμφάνειας. Οι μετρικές μπορούν να οριστούν ως εξής:

- **Pearson's Correlation Coefficient (CC)**: Ο Συντελεστής Συσχέτισης του Pearson, αποκαλούμενος και Γραμμικός Συντελεστής Συσχέτισης είναι μία στατιστική μέθοδος, η οποία χρησιμοποιείται σε διάφορα πεδία της επιστημής προκειμένου να μετρηθεί πόσο συσχετισμένες ή εξαρτώμενες είναι δύο μεταβλητές. Μπορεί να χρησιμοποιηθεί προκειμένου να ερμηνευθούν ο χάρτης εμφάνειας και ο συνεχής χάρτης προσοχής P και Q^D αντίστοιχα ως τυχαίες μεταβλητές και να μετρηθεί η γραμμική σχέση τους:

$$CC(P, Q^D) = \frac{\sigma(P, Q^D)}{\sigma(P) \times \sigma(Q^D)} \tag{2.7}$$

  όπου $\sigma(P, Q^D)$ είναι η συνδυακύμανση των P και Q^D. Η CC μετρική είναι συμμετρική από την άποψη ότι επηρεάζεται το ίδιο τόσο από τα λανθασμένα θετικά όσο και από τα λανθασμένα αρνητικά αποτελέσματα.

- **Normalized Scanpath Saliency (NSS)**: Η μετρική της Εμφάνειας του Κανονικοποιημένου Οπτικού Μονοπατιού, NSS προτάθηκε για το πρόβλημα της Εμφάνειας ως μία απλή μετρική της αντιστοιχίας μεταξύ των χαρτών Εμφάνειας και της πραγματικότητας, η οποία υπολογίζεται ως η μέση κανονικοποιημένη εμφάνεια σε σημεία προσοχής. Η NSS είναι ευαίσθητη στα λανθασμένα θετικά

αποτελέσματα, σε σχετικές διαφορές στην Εμφάνεια στο σύνολο της εικόνας και γενικά μονότονους μετασχηματισμούς. Αντίθετα είναι ιδιαίτερα ανθεκτική σε γραμμικούς μετασχηματισμούς, όπως για παράδειγμα η μετατόπιση της αντίθεσης. Για ένα χάρτη Εμφάνειας P και τον αντίστοιχο πραγματικό δυαδικό χάρτη προσοχής Q^B υπολογίζεται ως:

$$NSS(P, Q^B) = \frac{1}{N} \sum_j \sum_i \bar{P_{i,j}} \times Q^B_{i,j} \tag{2.8}$$

όπου $N = \sum_i Q^B_i$ και $\bar{P} = \frac{P - \mu(P)}{\sigma(P)}$, με i να είναι ο δείκτης του κάθε σημείου (pixel) της εικόνας και N ο συνολικός αριθμός των σημείων που υπάρχει εστίαση της προσοχής.

- **Area under ROC Curve (AUC)**: Δεδομένου του προβλήματος της Εμφάνειας, όπου ο στόχος είναι η πρόβλεψη των σημείων εστίασης της προσοχής, ένας χάρτης Εμφάνειας μπορεί να ερμηνευθεί ως ένας ταξινομητής του οποίου τα σημεία είναι η όχι σημεία εστίασης. Στην θεωρία των σημάτων, η καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη (Receiver Operating Characteristic-ROC) μετρά το συμβιβασμό (tradeoff) μεταξύ των αληθώς και λανθασμένων θετικών αποτελεσμάτων σε διάφορα κατώφλια. Η περιοχή κάτω από την καμπύλη (AUC) είναι η πιο γνωστή μετρική για την αξιολόγηση των χαρτών Εμφάνειας. Ο χάρτης εμφάνειας χρησιμοποιείται ως ένας ταξινομητής από σημεία εστίασης σε διάφορες τιμές κατωφλιών και η καμπύλη σχηματίζεται με την μέτρηση των ποσοστών των αληθώς και λανθασμένων θετικά αποτελεσμάτων κάτω από κάθε ταξινομητή. Οι διαφορετικές παραλλαγές και υλοποιήσεις της μετρικής αυτής διαφέρουν στον τρόπο που υπολογίζονται τα αληθώς και λανθασμένα θετικά σημεία.

- **AUC-Judd (AUC-J)**: Μία παραλλαγή της AUC μετρικής, η οποία προτάθηκε από τους Judd et al. στο [30] είναι η AUC-Judd. Για ένα δεδομένο κατώφλι, το ποσοστό των αληθώς θετικών σημείων είναι ο λόγος των αληθώς θετικών προς τον συνολικό αριθμό των σημείων εστίασης, όπου τα αληθώς θετικά σημεία είναι τιμές του χάρτη Εμφάνειας μεγαλύτερες από το κατώφλι σε σημεία με εστίαση προσοχής. Το ποσοστό των λανθασμένα θετικών σημείων είναι ο λόγος των λανθασμένα θετικών σημείων προς τον συνολικό αριθμό εστιασμένων σημείων σε ένα δεδομένο κατώφλι, όπου λανθασμένα θετικά είναι τα σημεία με τιμές μεγαλύτερες του κατωφλιού σε μη εστιασμένα σημεία.

- **Shuffled AUC (sAUC)**: Η Ανακατεμένη μετρική AUC δειγματοληπτεί τα αρνητικά σημεία από περιοχές εστιασμού από άλλες εικόνες αντί για τυχαία ομοιόμορφα δείγματα. Αυτό έχει ως αποτέλεσμα της δειγματοληψίας αρνητικών σημείων κυρίως από το κέντρο της εικόνας, καθώς παίρνοντας το μέσο όρο από πολλές εικόνες καταλήγει στην φυσική εμφάνιση μίας κεντρικής Γκαουσιανής κατανομής. Όσο περισσότερη έμφαση δίνει ένα μοντέλο στο κέντρο της εικόνας

για τις προβλέψεις του (π.χ. Center bias) τόσο χειρότερα αποδίδει με βάση την μετρική sAUC, καθώς αυτή δίνει μικρότερη βάση σε αυτές.

- **Similarity or histogram intersection (SIM)**: Η μετρική της Ομοιότητας, ή αλλιώς η Τομή των Ιστογραμμάτων μετρά την ομοιότητα μεταξύ δύο κατανομών, οι οποίες εκλαμβάνονται ως ιστογράμματα. Αρχικά προτάθηκε σαν μία μετρική για ταίριασμα εικόνων με βάση το χρώμα και το περιεχόμενο και έχει σήμερα γίνει δημοφιλής στο πεδίο της Εμφάνειας ως μία απλή σύγκριση μεταξύ ζευγαριών χαρτών Εμφάνειας. Η μετρική αυτή υπολογίζεται ως το άθροισμα των ελάχιστων τιμών σε κάθε σημείο της εικόνας, αφού οι χάρτες εισόδου έχουν κανονικοποιηθεί. Δεδομένου ενός χάρτη Εμφάνειας P και και του συνεχούς χάρτη προσοχής Q^D έχουμε:

$$SIM(P, Q^D) = \sum_i \min(P_i, Q_i^D) \tag{2.9}$$

όπου $\sum_i P_i = \sum_i Q_i^D = 1$. Για $SIM = 1$ έχουμε πως οι κατανομές είναι ίδιες, ενώ αντίθετα αν $SIM = 0$ τότε οι κατανομές δεν παρουσιάζουν καμία επικάλυψη.

Κεφάλαιο 3

Σύνολα Δεδομένων και Προεπεξεργασία

Στο κεφάλαιο αυτό θα γίνει αναφορά στα δεδομένα που χρησιμοποιήθηκαν για την εκπόνηση της Διπλωματικής Εργασίας, τόσο για τον αποτελεσματικό υπολογισμό του βάθους σε δισδιάστατες εικόνες όσο και για την βέλτιστη μοντελοποίηση του προβλήματος του Saliency. Επίσης θα αναφερθούμε στην απαραίτητη προεπεξεργασία που έγινε στα δεδομένα προκειμένου να χρησιμοποιηθούν στην Διπλωματική Εργασία.

3.1 Δεδομένα Βάθους

Για την εκπόνηση της διπλωματικής εργασίας, χρειάστηκε να γίνει εκτίμηση του σχετικού βάθους των αντικειμένων που απεικονίζονται σε μία δισδιάστατη εικόνα. Για το σκοπό αυτό χρησιμοποιήθηκαν και αξιολογήθηκαν ως προς την προσφορά τους στο πρόβλημα της Εμφάνειας, τρεις διαφορετικές μέθοδοι εξαγωγής βάθους από δισδιάστατες εικόνες με χρήση Πλήρως Συνελικτικών Νευρωνικών Δικτύων (CNNs).

Δεδομένα Εκπαίδευσης

NYU Depth Dataset V2

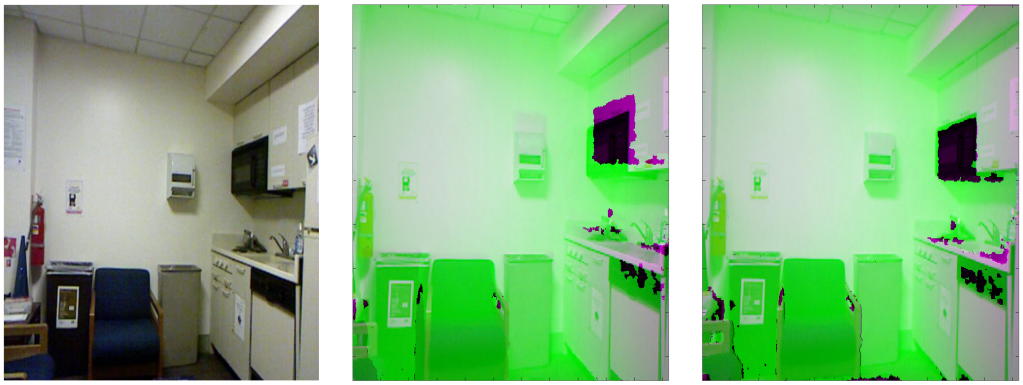
Η πρώτη μέθοδος που εξετάστηκε [22] απαιτούσε την υλοποίηση του μοντέλου και την εκπαίδευσή του απο την αρχή με βάση την δημοσίευση στην οποία προτάθηκε, καθώς ο κώδικας και το εκπαιδευμένο μοντέλο δεν είναι δημόσια διαθέσιμα. Για το σκοπό αυτό χρειάστηκε να δημιουργηθεί το κατάλληλο σύνολο δεδομένων και στην κατάλληλη μορφή ώστε να μπορέσει να εκτελεστεί η εκπαίδευση του μοντέλου. Το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του συγκεκριμένου μοντέλου είναι το NYU Depth Dataset V2 [49], το ίδιο που χρησιμοποιήθηκε και για την εκπαίδευση του μοντέλου στην αντίστοιχη δημοσίευση.

Αυτό το σύνολο δεδομένων περιέχει μία σειρά από βίντεο που απεικονίζουν πολλές διαφορετικές σκηνές εσωτερικού χώρου, καταγεγραμμένες τόσο από την RGB όσο και από τον αισθητήρα βάθους της Microsoft Kinect, μίας περιφερειακής συσκευής από αισθητήρες κίνησης. Οι σκηνές αυτές προέρχονται από 464 διαφορετικά σκηνικά από 3 πόλεις. Για την εκπαίδευση του δικτύου χρησιμοποιήθηκε ο επίσημος διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης και επαλήθευσης (train/test split), με τις 249 σκηνές να χρησιμοποιούνται για την εκπαίδευση και τις υπόλοιπες 215 σκηνές να χρησιμοποιούνται για την επαλήθευση. Από όλο το σύνολο των διαθέσιμων δεδομένων, μόνο ένας πολύ μικρός αριθμός 1449 δειγμάτων έχουν επεξεργαστεί την αρχική καταγραφή του αισθητήρα βάθους ώστε να μπορεί να χρησιμοποιηθεί και την έχουν αντιστοιχίσει με την RGB καταγραφή της κάμερας ώστε τα δεδομένα να είναι επισημειωμένα με το πραγματικό βάθος και να μπορούν να χρησιμοποιηθούν για την εκπαίδευση κάποιου δικτύου με επιβλεπόμενη μάθηση.

Στο σύνολό τους όμως τα δεδομένα αυτού του συνόλου αριθμούνται σε 407.024 δείγματα αντιστοιχίας RGB καταγραφής και πληροφορίας Βάθους, κάνοντας τα 1449 επισημειωμένα δείγματα ένα πολύ μικρό ποσοστό από την συνολική πληροφορία που μπορεί να προσφέρει το συγκεκριμένο σύνολο δεδομένων. Συνεπώς προέκυψε η ανάγκη προεπεξεργασίας ενός μεγαλύτερου υποσυνόλου των ακατέργαστων (raw) δεδομένων, προκειμένου αυτά να επισημειωθούν και να είναι σε κατάλληλη για την εκπαίδευση του μοντέλου μορφή. Για το σκοπό αυτό, στα πλαίσια της διπλωματικής, χρησιμοποιήθηκε ολόκληρο το σύνολο των ακατέργαστων δεδομένων, στο οποίο εφαρμόσαμε ομοιόμορφη χρονική δειγματοληψία από κάθε διαθέσιμη ακολουθία βίντεο, με αποτέλεσμα να προκύψουν συνολικά 35.698 δείγματα RGB-Βάθους αντιστοιχιών, διαστάσεων 680×480. Για την επεξεργασία τους χρησιμοποιήσαμε το επίσημο πακέτο εργαλείων που δημοσιεύθηκε για τον σκοπό αυτό από τους δημιουργούς του συνόλου δεδομένων. Τόσο τα δεδομένα όσο και τα εργαλεία επεξεργασίας έιναι διαθέσιμα στο σύνδεσμο: `https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html`.

Το διαθέσιμο πακέτο εργαλείων περιέχει χρήσιμες συναρτήσεις και αλγορίθμους για εκτέλεση σε περιβάλλον MATLAB, οι οποίες φροντίζουν για την διόρθωση συγκεκριμένων χαρακτηριστικών των δεδομένων τα οποία αποτελούν εμπόδιο στην χρήση των δεδομένων για την εκπαίδευση νευρωνικών δικτύων:

- **Σχετική μετατόπιση RGB - Βάθους:** Στο Σχήμα 3.1 βλέπουμε μία τυχαία εικόνα από το σύνολο δεδομένων, πάνω στην οποία έχει προβληθεί η αντίστοιχη ακατέργαστη καταγραφή του αισθητήρα βάθους. Τα σημεία που απεικονίζονται με πράσινο χρώμα, είναι τα σημεία στα οποία υπήρχε καταγραφή για την εκτίμηση του βάθους στο συγκεκριμένο σημείο της εικόνας. Τα σημεία που εμφανίζονται με πιο σκούρο χρώμα (μωβ-μαύρο) είναι τα σημεία στα οποία η τιμή του βάθους δεν καταγράφηκε από τον αισθητήρα και είναι συνεπώς άγνωστη. Συγκρίνοντας τις δύο εικόνες, βλέπουμε πως υπάρχει μία σχετική μετατόπιση της εικόνας και του Βάθους (Σχήμα 3.1(β)), η οποία διορθώθηκε (Σχήμα 3.1(γ)) με την χρήση του κατάλληλου αλγορίθμου του πακέτου εργαλειών, λαμβάνοντας υπόψιν και τις διάφορες παραμέτρους των αιθητήρων (παραμόρφωση, περιστροφή κ.ά.).

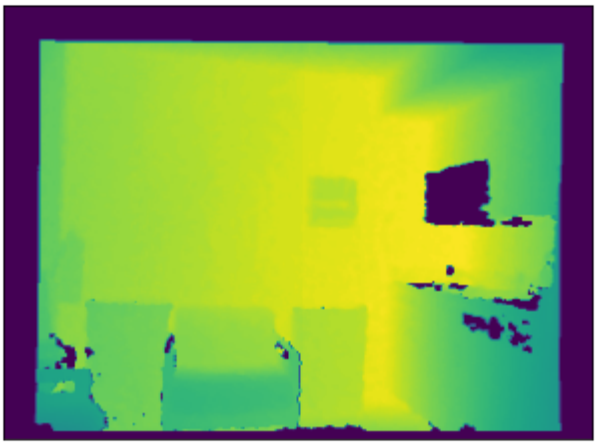


(α) (β) (γ)

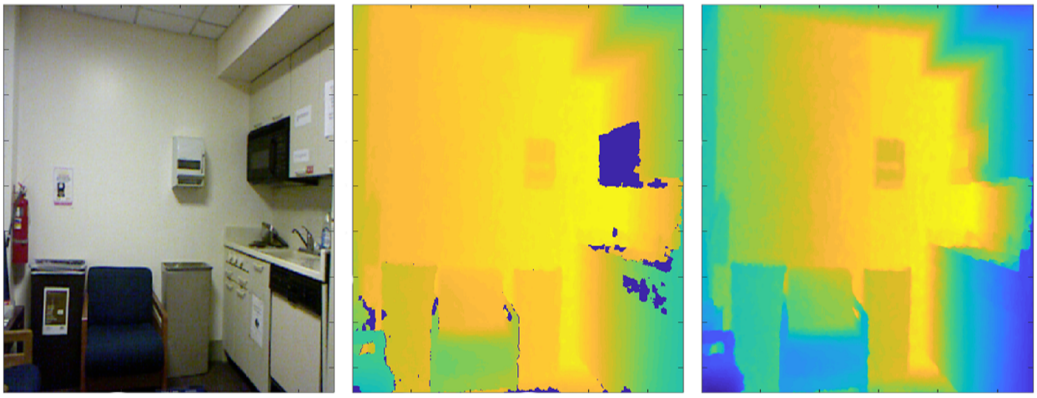
Σχήμα 3.1: (α) RGB εικόνα. (β) Παρουσία σχετικής μετατόπισης RGB εικόνας και καταγραφής Βάθους. (γ) Συγχρονισμός RGB εικόνας και βάθους με σωστή προβολή του βάθους πάνω στην εικόνα και εξάλειψη της σχετικής τους μετατόπισης.

- **Έλλειψη τιμών στα περιθώρια της εικόνας Βάθους:** Όπως φαίνεται στο Σχήμα 3.2, η καταγραφή του Βάθους περιβάλλεται από ένα περίγραμμα στα περιθώρια της, το οποίο δεν περιέχει χρήσιμη πληροφορία για τις τιμές Βάθους. Χρησιμοποιώντας την αντίστοιχη συνάρτηση του πακέτου εργαλειών υπολογίζεται μία «μάσκα», η οποία είναι πρακτικά μία εικόνα με δυαδικές τιμές, η οποία προσδιορίζει τα όρια στα οποία θα πρέπει να περιοριστεί η εικόνα και το αντίστοιχο βάθος (τα σημεία αυτά έχουν τιμή 1 στην «μάσκα»), αφού έχει γίνει η σωστή προβολή του βάθους πάνω στην εικόνα, ώστε να διατηρηθούν μόνο τα σημεία της εικόνας που περιέχουν χρήσιμη πληροφορία. Έτσι μετά την προβολή και στην σωστή περικοπή της εικόνας καταλήγουμε στο αποτέλεσμα του Σχήματος 3.1(γ).

- **Απουσία τιμών Βάθους:** Τέλος, για να ολοκληρωθεί η απαραίτητη προεπεξεργασία των δεδομένων και να είναι έτοιμα προς εκπαίδευση, απαιτείται η συμπλήρωση των ορισμένων τιμών που λείπουν από τις καταγραφές του αισθητήρα βάθους. Αυτό μπορεί να συμβεί λόγω διάφορων παραγόντων, μερικοί από τους οποίους είναι η παρουσία σκιάσεων στο χώρο, καθώς και η πολύ μεγάλη ή πολύ μικρή αντανάκλαση του φωτός από ορισμένα αντικείμενα της εικόνας (π.χ. καθρέφτες, τζάμια κ.ά.). Για την συμπλήρωση αυτών των τιμών, χρησιμοποιήσαμε τη μέθοδο της χρωματοποίησης που προτάθηκε στο [38] και περιέχεται ελαφρώς τροποποιημένη στο επίσημο πακέτο εργαλείων επεξεργασίας του συνόλου δεδομένων. Το αποτέλεσμα αυτής της μεθόδου φαίνεται στο Σχήμα 3.3, όπου βλέπουμε πως οι τιμές που απουσιάζουν αρχικά από το καταγεγραμμένο βάθος (β), έχουν συμπληρωθεί με αρκετή συνέπεια (γ) ώστε να προσεγγίζουν την πραγματικότητα.



Σχήμα 3.2: Περίγραμμα καταγραφής Βάθους με απουσία τιμών.



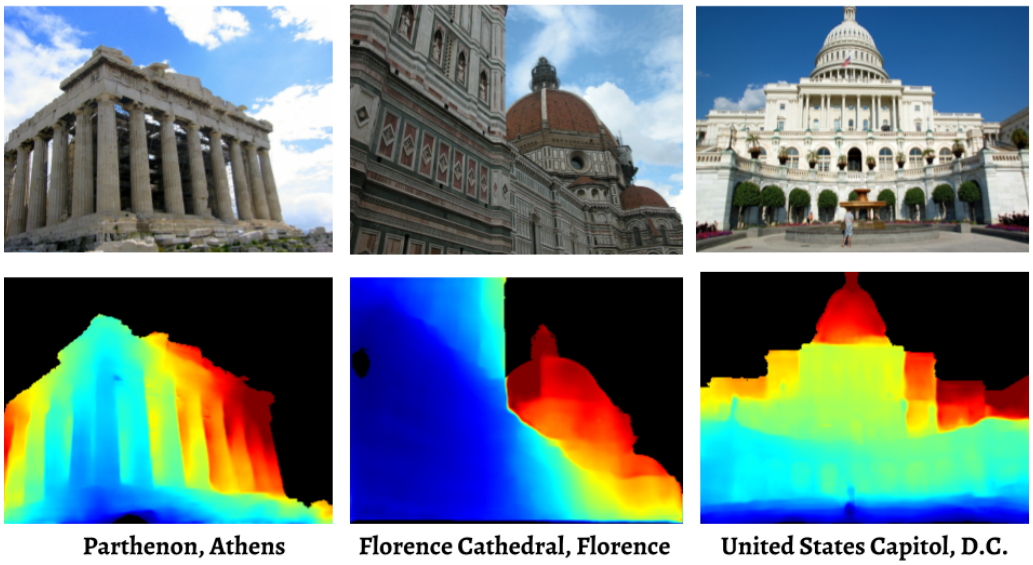
Σχήμα 3.3: (α) RGB εικόνα. (β) Χάρτης βάθους προβεβλημένος πάνω στην εικόνα με απουσιάζουσες τιμές (σκούρες μπλε περιοχές). (γ) Τελικός χάρτης βάθους με συμπληρωμένες τις απουσιάζουσες τιμές με χρήση της μεθόδου χρωματοποίησης [38].

MegaDepth Dataset

Η δεύτερη μέθοδος [39] που εξετάστηκε για την εκτίμηση του βάθους από δισδιάστατα δεδομένα και την χρησιμότητα αυτής της πληροφορίας στην μοντελοποίηση του προβλήματος της εμφάνειας, δημιούργησε ένα νέο σύνολο δεδομένων με αντιστοιχίες δισδιάστατων εκόνων και βάθους προκειμένου να εκπαιδεύσει το προτεινόμενο μοντέλο.

Το σύνολο αυτό των δεδομένων αποτελείται από εικόνες κατεβασμένες από το

Flickr, στις οποίες φαίνοντα καθαρά φωτογραφισμένα αξιοθέατα/ορόσημα διάφορων πόλεων. Για την ανακατασκευή κάθε εικόνας χρησιμοποιούνται state-of-the-art Structure-from-Motion (SfM) [63] και Multi-View Stereo (MVS) [64] μέθοδοι, από όπου προκύπτουν τόσο ένα SfM μοντέλο όσο και μία πυκνή αναπαράσταση του αντίστοιχου βάθους για κάθε ανακατασκευασμένη εικόνα. Παρόλαυτα η κάθε αναπαράσταση που προκύπτει περιέχει αρκετό θόρυβο και ορισμένες απουσιάζουσες ή ακραίες τιμές (outliers), τα οποία εμποδίζουν την σωστή εκπαίδευση των μοντέλων πάνω στο σύνολο αυτό. Για το λόγο αυτό γίνεται μία σειρά από βήματα προεπεξεργασίας από τους δημιουργούς του συνόλου πριν χρησιμοποιηθεί στην εκπαίδευση, όπως αναφέρονται λεπτομερώς στην αντίστοιχη δημοσίευση [39] και το τελικό αποτέλεσμα φαίνεται στο Σχήμα 3.4. Το σύνολο δεδομένων είναι δημόσια διαθέσιμο στον σύνδεσμο: `https://www.cs.cornell.edu/projects/megadepth/`.



Σχήμα 3.4: Δείγματα από το MegaDepth σύνολο δεδομένων.

ReDWeb, WSVD, DIML, 3D Movies Datasets

Η τρίτη μέθοδος εκτίμησης βάθους, αναφερόμενη ως MiDaS [56], που εξετάστηκε και αξιολογήθηκε ως προς την προσφορά της στην τρισδιάστατη μοντελοποίηση της οπτικής προσοχής χρησιμοποιεί ένα συνδυασμό από πέντε διαφορετικά σύνολα δεδομένων, προκειμένου να εμπλουτίσει τα δεδομένα εκπαίδευσης του μοντέλου και να προκύψει ένα μοντέλο με υψηλή ικανότητα γενίκευσης. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι το MegaDepth [39], όπως παρουσιάστηκε στην προηγούμενη ενότητα, καθώς και τα ReDWeb [82], WSVD [75], DIML [32] και 3D Movies [56] σύνολα δεδομένων.

Το σύνολο 3D Movies είναι ένα νέο σύνολο δεδομένων που δημιουργήθηκε από τους συγγραφείς του [56], προκειμένου να συμπληρώσει και να εμπλουτίσει το περιεχόμενο των υπόλοιπων τεσσάρων συνόλων δεδομένων που χρησιμοποιούνται για την εκπαίδευση του προτεινόμενου μοντέλου. Αποτελείται από υψηλής ποιότητας ακολουθίες εικόνων από βίντεο, τα οποία απεικονίζουν πολλές διαφορετικές σκηνές, που ποικίλουν από ανθρωποκεντρικές μέχρι σκηνές τοπίων και σε αντίθεση με το MegaDepth [39] σύνολο, δεν απαιτήθηκε κάποια επιπλέον προεπεξεργασία των δεδομένων. Ορισμένα δείγματα του συνόλου απεικονίζονται στο Σχήμα 3.5, όπου παρατηρείται και η ποικιλία της πληροφορίας που μπορεί να προσφέρει το σύνολο. Η μέθοδος αυτή εξαιτίας τόσο της αρχιτεκτονικής της αλλά περισσότερο εξαιτίας του πλούσιου περιεχομένου των δεδομένων εκπαίδευσης αποδείχθηκε με βάση τα πειράματά μας, η μέθοδος με τη μεγαλύτερη αποτελεσματικότητα στην εκτίμηση του βάθους απο δισδιάστατες εικόνες αλλά και στην εκτίμηση της εμφάνειας σε τρισδιάστατα βίντεο.

Σχήμα 3.5: Δείγματα από το 3D Movies σύνολο δεδομένων.

Δεδομένα Επαλήθευσης

Στις τρεις μεθόδους εκτίμησης βάθους από δισδιάστατες εικόνες που χρησιμοποιήθηκαν, εκτός απο τα αντίστοιχα σύνολα επαλήθευσης των συνόλων δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση, έγινε χρήση και κάποιων εντελώς διαφορετικών συνόλων, ώστε να εξεταστεί η αποτελεσματικότητα των μοντέλων σε νέα σύνολα δεδομένων και η δυνατότητά τους να γενικεύσουν ικανοποιητικά σε δεδομένα που τους ήταν τελείως άγνωστα κατά την εκπαίδευση. Μερικά από τα σύνολα αυτά δεδομένων, είναι τα σύνολα KITTI [45], DIW [8], Make3D [62], ETH3D [65], Sintel [2], TUM-RGBD [71], B3DO [27], SUN-RGBD [69], τα οποία φαίνονται και στο Σχήμα 3.6. Τα σύνολα αυτά καλύπτουν ένα ευρύ φάσμα διαφορετικού οπτικού περιεχομένου από τοπία και εξωτερικούς χώρους μέχρι ανθρωποκεντρικές σκηνές και εσωτερικούς χώρους.

3.2 Δεδομένα Παρακολούθησης Ματιού (Eye Tracking)

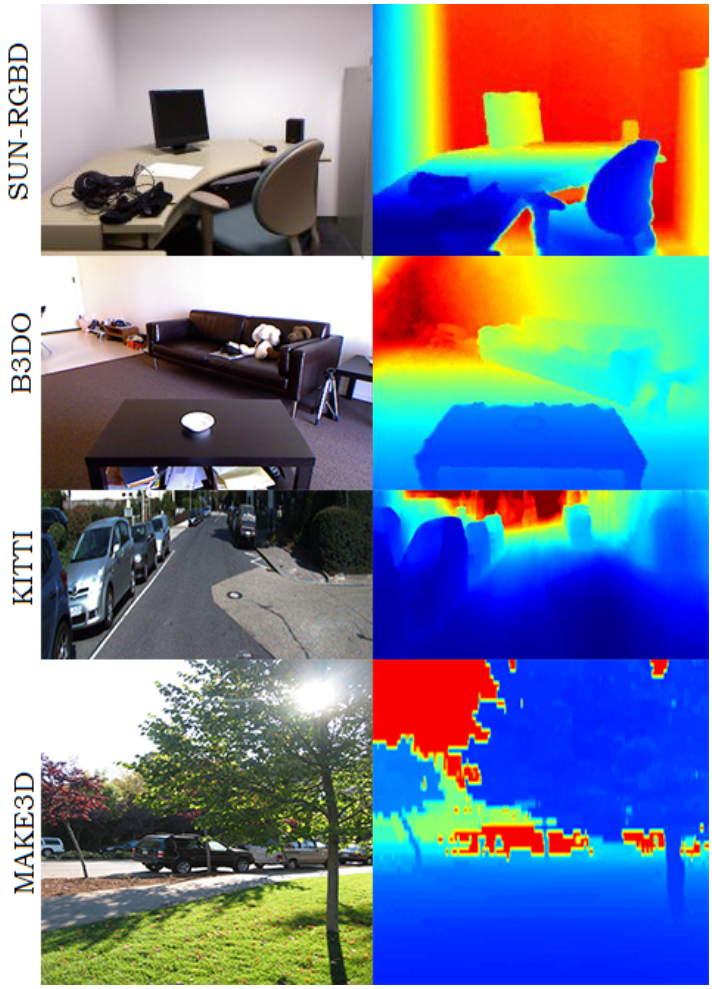
Στην ενότητα αυτή, παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση του δικτύου και την επίλυση του προβλήματος της μοντελοποίησης της ανθρώπινης οπτικής προσοχής. Για το σκοπό αυτό χρησιμοποιήθηκαν διάφορα σύνολα δεδομένων στα οποία περιέχονται ακολουθίες εικόνων από χρονικά σύντομα βίντεο, όπου απεικονίζεται πληθώρα διαφορετικών οπτικών σκηνών. Στις ακολουθίες αυτές, περιέχεται για κάθε καρέ και για κάθε εθελοντή που παρακολούθησε το βίντεο, η αντίστοιχη επισημείωση σχετικά με την τοποθεσία εστίασης της κόρης του ματιού (eye-tracking annotation) του ανθρώπου πάνω στην εικόνα την στιγμή που παρακολούθησε το βίντεο. Τυχαία δείγματα από όλα τα σύνολα δεδομένων απεικονίζονται στο Σχήμα 3.7 όπου αναδεικνύεται και η ποικιλομορφία των δεδομένων.

DIEM

Το DIEM [48] σύνολο δεδομένων αποτελείται από 84 βίντεο παντός περιεχομένου, τα οποία προέρχονται από δημόσια προσβάσιμες πηγές, όπως διαφημίσεις, ντοκιμαντέρ, τρέιλερ παιχνιδιών και ταινιών, βίντεο κλιπ, δελτία ειδήσεων καθώς και time-lapse υλικό. Η επισημείωση των δεδομένων έγινε συλλέγοντας δεδομένα και καταγράφοντας την κόρη του ματιού από 42 εθελοντές μέσω ενός Eyelink eye-tracker, με τυχαία σειρά αναπαραγωγής των βίντεο

AVAD

Το AVAD [47] σύνολο δεδομένων περιέχει 45 σύντομα κλιπ από βίντεο, διάρκειας περίπου 5-10 δευτερολέπτων με διάφορες οπτικοακουστικές σκηνές (π.χ. χορού, παιξίματος κάποιου μουσικού οργάνου, κελαιδήματος πτηνών κ.ά. Στην επισημείωση των δεδομένων συμμετείχαν 16 εθελοντές.



Σχήμα 3.6: Δείγματα από τα Δεδομένα Αξιολόγησης των τριών μεθόδων εξαγωγής του βάθους.

SumMe

Το SumMe [20, 73] σύνολο δεδομένων περιέχει 25 μη δομημένα βίντεο, δηλαδή περισσότερο ερασιτεχνικά βίντεο φτιαγμένα από απλούς χρήστες, τα οποία έχουν συλλεχθεί από διάφορες δημόσια διαθέσιμες πηγές. Τα δεδομένα σχετικά με την παρακολούθηση της κόρης του ματιού έχουν καταγραφεί [73] από 10 θεατές.

ETMD

Η ETMD [33, 73] βάση δεδομένων αποτελείται από 12 βίντεο από 6 διαφορετικές ταινίες του Χόλιγουντ. Τα eye-tracking δεδομένα συλλέχθηκαν [73] από συνολικά 10 εθελοντές μέσω ενός Eyelink eye-tracker. Τα αποσπάσματα αυτά των ταινιών

περιέχουν υψηλού επιπέδου πληροφορία με περίπλοκο σημασιολογικό περιεχόμενο.

Coutrot1 + Coutrot2

Οι βάσεις δεδομένων Coutrot είναι χωρισμένες σε Coutrot1 και Coutrot2: Η Coutrot1 βάση περιέχει 60 κλιπ από βίντεο με δυναμικές σκηνές χωρισμένες σε 4 διαφορετικές οπτικές κατηγορίες: ένα κινούμενο αντικείμενο, περισσότερα κινούμενα αντικείμενα, τοπία και πρόσωπα. Τα δεδομένα παρακολούθησης ματιού έχουν συλλεχθεί από 72 εθελοντές. Η Coutrot2 βάση αποτελείται από 15 σύντομα βίντεο πολύ περιορισμένου και στοχευμένου περιεχομένου, καθώς απεικονίζονται μόνο σκηνές με 4 ανθρώπους σε ένα δωμάτιο που διεξάγουν κάποια συζήτηση. Τα αντίστοιχα δεδομένα παρακολούθησης του ματιού καταγράφηκαν από 40 θεατές.

Σχήμα 3.7: Δείγματα από τα Δεδομένα Παρακολούθησης του Ματιού μαζί με την αντίστοιχη επισημείωση, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση του μοντέλου.

Hollywood-2

Το Hollywood-2 [43] περιέχει μία συλλογή από σύντομα κλιπ με σκηνές που διαδραματίστηκαν σε ταινίες του Χόλιγουντ. Το σύνολο αυτό δεδομένων δημιουργήθηκε αρχικά για το πρόβλημα της ανίχνευσης ανθρώπινων δράσεων, αλλά αργότερα αναπτύχθηκε και για την μοντελοποίηση του προβλήματος της Εμφάνειας, εφόσον συλλέχθηκαν τα δεδομένα παρακολούθησης του ματιού [44]. Τα δεδομένα εκπαίδευσης και επαλήθευσης αποτελούνται από 3100 και 3559 διαφορετικά, μη επικαλυπτόμενα κλιπ αντίστοιχα, κάθε ένα από τα οποία έχει παρακολουθηθεί από 12 εθελοντές.

UCF-Sports

Το UCF-Sports [58, 70] όμοια με το Hollywood-2 επίσης περιέχει σύντομα κλιπ τα οποία αρχικά συλλέχθηκαν για την επίλυση του προβλήματος της αναγνώρισης της ανθρώπινης δραστηριότητας. Αργότερα, συλλέχθηκαν τα δεδομένα παρακολούθησης ματιού από 19 θεατές. Το σύνολο δεδομένων έχει χωριστεί σε 104 και 48 μη επικαλυπτόμενα βίντεο εκπαίδευσης και επαλήθευσης αντίστοιχα, τα οποία απεικονίζουν σκηνές από διάφορες αθλητικές δραστηριότητες. Τα βίντεο αυτά επίσης ποικίλουν αρκετά στην χρονική τους διάρκεια, καθώς μπορεί κάποιο βίντεο να αποτελείται από 40 μέχρι 1 μόλις καρέ.

DHF1K

Το DHF1K [79] είναι ένα αρκετά μεγάλο σύνολο δεδομένων, το οποίο περιέχει βίντεο διαφορετικού μεγέθους (από 400 έως 1200 εικόνες ανά βίντεο) και πλούσιου περιεχομένου με διαφορετικές σκηνές. Αποτελείται από συνολικά 1000 βίντεο, 700 από τα οποία είναι δημόσια διαθέσιμα μαζί με τις αντίστοιχες επισημειώσεις, ενώ τα υπόλοιπα 300 έχουν παραμείνει μη προσβάσιμα και διατηρούνται από τους δημιουργούς του συνόλου για σκοπούς επαλήθευσης.

Κεφάλαιο 4

Εκπαίδευση Μοντέλων και Πειραματικά Αποτελέσματα Εξαγωγής Βάθους από Δισδιάστατες Εικόνες

Όπως έχει αναφερθεί, στόχος της διπλωματικής εργασίας είναι να ερευνηθεί η συνεισφορά του βάθους στο γενικότερο πρόβλημα της πρόβλεψης της εστίασης της ανθρώπινης προσοχής όταν αυτή διεγείρεται από οπτικά ερεθίσματα. Για το σκοπό αυτό, απαιτούνται δεδομένα, τα οποία περιέχουν τόσο την πληροφορία της εστίασης της προσοχής αλλά και την πληροφορία της απόστασης των απεικονιζόμενων αντικειμένων σε σχέση με την κάμερα. Από όσο γνωρίζουμε, εκτός από ένα σχετικά μικρό σύνολο δεδομένων αποτελούμενο από 54 βίντεο που αναπτύχθηκε από τους [37] και το οποίο όπως αναφέρεται περιέχει σκηνές στις οποίες το βάθος είναι βοηθητικό στην εκτίμηση της εμφάνειας, δεν υπάρχει κάποιο μεγαλύτερο σύνολο δεδομένων με πλούσιο και ποικίλο περιεχόμενο που να περιέχει τόσο τα δεδομένα εστίασης όσο και την πληροφορία του βάθους. Εφαρμόστηκαν λοιπόν κάποιες μέθοδοι εκτίμησης του βάθους από δισδιάστατες εικόνες, σε υπάρχοντα, ευρέως χρησιμοποιούμενα σύνολα δεδομένων παρακολούθησης ματιού αποτελούμενα από πολλές διαφορετικές σκηνές, έτσι ώστε αυτά να εμπλουτιστούν με αυτήν την πληροφορία με την μεγαλύτερη δυνατή ακρίβεια και να εξεταστεί η συνεισφορά του βάθους στη φύση.

Στο κεφάλαιο αυτό λοιπόν, θα παρουσιαστούν οι μέθοδοι που χρησιμοποιήθηκαν προκειμένου να πραγματοποιηθεί η εκτίμηση του βάθους από τις δισδιάστατες ακολουθίες εικόνων που απαρτίζουν τα βίντεο των συνόλων εκπαίδευσης που χρησιμοποιήθηκαν για την μοντελοποίηση του προβλήματος της Εμφάνειας. Θα γίνει αναφορά τόσο στις διαφορετικές αρχιτεκτονικές των δικτύων όσο και στα αποτελέσματα που επιτεύχθηκαν από κάθε μέθοδο και τα οποία οδήγησαν τελικά στην εφαρμογή της βέλτιστης μεθόδου, της μεθόδου MiDaS [56], για το πρόβλημα της εμφάνειας.

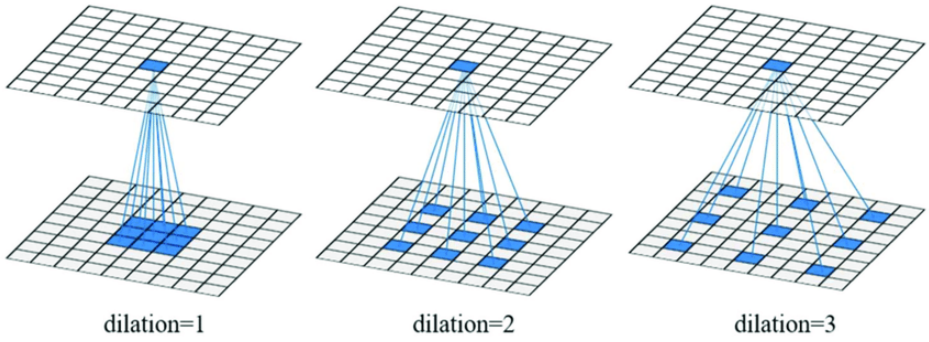
4.1 Αρχιτεκτονική Μοντέλων και Εκπαίδευση

Detail Preserving Depth Estimation from a Single Image Using Attention Guided Networks (Dilated Model) [22]

Σε αυτό το σημείο θα παρουσιαστεί η πρώτη μέθοδος που υλοποιήθηκε και η αρχιτεκτονική του μοντέλου που εκπαιδεύτηκε. Όπως φαίνεται και στην αντίστοιχη δημοσίευση [22] πρόκειται για ένα Πλήρως Συνελικτικό Νευρωνικό Δίκτυο (CNN), το οποίο χρησιμοποιεί ως βασικότερη πράξη, την πράξη της διεσταλμένης συνέλιξης (Dilated/Atrous Convolution) [6], η οποία υπολογίζεται ως:

$$(\mathbf{I} *_r \mathbf{w})(p) = \sum_t \mathbf{I}(p - rt) * \mathbf{w}(t) \tag{4.1}$$

όπου r είναι ο ρυθμός διαστολής. Η πράξη αυτή απεικονίζεται και γραφικά στο Σχήμα 4.1

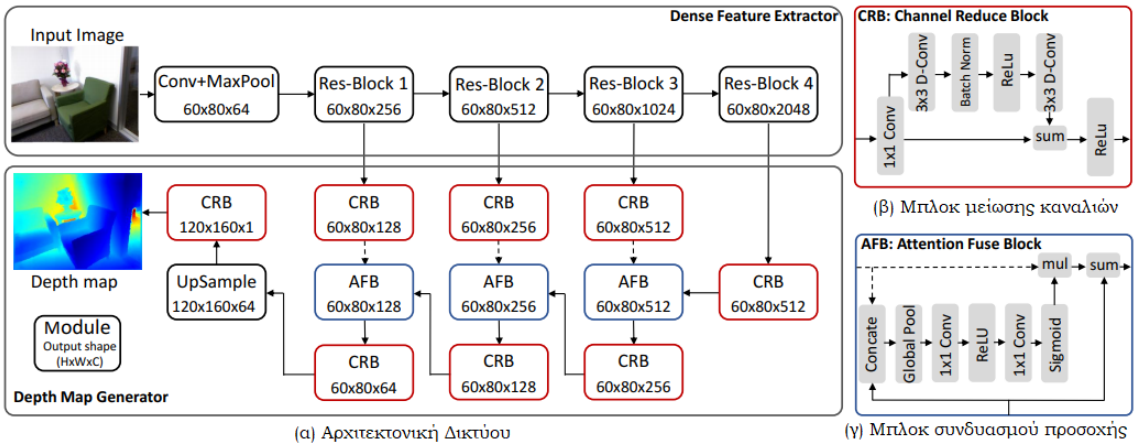


Σχήμα 4.1: Γραφική απεικόνιση της πράξης της δισδιάστατης διεσταλμένης συνέλιξης για ρυθμους διαστολής 1,2 και 3. Σχήμα από [11]

Όπως φαίνεται και στο Σχήμα 4.1, η επιπλέον παράμετρος του ρυθμού διαστολής αυτού του τύπου συνέλιξης επηρεάζει το εύρος του χωρικού πεδίου που θα συμμετέχει στον υπολογισμό της τιμής κάθε σημείου. Όσο μεγαλύτερος είναι ο ρυθμός διαστολής, τόσο μεγαλύτερο είναι και το πεδίο αυτό, με αποτέλεσμα για το εκάστοτε σημείο να λαμβάνονται υπόψιν οι τιμές ολοένα και πιο απομακρυσμένων γειτονικών σημείων. Αυτό έχει ως αποτέλεσμα να υπάρχει καλύτερη αντίληψη της συνολικής εικόνας και να συσχετίζονται αποτελεσματικά ακόμα και απομακρυσμένες χωρικά περιοχές, χωρίς να χρειάζεται να μειωθούν σημαντικά οι διαστάσεις της αρχικής εικόνας, όπως συμβαίνει στην κλασική μορφή της δισδιάστατης συνέλιξης.

Με βάση τα παραπάνω, η αρχιτεκτονική του συγκεκριμένου μοντέλου έχει το πλεονέκτημα πως μπορεί να εξάγει αποτελεσματικά χαρακτηριστικά από την εικόνα τόσο σε μικρή κλίμακα όσο και σε μεγαλύτερη κλίμακα χωρίς να χρειάζεται να μειώσει τις

διαστάσεις της αρχικής εικόνας σημαντικά. Αυτό οδηγεί στην αποτελεσματικότητα του δικτύου ως προς το πρόβλημα της πρόβλεψης του βάθους των απεικονιζόμενων αντικειμένων, αλλά παράλληλα και στην υψηλή ανάλυση και διατήρηση των λεπτομερειών της εικόνας σε σχέση με άλλες μεθόδους. Συνεπώς στις τελικές προβλέψεις μπορεί κανείς να διακρίνει ικανοποιητικά τα αρχικά αντικείμενα και τις λεπτομέρειες τους, όπως ακμές, γωνίες, περίγραμμα κ.λ.π. Η αρχιτεκτονική του δικτύου φαίνεται στο Σχήμα 4.2.



Σχήμα 4.2: Αρχιτεκτονική προτεινόμενου δικτύου στο [22]. Η αρχιτεκτονική του δικτύου βασίζεται στο ResNet-101, με ελαφρώς τροποποιημένα τα συνελικτικά μπλοκ ώστε να εφαρμόζουν διεσταλμένη συνέλιξη με διαφορετικούς ρυθμούς διαστολής. Το δίκτυο δέχεται ως είσοδο μία RGB εικόνα και παράγει μία πρόβλεψη του χάρτη βάθους της.

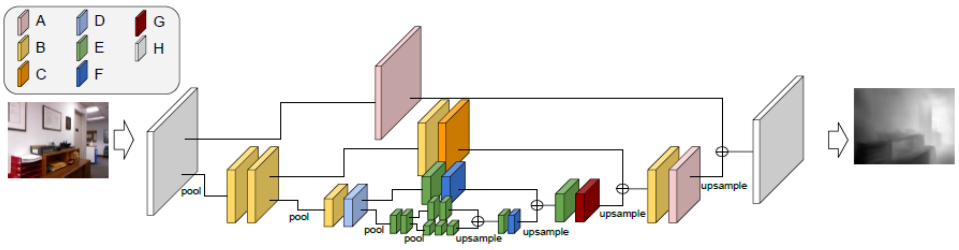
Το δίκτυο χρησιμοποιεί ως βάση το δίκτυο ResNet-101 [23], το οποίο τροποποιεί ελαφρώς ώστε σε κάθε συνελικτικό μπλοκ να χρησιμοποιεί διεσταλμένη συνέλιξη αντί για την παραδοσιακή, διαφορετικού ρυθμού διαστολής ανάλογα με το βάθος του μπλοκ στο δίκτυο. Η χρήση διαφορετικών ρυθμών διαστολής επιτρέπει την εξαγωγή χαρακτηριστικών σε πολλές διαφορετικές κλίμακες ενώ παράλληλα διατηρούνται οι διαστάσεις της εικόνας σε σημαντικό βαθμό, με την τελική πρόβλεψη να χρειάζεται μόνο μία διγραμμική παρεμβολή (Bilinear Interpolation) ώστε οι διαστάσεις της να γίνουν οι μισές της αρχικής εικόνας. Επίσης στην συνολική αρχιτεκτονική έχουν προστεθεί δύο συνελικτικά μπλοκ, το CRB και το AFB. Το CRB αναλαμβάνει την αποτελεσματική μείωση των υπολογισμένων καναλιών-χαρακτηριστικών της εικόνας εισόδου ώστε να καταλήξουμε στην τελική εικόνα εξόδου του δικτύου. Το AFB είναι επίσης ένα συνελικτικό μπλοκ, το οποίο μπορεί να ανιχνεύσει τις σημαντικότερες περιοχές των υπολογισμένων χαρακτηριστικών. Πρακτικά έχει τη δυνατότητα να μαθαίνει τη βαρύτητα των διαφορετικών καναλιών και να συγχωνεύει κανάλια διαδοχικών επιπέδων του δικτύου με παρόμοιο τρόπο όπως στο [24].

Στα πλαίσια της διπλωματικής, πραγματοποιήθηκε η εκπαίδευση του δικτύου πάνω στο σύνολο δεδομένων NYU Depth Dataset V2, αφού ολοκληρώθηκε η προεπεξεργασία των δεδομένων που παρουσιάστηκε αναλυτικά στην προηγούμενη ενότητα. Αφού έγινε η εκπαίδευση του δικτύου για περίπου 60 εποχές, εφαρμόσαμε το εκπαιδευμένο δίκτυο στα δεδομένα παρακολούθησης ματιού που παρουσιάστηκαν στην προηγούμενη ενότητα, προκειμένου να εξάγουμε από αυτά την πληροφορία του βάθους. Ενδεικτικά αποτελέσματα από αυτά τα σύνολα δεδομένων φαίνονται στα Σχήματα 4.6 και 4.7.

MegaDepth: Learning Single-View Depth Prediction from Internet Photos[39]

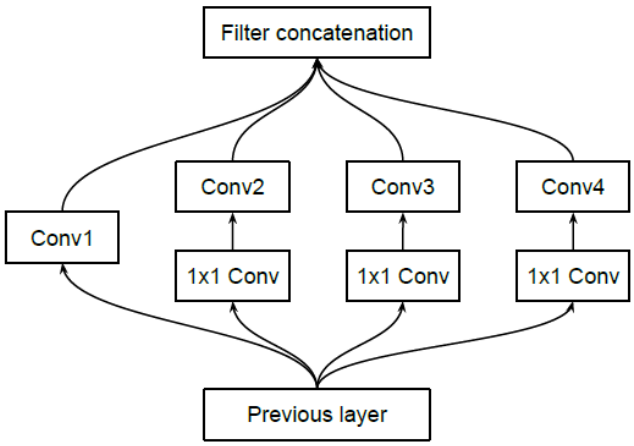
Η δεύτερη μέθοδος εξετάζει την αποτελεσματικότητα τριών διαφορετικών ήδη υπάρχοντων αρχιτεκτονικών στο πρόβλημα της εξαγωγής βάθους από δισδιάστατες εικόνες, την αρχιτεκτονική του VGG [17], του Hourglass Network [8] και του ResNet [23] καθώς αυτά εκπαιδεύονται στο προτεινόμενο σύνολο δεδομένων MegaDepth Dataset, όπως αυτό παρουσιάστηκε αναλυτικά στην προηγούμενη ενότητα.

Όπως φαίνεται στα αποτελέσματα των πειραμάτων, η εκπαίδευση αυτών των μοντέλων στο συγκεκριμένο σύνολο δεδομένων οδηγεί στην δημιουργία ακριβέστερων προβλέψεων όχι μονο στα δεδομένα επαλήθευσης αυτού του συνόλου αλλά και στα δεδομένα διαφορετικών συνόλων, σε σχέση με τις αντίστοιχες προβλέψεις των μοντέλων όταν αυτά εκπαιδεύονται σε κάποιο άλλο σύνολο. Το δίκτυο που χρησιμοποιήθηκε για την εξαγωγή των χαρτών βάθους σε όλα τα παρακάτω πειράματα είναι το Hourglass Network (MegaDepth(H)), το οποίο με βάση τους συγγραφείς έφερε τα ακριβέστερα αποτελέσματα από τις τρεις αρχιτεκτονικές που εξετάστηκαν. Η αρχιτεκτονική του δικτύου φαίνεται στο Σχήμα 4.3.



Σχήμα 4.3: Αρχιτεκτονική προτεινόμενου δικτύου στο [8]. Πρόκειται για ένα πλήρως συνελικτικό νευρωνικό δίκτυο (CNN) με πολυκλιμακωτή αρχιτεκτονική το οποίο χρησιμοποιεί ως βασική του δομική μονάδα μία ελαφρώς τροποποιημένη μορφή του συνελικτικού μπλοκ του δικτύου Inception για να παράξει τον χάρτη βάθους της εικόνας εισόδου.

Πρόκειται για ένα πλήρως συνελικτικό νευρωνικό δίκτυο (CNN), το οποίο χρησιμοποιεί ως βασική του μονάδα (μπλοκ A-G) μία ελαφρώς τροποποιημένη μορφή της μονάδας του Inception δικτύου [72], η οποία φαίνεται στο Σχήμα 4.4. Το στρώμα Η είναι απλά ένα συνελικτικό στρώμα με πυρήνα μεγέθους 3×3.



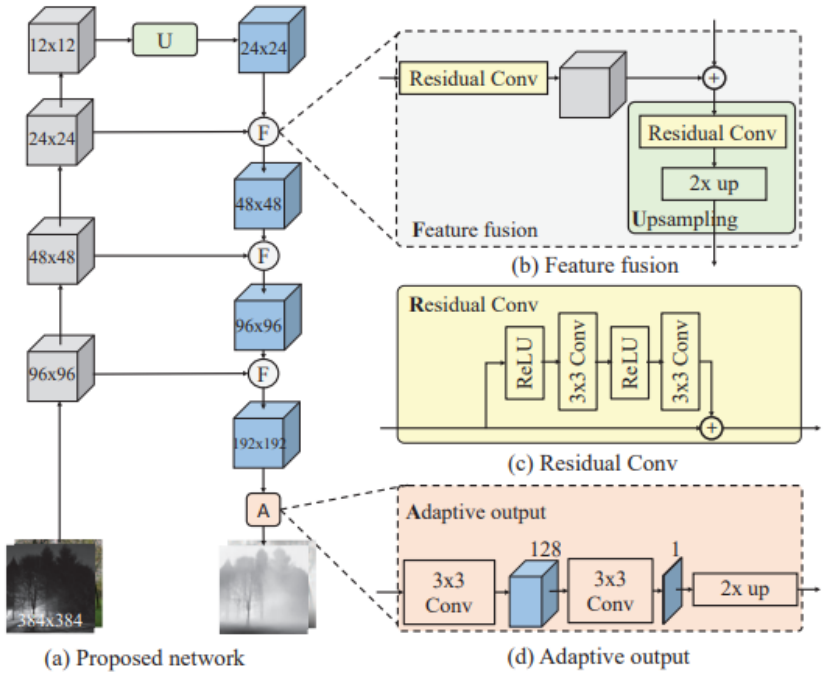
Σχήμα 4.4: Αρχιτεκτονική της βασικής μονάδας του δικτύου στο [8]

Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer (MiDaS)[56]

Στο σημείο αυτό θα παρουσιαστεί η τρίτη και τελευταία μέθοδος που εξετάστηκε ως προς την αποτελεσματικότητά της στην εκτίμηση του βάθους από τα δισδιάστατα δεδομένα. Η μέθοδος αυτή όπως και η προηγούμενη, δεν προτείνει κάποια νέα αρχιτεκτονική δικτύου, αλλά χρησιμοποιεί και εξετάζει ως προς την απόδοσή τους ήδη υπάρχοντα δίκτυα. Η βασική αρχιτεκτονική των δικτύων φαίνεται στο Σχήμα 4.5, όπου απεικονίζεται το δίκτυο το οποίο προτάθηκε από τους Xian *et al.* στο [82]. Το δίκτυο αυτό βασίζεται στην δομή του ResNet-50 [23] και υλοποιεί μία πολυκλιμακωτή (Multi-scale) αρχιτεκτονική, προκειμένου να εξάγει και να συνδυάζει χαρακτηριστικά σε διάφορες χωρικές κλίμακες.

Όπως βλέπουμε και στο Σχήμα 4.5, ο συνδυασμός αυτός υλοποιείται από κάποια μπλοκ συγχώνευσης χαρακτηριστικών (Feature Fusion). Τα μπλοκ αυτά είναι πλήρως συνελικτικά και χρησιμοποιούνται προκειμένου να συνδυάσουν αποτελεσματικά τα χαρακτηριστικά που έχουν υπολογιστεί στα διάφορα επίπεδα του δικτύου.

Η βασική συνεισφορά λοιπόν του [56] όσον αφορά το πρόβλημα της εκτίμησης του βάθους έγκειται κυρίως στον τρόπο εκπαίδευσης του μοντέλου, αλλά και στην αποτελεσματικότητα των διαφορετικών αρχιτεκτονικών κωδικοποιητή. Η γενικότερη



Σχήμα 4.5: Αρχιτεκτονική προτεινόμενου δικτύου στο [82]. Πρόκειται για ένα πλήρως συνελικτικό νευρωνικό δίκτυο (CNN) με πολυκλιμακωτή αρχιτεκτονική. Στα τελικά πειράματα χρησιμοποιήθηκε ως βάση η δομή του ResNeXt-101 [83].

αρχιτεκτονική του δικτύου διατηρείται στη μορφή που φαίνεται στο Σχήμα 4.5 για τους διαφορετικούς αποκωδικοποιητές, όμως διαφέρει το βάθος των δικτύων καθώς και η δομή των διάφορων συνελικτικών μπλοκ τα οποία υπολογίζουν τα χρήσιμα χαρακτηριστικά της εικόνας εισόδου. Συγκεκριμένα εξετάστηκαν ως αποκωδικοποιητές οι αρχιτεκτονικές των: ResNet-50, ResNet-101 [23], ResNeXt-101 [83] και DenseNet-161 [25].

Όσον αφορά την εκπαίδευση του μοντέλου, στόχος είναι να γίνει με τέτοιο τρόπο, ώστε κατά τη διάρκειά της το μοντέλο να «βλέπει» πολλά δεδομένα από διαφορετικά σύνολα και συνεπώς να προκύψει ένα δίκτυο το οποίο γενικεύει πολύ καλά και μπορεί να παράξει πολύ ακριβείς εκτιμήσεις ακόμα και σε δεδομένα που δεν έχει ξανασυναντήσει, αποφεύγοντας το πρόβλημα της Υπερ-προσαρμογής (Overfitting).

Στην προκειμένη μέθοδο συνεπώς εξετάζονται οι διάφοροι τρόποι εκπαίδευσης των διαφορετικών δικτύων ως προς τη χρήση και τον συνδυασμό των διαφορετικών συνόλων δεδομένων. Όπως αναφέρθηκε και στην προηγούμενη ενότητα, τα σύνολα εκπαίδευσης που χρησιμοποιήθηκαν στην συγκεκριμένη μέθοδο είναι τα: MegaDepth [39], ReDWeb [82], WSVD [75], DIML [32] και 3D Movies [56]. Κατά τα πειράματα εξετάζονται τόσο οι διαφορετικοί συνδυασμοί συνόλων δεδομένων καθώς και ο τρόπος συνδυασμού αυτών των δεδομένων, ξεκινώντας από έναν πολύ απλό συνδυασμό στον

οποίο κάθε πακέτο δεδομένων περιέχει την ίδια αναλογία δεδομένων από τα διαφορετικά σύνολα και καταλήγοντας σε έναν πιο δομημένο τρόπο ανάμιξης των δεδομένων, κατά τον οποίο η εκπαίδευση σε κάθε σύνολο δεδομένων θεωρείται ως μία ξεχωριστή εργασία. Στην δεύτερη μέθοδο τα δεδομένα συνδυάζονται έτσι ώστε τελικά το δίκτυο να φτάσει σε μία κατάσταση, κατά την οποία δεν είναι δυνατό το σφάλμα να μειωθεί περαιτέρω για κάποιο από τα σύνολα δεδομένων, χωρίς να αυξηθεί για τουλάχιστον ένα από τα υπόλοιπα σύνολα. Σε αυτή την κατάσταση γνωρίζουμε πως το μοντέλο έχει την καλύτερη δυνατή απόδοση για όλα τα σύνολα δεδομένων και συνεπώς έχει γενικεύσει καλά, παρόλο που δεν έχει την βέλτιστη δυνατή απόδοση σε κάθε σύνολο δεδομένων μεμονωμένα.

4.2 Πειραματικά Αποτελέσματα

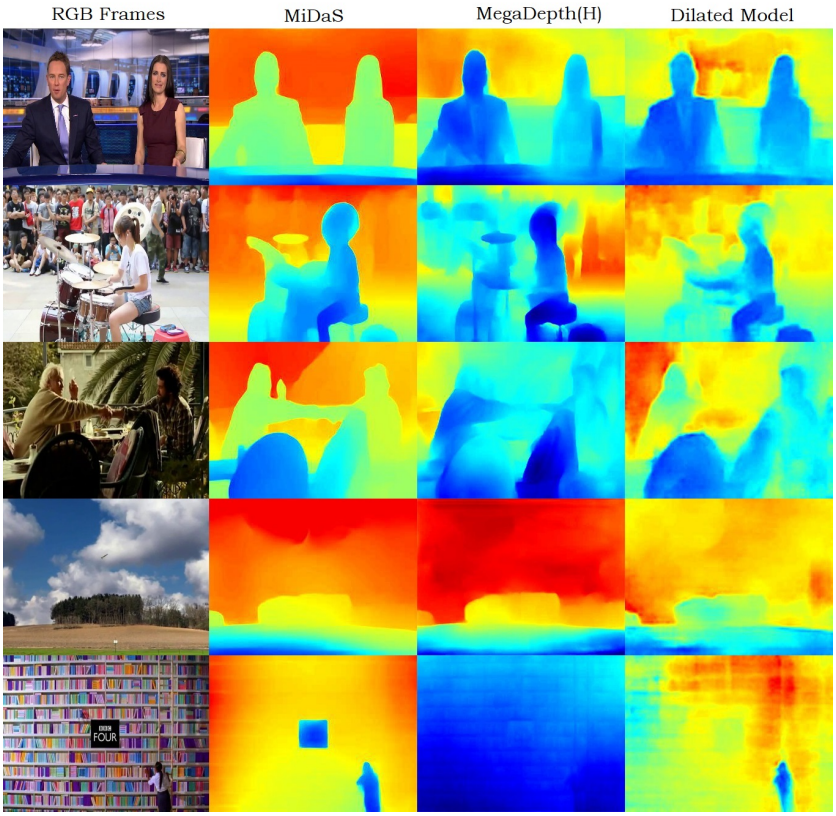
Στην ενότητα αυτή θα παρουσιαστούν και θα συγκριθούν οι προβλέψεις που προέκυψαν από την εφαρμογή των τριών παραπάνω μεθόδων σε κάποια ενδεικτικά δείγματα από τα δεδομένα παρακολούθησης ματιού. Όπως αναλύθηκε προηγουμένως, καμία από τις τρεις μεθόδους δεν έχει συναντήσει δείγματα από αυτά τα σύνολα δεδομένων κατά την εκπαίδευση της και συνεπώς εξετάζεται και η δυνατότητα των μοντέλων ως προς την γενίκευση.

Όπως βλέπουμε στα Σχήματα 4.6, 4.7, και τα τρία μοντέλα έχουν αρκετά καλή απόδοση στις περισσότερες περιπτώσεις, με το μοντέλο MiDaS [56] να υπερτερεί σημαντικά σε σύγκριση με τα άλλα δύο. Η καλύτερη απόδοση αυτού του μοντέλου οφείλεται αφενός στην αρχιτεκτονική του αλλά ακόμα σημαντικότερο ρόλο έπαιξε η εκπαίδευσή του, η οποία όπως αναφέρθηκε, έγινε συνδυάζοντας πολλά δεδομένα από διαφορετικά σύνολα. Αυτό έχει ως αποτέλεσμα την καλύτερη γενίκευση του μοντέλου και την δημιουργία ακριβέστερων εκτίμησεων βάθους ανεξαρτήτως σκηνικού. Επίσης παρατηρούμε πως σε αυτή τη μέθοδο επιτυγχάνεται καλύτερη σημασιολογική κατάτμηση της εικόνας (Semantic Segmentation), κάτι το οποίο επιτρέπει τον διαχωρισμό και την αναγνώριση των αντικειμένων στην εκτιμώμενη εικόνα βάθους διατηρώντας έτσι την σημασιολογική πληροφορία της αρχικής RGB εικόνας.

Αντίθετα η μέθοδος του Dilated Model [22], η οποία έχει εκπαιδευτεί αποκλειστικά σε δεδομένα με σκηνές εσωτερικού χώρου, παρατηρούμε πως έχει ιδιαίτερα κακή απόδοση όταν καλείται να εκτιμήσει το βάθος των απεικονιζόμενων αντικειμένων σε σκηνές εξωτερικού χώρου, χάνοντας και αρκετή από την σημασιολογική πληροφορία της εικόνας.

Στο δίκτυο του MegaDepth [39] βλέπουμε πως υπάρχει ένα μειονέκτημα συγκτικά με την δυνατότητα του μοντέλου να υπολογίσει το εύρος του βάθους των εικόνων, όπως για παράδειγμα στην πρώτη εικόνα του Σχήματος 4.7, όπου ο τοίχος στο παρασκήνιο έχει εκτιμηθεί σηματικά πιο κοντά από ότι είναι στην πραγματικότητα.

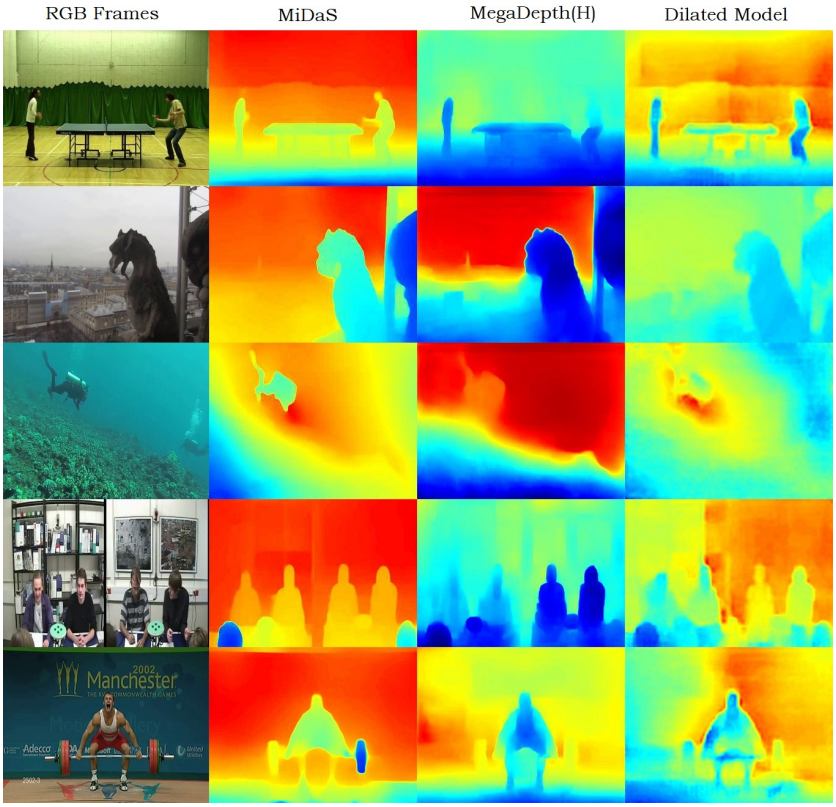
Έχοντας εξετάσει και τις τρεις μεθόδους, βλέπουμε πως σε όλες υπάρχουν αστοχίες. Για παράδειγμα βλέπουμε πως η τελευταία εικόνα του Σχήματος 4.6 δεν έχει εκτιμηθεί σωστά από καμία μέθοδο. Ακόμα και το δίκτυο MiDaS φαίνεται πως έχει εκτιμήσει λανθασμένα το μαύρο πλαίσιο που υπάρχει στο κέντρο της εικόνας. Πα-



Σχήμα 4.6: Ενδεικτικά ποτελέσματα των [22, 39, 56] σε Eye Tracking δεδομένα των συνόλων DIEM, Coutrot1, DHF1K.

ρόλ'αυτά παρατηρούμε πως η μέθοδος αυτή έχει τις λιγότερες αστοχίες στο σύνολο των δειγμάτων, κάτι που την καθιστά ίσως την καταλληλότερη μέθοδο για τον δικό μας σκοπό, ο οποίος συμπεριλαμβάνει την εκτίμηση του βάθους από πληθώρα διαφορετικών σκηνικών και συνεπώς απαιτεί ένα δίκτυο με καλή δυνατότητα γενίκευσης.

Η αξιολόγηση της απόδοσης και των αποτελεσμάτων των μοντέλων έγινε διαισθητικά, με βάση την δική μας αντίληψη και παρατηρώντας τις προβλέψεις των μοντέλων για διάφορα δείγματα των Eye Tracking δεδομένων. Η αξιολόγηση στην προκειμένη περίπτωση δεν θα μπορούσε να γίνει με κάποιο πιο δομημένο τρόπο, όπως για παράδειγμα με την σύγκριση συγκεκριμένων μετρικών, καθώς για κανένα σύνολο από τα Eye Tracking δεδομένα δεν έχουμε τους πραγματικούς χάρτες βάθους.



Σχήμα 4.7: Ενδεικτικά ποτελέσματα των [22, 39, 56] σε Eye Tracking δεδομένα των συνόλων DIEM, Coutrot1, Coutrot2, UCF Sports.

Κεφάλαιο 5

Εκπαίδευση Μοντέλων και Πειραματικά Αποτελέσματα Μοντελοποίησης Οπτικής Προσοχής

Στο κεφάλαιο αυτό θα παρουσιαστεί αναλυτικά το κυρίως μέρος και η συνεισφορά της διπλωματικής εργασίας. Αρχικά θα αναφερθούμε στη δημιουργία ενός ανταγωνιστικού μοντέλου για την εκτίμηση της εμφάνειας σε βίντεο με χρήση αποκλειστικά RGB δεδομένων. Η RGB πληροφορία είναι απαραίτητη στο πρόβλημα της Εμφάνειας, όπως και στα περισσότερα προβλήματα όρασης υπολογιστών. Αυτό επαληθεύεται και από τα πειράματα που παρουσιάζονται αργότερα στο κεφάλαιο, όπου φαίνεται πως το προτεινόμενο μοντέλο έχει σημαντικά μειωμένη απόδοση όταν καλείται να εκτιμήσει τους χάρτες εμφάνειας αποκλειστικά από τους χάρτες βάθους των δεδομένων εισόδου. Αυτή η συμπεριφορά είναι αναμενόμενη, καθώς τα RGB δεδομένα περιέχουν το μεγαλύτερο μέρος της οπτικής πληροφορίας όπως το σημασιολογικό περιεχόμενο, ακμές, γωνίες και άλλα χαρακτηριστικά τα οποία παίζουν σημαντικό ρόλο και επηρεάζουν σε μεγάλο βαθμό την ανθρώπινη αντίληψη και προσοχή.

Με βάση τα παραπάνω, αναγνωρίζουμε πως η RGB πληροφορία είναι αναντικατάστατη και κρίνεται απαραίτητη για την μοντελοποίηση του συγκεκριμένου προβλήματος. Παράλληλα όμως δημιουργείται το ερώτημα σχετικά με το αν και το ποια επιπλέον πληροφορία, αν συνδυαστεί με τα αρχικά RGB δεδομένα και ενσωματωθεί στην διαδικασία της εκπαίδευσης, μπορεί να αποδειχθεί ωφέλιμη για την εκτίμηση της εμφάνειας. Τέτοια πληροφορία μπορεί να είναι μία σχετικά απλούστερη πρότερη γνώση, όπως για παράδειγμα χάρτες που δηλώνουν την απεικόνιση προσώπων (face masks) σε μία περιοχή της εικόνας, όπως και κάποια πιο σύνθετη πληροφορία όπως η παρουσία ήχου στα βίντεο, η οπτική ροή ή η πληροφορία του βάθους. Παρόλο που τόσο ο ήχος αλλα και ακόμα περισσότερο η οπτική ροή έχουν εξεταστεί στο παρελθόν ως προς την συνεισφορά τους στο γενικότερο πρόβλημα της εμφάνειας σε βίντεο, το βάθος δεν

έχει προσεγγίσει το ίδιο ενδιαφέρον. Για το λόγο αυτό, στην παρούσα διπλωματική εργασία θα εξεταστεί η συνεισφορά του βάθους ως μία επιπλέον πληροφορία για την αποτελεσματική εκτίμηση της εμφάνειας. Στην δεύτερη λοιπόν ενότητα του κεφαλαίου παρουσιάζονται τα διάφορα πειράματα που πραγματοποιήθηκαν προκειμένου να εξεταστεί το αν και το πώς θα μπορούσε η επιπλέον αυτή πληροφορία να συνδυαστεί με τα αρχικά δεδομένα εισόδου ώστε να επιτευχθεί ακριβέστερη πρόβλεψη της εμφάνειας. Με βάση τα συμπεράσματα που προκύπτουν από τις πειραματικές διαδικασίες πάνω σε 9 διαφορετικά σύνολα δεδομένων παρακολούθησης ματιού, βλέπουμε πως στις περισσότερες περιπτώσεις το προτεινόμενο RGBD μοντέλο πετυχαίνει καλύτερα αποτελέσματα τόσο συγκριτικά με το RGB-μόνο μοντέλο, όσο και από άλλες state-of-the-art μεθόδους, υποδεικνύοντας πως το βάθος μπορεί πράγματι να συνεισφέρει στην εκτίμηση της εμφάνειας.

5.1 Δισδιάστατο Οπτικό Μοντέλο (RGB)

Στην ενότητα αυτή θα παρουσιαστεί η δισδιάστατη προσέγγιση του προβλήματος, με την έννοια ότι σαν σήματα εισόδου χρησιμοποιούνται μόνο οι RGB ακολουθίες εικόνων από τα βίντεο. Όπως φαίνεται στο Σχήμα 5.1, πρόκεται για ένα Τρισδιάστατο Πλήρως Συνελικτικό Νευρωνικό Δίκτυο (3D Fully-Convolutional Neural Network), το οποίο αποτελείται από έναν Κωδικοποιητή (Encoder), από κάποιες Βαθιά Επιβλεπόμενες Μονάδες Προσοχής (Deeply Supervised Attention Modules (DSAM)) καθώς και από έναν Αποκωδικοποιητή (Decoder). Η αρχιτεκτονική του μοντέλου και συγκεκριμένα ο κωδικοποιητής βασίστηκε τόσο στο SUSiNet [34] όσο και στο STAViS [74]. Το SUSiNet αποτελεί ένα τρισδιάστατο πλήρως συνελικτικό δίκτυο που προσπαθεί να μοντελοποιήσει συγχρόνως τρία προβλήματα της όρασης υπολογιστών: Εκτίμηση Εμφάνειας, Αναγνώριση Δράσεων και Περίληψη Βίντεο. Το STAViS [74], το οποίο επίσης βασίστηκε στο SUSiNet, καλείται να αντιμετωπίσει το πρόβλημα της Εμφάνειας πάνω σε οπτικοακουστικά δεδομένα, συμπεριλαμβάνοντας και την πληροφορία του ήχου στην διαδικασία της μοντελοποίησης.

Κωδικοποιητής και DSAM

Ο Κωδικοποιητής είναι υπεύθυνος για την εξαγωγή των χρήσιμων χωροχρονικών χαρακτηριστικών από τις ακολουθίες εικόνων σε διάφορες χωροχρονικές κλίμακες. Η αρχιτεκτονική του Κωδικοποιητή επεκτείνει την τρισδιάστατη εκδοχή του ResNet-50, το οποίο είχε αρχικά προταθεί για την επίλυση του προβλήματος της αναγνώρισης και κατηγοριοποίησης δράσεων. Το δίκτυο αποτελείται από 4 τρισδιάστατα πλήρως συνελικτικά μπλοκ, τα οποία εξάγουν χωροχρονικά χαρακτηριστικά σε διαφορετικές κλίμακες των εικόνων εισόδου, τα οποία αντιπροσωπεύονται από τα X^1, X^2, X^3, X^4. Κάθε έξοδος X^m των συνελικτικών μπλοκ βελτιστοποιείται αρχικά χρησιμοποιώντας έναν μηχανισμό προσοχής και στη συνέχεια προωθείται στο επόμενο συνελικτικό μπλοκ. Για το σκοπό αυτό χρησιμοποιούμε τις Βαθιά Επιβλεπόμενες Μονάδες προσοχής, των οποίων η δομή απεικονίζεται στο Σχήμα 5.2.

Η Βαθιά Επίβλεψη έχει προηγουμένως χρησιμοποιηθεί στην Ανίχνευση Ακμών [84], Κατάτμηση Αντικειμένων [4] και στο πρόβλημα της στατικής Εμφάνειας [76]. Ο ρόλος των μονάδων αυτών είναι τριπλός: Χρησιμοποιούνται για την βελτίωση των αναπαραστάσεων των υπολογισμένων χωρικών χαρακτηριστικών, για να παρέχουν τους πολυ-επίπεδους χάρτες ενεργοποίησης A^m, οι οποίοι θα χρησιμοποιηθούν στη συνάρτηση σφάλματος προκειμένου να υπολογιστεί η απόκλιση των προβλέψεων και των πραγματικών τιμών και τέλος για να υπολογίσουν τους χάρτες εμφάνειας S^m, οι οποίοι θα χρησιμοποιηθούν ως είσοδοι στη μονάδα του Αποκωδικοποιητή προκειμένου να αποκτήσουμε έναν τελικό χάρτη εμφάνειας. Συνεπώς οι παράμετροι W_{am}^m των Μονάδων Βαθιάς Επιβλεπόμενης Προσοχής εκπαιδεύονται κανονικά μαζί με όλες τις υπόλοιπες παραμέτρους του δικτύου.

Το Σχήμα 5.2 απεικονίζει την αρχιτεκτονική της Μονάδας DSAM στο επίπεδο m. Συμπεριλαμβάνει ένα στρώμα Average Pooling στην διάσταση του χρόνου, προκειμένου από το σύνολο των χαρακτηριστικών για όλες τις εικόνες της ακολουθίας να προκύψει μία δισδιάστατη αναπαράσταση. Η έξοδος αυτού του στρώματος κατευθύνεται στη συνέχεια σε δύο διαφορετικά μονοπάτια μέσα στη μονάδα. Το ένα μονοπάτι αποτελείται από ένα στρώμα δισδιάστατης (χωρικής) συνέλιξης το οποίο παρέχει τους 64-ων χαρακτηριστικών χάρτες εμφάνειας S^m. Το άλλο μονοπάτι αποτελείται από δύο συνελικτικά στρώματα τα οποία εν τέλει υπολογίζουν έναν χάρτη ενεργοποίησης A^m. Ένας χωρικός Softmax μετασχηματισμός, ο οποίος εφαρμόζεται στον χάρτη ενεργοποίησης A^m παράγει τον χάρτη προσοχής M^m:

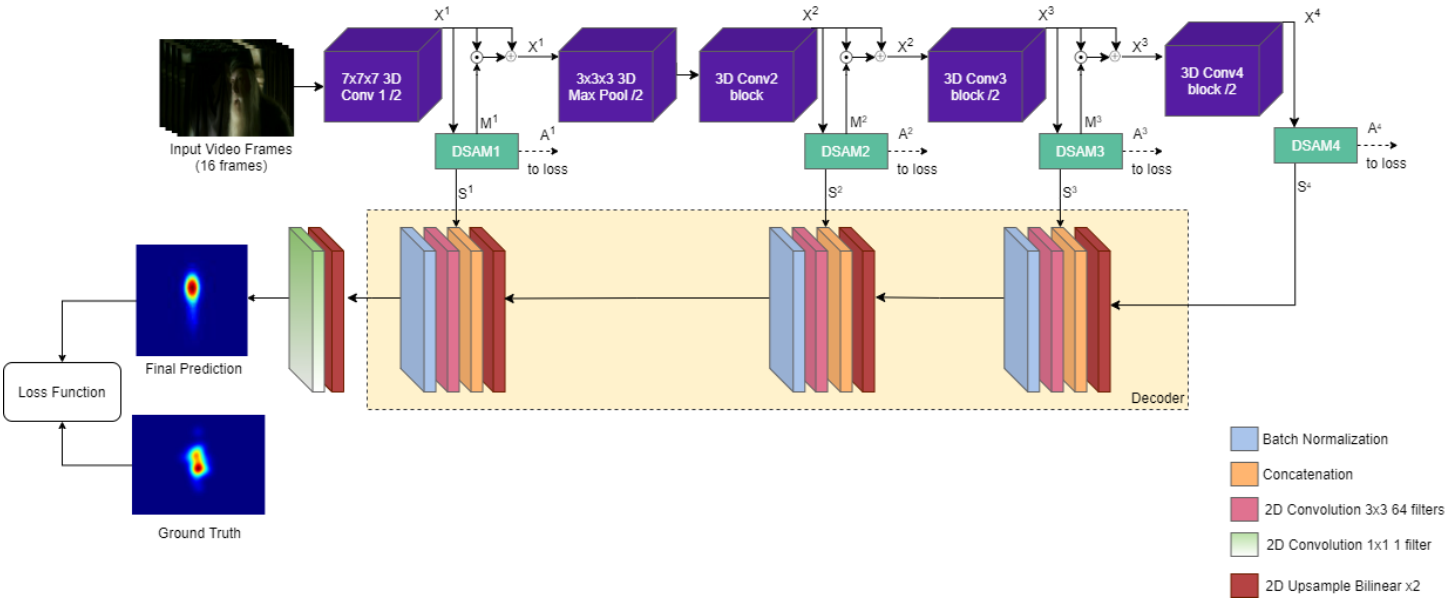
$$M^m(x,y) = \frac{exp(A^m(x,y))}{\sum_x \sum_y exp(A^m(x,y))} \tag{5.1}$$

Τέλος, οι διαστάσεις του χάρτη ενεργοποίησης A^m διπλασιάζονται προκειμένου να φτάσουν το μέγεθος των αρχικών εικόνων εισόδου χρησιμοποιώντας την πράξη της Ανάστροφης Δισδιάστατης Συνέλιξης (Transposed Convolution). Αυτός ο τύπος συνέλιξης πραγματοποιεί ουσιαστικά την αντίστροφη διαδικασία από την κλασική συνέλιξη, προσπαθεί δηλαδή να εκτιμήσει την αρχική εικόνα από την οποία έχει προκύψει κατόπιν εφαρμογής κλασικής συνέλιξης η εικόνα εισόδου. Έτσι σε αντίθεση με τον παραδοσιακό τρόπο αναβάθμισης των διαστάσεων μίας εικόνας, π.χ. με διγραμμική παρεμβολή (Bilinear Interpolation), η Ανάστροφη Συνέλιξη είναι ένας εναλλακτικός, παραμετροποιημένος και συνεπώς εκπαιδεύσιμος τρόπος αύξησης των διαστάσεων μίας εικόνας, ο οποίος εξαιτίας της παραμετροποίησής του μπορεί να επιφέρει καλύτερα αποτελέσματα.

Η έξοδος X^m κάθε συνελικτικού μπλοκ του δικτύου στο επίπεδο m πολλαπλασιάζεται ανά στοιχείο με τον χάρτη προσοχής M^m και προστίθεται στην αρχική της τιμή προκειμένου να βελτιωθούν οι πιο σημαντικές περιοχές της και να προκύψει η είσοδος του επόμενου συνελικτικού μπλοκ $\hat{X}^m$:

$$\tilde{X}^m = (1 + M^m) \odot X^m, m = 1, ..., 4 \tag{5.2}$$

όπου με $\odot$ συμβολίζεται ο ανά στοιχείο πολλαπλασιασμός των πινάκων/χαρτών.



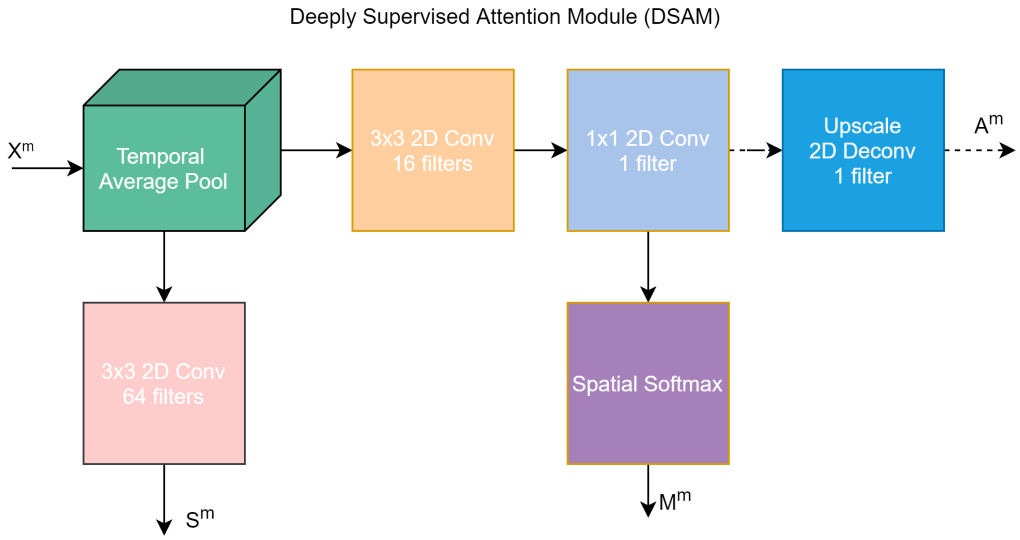
Σχήμα 5.1: ViDaS-RGB [14]: Αρχιτεκτονική του RGB προτεινόμενου μοντέλου. Το δίκτυο δέχεται ως είσοδο μία ακολυθία από 16 RGB εικόνες, υπολογίζει τις χωρο-χρονικές αναπαραστάσεις της μέσα από τα τρισδιάστατα συνελικτικά μπλοκ καθώς και μέσα από τον αποκωδικοποιητή (κίτρινο πλαίσιο) και τελικά παράγει την πρόβλεψη για τον χάρτη εμφάνειας της μεσαίας εικόνας (9ης) της ακολουθίας.

Αποκωδικοποιητής

Η μονάδα Αποκωδικοποιητή του μοντέλου φαίνεται στο Σχήμα 5.1 (κίτρινο πλαίσιο). Πρόκειται ξανά για μία πλήρως συνελικτική μονάδα επεξεργασίας των δισδιάστατων εικόνων που προκύπτουν από τα DSAM. Ο Αποκωδικοποιητής ακολουθεί μία αρχιτεκτονική παρόμοια με αυτή του U-Net [59], υπό την έννοια ότι σταδιακά συγχωνεύει χαρακτηριστικά μικρότερης κλίμακας με χαρακτηριστικά μεγαλύτερης κλίμακας, τα οποία υπολογίζονται βαθύτερα στο δίκτυο. Αποτελείται από τρία δισδιάτατα πλήρως συνελικτικά μπλοκ, τα οποία χρησιμοποιούνται ώστε να συγχωνεύουν αποτελεσματικά τα διαφορετικά χαρακτηριστικά που εξήγαγε ο Κωδικοποιητής του δικτύου στις διάφορες κλίμακες.

Το πρώτο συνελικτικό μπλοκ του Αποκωδικοποιητή παίρνει σαν είσοδο τις εξόδους S^3, S^4 των δύο τελευταίων DSAM του Κωδικοποιητή. Κατόπιν, τα άλλα δύο συνελικτικά μπλοκ του Αποκωδικοποιητή παίρνουν ως είσοδο τους χάρτες εμφάνειας S^m του αντίστοιχου επιπέδου καθώς και την έξοδο του προηγούμενου μπλοκ. Σε κάθε μπλοκ, οι δύο είσοδοι αφού έρθουν στις ίδιες διαστάσεις εφαρμόζοντας διγραμμική παρεμβολή, συνδέονται σειριακά και περνάνε μαζί μέσα από ένα στρώμα δισδιάστατης

συνέλιξης ακολουθούμενης από ένα στρώμα Κανονικοποίησης Παρτίδας (Batch Normalization), προκειμένου να επιταχυνθεί η διαδικασία της εκπαίδευσης του δικτύου και να αποφευχθεί το πρόβλημα των παραγώγων που εκρήγνυνται.

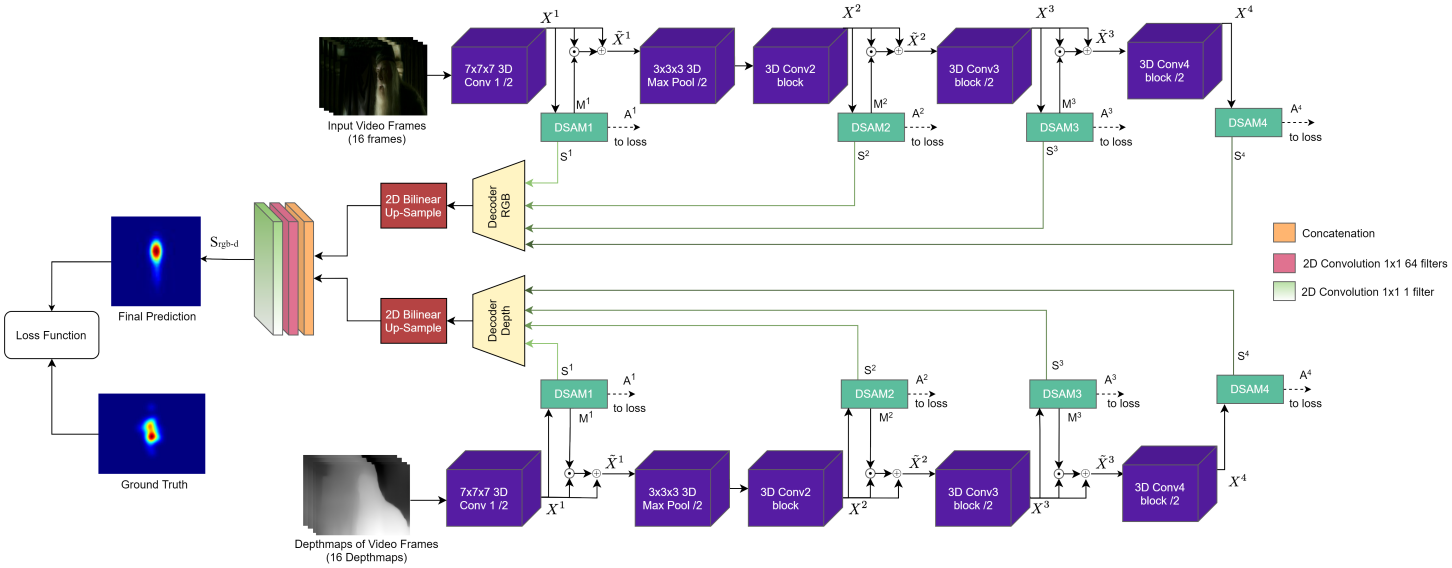


Σχήμα 5.2: Αρχιτεκτονική των Βαθιά Επιβλεπόμενων Μονάδων Προσοχής (DSAM).

5.2 Τρισδιάστατο Οπτικό Μοντέλο (RGB-D)

Στην ενότητα αυτή θα παρουσιαστεί η τεσσάρων διαστάσεων (τρισδιάστατα δεδομένα και χρονική διάσταση) εκδοχή του μοντέλου που αναλύθηκε πορηγουμένως, η οποία πρακτικά επεκτείνει την αρχιτεκτονική του προκειμένου να ληφθεί υπ'όψιν και η πληροφορία του βάθους για τις ακολουθίες εικόνων των διάφορων βίντεο. Προκειμένου να γίνει η επεξεργασία του βάθους, το αρχικό δίκτυο επεκτείνεται, προσθέτοντας μία ακόμα, πανομοιότυπη οπτική ροή, η οποία δέχεται σαν είσοδο τους χάρτες βάθους για τις αρχικές εικόνες εισόδου.

Όπως φαίνεται και στο Σχήμα 5.3, το συνολικό δίκτυο αποτελείται από δύο παρόμοιας αρχιτεκτονικής οπτικές ροές, κάθε μία από τις οποίες εξάγει και αποκωδικοποιεί την αντίστοιχη είσοδό της ανεξάρτητα από την άλλη. Όλα τα μπλοκ και οι μονάδες επεξεργασίας, όπως τα DSAM και ο Αποκωδικοποιητής έχουν διπλασιαστεί. Βλέπουμε πως οι δύο ροές του δικτύου δεν αλληλεπιδρούν μεταξύ τους παρά μόνο στο τελικό επίπεδο, όπου κάθε ροή έχει παράξει τους δικούς της χάρτες Εμφάνειας S_{rgb} και S_d, μεγέθους 64-ων χαρακτηριστικών ο κάθε ένας, οι οποίοι τελικά συνδέονται σειριακά και περνάνε από δυο στρώματα δισδιάστατης συνέλιξης προκειμένου να



Σχήμα 5.3: ViDaS-RGBD [14]: Αρχιτεκτονική του RGBD προτεινόμενου μοντέλου. Το δίκτυο δέχεται ως είσοδο μία ακολυθία από 16 RGB εικόνες και τους αντίστοιχους χάρτες βάθους τους, υπολογίζει τις αναπαραστάσεις τους μέσα από κάθε ροή και τελικά παράγει την ενιαία πρόβλεψη για τον χάρτη εμφάνειας της μεσαίας εικόνας (9ης) της ακολουθίας.

προκύψει ένας ενιαίος τελικός χάρτης Εμφάνειας S_{rgb-d}. Με τον τρόπο αυτό, διατηρώντας τις δύο ροές του δικτύου ανεξάρτητες μέχρι ένα επίπεδο, επιτρέπουμε σε κάθε ροή να μάθει ξεχωριστά τις κατάλληλες αναπαραστάσεις χαρακτηριστικών από κάθε είσοδο και το συνολικό δίκτυο μπορεί να εκπαιδευτεί καλύτερα ώστε να μαθαίνει πότε η πληροφορία του βάθους μπορεί να είναι χρήσιμη για κάθε διαφορετική είσοδο.

Συνάρτηση Σφάλματος

Για την εκπαίδευση του μοντέλου, υλοποιήθηκε μία προσαρμοσμένη στο πρόβλημα συνάρτηση σφάλματος $\mathcal{L}$, η οποία υπολογίζεται συνδυάζοντας τρία διαφορετικά σφάλματα. Για τον υπολογισμό αυτών των σφαλμάτων και εν τέλει το συνολικό σφάλμα, χρησιμοποιούμε τον πραγματικό χάρτη εμφάνειας Y, τόσο στη διακριτή, όσο και στην συνεχή του μορφή, οι οποίες συμβολίζονται με Y_b και Y_c αντίστοιχα. Ο χάρτης αυτός συγκρίνεται όχι μόνο με τον εκτιμώμενο χάρτη εμφάνειας S_{rgb-d} του δικτύου αλλά και με κάθε έναν από τους τέσσερις χάρτες ενεργοποίησης $A^m, m = 1, 2, 3, 4$ τόσο της RGB όσο και της ροής Βάθους του δικτύου, οι οποίοι συμβολίζονται με A^m_{rgb} και A^m_d αντίστοιχα. Τα σφάλματα που προκύπτουν από την σύγκριση του Y με τα S_{rgb-d}, A^m_{rgb} και A^m_d συμβολίζονται με $\mathcal{L}_{sal}$, $\mathcal{L}_{rgb}$ και $\mathcal{L}_d$ αντίστοιχα:

$$\mathcal{L} = \mathcal{L}_{sal}(S_{rgb-d}, Y) + (1-\epsilon)\sum_{m=1}^{4}\mathcal{L}_{rgb}(A^m_{rgb}, Y) + (1-\epsilon)\sum_{m=1}^{4}\mathcal{L}_{d}(A^m_{d}, Y) \quad (5.3)$$

όπου ϵ είναι μία παράμετρος μείωσης, η οποία σε κάθε εποχή εκπαίδευσης ορίζεται ως $\epsilon = \frac{currentepoch}{totalnumberofepochs}$. Για τα σφάλματα $\mathcal{L}_{sal}, \mathcal{L}_{rgb}$ και $\mathcal{L}_d$, υπολογίζουμε τρεις διαφορετικές μετρικές. Πρώτα υπολογίζουμε το σφάλμα διασταυρωμένης εντροπίας (Cross-entropy Loss (CE)) μεταξύ των υπολογισμένων χαρτών M, όπου $M = S_{rgb-d}, A^m_{rgb}, A^m_d, m = 1, ..., 4$ και του συνεχούς πραγματικού χάρτη εμφάνειας Y_c ο οποίος αποκτάται μετά από την συνέλιξη του δυαδικού χάρτη εμφάνειας Y_b των δεδομένων παρακολούθησης ματιού με έναν Γκαουσιανό πυρήνα:

$$\mathcal{L}_{CE}(M, Y_c) = -\sum_{x,y} Y_c(x,y)\cdot\log M(x,y) + (1-Y_c(x,y))\cdot(1-\log M(x,y)) \quad (5.4)$$

Η δεύτερη μετρική που υπολογίζεται είναι ο γραμμικός συντελεστής συσχέτισης (Linear Correlation Coefficient (CC)) μεταξύ των χαρτών M και του συνεχούς χάρτη εμφάνειας Y_c. Η συγκεκριμένη μετρική αντιμετωπίζει τον πραγματικό και τον εκτιμώμενο χάρτη ως τυχαίες μεταβλητές και χρησιμοποιεί στην συνδιακύμανσή τους *cov* και την τυπική απόκλιση ρ προκειμένου να υπολογίζει την συσχέτισή τους:

$$\mathcal{L}_{CC}(M, Y_c) = -\frac{cov(M, Y_c)}{\rho(M)\cdot\rho(Y_c)} \quad (5.5)$$

Η τελευταία μετρική που υπολογίζεται για το συνολικό σφάλμα εμφάνειας του μοντέλου είναι η μετρική της Εμφάνειας του Κανονικοποιημένου Οπτικού Μονοπατιού (Normalized Scanpath Saliency (NSS)) μεταξύ των χαρτών M και του δυαδικού πραγματικού χάρτη εμφάνειας Y_b:

$$\mathcal{L}_{NSS}(M, Y_b) = -\frac{1}{N_b}\sum_{x,y}\tilde{M}(x,y)\cdot Y_b(x,y) \quad (5.6)$$

όπου $\tilde{M}(x,y) = \frac{M(x,y)-\bar{M}(x,y)}{\rho(M(x,y))}$, ο κανονικοποιημένος σε μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση χάρτης M και $N_b = \sum_{x,y} Y_b(x,y)$, ο συνολικός αριθμός διακριτών σημείων εστίασης στον δυαδικό χάρτη Y_b. Τα τρία αρχικά σφάλματα μπορούν να γραφτούν ως εξής:

$$\mathcal{L}_{sal}(S_{rgb-d}, Y) = w_1\mathcal{L}_{CE}(S_{rgb-d}, Y_c) + w_2\mathcal{L}_{CC}(S_{rgb-d}, Y_c) + w_3\mathcal{L}_{NSS}(S_{rgb-d}, Y_b) \quad (5.7)$$

$$\mathcal{L}_{rgb}(A^m_{rgb}, Y) = w_1\mathcal{L}_{CE}(A^m_{rgb}, Y_c) + w_2\mathcal{L}_{CC}(A^m_{rgb}, Y_c) + w_3\mathcal{L}_{NSS}(A^m_{rgb}, Y_b) \quad (5.8)$$

$$\mathcal{L}_{d}(A^m_{d}, Y) = w_1\mathcal{L}_{CE}(A^m_{d}, Y_c) + w_2\mathcal{L}_{CC}(A^m_{d}, Y_c) + w_3\mathcal{L}_{NSS}(A^m_{d}, Y_b) \quad (5.9)$$

όπου w_1, w_2, w_3 είναι τα βάρη που χρησιμοποιούνται για την απόκτηση του σταθμισμένου αθροίσματος των τριών μετρικών και τον υπολογισμό του συνολικού σφάλματος.

5.3 Εκπαίδευση και Πειράματα

Εκπαίδευση

Όπως αναφέρθηκε η αρχιτεκτονική και των δύο ροών του δικτύου βασίζεται στην τρισδιάστατη εκδοχή του ResNet-50, η οποία έχει δείξει ανταγωνιστική απόδοση σε σύγκριση με άλλες αρχιτεκτονικές βαθιών νευρωνικών δικτύων σε προβλήματα αναγνώρισης δράσεων, όσον αφορά όχι μόνο την ακρίβεια αλλά και το υπολογιστικό κόστος. Ως αρχικό σημείο, για τις εκπαιδεύσιμες παραμέτρους $\mathbf{W}_{rgb}, \mathbf{W}_d$ των δύο ροών αντίστοιχα, χρησιμοποιήσαμε τα βάρη του προεκπαιδευμένου μοντέλου στην βάση δεδομένων Kinetics 400 [31]. Για την εκπαίδευση του δικτύου εφαρμόζουμε τον αλγόριθμο Στοχαστικής Καθόδου Κλίσης (Stochastic Gradient Descent) με παράγοντα Momentum ίσο με 0.9 και έναν όρο κανονικοποίησης των βαρών ίσο με 10^{-5}. Τα δεδομένα εισάγονται στο δίκτυο σε παρτίδες (Batches) των 128 δειγμάτων, ενώ για τον ρυθμό εκπαίδευσης εφαρμόζουμε την μέθοδο του Πολυ-βηματικού ρυθμού (Multi-step Learning Rate).

Τα στρώματα των Μονάδων Βαθιά Επιβλεπόμενης Προσοχής και οι μονάδες των Αποκωδικοποιητών εκπαιδεύονται χρησιμοποιώντας έναν αρχικό ρυθμό εκπαίδευσης 10^{-4} ενώ οι δύο ροές των Κωδικοποιητών εκπαιδεύονται χρησιμοποιώντας έναν αρχικό ρυθμό εκπαίδευσης 10^{-3}. Το δίκτυο εκπαιδεύεται για 60 εποχές. Οι χωρικές διαστάσεις των δεδομένων εισόδου μειώνονται σε 112x112 και εφαρμόζεται ένα κυλιόμενο παράθυρο μεγέθους 16 εικόνων σε κάθε βίντεο. Συνεπώς κάθε παρτίδα εισόδου του δικτύου αποτελείται από 128 δείγματα, το κάθε ένα από τα οποία είναι 16 διαδοχικά καρέ ενός βίντεο και η τελική πρόβλεψη του δικτύου αφορά το μεσαίο καρέ.

Για την επαύξηση των διαθέσιμων δεδομένων εισόδου (Data Augmentation) με βάση τα δείγματα στα Δεδομένα Παρακολούθησης του Ματιού, εφαρμόζουμε τυχαίο οριζόντιο γύρισμα με πιθανότητα $P = 0.5$ τόσο στην αρχική εικόνα, στην αντίστοιχη εικόνα βάθους αλλά και στον πραγματικό χάρτη εμφάνειας. Δεν εφαρμόζουμε κανέναν άλλο μετασχηματισμό. Τέλος τα βάρη w_1, w_2, w_3 της συνάρτησης σφάλματος τέθηκαν ίσα με 0.1, 2 και 1 αντίστοιχα κατόπιν πειραματισμού.

Πειράματα

Προκειμένου να καταλήξουμε στο αποδοτικότερο μοντέλο και να γίνει η βέλτιστη δυνατή μοντελοποίηση του προβλήματος εξετάστηκαν τόσο η αρχιτεκτονική του μοντέλου και οι υπερπαράμετροί του όσο και οι διαφορετικές μέθοδοι εξαγωγής βάθους από τα δεδομένα όπως αυτές παρουσιάστηκαν αναλυτικά στο προηγούμενο κεφάλαιο. Τα αποτελέσματα των διαφορετικών πειραμάτων φαίνονται στον Πίνακα 5.1. Για τις διάφορες δοκιμές που αναφέρονται στον Πίνακα 5.1 χρησιμοποιήθηκαν τα εξής 6 σύνολα δεδομένων: DIEM, AVAD, Coutrot1, Coutrot2, SumMe και ETMD και καταγράφηκαν στον Πίνακα ο μέσος όρος από το σύνολο των συνόλων δεδομένων για κάθε μετρική. Η εξαγωγή του βάθους είναι μέρος των πειραμάτων όπως αναφέρθηκε και εξετάζεται στις τρεις τελευταίες γραμμές του Πίνακα. Για όλα τα υπόλοιπα πειράματα

το βάθος είχε εξαχθεί χρησιμοποποιώντας την μέθοδο του MiDaS [56].

Όσον αφορά τις γραμμές του πίνακα, αρχικά παρατηρούμε ότι υπάρχουν τρία διαφορετικά προθέματα: RGB, Depth και RGBD, τα οποία αντιστοιχούν στην εκπαίδευση του μοντέλου μόνο σε RGB δεδομένα, μόνο σε δεδομένα Βάθους και στην εκπαίδευση του μοντέλου και στις δυο πληροφορίες αντίστοιχα. Ο αριθμός που υπάρχει αμέσως μετά από αυτό το πρόθεμα (16 ή 64) αντιστοιχεί στον αριθμό των χαρακτηριστικών στους χάρτες Εμφάνειας S^m που προκύπτουν από τα DSAM. Ο όρος SC αντιστοιχεί στον αγγλικό όρο Simple Concatenation και πρακτικά αντιστοιχεί σε μία πιο απλουστευμένη μορφή του μοντέλου που παρουσιάστηκε, όπου δεν έχει χρησιμοποιηθεί η μονάδα του αποκωδικοποιητή, αλλά οι έξοδοι S^m των DSAM εξισώνουν τις διαστάσεις τους χρησιμοποιώντας διγραμμική παρεμβολή (Bilinear Interpolation) και συνδέονται σειριακά ώστε να περάσουν μαζί από ένα στρώμα δισδιάστατης συνέλιξης και να παραχθεί η τελική πρόβλεψη του μοντέλου.

Ο όρος MF που εμφανίζεται σε όλες τις γραμμές του Πίνακα 5.1 αντιστοιχεί στον αγγλικό όρο Multi-scale Fusion, ο οποίος πρακτικά σηματοδοτεί την χρήση της μονάδας του Αποκωδικοποιητή σε κάθε οπτική ροή, ο οποίος όπως αναφέρθηκε πραγματοποιεί πολυ-κλιμακωτή συγχώνευση των χαρτών Εμφάνειας S^m. Όσον αφορά τους όρους ADD, CON αφορούν την RGBD μέθοδο και πρακτικά είναι τα πειράματα στα οποία εξετάστηκε η αλληλεπίδραση των δύο οπτικών ροών σε διάφορα επίπεδα χωροχρονικής κλίμακας. Συγκεκριμένα το ADD αναφέρεται στην ανά στοιχείο προσθήκη των εξόδων S^m των αντίστοιχων DSAM στο m-οστό επίπεδο και την επεξεργασία τους από μία ενιαία μονάδα Αποκωδικοποιητή και για τις δύο ροές. Αντίστοιχα ο όρος CON αναφέρεται στη σειριακή σύνδεση των χαρτών S^m των DSAM και ξανά την επεξεργασία τους από έναν κοινό Αποκωδικοποιητή και για τις δύο ροές.

Όσον αφορά τον συνδυασμό των δύο οπτικών ροών έχουμε επίσης τον όρο CLL, ο οποίος αναφέρεται στην μέθοδο Concatenation in Last Layer, η οποία πρακτικά αποτελεί και το προτεινόμενο μοντέλο, όπου οι δύο ροές επεξεργάζονται την κάθε είσοδο ξεχωριστά χωρίς κάποια αλληλεπίδραση στα ενδιάμεσα στάδια και οι έξοδοι των δυο ξεχωριστών Αποκωδικοποιητών κάθε ροής συγχωνεύονται και συνδυάζονται μέσω μίας δισδιάστατης συνέλιξης μόνο στο τελευταίο επίπεδο του δικτύου όπου και παράγεται η τελική πρόβλεψη.

Τέλος βλέπουμε ότι οι τρεις τελευταίες γραμμές του Πίνακα 5.1 έχουν επίσης το επίθεμα MID, MEG ή DIL, το οποίο αφορά την μέθοδο εξαγωγής βάθους που χρησιμοποιήθηκε σε κάθε περίπτωση: MID αναφέρεται στη μέθοδο του MiDaS [56], MEG αναφέρεται στη μέθοδο του MegaDepth [39] και DIL αναφέρεται στη μέθοδο του Dilated μοντέλου [22].

Από τα αποτελέσματα των πειραμάτων βλέπουμε πως οι μέθοδοι που χρησιμοποιούν αποκλειστικά την πληροφορία του βάθους έχουν σημαντικά χειρότερη απόδοση, καθώς η RGB πληροφορία είναι ιδιαίτερα σημαντική στην εξαγωγή των διάφορων χαρακτηριστικών. Επίσης βλέπουμε πως την συνολικά καλύτερη απόδοση μπορούμε να την πετύχουμε χρησιμοποιώντας και τις δύο ροές του μοντέλου εκμεταλλευόμενοι όλη την διαθέσιμη πληροφορία. Όσον αφορά την αρχιτεκτονική του δικτύου, είναι εμφανές πως η χρήση του Αποκωδικοποιητή επιφέρει σημαντική βελτίωση σε όλες τις

Method \ Dataset	Overall				
	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
RGB16SC	0.5458	2.7523	0.9025	0.6478	0.4296
Depth16SC	0.4875	2.3494	0.89015	0.6196	0.3890
RGBD16SC	0.5473	2.7041	0.9022	0.6485	0.4313
RGB16MF	0.5859	3.0548	0.9121	0.6584	0.4623
RGB64MF	0.5950	3.1665	0.9149	0.6624	0.4691
Depth64MF	0.5063	2.0174	0.8954	0.6276	0.4107
RGBD64MF_ADD	0.6002	3.2074	0.9157	0.6643	0.4689
RGBD64MF_CON	0.5956	3.1333	0.9152	0.6636	0.4674
RGBD64MF_CLL_MID	**0.6041**	**3.2253**	**0.9166**	**0.6654**	**0.4716**
RGBD64MF_CLL_MEG	0.5968	3.1776	0.9156	0.6640	0.4667
RGBD64MF_CLL_DIL	0.5944	3.1798	0.9147	0.6626	0.4672

Πίνακας 5.1: Πειραματικές δοκιμές: Εξετάζονται διαφοτετικοί τρόποι συγχώνευσης, ο αριθμός των χαρακτηριστικών στους χάρτες Εμφάνειας S^m και οι τρεις μέθοδοι εκτίμησης του βάθους.

μετρικές, αποδίδοντας αρκετά πιο ακριβείς προβλέψεις.

5.4 Αποτελέσματα και Σύγκριση

Στην ενότητα αυτή θα παρουσιαστούν αναλυτικότερα τα αποτελέσματα του προτεινόμενου μοντέλου και η σύγκριση αυτών τόσο σχετικά με την RGB και RGBD μορφή του αλλά και με άλλα state-of-the-art μοντέλα, δηλαδή μοντέλα τα οποία παρουσιάζουν ιδιαίτερα αποδοτική μοντελοποίηση του συγκεκριμένου προβλήματος. Η σύγκριση αυτή όσον αφορά τις μετρικές του προβλήματος της Εμφάνειας παρουσιάζονται στους Πίνακες 5.2, 5.3, 5.4 για όλα τα Σύνολα Δεδομένων Παρακολούθησης Ματιού. Επίσης στα Σχήματα 5.7, 5.8, 5.9 φαίνονται οι προβλέψεις των διάφορων μοντέλων σε κάποια τυχαία δείγματα των συνόλων δεδομένων και η σύγκριση αυτών. Τέλος στα Σχήματα 5.5, 5.6 και 5.4 φαίνεται αναλυτικότερα η σύγκριση των προβλέψεων του μοντέλου και της ακρίβειας αυτών με και χωρίς την χρήση της πληροφορίας του βάθους.

Όσον αφορά τους Πίνακες 5.2, 5.3, 5.4, βλέπουμε πως έγιναν εκτενείς συγκρίσεις με άλλα 11 μοντέλα πάνω σε 9 διαφορετικά σύνολα δεδομένων παρακολούθησης ματιού, για τις 5 διαφορετικές μετρικές. Η σύγκριση αυτή γίνεται με διάφορες εκδοχές του προτεινόμενου μοντέλου τόσο όσον αφορά την χρήση ή μη της πληροφορίας του βάθους όσο και τα σύνολα δεδομένων που χρησιμοποιήθηκαν κάθε φορά για την εκπαίδευσή του. Συγκεκριμένα το αναγνωριστικό [S] προκύπτει από τον όρο Spatial και δηλώνει πως τα συγκεκριμένα μοντέλα επεξεργάζονται αποκλειστικά την χωρική πληροφορία. Αντίθετα το αναγνωριστικό [ST] προκύπτει από τον όρο Spatio-Temporal και δηλώνει πως τα συγκεκριμένα μοντέλα λαμβάνουν υπόψιν τους τόσο τα χωρικά

όσο και τα χρονικά χαρακτηριστικά της εισόδου προκειμένου να παράξουν κάποια πρόβλεψη. Τέλος το αναγνωριστικό [STD] προκύπτει από τον όρο Spatio-Temporal Depth και αναφέρεται σε μοντέλα που επεξεργάζονται την χωροχρονική πληροφορία αλλά και την διαθέσιμη πληροφορία του βάθους. Επίσης για το προτεινόμενο μοντέλο (ViDaS), βλέπουμε πως υπάρχουν 5 διαφορετικά αναγνωριστικά στο τέλος της μεθόδου: tDHF1K, tHOLLY, tUCF, tUHD και tSTAViS, τα οποία αναφέρονται στα διαφορετικά σύνολα δεδομένων που χρησιμοποιήθηκαν κάθε φορά για την εκπαίδευση.

Τα tDHF1K, tHOLLY, tUCF, δηλώνουν πως το μοντέλο εκπαιδεύτηκε αποκλειστικά σε κάποιο από αυτά τα σύνολα δεδομένων (DHF1K, Hollywood 2 ή UCF Sports αντίστοιχα) . Το tUHD δηλώνει πως το μοντέλο εκπαιδεύτηκε σε έναν συνδυασμό των τριών αυτών συνόλων δεδομένων. Τέλος το tSTAViS δηλώνει πως το μοντέλο εκπαιδεύτηκε στα σύνολα δεδομένων DIEM, AVAD, Coutrot1, Coutrot2, SumMe και ETMD όπως αυτά συνδυάστηκαν και χρησιμοποιήθηκαν για την εκπαίδευση του STAViS [74]. Εξαιτίας όλων αυτών των διαφορετικών σχημάτων εκπαίδευσης έχουμε τη δυνατότητα να κάνουμε δίκαιες συγκρίσεις του προτεινόμενου μοντέλου και άλλων των άλλων δικτύων και να συγρίνουμε την απόδοσή τους τόσο σε γνωστά δεδομένα αλλά και σε τελείως καινούρια για τα μοντέλα δεδομένα, εξετάζοντας έτσι και την δυνατότητα των μοντέλων να γενικεύουν.

Όσον αφορά τον Πίνακα 5.2 βλέπουμε πως το προτεινόμενο μοντέλο πετυχαίνει ανταγωνιστική απόδοση και στα τρια σύνολα δεδομένων, ενώ στα Hollywood 2, UCF Sports πετυχαίνει την υψηλότερη απόδοση στις περισσότερες μετρικές. Βλέπουμε πως σε ορισμένες περιπτώσεις τα μοντέλα TASED και Unisal παρουσιάζουν καλύτερη απόδοση σε κάποιες μετρικές. Αυτό πιθανώς να οφείλεται στο γεγονός πως και τα δύο αυτά μοντέλα εκπαιδεύονται μεταξύ άλλων και σε σύνολα δεδομένων παρακολούθησης ματιού με στατικές εικόνες όπως το SALICON [29], κάτι το οποίο φαίνεται πως ενισχύει σε κάποιο βαθμό την απόδοσή τους.

Method \ Dataset	DHF1K					Hollywood-2					UCFsports				
	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
ViDaS [STD] tDHF1K	0.4891	2.75	0.9071	0.6924	0.3790	0.5982	2.82	0.9128	0.5396	0.4903	0.5673	2.73	0.8981	0.5972	0.4647
ViDaS [ST] tDHF1K	0.4864	2.73	0.9079	0.6913	0.3787	0.5901	2.78	0.9106	0.5381	0.4887	0.5538	2.66	0.8985	0.5963	0.4565
ViDaS [STD] tHOLLY	0.4366	2.42	0.8913	0.6728	0.3399	0.6457	2.96	0.9173	0.5409	0.5262	0.5462	2.57	0.8801	0.5959	0.4549
ViDaS [ST] tHOLLY	0.4338	2.40	0.8926	0.6702	0.3420	0.6425	2.94	0.9168	0.5402	0.5255	0.5400	2.53	0.8844	0.5915	0.4537
ViDaS [STD] tUCF	0.4074	2.27	0.8815	0.6712	0.3109	0.4747	2.06	0.8728	0.5219	0.3781	0.6360	3.26	**0.9160**	0.6477	0.5241
ViDaS [ST] tUCF	0.3928	2.20	0.8764	0.6622	0.3129	0.4471	1.92	0.8609	0.5188	0.3725	0.6317	3.23	0.9124	0.6467	0.5242
ViDaS [STD] tUHD	0.4778	2.69	0.9058	0.6876	0.3724	0.6462	3.00	0.9184	0.5427	0.5283	**0.6463**	3.26	0.9111	0.6383	**0.5331**
ViDaS [ST] tUHD	0.4798	2.70	0.9056	0.6869	0.3743	**0.6653**	2.90	0.9146	0.5297	**0.5397**	0.6272	3.12	0.9089	0.6310	0.5194
ViDaS [STD] tSTAViS	0.4736	2.64	0.9047	0.6924	0.3583	0.6202	2.77	0.9131	0.5324	0.5003	0.5818	2.78	0.9020	0.6189	0.4664
ViDaS [ST] tSTAViS	0.4693	2.61	0.9034	0.6884	0.3582	0.6141	2.75	0.9123	0.5312	0.4993	0.5762	2.72	0.8954	0.6082	0.4662
DeepNet [51] [S]	0.2969	1.58	0.8421	0.6432	0.1878	0.4163	1.89	0.8717	0.5416	0.2851	0.4121	1.89	0.8609	0.6162	0.2844
DVA [77] [S]	0.3592	2.06	0.8609	0.6572	0.2462	0.4644	2.44	0.8806	0.5529	0.3500	0.4495	2.37	0.8706	0.6207	0.3288
SAM [9] [S]	0.3684	2.12	0.8680	0.6562	0.2918	0.4798	2.61	0.8858	0.5552	0.4009	0.4941	2.75	0.8854	0.6272	0.4036
SalGAN [50] [S]	0.3533	1.95	0.8626	0.6732	0.2515	0.4534	2.19	0.8761	0.5540	0.3475	0.4388	2.10	0.8674	0.6240	0.3254
ACLNet [78, 80] [ST]	0.4167	2.30	0.8883	0.6523	0.3008	0.5954	3.06	0.9179	0.5428	0.4855	0.5070	2.54	0.8977	0.5908	0.4058
DeepVS [28] [ST]	0.3500	1.97	0.8561	0.6405	0.2622	0.4769	2.48	0.8883	0.5481	0.3857	0.4550	2.31	0.8703	0.6136	0.3682
TASED [46] [ST]	**0.5142**	**2.87**	**0.9130**	**0.7123**	0.3592	0.5622	2.77	0.9138	0.5397	0.4372	0.4943	2.31	0.8884	0.5528	0.4027
Unisal [15] [ST]	0.4778	2.75	0.8994	0.6759	0.3815	0.6158	**3.40**	**0.9217**	**0.5739**	0.4961	0.6254	**3.38**	0.9117	**0.6536**	0.5104
STRANet [36] [ST]	0.4617	2.58	0.8971	0.6727	0.3568	0.6010	3.19	0.8735	0.5520	0.4922	0.5635	2.85	0.9067	0.6135	0.4639
SALEMA [40] [ST]	0.4939	2.86	0.9064	0.6866	**0.3919**	0.5531	2.98	0.9085	0.5545	0.4622	0.5551	2.93	0.9014	0.6218	0.4614
STAViS [ST] [74]	0.4312	2.35	0.8936	0.6789	0.3139	0.5898	2.59	0.9085	0.5303	0.4684	0.5376	2.44	0.8906	0.6051	0.4239

Πίνακας 5.2: Αποτελέσματα αξιολόγησης των εκτιμήσεων της Εμφάνειας στο σετ δεδομένων επαλήθευσης της DHF1K βάσης καθώς και στις βάσεις UCF Sports και Hollywood 2. Τα αποτελέσματα της προτεινόμενης μεθόδου (ViDaS [STD]) και της RGB μόνο μεθόδου ([ST]) εμφανίζονται για διαφορετικά σχήματα εκπαίδευσης.

Η TASED, η οποία παρουσιάζει καλύτερη απόδοση από ότι το προτεινόμενο μοντέλο στο σύνολο DHF1K, εφαρμόζει κυλιόμενο παράθυρο μεγέθους 36 εικόνων ανά είσοδο, σε αντίθεση με το προτεινόμενο δίκτυο που εφαρμόζει παράθυρο μεγέθους 16 καρέ. Συνολικά συμπεραίνουμε πως το προτεινόμενο μοντέλο πετυχαίνει ιδιαίτερα ανταγωνιστική απόδοση σε σχέση με τα υπόλοιπα μοντέλα για όλα τα σύνολα δεδομένων ακόμα και στις περιπτώσεις που τα δεδομένα αυτά είναι άγνωστα στο μοντέλο ανάλογα με το κάθε σχήμα εκπαίδευσης.

Όσον αφορά τους Πίνακες 5.3, 5.4 παρατηρούμε πως με εξαίρεση ελάχιστες περιπτώσεις, το προτεινόμενο δίκτυο παρουσιάζει την καλύτερη απόδοση στις περισσότερες μετρικές και πως αυτό συμβαίνει για διαφορετικά σχήματα εκπαίδευσης του δικτύου και όχι απαραίτητα για το σχήμα tSTAViS, το οποίο είναι και στο σχήμα εκπαίδευσης που αποτελείται από αυτά τα 6 σύνολα δεδομένων, κάτι το οποίο υπογραμμίζει την δυνατότητα του προτεινόμενου μοντέλου να γενικεύει καλά σε άγνωστα δεδομένα σε σύγκριση με άλλες μεθόδους.

Παρατηρούμε χαρακτηριστικά την ιδιαίτερα χαμηλή επίδοση των ACLNet, Unisal, SALEMA στις βάσεις δεδομένων Coutrot1, Coutrot2, στις οποίες αντίθετα παρατηρούμε το προτεινόμενο δίκτυο να αποδίδει σημαντικά καλύτερα, ακόμα και όταν δεν έχει εκπαιδευτεί σε παρόμοια δεδομένα (π.χ. tUHD, tDHF1K). Αντίστοιχα για τις βάσεις δεδομένων AVAD, DIEM, SumMe τα μοντέλα ACLNet, SALEMA φαίνεται να έχουν χαμηλή απόδοση.

Method \ Dataset	DIEM					Coutrot1					Coutrot2				
	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
ViDaS [STD] tDHF1K	0.5772	2.28	0.8819	0.6540	0.4896	0.4662	2.08	0.8657	0.5720	0.3927	0.5331	3.77	0.9304	0.6389	0.3921
ViDaS [ST] tDHF1K	0.5609	2.22	0.8826	0.6515	0.4839	0.4600	2.06	0.8649	0.5740	0.3918	0.5672	3.94	0.9250	0.6569	0.3881
ViDaS [STD] tHOLLY	0.5640	2.23	0.8794	0.6504	0.4763	0.4621	2.08	0.8594	0.5704	0.3927	0.4910	3.31	0.9224	0.6464	0.3652
ViDaS [ST] tHOLLY	0.5585	2.20	0.8766	0.6459	0.4761	0.4526	2.03	0.8565	0.5650	0.3872	0.4742	3.19	0.9137	0.6451	0.3420
ViDaS [STD] tUCF	0.4732	1.88	0.8506	0.6386	0.4071	0.3844	1.69	0.8403	0.5656	0.3358	0.4959	3.39	0.9295	0.6795	0.3190
ViDaS [ST] tUCF	0.4651	1.83	0.8561	0.6357	0.4067	0.3730	1.68	0.8331	0.5737	0.3358	0.4566	3.02	0.9043	0.6791	0.2886
ViDaS [STD] tUHD	0.5693	2.26	0.8834	0.6521	0.4829	0.4791	2.18	0.8635	0.5759	0.4035	0.4258	3.01	0.9250	0.6329	0.3437
ViDaS [ST] tUHD	0.5721	2.27	0.8834	0.6551	0.4870	0.4877	2.22	0.8705	0.5787	0.4076	0.5295	3.73	0.9351	0.6555	0.3894
ViDaS [STD] tSTAViS	**0.6387**	**2.50**	**0.8995**	**0.6848**	**0.5278**	**0.5256**	**2.37**	**0.8786**	**0.5891**	**0.4300**	**0.7701**	**5.71**	**0.9633**	0.7147	0.5577
ViDaS [ST] tSTAViS	0.6342	2.48	0.8963	0.6834	0.5254	0.5056	2.28	0.8754	0.5841	0.4196	0.7679	5.64	0.9627	**0.7150**	**0.5589**
DeepNet [51] [S]	0.4075	1.52	0.8321	0.6227	0.3183	0.3402	1.41	0.8248	0.5597	0.2732	0.3012	1.82	0.8966	0.6000	0.2019
DVA [77] [S]	0.4779	1.97	0.8547	0.641	0.3785	0.4306	2.07	0.8531	0.5783	0.3324	0.4634	3.45	0.9328	0.6324	0.2742
SAM [9] [S]	0.4930	2.05	0.8592	0.6446	0.4261	0.4329	2.11	0.8571	0.5768	0.3672	0.4194	3.02	0.9320	0.6152	0.3041
SalGAN [50] [S]	0.4868	1.89	0.8570	0.6609	0.3931	0.4161	1.85	0.8536	0.5799	0.3321	0.4398	2.96	0.9331	0.6183	0.2909
ACLNet [78, 80] [ST]	0.5229	2.02	0.8690	0.6221	0.4279	0.4253	1.92	0.8502	0.5429	0.3612	0.4485	3.16	0.9267	0.5943	0.3229
DeepVS [28] [ST]	0.4523	1.86	0.8406	0.6256	0.3923	0.3595	1.77	0.8306	0.5617	0.3174	0.4494	3.79	0.9255	0.6469	0.2590
TASED [46] [ST]	0.5579	2.16	0.8812	0.6579	0.4615	0.4799	2.18	0.8676	0.5808	0.3884	0.4375	3.17	0.9216	0.6118	0.3142
Unisal [15] [ST]	0.5711	2.36	0.8789	0.6435	0.4822	0.4248	2.07	0.8489	0.5642	0.3714	0.3647	2.82	0.9301	0.5986	0.3012
SALEMA [40] [ST]	0.5180	2.13	0.8638	0.6320	0.4515	0.4334	2.05	0.8505	0.5608	0.3747	0.4671	3.67	0.9273	0.6162	0.3402
STAViS [ST] [74]	0.5665	2.19	0.8792	0.6648	0.4719	0.4587	1.99	0.8617	0.5764	0.3842	0.6529	4.19	0.9405	0.6895	0.4470
STAViS [STA] [74]	0.5795	2.26	0.8838	0.6741	0.4824	0.4722	2.11	0.8686	0.5847	0.3935	0.7349	5.28	0.9581	0.7106	0.5111

Πίνακας 5.3: Αξιολόγηση των αποτελεσμάτων της Εμφάνειας στις βάσεις δεδομένων DIEM, Coutrot1 και Coutrot2.Τα αποτελέσματα της προτεινόμενης μεθόδου (ViDaS [STD]) και της RGB μόνο μεθόδου ([ST]) εμφανίζονται για διαφορετικά σχήματα εκπαίδευσης.

Σε σχέση με τη συνεισφορά του βάθους στην μοντελοποίηση του προβλήματος της Εμφάνειας, παρατηρούμε πως για τα περισσότερα σύνολα δεδομένων, η ενσωμάτωση της πληροφορίας αυτής στην διαδικασία της εκπαίδευσης λειτουργεί βοηθητικά και οδηγεί σε ακριβέστερες προβλέψεις και καλύτερη απόδοση στις περισσότερες μετρικές. Από τα αποτελέσματα παραητρούμε πως η συνεισφορά του βάθους είναι ακόμη μεγαλύτερη όταν το μοντέλο εξετάζεται ως προς την απόδοσή του σε άγνωστα δεδομένα που δεν έχει ξανασυναντήσει, κάτι το οποίο υπογραμμίζει την σημασία του βάθους στην δυνατότητα γενίκευσης του μοντέλου.

Στα Σχήματα 5.4, 5.5, 5.6 εξετάζεται και οπτικά η συνεισφορά της πληροφορίας του βάθους στην μοντελοποίηση του προβλήματος με την πάροδο του χρόνου. Όλα τα βίντεο των Σχημάτων αυτών είναι επιλεγμένα με τέτοιο τρόπο ώστε να περιέχουν πολλά επίπεδα βάθους προκειμένου να γίνει αισθητή η συνεισφορά του. Σε κάθε Σχήμα βλέπουμε 4 καρέ από βίντεο, τα οποία έχουν δειγματοληφθεί σε σχετικά κοντινά και ίσα χρονικά διαστήματα μεταξύ τους, τα οποία παρουσιάζονται μαζί με τις πραγματικές τιμές αλλά και τις αντίστοιχες προβλέψεις για τους χάρτες Εμφάνειας για τις δύο μορφές του προτεινόμενου μοντέλου (RGB και RGBD). Στα Σχήματα 5.5, 5.6 βλέπουμε επίσης και τους αντίστοιχους χάρτες βάθους όπως αυτοί υπολογίστηκαν με τη μέθοδο του MiDas [56] για τις RGB εικόνες.

Method \ Dataset	AVAD					SumMe					ETMD				
	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
ViDaS [STD] tDHF1K	0.6270	3.45	0.9204	0.5945	0.4941	0.4352	2.18	0.8916	0.6492	0.3602	0.5252	2.74	0.9247	0.7000	0.4176
ViDaS [ST] tDHF1K	0.6329	3.46	0.9227	0.5951	**0.4989**	0.4417	2.20	0.8938	0.6489	0.3655	0.5108	2.66	0.9215	0.6921	0.4107
ViDaS [STD] tHOLLY	0.6088	3.28	0.9191	0.5943	0.4799	0.4092	2.01	0.8801	0.6412	0.3436	0.5378	2.78	0.9253	0.7109	0.4221
ViDaS [ST] tHOLLY	0.6091	3.26	0.9176	0.5901	0.4821	0.3942	1.92	0.8760	0.6291	0.3367	0.5287	2.72	0.9236	0.7040	0.4181
ViDaS [STD] tUCF	0.5389	2.86	0.9124	0.5917	0.4064	0.3601	1.84	0.8581	0.6342	0.3004	0.4234	2.20	0.8962	0.6900	0.3188
ViDaS [ST] tUCF	0.4755	2.46	0.9031	0.5806	0.3793	0.3417	1.75	0.8542	0.6317	0.3014	0.4025	2.08	0.8893	0.6848	0.3181
ViDaS [STD] tUHD	0.6374	3.51	0.9241	0.5960	0.4983	0.4249	2.12	0.8891	0.6478	0.3550	0.5450	2.84	0.9284	0.7146	0.4283
ViDaS [ST] tUHD	0.6270	3.41	0.9216	0.5933	0.4924	0.4240	2.11	0.8886	0.6444	0.3552	0.5404	2.80	0.9268	0.7128	0.4240
ViDaS [STD] tSTAViS	**0.6481**	3.45	**0.9262**	**0.5991**	0.4976	**0.4541**	2.24	**0.8973**	0.6675	0.3627	**0.5882**	**3.07**	**0.9349**	0.7374	**0.4537**
ViDaS [ST] tSTAViS	0.6371	3.39	0.9242	0.5955	0.4988	0.4458	2.20	0.8971	0.6645	0.3617	0.5791	3.02	0.9336	0.7321	0.4499
DeepNet [51] [S]	0.3831	1.85	0.8690	0.5616	0.2564	0.3320	1.55	0.8488	0.6451	0.2274	0.3879	1.90	0.8897	0.6992	0.2253
DVA [77] [S]	0.5247	3.00	0.8887	0.5820	0.3633	0.3983	2.14	0.8681	0.6686	0.2811	0.4965	2.72	0.9039	0.7288	0.3165
SAM [9] [S]	0.5279	2.99	0.9025	0.5777	0.4244	0.4041	2.21	0.8717	0.6728	0.3272	0.5068	2.78	0.9073	0.7310	0.3790
SalGAN [50] [S]	0.4912	2.55	0.8865	0.5799	0.3608	0.3978	1.97	0.8754	**0.6882**	0.2897	0.4765	2.46	0.9035	**0.7463**	0.3117
ACLNet [78, 80] [ST]	0.5809	3.17	0.9053	0.5600	0.4463	0.3795	1.79	0.8687	0.6092	0.2965	0.4771	2.36	0.9152	0.6752	0.3290
DeepVS [28] [ST]	0.5281	3.01	0.8968	0.5858	0.3914	0.3172	1.62	0.8422	0.6120	0.2622	0.4616	2.48	0.9041	0.6861	0.3495
TASED [46] [ST]	0.6006	3.16	0.9146	0.5898	0.4395	0.4288	2.10	0.8840	0.6570	0.3337	0.5093	2.63	0.9164	0.7117	0.3660
Unisal [15] [ST]	0.6220	**3.69**	0.9143	0.5924	0.4969	0.4459	**2.37**	0.8899	0.6480	**0.3725**	0.5432	2.96	0.9275	0.7093	0.4287
SALEMA [40] [ST]	0.5500	3.17	0.9067	0.5738	0.4441	0.4073	2.10	0.8772	0.6290	0.3440	0.5108	2.76	0.9192	0.6955	0.4057
STAViS [ST] [74]	0.6041	3.07	0.9157	0.5900	0.4431	0.4180	1.98	0.8848	0.6477	0.3325	0.5602	2.84	0.9290	0.7278	0.4121
STAViS [STA] [74]	0.6086	3.18	0.9196	0.5936	0.4578	0.4220	2.04	0.8883	0.6562	0.3373	0.5690	2.94	0.9316	0.7317	0.4251

Πίνακας 5.4: Αξιολόγηση των αποτελεσμάτων της Εμφάνειας στις βάσεις δεδομένων AVAD, SumMe και ETMD.Τα αποτελέσματα της προτεινόμενης μεθόδου (ViDaS [STD]) και της RGB μόνο μεθόδου ([ST]) εμφανίζονται για διαφορετικά σχήματα εκπαίδευσης.

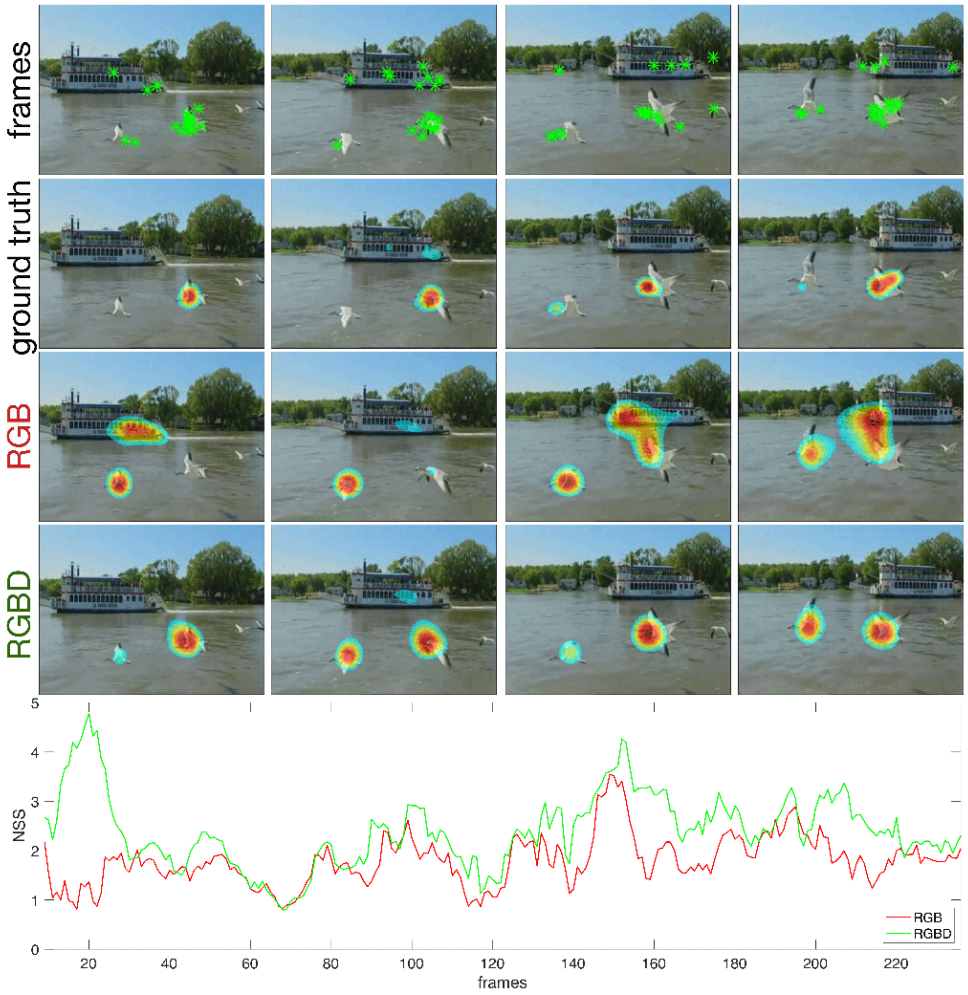
Συγκεκριμένα στο Σχήμα 5.4 βλέπουμε επίσης ένα διάγραμμα στο οποίο απεικονίζεται η τιμή της μετρικής NSS κατά την πάροδο του χρόνου για τα καρέ του συγκεκριμένου βίντεο. Βλέπουμε πως το συγκεκριμένο βίντεο απεικονίζει περιεχόμενο με πολλά επίπεδα βάθους (γλάροι, ποτάμι, πλοίο, δέντρα κ.λ.π.) και συνεπώς το βάθος παίζει ακόμα σημαντικότερο ρόλο στις προβλέψεις για τη συγκεκριμένη σκηνή. Για τα απεικονιζόμενα καρέ βλέπουμε πως οι προβλέψεις του RGBD μοντέλου είναι σημαντικά πιο ακριβείς και συγκετρωμένες στην πραγματική τιμή. Παρατηρούμε χαρακτηστικά πως η πληροφορία του βάθους δίνει την δυνατότητα στο μοντέλο να δώσει περισσότερη προσοχή στα κοντινότερα αντικείμενα της εικόνας (δηλαδή τους γλάρους) και να αγνοήσει τις λανθασμένα θετικές περιοχές που προέβλεψε το RGB μοντέλο. Κατά συνέπεια βλέπουμε πως η μετρική NSS είναι σε όλα τα σημεία υψηλότερη στην RGBD εκδοχή του προτεινόμενου μοντέλου, υποδεικνύοντας την συνεισφορά της πληροφορίας του βάθους, με τη μεγαλύτερη απόκλιση να εμφανίζεται στα πρώτα καρέ του βίντεο, πιθανώς γιατί σε αυτά το μοντέλο δεν έχει ακόμα εξάγει ιδιαίτερα χρήσιμα χρονικά χαρακτηριστικά αλλά βασίζεται περισσότερο στα χωρικά χαρακτηριστικά των εικόνων, αντιλαμβάνοντάς τες ως πιο στατικές.

Στα Σχήματα 5.5, 5.6 βλέπουμε ξανά επιλεγμένα καρέ από δύο διαφορετικά σκηνικά τα οποία απέχουν κατά ίσα χρονικά διαστήματα μεταξύ τους, καθώς και τους αντίστοιχους χάρτες βάθους για κάθε εικόνα. Όπως φαίνεται και απο τους χάρτες βάθους, και στις δύο περιπτώσεις πρόκειται για σκηνικά όπου το σχετικό βάθος των εικόνων αλλάζει με την πάροδο του χρόνου, με έναν άντρα να κατευθύνεται από πιο μακριά προς την κάμερα.

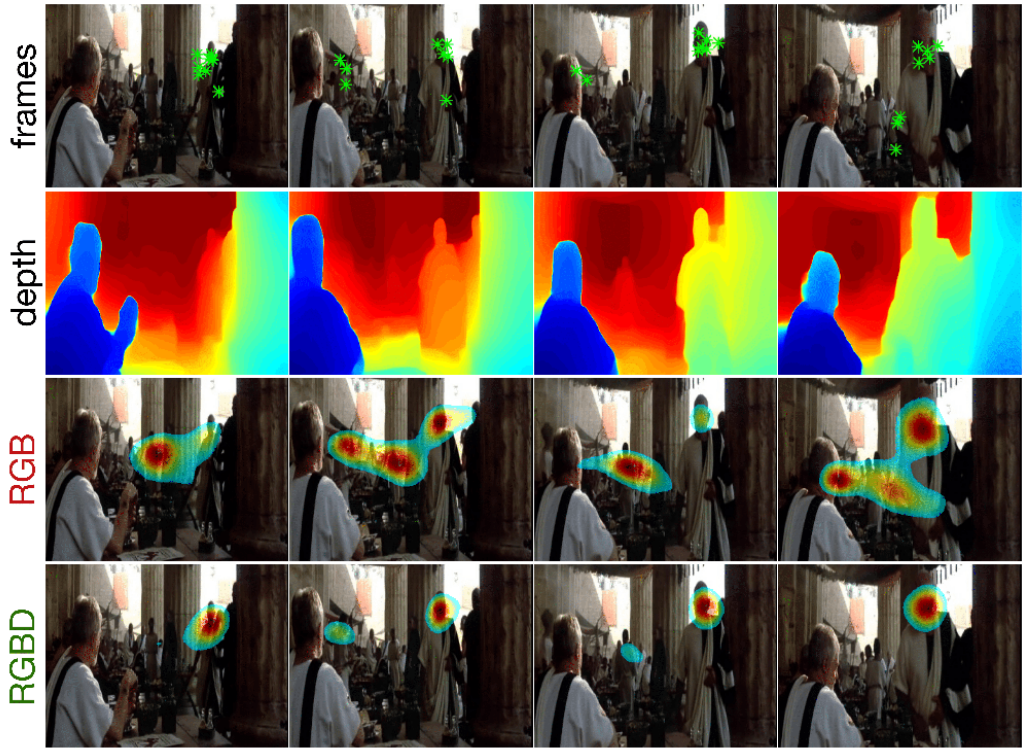
Και στις δύο περιτπώσεις βλέπουμε πως η RGBD εκδοχή επιφέρει ακριβέστερες και πιο στοχευμένες προβλέψεις, εστιασμένες περισσότερο στην πραγματική τιμή, αποκλείοντας λανθασμένα θετικά και αναγνωρίζοντας λανθασμένα αρνητικές περιοχές του RGB μοντέλου. Στο σημείο αυτό και με βάση αυτά τα δύο Σχήματα, αξίζει να αναφερθεί πως όπως παρατηρούμε, η πληροφορία του βάθους φαίνεται να λειτουργεί βοηθητικά και να συνεισφέρει ακόμα και στις περιπτώσεις όπου το αντικείμενο το οποίο τραβά την προσοχή του ανθρώπου βρίσκεται σε απομακρυσμένο σημείο συγκριτικά με τα υπόλοιπα. Για παράδειγμα στο Σχήμα 5.5 βλέπουμε πως σε όλες τις εικόνες με έμφαση στις πρώτες δύο στήλες, η περιοχή στην οποία επικετρώνεται η προσοχή του ανθρώπου δεν είναι η κοντινότερη στην κάμερα περιοχή. Παρόλαυτα βλέπουμε πως η πληροφορία του βάθους βοηθά αποτελεσματικά στην ακριβέστερη εκτίμηση του χάρτη Εμφάνειας.

Τέλος παρουσιάζονται τα Σχήματα 5.7, 5.8, 5.9, στα οποία συγκρίνονται οι εκτιμήσεις του χάρτη Εμφάνειας του προτεινόμενου RGBD μοντέλου, με τις εκτιμήσεις άλλων μεθόδων για πολλά τυχαία δείγματα από τα διάφορα σύνολα δεδομένων παρακολούθησης ματιού. Βλέπουμε πως η προτεινόμενη μέθοδος παρουσιάζει ακριβέστερες προβλέψεις τόσο σε δείγματα με σημαντική πληροφορία βάθους, όσο και σε δείγματα όπου το βάθος έχει μικρότερο εύρος και πιθανώς να παίζει μικρότερο ρόλο. Μεταξύ άλλων, από τα εικονιζόμενα αποτελέσματα προκύπτει πως το προτεινόμενο μοντέλο μπορεί να παράξει προβλέψεις καλύτερα εστιασμένες γύρω από την περιοχή ενδιαφέροντος μειώνοντας τα λανθασμένα θετικά σημεία της πρόβλεψης. Επίσης φαίνεται πως

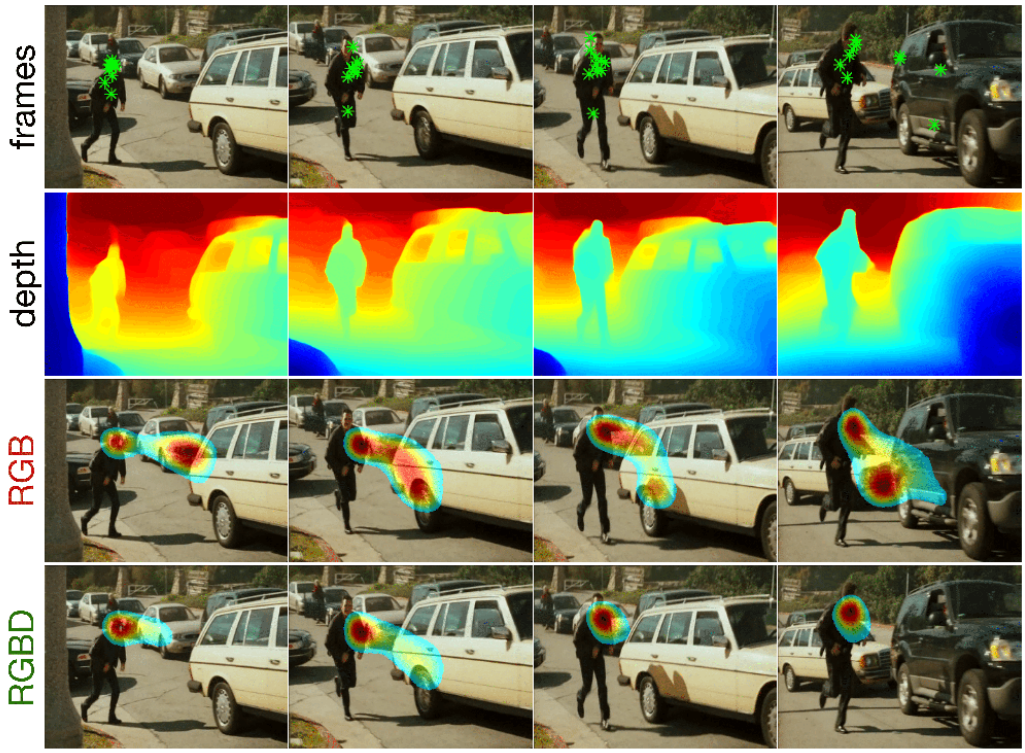
σε πολλά δείγματα, τα περισσότερα μοντέλα παρουσιάζουν μία προκατάληψη υπέρ των περιοχών που εμφανίζονται πρόσωπα (Face Bias) και αγνοούν τις υπόλοιπες περιοχές ενδιαφέροντος. Αντίθετα το προτεινόμενο μοντέλο φαίνεται πως έχει καταφέρει να ξεπεράσει αυτήν την προκατάληψη και να εντοπίσει την πραγματικά σημαντική περιοχή της εικόνας (π.χ. πρώτη και τρίτη στήλη των Σχημάτων 5.8, 5.9).



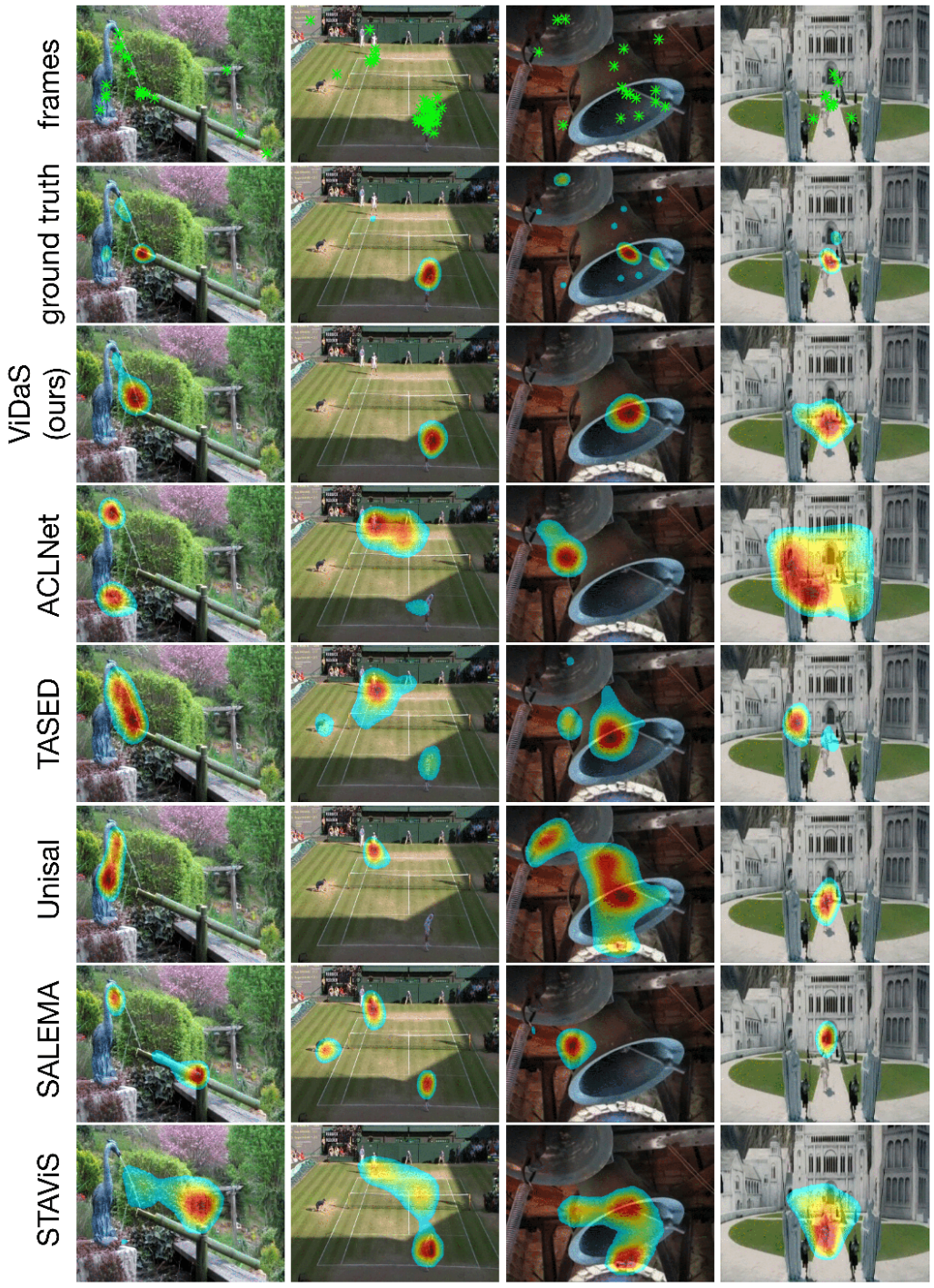
Σχήμα 5.4: Τυχαία δείγματα της βάσης δεδομένων Coutrot1 μαζί με τα δεδομένα παρακολούθησης ματιού, τους αντίστοιχους πραγματικούς χάρτες Εμφάνειας, τις αντίστοιχες εκτιμήσεις του RGB και RGBD προτεινόμενου μοντέλου καθώς και την NSS καμπύλη για την πάροδο του χρόνου.



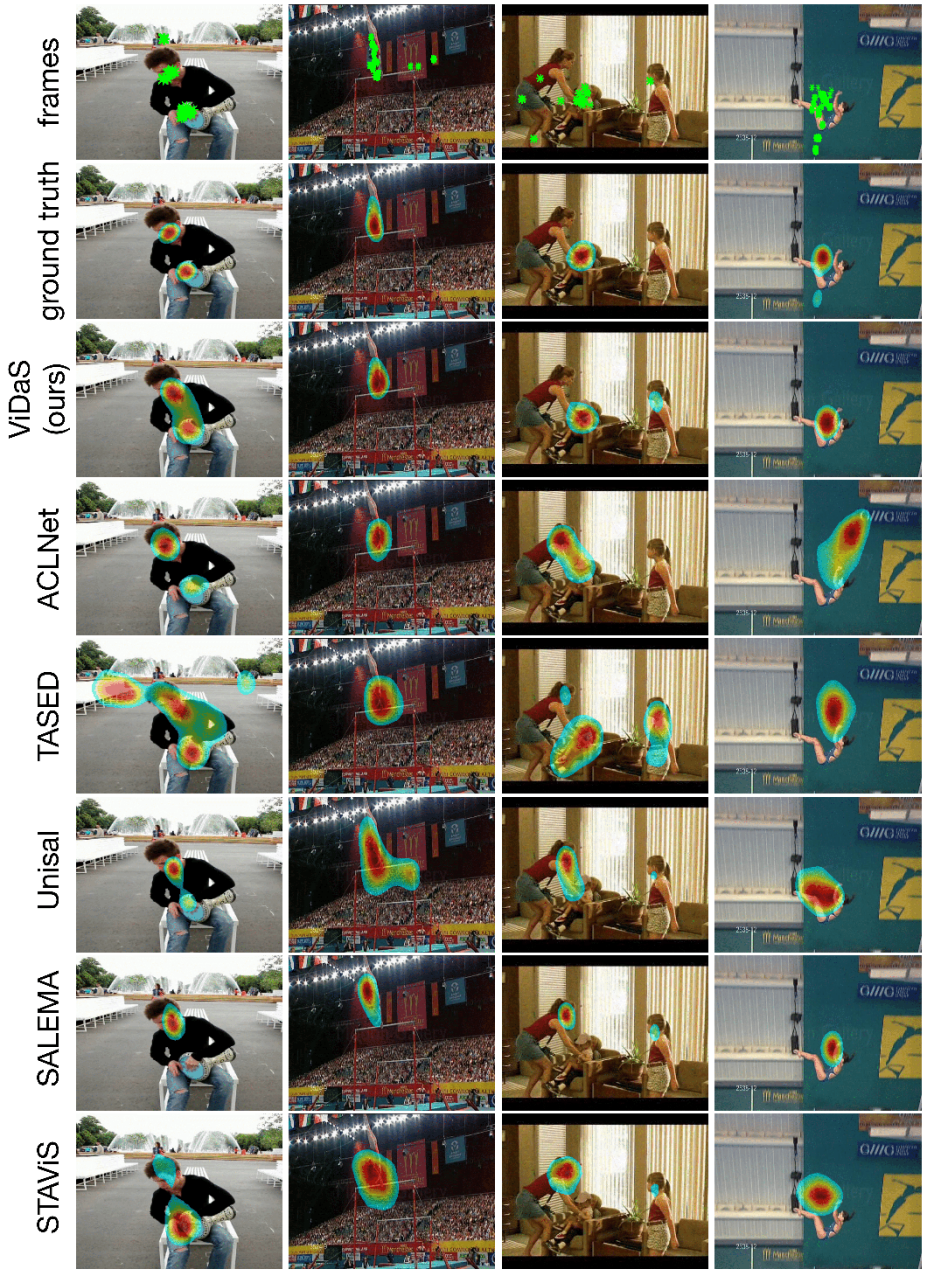
Σχήμα 5.5: Καρέ μαζί με τα δεδομένα παρακολούθησης ματιού καθώς και τους αντίστοιχους χάρτες βάθους από μία ταινία του Χόλιγουντ. Στην τρίτη γραμμή φαίνονται οι εκτιμήσεις των χαρτών Εμφάνειας του RGB μοντέλου, ενώ στην τελευταία γραμμή φαίνονται οι εκτιμήσεις του προτεινόμενου RGBD μοντέλου, το οποίο φαίνεται να επιτυγχάνει καλύτερα στην πρόβλεψη της ανθρώπινης προσοχής.



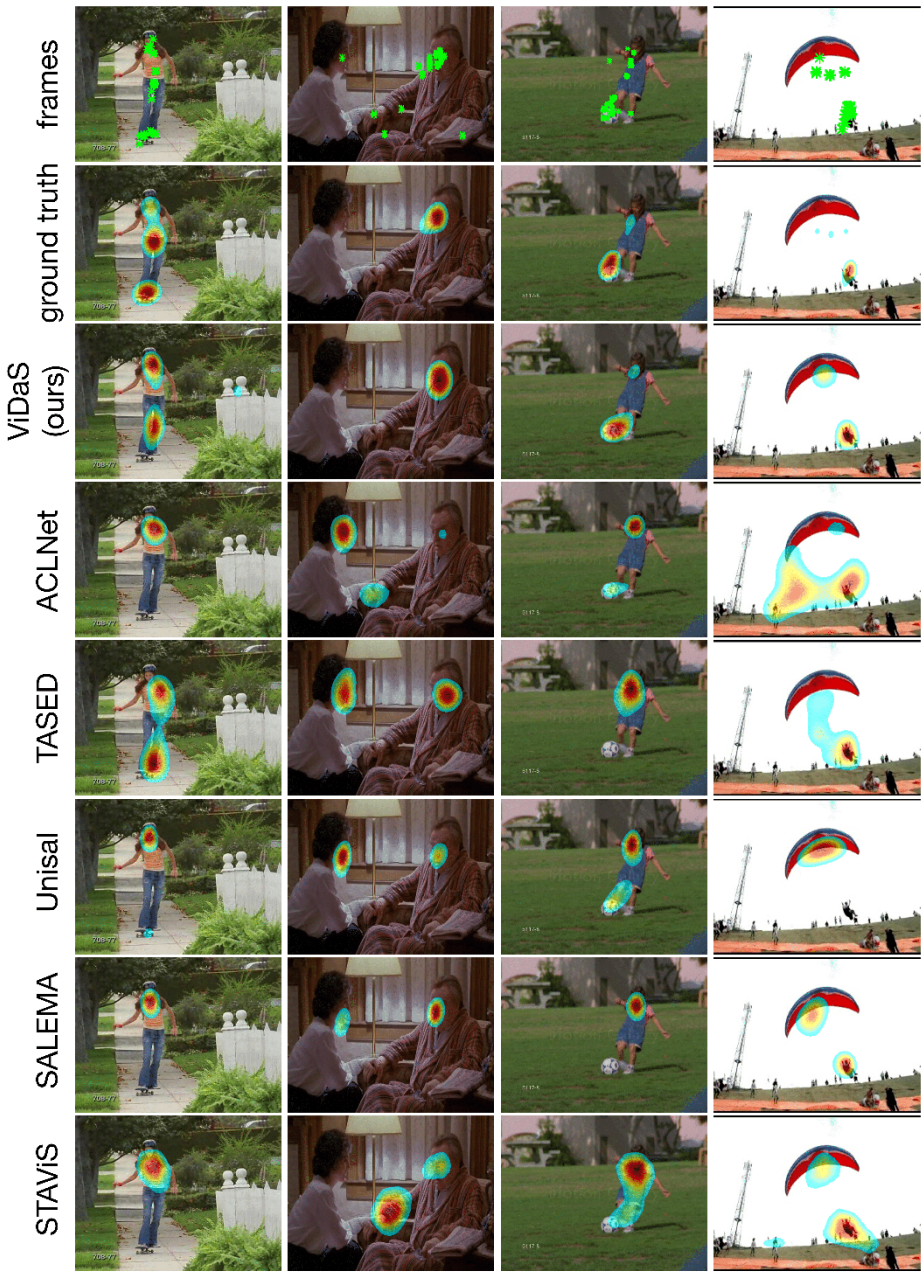
Σχήμα 5.6: Οι RGB ακολουθίες εικόνων από μία ταινία του Χόλιγουντ από το σύνολο δεδομένων ETMD μαζί με τα δεδομένα παρακολούθησης ματιού (1η γραμμή) και τους αντίστοιχους υπολογισμένους χάρτες βάθους (2η γραμμή). Είναι εμφανές πως οι προβλέψεις του RGBD μοντέλου (4η γραμμή) είναι πιο ακριβείς από αυτές της RGB μόνο εκδοχής (3η γραμμή).



Σχήμα 5.7: Τυχαία δείγματα από διάφορες βάσεις δεδομένων μαζί με τα δεδομένα παρακολούθησης ματιού,τους πραγματικούς χάρτες Εμφάνειας, τις εκτιμήσεις του προτεινόμενου RGBD μοντέλου και τις εκτίμησεις αρκετών άλλων state-of-the-art μεθόδων.



Σχήμα 5.8: Οι RGB ακολουθίες εικόνων από τα AVAD, UCF Sports και Hollywood 2 σύνολα δεδομένων μαζί με τα δεδομένα παρακολούθησης ματιού (1η γραμμή) και τους πραγματικούς χάρτες Εμφάνειας (2η γραμμή). Η 3η γραμμή απεικονίζει τις προβλέψεις του προτεινόμενου RGBD μοντέλου, ενώ οι υπόλοιπες 5 γραμμές απεικονίζουν τις προβλέψεις από άλλα 5 state-of-the-art μοντέλα: ACLNet, TASED, Unisal, SALEMA και STAViS.



Σχήμα 5.9: Οι RGB ακολουθίες εικόνων από τα DHF1K, UCF Sports και Hollywood 2 βάσεις δεδομένων μαζί με τα δεδομένα παρακολούθησης ματιού (1η γραμμή) και τους πραγματικούς χάρτες Εμφάνειας (2η γραμμή). Η 3η γραμμή απεικονίζει τις προβλέψεις του προτεινόμενου RGBD μοντέλου, ενώ οι υπόλοιπες 5 γραμμές απεικονίζουν τις προβλέψεις από άλλα 5 state-of-the-art μοντέλα: ACLNet, TASED, Unisal, SALEMA και STAViS.

Κεφάλαιο 6

Συνεισφορά, Συμπεράσματα και Μελλοντική Έρευνα

Στην παρούσα Διπλωματική Εργασία εξετάστηκε το πρόβλημα της Μοντελοποιήσης της Ανθρώπινης Προσοχής σε βίντεο, τόσο σε Δισδιάστατα όσο και σε Τρισδιάστατα Δεδομένα, χρησιμοποιώντας τεχνικές Βαθιάς Μάθησης και Νευρωνικών Δικτύων. Για το σκοπό αυτό χρησιμοποιήθηκαν 9 διαφορετικά Σύνολα Δεδομένων Παρακολούθησης Ματιού (Eye-tracking Data) και έγινε σύγκριση με άλλες 11 διαφορετικές state-of-the-art μεθόδους, όπου φαίνεται πως η προτεινόμενη μέθοδος πετυχαίνει ανταγωνιστική αν όχι καλύτερη απόδοση στις περισσότερες περιπτώσεις.

6.1 Συνεισφορά και Συμπεράσματα

Η συνεισφορά της Διπλωματικής Εργασίας μπορεί να συνοψιστεί στα παρακάτω σημεία:

- Επισημείωση μεγάλου μέρους των δεδομένων του NYU Depth Dataset V2 με τους αντίστοιχους χάρτες βάθους.
- Σύγκριση 3 διαφορετικών μεθόδων υπολογισμού βάθους από δισδιάστατες εικόνες.
- Εμπλουτισμός 9 υπάρχοντων συνόλων δεδομένων παρακολούθησης ματιού (Eye-tracking datasets) με τους αντίστοιχους χάρτες βάθους και από τις 3 μεθόδους.
- Σχεδιασμός και υλοποίηση ενός μοντέλου πρόβλεψης της ανθρώπινης προσοχής σε δισδιάστατα βίντεο, χρησιμοποιώντας μόνο τα RGB δεδομένα.
- Σχεδιασμός και υλοποίηση ενός μοντέλου δύο ροών για την πρόβλεψη της ανθρώπινης προσοχής σε τρισδιάστατα βίντεο, ενσωματώνοντας και την πληροφορία του βάθους σε μία δεύτερη οπτική ροή και συγχωνεύοντας τις εξόδους των δύο ροών.

- Σύγκριση των δύο μοντέλων μεταξύ τους και εξαγωγή συμπερασμάτων σχετικά με την συνεισφορά του βάθους στο συγκεκριμένο πρόβλημα.
- Σύγκριση των μοντέλων με άλλα 11 state-of-the-art μοντέλα σε 9 διαφορετικά σύνολα δεδομένων και εξαγωγή συμπερασμάτων.

Με βάση τα πειραματικά αποτελέσματα που παρουσιάστηκαν, τόσο όσον αφορά τις μετρικές αξιολόγησης αλλά και τα Σχήματα, μπορούμε να συμπαιράνουμε τα εξής:

- Η πληροφορία του βάθους συμβάλλει πράγματι στην εκτίμηση μίας ακριβέστερης και χωρικά πιο στοχευμένης πρόβλεψης της ανθρώπινης προσοχής σε βίντεο και στην υλοποίηση ενός πιο γενικευμένου μοντέλου που αποδίδει αρκετά καλά σε νέα δεδομένα.
- Το βάθος μπορεί να λειτουργήσει συνδυαστικά με την RGB πληροφορία ώστε να βελτιώσει τις προβλέψεις, ενώ από μόνο του ως είσοδος δεν είναι αρκετό.
- Η μέθοδος υπολογισμού και η μορφή των χαρτών βάθους παίζει σημαντικό ρόλο στην συνεισφορά αυτής της επιπρόσθετης πληροφορίας στην εκτίμηση της ανθρώπινης προσοχής.
- Στην προτεινόμενη αρχιτεκτονική το βάθος ως πληροφορία λειτουργεί βοηθητικά στις περιπτώσεις όπου αυτό είναι εφικτό, ενώ δεν αποτρέπει τις ακριβείς προβλέψεις στις περιπτώσεις όπου η περιοχή της εικόνας στην οποία ο άνθρωπος εστίασε την προσοχή του είναι σε μεγαλύτερη σχετική απόσταση.
- Από τα πειράματα σχετικά με τις διαφορετικές αρχιτεκτονικές του συνολικού μοντέλου και της αλληλεπίδρασης των δύο ροών, προκύπτει πως με την προτεινόμενη προσέγγιση επιτυγχάνονται τα βέλτιστα αποτελέσματα, καθώς οι δύο ροές έχουν τη δυνατότητα να μάθουν ανεξάρτητα τις αναπαραστάσεις από τις δύο εισόδους αντίστοιχα και το συνολικό μοντέλο μπορεί να μάθει πότε να χρησιμοποιεί την πληροφορία του βάθους, αλλά και πότε αυτή δεν συνεισφέρει στην εκτίμηση.

6.2 Μελλοντικές Επεκτάσεις

Με βάση την έρευνα και τα αποτελέσματα της παρούσας Διπλωματικής Εργασίας, μπορούμε να προτείνουμε τις εξής μελλοντικές κατευθύνσεις:

- **Εφαρμογή του εκπαιδευμένου μοντέλου στην ρομποτική.** Η προτεινόμενη μέθοδος θα μπορούσε να εφαρμοστεί σε κάποιο ρομποτικό σύστημα, προκειμένου αυτό να μπορέσει να προσομοιώσει τον άνθρωπο όσον αφορά την εστίαση της οπτικής του προσοχής, με αποτέλεσμα να εστιάζει τους αισθητήρες του στο σημαντικότερο αντικείμενο/ γεγονός κ.λ.π.

- **Δοκιμή ακόμα περισσότερων μεθόδων συγχώνευσης των δύο ροών.** Όπως παρουσιάστηκε, στην παρούσα Διπλωματική Εργασία, πραγματοποιήθηκε εκτεταμένη έρευνα σχετικά με την αρχιτεκτονική του μοντέλου και την αλληλεπίδραση των δύο ροών. Παρόλαυτα υπάρχει πάντα χώρος για περισσότερα πειράματα και νέες δοκιμές που μπορεί να επιδόσουν ακόμα πιο αποτελεσματικά.

- **Δοκιμή διαφορετικών μεθόδων εξαγωγής του βάθους.** Στην μέθοδο που παρουσιάστηκε, εξετάστηκαν 3 διαφορετικές μέθοδοι εξαγωγής βάθους, οι οποίες βασίζονται σε μοντέλα βαθιάς μάθησης. Θα ήταν ενδιαφέρον να ερευνηθούν και εναλλακτικοί τρόποι υπολογισμού του βάθους, όπως αισθητήρες βάθους ή μέθοδοι βασισμένοι στην παράλλαξη (Disparity), οι οποίοι αποτέλουν και πιο γενικευμένες μεθόδους που έχουν την ίδια αποτελεσματικότητα ανεξαρτήτως δεδομένων.

- **Επέκταση της πληροφορίας εισόδου του μοντέλου.** Στην παρούσα Διπλωματική Εργασία εξετάσαμε την συνεισφορά του βάθους στο πρόβλημα της Εμφάνειας επεκτείνοντας το μοντέλο και ενσωματώνοντας την πληροφορία αυτή. Έτσι θα μπορούσε να εξεταστεί και η συνεισφορά άλλου είδους πληροφορίας, όπως ο ήχος, η οπτική ροή, η ανίχνευση προσώπου κ.ά. τόσο συνεργατικά με το βάθος όσο και ανεξάρτητα.

Βιβλιογραφία

[1] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1688–1698, 2017.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision – ECCV 2012. ECCV 2012*, ser. LNCS, vol. 7577. Springer, Berlin, Heidelberg, Oct. 2012, pp. 611–625.

[3] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[4] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 221–230.

[5] Q. Chang, S. Zhu, and L. Zhu, "Temporal-spatial feature pyramid for video saliency detection," *arXiv preprint arXiv:2105.04213*, 2021.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[7] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D salient object detection via 3D convolutional neural networks," *arXiv preprint arXiv:2101.10241*, 2021.

[8] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single-image depth perception in the wild," *NIPS*, vol. 29, pp. 730–738, 2016.

[9] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[10] A. Coutrot and N. Guyader, "Learning a time-dependent master saliency map from eye-tracking data in videos," *arXiv preprint arXiv:1702.00714*, 2017.

[11] X. Cui, K. Zheng, L. Gao, D. Yang, and J. Ren, "Multiscale spatial-spectral convolutional network with image-based framework for hyperspectral imagery classification," *Remote Sensing*, vol. 11, no. 19, p. 2220, Sept 2019.

[12] I. Dabbura, "Gradient descent algorithm and its variants," 2017, available at https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3.

[13] T. Dang, C. Papachristos, and K. Alexis, "Visual saliency-aware receding horizon autonomous exploration with application to aerial robotics," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2526–2533.

[14] I. Diamanti, A. Tsiami, P. Koutras, and P. Maragos, "ViDaS: Video depth-aware saliency network," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, submitted.

[15] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020*, ser. LNCS, vol. 12350. Springer, Cham, 2020, pp. 419–435.

[16] O. Duerr, B. Sick, and E. Murina, *Probabilistic Deep Learning: With Python, Keras and TensorFlow Probability.* Manning Publications, 2020.

[17] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2650–2658.

[18] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, and Y. Avrithis, "Movie summarization based on audiovisual saliency detection," in *2008 15th IEEE International Conference on Image Processing*, 2008, pp. 2528–2531.

[19] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[20] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[21] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2013.

[22] Z. Hao, Y. Li, S. You, and F. Lu, "Detail preserving depth estimation from a single image using attention guided networks," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 304–313.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[26] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "Vinet: Pushing the limits of visual modality for audio-visual saliency prediction," *arXiv preprint arXiv:2012.06170*, 2020.

[27] A. Janoch *et al.*, "A category-level 3d object dataset: Putting the kinect to work," in *Fossati A., Gall J., Grabner H., Ren X., Konolige K. (eds) Consumer Depth Cameras for Computer Vision. Advances in Computer Vision and Pattern Recognition.* Springer, London, 2013, pp. 141–165.

[28] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 602–617.

[29] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[30] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 2106–2113.

[31] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[32] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4131–4144, 2018.

[33] P. Koutras and P. Maragos, "A perceptually based spatio-temporal computational framework for visual saliency estimation," *Signal Processing: Image Communication*, vol. 38, pp. 15–31, 2015.

[34] ——, "Susinet: See, understand and summarize it," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[35] C. Kulkarni, "Learning rate tuning and optimizing," 2018, available at https://medium.com/@ck2886/learning-rate-tuning-and-optimizing-d03e042d0500.

[36] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2019.

[37] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, and R. Raskar, "Learning gaze transitions from depth to improve video saliency estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1698–1707.

[38] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization." *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 689–694, 2004.

[39] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] P. Linardos, E. Mohedano, J. J. Nieto, K. McGuinness, X. Giro-i Nieto, and N. E. O'Connor, "Simple vs complex temporal recurrences for video saliency prediction," in *British Machine Vision Conf. (BMVC)*, 2019.

[41] V. Lyudvichenko, M. Erofeev, Y. Gitman, and D. Vatolin, "A semiautomatic saliency model and its application to video compression," in *Proceedings of the IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2017, pp. 403–410.

[42] V. Lyudvichenko, M. Erofeev, A. Ploshkin, and D. Vatolin, "Improving video compression with deep visual-attention models," in *Proceedings of the International Conference on Intelligent Medicine and Image Processing*, 2019, pp. 88–94.

[43] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[44] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1408–1424, 2014.

[45] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[46] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2394–2403.

[47] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 1, pp. 1–23, 2016.

[48] P. Mital, T. Smith, R. L. Hill, and J. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, pp. 5–24, 2010.

[49] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2012.

[50] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[51] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 598–606.

[52] G. Pantazis, G. Dimas, and D. K. Iakovidis, "Salsum: Saliency-based video summarization using generative adversarial networks," *arXiv preprint arXiv:2011.10432*, 2020.

[53] H. Park and J.-H. Son, "Machine learning techniques for thz imaging and time-domain spectroscopy," *Sensors*, vol. 21, no. 4, 2021.

[54] L. Pauly, H. Peel, S. Luo, D. Hogg, and R. Fuentes, "Deeper networks for pavement crack detection," in *Proceedings of the 34th International Symposium on Automation and Robotics in Construction (ISARC)*, July 2017, pp. 479–485.

[55] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7254–7263.

[56] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[57] K. Rapantzikos, G. Evangelopoulos, P. Maragos, and Y. Avrithis, "An audio-visual saliency model for movie summarization," in *2007 IEEE 9th Workshop on Multimedia Signal Processing*, 2007, pp. 320–323.

[58] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[60] S. Rowe, "Introduction to neurons in neural networks," 2019, available at https://medium.com/artificial-neural-networks/introduction-to-neurons-in-neural-networks-71828d040a65.

[61] S. Saha, "A comprehensive guide to convolutional neural networks," 2018, available at https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[62] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2008.

[63] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[64] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[65] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[66] S. Sharma, "Activation functions in neural networks," 2017, available at https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6.

[67] P. Shi, M. Billeter, and E. Eisemann, "Salientgaze: Saliency-based gaze correction in virtual reality," *Computers & Graphics*, vol. 91, pp. 83–94, 2020.

[68] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in vr: How do people explore virtual environments?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

[69] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 567–576.

[70] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer vision in sports*, vol. 71. Springer, 2014, pp. 181–208.

[71] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2012, pp. 573–580.

[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[73] A. Tsiami, P. Koutras, A. Katsamanis, A. Vatakis, and P. Maragos, "A behaviorally inspired fusion approach for computational audiovisual saliency modeling," *Signal Processing: Image Communication*, vol. 76, pp. 186 – 200, 2019.

[74] A. Tsiami, P. Koutras, and P. Maragos, "STAViS: Spatio-temporal audiovisual saliency network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[75] C. Wang, S. Lucey, F. Perazzi, and O. Wang, "Web stereo video supervision for depth prediction from dynamic scenes," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 348–357.

[76] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2017.

[77] ——, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

[78] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4894–4903.

[79] ——, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[80] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 220–237, 2019.

[81] X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang, "Salsac: a video saliency prediction model with shuffled attentions and correlation-based ConvLSTM," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 410–12 417.

[82] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[83] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1492–1500.

[84] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1395–1403.

[85] X. Yuan, J. Yue, and Y. Zhang, "Rgb-d saliency detection: Dataset and algorithm for robot vision," in *Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2018, pp. 1028–1033.

[86] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate RGB-D saliency detection," in *Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020*, ser. LNCS, vol. 12373. Springer, 2020, pp. 374–390.

[87] Y. Zhang, M. Jiang, and Q. Zhao, "Saliency prediction with external knowledge," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 484–493.

[88] T. Zhou, D.-P. Fan, M.-M. Cheng *et al.*, "RGB-D salient object detection: A survey," in *Computational Visual Media*, vol. 7. Springer, 2021, pp. 37–69.

[89] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "Pdnet: Prior-model guided depth-enhanced network for salient object detection," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 199–204.