Σύνθεση κίνησης σε βίντεο μέσω εκμάθησης τοπικών μετασχηματισμών

Η Μεταπτυχιακή Διπλωματική Εργασία

υποβάλλεται στην ορισθείσα

από τη Συνέλευση

του Τμήματος Μηχανικών Η/Υ και Πληροφορικής

Εξεταστική Επιτροπή

από την

Βιργινία Τάγκα

ως μέρος των υποχρεώσεων για την απόκτηση του

ΔΙΠΛΩΜΑΤΟΣ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΣΤΗ ΜΗΧΑΝΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΜΕ ΕΙΔΙΚΕΥΣΗ

ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΚΑΙ ΜΗΧΑΝΙΚΗ ΔΕΔΟΜΕΝΩΝ

Πανεπιστήμιο Ιωαννίνων

Πολυτεχνική Σχολή

Ιωάννινα 2021

Εξεταστική Επιτροπή:

- **Νίκου Χριστόφορος**, Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων (Επιβλέπων)
- **Βρίγκας Μιχαήλ**, Επίκ. Καθηγητής, Τμήμα Επικοινωνίας και Ψηφιακών Μέσων, Πανεπιστήμιο Δυτικής Μακεδονίας
- **Φούντος Ιωάννης**, Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων

Αφιερωση

Στους γονείς μου και στα αδέρφια μου, Λευτέρη και Αλέξανδρο.

Περιεχομενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περιληψη

Βιργινία Τάγκα, Δ.Μ.Σ. στη Μηχανική Δεδομένων και Υπολογιστικών Συστημάτων, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων, 2021.
Σύνθεση κίνησης σε βίντεο μέσω εκμάθησης τοπικών μετασχηματισμών.
Επιβλέπων: Χριστόφορος Νίκου, Καθηγητής.

Η σύνθεση κίνησης σε βίντεο (animation) αποτελεί μία σημαντική εξέλιξη στον τομέα της υπολογιστικής όρασης με εφαρμογές, μεταξύ άλλων στον κινηματογράφο, σε βιντεοπαιχνίδια, γραφικά και οπτικά εφέ. Το πρόβλημα που πραγματεύεται η εργασία αυτή αφορά στην αναπαράσταση της κίνησης σε βίντεο σύμφωνα με τοπικούς μετασχηματισμούς που εφαρμόζονται σε συγκεκριμένα χαρακτηριστικά σημεία που έχουν εξαχθεί από τις εικόνες. Τα σημεία αυτά σχετίζονται με συγκεκριμένα μέρη του ανθρώπινου σώματος ή του προσώπου με συνέπεια η εφαρμογή να επικεντρώνεται σε κινήσεις του ανθρώπινου σώματος και σε εκφράσεις του προσώπου. Η πρόκληση εδώ είναι πως δεν έχουμε καμία εκ των προτέρων πληροφορία για τα δεδομένα και προσπαθούμε να υλοποιήσουμε αναπαράσταση κινήσεων στηριζόμενοι σε βίντεο που λειτουργούν ως οδηγοί. Πιο συγκεκριμένα, οι εικόνες του βίντεο μετασχηματίζονται σύμφωνα με τις κινήσεις των αντικειμένων στις αντίστοιχες εικόνες ενός βίντεο οδηγού. Αρχικά, αφού υπολογίσουμε τις συντεταγμένες των χαρακτηριστικών σημείων του σώματος ή του προσώπου μέσω ενός προεκπαιδευμένου μοντέλου, εισάγουμε ένα συνελικτικό νευρωνικό δίκτυο που υπολογίζει μέσω της οπτικής ροής ένα πυκνό πεδίο κίνησης το οποίο υποδεικνύει τον τρόπο που μετακινούνται τα χαρακτηριστικά σημεία και δημιουργεί μία μάσκα που φανερώνει ποια τμήματα των καρέ μπορούν να ανακατασκευαστούν μέσω γεωμετρικών μετασχηματισμών. Επιπλέον, υποδηλώνει τα τμήματα που δεν μπορούν να ανακτηθούν από την αρχική εικόνα και πρέπει να εκτιμηθούν. Στη συνέχεια, εκπαιδεύουμε ένα ανταγωνιστικό νευρωνικό δίκτυο το οποίο σύμφωνα με τις πληροφορίες που λαμ-

βάνει από τις προηγούμενες εξόδους του δικτύου, παράγει νέες εικόνες με τέτοιο τρόπο ώστε να μοιάζουν όσο γίνεται περισσότερο σε εικόνες στόχους.

Extended Abstract

Virginia Tagka, M.Sc. in Data and Computer Systems Engineering, Department of Computer Science and Engineering, School of Engineering, University of Ioannina, Greece, 2021.
Image animation by learning keypoints local transformations.
Advisor: Christoforos Nikou, Professor.

Video animation is one of the most important developments in Computer Vision field with various applications such as in cinema, video games, graphic and visual effects. In this project, we try to solve the problem of motion representation in videos according to local transformations applied to specific keypoints extracted from the images. These keypoints are related to specific parts of the human body or face, so the problem is focused on human body movements and facial expressions. The challenge is that we do not have any prior information about the data and we try to implement motion representation based on videos that act as guides. More specifically, the video images are transformed according to the movements of the objects in the corresponding images of a video guide. First, after calculating the coordinates of the keypoints of the body or face through a pre-trained model, we introduce a convolutional neural network that estimates a dense motion field through optical flow. The dense motion field indicates keypoints' movements and creates a mask that reveals which parts of the frames can be reconstructed through geometric transformations. In addition, it indicates the parts that can not be retrieved from the original image and should be evaluated and impainted. Next, we train a generative adversarial network which takes into account information from previous network outputs and generates new images that resemble as much as possible with the target frames.

Κεφαλαιο 1

Εισαγωγη

Με την ταχύτατη τεχνολογική ανάπτυξη και την εξέλιξη στον χώρο της Υπολογιστικής Όρασης δημιουργούνται συνεχώς νέες προκλήσεις. Ένα σημαντικό πρόβλημα αποτελεί η αναπαράσταση κίνησης εικόνων, γνωστή και ως animation που έχει κατακτήσει τον χώρο των γραφικών και οπτικών εφέ και αποκτά όλο και μεγαλύτερο ενδιαφέρον, ειδικά στον χώρο της παραγωγής ταινιών και βιντεοπαιχνιδιών. Ο σκοπός της παρούσας εργασίας είναι να βρεθεί ένα μοντέλο, ικανό να αναπαράγει νέα βίντεο στα οποία, το αντικείμενο μελέτης θα κινείται με τον ίδιο τρόπο όπως κινείται ένα διαφορετικό αντικείμενο, της ίδιας όμως κατηγορίας, σε ένα διαφορετικό βίντεο. Με άλλα λόγια, έχοντας μία εικόνα-καρέ και ένα βίντεο που λειτουργεί ως οδηγός, προσπαθούμε να δημιουργήσουμε νέες εικόνες μετασχηματίζοντας την εικόνα-καρέ σύμφωνα με τις κινήσεις του αντικειμένου σε κάθε καρέ του βίντεο-οδηγού. Έτσι, ενώνοντας όλες τις εικόνες αποκτάμε ένα νέο βίντεο που αναπαριστά τις κινήσεις του βίντεο-οδηγού κρατώντας όμως το περιεχόμενο της εικόνας-καρέ (αντικείμενο και φόντο). Τα αντικείμενα της μελέτης μας περιλαμβάνουν ανθρώπινα σώματα και μορφασμούς του προσώπου. Έτσι, για παράδειγμα, έχοντας την εικόνα ενός ανθρώπου που στέκεται όρθιος και ένα βίντεο με έναν άνθρωπο που χορεύει, αναλύουμε τον τρόπο με τον οποίο θα δημιουργηθούν νέες εικόνες όπου ο ακίνητος άνθρωπος θα αποκτήσει διαφορετικές στάσεις σώματος και τελικά θα παραχθεί ένα βίντεο όπου θα τον δείχνει να χορεύει με τον ίδιο τρόπο που χορεύει το αντικείμενο στο βίντεο οδηγό. Μία σημαντική έλλειψη είναι η απουσία πληροφοριών για τα αντικείμενα μελέτης των δεδομένων μας. Στις μέρες μας, στα περισσότερα προβλήματα είναι δύσκολο να βρεθούν groudtruth δεδομένα που να συμβάλλουν

σημαντικά στην απόδοση της εκπαίδευσης των νευρωνικών δικτύων. Έχει δημιουργηθεί έτσι, η ανάγκη εύρεσης εναλλακτικών λύσεων και προσπάθειας εκπαίδευσης των νευρωνικών δίκτύων με μη επιβλέποντα τρόπο. Στην παρούσα εργασία, έχουν αναπτυχθεί συνελικτικά νευρωνικά δίκτυα και ένα γενετικό ανταγωνιστικό δίκτυο το οποίο εκπαιδεύεται για να δημιουργήσει νέες εικόνες που θα αποτελέσουν το animation βίντεο του στόχου μας.

Το περιεχόμενο της διατριβής συνεχίζεται με 4 κεφάλαια. Στο Κεφάλαιο 2 ακολουθεί μια γενική περιγραφή και θεωρητική ανάλυση των Νευρωνικών Δικτύων και των Βαθιών Νευρωνικών Δικτύων όπως τα Συνελικτικά Νευρωνικά Δίκτυα και τα Γενετικά Ανταγωνιστικά Δίκτυα. Στο Κεφάλαιο 3 παρατίθενται η σχετική έρευνα που έχει υλοποιηθεί μέσω δημοσιεύσεων και μελετών στον κλάδο της δημιουργίας βίντεο και της αναπαράστασης κίνησης εικόνων. Στο Κεφάλαιο 4 περιγράφεται αναλυτικά η μεθοδολογία που ακολουθήσαμε για την πραγματοποίηση της μελέτης μας. Τέλος, στο Κεφάλαιο 5 καταγράφονται ορισμένες λεπτομέρειες της υλοποίησης και αναλύονται τα αποτελέσματα των πειραμάτων τόσο ποσοτικά όσο και ποιοτικά.

Κεφαλαιο 2

Νευρωνικα Δικτυα

2.1 Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) αρχικά προτάθηκαν ως ένα μαθηματικό μοντέλο προσομοίωσης της πολύπλοκης λειτουργίας του ανθρώπινου εγκεφάλου. Η δομή του εγκεφάλου είναι τέτοια ώστε να επιτρέπει την παράλληλη επεξεργασία δεδομένων και τη δυνατότητα συνεχούς μάθησης μέσω της αλληλεπίδρασης με το περιβάλλον. Τα δύο αυτά βασικά χαρακτηριστικά συμβάλλουν στην ικανότητα, αφενός, να εκτελεί δύσκολα καθήκοντα, όπως ταχύτατη αναγνώριση προτύπων, ταξινόμηση κ.ά., αφετέρου, να εξελίσσεται συνεχώς, μαθαίνοντας από το περιβάλλον του κατά την αλληλεπίδρασή του με αυτό. Πρόκειται λοιπόν για ένα αφηρημένο αλγοριθμικό κατασκεύασμα το οποίο εμπίπτει στον τομέα της υπολογιστικής νοημοσύνης και στοχεύει στην επίλυση κάποιου υπολογιστικού προβλήματος.

Η δομή του τεχνητού νευρωνικού δικτύου μιμείται κατά το δυνατό εκείνη του βιολογικού νευρωνικού δικτύου, ώστε να εμφανίζει παρόμοιες ιδιότητες. Κατ' αναλογία επομένως με ένα δίκτυο νευρώνων εγκεφάλου, ένα τεχνητό δίκτυο αποτελείται από ένα σύνολο τεχνητών νευρώνων που αλληλεπιδρούν, συνδεόμενοι μεταξύ τους με τις λεγόμενες συνάψεις (synapses). Ο βαθμός αλληλεπίδρασης είναι διαφορετικός

για κάθε ζεύγος νευρώνων και καθορίζεται από τα λεγόμενα βάρη (weights). Συγκεκριμένα, καθώς το νευρωνικό δίκτυο αλληλεπιδρά με το περιβάλλον και μαθαίνει από αυτό, τα βάρη μεταβάλλονται συνεχώς, ενδυναμώνοντας ή αποδυναμώνοντας την ισχύ του κάθε δεσμού. Όλη η εμπειρική γνώση που αποκτά επομένως το νευρωνικό δίκτυο από το περιβάλλον κωδικοποιείται στα βάρη. Αυτά αποτελούν το χαρακτηριστικό εκείνο που δίνει στο δίκυο την ικανότητα για εξέλιξη και προσαρμογή στο περιβάλλον.

Αναλυτικά, η πιο κοινή δομή ενός νευρωνικού δικτύου αποτελείται από τρεις ομάδες ή στρώματα (layers): ένα στρώμα μονάδων «εισόδου» (input layer), που συνδέεται με ένα στρώμα «κρυφών» μονάδων (hidden layer), το οποίο συνδέεται με ένα επίπεδο μονάδων «εξόδου» (hidden layer) (Σχήμα 2.1)[1].

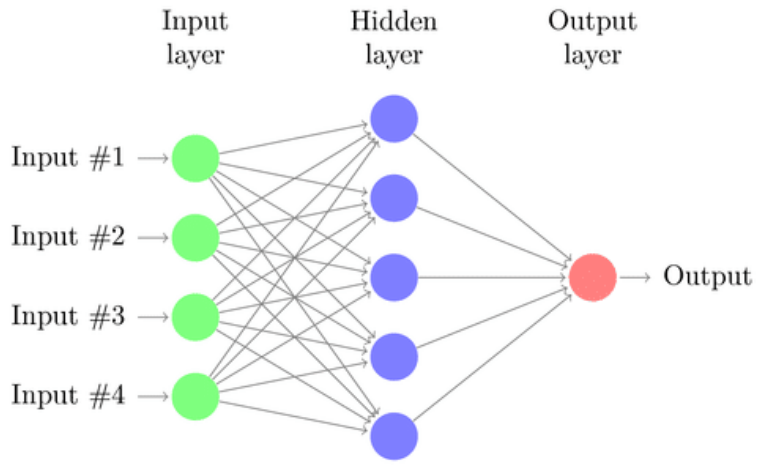
- **Επίπεδο Εισόδου**
  Η δραστηριότητα των μονάδων εισαγωγής αντιπροσωπεύει τις πρώτες πληροφορίες που τροφοδοτούνται στο δίκτυο.

- **Κρυφό Επίπεδο**
  Η δραστηριότητα κάθε κρυφής μονάδας καθορίζεται από τις μονάδες εισόδου και τα βάρη στις συνδέσεις μεταξύ της εισόδου και των κρυφών μονάδων. Τα κρυφά επίπεδα είναι συνήθως περισσότερα από ένα μιας και τα νευρωνικά δίκτυα έχουν να επιλύσουν αρκετά περίπλοκα προβλήματα. (Multilayer Neural Networks)

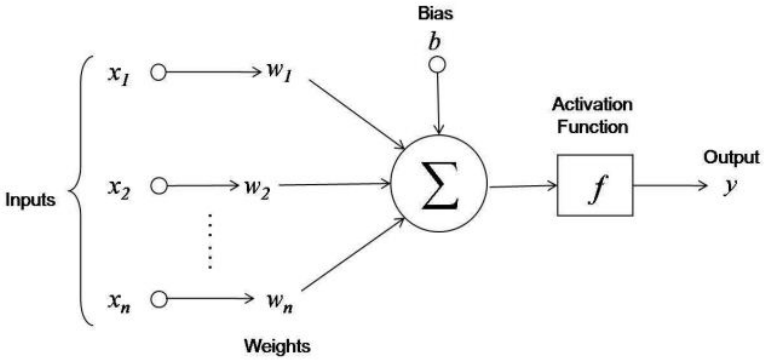
- **Επίπεδο Εξόδου**
  Η συμπεριφορά των μονάδων εξόδου εξαρτάται από τις κρυφές μονάδες και τα βάρη μεταξύ των κρυφών και των μονάδων εξόδου.

Κάθε νευρώνας ή κόμβος αποτελεί τη βασική μονάδα υπολογισμού σε ένα νευρωνικό Λαμβάνει είσοδο από ορισμένους άλλους κόμβους ή από εξωτερική πηγή και υπολογίζει μια έξοδο. Κάθε είσοδος έχει ένα σχετικό βάρος, το οποίο εκχωρείται βάσει της σχετικής σημασίας του για άλλες εισόδους. Στον κόμβο εφαρμόζεται μια συνάρτηση ενεργοποίσης f στο σταθμισμένο άθροισμα των εισόδων του, όπως φαίνεται στο Σχήμα 2.2[2]. Το δίκτυο λαμβάνει αριθμητικές εισόδους $x_1 ... x_n$ με βάρη $w_1 ... w_n$ αντίστοιχα,τα οποία σχετίζονται με αυτές τις εισόδους. Τα βάρη αποφασίζουν πόση επιρροή θα έχει κάθε είσοδος x_i στην έξοδο. Τα bias, τα οποία είναι

[1]Πηγή: https://jintensivecare.biomedcentral.com/articles/10.1186/s40560-019-0393-1
[2]Πηγή: https://qichaozhao.github.io/potato-lemon-2/



Σχήμα 2.1: Νευρωνικό Δίκτυο εμπρόσθιας διάδοσης



Σχήμα 2.2: Τεχνητός Νευρώνας

σταθερά, αποτελούν μια επιπλέον είσοδο στο επόμενο επίπεδο που έχουν πάντα την τιμή 1. Τα bias b δεν επηρεάζονται από το προηγούμενο επίπεδο (δεν έχουν εισερχόμενες συνδέσεις) αλλά έχουν εξερχόμενες συνδέσεις με τα δικά τους βάρη. Η τιμή 1 εγγυάται ότι ακόμη και όταν όλες οι είσοδοι είναι μηδενικές, θα εξακολουθεί να υπάρχει ενεργοποίηση στον νευρώνα.

Η συνάρτηση ενεργοποίησης f είναι μία μη γραμμική συνάρτηση και έχει ως σκοπό να εισαγάγει τη μη γραμμικότητα στην έξοδο ενός νευρώνα και να επιτρέπει την επίλυση μη γραμμικών προβλημάτων. Ένα νευρωνικό δίκτυο χωρίς τη συνάρτηση ενεργοποίησης είναι ουσιαστικά απλώς ένα μοντέλο γραμμικής παλινδρόμησης. Η συνάρτηση ενεργοποίησης για ενα πολυεπίπεδο νευρωνικό δίκτυο με L κρυφά επίπεδα ορίζεται ως εξής:

$$h^{(k)}(x) = g(b^{(k)} + W^{(k)} h^{(k-1)}(x)), \quad k \in [1, L] \subseteq \mathbb{N} \tag{2.1}$$

όπου $g(\cdot)$ η συνάρτηση ενεργοποίησης, $W^{(k)}$ τα βάρη, $b^{(k)}$ τα bias και $h^{(k-1)}(x)$ η συνάρτηση ενεργοποίησης του προηγούμενου επιπέδου $k-1$ με

$$h^{(k-1)}(x) = g(b^{(k-1)} + \sum_{i=1}^{n} w_i^{(k-1)} x_i) = f(x), \quad n \in \mathbb{N} \tag{2.2}$$

Για $k = L+1$, η ενεργοποίηση στο επίπεδο εξόδου είναι:

$$h^{(L+1)}(x) = o(b^{(L+1)} + W^{(L+1)} h^{(L)}(x)) \tag{2.3}$$

όπου $o(\cdot)$ η συνάρτηση ενεργοποίησης του τελευταίου επιπέδου εξόδου.

Μερικές συναρτήσεις ενεργοποίησης είναι:

- **Σιγμοειδής (Sigmoid)**
  Δέχεται ως είσοδο μια πραγματική τιμή και την μετατρέπει σε έυρος μεταξύ 0 και 1.
  $$\sigma(x) = \frac{1}{1+e^{-x}}, \quad x \in \mathbb{R} \tag{2.4}$$

- **Softmax**
  Σε προβλήματα ταξινόμησης, χρησιμοποιούμε γενικά την softmax ως συνάρτηση ενεργοποίησης στο επίπεδο εξόδου του νευρωνικού δικτύου για να διασφαλίσουμε ότι κάθε έξοδος αντιπροσωπεύει μία πιθανότητα και ότι συνολικά όλες οι έξοδοι αθροίζουν στην μονάδα. Με άλλα λόγια, η συνάρτηση Softmax παίρνει ένα διάνυσμα αυθαίρετων πραγματικών τιμών και τις μετατρέπει σε ένα διάνυσμα τιμών μεταξύ 0 και 1 που το συνολικό τους άθροισμα δίνει τη μονάδα.

- **Συνεφαπτομένη (Tanh)**
  Δέχεται ως είσοδο μια πραγματική τιμή και την μετατρέπει σε έυρος μεταξύ -1 και 1.
  $$g(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1, \quad x \in \mathbb{R} \tag{2.5}$$

- **Rectified Linear Unit (ReLU)**
  Δέχεται ως είσοδο μια πραγματική τιμή και αντικαθιστά τις αρνητικές τιμές με μηδέν. Θέτει δηλαδή, ως κατώτατο όριο το 0.
  $$g(x) = \max(0, x), \quad x \in \mathbb{R} \tag{2.6}$$

- **Leaky ReLU**

  Η Leaky ReLU είναι μια βελτιωμένη έκδοση της ReLU. Για οποιαδήποτε αρνητική τιμή του x, αυτή η συνάρτηση επιστρέφει μια πολύ μικρή αρνητική τιμή. Ως εκ τούτου, δεν εμφανίζονται πλέον νεκροί νευρώνες στην περιοχή.

$$g(x) = \max(0.01x, x), \quad x \in \mathbb{R} \tag{2.7}$$

Η επιλογή της κατάλληλης συνάρτησης ενεργοποίησης δεν είναι προφανής και πολλές φορές μπορεί να αποτελέσει δύσκολο πρόβλημα. Σε προβλήματα πολλαπλής δυαδικής ταξινόμησης χρησιμοποιείται συνήθως η σιγμοειδής συνάρτηση, ενώ σε προβλήματα ταξινόμησης πολλών κατηγοριών προτιμάται η softmax συνάρτηση.

Κάθε νευρωνικό δίκτυο έχει δύο μέρη, την εμπρόσθια διάδοση (forward propagation) κατά την οποία διαδίδεται η πληροφορία από το πρώτο επίπεδο εισόδου στα επόμενα επίπεδα και την οπισθοδιάδοση (back propagation). Για να εκπαιδευτεί ένα νευρωνικό δίκτυο και να βελτιώσει τις προβλέψεις του απαιτείται η σωστή ρύθμιση των παραμέτρων του, δηλαδή των βαρών w και των bias b. Η διαδικασία αυτή επιτυγχάνεται μέσω της ελαχιστοποίησης της συνάρτησης απώλειας που συχνά εκφράζεται ως η αρνητική λογαριθμική πιθανότητα :

$$E(\theta) = l(f(x^{(t)}; \theta), y^{(t)}) = -\sum log f(x)_y = -\sum_{i=1}^{N} log(p(y_i|x_i; \theta)), \tag{2.8}$$

όπου $\theta \equiv \{w^{(1)}, b^{(1)}...w^{(L+1)}, b^{(L+1)}\}$ οι παράμετροι, $(x^{(t)}, y^{(t)})$ κάθε δεδομένο εκπαίδευσης με τιμή εισόδου x και αντίστοιχη τιμή στόχου (target) y_i και $f(x)_y$ η πιθανότητα, δοθέντος μιας εισόδου x_i, να ανήκει στην κατηγορία y. Αυτή η πιθανότητα υπολογίζεται από το επίπεδο εξόδου ενός νευρωνικού δικτύου για ταξινόμηση. Στη συνέχεια, αφού υπολογιστεί το συνολικό σφάλμα στο επίπεδο εξόδου, προωθείται προς τα πίσω στο δίκτυο μέσω της οπισθοδιάδοσης back propagation για να υπολογιστούν οι κλίσεις (gradients). Για τον υπολογισμό τους χρησιμοποιούνται μέθοδοι βελτιστοποίησης όπως το Gradient Descent και μέσω αυτού ρυθμίζονται όλα τα βάρη του δικτύου ώστε να ελαχιστοποιηθεί η συνάρτηση απώλειας. Οι παράμετροι w και b αρχικοποιούνται με τυχαίες τιμές.

Ο αλγόριθμος του Gradient Descent χρησιμοποιείται για να βρεθεί η διεύθυνση και ο ρυθμός ανανέωσης των παραμέτρων θ:

$$\theta^{(\tau+1)} = \theta^{(\tau)} - \eta \nabla E(\theta^{(\tau)}), \tag{2.9}$$

όπου η ο ρυθμός εκμάθησης και τ η δεδομένη επανάληψη. Η συνάρτηση απώλειας καθορίζεται ως προς το σετ εκπαίδευσης και έτσι κάθε βήμα απαιτεί την προσπέλαση ολόκληρου του σετ εκπαίδευσης για να υπολογιστεί η κλίση ∇E. Μετά από κάθε ανανέωση, οι κλίσεις υπολογίζονται ξανά για το νέο διάνυσμα βάρους w και η διαδικασία επαναλαμβάνεται έως ότου τα αποτελέσματα να ταυτίζονται σχεδόν με τα επιθυμητές τιμές στόχου. Οι τεχνικές που χρησιμοποιούν με τη μία, όλο το σετ δεδομένων λέγονται μέθοδοι batch. Το mini-batch Gradient Descent είναι μια παραλλαγή του αλγορίθμου Gradient Descent που χωρίζει το σύνολο δεδομένων εκπαίδευσης σε μικρές παρτίδες που χρησιμοποιούνται για τον υπολογισμό του σφάλματος του μοντέλου και την ενημέρωση των παραμέτρων. Είναι η πιο συνηθισμένη εφαρμογή της κλίσης καθόδου που χρησιμοποιείται στον τομέα της μηχανικής μάθησης.

2.2 Βαθιά Νευρωνικά Δίκτυα

Με το πέρασμα των χρόνων και την τεχνολογική εξέλιξη, τα προβλήματα γίνονται ολοένα και πιο περίπλοκα. Η εκμάθηση των νευρωνικών δικτύων γίνεται βαθύτερη, και έτσι απαιτείται η κατασκευή μεγαλύτερων και πιο βαθιών νευρωνικών δικτύων. Τα Βαθιά Νευρωνικά Δίκτυα (Deep Neural Networks DNN) ακολουθούν το σχεδιασμό ενός απλού νευρωνικού δικτύου μόνο που περιλαμβάνουν περισσότερα από ένα ή δύο κρυφά επίπεδα. Ο αριθμός των κρυφών επιπέδων τους, καθορίζει και το βάθος των νευρωνικών δικτύων. Τα βαθιά νευρωνικά δίκτυα καταφέρνουν να επεξεργαστούν και να διεξάγουν περισσότερα χαρακτηριστικά των δεδομένων εισόδου και να επιλύσουν με αυτό το τρόπο διεργασίες με μεγαλύτερη ευκολία.

2.2.1 Συνελικτικά Νευρωνικά Δίκτυα

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks) είναι νευρωνικά δίκτυα που χρησιμοποιούνται κυρίως για την ταξινόμηση εικόνων, την ομαδοποίηση εικόνων κατά ομοιότητα και την αναγνώριση αντικειμένων. Για παράδειγμα, τα συνελικτικά νευρικά δίκτυα (CNN) χρησιμοποιούνται για την αναγνώριση προσώπων, ατόμων, οδών και πολλών άλλων πτυχών των οπτικών δεδομένων. Με άλλα λόγια πρόκειται για αλγορίθμους βαθιάς μηχανικής μάθησης που χρησιμοποιούνται στον τομέα της Υπολογιστικής Όρασης και την επεξεργασίας εικόνων.

Στο πλαίσιο ενός συνελικτικού νευρωνικού δικτύου, η συνέλιξη είναι μια γραμμική λειτουργία που περιλαμβάνει τον πολλαπλασιασμό ενός συνόλου βαρών με την είσοδο, σαν ένα παραδοσιακό νευρωνικό δίκτυο. Δεδομένου ότι η τεχνική σχεδιάστηκε για δισδιάστατη είσοδο, ο πολλαπλασιασμός πραγματοποιείται μεταξύ μιας σειράς δεδομένων εισόδου και μιας δισδιάστατης σειράς βαρών, που ονομάζεται φίλτρο ή πυρήνας (kernel). Η χρήση φίλτρου μικρότερου από την είσοδο είναι σκόπιμη καθώς επιτρέπει στο ίδιο το φίλτρο να πολλαπλασιαστεί με τη ακολουθία εισόδου πολλές φορές σε διαφορετικά σημεία της εισόδου. Συγκεκριμένα, το φίλτρο εφαρμόζεται συστηματικά σε κάθε επικαλυπτόμενο μέρος ή σε τμήμα μέγεθους φίλτρου των δεδομένων εισαγωγής, από αριστερά προς τα δεξιά, από πάνω προς τα κάτω. Εάν το φίλτρο έχει σχεδιαστεί για να ανιχνεύει έναν συγκεκριμένο τύπο χαρακτηριστικού στην είσοδο, τότε η εφαρμογή αυτού του φίλτρου συστηματικά σε ολόκληρη την εικόνα εισόδου,του επιτρέπει να ανακαλύψει αυτό το χαρακτηριστικό οπουδήποτε στην εικόνα. Η έξοδος από τον πολλαπλασιασμό του φίλτρου με τη δεδομένα εισόδου μία φορά είναι μία τιμή. Καθώς το φίλτρο εφαρμόζεται πολλές φορές στη ακολουθία εισόδου, το αποτέλεσμα είναι ένας δισδιάστατος πίνακας τιμών εξόδου που αντιπροσωπεύουν ένα φιλτράρισμα της εισόδου. Ως εκ τούτου, ο δισδιάστατος πίνακας εξόδου από αυτήν τη λειτουργία ονομάζεται ”χάρτης χαρακτηριστικών (feature map)”.

Η καινοτομία των συνελικτικών νευρωνικών δικτύων είναι η ικανότητα αυτόματης εκμάθησης ενός μεγάλου αριθμού φίλτρων σε παράλληλο χρόνο, ειδικά για ένα σύνολο δεδομένων εκπαίδευσης υπό τους περιορισμούς ενός συγκεκριμένου προβλήματος μοντελοποίησης, όπως η ταξινόμηση εικόνας. Το αποτέλεσμα περιλαμβάνει υψηλής τάξης χαρακτηριστικά που μπορούν να εντοπιστούν οπουδήποτε στις εικόνες εισόδου.

Αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου

Σε ένα συνελικτικό νευρωνικό δίκτυο, τα επίπεδα (layers) αποτελούνται από νευρώνες οργανωμένους σε τρεις διαστάσεις, τη χωρική διάσταση του δεδομένου εισόδου (ύψος και πλάτος) και το βάθος. Το βάθος δεν αναφέρεται στον συνολικό αριθμό επιπέδων εντός του δικτύου, αλλά στην τρίτη διάσταση του όγκου ενεργοποίησης. Σε αντίθεση με τα τυπικά νευρωνικά, οι νευρώνες σε οποιοδήποτε επίπεδο εντός του δικτύου συνδέονται μόνο σε μια μικρή περιοχή του επιπέδου που προηγείται.

Τα CNN αποτελούνται από τρεις βασικούς τύπους επιπέδων:

1. **Συνελικτικά Επίπεδα (Convolutional Layers)**

   Είναι το πρώτο επίπεδο που χρησιμοποιείται για την εξαγωγή των διαφόρων χαρακτηριστικών από τις εικόνες εισόδου. Σε αυτό το επίπεδο, η μαθηματική λειτουργία της συνέλιξης εκτελείται μεταξύ της εικόνας εισόδου και ενός φίλτρου συγκεκριμένου μεγέθους MxM. Σύροντας το φίλτρο πάνω από την εικόνα εισόδου, το αποτέλεσμα (εσωτερικό γινόμενο) λαμβάνεται μεταξύ του φίλτρου και των τμημάτων της εικόνας εισόδου σε σχέση με το μέγεθος του φίλτρου (MxM). Ο χάρτης χαρακτηριστικών που προκύπτει ως έξοδος μας δίνει πληροφορίες σχετικά με την εικόνα, όπως οι γωνίες και οι άκρες. Αργότερα, αυτός ο χάρτης τροφοδοτείται σε άλλα επίπεδα για να μάθει περισσότερα χαρακτηριστικά της εικόνας εισόδου. Μετά από το συνελικτικό επίπεδο ακολουθεί η διορθωμένη γραμμική μονάδα (ReLU), η οποία εφαρμόζει μια συνάρτηση ενεργοποίησης, όπως η σιγμοειδής, στην έξοδο ενεργοποίησης που παράγεται από το προηγούμενο επίπεδο. Ο σκοπός της εφαρμογής του ReLU είναι να αυξήσει τη μη γραμμικότητα στις εικόνες.

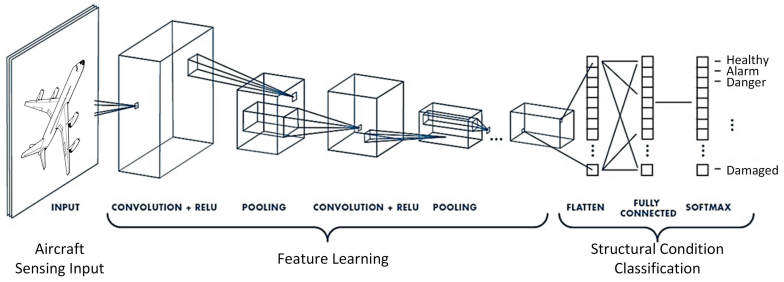
2. **Επίπεδα Ομαδοποίησης (Pooling Layers)**

   Στις περισσότερες περιπτώσεις, ένα συνελικτικό επίπεδο ακολουθείται από ένα επίπεδο ομαδοποίησης. Ο πρωταρχικός στόχος αυτού του επιπέδου είναι η μείωση του μεγέθους του χάρτη χαρακτηριστικών που είχε υποστεί συνέλιξη ώστε να μειωθεί το υπολογιστικό κόστος. Η διαδικασία αυτή πραγματοποιείται μειώνοντας τις συνδέσεις μεταξύ των επιπέδων και λειτουργεί ανεξάρτητα σε κάθε χάρτη χαρακτηριστικών. Ανάλογα με τη μέθοδο που χρησιμοποιείται, υπάρχουν διάφοροι τύποι λειτουργιών Pooling.

   - **Μέγιστη Ομαδοποίηση (Max Pooling)**
     Επιστρέφει τη μέγιστη τιμή από το τμήμα της εικόνας που καλύπτεται από το φίλτρο.
   - **Μέσο Ομαδοποίηση (Average Pooling)**
     Επιστρέφει τον μέσο όρο όλων των τιμών από το τμήμα της εικόνας που καλύπτεται από το φίλτρο.

   Το επίπεδο ομαδοποίησης pooling χρησιμεύει συνήθως ως γέφυρα μεταξύ του συνελικτικού επιπέδου και του πλήρως συνδεδεμένου επιπέδου.

3. **Πλήρως Συνδεδεμένα Επίπεδα (Fully Connected Layers)**



Σχήμα 2.3: Συνελικτικό Νευρωνικό Δίκτυο

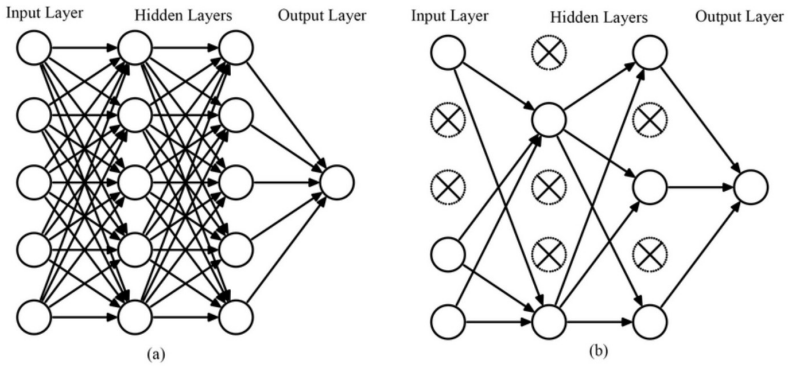
Το πλήρως συνδεδεμένο επίπεδο αποτελείται από τα βάρη και τα bias μαζί με τους νευρώνες και χρησιμοποιείται για τη σύνδεση των νευρώνων μεταξύ δύο διαφορετικών επιπέδων. Αυτά τα επίπεδα συνήθως τοποθετούνται πριν από το επίπεδο εξόδου και σχηματίζουν τα τελευταία επίπεδα μιας αρχιτεκτονικής CNN. Η εικόνα εισόδου από τα προηγούμενα επίπεδα 'απλώνεται' σε ένα διάνυσμα στήλης και έπειτα τροφοδοτείται στο πλήρως συνδεδεμένο επίπεδο. Το διάνυσμα αυτό υποβάλλεται στη συνέχεια σε λίγα ακόμη επίπεδα FC όπου συνήθως πραγματοποιούνται οι λειτουργίες των μαθηματικών συναρτήσεων. Σε αυτό το στάδιο, ξεκινά και η διαδικασία ταξινόμησης. (Σχήμα 2.3[3])

Εκτός από αυτά τα τρία επίπεδα, υπάρχουν δύο ακόμη σημαντικές παράμετροι που είναι το επίπεδο Αποκοπής Νευρώνων (Dropout Layer) και οι συναρτήσεις ενεργοποίησης:

1. **Επίπεδο Αποκοπής Νευρώνων (Dropout Layer)**
   Συνήθως, όταν όλα τα χαρακτηριστικά είναι συνδεδεμένα με το πλήρως συνδεδεμένο επίπεδο, μπορεί να προκληθεί υπερεκπαίδευση (overfitting) στο σύνολο δεδομένων εκπαίδευσης. Η υπερεκπαίδευση συμβαίνει όταν ένα συγκεκριμένο μοντέλο λειτουργεί τόσο καλά στα δεδομένα εκπαίδευσης προκαλώντας αρνητικό αντίκτυπο στην απόδοση του μοντέλου όταν χρησιμοποιείται σε νέα δεδομένα και σε δεδομένα που δεν έχουμε ξαναμελετήσει (unseen data). Για να ξεπεραστεί αυτό το πρόβλημα, χρησιμοποιείται ένα επίπεδο αποκοπής όπου μερικοί νευρώνες αποκόπτονται από το νευρωνικό δίκτυο κατά τη διάρκεια της εκπαίδευσης με αποτέλεσμα να μειώνεται το μέγεθος του μοντέλου. Η τεχνική αυτή φαίνεται στο Σχήμα 2.4 [4].

[3]Πηγή: https://www.mdpi.com/1424-8220/19/22/4933/xml

[4]Πηγή: https://www.researchgate.net/figure/Dropout-Strategy-a-A-standard-neural-network-b-Applying-dropout-to-the-neural_fig3_340700034



Σχήμα 2.4: Τεχνική Dropout (a) Τυπικό πλήρως συνδεδεμένο νευρωνικό δίκτυο, (b) Το ίδιο νευρωνικό δίκτυο εφαρμόζοντας dropout, διαγράφοντας δηλαδή τους σημαδεμένους κόμβους

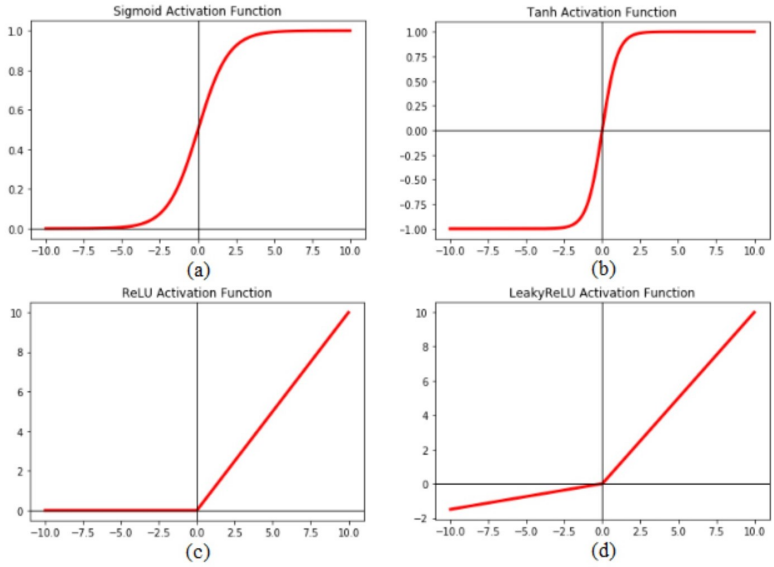
2. **Συναρτήσεις ενεργοποίησης**

   Χρησιμοποιούνται για να μάθουν και να προσεγγίσουν κάθε είδους περίπλοκη σχέση μεταξύ μεταβλητών του δικτύου. Αποφασίζουν δηλαδή, ποιες πληροφορίες του μοντέλου θα πρέπει να ενεργοποιηθούν κατά την εμπρόσθια κατεύθυνση και ποιες όχι, στο τέλος του δικτύου. Επίσης, προσθέτουν μη γραμμικότητα στο δίκτυο. Όπως ήδη αναφέραμε, υπάρχουν πολλές συναρτήσεις ενεργοποίησης, όπως οι λειτουργίες ReLU, Softmax, tanH και Sigmoid. Κάθε μία από αυτές τις συναρτήσεις έχει συγκεκριμένη χρήση. Για ένα μοντέλο δυαδικής ταξινόμησης CNN, οι λειτουργίες sigmoid και softmax προτιμώνται για μια ταξινόμηση πολλαπλών κατηγοριών, που συνήθως χρησιμοποιούμε το softmax. Ορισμένες από αυτές απεικονίζονται στο Σχήμα 2.5[5].

Δίκτυο U-Net

Τα τελευταία χρόνια, τα βαθιά συνελικτικά δίκτυα έχουν ξεπεράσει την αιχμή της επιστήμης σε πολλά προβλήματα οπτικής αναγνώρισης. Αν και τα συνελικτικά νευρωνικά δίκτυα χρησιμοποιούνται ήδη για μεγάλο χρονικό διάστημα, η επιτυχία τους ήταν περιορισμένη λόγω του ελλιπούς μεγέθους των διαθέσιμων σετ εκπαίδευσης και της ανεπαρκούς σχεδίασης των νευρωνικών δικτύων που χρησιμοποιούνταν για αυτό το σκοπό. Τα συνελικτικά νευρωνικά δίκτυα χρησιμοποιούνται συνήθως σε προβλήματα ταξινόμησης, όπου η έξοδος μιας εικόνας είναι μια ετικέτα μιας κατη-

[5]Πηγή: https://www.researchgate.net/figure/Plot-of-different-activation-functions-a-Sigmoid-activation-function-b-Tanh_fig4_339991922



Σχήμα 2.5: Γραφικές Παραστάσεις Συναρτήσεων Ενεργοποίησης

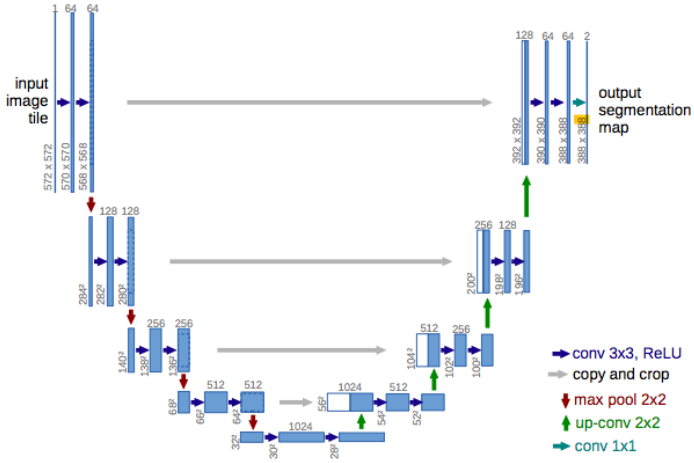
γορίας. Ωστόσο, σε πολλές διεργασίες που χειρίζονται οπτικά δεδομένα όπως εικόνες και βίντεο και ειδικότερα εικόνες βιοιατρικών δεδομένων, η επιθυμητή έξοδος πρέπει να περιλαμβάνει τον εντοπισμό (localization) του αντικειμένου στην εικόνα, μια διαδικασία που περιορίζει το αντικείμενο μελέτης σε μια συγκεκριμένη θέση στην εικόνα. Το ερευνητικό ενδιαφέρον, λοιπόν, επικεντρώθηκε όχι μόνο στην αναγνώριση του είδους του αντικειμένου και την ταξινόμηση του αλλά και στο ”που” βρίσκεται το αντικείμενο αυτό. Πρόκειται δηλαδή, για μια κατηγοριοποιημένη ετικέτα που υποτίθεται πως πρέπει να εκχωρηθεί σε κάθε εικονοστοιχείο της εικόνας.

Στην περίπτωση, όμως, των βιοιατρικών δεδομένων, ήταν σχεδόν αδύνατο να συλλεχθούν χιλιάδες εικόνες ως σετ δεδομένων εκπαίδευσης για τα συνελικτικά δίκτυα. Ως εκ τούτου, ο Ciresan και η ομάδα του [1] για να προβλέψουν την κατηγορία κάθε εικονοστοιχείου της εικόνας, εκπαίδευσαν ένα δίκτυο με μια τεχνική συρόμενου παραθύρου, παρέχοντας ως είσοδο στο δίκτυο μια τοπική περιοχή (patch) γύρω από αυτό το εικονοστοιχείο. Με τον τρόπο αυτό, κάθε τοπική περιοχή λειτουργούσε ως δεδομένο εκπαίδευσης και έτσι ο αριθμός τους ήταν πολύ μεγαλύτερος από τον αριθμό των εικόνων εκπάίδευσης, καθώς σε κάθε εικόνα μπορούσαν να διεξαχθούν δεκάδες τοπικά ”τμήματα”. Προφανώς, η στρατηγική αυτή είχε μειονεκτήματα. Αρχικά, η εκπαίδευση ήταν αρκετά αργή επειδή το δίκτυο έπρεπε να εκτελείται ξεχωριστά για κάθε τοπικό τμήμα της εικόνας και πολλές φορές τα τμήματα αυτά επικαλύπτονταν προκαλώντας περιττούς υπολογισμούς για το δίκτυο. Επιπλέον, έπρεπε να βρεθεί ισορροπία μεταξύ ακρίβειας του εντοπισμού και

των τοπικών τμημάτων. Μεγαλύτερα τμήματα απαιτούν περισσότερα pooling επίπεδα που μειώνουν την ακρίβεια εντοπισμού, ενώ μικρότερα τμήματα επιτρέπουν στο δίκτυο να επεξεργάζεται λιγότερο περιεχόμενο και να χάνει ίσως σημαντικές πληροφορίες. Για να αντιμετωπιστούν αυτά τα θέματα, εισήγαγαν το U-Net, ένα πλήρως συνελικτικό δίκτυο, το οποίο λειτουργεί με πολύ λίγες εικόνες εκπαίδευσης και αποδίδει πιο ακριβείς τμηματοποιήσεις από τις αρχιτεκτονικές που είχαν παρουσιαστεί μέχρι τότε. Αρχικά, το U-Net αναπτύχθηκε για την επεξεργασία και την τμηματοποίηση βιοϊατρικών εικόνων.

Η κύρια ιδέα για την σχεδίαση του δικτύου είναι ο εφοδιασμός ενός συνηθισμένου συμβαλλόμενου δικτύου με διαδοχικά επίπεδα, όπου οι τελεστές ομαδοποίησης (pooling) αντικαθίστανται από τελεστές επέκτασης (upsampling). Ως εκ τούτου, αυτά τα επίπεδα αυξάνουν την ανάλυση της εξόδου. Για την τοπικό εντοπισμό, χαρακτηριστικά στοιχεία υψηλής ανάλυσης από τη διαδρομή σύμβασης συνδυάζονται με την έξοδο που έχει επεκταθεί μέσω της επέκτασης. Έπειτα, διαδοχικά συνελικτικά στρώματα μπορούν να το εκπαιδεύσουν ώστε να συγκεντρώνει ένα πιο ακριβές αποτέλεσμα βάσει αυτών των πληροφοριών. Μια σημαντική τροποποίηση στην αρχιτεκτονική αυτή είναι πως στο κομμάτι της επέκτασης υπάρχει μεγάλος αριθμός καναλιών χαρακτηριστικών, που επιτρέπουν στο δίκτυο να διαδόσει σχετικές πληροφορίες σε επίπεδα υψηλότερης ανάλυσης. Ως συνέπεια,η επεκτατική διαδρομή είναι λίγο πολύ συμμετρική με τη διαδρομή συστολής και αποδίδει μια αρχιτεκτονική σε σχήμα U. Το δίκτυο δεν έχει πλήρως συνδεδεμένα επίπεδα και χρησιμοποιεί μόνο το έγκυρο μέρος κάθε συνέλιξης, δηλαδή τον χάρτη τμηματοποίησης που περιέχει μόνο τα εικονοστοιχεία, για τα οποία το πλήρες περιεχόμενο είναι διαθέσιμο στην εικόνα εισαγωγής. Αυτή η στρατηγική επιτρέπει την απρόσκοπτη τμηματοποίηση αυθαίρετων μεγάλων διαστάσεων εικόνων από μια στρατηγική επικάλυψης πλακιδίων. Για να προβλεφθούν τα εικονοστοιχεία στην περιοχή των συνόρων της εικόνας, τα τμήματα που λείπουν συμπληρώνονται από τον κατοπτρισμό της εικόνας εισόδου. Αυτή η στρατηγική πλακιδίων είναι σημαντική για την εφαρμογή του δικτύου σε μεγάλες εικόνες, δεδομένου ότι διαφορετικά, η ανάλυση θα περιοριζόταν από τη μνήμη GPU.

Η αρχιτεκτονική του δικτύου U-Net περιέχει δύο μονοπάτια. Το πρώτο μονοπάτι στα αριστερά περιλαμβάνει την διαδρομή συστολής (που ονομάζεται επίσης κωδικοποιητής) η οποία χρησιμοποιείται για την καταγραφή του περιεχομένου στην εικόνα. Ο κωδικοποιητής είναι απλώς μια τυπική στοίβα συνελικτικών, ReLU και



Σχήμα 2.6: Αρχιτεκτονική Δικτύου U-Net

Max Pooling επιπέδων. Το δεύτερο μονοπάτι στα δεξία περιλαμβάνει τη συμμετρική διαδρομή επέκτασης (ονομάζεται επίσης και ο αποκωδικοποιητής) η οποία χρησιμοποιείται για τον ακριβή εντοπισμό των αντικειμένων χρησιμοποιώντας αντίστροφες συνελίξεις. Η αρχιτεκτονική του δικτύου απεικονίζεται στο Σχήμα 2.6[6]. Αποτελείται από την επαναλαμβανόμενη εφαρμογή δύο 3X3 συνελίξεων,όπου η κάθε μία ακολουθείται από μια διορθωμένη γραμμική μονάδα ReLU και μια 2X2 μέγιστη pooling λειτουργία με άλμα 2 για συρρίκνωση (downsampling). Η συρρίκνωση μειώνει τις διαστάσεις των εικόνων και με αυτό τον τρόπο επιτυχάνεται πιο γρήγορη επεξεργασία των δεδομένων. Επιπλέον, τα δεδομένα καταναλώνουν λιγότερο χώρο στη μνήμη και γίνονται πιο διαχειρίσιμα για την GPU. Σε κάθε βήμα συρρίκνωσης διπλασιάζεται ο αριθμός των καναλιών των χαρακτηριστικών. Κάθε βήμα στο μονοπάτι επέκτασης αποτελείται από μια επέκταση του χάρτη χαρακτηριστικών. Η επέκταση (upsampling) είναι η αντίστροφη διαδικασία της συρρίκνωσης και αυξάνει τις διαστάσεις των εικόνων. Προσθέτει δηλαδή γραμμές και στήλες στην εικόνα συμπληρώνοντας τα κενά. Στη συνέχεια ακολουθεί μια 2X2 συνέλιξη που διαιρεί στο μισό τον αριθμό των καναλιών των χαρακτηριστικών. Έπειτα, ακολουθεί μια συνένωση με τον αντίστοιχο κομμένο χάρτη χαρακτηριστικών από το μονοπάτι της συρρίκνωσης και δύο 3X3 συνελίξεις, η κάθε μία ακολουθούμενη από ένα επίπεδο ReLU. Η περικοπή στον χάρτη χαρακτηριστικών είναι απαραίτητη λόγω της απώλειας εικονοστοιχείων στο περίγραμμα της εικόνας σε κάθε συνέλιξη. Στο τελικό επίπεδο χρησιμοποιείται μια 1X1 συνέλιξη για να αντιστοιχήσει κάθε διάνυσμα χα-

[6]Πηγή: https://www.programmersought.com/article/56616823825/

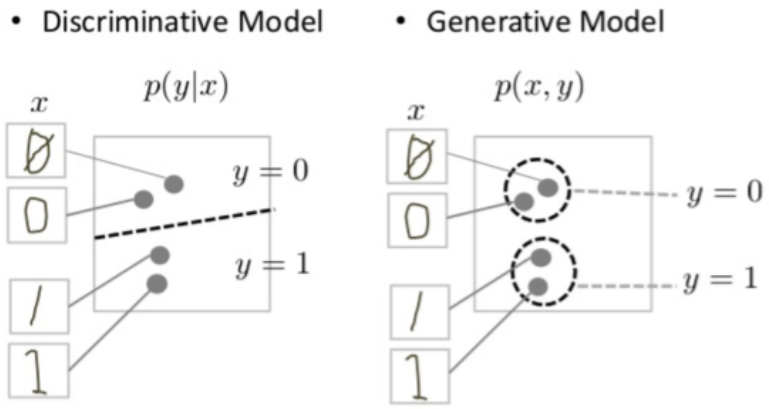
ρακτηριστικών στοιχείων 64 συνιστωσών στον επιθυμητό αριθμό τάξεων. Συνολικά το δίκτυο έχει 23 συνελικτικά επίπεδα.

2.2.2 Γενετικά Ανταγωνιστικά Δίκτυα

Μία από τις σημαντικότερες προκλήσεις στην στατιστική επεξεργασία σημάτων και στη μηχανική μάθηση είναι ο τρόπος απόκτησης ενός γενετικού μοντέλου που μπορεί να παράγει δείγματα από κατανομές δεδομένων μεγάλης κλίμακας, όπως εικόνες και ηχητικά σήματα. Τα γενετικά μοντέλα (generative models) είναι μοντέλα της μηχανικής μάθησης που επικεντρώνονται στην κατανομή μεμονωμένων τάξεων σε ένα σύνολο δεδομένων και οι αλγόριθμοι μάθησης που χρησιμοποιούν, τείνουν να μοντελοποιούν τα υποκείμενα μοτίβα ή την κατανομή των δεδομένων. Δοθέντος ενός συνόλου δεδομένων X και των αντίστοιχων ετικέτων τους Y, τα γενετικά μοντέλα υπολογίζουν την από κοινού πιθανότητα $p(X, Y)$ ή απλώς την $p(X)$ αν δεν υπάρχουν ετικέτες. Παράγουν δεδομένα με σεβασμό σε κάποια συγκεκριμένη συνάρτηση πιθανότητας. Για παράδειγμα, ένα γενετικό μοντέλο μπορεί να προβλέψει την επόμενη λέξη που ακολουθεί σε μια πρόταση, επειδή έχει αναθέσει στην ακολουθία των λέξεων κάποια κατανομή πιθανότητας. Αν και αυτά τα μοντέλα μπορούν να δημιουργήσουν νέα δεδομένα, έχουν ένα μεγάλο μειονέκτημα. Η παρουσία ακραίων τιμών μπορεί να τα επηρεάσει σε σημαντικό βαθμό. Από την άλλη πλευρά, τα μοντέλα διάκρισης (discriminative models) αποτελούν μια κατηγορία μοντέλων που χρησιμοποιούνται για ταξινόμηση. Υπολογίζουν την δεσμευμένη πιθανότητα

$$p(Y|X) = \frac{p(Y, X)}{p(X)}, \tag{2.10}$$

δηλαδή δοθέντος X, υπολογίζουν την πιθανότητα να ανήκει το δεδομένο στην κατηγορία Υ. Αυτά διακρίνουν το όρια μεταξύ των τάξεων / ετικετών σε ένα σύνολο δεδομένων. Σε αντίθεση με τα γενετικά μοντέλα, ο στόχος τους είναι να ανακαλύψουν το όριο απόφασης που χωρίζει τη μια τάξη από την άλλη. Φυσικά, κανένα είδος μοντέλου δεν θα υπολογίσει αριθμούς ως πιθανότητες, θα μοντελοποιήσει όμως την κατανομή των δεδομένων που την μιμούνται. Έτσι, ενώ ένα γενετικό μοντέλο τείνει να μοντελοποιεί την από κοινού συνάρτηση πιθανότητας των δεδομένων και είναι ικανό να δημιουργεί νέα δεδομένα χρησιμοποιώντας εκτιμήσεις πιθανότητας, τα μοντέλα διάκρισης χωρίζουν τις κατηγορίες τους μοντελοποιώντας την δεσμευμένη πιθανότητα και δεν κάνουν υποθέσεις σχετικά με τα σημεία δεδομένων. Επίσης τα μοντέλα διάκρισης, δεν μπορούν να δημιουργήσουν νέα δεδομένα. Στο σχήμα 2.7[7]



Σχήμα 2.7: Μοντέλο Διάκρισης και Γενετικό Μοντέλο

περιγράφεται το γενετικό μοντέλο και το μοντέλο διάκρισης για δύο χειρόγραφα αριθμητικά ψηφία, το 0 και 1. Το μοντέλο διάκρισης προσπαθεί να εντοπίσει τη διαφορά μεταξύ χειρόγραφων 0 και 1, σχεδιάζοντας μια γραμμή ως όριο στο χώρο δεδομένων. Εάν βρει τη σωστή γραμμή περιθωρίου, μπορεί να διακρίνει τα 0 από τα 1 χωρίς να χρειάζεται να μοντελοποιήσει ακριβώς το πού βρίσκονται τα σημεία στο χώρο δεδομένων και στις δύο πλευρές της γραμμής. Αντίθετα, το γενετικό μοντέλο προσπαθεί να δημιουργήσει τα ψηφία 1 και 0 όσο πιο ρεαλιστικά γίνεται, ώστε να πλησιάζουν τα αντίστοιχα πραγματικά ψηφία στο χώρο δεδομένων. Πρέπει να μοντελοποιήσει την κατανομή σε όλο το χώρο δεδομένων. Τα μοντέλα διάκρισης έχουν το πλεονέκτημα ότι είναι πιο ισχυρά έναντι των ακραίων τιμών, σε αντίθεση με τα γενετικά μοντέλα. Ωστόσο, ένα σημαντικό μειονέκτημα είναι το πρόβλημα εσφαλμένης ταξινόμησης, δηλαδή, η λανθασμένη κατηγοριοποίηση ενός δεδομένου. Μια άλλη βασική διαφορά μεταξύ αυτών των δύο τύπων μοντέλων είναι ότι ενώ το γενετικό μοντέλο επικεντρώνεται στην εξήγηση του τρόπου με τον οποίο δημιουργήθηκαν τα δεδομένα, το μοντέλο διάκρισης επικεντρώνεται στην πρόβλεψη ετικετών των δεδομένων.

Ένας αποτελεσματικός τρόπος εκπαίδευσης τέτοιων μοντέλων προσφέρεται μέσω των ανταγωνιστικών γενετικών δικτύων. Τα ανταγωνιστικά γενετικά δίκτυα (Generative Adversarial Networks - GANs) προτάθηκαν από τον Ian Goodfellow και την ομάδα του το 2014 [2] και αποτελούν μια κατηγορία αλγορίθμων ημιελεγχόμενης μάθησης και μάθησης χωρίς επίβλεψη που χρησιμοποιούνται για την παραγωγή εικόνων και βίντεο, τον σχηματισμό συμβόλων, την δημιουργία ηχητικών σημάτων και

[7]Πηγή: https://developers.google.com/machine-learning/gan/generative

ομιλιών κ.ο.κ. Κατά αυτή τη προσέγγιση, δύο νευρωνικά δίκτυα ανταγωνίζονται μεταξύ τους σε ένα παιχνίδι (παιχνίδι μηδενικού αθροίσματος κατά τη θεωρία παιγνίων). Το ένα δίκτυο είναι η γεννήτρια δεδομένων (generator), το γενετικό μοντέλο το οποίο έχει ως σκοπό να παράξει δεδομένα όσο πιο αληθοφανή γίνεται ενώ το άλλο δίκτυο είναι ο διαχωριστής (discriminator) ο οποίος προσπαθεί να ξεχωρίσει τα παραγόμενα δεδομένα της γεννήτριας από τα αληθινά δεδομένα (groudtruth). Τα δύο αντίπαλα δίκτυα παρομοιάζονται με μια γεννήτρια πλαστών χαρτονομισμάτων και ένα μηχάνημα εντοπισμού των πλαστών που λειτουργεί ως ο διαχωριστής των ψεύτικων χαρτονομισμάτων από τα γνήσια. Στη συνέχεια, τα δύο δίκτυα μπορούν να εκπαιδευτούν ταυτόχρονα έχοντας αντίθετους στόχους:

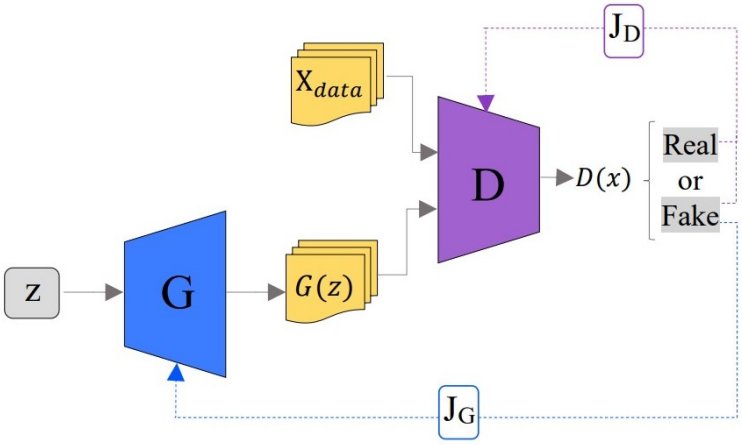
- **Γεννήτρια**

  Ο στόχος της γεννήτριας είναι να ξεγελάσει τον διαχωριστή, και εκπαιδεύεται με τέτοιο τρόπο ώστε να ελαχιστοποιήσει το τελικό σφάλμα ταξινόμησης (μεταξύ πραγματικών και παραγόμενων δεδομένων). Στη γεννήτρια εισάγεται ένα διάνυσμα θορύβου z με μία εκ των προτέρων πιθανότητα $p_z(z)$ η οποία στη συνέχεια αντιπροσωπεύει μια αντιστοίχιση στο χώρο δεδομένων ως $G(z;\theta_g)$, όπου G παραγωγίσιμη συνάρτηση που αντιπροσωπεύεται από ένα πολυεπίπεδο perceptron με παραμέτρους θ_g. Η γεννήτρια επιστρέφει μια έξοδο $x_g = G(z)$, η οποία μετά το πέρας της εκπαίδευσης θα πρέπει να ακολουθεί την επιθυμητή κατανομή πιθανότητας.

- **Διαχωριστής**

  Ο στόχος του διαχωριστή είναι να ανιχνεύσει τα ψεύτικα δεδομένα που παράγει η γεννήτρια. Ο διαχωριστής ορίζεται ως ένα δεύτερο πολυεπίπεδο perceptron $D(x;\theta_d)$. Ο διαχωριστής δέχεται ως είσοδο ένα δεδομένο x που είναι είτε γνήσιο x_t με πιθανότητα p_t είτε ένα ψεύτικο x_g που έχει παραχθεί από την γεννήτρια με πιθανότητα p_g. Η έξοδος του διαχωριστή είναι η πιθανότητα $D(x)$ που φανερώνει την πιθανότητα του x να είναι αληθινό δεδομένο.

Η minimax απώλεια του γενετικού ανταγωνιστικού δίκτυου θεωρείται ως στρατηγική βελτιστοποίησης σε παιχνίδια δύο παικτών, όπου κάθε παίκτης μειώνει τις απώλειές του ή αυξάνει το κόστος του άλλου παίκτη. Στο GAN, η γεννήτρια και ο διαχωριστής αντιπροσωπεύουν τους δύο παίκτες, που με τη σειρά τους ενημερώνουν το βάρος του δικτύου τους. Το minimax λοιπόν, αναφέρεται στην ελαχιστοποίηση της απώλειας στη γεννήτρια και στη μεγιστοποίηση της απώλειας στον διαχωριστή



Σχήμα 2.8: Αρχιτεκτονική Γενετικού Ανταγωνιστικού Δικτύου

με συνάρτηση τιμής $V(G, D)$:

$$\min_{G} \max_{D} V(G, D) = \mathbb{E}_{x \sim p_t}[log(D(x)] + \mathbb{E}_{z \sim p_z}[log(1 - D(G(z))] \tag{2.11}$$

Με άλλα λόγια, ο διαχωριστής προσπαθεί να μεγιστοποιήσει την πιθανότητα εκχώρησης σωστών ετικετών στα δεδομένα. Αντίθετα, η γεννήτρια επιδιώκει να δημιουργήσει μια σειρά δειγμάτων κοντά στην πραγματική κατανομή δεδομένων για να ελαχιστοποιήσει την εγκάρσια εντροπία cross-entropy. Στην πραγματικότητα, όμως, κατά τα πρώτα στάδια της εκπαίδευσης όπου η γεννήτρια είναι ανεπαρκής, ο διαχωριστής μπορεί να απορρίψει έμπιστα δεδομένα επειδή είναι σαφώς διαφορετικά από τα δεδομένα εκπαίδευσης. Έτσι, αντί να εκπαιδεύουμε την γεννήτρια να ελαχιστοποιήσει την ποσότητα $log(1 - D(G(z))$, την εκπαιδεύουμε να μεγιστοποιήσει την $log(D(G(z))$. Με αυτό τον τρόπο παρέχονται πολύ πιο ισχυρές κλίσεις (gradients) στην αρχική φάση της εκπαίδευσης.

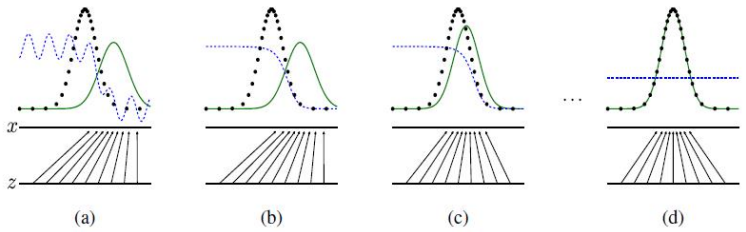
Η αρχιτεκτονική του GAN απεικονίζεται στο Σχήμα 2.8[8]. Τα X_{data} είναι τα πραγματικά δείγματα στο σύνολο δεδομένων εκπαίδευσης και τα $G(Z)$ τα ψεύτικα δεδομένα που παράγονται από τη γεννήτρια G μέσω ενός διανύσματος θορύβου σταθερού μήκους z (τυχαίο διάνυσμα που προέρχεται από ομοιόμορφη ή κανονική Gaussian κατανομή). Για να δημιουργηθεί μια νέα εικόνα από την γεννήτρια δεν απαιτείται να εισαχθεί μια αρχική εικόνα, παρά μόνο ένα διάνυσμα τυχαίων τιμών. Ο διαχωριστής D υπολογίζει την πιθανότητα να είναι γνήσια ή πλαστά τα δεδομένα που λαμβάνει ως είσοδο. Μετά την εκπαίδευση, τα σημεία αυτού του πολυδιάστατου διανύσματος ταιριάζουν με τα σημεία στον τομέα του προβλήματος, με

[8]Πηγή: https://www.semanticscholar.org/paper/Generative-Adversarial-Networks-(GANs)%3A-An-Overview-Salehi-Chalechale/fb1456a12630d97f25b80809865847b70c116ebb

αποτέλεσμα να επιτυγχάνεται μια συμπιεσμένη αναπαράσταση της κατανομής που ακολουθούν τα δεδομένα. Αυτός ο διανυσματικός χώρος είναι γνωστός ως λανθάνων χώρος ή διανυσματικός χώρος που αποτελείται από λανθάνουσες μεταβλητές. Οι λανθάνουσες μεταβλητές περιλαμβάνουν σημαντικές, αλλά όχι άμεσα παρατηρήσιμες μεταβλητές στον τομέα. Τα μοντέλα μηχανικής μάθησης μπορούν να μάθουν τον στατιστικό λανθάνοντα χώρο των εικόνων και στη συνέχεια να δημιουργήσουν μια σειρά από νέες εικόνες με προδιαγραφές παρόμοιες με αυτές των πραγματικών δειγμάτων αυτού του χώρου. Έπειτα, ο διαχωριστής ενεργεί ως δυαδικός ταξινομητής και διαφοροποιεί τα πλαστά δεδομένα $G(z)$ από τα πραγματικά X_{data}

Εάν η είσοδος προέρχεται από τα X_{data}, ο διαχωριστής την ταξινομεί ως πραγματικό δεδομένο και επιστρέφει μια αριθμητική τιμή κοντά στο 1. Διαφορετικά, εάν η είσοδος προέρχεται από δεδομένα που παράγονται από τη γεννήτρια, ο διαχωριστής την χαρακτηρίζει ως ψεύτικο δεδομένο και επιστρέφει μια αριθμητική τιμή κοντά στο 0. Επομένως, τα δύο δίκτυα χρειάζονται δύο συναρτήσεις απώλειας για να ενημερωθούν και να εκπαιδευτούν. Ο διαχωριστής έχει την συνάρτηση απώλειας J_D και η γεννήτρια τη συνάρτηση απώλειας J_G. Κατά τη διάρκεια εκπαίδευσης του διαχωριστή, ο διαχωριστής ταξινομεί τόσο τα πραγματικά όσο και τα πλαστά δεδομένα από τη γεννήτρια. Στη συνέχεια, η απώλεια του διαχωριστή, τον τιμωρεί όταν ταξινομεί ένα πραγματικό δεδομένο ως ψεύτικο ή όταν ταξινομεί ένα πλαστό δεδομένο ως αληθινό. Έπειτα, ο διαχωριστής ενημερώνει τα βάρη του μέσω της οπισθοδιάδοσης (backpropagation) από την απώλεια του προς τα πίσω στο δίκτυο του διαχωριστή. Το μέρος της γεννήτριας ενός GAN μαθαίνει να δημιουργεί πλαστά δεδομένα ενσωματώνοντας σχόλια από το διακριτικό. Μαθαίνει να κάνει τον διαχωριστή να ταξινομεί την παραγωγή του ως πραγματικό. Η εκπαίδευση της γεννήτριας απαιτεί αυστηρότερη ενσωμάτωση μεταξύ της γεννήτριας και του διαχωριστή, αφού η γεννήτρια λαμβάνει πληροφορίες από τον διαχωριστή ως ανατροφοδότηση.

Για να εκπαιδεύσουμε ένα νευρωνικό δίκτυο, αλλάζουμε τα βάρη του δικτύου για να μειώσουμε το σφάλμα ή την απώλεια της εξόδου του. Στο GAN, ωστόσο, η γεννήτρια δεν συνδέεται άμεσα με την απώλεια που προσπαθούμε να επηρεάσουμε. Η γεννήτρια τροφοδοτεί το δίκτυο του διαχωριστή και αυτός παράγει την έξοδο που προσπαθούμε να μεταποιήσουμε. Η απώλεια της γεννήτριας τιμωρεί τη γεννήτρια για την παραγωγή δείγματος που ο διαχωριστής χαρακτηρίζει ως ψεύτικο. Η οπισθοδιάδοση προσαρμόζει κάθε βάρος στη σωστή κατεύθυνση υπολογίζοντας τις κλίσεις (gradients), την επίδραση του βάρους στην έξοδο, το πώς δηλαδή θα αλ-



Σχήμα 2.9: Βήματα Εκπαίδευσης ενός Γενετικούννητικού Ανταγωνιστικού Δικτύου

λάξει η έξοδος εάν αλλάξει το βάρος. Όμως, η επίδραση του βάρους της γεννήτριας εξαρτάται από την επίδραση του βάρους του διαχωριστή που τον τροφοδοτεί. Έτσι, η οπισθοδιάδοση ξεκινά από την έξοδο και ρέει προς τα πίσω μέσω του διαχωριστή στη γεννήτρια. Ταυτόχρονα, δεν θέλουμε να αλλάξει ο διαχωριστής κατά τη διάρκεια της εκπαίδευσης της γεννήτριας.

Η εκπαίδευση ενός γενετικού ανταγωνιστικού δικτύου είναι πολύ δύσκολη υπόθεση, καθώς ο αλγόριθμος πρέπει να αντισταθμίσει δύο διαφορετικά δίκτυα εκπαίδευσης. Κατά την φάση της εκπαίδευσης του διαχωριστή, διατηρούμε τη γεννήτρια σταθερή. Καθώς ο διαχωριστής προσπαθεί να καταλάβει πώς να διακρίνει τα πραγματικά δεδομένα από τα ψεύτικα, πρέπει να μάθει πώς να αναγνωρίζει τα ελαττώματα της γεννήτριας. Αυτό είναι ένα διαφορετικό πρόβλημα για μια καλά εκπαιδευμένη γεννήτρια από ότι για μια μη εκπαιδευμένη γεννήτρια που παράγει τυχαία έξοδο. Παρομοίως, διατηρούμε τον διαχωριστή σταθερό κατά τη φάση εκπαίδευσης της γεννήτριας. Διαφορετικά, η γεννήτρια θα προσπαθούσε να χτυπήσει έναν κινούμενο στόχο, τον οποίο δεν θα τον πετύχαινε σχεδόν ποτέ. Καθώς η γεννήτρια βελτιώνεται με την εκπαίδευση, η απόδοση του διαχωριστή επιδεινώνεται επειδή ο ίδιος δεν μπορεί εύκολα να διακρίνει τη διαφορά μεταξύ πραγματικού και ψεύτικου. Εάν η γεννήτρια επιτύχει τέλεια αποτελέσματα που είναι σχεδόν πανομοιότυπα με τα πραγματικά, τότε ο διαχωριστής έχει ακρίβεια 50%. Η πιθανότητα δηλαδή να είναι το δεδομένο είτε αληθινό είτε ψεύτικο είναι $\frac{1}{2}$. Αυτή η εξέλιξη θέτει ένα πρόβλημα για τη σύγκλιση στο σύνολο του GAN: Η ανατροφοδότηση του διαχωριστή γίνεται λιγότερο σημαντική με την πάροδο του χρόνου. Εάν το GAN συνεχίσει να εκπαιδεύεται πέρα από το σημείο που ο διαχωριστής δίνει μια εντελώς τυχαία ανατροφοδότηση, τότε η γεννήτρια αρχίζει να εκπαιδεύεται με ανεπιθύμητη ανατροφοδότηση και η ποιότητα της μπορεί να υποβαθμιστεί. Για ένα GAN, η σύγκλιση είναι συχνά μια ασταθής, παρά μόνιμη, κατάσταση.

Το Σχήμα 2.9[9] απεικονίζει διάφορα βήματα της ταυτόχρονης εκπαίδευσης της γεννήτριας και του διαχωριστή σε ένα GAN ως παράδειγμα. Η κάτω οριζόντια γραμμή υποδεικνύει τον τομέα από τον οποίο γίνεται δειγματοληψία των δεδομένων z, που σε αυτήν την περίπτωση προέρχονται από την ομοιόμορφη κατανομή. Η πάνω οριζόντια γραμμή είναι μέρος του τομέα των πραγματικών δεδομένων x. Τα βέλη προς τα πάνω δείχνουν πώς η αντιστοίχιση $x = G(z)$ επιβάλλει την μη ομοιόμορφη κατανομή p_g στα μετασχηματισμένα δείγματα. Στη φάση (a), το GAN εκπαιδεύεται ενημερώνοντας ταυτόχρονα την κατανομή του διαχωριστή (μπλε, διακεκομμένη γραμμή) έτσι ώστε να διακρίνει μεταξύ δειγμάτων από την πραγματική κατανομή δεδομένων (μαύρη, διακεκομμένη γραμμή) και την παραγόμενη διανομή δεδομένων από την γεννήτρια (πράσινη,συμπαγής γραμμή). Στη φάση (b), ο διαχωριστής εκπαιδεύτηκε να κάνει διακρίσεις μεταξύ πραγματικών και ψεύτικων δεδομένων, και εκτελεί εύκολα το καθήκον του. Στη φάση (c), η διαδικασία εκπαίδευσης του διαχωριστή διακόπτεται και εκπαιδεύεται μόνο η γεννήτρια ώστε να φέρει την ψεύτικη κατανομή δεδομένων πιο κοντά στην πραγματική κατανομή. Αυτές οι ενημερώσεις συνεχίζονται έως ότου ο διαχωριστής δεν ξεχωρίζει πλέον τα δεδομένα (φάση (d)).

Αξίζει να σημειωθεί ότι η διαδικασία εκπαίδευσης των GAN δεν είναι τόσο απλή όσο η διαδικασία που παρουσιάζεται στο Σχήμα 2.9. Η ψεύτικη κατανομή δεδομένων επικαλύπτεται πλήρως με την πραγματική κατανομή υπό ιδανικές συνθήκες, ενώ υπάρχουν διάφορες προκλήσεις στην πράξη.

[9]Πηγή: https://arxiv.org/pdf/1406.2661.pdf

Κεφαλαιο 3

Σχετικη ΄Ερευνα

3.1 Δημιουργία Βίντεο

Παλαιότερες έρευνες στον κλάδο της βαθιάς δημιουργίας βίντεο μελετούσαν τρόπους με τους οποίους χωροχρονικά νευρωνικά δίκτυα θα κατάφερναν να δημιουργήσουν καρέ των βίντεο από διανύσματα θορύβου [3, 4]. Πρότειναν ένα γενετικό ανταγωνιστικό νευρωνικό δίκτυο με χωροχρονική αρχιτεκτονική συνελικτικού δικτύου που ξεχώριζε το προσκήνιο από το φόντο. Πρόκειται για ένα μοντέλο δυναμικής σκηνής όπου θα μπορούσε να προβλέπει εύλογες μελλοντικές στατικές εικόνες και να αναγνωρίζει κινήσεις και δράσεις με ελάχιστη επίβλεψη. [4] Σε άλλη παρόμοια έρευνα, χρησιμοποιήθηκε ένα γενετικό μοντέλο, το χρονικό γενετικό ανταγωνιστικό δίκτυο, το οποίο μπορούσε να μάθει μια σημασιολογική αναπαράσταση των βίντεο χωρίς ετικέτα και να δημιουργεί νέα βίντεο. Σε αντίθεση με τα υπάρχοντα GAN που δημιουργούν βίντεο με μία μόνο γεννήτρια, αυτό το μοντέλο εκμεταλλεύεται δύο διαφορετικούς τύπους γεννητριών, μία χρονική γεννήτρια και μια γεννήτρια εικόνων. Η χρονική γεννήτρια παίρνει μια μόνο λανθάνουσα μεταβλητή ως είσοδο και εξάγει ένα σύνολο από λανθάνουσες μεταβλητές, καθεμία από τις οποίες αντιστοιχεί σε μια εικόνα καρέ σε ένα βίντεο. Η γεννήτρια εικόνας μετατρέπει ένα σύνολο από τέτοιες λανθάνουσες μεταβλητές σε ένα βίντεο. [3]

Σε πιο πρόσφατες αναζητήσεις, αρκετές προσεγγίσεις αντιμετώπισαν το πρόβλημα της δημιουργίας βίντεο υπό όρους (Conditional Video Generation). Για παράδειγμα, ο Wang και η ομάδα του [5] συνδυάζουν ένα επαναλαμβανόμενο νευρωνικό δίκτυο μαζί με έναν Παραλλαγμένο Αυτόματο Κωδικοποιητή για να δημιουργήσει βίντεο με πορτραίτα ανθρώπων. Λαμβάνοντας υπόψη μια εικόνα εισόδου ενός ουδέτερου προσώπου, μπορεί να δημιουργήσει πολλά βίντεο χαμόγελου με διακριτικά χαρακτηριστικά. Έτσι, προτείνει μια αρχιτεκτονική βαθιάς μάθησης με το όνομα Conditional Multi-Mode Network (CMM-Net), το οποίο εκμεταλλεύεται τα χαρακτηριστικά σημεία του προσώπου (μάτια, μύτη, στόμα κλπ) για τη δημιουργία ακολουθιών χαμόγελου. Τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks), είναι ένας τύπος τεχνητών νευρωνικών δικτύων που έχουν σχεδιαστεί για να αναγνωρίζουν μοτίβα σε ακολουθίες δεδομένων, όπως χρονοσειρές δεδομένων που προέρχονται από αισθητήρες, κείμενα, μουσική, χειρόγραφα, ομιλίες κ.ά.. Αυτό που διαφοροποιεί τα επαναλαμβανόμενα νευρωνικά δίκτυα από άλλα δίκτυα είναι ότι λαμβάνουν υπόψη τον χρόνο και την ακολουθία, έχουν δηλαδή μια διάσταση του χρόνου.

Ο αυτόματος κωδικοποιητής (Autoencoder) είναι ένας τύπος τεχνητού νευρωνικού δικτύου που χρησιμοποιείται για την εκμάθηση αποτελεσματικών κωδικοποιήσεων των δεδομένων με έναν μη εποπτευόμενο τρόπο. Ο στόχος ενός αυτόματου κωδικοποιητή είναι να μάθει μια αναπαράσταση (κωδικοποίηση) για ένα σύνολο δεδομένων, συνήθως για μείωση των διαστάσεων, εκπαιδεύοντας το δίκτυο να αγνοήσει το σήματα θορύβου. Μαζί με την πλευρά μείωσης των διαστάσεων, μαθαίνει και μια πλευρά ανακατασκευής, όπου ο αυτόματος κωδικοποιητής προσπαθεί να δημιουργήσει από τη περιορισμένη κωδικοποίηση μια αναπαράσταση όσο το δυνατόν πιο κοντά στην αρχική του είσοδο. Οι αυτόματοι κωδικοποιητές εφαρμόζονται σε πολλά προβλήματα, από την αναγνώριση προσώπου έως την απόκτηση της σημασιολογικής σημασίας των λέξεων. Υπάρχουν πολλές διαφορετικές παραλλαγές της γενικής αρχιτεκτονικής του αυτόματου κωδικοποιητή με στόχο να διασφαλιστεί ότι η συμπιεσμένη αναπαράσταση αντιπροσωπεύει ουσιαστικά χαρακτηριστικά της αρχικής εισαγωγής δεδομένων. Μία παραλλαγή αυτών είναι οι παραλλαγμένοι αυτόματοι κωδικοποιητές (Variatonal Autoencoders). Οι παραλλαγμένοι αυτόματοι κωδικοποιητές είναι ένα είδος βαθιών γενετικών μοντέλων όπως και τα γενετικά ανταγωνιστικά δίκτυα. Πρόκειται για αυτόματους κωδικοποιητές των οποίων η κατανομή κωδικοποιήσεων κανονικοποιείται κατά τη διάρκεια της

εκπαίδευσης, προκειμένου να διασφαλιστεί ότι ο λανθάνων χώρος του έχει καλές ιδιότητες που επιτρέπουν την γενετική διαδικασία.

Τα οπτικά σήματα σε ένα βίντεο μπορούν να χωριστούν σε περιεχόμενο και κίνηση. Ενώ το περιεχόμενο καθορίζει ποια αντικείμενα βρίσκονται στο βίντεο, η κίνηση περιγράφει τη δυναμική τους. Με βάση αυτό, ο Tulyakov και οι συνεργάτες του εισήγαγαν το MoCoGan [6], μία επαναλαμβανόμενη αρχικτεκτονική που εκπαιδεύτηκε με ανταγωνιστικό τρόπο, για να συνθέσει βίντεο από θορύβους, κατηγοριοποιημένες ετικέτες ή στατικές εικόνες. Πρόκειται για ένα αποσυντιθέμενο γενετικό ανταγωνιστικό δίκτυο που δημιουργεί βίντεο αντιστοιχώντας μια ακολουθία τυχαίων διανυσμάτων σε μια ακολουθία καρέ βίντεο. Κάθε τυχαίο διάνυσμα αποτελείται από ένα μέρος περιεχομένου και ένα μέρος κίνησης. Ενώ το μέρος περιεχομένου διατηρείται σταθερό, το μέρος κίνησης υποδηλώνεται ως μια στοχαστική διαδικασία. Το MoCoGan επιτρέπει σε κάποιον να δημιουργεί βίντεο με ίδιο περιεχόμενο και διαφορετική κίνηση, καθώς και βίντεο με διαφορετικό περιεχόμενο και ίδια κίνηση.

Μία άλλη συνηθισμένη περίπτωση δημιουργίας βίντεο υπό όρους είναι το πρόβλημα της μελλοντικής πρόβλεψης καρέ, κατά την οποία το παραγόμενο βίντεο ρυθμίζεται σύμφωνα με το αρχικό καρέ. Για να εκμάθηση της κίνησης ενός φυσικού αντικειμένου που δεν περιλαμβάνει ετικέτες, αναπτύχθηκαν μοντέλα πρόβλεψης βίντεο με κίνηση, τα οποία μοντελοποιούν την κίνηση των εικονοστοιχείων, προβλέποντας μια κατανομή που ακολουθούν οι κινήσεις των εικονοστοιχείων από προηγούμενα καρέ [7] Με την ίδια λογική, αντί να προβλεφθούν άμεσα τα εικονοστοιχεία σε ένα καρέ, προβλέπονται οι μετασχηματισμοί που απαιτούνται για τη δημιουργία του επόμενου καρέ σε μια ακολουθία, σύμφωνα με τους μετασχηματισμούς των προηγούμενων καρέ. Αυτό οδήγησε σε ευκρινέστερα αποτελέσματα, ενώ χρησιμοποιήθηκε μικρότερο μοντέλο πρόβλεψης [8]. Επιπλέον, για τον ίδιο σκοπό, προτάθηκαν δύο βαθιές αρχιτεκτονικές νευρωνικών δικτύων που στηρίζονταν σε κωδικοποίηση, μετασχηματισμό υπό όρους και αποκωδικοποίηση. Οι αρχιτεκτονικές αυτές βασίζονταν σε συνελικτικά νευρωνκά δίκτυα και επαναλαμβανόμενα νευρωνικά δίκτυα [9].

Για την εκμάθηση αναπαραστάσεων των ακολουθιών βίντεο ο Srivastava και οι συνεργάτες του, χρησιμοποίησαν δίκτυα μακροπρόθεσμης μνήμης Long Short-Term Memory (LSTMs) [10]. Το μοντέλο τους χρησιμοποιούσε έναν κωδικοποιητή LSTM για να αντιστοιχίσει μια ακολουθία εισόδου εικονοστοιχείων σε μια αναπαράσταση

σταθερού μήκους. Αυτή η αναπαράσταση αποκωδικοποιείται με χρήση ενός ή πολλαπλών αποκωδικοποιητών LSTM για την εκτέλεση διαφορετικών εργασιών, όπως η ανακατασκευή της ακολουθίας εισόδου ή η πρόβλεψη της μελλοντικής ακολουθίας. Οι μονάδες ή τα μπλοκ μακροπρόθεσμης μνήμης LSTM αποτελούν μέρος της δομής ενός επαναλαμβανόμενου νευρωνικού δικτύου. Τα επαναλαμβανόμενα νευρωνικά δίκτυα χρησιμοποιούν μακροπρόθεσμα μπλοκ μνήμης για να παρέχουν ένα πλαίσιο, σύμφωνα με το οποίο το πρόγραμμα λαμβάνει εισόδους και δημιουργεί εξόδους. Το μακροπρόθεσμο μπλοκ μνήμης είναι μια πολύπλοκη μονάδα με διάφορα στοιχεία όπως σταθμισμένες εισόδους, συναρτήσεις ενεργοποίησης, εισόδους από προηγούμενα μπλοκ και ενδεχόμενες εξόδους. Η μονάδα ονομάζεται μακροπρόθεσμο μπλοκ μνήμης επειδή το πρόγραμμα χρησιμοποιεί μια δομή που βασίζεται σε διαδικασίες βραχυπρόθεσμης μνήμης για τη δημιουργία μακροπρόθεσμης μνήμης. Αυτά τα συστήματα χρησιμοποιούνται συχνά, για παράδειγμα, στην επεξεργασία φυσικής γλώσσας. Το επαναλαμβανόμενο νευρωνικό δίκτυο χρησιμοποιεί τα μακροπρόθεσμα μπλοκ μνήμης για να πάρει μια συγκεκριμένη λέξη ή φωνή και να την αξιολογήσει ως μια συμβολοσειρά, όπου η μνήμη μπορεί να είναι χρήσιμη για την ταξινόμηση και την κατηγοριοποίηση αυτών των τύπων εισόδων.

Επίσης, μια ακόμη έρευνα έγινε από τον Zhao και την ομάδα του [11], οι οποίοι πρότειναν ένα μοντέλο με το οποίο σύντομες κινήσεις υψηλής ποιότητας μπορούν να δημιουργηθούν από χωροχρονικά γενετικά δίκτυα που αξιοποιούν τη χρονική γνώση από τα δεδομένα εκπαίδευσης. Πρότειναν ένα γενετικό πλαίσιο δύο σταδίων όπου τα βίντεο δημιουργούνται από κατασκευάσματα και στη συνέχεια τελειοποιούνται με χρονικά σήματα. Για να μοντελοποιήσουν τις κινήσεις πιο αποτελεσματικά και να αποφύγουν την εκμάθηση άσχετων λεπτομερειών κίνησης, εκπαίδευσαν τα δίκτυα με τέτοιο τρόπο ώστε να μάθουν την υπολειπόμενη κίνηση μεταξύ του τρέχοντος και του μελλοντικού καρέ. Σε αυτές τις μεθόδους, ρεαλιστικές προβλέψεις μπορούν να επιτευχθούν μέσω απλού μετασχηματισμού (σκέβρωσης) του αρχικού καρέ [12, 7, 8].

3.2 Αναπαράσταση Κίνησης Εικόνων

Η δημιουργία βίντεο με κίνηση αντικειμένων σε ακίνητες εικόνες έχει πλέον αμέτρητες εφαρμογές σε κλάδους παραγωγής ταινιών, φωτογραφίας και ηλεκτρονικού εμπορίου κ.ά. Πιο συγκεκριμένα, η αναπαράσταση κίνησης σε εικόνα αναφέρεται

στην διαδικασία της αυτόματης σύνθεσης βίντεο συνδυάζοντας την εμφάνιση που διεξάγεται από μια εικόνα που λειτουργεί ως πηγή με μοτίβα κίνησης που προέρχονται από ένα βίντεο που λειτουργεί ως οδηγός. Παραδοσιακές προσεγγίσεις για αναπαράσταση κίνησης σε εικόνα και επαναπροσδιορισμό βίντεο [13, 14, 15] σχεδιάστηκαν για συγκεκριμένους τομείς όπως ανθρώπινα πρόσωπα [16, 17] ,ανθρώπινες σιλουέτες [18, 19, 20] και χειρονομίες [21] και απαιτούσαν ισχυρές εκ των προτέρων γνώσεις για τα αντικείμενα που επρόκειτο να αναπαρασταθούν.

Οι τομείς των γραφικών υπολογιστών και της υπολογιστικής όρασης έχουν αφιερώσει μακροχρόνιες προσπάθειες στην κατασκευή υπολογιστικών εργαλείων για την ανακατασκευή, παρακολούθηση και ανάλυση ανθρώπινων προσώπων με βάση την οπτική είσοδο. Τα τελευταία χρόνια έχει σημειωθεί πολύ μεγάλη τεχνολογική πρόοδος, που οδήγησε σε νέους σχυρούς αλγόριθμους που επιτυγχάνουν εντυπωσιακά αποτελέσματα ακόμη και στην πολύ δύσκολη περίπτωση της ανακατασκευής από μία RGB κάμερα. Το εύρος των εφαρμογών είναι τεράστιο και αυξάνεται σταθερά καθώς αυτές οι τεχνολογίες βελτιώνονται συνεχώς ως προς την ταχύτητα, την ακρίβεια και την ευκολία χρήσης. Με αυτό το κίνητρο και για την αναπαράσταση κίνησης σε πρόσωπα σε πραγματικό χρόνο, ο Zollhofer και η ομάδα του [17] παρήγαγε ρεαλιστικά αποτελέσματα βασιζόμενος σε τρισδιάστατα μορφοποιημένα μοντέλα του προσώπου που χρησιμοποιούσαν αλγόριθμους ανακατασκευής με βάση τη βελτιστοποίηση. Χρησιμοποίησε πλήρως τις εκ των προτέρων πληροφορίες που διέθετε για να αντιμετωπίσει με τον καλύτερο τρόπο τα προβλήματα της περιορισμένης μονοφθαλμικής (monocular) ανακατασκευής και δοκίμασε τεχνικές βελτιστοποίησης για ανάκτηση πυκνών, φωτομετρικών τρισδιάστατων μοντέλων προσώπου από μονοφθαλμικά δισδιάστατα δεδομένα. Ωστόσο, σε πολλές εφαρμογές, τέτοια μοντέλα δεν είναι διαθέσιμα.

Η αναπαράσταση κίνησης σε εικόνα μπορεί επίσης να αντιμετωπιστεί ως πρόβλημα μεταφοράς από έναν οπτικό τομέα σε έναν άλλο. Ο Isola [22] διερεύνησε υπό όρους ανταγωνιστικά νευρωνικά δίκτυα για χρήση σε προβλήματα μεταφοράς εικόνας-προς-εικόνα γενικού σκοπού. Αυτά τα δίκτυα όχι μόνο μαθαίνουν την αντιστοίχιση από την εικόνα εισόδου στην εικόνα εξόδου, αλλά επίσης μαθαίνουν μια συνάρτηση απώλειας για να εκπαιδεύσουν αυτήν την αντιστοίχιση. Με αυτό τον τρόπο, η ίδια γενική προσέγγιση μπορεί να εφαρμοστεί και σε προβλήματα που παραδοσιακά θα απαιτούσαν πολύ διαφορετικές συναρτήσεις απώλειας. Η μέθοδος του Isola είναι αποτελεσματική στη σύνθεση φωτογραφιών από χάρτες

ετικετών, στην ανακατασκευή αντικειμένων από χάρτες ακμών, στο χρωματισμό εικόνων, και σε πληθώρα άλλων άλλων εργασιών. Ο Wang [20], στηρίχτηκε στο πλαίσιο μεταφοράς εικόνα-προς-εικόνα του Isola και μετέφερε την ανθρώπινη κίνηση επεκτείνοντας το σε ένα πρόβλημα μεταφοράς βίντεο-προς-βίντεο. Στόχος του ήταν να μάθει μια συνάρτηση αντιστοίχισης από ένα βίντεο εισόδου που λειτουργεί ως πηγή, όπως για παράδειγμα μια ακολουθία μασκών σημασιολογικής τμηματοποίησης (semantic segmentation) σε ένα φωτορεαλιστικό βίντεο εξόδου που απεικονίζει ακριβώς το περιεχόμενο του βίντεο πηγής. Έτσι, πρότεινε μια προσέγγιση σύνθεσης βίντεο-προς-βίντεο στο πλαίσιο της γενετικής ανταγωνιστικότητας. Μέσα από προσεκτικά σχεδιασμένες γεννήτριες και διαχωριστές, σε συνδυασμό με έναν χωροχρονικό ανταγωνιστικό στόχο, κατάφερε να πετύχει χρονικά συναφή ,φωτορεαλιστικά βίντεο υψηλής ανάλυσης σε ένα διαφορετικό σύνολο μορφών εισόδου, όπως μάσκες τμηματοποίησης, σκίτσα και πόζες.

Ομοίως, οι Bansal και οι συνεργάτες του [23] επέκτειναν αυτές τις μεθόδους με υπό όρους γενετικά ανταγωνιστικά δίκτυα, τα Recycle-GAN, ενσωματώνοντας χωροχρονικά στοιχεία για τη βελτίωση της μεταφοράς βίντεο-προς-βίντεο από τον έναν τομέα στον άλλον. Παρουσίασαν μια προσέγγιση βάσει δεδομένων για μη επιβλεπόμενο επαναπροσδιορισμό βίντεο που μεταφέρει περιεχόμενο από έναν τομέα σε έναν άλλο, διατηρώντας ταυτόχρονα το στυλ του τομέα στον οποίον μεταφέρεται. Για παράδειγμα, εάν το περιεχόμενο της ομιλίας ενός ατόμου Α επρόκειτο να μεταφερθεί σε ένα άτομο Β,τότε η παραγόμενη ομιλία πρέπει να ακολουθεί το στυλ του ατόμου Β (χροιά φωνής). Η προσέγγιση αυτή συνδύαζε τόσο χωρικές όσο και χρονικές πληροφορίες μαζί με ανταγωνιστικές απώλειες για την μεταφορά περιεχομένου και τη διατήρηση στυλ. Όμως, τέτοιες προσεγγίσεις απαιτούν ώρες βίντεο από αυτό το άτομο με ετικέτες σημασιολογικών πληροφοριών, και ως εκ τούτου πρέπει να επανεκπαιδεύονται για κάθε άτομο.

Προτάθηκαν αρκετές προσεγγίσεις που δεν απαιτούν εκ των προτέρων πληροφορίες σχετικά με το αντικείμενο. Η Wiles και η ομάδα της πρότεινε το X2Face [24], ένα μοντέλο νευρωνικού δικτύου, το οποίο μπορεί να ρυθμίζει ένα πρόσωπο-πηγή (που καθορίζεται από ένα ή περισσότερα καρέ) χρησιμοποιώντας ένα άλλο πρόσωπο ως καρέ-οδηγός για να δημιουργήσει ένα νέο καρέ το οποίο θα έχει την ίδια ταυτότητα με το καρέ-πηγή αλλά η στάση και η έκφραση του προσώπου θα είναι παρόμοια με αυτή του καρέ-οδηγού. Το δίκτυο μπορεί να εκπαιδευτεί με πλήρη αυτο-επίβλεψη χρησιμοποιώντας μια μεγάλη συλλογή δεδομένων βίντεο. Το

X2Face χρησιμοποιεί ένα πυκνό πεδίο κίνησης για τη δημιουργία βίντεο εξόδου μέσω παραμόρφωσης της εικόνας (σκέβρωσης).

Ο Siarohin και η ομάδα του [25] εισήγαγε το Monkey-Net, ένα αυτο-εποπτευόμενο πλαίσιο για αναπαράσταση κίνησης αυθαίρετων αντικειμένων χρησιμοποιώντας αραιές τροχιές χαρακτηριστικών σημείων. Λαμβάνοντας υπόψη μια εικόνα εισόδου με ένα αντικείμενο στόχου και μια ακολουθία ενός βίντεο-οδηγού που απεικονίζει ένα κινούμενο αντικείμενο, το Monkey-Net δημιουργεί ένα βίντεο στο οποίο το αντικείμενο στόχου κινείται σύμφωνα με την ακολουθία του βίντεο-οδηγού. Αυτό επιτυγχάνεται μέσω μιας βαθιάς αρχιτεκτονικής που αποσυνθέτει την εμφάνιση(στάση) και την κίνηση. Το πλαίσιο τους αποτελείται από τρεις κύριες ενότητες: (α) έναν ανιχνευτή χαρακτηριστικών σημείων που εκπαιδεύεται χωρίς επίβλεψη ώστε να εξάγει τα χαρακατηριστικά σημεία, (β) ένα δίκτυο πρόβλεψης πυκνής κίνησης για τη δημιουργία πυκνών χαρτών θερμότητας από αραιά χαρ/κά σημεία, προκειμένου να κωδικοποιηθούν καλύτερα οι πληροφορίες κίνησης και (γ) ένα δίκτυο μεταφοράς κίνησης, το οποίο χρησιμοποιεί τους χάρτες θερμότητας κίνησης και τις πληροφορίες εμφάνισης που εξάγονται από την εικόνα εισόδου για να συνθέσει τα τελικά καρέ.

Σε πιο πρόσφατη έρευνα, ο Siarohin και η ομάδα του, βασισμένοι στα προηγούμενα μοντέλα, πρότειναν το μοντέλο First Order Motion [26] με αρκετά κοινή αρχιτεκτονική με το Monkey-Net. Η προσέγγιση τους δεν βασίζεται ούτε σε ετικέτες, ούτε σε εκ των προτέρων πληροφορίες για τα κινούμενα αντικείμενα, ούτε για συγκεκριμένες εκπαιδευτικές διαδικασίες για κάθε ένα αντικείμενο. Επιπλέον, μπορεί να εφαρμοστεί σε οποιοδήποτε αντικείμενο της ίδιας κατηγορίας(π.χ. πρόσωπα, ανθρώπινα σώματα κ.λπ.) Όπως και στο Monkey-Net, έτσι και εδώ, υπολογίζονται αραιές τροχιές χαρ/κών σημείων με μη-επιβλεπόμενο τρόπο. Ωστόσο,σε αντίθεση με το Monkey-net, μοντελοποιούν την κίνηση των αντικειμένων στη γειτονιά κάθε χαρ/κού σημείου μέσω τοπικών αφινικών μετασχηματισμών και όχι μόνο στα ίδια τα σημεία. Επίσης, παρουσιάζει μια γεννήτρια, η οποία υιοθετεί μια μάσκα απόκλισης για να υποδείξει τα τμήματα των αντικειμένων που μπορεί να παράξει μέσω παραμόρφωσης της εικόνας ή τα τμήματα που δεν είναι ορατά στην εικόνα-πηγή και θα πρέπει να τα ζωγραφίσει. Ακόμη, επεκτείνει την απώλεια συμμετρίας (equivariance) που χρησιμοποιείται συνήθως για εκπαίδευση ανιχνευτών χαρακτηριστικών σημείων [27, 28] για τη βελτίωση της εκτίμησης των τοπικών αφινικών μετασχηματισμών.

Στην εργασία αυτή, βασιζόμαστε στο μοντέλο First Order Motion του Siarohin

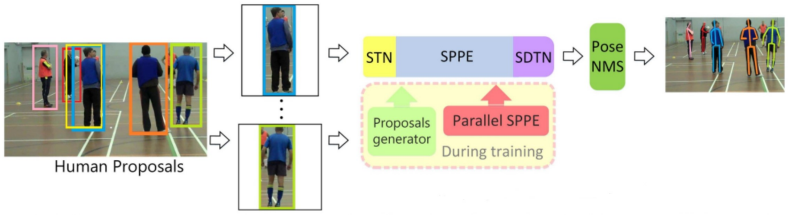
και παρουσιάζουμε ένα βελτιωμένο μοντέλο με πολλά κοινά χαρακτηριστικά. Στο FOMM μοντέλο του Siarohin που δοκιμάσαμε, τα αποτελέσματα δεν ήταν τόσο ικανοποιητικά ως προς την εκπαίδευση του ανιχνευτή των χαρακτηριστικών σημείων, με αποτέλεσμα να ανιχνεύει λανθασμένες θέσεις των χαρακτηριστικών σημείων που προκαλούσαν προβλήματα στην σωστή αναπαράσταση κίνησης των αντικειμένων. Έτσι, για να διορθώσουμε το πρόβλημα αυτό και να βελτιώσουμε τα αποτελέσματα μας, έχουμε αφαιρέσει εντελώς τον Ανιχνευτή των χαρακτηριστικών σημείων που χρησιμοποιεί ο Siarohin και τον έχουμε αντικαταστήσει με ένα προεκπαιδευμένο μοντέλο που υπολογίζει τις πόζες και τις στάσεις του σώματος και τους μορφασμούς του προσώπου. Το προεκπαιδευμένο μοντέλο δεν χρειάζεται καμία εκ των προτέρων πληροφορία για τα δεδομένα εισόδου, ούτε ετικέτες και ταμπέλες και μπορεί να χρησιμοποιηθεί για οποιοδήποτε σετ δεδομένων που περιλαμβάνει σώματα ανθρώπων και πρόσωπα. Μέσα από αυτό υπολογίζουμε τα χαρακτηριστικά σημεία για το σώμα και το πρόσωπο. Κάθε χαρ/κό σημείο αντικατοπτρίζει ένα συγκεκριμένο σημείο του αντικειμένου μελέτης. Επιπλέον, έχουν αφαιρεθεί και οι υπολογισμοί που απαιτούσε το μοντέλο του Siarohin σχετικά με την απώλεια συμμετρίας (equivariance), μιας και τώρα δεν εκπαιδεύουμε κάποιον ανιχνευτή χαρακτηριστικών σημείων. Το μοντέλο μας είναι πιο απλοποιημένο και στην ουσία περιορίζεται στην εκπαίδευση μόνο ενός γενετικού ανταγωνιστικού νευρωνικου δικτύου. Στην επόμενη ενότητα παρουσιάζεται αναλυτικά η μεθοδολογία του μοντέλου μας.

Κεφαλαιο 4

Μεθοδολογια

4.1 Υπολογισμός χαρακτηριστικών σημείων

Ο σκοπός μας στην παρούσα εργασία είναι να δημιουργήσουμε μία νέα ακολουθία καρέ από μία εικόνα-πηγή, στην οποία το αντικείμενο του στόχου μας κινείται σύμφωνα με τις κινήσεις ενός βίντεο-οδηγού. Στο συγκεκριμένο πρόβλημα, δεν έχουμε καμία εκ των προτέρων πληροφορία ούτε groundtruth δεδομένα για το συγκεκριμένο αντικείμενο που μελετάμε. Τα αντικείμενα της μελέτης μας είναι ολόκληρα σώματα ανθρώπων και τα πρόσωπα τους. Για να υποστηρίξουμε τις σύνθετες κινήσεις του αντικειμένου μελέτης, χρησιμοποιούμε μια αναπαράσταση που αποτελείται από ένα σετ χαρακτηριστικών σημείων σε συνδυασμό με τους τοπικούς αφινικούς μετασχηματισμούς τους. Η πρώτη φάση της μεθόδου μας για την αναπαράσταση κίνησης είναι ο υπολογισμός των χαρακτηριστικών σημείων που αντιπροσωπεύουν συγκεκριμένα σημεία του σώματος ή του προσώπου (π.χ. χέρι, ώμος, μάτια κ.λ.π.). Για το σκοπό αυτό έχουμε χρησιμοποιήσει ένα προεκπαιδευμένο μοντέλο που αναγνωρίζει και υπολογίζει σε μία εικόνα, τις στάσεις πολλαπλών ανθρώπινων σωμάτων (Regional Multi-Person Pose Estimation **RMPE**) [29] ακόμα κι αν τα πλαίσια οριοθέτησης των αντικειμένων είναι ανακριβή. Αυτές οι στάσεις έχουν σχήμα σκελετού που δημιουργείται από την ένωση των χαρακτηριστικών σημείων με οριζόντιες γραμμές



Σχήμα 4.1: Σχεδίαση Μοντέλου RMPE

όπως φαίνεται στην τελική εικόνα εξόδου του Σχήματος 4.1[1]. Το μοντέλο αυτό είναι γενικό και μπορεί να εφαρμοστεί σε διαφορετικούς ανιχνευτές ανθρώπων (human detectors) και σε υπολογισμό στάσης ενός ατόμου.

Η σχεδίαση αυτού του μοντέλου αποτελείται από τρία μέρη:

1. Συμμετρικό δίκτυο χωρικών μετασχηματισμών (Symmetric Spatial Transformer Network **STN**)

2. Δίκτυο μη μέγιστης καταστολής παραμετρικής στάσης (Parametric Pose Non-Maximum Suppression **Pose NMS**)

3. Γεννήτρια ανθρώπινων προσομοιωμάτων (σκελετών) καθοδηγούμενη από τις στάσεις (Pose Guided Proposal Generator **PGPG**)

Στο συγκεκριμένο μοντέλο έχει χρησιμοποιηθεί ο βασισμένος στο VGG SSD-512, ως ανιχνευτής ανθρώπων, καθώς έχει εξαιρετική απόδοση στην αναγνώριση αντικειμένων. Επίσης χρησιμοποιήθηκε το stacked hourglass μοντέλο ως εκτιμητής στάσης ενός ατόμου (SPPE) και το δίκτυο ResNet-18 ως δίκτυο εντοπισμού.

Το μοντέλο, όπως φαίνεται και στο σχήμα 4.1 περιγράφεται εν συντομία ως εξής: Αρχικά, ο ανιχνευτής ατόμων βρίσκει τα πλαίσια οριοθέτησης των ανθρώπων, τα οποία εισάγονται στο STN. Μέσα στο δίκτυο αυτό, εξάγονται ανθρώπινες προσομοιώσεις υψηλής ποιότητας (proposals), μέσω δισδιάστατων affine μετασχηματισμών. Στη συνέχεια, προσκολάται το SPPE και η στάση που δίνεται ως έξοδος σχεδιάζεται στην αρχική εικόνα των ανθρώπινων ομοιωμάτων. Το SPPE έχει εκπαιδευτεί σε εικόνες ενός ατόμου και παρουσιάζει δυσκολία στη διαχείριση των σφαλμάτων εντοπισμού πολλών ανθρώπων. Για αυτό το λόγο, προηγείται το STN, για να ενισχύσει την αποδοτικότητα του SPPE, όταν γίνονται εσφαλμένοι υπολογισμοί στάσεων. Το τρίπτυχο ολοκληρώνεται με ένα χωρικό δίκτυο αντίστροφου

[1]Πηγή: https://openaccess.thecvf.com/content_ICCV_2017/papers/Fang_RMPE_Regional_Multi-Person_ICCV_2017_paper.pdf

μετασχηματισμού (Spatial De-Transformer Network SDTN), το οποίο ξανασχεδιάζει την υπολογισμένη ανθρώπινη πόζα με τις συνταναγμένες της αρχικής εικόνας. Το SDTN πραγματοποιεί αντίστροφη λειτουργία από αυτή του STN. Ένα παράλληλο SPPE χρησιμοποιήθηκε ως ένας επιπλέον κανονικοποιητής για να βοηθήσει το STN να βελτιώσει την αποδοτικότητα του. Ο PGPG σχεδιάστηκε για να αυξηθούν τα ήδη υπάρχοντα δεδομένα εκπαίδευσης. Τέλος, προστίθεται το παραμετρικό Pose-NMS , το οποίο εξαλείφει τυχόν περιττές ανιχνεύσεις στάσεων.

4.2 Σύνθεση κίνησης σε βίντεο μέσω εκμάθησης τοπικών μεταχηματισμών

Ο σχεδιασμός του μοντέλου μας συνεχίζεται με τη δημιουργία ενός δικτύου κίνησης που δέχεται ως είσοδο την αναπαράσταση των κινήσεων και υπολογίζει την οπτική ροή μεταξύ του καρέ-οδηγού από το βίντεο-οδηγό και της εικόνας-πηγής. Επιπλέον, το δίκτυο κίνησης υπολογίζει έναν χάρτη απόκλισης, ο οποίος υποδεικνύει τα μέρη της εικόνας τα οποία δεν είναι εμφανή και πρέπει να τα υπολογίσει. Στη συνέχεια, προσθέτουμε στο μοντέλο μας ένα γενετικό ανταγωνιστικό δίκτυο, το οποίο θα δημιουργήσει την τελική εικόνα για κάθε καρέ του βίντεο. Στις παρακάτω ενότητες, παραθέτονται αναλυτικά οι αρχιτεκτονικές των δικτύων που χρησιμοποιήθηκαν.

4.2.1 Πυκνό Πεδίο Κίνησης (Dense Motion Field)

Κατά τη διάρκεια της εκπαίδευσης, εισάγονται στο μοντέλο ζεύγη εικόνων διάστασης HxW που αποτελούν δύο τυχαία καρέ από το ίδιο βίντεο. Το ένα καρέ δηλώνεται ως εικόνα-πηγή $\mathbf{S} \in \mathbb{R}^{3xHxW}$ και το άλλο ως εικόνα-οδηγός $\mathbf{D} \in \mathbb{R}^{3xHxW}$. Με αυτό τον τρόπο η εικόνα-οδηγός λειτουργεί ως groundtruth δεδομένο και ο σκοπός μας είναι να εκπαιδευτεί το δίκτυο ώστε να δημιουργεί από την εικόνα-πηγή ένα νέο καρέ που θα είναι ίδιο με το groundtruth (εικόνα-οδηγός). Αφού υπολογιστούν τα χαρακτηριστικά σημεία για τα δύο καρέ από τον αλγόριθμο που αναλύσαμε στην προηγούμενη ενότητα , εισάγονται και αυτά στο δίκτυο κίνησης για να περιγράψουν τις σχετικές κινήσεις των εικονοστοιχείων. Το δίκτυο κίνησης μοντελοποιείται από μια συνάρτηση $\mathcal{T}_{S \leftarrow D} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ που αντιστοιχεί κάθε εικονοστοιχείο στο καρέ $\mathbf{D}$ με την αντίστοιχη θέση του στο καρέ $\mathbf{S}$. Η συνάρτηση $\mathcal{T}_{S \leftarrow D}$ αναφέρεται και ως

οπισθοδρομική οπτική ροή (backward optical flow) από το καρέ-οδηγό D προς το καρέ-πηγή S.

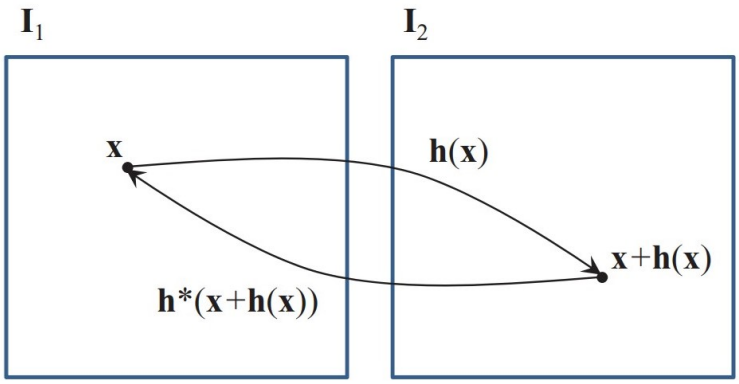
Οπτική Ροή

Η οπτική ροή περιγράφει το μοτίβο της φαινομενικής κίνησης των αντικειμένων σε μια ακολουθία εικόνων που προκαλείται από τη σχετική κίνηση μεταξύ ενός παρατηρητή και μιας σκηνής [30]. Η οπτική ροή μπορεί επίσης να οριστεί ως η κατανομή φαινομένων ταχύτητας κίνησης του μοτίβου φωτεινότητας σε μια εικόνα. Στην εργασία μας χρησιμοποιούμε οπτική ροή προς τα πίσω, αντί για εμπρόσθια οπτική ροή, καθώς μπορεί να εφαρμοστεί αποτελεσματικά σε διαφοροποιήσιμα προβλήματα με τη χρήση διγραμμικής παρεμβολής (bilinear interpolation). Η παρεμβολή εικόνας χρησιμοποιείται συνήθως σε διαδικασίες μεγέθυνσης ή σμίκρυνσης εικόνων και ορίζεται ως η διαδικασία χρήσης γνωστών δεδομένων προκειμένου να υπολογίσουμε τιμές σε άγνωστες θέσεις. Η διγραμμική παρεμβολή είναι μια μέθοδος δειγματοληψίας που χρησιμοποιεί τον σταθμισμένο μέσο όρο από τις τέσσερις πλησιέστερες τιμές εικονοστοιχείων για τον υπολογισμό της τιμής ενός νέου εικονοστοιχείου. Ο υπολογισμός της οπισθοδρομικής οπτικής ροής, είναι απαραίτητος σε τέτοιου είδους μεθόδους και έχει αποδειχθεί ότι παρέχει καλύτερα αποτελέσματα από τις λύσεις μονής-κατεύθυνσης.

Πιο πρόσφατα, η οπισθοδρομική οπτική ροή έχει επίσης χρησιμοποιηθεί σε μεθόδους όπου είναι απαραίτητη η παρακολούθηση της ροής σε προηγούμενες χρονικές στιγμές. Εάν έχουμε μια μεγάλη ακολουθία εικόνων καρέ, τότε είναι ενδιαφέρον να αποκτήσουμε συνεχείς και συνεπείς ροές μεταξύ των καρέ. Η ροή επιτρέπει την αναζήτηση των αντιστοιχιών στα προηγούμενα καρέ και τη δημιουργία χωροχρονικών περιορισμών που αποδίδουν χρονικά πεδία συνεχούς ροής. Εάν γνωρίζουμε την οπτική ροή, τότε είναι δυνατόν να εκτιμηθούν οι αντιστοιχίες προς τα πίσω με ορισμένους περιορισμούς. Στην πραγματικότητα, ο υπολογισμός της αντίστροφης οπτικής ροής μπορεί να πραγματοποιηθεί εύκολα στην περίπτωση των ένα-προς-ένα μετασχηματισμών. Ωστόσο, πρέπει να ασχοληθούμε με τη διακριτή φύση των εικόνων που καθιστούν δύσκολη την λήψη αντίστροφων χαρτών με ακρίβεια. Η οπτική ροή εκφράζεται από ένα διανυσματικό πεδίο

$$h(x) = (u(x), v(x)), \tag{4.1}$$

που θέτει σε αντιστοιχία τα εικονοστοιχεία της εικόνας-πηγής και της εικόνας στό-



Σχήμα 4.2: Σχέση μεταξύ εμπρόσθιας και οπισθοδρομικής οπτικής ροής

χου. Η οπισθοδρομική οπτική ροή

$$h^*(x) = (u^*(x), v^*(x)), \tag{4.2}$$

είναι η αντίστροφη της $h(x)$ και και θέτει σε αντιστοιχία τα εικονοστοιχεία της εικόνας στόχου στην εικόνα-πηγή. Η σχέση μεταξύ της εμπρόσθιας και οπισθοδρομικής οπτικής ροής δίνεται από τη σχέση

$$h(x) = -h^*(x + h(x)) \tag{4.3}$$

ή

$$h^*(x) = -h(x + h^*(x)) \tag{4.4}$$

Στο Σχήμα 4.2[2] απεικονίζεται η σχέση μεταξύ εμπρόσθιας και οπισθοδρομικής οπτικής ροής σε μια ακολουθία δύο εικόνων I_1 και I_2. Εάν ακολουθήσουμε τις διαδρομές των δύο ροών, πρέπει να φτάσουμε στην αρχική θέση. Η αντιστοιχία των θέσεων δίνεται από την συνάρτηση μετασχηματισμού

$$x \leftarrow x + h(x) \tag{4.5}$$

Για να μπορούμε να επεξεργαζόμαστε ανεξάρτητα τα καρέ D και S, υποθέτουμε ότι υπάρχει ένα αφηρημένο καρέ αναφοράς R. Το καρέ R είναι μια αφηρημένη έννοια και δεν υπολογίζεται ποτέ ρητά ούτε απεικονίζεται. Ο λόγος που το αναφέρουμε είναι επειδή κατά τη διάρκεια του τελικού ελέγχου, το μοντέλο λαμβάνει ζεύγη μιας

[2]Πηγή: https://www.researchgate.net/figure/Relation-between-the-forward-hx-and-backward-h-x-optical-flows_fig1_258547557

εικόνας-πηγής και ενός καρέ-οδηγού που λαμβάνονται από διαφορετικά βίντεο και τα οποία μπορεί να είναι πολύ διαφορετικά οπτικά.

Στο πρώτο βήμα, προσεγγίζουμε και τους δύο μετασχηματισμούς από σύνολα αραιών τροχιών, που λαμβάνονται από τις μετακινήσεις των χαρακτηριστικών σημείων που έχουμε ήδη υπολογίσει. Όπως φαίνεται από τον Siarohin [26], μια τόσο αραιή αναπαράσταση κίνησης είναι κατάλληλη για κινούμενη εικόνα, καθώς κατά τη διάρκεια της τελικής δοκιμής, τα χαρακτηριστικά σημεία της εικόνας-πηγής μπορούν να μετακινηθούν σύμφωνα με τις τροχιές των σημείων στο βίντεο-οδηγό.

Κατά το δεύτερο βήμα, το δίκτυο πυκνής κίνησης συνδυάζει τις τοπικές προσεγγίσεις για την εύρεση του πεδίου πυκνής κίνησης $\hat{\mathcal{T}}_{S\leftarrow D}$. Επιπλέον, εκτός από το πεδίο πυκνής κίνησης, αυτό το δίκτυο εξάγει μια μάσκα απόκλισης $\hat{\mathcal{O}}_{S\leftarrow D}$ που δείχνει ποια τμήματα εικόνας του καρέ-οδηγού D μπορούν να ανακατασκευαστούν από στρέβλωση της εικόνας-πηγής S και ποια μέρη πρέπει να σχεδιαστούν με βάση τα περιβάλλοντα στοιχεία. Τέλος, η γεννήτρια παράγει μια νέα εικόνα του αντικειμένου της εικόνας-πηγής που κινείται όπως και στο βίντεο-οδηγό. Εδώ, χρησιμοποιούμε ένα δίκτυο γεννητριών G που στρεβλώνει την εικόνα-πηγή σύμφωνα με το $\hat{\mathcal{T}}_{S\leftarrow D}$. και χρωματίζει τα μέρη της εικόνας που αποκλείονται στην αρχική εικόνα. Στις παρακάτω ενότητες αναλύουμε λεπτομερώς καθένα από αυτά τα βήματα και τη διαδικασία εκπαίδευσης.

Το πεδίο υπολογισμού κίνησης υπολογίζει την οπισθοδρομική οπτική ροή $\mathcal{T}_{S\leftarrow D}$ από ένα καρέ-οδηγό προς μια εικόνα-πηγή S. Ο υπολογισμός του μετασχηματισμού, $\mathcal{T}_{S\leftarrow D}$ αποτελείται από δύο επιμέρους υπολογισμούς $\mathcal{T}_{S\leftarrow R}$ και $\mathcal{T}_{R\leftarrow D}$. Δοθέντος ενός καρέ X, υπολογίζουμε κάθε μετασχηματισμό $\mathcal{T}_{X\leftarrow R}$ στα χαρακτηριστικά σημεία. Υποθέτουμε ότι $p_1, ..., p_K$ είναι οι συντεταγμένες των χαρακτηριστικών σημείων στο καρέ αναφοράς R. Η συνάρτηση κίνησης $\mathcal{T}_{X\leftarrow R}$ αναπαρίσταται από τις τιμές της σε κάθε χαρακτηριστικό σημείο p_k :

$$\mathcal{T}_{X\leftarrow R}(p) = \mathcal{T}_{X\leftarrow R}(p_k) + \left(\frac{d}{dp}\mathcal{T}_{X\leftarrow R}(p)|_{p=p_k}\right)(p - p_k) + o(||p - p_k||) \tag{4.6}$$

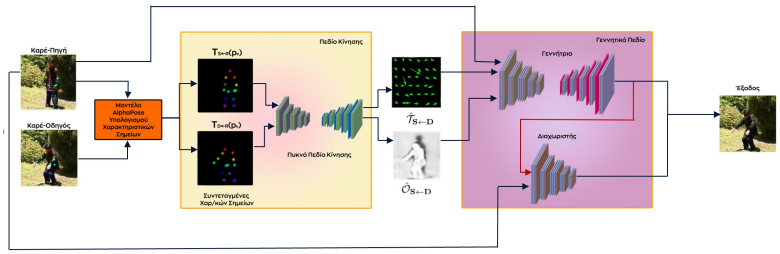
Επιπλέον, για να υπολογιστεί ο μετασχηματισμός $\mathcal{T}_{R\leftarrow X} = \mathcal{T}^{-1}_{X\leftarrow R}$ υποθέτουμε ότι ο $\mathcal{T}_{X\leftarrow R}$ είναι τοπικά ένα-προς-ένα σε κάθε χαρακτηριστικό σημείο και στη γειτονιά αυτού. Ο σκοπός μας είναι να υπολογίσουμε τον μετασχηματισμό $\mathcal{T}_{S\leftarrow D}$ κοντά στο χαρακτηριστικό σημείο z_k στο καρέ D, με την προϋπόθεση ότι z_k είναι η τοποθεσία του εικονοστοιχείου που αντιστοιχεί στην τοποθεσία p_k στο καρέ R. Αρχικά, υπολογίζουμε τον μετασχηματισμό $\mathcal{T}_{R\leftarrow D}$ κοντά στο χαρακτηριστικό σημείο

z_k στο καρέ-οδηγό D. Στη συνέχεια, υπολογίζεται μέσω του αναπτύγματος Taylor ο μετασχηματισμός:

$$\mathcal{T}_{S\leftarrow D}(z) \approx \mathcal{T}_{S\leftarrow R}(p_k) + (z - \mathcal{T}_{D\leftarrow R}(p_k)) \tag{4.7}$$

Στην πραγματικότητα, οι μετασχηματισμοί $\mathcal{T}_{S\leftarrow R}$ και $\mathcal{T}_{D\leftarrow R}$ αντιπροσωπεύουν τα χαρακτηριστικά σημεία που έχουν ηδη υπολογιστεί. Έτσι, $p_k = \mathcal{T}_{R\leftarrow D}(z_k)$. Έπειτα, υπολογίζεται ο μετασχηματισμός $\mathcal{T}_{S\leftarrow R}$ κοντά στο σημείο p_k στο καρέ αναφοράς R. Τελικά, αποκτάμε τον επιθυμητό μετασχηματισμό $\mathcal{T}_{S\leftarrow D}$:

$$\mathcal{T}_{S\leftarrow D} = \mathcal{T}_{S\leftarrow R} \circ \mathcal{T}_{R\leftarrow D} = \mathcal{T}_{S\leftarrow R} \circ \mathcal{T}^{-1}_{D\leftarrow R} \tag{4.8}$$



Σχήμα 4.3: Αρχιτεκτονική Δικτύου Αναπαράστασης Κίνησης

Αρχιτεκτονική Μοντέλου

Χρησιμοποιούμε ένα συνελικτικό νευρωνικό δίκτυο για την εκτίμηση του μετασχηματισμού $\hat{\mathcal{T}}_{S\leftarrow D}$ στα χαρακτηριστικά σημεία που έχουν ήδη υπολογιστεί για το αντικείμενο του καρέ-οδηγού και της εικόνας-πηγής. Όμως, από τη στιγμή που η οπτική ροή $\hat{\mathcal{T}}_{S\leftarrow D}$ αντιστοιχεί την τοποθεσία κάθε εικονοστοιχείου στο καρέ-οδηγό D με την αντίστοιχη τοποθεσία του στην εικόνα-πηγή S, τα τοπικά μοτίβα που περιλαμβάνονται στο πεδίο $\hat{\mathcal{T}}_{S\leftarrow D}$, όπως οι ακμές και οι υφές, είναι ευθυγραμμισμένες εικονοστοιχείο-προς-εικονοστοιχείο με το καρέ-οδηγό D αλλά όχι και με την εικόνα-πηγή S. Έτσι, για να διευκολύνουμε το δίκτυο να πραγματοποιήσει καλύτερες προβλέψεις του $\hat{\mathcal{T}}_{S\leftarrow D}$ από την εικόνα-πηγή S, στρεβλώνουμε το καρέ-πηγή S, σύμφωνα με τους τοπικούς μετασχηματισμούς που υπολογίζονται από την εξίσωση 4.7. Με αυτόν τον τρόπο, αποκτάμε K μετασχηματισμένες εικόνες $S^1, ..., S^K$ όπου η κάθε μία εικόνα ευθυγραμμίζεται με την οπτική ροή $\hat{\mathcal{T}}_{S\leftarrow D}$ σε κάθε χαρακτηριστικό σημείο.

Τελευταία αλλά εξίσου σημαντική είναι η διατήρηση μιας επιπλέον εικόνας $S^0 = S$ που ταυτίζεται με την εικόνα-πηγή και αντιπροσωπεύει το φόντο του αντικειμένου.

Για κάθε χαρακτηριστικό σημείο υπολογίζονται οι χάρτες θερμότητας (heatmaps) H_k οι οποίοι υποδεικνύουν στο δίκτυο κίνησης την τοποθεσία κάθε μετασχηματισμού. Κάθε χάρτης θερμότητας $H_k(z)$ αντιπροσωπεύει τη διαφορά άλλων δύο χαρτών που αφορούν τους μετασχηματισμούς $T_{D\leftarrow R}(p_k)$ για το καρέ-οδηγό και $T_{S\leftarrow R}(p_k)$ για το καρέ-πηγή:

$$H_k(z) = exp\left(\frac{(T_{D\leftarrow R}(p_k) - z)^2}{\sigma}\right) - exp\left(\frac{(T_{S\leftarrow R}(p_k) - z)^2}{\sigma}\right) \tag{4.9}$$

Η διακύμανση που έχει χρησιμοποιηθεί σε όλα τα πειράματα είναι $\sigma = 0.01$

Στην συνέχεια, οι χάρτες θερμότητας H_k και οι μετασχηματισμένες εικόνες $S^0, ..., S^K$ συγχωνεύονται και εισάγονται σε ένα δίκτυο U-Net. Ο υπολογισμός του πεδίου $\hat{\mathcal{T}}_{S\leftarrow D}$ είναι εμπνευσμένος από το μοντέλο Monkey-Net [25]. Υποθέτουμε ότι ένα αντικείμενο αποτελείται από K ακέραια μέρη και πως κάθε μέρος της εικόνας κινείται σύμφωνα με την εξίσωση 4.7. Ως εκ τούτου, υπολογίζουμε K + 1 μάσκες M_k, k = 0,...,K που υποδηλώνουν που πραγματοτοποείται κάθε τοπικός μετασχηματισμός.

Το τελικό πυκνό πεδίο πρόβλεψης κίνησης $\hat{\mathcal{T}}_{S\leftarrow D}(z)$ δίνεται από τη σχέση:

$$\hat{\mathcal{T}}_{S\leftarrow D}(z) = M_0 z + \sum_{k=1}^{K} M_k(\mathcal{T}_{S\leftarrow R}(p_k) + (z - \mathcal{T}_{D\leftarrow R}(p_k)) \tag{4.10}$$

Ο όρος $M_0 z$ περιλαμβάνεται για την μοντελοποίηση τμημάτων της εικόνας που είναι ακίνητα και σταθερά όπως το φόντο.

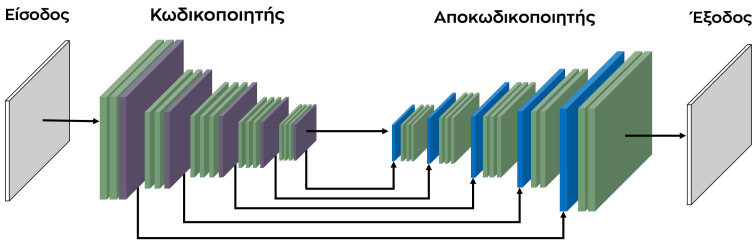
Συγκεντρωτικά, το **πυκνό πεδίο κίνησης**:

1. Δημιουργεί γκαουσιανές αναπαραστάσεις των χαρακτηριστικών σημείων της εικόνας πηγής με μορφή χάρτη θερμότητας (heatmap) (4.9) μετατρέποντας τις συντεταγμένες των χαρακτηριστικών σημείων της εικόνας-πηγής και του καρέ-οδηγού αντίστοιχα.

2. Δημιουργεί αραιές κινήσεις (4.7) από την εικόνα-πηγή και τις συντεταγμένες των χαρακτηριστικών σημείων του καρέ-πηγής και του καρέ-οδηγού.

3. Σύμφωνα με αυτές τις αραιές κινήσεις δημιουργεί μια νέα εικόνα που αποτελεί παραμόρφωση της εικόνας-πηγής (4.10)

4. Συνδυάζει την αναπαράσταση του χάρτη θερμότητας και την παραμορφωμένη εικόνα σε μία, και στην συνέχεια την εισάγει στο δίκτυο Κωδικοποιητή-Αποκωδικοποιητή Hourglass.



Σχήμα 4.4: Γενική Σχεδίαση Αρχιτεκτονικής Hourglass

Η **Αρχιτεκτονική Hourglass** που μοντελοποιείται μέσα στο δίκτυο πυκνής κίνησης και φαίνεται στο Σχήμα 4.4[3] είναι βασισμένη στο δίκτυο Unet και περιγράφεται ως εξής:

- **Είσοδος**
  Αρχικά, το δίκτυο δέχεται ως είσοδο τις αναπαραστάσεις χαρτών θερμότητας των χαρ/κών σημείων (4.9) και τις αραιές κινήσεις τους (4.7)

- **Κωδικοποιητής (Encoder)**
  Ο κωδικοποιητής υλοποιείται με ένα συνελικτικό νευρωνικό δίκτυο που περιέχει 5 μπλοκ συρρίκνωσης (downsampling) όπου το καθένα αποτελείται από τα διαδοχικά επίπεδα:

  1. Συνελικτικό Επίπεδο
  2. Επίπεδο Κανονικοποίησης Παρτίδας **(Batch Normalization)**
     Η κανονικοποίηση είναι ένα εργαλείο προεπεξεργασίας δεδομένων που χρησιμοποιείται για να φέρει τα αριθμητικά δεδομένα σε κοινή κλίμακα χωρίς να παραμορφώνεται το σχήμα τους. Ένας βασικός λόγος που κανονικοποιούμε είναι για να διασφαλίσουμε ότι το μοντέλο μας μπορεί να γενικευτεί κατάλληλα. Η κανονικοποίηση παρτίδας είναι μια διαδικασία που καθιστά τα νευρωνικά δίκτυα ταχύτερα και πιο σταθερά, προσθέτοντας επιπλέον στρώματα σε ένα βαθύ νευρωνικό δίκτυο. Το νέο επίπεδο κανονικοποιεί και ομαλοποιεί την είσοδο ενός επιπέδου που προέρχε-

[3]Πηγή: https://towardsdatascience.com/using-hourglass-networks-to-understand-human-poses-1e40e349fa15

ται από ένα προηγούμενο επίπεδο. Με την κανονικοποίηση παρτίδας ένα τυπικό νευρωνικό δίκτυο εκπαιδεύεται χρησιμοποιώντας μία ομάδα συνόλου δεδομένων εισόδου που ονομάζεται παρτίδα (batch). Έτσι, η διαδικασία ομαλοποίησης στην κανονικοποίηση παρτίδας πραγματοποιείται σε παρτίδες και όχι ως μία είσοδος.

3. Επίπεδο ReLU
4. Επίπεδο Μέσου Pooling

- **Αποκωδικοποιητής (Decoder)**
  Ο αποκωδικοποιητής υλοποιείται επίσης από ένα συνελικτικό νευρωνικό δίκτυο που περιέχει 5 μπλοκ επέκτασης (upsampling) όπου το καθένα αποτελείται από τα διαδοχικά επίπεδα:

  1. Συνελικτικό Επίπεδο
  2. Επίπεδο Κανονικοποίησης Παρτίδας (Batch Normalization)
  3. Επίπεδο ReLU

- **Έξοδος**
  Το δίκτυο επιστρέφει ως έξοδο μια εικόνα πρόβλεψης

Στην συνέχεια, το δίκτυο κίνησης εφαρμόζει διαδοχικά στην εικόνα πρόβλεψης ένα ακόμα επίπεδο δισδιάσταστης συνέλιξης και ένα επίπεδο softmax για να υπολογίσει μία μάσκα απόκλισης $\hat{\mathcal{O}}_{S\leftarrow D}$ που δείχνει ποια τμήματα εικόνας του καρέ-οδηγού D μπορούν να ανακατασκευαστούν από στρέβλωση της εικόνας-πηγής S και ποια μέρη πρέπει να σχεδιαστούν με βάση τα περιβάλλοντα στοιχεία.

Όπως αναφέρθηκε ήδη, η εικόνα-πηγή S δεν ειναι αντιστοιχισμένη εικονοστοιχείο-προς-εικονοστοιχείο με την εικόνα $\hat{D}$ που θα αναπαραχθεί. Για να μπορέσουμε να αντιμετωπίσουμε αυτήν την κακή ευθυγράμμιση, χρησιμοποιούμε μια τεχνική στρέβλωσης παρόμοια με [31, 25]. Για την ακρίβεια, εφαρμόζοντας 2 μπλοκ συρρίκνωσης στην πρόβλεψη που εκτιμήθηκε από το Hourglass μοντέλο, λαμβάνουμε έναν χάρτη χαρακτηριστικών $\xi \in R^{HxW}$ διάστασης HxW. Στην συνέχεια, στρεβλώνουμε τον χάρτη ξ σύμφωνα με τις κινήσεις $\hat{\mathcal{T}}_{S\leftarrow D}$. Με την παρουσία των αποκλίσεων στην εικόνα-πηγή S, η οπτική ροή ίσως να μην είναι ικανή να παράγει την εικόνα $\hat{D}$. Πράγματι, υπάρχουν τμήματα στην εικόνα-πηγή S που δεν μπορούν να ανακτηθούν από απλή στρέβλωση και μετασχηματισμό της εικόνας και για αυτό πρέπει να ζωγραφιστούν.

Συνεπώς, εκτιμάται ένας χάρτης απόκλισης $\hat{\mathcal{O}}_{S\leftarrow D} \in [0,1]^{H'xW'}$ που φανερώνει τα σημεία που πρέπει να σχεδιαστούν. Η μάσκα απόκλισης περιορίζει την επίδραση ορισμένων χαρακτηριστικών που αντιστοιχίζονται στα αποκλισμένα μέρη. Έτσι, ο μετασχηματισμένος χάρτης χαρακτηριστικών περιγράφεται από την εξίσωση:

$$\xi' = \hat{\mathcal{O}}_{S\leftarrow D} \odot f_w(\xi, \hat{\mathcal{T}}_{S\leftarrow D}), \tag{4.11}$$

όπου $f_w(\cdot,\cdot)$ αναφέρεται στην οπισθο-στρέβλωση (back-warping) και το $\odot$ αναφέρεται στο γινόμενο Hadamard. Για τον υπολογισμό της μάσκας απόκλισης, προστίθεται ένα κανάλι στο τελευταίο επίπεδο του δικτύου πυκνής κίνησης. Τέλος, ο μετασχηματισμένος χάρτης χαρακτηριστικών εισάγεται στα επίπεδα δικτύου του γενετικού πεδίου που θα αναλύσουμε στην επόμενη ενότητα, για να αποδώσει την τελική επιθυμητή εικόνα-καρέ.

4.2.2 Γενετικό Πεδίο (Generation Module)

Στην συνέχεια της μεθοδολογίας και της αρχιτεκτονικής του δικτύου μας, εισάγεται ένα Ανταγωνιστικό Γενετικό Νευρωνικό Δίκτυο με γεννήτρια και διαχωριστή.

Γεννήτρια

Η γεννήτρια δέχεται ως είσοδο την εικόνα-πηγή και τις συντεταγμένες των χαρακτηριστικών σημείων του αντικειμένου της εικόνας-πηγής και του καρέ-οδηγού. Επιπλέον, εισάγονται οι έξοδοι του πυκνού πεδίου κίνησης, δηλαδή, οι μετασχηματισμοί σύμφωνα με τους οποίους θα μεταβληθούν οι συντεταγμένες των χαρακτηριστικών σημείων και η μάσκα απόκλισης που θα υποδείξει τα σημεία που πρέπει να ζωγραφιστούν. Η γεννήτρια περιγράφει ένα νευρωνικό δίκτυο που βασίζεται στην αρχιτεκτονική Johnson και περιλαμβάνει τρία μέρη:

- **Κωδικοποιητής (Μέρος Συρρίκνωσης Downsampling)**
  Ο κωδικοποιητής αποτελείται αρχικά από ένα απλό μπλοκ για διατήρηση χωρικής ανάλυσης με τρία διαδοχικά επίπεδα:

  1. Συνελικτικό Επίπεδο
  2. Επίπεδο Κανονικοποίησης Παρτίδας
  3. Επίπεδο ReLu

Στην συνέχεια, προστίθενται δύο μπλοκ συρρίκνωσης με 4 διαδοχικά επίπεδα:

1. Συνελικτικό Επίπεδο
2. Επίπεδο Κανονικοποίησης Παρτίδας
3. Επίπεδο ReLu
4. Επίπεδο Μέσου Pooling

Η έξοδος του κωδικοποιητή αποτελεί τον χάρτη χαρακτηριστικών ξ που αναλύσαμε σε προηγούμενη ενότητα.

- **Μετασχηματιστής**
  Ο μετασχηματιστής πραγματοποιεί σκέβρωση του χάρτη χαρακτηριστικών ξ σύμφωνα με τις μετατοπίσεις των συντεταγμένων των χαρακτηριστικών σημείων του αντικειμένου. Επιπλέον, μέσω του χάρτη απόκλισης υπολογίζει τον μετασχηματισμένο χάρτη χαρακτηριστικών ξ' και τροποποεί την πρόβλεψη σύμφωνα με αυτόν. Με άλλα λόγια σχεδιάζει τα τμήματα της εικόνας που δεν μπορούν να ανακτηθούν με απλή σκέβρωση. Έτσι, δημιουργεί τη νέα εικόνα με βάση αυτούς τους μετασχηματισμούς.

- **Αποκωδικοποιητής (Μέρος Επέκτασης Upsampling)**
  Στη συνέχεια, η μετασχηματισμένη πρόβλεψη εικόνας εισάγεται μέσα στον αποκωδικοποιητή ο οποίος στηρίζεται στο ακολουθιακό μοντέλο και αποτελείται από 6 residual μπλοκ για διατήριση χωρικής ανάλυσης με 3 διαδοχικά επίπεδα:

  1. Επίπεδο Κανονικοποίησης Παρτίδας
  2. Επίπεδο ReLu
  3. Συνελικτικό Επίπεδο

  Επιπλέον, εισάγονται 2 upsampling μπλοκ επέκτασης με 3 διαδοχικά επίπεδα:

  1. Συνελικτικό Επίπεδο
  2. Επίπεδο Κανονικοποίησης Παρτίδας
  3. Επίπεδο ReLu

  Τέλος, προστίθεται ένα ακόμα συνελικτικό επίπεδο και ένα τελικό σιγμοειδές επίπεδο (sigmoid).

Ως αποτέλεσμα, λαμβάνουμε την τελική εικόνα πρόβλεψης αναπαράστασης κίνησης του αντικειμένου. Στην πραγματικότητα, αυτή η εικόνα περιλαμβάνει το αντικείμενο της εικόνας-πηγής μετασχηματισμένο με τέτοιο τρόπο ώστε να έχει την ίδια στάση ή πόζα με το αντικείμενο του καρέ-οδηγού.

Διαχωριστής

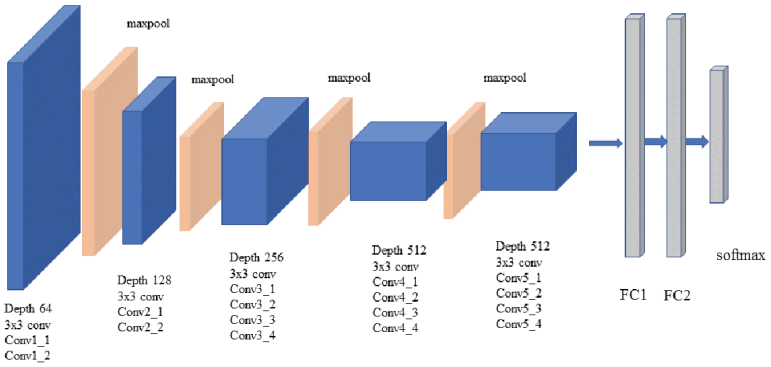
Η αρχιτεκτονική του διαχωριστή στηρίζεται στην έρευνα του Isola και της ομάδας του [22]. Ο διαχωριστής δημιουργεί χάρτες θερμότητας για τις γκαουσιανές αναπαραστάσεις των συντεταγμένων των χαρακτηριστικών σημείων του αντικειμένου. Έπειτα ενώνει την εικόνα πηγή με τους χάρτες θερμότητας και στην έξοδο αυτή προσθέτει 4 μπλοκ συρρίκνωσης με 4 διαδοχικά επίπεδα:

1. Συνελικτικό Επίπεδο
2. Επίπεδο Instance Κανονικοποίησης **(Instance Normalization)**
   Πρόκειται για άλλο ένα είδος κανονικοποίησης δεδομένων. Σε αυτήν την περίπτωση κανονικοποιείται κάθε δεδομένο της παρτίδας ξεχωριστά. ο μέσος όρος και η διακύμανση υπολογίζονται για κάθε μεμονωμένο κανάλι για κάθε μεμονωμένο δείγμα στις χωρικές διαστάσεις. Αντίθετα, στην κανονικοποίηση παρτίδας η μέση τιμή και η διακύμανση υπολογίζονται για κάθε μεμονωμένο κανάλι σε όλα τα δείγματα στις χωρικές διαστάσεις.
3. Επίπεδο leaky ReLU
4. Επίπεδο Μέσου Pooling

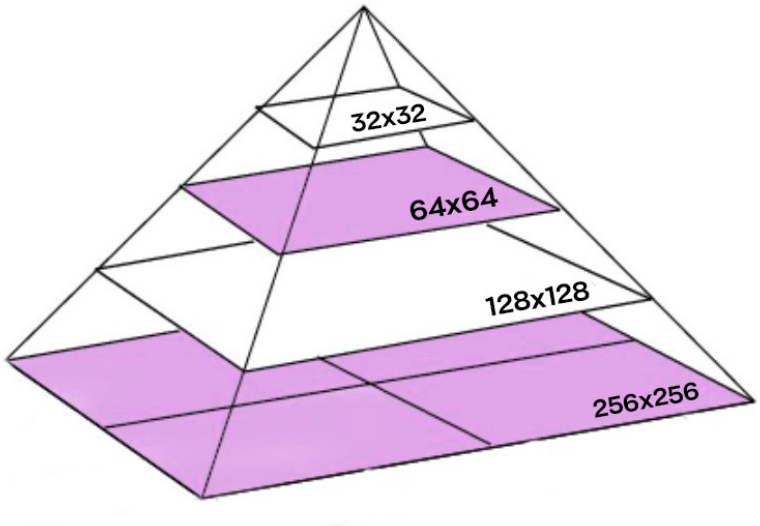
Τέλος προστίθεται ένα τελικό συνελικτικό επίπεδο και έτσι αποκτάμε το χάρτη πρόβλεψης. Ο σκοπός είναι να εκπαιδευτεί το ανταγωνιστικό δίκτυο με τέτοιο τρόπο ώστε να 'ξεγελάει' τον διαχωριστή και να αναγνωρίζει τις παραγόμενες εικόνες από την γεννήτρια ως αληθινές. Φυσικά, και η γεννήτρια πρέπει να εκπαιδευτεί όσο καλύτερα γίνεται ώστε να δημιουργεί πειστικές εικόνες που είναι σχεδόν πανομοιότυπες με τις εικόνες στόχου που έχουν οριστεί ως groundtruth. Στο Σχήμα 4.3 απεικονίζεται η αρχιτεκτονική του μοντέλου μας.

Συναρτήσεις Απώλειας

Για την εκπαίδευση ολόκληρου του δικτύου συνδυάζονται 4 είδη απώλειας:



Σχήμα 4.5: Αρχιτεκτονική Δικτύου VGG - 19 επιπέδων



Σχήμα 4.6: Πυραμίδα εικόνων με 4 διαφορετικές αναλύσεις

1. **Απώλεια Perceptual**

   Πρόκειται για την απώλεια ανακατασκευής που βασίζεται στην απώλεια perceptual του Johnson [32] ως κύρια καθοδηγήτρια απώλεια. Οι perceptual συναρτήσεις απώλειας στηρίζονται σε χαρακτηριστικά υψηλής τάξεως που διεξάγονται από προεκπαιδευμένα δίκτυα απωλειών που χρησιμοποιούνται για ταξινόμηση εικόνων. Αυτές οι perceptual συναρτήσεις απώλειας αποτελούν από μόνες τους βαθειά συνελικτικά νευρωνικά δίκτυα. Κατά τη διάρκεια της εκπαίδευσης οι perceptual απώλειες μετρούν ομοιότητες εικόνων οι οποίες είναι περισσότερο αξιόπιστες από ότι οι απώλειες που υπολογίζονται ανά εικονοστοιχείο. Στο δικό μας μοντέλο χρησιμοποιούμε το VGG δίκτυο 19-επιπέδων, βασισμένο στην υλοποίηση του Wang [20].

   Το δίκτυο VGG-19 είναι ένα μοντέλο συνελικτικού νευρωνικού δικτύου που

χρησιμοποιείται για ταξινόμηση εικόνων και η αρχιτεκτονική του απεικονίζεται στο Σχήμα 4.5[4]

Η απώλεια ανακατασκευής ορίζεται ως:

$$\mathcal{L}_{rec}(\hat{D}, D) = \sum_{i=1}^{I} ||N_i(\hat{D}) - N_i(D)||, \tag{4.12}$$

όπου D το καρέ-οδηγός εισόδου, $\hat{D}$ το αντίστοιχο ανακατασκευασμένο καρέ, $N_i(\cdot)$ το i-οστό κανάλι χαρακτηριστικών που έχει διεξαχθεί από ένα συγκεκριμένο VGG-19 επίπεδο και I ο αριθμός των καναλιών των χαρ/κών σε αυτό το επίπεδο. Επιπλέον, η απώλεια αυτή χρησιμοποιείται για μια σειρά από 4 διαφορετικές αναλύσεις εικόνας που σχηματίζουν μια πυραμίδα και λαμβάνονται μέσω συρρίκνωσης των D και $\hat{D}$. Οι κλίμακες που χρησιμοποιήθηκαν είναι 1, 0.5, 0.25 και 0.125. Αυτό σημαίνει ότι η απώλεια ανακατασκευής θα υπολογιστεί σε εικόνες που θα έχουν ανάλυση 256x256, 128x128, 64x64 και 32x32 αντίστοιχα. [33, 34], όπως φαίνεται και στο Σχήμα 4.6

2. **Απώλεια Ταιριάσματος Χαρακτηριστικών (Feature Matching)**
   Η απώλεια ταιριάσματος χαρ/κών βοηθάει στην σταθεροποίηση της εκπαίδευσης και βασίζεται στον διαχωριστή. Θεωρούμε δύο εικόνες I και I' που αποτελούν εικόνες καρέ που προέρχονται από το ίδιο βίντεο και $\hat{I}'$ η παραγόμενη εικόνα. Αυτή η απώλεια ενθαρρύνει την παραγόμενη εικόνα $\hat{I}'$ και το καρέ I' να έχουν παρόμοιες αναπαραστάσεις χαρ/κών. Οι αναπαραστάσεις χαρ/κών που αναπτύχθηκαν για να υπολογιστεί αυτή η απώλεια αποτελούν τα ενδιάμεσα επίπεδα του διαχωριστή D. Η απώλεια ταιριάσματος χαρακτηριστικών δίνεται από τη εξίσωση:

$$\mathcal{L}_{FM} = \mathbb{E}_{(I,I')}[||D_i(\hat{I}' \oplus H') - D_i(I' \oplus H')||_1], \tag{4.13}$$

   όπου D_i φανερώνει το i-οστό επίπεδο διεξαγωγής χαρακτηριστικών του διαχωριστή D. Η σημαντική διαφορά της απώλειας ταιριάσματος χαρακτηριστικών με τις perceptual απώλειες είναι πως δεν απαιτεί τη χρήση κάποιου εξωτερικού προεκπαιδευμένου δικτυου όπως για παράδειγμα το VGG-19 που χρησιμοποιήσαμε. Ως D_0 θεωρείται η είσοδος του διαχωριστή.

3. **Απώλεια Διαχωριστή**
   Ακολουθώντας τις πρόσφατες εξελίξεις στην αναπαραγωγή και δημιουργία

[4]Πηγή: https://morioh.com/p/383582dc31a6

εικόνων, συνδυάζουμε την ανταγωνιστική απώλεια και την απώλεια ταιριάσματος χαρ/κών ώστε να μάθουμε να ανακατασκευάσουμε το νέο καρέ. Πιο συγκεκριμένα, χρησιμοποιούμε το δίκτυο του διαχωριστή D που λαμβάνει ως είσοδο τον χάρτη θερμότητας H' των χαρακτηριστικών σημείων στην εικόνα I' συνενωμένο είτε με την αληθινή εικόνα $'$ είτε την παραγόμενη εικόνα $\hat{I}'$. Η γεννήτρια πρέπει να εκπαιδευτεί ώστε να είναι ικανή να ανακατασκευάσει το καρέ I' από τις συντεταγμένες των χαρακτηριστικών σημείων και την εικόνα I. Αναπτύσουμε τη φόρμουλα ενός ανταγωνιστικού δικτύου ελαχίστων τετραγώνων [29] που οδηγεί στην απώλεια που χρησιμοποιείται για την εκπαίδευση του διαχωριστή:

$$\mathcal{L}_{GAN}^{D}(D) = \mathbb{E}_{I'}[(1 - D(I' \oplus H'))^2] + \mathbb{E}_{(I,I')}[D(\hat{I}' \oplus H')^2], \tag{4.14}$$

όπου $\oplus$ φανερώνει την συνένωση κατά τον άξονα των καναλιών.

4. **Απώλεια Γεννήτριας**
   Παρόμοια, έχουμε την απώλεια που χρησιμοποιείται για την εκπαίδευση της γεννήτριας:
   $$\mathcal{L}_{GAN}^{G}(G) = \mathbb{E}_{(I,I')}[(D(\hat{I}' \oplus H') - 1)^2] \tag{4.15}$$

Τέλος, η συνολική απώλεια υπολογίζεται ως το άθροισμα των επιμέρους απωλειών:

$$\mathcal{L}_{tot} = \lambda\mathcal{L}_{rec} + \lambda\mathcal{L}_{FM} + \mathcal{L}_{GAN}^{G} + \mathcal{L}_{GAN}^{D}, \tag{4.16}$$

όπου $\lambda = 10$ σταθερός συντελεστής στα πειράματα μας σύμφωνα με [35].

Κεφαλαιο 5

Υλοποιηση και Αποτελεσματα

5.1 Λεπτομέρειες Υπολογισμών

5.1.1 Ρυθμίσεις Πειραμάτων

Όλα τα πειράματά μας πραγματοποιήθηκαν σε μια πλατφόρμα εξοπλισμένη με οκταπύρηνο επεξεργαστή AMD Ryzen 5 2400G με Radeon Vega Graphics με 15.6 GB μνήμης RAM σε 1517.709 MHh και μία κάρτα γραφικών GPU Nvidia Titan V με 12 GB μνήμης. Ο μέσος χρόνος για κάθε πείραμα ήταν περίπου 5 ημέρες για τα σετ δεδομένων ”Tai-Chi-HD” και ”Fashion Video” και 8 ημέρες για το σετ δεδομένων ”VoxCeleb”.

Για την εκπαίδευση του νευρωνικού δικτύου και συγκεκριμένα την εκπαίδευση της γεννήτριας και του διαχωριστή χρησιμοποιήθηκε βελτιστοποιητής Adam με ρυθμό εκπαίδευσης 2e-4 και μέγεθος παρτίδας (batch size) 8. Ο **Adam** είναι ένας αλγόριθμος βελτιστοποίησης που μπορεί να χρησιμοποιηθεί αντί του κλασικού αλγορίθμου Στοχαστικής Κλίσης Κατάβασης (Stochastic Gradient Descent) για την επαναληπτική ανανέωση των βαρών του δικτύου με βάση τα δεδομένα εκπαίδευσης. Η Στοχαστική Κλίση Κατάβασης διατηρεί ένα μόνο ρυθμό εκμάθησης (που ονομάζεται

alpha) για όλες τις ανανεώσεις βαρών και έτσι ο ρυθμός δεν μεταβάλλεται κατά τη διάρκεια της εκπαίδευσης. Αντίθετα, η μέθοδος βελτιστοποίησης Adam υπολογίζει μεμονωμένους προσαρμόσιμους ρυθμούς μάθησης για διαφορετικές παραμέτρους από εκτιμήσεις της πρώτης και δεύτερης φάσης των κλίσεων (gradients).

Συγκεκριμένα, ο αλγόριθμος υπολογίζει έναν εκθετικό μεταβαλλόμενο μέσο όρο της κλίσης και της τετραγωνικής κλίσης, και οι παράμετροι beta 1 και beta 2 ελέγχουν τους ρυθμούς καθυστέρησης (decay rate) αυτών των μεταβαλλόμενων μέσων όρων. Στα πειράματα μας, έχουμε εφαρμόσει τιμές beta1 = 0.5 και beta2 = 0.999 που αντιπροσωπέυουν τους εκθετικούς ρυθμούς καθυστέρησης για τις εκτιμήσεις της πρώτης φάσης και της δεύτερης φάσης, αντίστοιχα. Εφαρμόσαμε πτώση ρυθμού εκπαίδευσης κατά 10 φορές μετά από την 60η εποχή και την 90η εποχή (epoch) σε κάθε σετ δεδομένων. Οι αριθμοί των εποχών περιγράφονται παρακάτω αναλυτικά για κάθε σετ δεδομένων.

5.1.2 Σετ Δεδομένων

Εκπαιδεύσαμε και ελέγξαμε τρία διαφορετικά σετ δεδομένων. Όλα τα σετ δεδομένων έχουν υποστεί προεπεξεργασία και έχουν κοπεί ώστε οι διαστάσεις τους να είναι 256x256. Τα δύο σετ αναφέρονται σε ολόκληρα ανθρώπινα σώματα (Tai-Chi-HD και Fashion Video) και το VoxCeleb μόνο σε πρόσωπα.

Δεδομένα Taichi

Τα βίντεο του Tai-Chi-HD έχουν συλλεχθεί από το YouTube. Περιλαμβάνονται 252 βίντεο για εκπαίδευση και 28 βίντεο για έλεγχο. Ορισμένα βίντεο από αυτά έχουν χωριστεί σε διαφορετικά κλιπ και έτσι τελικά έχουμε αποκτήσει 1072 βίντεο εκπαίδευσης και 86 βίντεο ελέγχου όπου το μεγεθος τους κυμαίνεται μεταξύ 128 και 1024 καρέ. Για να ενισχυθεί το μέγεθος του σετ δεδομένων που προορίζεται για εκπαίδευση και για να έχουμε καλύτερη απόδοση εισόδου-εξόδου, πολλαπλασιάζουμε το σετ επί 150 φορές. Αυτό σημαίνει ότι κάθε βίντεο αντιγράφεται άλλες 150 φορές στο τελικό σετ εκπαίδευσης και έτσι συνολικά ο αριθμός των δεδομένων αυξάνεται σημαντικά και κάθε εποχή γίνεται πιο ισχυρή. Ο αριθμός των εποχών που έχουμε ορίσει για την εκπαίδευση του σετ είναι 150. Το σετ δεδομένων Tai-Chi-HD είναι δημοσίως διαθέσιμο για ερευνητική χρήση. Τα χαρακτηριστικά σημεία που έχουμε υπολογίσει είναι 13 και οι θέσεις τους φαίνονται στο Σχήμα 5.1.

Σχήμα 5.1: Θέσεις χαρακτηριστικών σημείων στο σετ δεδομένων Tai-Chi-HD

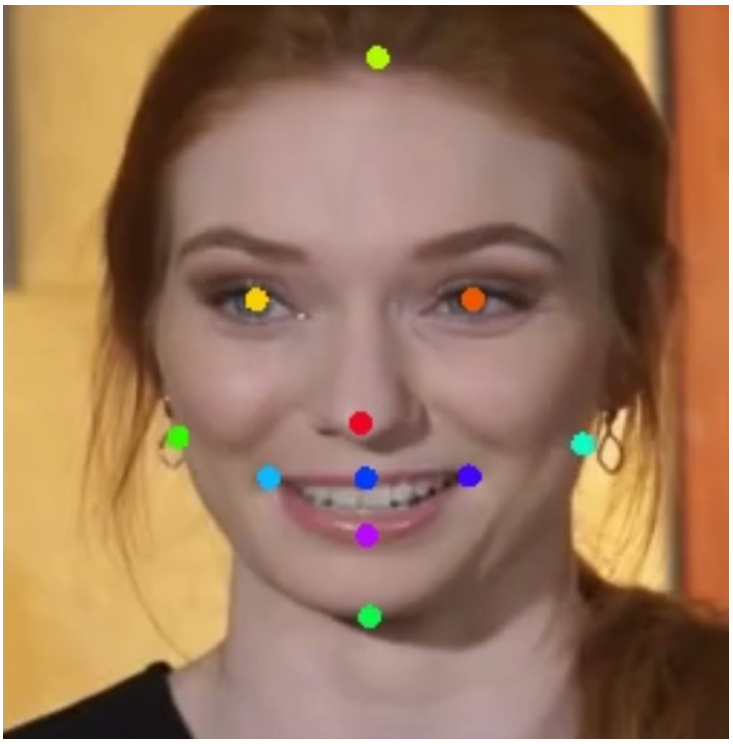
Δεδομένα Fashion

Το σετ Fashion Dataset εισήχθη αρχικά για την BMVC2019 δημοσίευση με τίτλο ”DwNet:Dense warp-based network for pose-guided human video generation”. Περιλαμβάνονται 499 βίντεο για εκπαίδευση και 99 βίντεο για έλεγχο. Για να ενισχυθεί το μέγεθος του σετ δεδομένων που προορίζεται για εκπαίδευση και για να έχουμε καλύτερη απόδοση εισόδου-εξόδου, πολλαπλασιάζουμε το σετ επί 50 φορές. Αυτό σημαίνει ότι κάθε βίντεο αντιγράφεται άλλες 50 φορές στο τελικό σετ εκπαίδευσης και έτσι συνολικά ο αριθμός των δεδομένων αυξάνεται σημαντικά και κάθε εποχή γίνεται πιο ισχυρή. Ο αριθμός των εποχών που έχουμε ορίσει για την εκπαίδευση του σετ είναι 100. Το σετ δεδομένων Fashion Video είναι δημοσίως διαθέσιμο για ερευνητική χρήση. Τα χαρακτηριστικά σημεία που έχουμε υπολογίσει είναι 13 και οι θέσεις τους φαίνονται στο Σχήμα 5.2.

Σχήμα 5.2: Θέσεις χαρακτηριστικών σημείων στο σετ δεδομένων Fashion Video

Δεδομένα VoxCeleb

Τα βίντεο του VoxCeleb έχουν συλλεχθεί από το YouTube και πρόκειται για σετ δεδομένων με 22496 βίντεο. Ορισμένα βίντεο από αυτά έχουν χωριστεί σε διαφορετικά κλιπ και έτσι τελικά έχουμε αποκτήσει 831 βίντεο εκπαίδευσης και 506 βίντεο ελέγχου όπου το μεγεθος τους κυμαίνεται μεταξύ 128 και 1024 καρέ. Για να ενισχυθεί το μέγεθος του σετ δεδομένων που προορίζεται για εκπαίδευση και για να έχουμε καλύτερη απόδοση εισόδου-εξόδου, πολλαπλασιάζουμε το σετ επί 75 φορές. Αυτό σημαίνει ότι κάθε βίντεο αντιγράφεται άλλες 75 φορές στο τελικό σετ εκπαίδευσης και έτσι συνολικά ο αριθμός των δεδομένων αυξάνεται σημαντικά και κάθε εποχή γίνεται πιο ισχυρή. Ο αριθμός των εποχών που έχουμε ορίσει για την εκπαίδευση του σετ είναι 150. Το σετ δεδομένων VoxCeleb είναι δημοσίως διαθέσιμο για ερευνητική χρήση. Τα χαρακτηριστικά σημεία που έχουμε υπολογίσει είναι 11 και οι θέσεις τους φαίνονται στο Σχήμα 5.3.



Σχήμα 5.3: Θέσεις χαρακτηριστικών σημείων στο σετ δεδομένων VoxCeleb

5.1.3 Μετρικές Υπολογισμών

Ο υπολογισμός της ποιότητας των εικόνων-καρέ δεν είναι προφανής καθώς δεν υπάρχουν groundtruth δεδομένα ώστε να υπάρχει σύγκριση. Ακολουθείται το πρωτόκολλο υπολογισμού που χρησιμοποιεί ο Siarohin [26]. Στην περίπτωση μας, ανακατασκευάζουμε τα βίντεο εισόδου συνδυάζοντας το πρώτο καρέ των βίντεο με τις αναπαραστάσεις των κινήσεων σε κάθε επόμενο καρέ. Με άλλα λόγια, αν υποθέσουμε ότι ένα βίντεο αποτελείται από 100 καρέ, ανακατασκευάζουμε το πρώτο καρέ σύμφωνα με τις κινήσεις των καρέ που ακολουθούν, θεωρώντας κάθε ένα από αυτά τα επόμενα 99 καρέ ως groundtruth. Άρα, ο σκοπός μας είναι να φτιάξουμε από το πρώτο καρέ, άλλα 99 καινούρια καρέ που θα είναι σχεδόν πανομοιότυπα με τα αντίστοιχα groundtruth τους. Οι μετρικές που χρησιμοποιήθηκαν είναι:

- **Απώλεια εικονοστοιχείου ή L1**
  Είναι η απώλεια που βασίζεται σε χαμηλού επιπέδου πληροφορίες των εικονοστοιχείων. Πρόκειται για την κανονικοποιημένη ευκλείδια απόσταση μεταξύ της τελικής εικόνας πρόβλεψης $\hat{y}$ και της εικόνας στόχου y. Εάν έχουν και οι

δύο το ίδιο σχήμα CxHxW τότε:

$$L_1 = \lambda_{pixel}(\hat{y}, y) = ||\hat{y} - y||^2{}_2/CHW \tag{5.1}$$

Αυτή η απώλεια χρησιμοποιείται μόνο όταν έχουμε groundtruth πληροφορία. Στη συγκεκριμένη περίπτωση επειδή δεν υπάρχουν τέτοια δεδομένα, χρησιμοποιείται ως groundtruth κάθε καρέ κάθε βίντεο και ο στόχος είναι να ανακατασκευαστεί το πρώτο καρέ κάθε βίντεο σύμφωνα με κάθε επόμενο καρέ του ίδιου βίντεο που λειτουργεί ως groundtruth.

- **Μέση Ευκλείδια Απόσταση AED**
  Η Μέση Ευκλείδια Απόσταση μεταξύ groundtruth και αναπαράστασης παραγόμενου καρέ. Υπολογίζεται αν η κίνηση είναι ίδια με αυτή του βίντεο στόχου και αν διατηρείται η ταυτότητα του αντικειμένου.

- **Δείκτης Ομοιότητας (Structural Similarity Index)**
  Ο δείκτης αυτός μετρά στην πραγματικότητα την αντιληπτική διαφορά μεταξύ δύο παρόμοιων εικόνων. Ο δείκτης SSIM μπορεί να θεωρηθεί ως ποιοτικό μέτρο, με την προϋπόθεση ότι η άλλη εικόνα θεωρείται άριστης ποιότητας και ανάλυσης. Υπολογίζει την υποβάθμιση της ανάλυσης της εικόνας που προκαλείται από επεξεργασία, όπως συμπίεση δεδομένων ή από απώλειες στη μετάδοση δεδομένων. Πρόκειται για μια πλήρη μέτρηση αναφοράς που απαιτεί δύο εικόνες από την ίδια λήψη - μια εικόνα αναφοράς και μια επεξεργασμένη εικόνα. Ο δείκτης αυτός είναι πιο γνωστός στον χώρο παραγωγής βίντεο, αλλά έχει ισχυρές εφαρμογές και στην φωτογραφία. Τιμή του δείκτη κοντά στην μονάδα φανερώνει πως οι δύο εικόνες είναι σχεδόν πανομοιότυπες. Ο υπολογισμός του αποτελείται από έναν συνδυασμό 3 συνιστωσών, της φωτεινότητας (luminance), της αντίθεσης (contrast) και της δομής (structure). Έχουμε:

  1. **Φωτεινότητα**
  $$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \tag{5.2}$$
  όπου x, y εικόνες ίδιας διάστασης, μ_x, μ_y μέσες τιμές των x και y αντίστοιχα και $c_1 = 0.01L$, L το δυναμικό εύρος των τιμών των εικονοστοιχείων.

  2. **Αντίθεση**
  $$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{5.3}$$

όπου σ_x, σ_y οι διακυμάνσεις των x και y αντίστοιχα και $c_2 = 0.03L$, L το δυναμικό εύρος των τιμών των εικονοστοιχείων.

3. **Δομή**

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x \sigma_y + c_3}, \tag{5.4}$$

όπου σ_{xy} η συνδιακύμανση των x και y και $c_3 = c_2/2$

Τελικά:

$$SSIM(x, y) = [l(x, y)c(x, y)s(x, y)] \tag{5.5}$$

5.2 Αποτελέσματα και Μελέτη Ablation

Συγκρίνουμε τις ακόλουθες διαφορετικές εκδοχές του μοντέλου μας, όπως και στο [26].

- **Baseline**
  Πρόκειται για το πιο απλό μοντέλο που εκπαιδεύεται χωρίς τη μάσκα απόκλισης ($\mathcal{O}_{S \leftarrow D} = 1$ στην εξίσωση 4.11) και με απώλεια ανακατασκευής $\mathcal{L}_{rec}$ μόνο στην μέγιστη ανάλυση εικόνας (256x256).

- **Baseline With Occlusion Mask**
  Πρόκειται για το ίδιο μοντέλο με το Baseline, μόνο που τώρα έχει προστεθεί και η μάσκα απόκλισης.

- **Pyramid**
  Πρόκειται για το ίδιο μοντέλο με το Baseline, μόνο που τώρα έχουν προστεθεί και οι 4 διαφορετικές αναλύσεις εικόνας (256x256, 128x128, 64x64, 32x32).

- **Full**
  Πρόκειται για το πλήρες μοντέλο που περιλαμβάνει μάσκα απόκλισης και πυραμίδα 4 αναλύσεων εικόνας.

Γκαουσιανό φιλτράρισμα των βίντεο στον τομέα του χωρο-χρόνου

Για να βελτιώσουμε κυρίως τα οπτικά αποτελέσματα των πειραμάτων μας έχουμε εφαρμόσει μία ακόμη εκδοχή στα πειράματα μας ως διαδικασία μεταεπεξεργασίας

των βίντεο. Η διαδικασία αυτή είναι ανεξάρτητη από την εκπαίδευση του νευρωνικού δικτύου και εφαρμόζεται στο τέλος. Έτσι, χρησιμοποιήσαμε ένα χαμηλοπερατό γκαουσιανό φίλτρο (low-pass spatio-temporal filter) στα βίντεο στον άξονα του χρόνου:

$$\mathcal{G}(x, y, t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x^2+y^2+t^2)}{2\sigma^2}}, \tag{5.6}$$

όπου σ η τυπική απόκλιση. Στην πραγματικότητα, η εφαρμογή του φίλτρου γίνεται μέσω συνέλιξης της ακολουθίας των τιμών κάθε εικονοστοιχείου ξεχωριστά από όλα τα καρέ (δηλαδή σε βάθος χρόνου). Όσο μεγαλύτερο είναι το σ και το μέγεθος του φίλτρου τόσο πιο έντονη θα είναι και η επεξεργασία και το θόλωμα της κίνησης. Το γκαουσιανό φίλτρο αφαιρεί τους θορύβους και θολώνει τις κινήσεις του αντικειμένου μελέτης καθώς περνάει ο χρόνος. Προσφέρει μια πιο μαλακή κίνηση στα αντικείμενα των βίντεο και λειτουργεί ως σταθεροποιητής κίνησης εξαλείφοντας το ''τρέμουλο'' της εικόνας. Στα παρακάτω σχήματα φαίνονται ενδεικτικά μερικές διαφορετικές τιμές των σ και των μεγεθών φίλτρου που δοκιμάσαμε και τα αντίστοιχα αποτελέσματα τους στο σετ δεδομένων Tai-Chi-HD.

Σχήμα 5.4: Καρέ με εφαρμογή χαμηλοπερατού γκαουσιανού φίλτρου με μέγεθος φίλτρου $f = 3$ και $\sigma = 2.5$

Σχήμα 5.5: Καρέ με εφαρμογή χαμηλοπερατού γκαουσιανού φίλτρου με μέγεθος φίλτρου $f = 10$ και $\sigma = 2.5$

Σχήμα 5.6: Καρέ με εφαρμογή χαμηλοπερατού γκαουσιανού φίλτρου με μέγεθος φίλτρου $f = 20$ και $\sigma = 5$

Παρατηρούμε πως με μεγάλες τιμές μεγέθους φίλτρου και σ το θόλωμα στο χέρι του άνδρα είναι ακόμη πιο έντονο. Τα αποτελέσματα είναι εμφανή κατά την αναπαραγωγή των βίντεο, όπου αποτυπώνεται κανονικά η κίνηση. Η καλύτερη επιλογή ήταν το μέγεθος φίλτρου = 3 και $\sigma = 2.5$ με τα οποία καταφέραμε σταθεροποίηση

της εικόνας και των κινήσεων με ελάχιστο θόλωμα.

5.2.1 Σετ Δεδομένων Tai-Chi-HD

Στους επόμενους πίνακες ως Gaussian εκδοχή εννοούμε την εφαρμογή γκαουσιανού φίλτρου με μέγεθος φίλτρου $f = 3$ και $\sigma = 2.5$

Πίνακας 5.1: Αποτελέσματα δεδομένων Tai-Chi-HD με και χωρίς γκαουσιανό φίλτρο.

	L_1	AED	SSIM
Gaussian	0.0563	0.1550	0.7650
Non-Gaussian	0.0579	0.1552	0.7546

Πίνακας 5.2: **Μελέτη Ablation** για δεδομένα Tai-Chi-HD με γκαουσιανό φίλτρο

Με Gaussian φίλτρο			
Model	L_1	**AED**	**SSIM**
Baseline	0.0591	0.1579	0.7626
Baseline + Occl.	0.0575	0.1553	0.7665
Pyramid	0.0566	0.1561	0.761
Full	0.0563	0.1550	0.7650

Πίνακας 5.3: **Μελέτη Ablation** για δεδομένα Tai-Chi-HD χωρίς γκαουσιανό φίλτρο

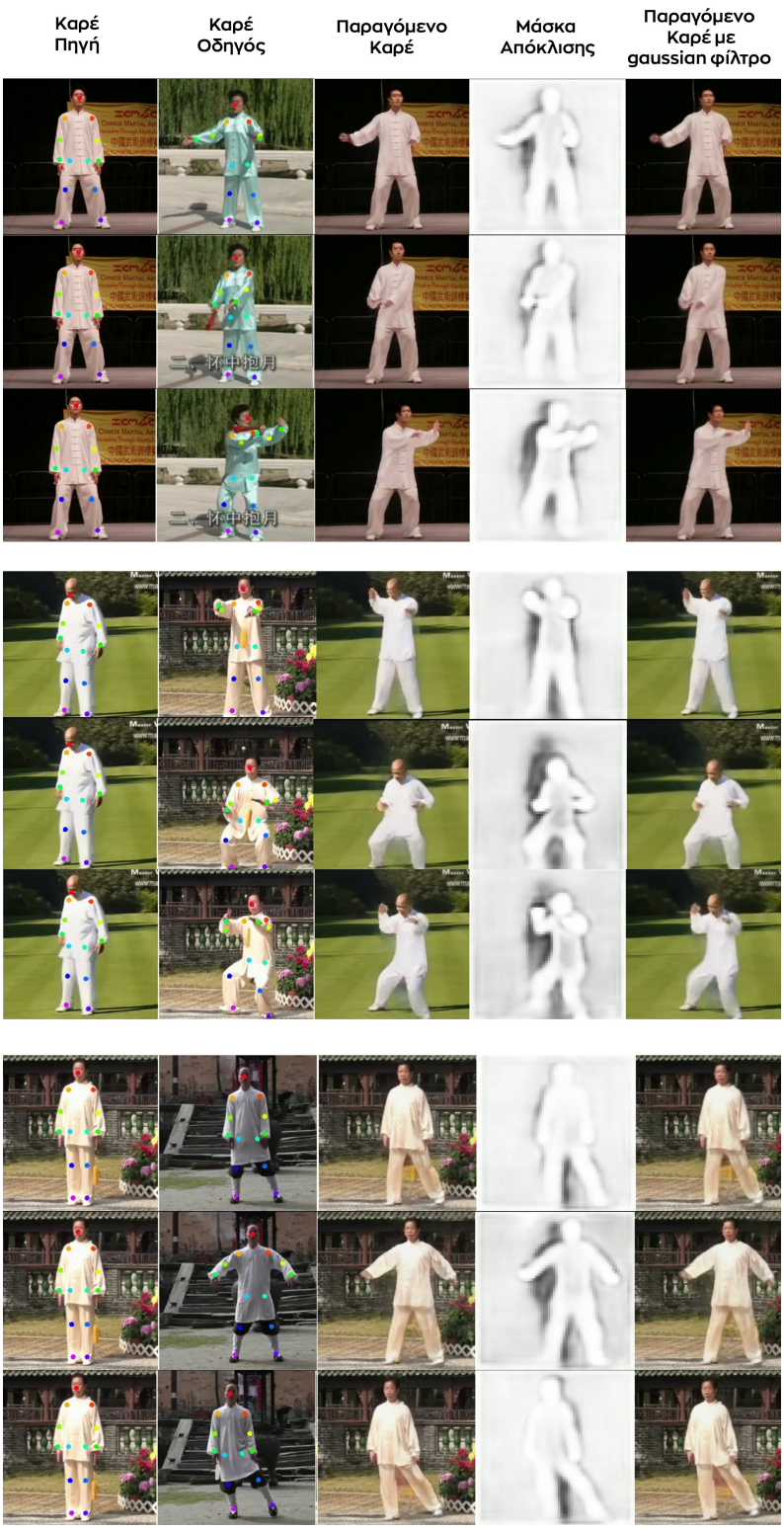
Χωρίς Gaussian Φίλτρο			
Model	L_1	**AED**	**SSIM**
Baseline	0.0607	0.1582	0.7525
Baseline + Occl.	0.0591	0.1564	0.7567
Pyramid	0.0583	0.154	0.7498
Full	0.0579	0.1552	0.7546

Παρατηρούμε από την ablation μελέτη πως τα καλύτερα αποτελέσματα για αυτό το σετ δεδομένων παρέχει το πλήρες μοντέλο, αν και οι διαφορές μεταξύ των 4 εκδοχών είναι αρκετά μικρές. Οι καλύτερες τιμές βρέθηκαν στην γκαουσιανή εκδοχή

του πλήρους μοντέλου με $L_1 = 0.0563, AED = 0.1550, SSIM = 0.7650$. Στον πίνακα 5.4 καταγράφονται οι τιμές των 4 απωλειών κατά τη διάρκεια της εκπαίδευσης σε διαφορετικές εποχές (epochs). Η βασική καθοδηγήτρια απώλεια perceptual μειώνεται με την πάροδο των εποχών, η απώλεια της γεννήτριας αυξάνεται και του διαχωριστή μειώνεται, όπως ακριβώς έχουμε αναλύσει και στην θεωρία των ανταγωνιστικών δικτύων.

Πίνακας 5.4: **Τιμές απωλειών** κατά τη διάρκεια της εκπαίδευσης του πλήρους μοντέλου (Full) του σετ δεδομένων Tai-Chi-HD

Epoch	Perceptual	Feature Matching	Generator Loss	Discriminator Loss
1	139.8005	2	0.4580	0.3489
50	87.8005	1.5883	0.6745	0.2186
100	81.3353	1.6736	0.7325	0.1812
150	77.6425	1.7175	0.7552	0.1660



Σχήμα 5.7: Ποιοτικά Αποτελέσματα Πλήρους μοντέλου για Tai-Chi-HD

Στον πίνακα 5.7 απεικονίζονται τα ποιοτικά αποτελέσματα σε 3 διαφορετικές χρονικές στιγμές για τρία βίντεο. Τα μαύρα σημεία στην μάσκα απόκλισης φανερώνουν τα τμήματα που δεν μπορούν να ανακτηθούν από την εικόνα-πηγή και πρέπει

να ζωγραφιστούν.

5.2.2 Σετ Δεδομένων Fashion Video

Πίνακας 5.5: Αποτελέσματα δεδομένων Fashion Video με και χωρίς γκαουσιανό φίλτρο.

	L_1	AED	SSIM
Gaussian	0.0226	0.0988	0.9238
Non-Gaussian	0.0248	0.097	0.9152

Πίνακας 5.6: **Μελέτη Ablation** για δεδομένα Fashion Video με γκαουσιανό φίλτρο

Με Gaussian Φίλτρο			
Model	L_1	**AED**	**SSIM**
Baseline	0.0249	0.103	0.9216
Baseline + Occl.	0.0246	0.0995	0.9228
Pyramid	0.0241	0.099	0.9232
Full	0.0226	0.0988	0.9238

Πίνακας 5.7: **Μελέτη Ablation** για δεδομένα Fashion Video χωρίς γκαουσιανό φίλτρο

Χωρίς Gaussian Φίλτρο			
Model	L_1	**AED**	**SSIM**
Baseline	0.0265	0.12	0.9125
Baseline + Occl.	0.0262	0.116	0.9137
Pyramid	0.0257	0.098	0.9143
Full	0.0248	0.097	0.9152

Παρατηρούμε από την ablation μελέτη πως τα καλύτερα αποτελέσματα για αυτό το σετ δεδομένων παρέχει το πλήρες μοντέλο, αν και οι διαφορές μεταξύ των 4 εκδοχών είναι αρκετά μικρές. Οι καλύτερες τιμές βρέθηκαν στην γκαουσιανή εκδοχή του πλήρους μοντέλου με $L_1 = 0.0226, AED = 0.0988, SSIM = 0.9238$. Η τιμή του

δείκτη SSIM πλησιάζει σχεδόν τη μονάδα, γεγονός που σημαίνει πως οι παραγόμενες εικόνες είναι σχεδόν πανομοιότυπες με τις εικόνες στόχου. Στον πίνακα 5.8 καταγράφονται οι τιμές των 4 απωλειών κατά τη διάρκεια της εκπαίδευσης σε διαφορετικές εποχές (epochs). Η βασική καθοδηγήτρια απώλεια perceptual μειώνεται με την πάροδο των εποχών, η απώλεια της γεννήτριας αυξάνεται και του διαχωριστή μειώνεται, όπως είδαμε και προηγουμένως.

Πίνακας 5.8: **Τιμές απωλειών** κατά τη διάρκεια της εκπαίδευσης του πλήρους μοντέλου (Full) του σετ δεδομένων Fashion Video

Epoch	Perceptual	Feature Matching	Generator Loss	Discriminator Loss
1	87.7881	3.2088	0.5764	0.2683
50	48.1505	1.5899	0.8925	0.07
100	41.8825	1.6063	0.9973	0.0019

Σχήμα 5.8: Ποιοτικά Αποτελέσματα Πλήρους μοντέλου για Fashion Video

Στον πίνακα 5.8 απεικονίζονται τα ποιοτικά αποτελέσματα σε 3 διαφορετικές χρονικές στιγμές για τρία βίντεο.

5.2.3 Σετ Δεδομένων VoxCeleb

Πίνακας 5.9: Αποτελέσματα δεδομένων VoxCeleb με και χωρίς γκαουσιανό φίλτρο.

	L_1	AED	SSIM
Gaussian	0.0540	0.1731	0.8275
Non-Gaussian	0.0565	0.1735	0.8071

Πίνακας 5.10: **Μελέτη Ablation** για δεδομένα VoxCeleb με γκαουσιανό φίλτρο

Με Gaussian Φίλτρο			
Model	L_1	**AED**	**SSIM**
Baseline	0.0562	0.1740	0.8197
Baseline + Occl.	0.0551	0.1733	0.8277
Pyramid	0.0559	0.1738	0.8234
Full	0.054	0.1731	0.8275

Πίνακας 5.11: **Μελέτη Ablation** για δεδομένα VoxCeleb με γκαουσιανό φίλτρο

Χωρίς Gaussian Φίλτρο			
Model	L_1	**AED**	**SSIM**
Baseline	0.0577	0.1752	0.794
Baseline + Occl.	0.0569	0.1746	0.8032
Pyramid	0.0571	0.175	0.801
Full	0.0565	0.1735	0.8071

Παρατηρούμε από την ablation μελέτη πως τα καλύτερα αποτελέσματα για αυτό το σετ δεδομένων παρέχει το πλήρες μοντέλο, αν και οι διαφορές μεταξύ των 4 εκδοχών είναι αρκετά μικρές. Οι καλύτερες τιμές βρέθηκαν στην γκαουσιανή εκδοχή του πλήρους μοντέλου με $L_1 = 0.054, AED = 0.1731, SSIM = 0.8275$. Η τιμή του δείκτη SSIM είναι αρκετά υψηλή, γεγονός που σημαίνει πως και σε αυτό το σετ δεδομένων οι παραγόμενες εικόνες είναι σχεδόν πανομοιότυπες με τις εικόνες στόχου. Στον πίνακα 5.12 καταγράφονται οι τιμές των 4 απωλειών κατά τη διάρκεια

της εκπαίδευσης σε διαφορετικές εποχές (epochs). Η βασική καθοδηγήτρια απώλεια perceptual μειώνεται με την πάροδο των εποχών, η απώλεια της γεννήτριας αυξάνεται και του διαχωριστή μειώνεται, όπως είδαμε και προηγουμένως.

Πίνακας 5.12: **Τιμές απωλειών** κατά τη διάρκεια της εκπαίδευσης του πλήρους μοντέλου (Full) του σετ δεδομένων VoxCeleb

Epoch	Perceptual	Feature Matching	Generator Loss	Discriminator Loss
1	131.4044	2.0965	0.4645	0.3505
50	87.9588	2.2603	0.7331	0.1780
100	81.0594	2.3691	0.7939	0.1391
150	77.4816	2.53	0.8504	0.1028

Σχήμα 5.9: Ποιοτικά Αποτελέσματα Πλήρους μοντέλου για VoxCeleb

Στον πίνακα 5.9 απεικονίζονται τα ποιοτικά αποτελέσματα σε 3 διαφορετικές χρονικές στιγμές για τρία βίντεο. Τα μαύρα σημεία στην μάσκα απόκλισης φανερώνουν τα τμήματα που δεν μπορούν να ανακτηθούν από την εικόνα-πηγή και πρέπει

να ζωγραφιστούν.

5.3 Σύγκριση Αποτελεσμάτων και Τελικά Συμπεράσματα

Πίνακας 5.13: Σύγκριση Μοντέλων

	Tai-Chi-HD		**VoxCeleb**	
	L_1	**AED**	L_1	**AED**
Monkey-Net [25]	0.077	0.228	0.049	0.199
FOMM [26]	0.063	0.179	0.043	0.140
Ours	0.056	0.155	0.054	0.173

Σύμφωνα με τον πίνακα 5.13 παρατηρούμε πως το πλήρες μοντέλο μας ανταποκρίνεται καλύτερα από το "FOMM First Order Motion Model" [26] σε ότι αφορά το σετ δεδομένων Tai-Chi-HD που περιλαμβάνει κινήσεις ολόκληρου του ανθρώπινου σώματος. Συγκεκριμένα, τόσο η απώλεια εικονοστοιχείου L_1 όσο και και η Μέση Ευκλείδια απόσταση AED που υπολογίσαμε είναι μικρότερες από τις αντίστοιχες τιμές του FOMM και του Monkey-Net. Όσον αφορά το σετ δεδομένων προσώπων τα αποτελέσματα του FOMM ήταν ελάχιστα καλύτερα από τα δικά μας, χωρίς αυτό όμως να σημαίνει πως τα αποτελέσματα μας δεν ήταν αρκετά ικανοποιητικά. Τα οπτικά αποτελέσματα των παραπάνω πινάκων απέδειξαν πως το μοντέλο μας μπορεί να πραγματοποιήσει αναπαράσταση κίνησης με αξιόλογο αποτέλεσμα. Αποδείξαμε έτσι ότι μπορούμε με ένα αρκετά πιο απλοποιημένο μοντέλο από το FOMM να συνθέσουμε βίντεο με αναπαραστάσεις κινήσεων. Αξίζει να σημειωθεί πως συγκρητικά με το μοντέλο FOMM, το δικό μας δεν περιείχε κανένα είδος δικτύου διεξαγωγής χαρακτηριστικών σημείων που να χρειάστηκε εκπαίδευση για αυτό το σκοπό. Η εκπαίδευση περισορίστηκε μόνο στο ανταγωνιστικό δίκτυο και έτσι η συνολική πολυπλοκότητα του αλγορίθμου μας μειώθηκε αρκετά. Επιπλέον, ο Siarohin και η ομάδα του υπολόγισαν τους τοπικούς μετασχηματισμούς των κινήσεων όχι μόνο στις θέσεις των χαρακτηριστικών σημείων, αλλά και στην γειτονιά αυτών, χρησιμοποιώντας Ιακωβιανούς πίνακες (Jacobians). Στο δικό μας μοντέλο δεν χρησιμοποιήθηκαν ούτε υπολογίστηκαν πίνακες Τζακόμπι και βασιστήκαμε μόνο στις θέσεις των συντεταγμένων των χαρακτηριστικών σημείων (keypoints). Διαπι-

στώσαμε επίσης, πως με την εφαρμογή του χαμηλοπερατού γκαουσιανού φίλτρου στον τομέα του χρόνου, τα αποτελέσματα βελτιώθηκαν ακόμη περισσότερο, τόσο ποσοτικά όσο και ποιοτικά. Τα βίντεο σταθεροποιήθηκαν και απέκτησαν πιο λεία και ομαλή κίνηση χωρίς τρεμούλιασμα κινήσεων.

Βιβλιογραφια

[1] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241, (available on arXiv:1505.04597 [cs.CV]). [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[3] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2849–2858.

[4] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/04025959b191f8f9de3f924f0940515f-Paper.pdf

[5] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1526–1535.

[7] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/d9d4f495e875a2e075a1a4a6e1b9770f-Paper.pdf

[8] J. R. van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," *CoRR*, vol. abs/1701.08435, 2017. [Online]. Available: http://arxiv.org/abs/1701.08435

[9] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-Conditional Video Prediction using Deep Networks in Atari Games," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2845–2853.

[10] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 843–852. [Online]. Available: http://proceedings.mlr.press/v37/srivastava15.html

[11] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. Metaxas, "Learning to forecast and refine residual motion for image-to-video generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[12] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," *CoRR*, vol. abs/1710.11252, 2017. [Online]. Available: http://arxiv.org/abs/1710.11252

[13] C. Cao, Q. Hou, and K. Zhou, "K.: Displaced dynamic expression regression for real-time facial tracking and animation," in *In ACM TOG (Proc. SIGGRAPH) (2014*.

[14] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," *CoRR*, vol. abs/1902.08900, 2019. [Online]. Available: http://arxiv.org/abs/1902.08900

[15] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] E. Zakharov, A. Shysheya, E. Burkov, and V. S. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," *CoRR*, vol. abs/1905.08233, 2019. [Online]. Available: http://arxiv.org/abs/1905.08233

[17] M. Zollhöfer, J. Thies, D. Bradley, P. Garrido, T. Beeler, P. Péerez, M. Stamminger, M. Nießner, and C. Theobalt, "State of the art on monocular 3d face reconstruction, tracking, and applications," 2018.

[18] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," *CoRR*, vol. abs/1808.07371, 2018. [Online]. Available: http://arxiv.org/abs/1808.07371

[19] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, A. Vakhitov, and V. Lempitsky, "Textured neural avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/d86ea612dec96096c5e0fcc8dd42ab6d-Paper.pdf

[21] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "Gesturegan for hand gesture-to-gesture translation in the wild," *CoRR*, vol. abs/1808.04859, 2018. [Online]. Available: http://arxiv.org/abs/1808.04859

[22] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[23] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," *CoRR*, vol. abs/1808.05174, 2018. [Online]. Available: http://arxiv.org/abs/1808.05174

[24] O. Wiles, A. S. Koepke, and A. Zisserman, "X2face: A network for controlling face generation by using images, audio, and pose codes," *CoRR*, vol. abs/1807.10550, 2018. [Online]. Available: http://arxiv.org/abs/1807.10550

[25] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," *CoRR*, vol. abs/1812.08861, 2018. [Online]. Available: http://arxiv.org/abs/1812.08861

[26] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf

[27] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks through conditional image generation," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/1f36c15d6a3d18d52e8d493bc8187cb9-Paper.pdf

[28] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Learning to forecast and refine residual motion for image-to-video generation," *CoRR*, vol. abs/1807.09951, 2018. [Online]. Available: http://arxiv.org/abs/1807.09951

[29] H. Fang, S. Xie, and C. Lu, "RMPE: regional multi-person pose estimation," *CoRR*, vol. abs/1612.00137, 2016. [Online]. Available: http://arxiv.org/abs/1612.00137

[30] J. Sánchez, A. Nuez, and N. Monzón, "Computing inverse optical flow," *CTIM Technical Report*, 01 2013.

[31] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. S. Lempitsky, "Coordinate-based texture inpainting for pose-guided image generation," *CoRR*, vol. abs/1811.11459, 2018. [Online]. Available: http://arxiv.org/abs/1811.11459

[32] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016. [Online]. Available: http://arxiv.org/abs/1603.08155

[33] H. Tang, D. Xu, W. Wang, Y. Yan, and N. Sebe, "Dual generator generative adversarial networks for multi-domain image-to-image translation," *CoRR*, vol. abs/1901.04604, 2019. [Online]. Available: http://arxiv.org/abs/1901.04604

[34] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.

[35] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *CoRR*, vol. abs/1711.11585, 2017. [Online]. Available: http://arxiv.org/abs/1711.11585

Σύντομο Βιογραφικο

Η Τάγκα Βιργινία γεννήθηκε στα Ιωάννινα το 1995 και έχει καταγωγή από το Κουκούλι Ζαγορίου. Είναι πτυχιούχος του τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων από το οποίο αποφοίτησε το 2018. Το 2019 ξεκίνησε τις μεταπτυχιακές σπουδές της στο τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής του Πανεπιστημίου Ιωαννίνων, υπό την επίβλεψη του καθηγητή Χριστόφορου Νίκου. Τα ενδιαφέροντα της σχετίζονται με την επεξεργασία εικόνας, τη γραφιστική σχεδίαση καθώς επίσης και με τη στατιστική ανάλυση δεδομένων.