Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
τομεας σηματων, ελεγχου και ρομποτικης

3Δ Ανακατασκευή Ανθρώπινου Σχήματος και Πόζας απο Τρικαναλικές Εικόνες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Σωτήριου Καραπιπέρη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Συν-επιβλέπων: Δρ. Παναγιώτης Φιλντίσης
Μεταδιδακτορικός Ερευνητής Ε.Μ.Π.

Συν-επιβλέπων: Δρ Γεώργιος Ρετσινάς
Μεταδιδακτορικός Ερευνητής Ε.Μ.Π.

ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ

Αθήνα, Οκτώβριος 2023

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

3Δ Ανακατασκευή Ανθρώπινου Σχήματος και Πόζας απο Τρικαναλικές Εικόνες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Σωτήριου Καραπιπέρη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Συν-επιβλέπων: Δρ. Παναγιώτης Φιλντίσης
Μεταδιδακτορικός Ερευνητής Ε.Μ.Π.

Συν-επιβλέπων: Δρ Γεώργιος Ρετσινάς
Μεταδιδακτορικός Ερευνητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20^{η} Οκτωβρίου, 2023.

.....	.....	.....
Πέτρος Μαραγκός	Γεράσιμος Ποταμιάνος	Κών/νος Τζαφέστας
Καθηγητής Ε.Μ.Π.	Αν. Καθ. Πάν/μιο Θεσσαλίας.	Αν. Καθ. Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

.....

Σωτηριος Καραπιπερης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Ένα απο τα σημαντικότερα προβλήματα της όρασης υπολογιστών αποτελεί η εκτίμηση ανθρώπινης πόζας απο τρικαναλικές εικόνες. Η συγκεκριμένη λειτουργία είναι απαραίτητη σε σύγχρονα αυτόνομα συστήματα που πρέπει να αντιλαμβάνονται το περιβάλλον γύρω τους ώστε να παίρνουν αποφάσεις αλλα και γενικότερα βρίσκει εφαρμογή σε πεδία οπως η αλληλεπίδραση ανθρώπου μηχανής και η εικονική/επαυξημένη πραγματικότητα. Στην παρούσα εργασία ασχολούμαστε με το υποπρόβλημα της τρισδιάστατης ανακατασκευής πόζας και σχήματος ανθρώπινου σώματος απο μία όψη μέσω RGB εικόνων. Το πρόβλημα παρουσιάζει ιδιαίτερη δυσκολία καθώς απαιτεί αφενώς τον εντοπισμό των βασικών αρθρώσεων στο πεδίο της εικόνας και αφετέρου την εκτίμηση του τρόπου με τον οποίο "ξεδιπλώνονται" αυτές οι αρθρώσεις στον τρισδιάστατο χώρο. Το παραπάνω πρόβλημα ανήκει στην ευρύτερη κατηγορία των αντίστροφών προβλημάτων καθώς προσπαθούμε να αντιστρέψουμε μια γνωστή διεργασία, την προβολή της γεωμετρίας ενός ανθρώπου στο επίπεδο της εικόνας.

Πιο συγκεκριμένα ασχολούμαστε τόσο με την στατική εκδοχή του προβλήματος οπου λαμβάνουμε σαν είσοδο μια εικόνα όσο και με την δυναμική εκδοχή του οπου η είσοδος είναι μια ακολουθία απο εικόνες(βίντεο). Αφου παρουσιάσουμε το απαραίτητο θεωρητικο υπόβαθρο και την σχετική βιβλιογραφία, εκκινούμε την εργασία μας υλοποιώντας και εκπαιδεύοντας δημοφιλή μοντέλα και απο τις δύο κατηγορίες της βιβλιογραφίας.

Στην συνέχεια προτείνουμε τροποποιήσεις στα βασικά μοντέλα με αφορμή τις εξαρτήσεις που υπάρχουν τόσο ανάμεσα στα δεδομένα πόζας και σχήματος όσο και στα δεδομένα πόζας μεταξύ των χρονικών στιγμών. Χρησιμοποιούμε μια πληθώρα απο διαφορετικά σύνολα δεδομένων και εκπαιδεύουμε τα μοντέλα σε διαφορετικά σενάρια ώστε να δείξουμε την επίδραση κάθε υποσυστήματος ενώ παράλληλα δείχνουμε πειραματικά οτι τα μοντέλα που προτείνουμε επιτυγχάνουν απόδοση όμοια ή και καλύτερη απο το βασικό μοντέλο.

Τέλος παρουσιάζουμε τα συμπεράσματα μας και προτείνουμε τρόπους επέκτασης της εργασίας.

Λέξεις Κλειδιά — Μηχανική Μάθηση, Βαθιά Μάθηση, Όραση Υπολογιστών, 3Δ Ανακατασκευή Ανθρώπινης Πόζας και Σχήματος.

Abstract

3D Human pose and shape estimation belongs to the most fundamental problems of computer vision. This functionality is needed on synchronous autonomous systems to be able to understand their enviroment and make critical decision. Futhermore this technology is widely used on other fields such as Human-Computer interaction and Virual/Augmented reality. In this work we concern ourselves with the subproblem of 3D human pose and shape estimation from single view RGB input. This problem is especially challenging since the model should not only be able to recognize the locations of joints on the image plane but should also be able to reason about the "unfolding" of these points on the three dimensional space. It belongs to the family of inverse problems since we try to inverse a forward process, in this case the projection of human geometry on the image plane.

More specifically we concern ourselves with both the static version of the problem, where our goal is to regress the 3D pose and shape from single RGB images, and the dynamic version where we use RGB video as input to the system. Initially we present the theoretical background regarding the problem and the relevant part of the literature. Subsequently we start our work by implementing popular models from the both categories.

Following this, we propose a number of modifications upon the baseline models by observing and explicitly modeling the dependence between both the pose and shape parameters and the pose parameters across time. We train our models on a plethora of different datasets under various settings to investigate the effect of subsystems on the performance of our models and experimentally show that our models achieve on par or better performance against the baselines.

At last we conclude our findings and propose ways to extend this work.

Keywords — Machine Learning, Deep Learning, Computer Vision, 3D Human Pose and Shape reconstruction.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό που μου έδωσε την ευκαιρία να εκπονήσω την παρούσα διπλωματική εργασία στο εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων. Τα μαθήματα του ήταν η αφορμή να ασχοληθώ με τον κόσμο της Επεξεργασίας Σήματος και της Όρασης Υπολογιστών. Στην συνέχεια θα ήθελα να ευχαριστήσω τους συνεπιβλέποντες μεταδιδακτορικούς ερευνητές του εργαστηρίου Δρ Παναγιώτη Φιλντίση και Δρ. Γεώργιο Ρετσινά για όλη τους την βοήθεια στην εκπόνηση της διπλωματικής αλλα κυρίως γιατι με βοήθησαν να καταλάβω τον τρόπο σκέψης ενός ερευνητή μέσα απο τις παρατηρήσεις τους. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου και τους φίλους για την στηριξή τους καθ'όλη την διάρκεια των σπουδών μου.

Σωτήριος Καραπιπέρης
Οκτώβριος 2023

Περιεχόμενα

Λίστα Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Μηχανική Μάθηση

Μία από τις μεγαλύτερες τεχνολογικές προκλήσεις του σύγχρονου κόσμου αποτελεί η δημιουργία αυτόνομων συστημάτων ικανών να αντιλαμβάνονται το περιβάλλον τους και να παίρνουν αποφάσεις το ίδιο απόδοτικά με τους ανθρώπους. Για μεγάλο μέρος των προβλημάτων γνωρίζουμε πως θα πρέπει να συμπεριφέρεται ένα τέτοιο σύστημα δεδομένης μιας εισόδου. Ωστόσο στην πλειοψηφία των προβλημάτων ενδιαφέροντος η δημιουργία μιας λίστας από κανόνες όπου το σύστημα θα ακολουθεί δεν αποτελεί εφικτή λύση καθώς είτε το πρόβλημα δεν είναι εύκολο να περιγραφεί με την μορφή κανόνων είτε η λίστα με τους κανόνες πρέπει να είναι μη πεπερασμένη. Ετσι προκύπτει η ανάγκη για δημιουργία μοντέλων και αλγορίθμων που μεταβάλλουν τις παραμέτρους των μοντέλων με την χρήση δεδομένων έως ότου το μοντέλο να αναγνωρίζει τα μοτιβα που κυβερνούν τις σχέσεις εισόδου-εξόδου. Η μελέτη τέτοιων μοντέλων και αλγορίθμων αποτελεί το πεδίο της μηχανικής μάθησης.

Αν και οι ερευνητές της επιστήμης υπολογιστών μελετούν και αναπτύσουν αλγόριθμους μηχανικής μάθησης ήδη από το μέσα του προηγούμενου αιώνα, η απουσία της απαραίτητης υπολογιστικής ικανότητας είχε προσφέρει περιορισμένα εφαρμοσμένα αποτελεσματα. Ωστόσο η ραγδαία τεχνολογική ανάπτυξη στον χώρο των ηλεκτρολογικών υλικών τις τελευταίες τρεις δεκαετίες, η οποία έχει αυξήσει την διαθέσιμη υπολογιστική ικανότητα των υπολογιστών, έχει δώσει την δυνατότητα στους ερευνητές να εφαρμόσουν τους παραπάνω αλγορίθμους επιτυγχάνοντας τα πρώτα αποτελέσματα, αποδεικνύοντας τις δυνατότητες της μηχανικής μάθησης.

1.2 Βαθιά Μάθηση

Οι πρώτοι αλγόριθμοι μηχανικής μάθησης, εκτός από την χρήση δεδομένων, βασίζονταν στην χρήση μεθόδων που μετασχηματιζουν τα δεδομένα από την αρχική μορφής τους σε μία χρήσιμη αναπαράσταση για τους αλγορίθμους μηχανικής μάθησης, οι οποίες προκειμένου να αναπτυχθούν χρειάζονταν τεχνικές γνώσεις γύρω από το συγκεκριμένο πρόβλημα. Ωστόσο η παρατήρηση ότι αυτές οι αναπαραστάσεις μπορούν επίσης να είναι αποτέλεσμα ενός μοντέλου μηχανικής μάθησης σε συνδυασμό με την αύξηση των διαθέσιμων δεδομένων και την ιδέα της εκπαίδευσης των μοντέλων σε κάρτες γραφικών οδήγησαν στην δημιουργία του αντικειμένου της Βαθιάς Μάθησης. Τα μοντέλα που ανήκουν σε αυτή την κατηγορία δεν απαιτούν τον μετασχηματισμο των δεδομένων σε κάποια συγκεκριμένη αναπαράσταση, αντιθέτως η εξαγωγή της κατάλληλης αναπαράστασης έχει ενσωματωθεί σαν μέρος του προβλήματος.

Δεδομένου ότι τα αποτελέσματα των μοντέλων βαθιάς μάθησης είναι συνήθως ανώτερα των προηγούμενων μεθόδων, η βαθιά μάθηση χρησιμοποιείται πλέον σαν εργαλείο σε πολλά πρακτικά πεδία οπως η Όραση Υπολογιστών, τα Γραφικά Υπολογιστών και η Επεξεργασία Φυσικής Γλώσσας και Ομιλίας.

1.3 Οραση Υπολογιστών

Το πεδίο της όρασης υπολογιστών ασχολείται με την μελέτη και ανάπτυξη μοντέλων και αλγορίθμων για την αντιμετώπιση προβλημάτων που ένας άνθρωπος προκειμένου να τα λύσει θα βασιζόταν στην όραση του. Στόχος είναι η δημιουργία συστημάτων ικανών να αντιλαμβάνονται το περιβάλλον μέσω σημάτων εικόνας και περιλαμβάνει μια πληθώρα από προβλήματα οπως κατηγοριοποίηση, κατάτμηση εικόνων κ.α. Αποτελεί ένα από τα πρώτα πεδία που επωφελήθηκε από την χρήση της βαθιάς μάθησης καθως το 2012 οι ερευνητές [KSH12] απέδειξαν την ανωτερότητα των βαθιών μοντέλων οταν κατάφεραν να αυξήσουν τον αριθμό των στρώσεων και να εκπαιδεύσουν επιτυχημένα ένα νευρωνικό δίκτυο αναγνώρισης εικόνων σημειώνοντας την καλύτερη επίδοση στον διαγωνισμό ImageNet2012. Εκ τοτε η πλειοψηφία των υποπροβλημάτων που περιλαμβάνει το πεδίο της Όρασης Υπολογιστών κάνουν χρήση τεχνικών βαθιάς μάθησης προκειμένου να εντοπίσουν μοτίβα στις εικόνες εισόδου και να παράξουν την αντίστοιχη έξοδο.

Ένα από τα σημαντικότερα υποπροβλήματα της όρασης υπολογιστών είναι η εκτίμηση και ανακατασκευή της ανθρώπινης πόζας από εικόνες. Είναι απαραίτητη τόσο για την καλύτερη κατανόηση του περιβάλλοντος από το σύστημα όσο και για την καλύτερη αλληλεπίδραση ανθρώπου-υπολογιστή.

1.4 Εκτίμηση και Ανακατασκευή Πόζας και Σχήματος

Το πρόβλημα της εκτίμησης και ανακατασκευής ανθρώπινης πόζας μπορεί να κατηγοριοποιηθεί τόσο ως προς το είδος του σήματος εξόδου του μοντέλου όσο και ως προς το ειδος του σήματος εισόδου. Με βάση το είδος του σήματος εισόδου το πρόβλημα κατηγοριοποιείται σε:

- Εκτίμηση πόζας απο μία όψη: Στη συγκεκριμένο σενάριο η μοναδική είσοδος στο μοντέλο είναι μια τρικαναλική/μονοκαναλική εικόνα.
- Εκτίμηση πόζας απο πολλαπλές όψεις: Στην προκειμένη περίπτωση στο μοντέλο δίνονται σαν είσοδοι πολλές εικόνες του ίδιο υποκειμένου από διαφορετικές οπτικές γωνίες.

Με βάση το είδος του σήματος εξόδου το πρόβλημα κατηγοριοποιείται σε:

- 2Δ εκτίμηση Πόζας: Στο συγκεκριμένο σενάριο το μοντέλο παράγει τις θέσεις των αρθρώσεων πάνω στο δισδιάστατο επίπεδο της εικόνας.
- 3Δ εκτίμηση Πόζας: Στην προκειμένη περίπτωση το μοντέλο εκτιμά της θέσεις των αρθρώσεων στον τρισδιάστατο χώρο είτε ως προς ένα καθολικό σύστημα αναφοράς είτε ως προς το σύστημα αναφοράς της κάμερας κάνοντας παράλληλα μια εκτίμηση για την θέση και τον προσανατολισμό της κάμερας.

Παραδείγματα από τις διαφορετικές κατηγορίες προβλημάτων εκτίμησης πόζας φαίνονται στην Εικόνα 1.4.1

Στην παρούσα διπλωματική ασχολούμαστε εν μέρη με το πρόβλημα της 3Δ εκτίμησης πόζας απο μια όψη. Το συγκεκριμένο σενάριο κατέχει εναν επιπλέον συντελεστή δυσκολίας καθως ακόμα και αν γνωρίζουμε τις θέσεις των αρθρώσεων στο πεδιο της εικόνας το μοντέλο καλείται να λύσει ενα "αντίστροφο" πρόβλημα, να εκτιμήσει το "ξεδίπλωμα" των σημείων στον χώρο. Το συγκεκριμένο πρόβλημα είναι το αντίστροφο πρόβλημα της προβολής ένα συνόλου σημείων του χώρου σε ενα συγκεκριμένο επίπεδο και ανήκει στην γενικότερη κατηγορία των "αντίστροφων" προβλημάτων.

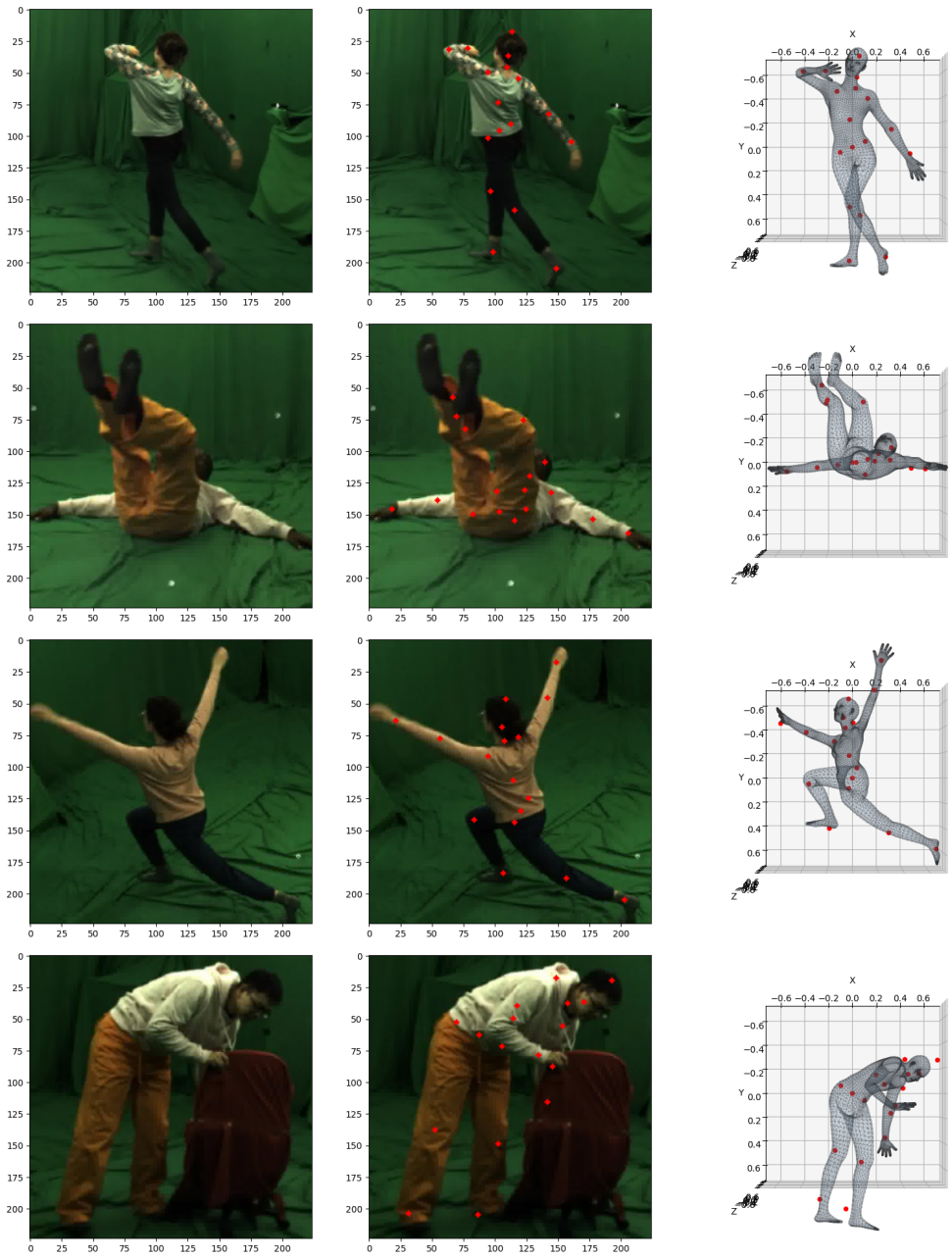
Η έξοδος ενός μοντέλου εκτίμησης τρισδιάστατης παράγει έναν σκελετό στον χώρο ο οποίος προκύπτει ενώνοντας τα σημεία των αρθρώσεων με ευθύγραμμα τμήματα. Ωστόσο αυτή η αφαιρετική αναπαράσταση δεν φτάνει για να περιγράψει ικανοποιητικά έναν άνθρωπο στον χώρο καθώς πολλοί σωματότυποι αντιστοιχούν στον ίδιο σκελετο. Μια πιο ικανοποιητική αναπαράσταση θα ήταν ένα πλέγμα από κόμβους στο χώρο που θα περιγράφει ακριβώς το σχήμα του ανθρώπινου σώματος στη δεδομένη πόζα. Για τον σκοπό αυτό έχουν αναπτυχθεί παραμετρικά μοντέλα του ανρθωπίνου σώματος, τα οποία λαμβάνουν σαν είσοδο 2 ειδών παραμέτρους χαμηλής διάστασης, πόζας και σχήματος, και παράγουν το πλέγμα.

Ολοκληρώνοντας, στην παρούσα διπλωματική θα αντιμετωπίσουμε το πρόβλημα της 3Δ ανακατασκευής πόζας και σχήματος απο μια οψη χρησιμοποιώντας μοντέλα βαθιάς μάθησης και παραμετρικά μοντέλα του ανθρωπίνου σώματος. Πιο συγκεκριμένα θα εκπαιδεύσουμε τα μοντέλα βαθιάς μάθησης ώστε να εκτιμούν απο την εικόνα εισόδου τις παραμέτρους πόζας και σχήματος, οι οποίες θα δίνονται σαν εισοδοι στο παραμετρικό μοντέλο ώστε να παράξει το τελικό πλέγμα.

1.5 Στόχοι και Οργάνωση Εργασίας

Στόχος της παρούσας εργασίας είναι υλοποίηση και ανάλυση δημοφιλών μοντέλων απο την σχετική βιβλιογραφία, η επέκταση τους με βασή τις παρατηρήσεις μας για τις εξαρτήσεις μεταξύ των παραμέτρων πόζας και σχηματος και η σύγκριση τους ώστε να προκύψουν χρήσιμα συμπεράσματα. Στην συνέχεια εξετάζουμε το πρόβλημα στην χρονική του διάσταση επεκτείνοντας ενα δημοφιλές μοντέλο από την βιβλιογραφία με βάση τα συμπεράσματα μας. Δείχνουμε πειραματικά οτι οταν στον σχεδιασμό της αρχιτεκτονικής μοντελοποιούνται ρητά οι εξαρτήσεις και οι περιορισμοί μεταξύ των παραμέτρων η απόδοση του μοντέλου βελτιώνεται ενώ παράλληλα γίνεται ευκολότερος ο αντιμετωπισμός των αποτυχιών του μοντέλου. Τέλος σκοπός είναι η δημιουργία ενος αποθετήριου κώδικα που παρέχει τα υλοποιημένα μοντέλα της παρούσας διπλωματικής και διευκολύνει την επέκταση του με νέα μοντέλα της βιβλιογραφίας.

Όσον αφορά στην οργάνωση της εργασίας, αρχικά παρουσιάστηκε μια εισαγωγή στο αντικείμενο μελέτης. Στην συνέχεια θα δοθεί το θεωρητικό υπόβαθρο το οποίο είναι χρήσιμο για την κατανόηση της παρουσίασης της βιβλιογραφίας που ακολουθεί. Στα τελευταία τρία κεφάλαια παρουσιάζουμε την προτεινόμενη μεθοδολογία, την πειραματική διαδικασία και τέλος τα συμπεράσματα μας. Πιο αναλυτικά:



Εικόνα 1.4.1: Παραδείγματα από το σύνολο δεδομένων MPI-INF-3DHP [Meh+17](1η στήλη). Το αντικείμενο του προβλήματος εκτίμησης δισδιάστατης πόζας φαίνεται στην 2η στήλη οπου στόχος είναι η εύρεση των αρθρώσεων στο πεδίο της εικόνας ενώ η 3η στήλη παρουσιάζει την επιθυμητή έξοδο ενός μοντέλου εκτίμησης τρισδιάστατης πόζας και σχήματος. Στα παραπάνω παραδείγματα φαίνεται η δυσκολία του προβλήματος, καθώς πολλές φορές είτε οι αρθρώσεις είτε αντικείμενα αποκρύπτουν μέρος του σώματος του υποκειμένου.

- Στο κεφάλαιο 1 παρουσιάστηκε μια σύντομη εισαγωγή στο αντικείμενο μελέτης και στο ευρύτερο ερευνητικό πεδιο στο οποίο ανήκει.
- Στο κεφάλαιο 2 παρουσιάζεται το απαραίτητο θεωρητικό υπόβαθρο το οποίο είναι απαραίτητο προκειμένου να κατανοηθούν τα πειράματα
- Στο κεφάλαιο 3 παρουσιάζεται η σχετική βιβλιογραφία δίνοντας έμφαση στο μοντέλο το οποίο θα χρησιμοποιηθεί ώς το αρχικό μας μοντέλο και το οποίο θα αποτελέσει το μέτρο σύγκρισης.

- Στο κεφάλαιο 4 παρουσιάζονται οι επεκτάσεις των βασικών μοντέλων που προτείνουμε.
- Στο κεφάλαιο 5 παρουσιάζεται η πειραματική διαδικασία με τα αντίστοιχα ποσοτικά και ποιοτικά πειράματα.
- Τέλος στο κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα μας, οι περιορισμοί της ερευνητικής εργασίας και πιθανοί τρόποι επέκτασης της.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Μηχανική Μάθηση

Οπως αναφέρθηκε στην κεφάλαιο 1 η Μηχανική Μάθηση ασχολείται με την μελέτη και υλοποίηση αλγορίθμων που έχουν την ικανότητα να αναγνωρίζουν μοτίβα στα δεδομένα ωστε να επιλύουν προβλήματα για τα οποία δεν έχουν προγραμματιστεί ρητά. Η ικανότητα των αλγορίθμων να μεταβάλλουν τις παραμέτρους τους σύμφωνα με τα δεδομένα ονομάζεται Μάθηση ενώ ενας πιο σαφής ορισμός της δόθηκε από τον Tom Mitchel το 1997 που όρισε την διαδικασία μάθησης ως: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E».

Στον προηγούμενο ορισμό η εμπειρία είναι συνήθως το σύνολο δεδομένων, ωστόσο η μορφή με την οποία υπάρχουν τα δεδομένα δημιουργούν διαφορετικά σενάρια μάθησης. Πιο συγκεκριμένα υπάρχουν τρία κυρίαρχα σενάρια:

- Επιβλεπόμενη Μάθηση: Στο συγκεκριμένο σενάριο το σύνολο δεδομένων αποτελείται από ζεύγη εισόδου-εξόδου και στόχος του μοντέλου είναι η αναγνώριση της λανθάνουσας σχέσης που τα διέπει.
- Μη Επιβλεπόμενη Μάθηση: Σε αυτή την περίπτωση το σύνολο δεδομένων αποτελείται μόνο από δείγματα εισόδου χωρίς να υπάρχουν επισημειώσεις για αυτα και στόχος είναι η μελέτη της δομής του χώρου στον οποίο ανήκουν.
- Ενισχυτική Μάθηση: Στην προκειμένη περίπτωση η εμπειρία δεν δίνεται στο μοντέλο σαν σύνολο δεδομένων, αντιθέτως το μοντέλο έχει την δυνατότητα να αλληλεπιδράσει με τον περιβάλλον στο οποίο ανήκει και ανάλογα με την ανταμοιβή που θα λάβει πρέπει να εξάγει πολιτικές συμπεριφοράς.

Η παρούσα εργασία ασχολείται με τα 2 πρώτα είδη τα οποία θα αναλυθούν παρακάτω συνοπτικά.

2.1.1 Επιβλεπόμενη Μάθηση

Όπως αναφέρθηκε παραπάνω, στο σενάριο της επιβλεπόμενης μάθησης η εμπειρία E δίνεται στο μοντέλο ως ένα σύνολο δειγμάτων όπου κάθε δείγμα είναι ένα ζεύγος εισόδου-εξόδου. Συνήθως η είσοδος είναι το κατεργασμένο ή ακατέργαστο δεδομένο ενώ η έξοδος αποτελεί τον μετασχηματισμό του δειγματος εισόδου, τον οποίο το μοντέλο πρέπει να ανιχνεύσει. Το είδος της εξόδου ποικίλει ανάλογα με την κλάση εργασιών T, π.χ. για την περίπτωση όπου η κλάση T είναι η κατηγοριοποίηση εικόνας, τότε η είσοδος είναι μια εικόνα είτε με την μορφή ενός πίνακα από εικονοστοιχεία είτε με μια άλλη αναπαράσταση και έξοδος είναι η κατηγορία στην οποία ανήκει η συγκεκριμένη εικόνα. Το μέτρο P στην προκειμένη περίπτωση θα ήταν ευστοχία του μοντέλου στην κατηγοριοποίηση του συνόλου εικόνων.

Ανάλογα με το είδος της εξόδου διακρίνονται 2 υποκατηγορίες επιβλεπόμενης μάθησης:

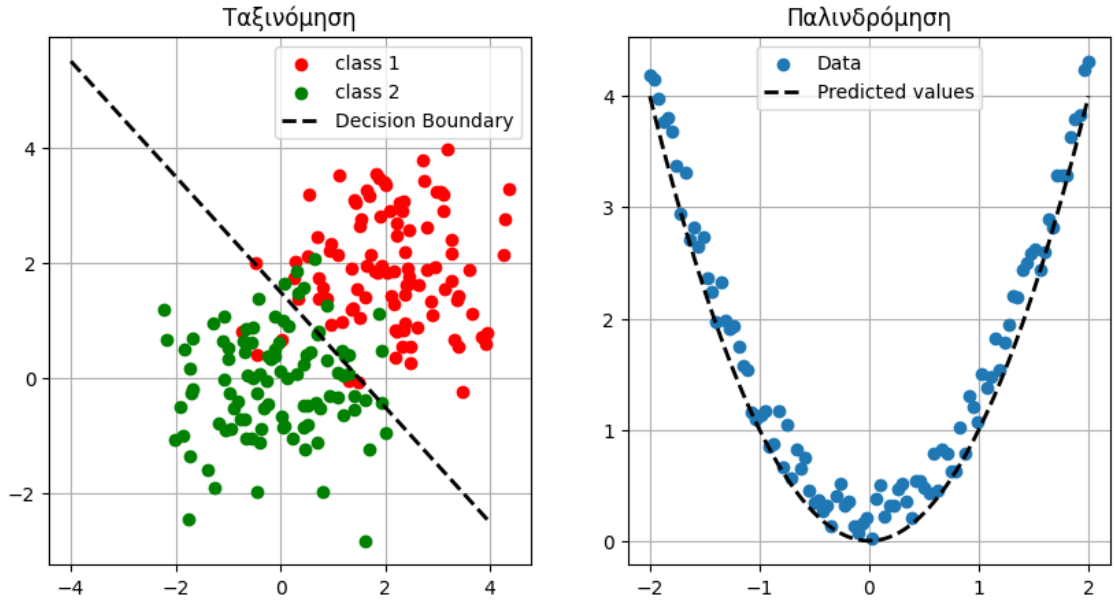
- Ταξινόμηση: Σε αυτή την κατηγορία σκοπός είναι η κατηγοριοποίηση/ταξινόμηση κάθε δεδομένου εισόδου σε μια από τις διαθέσιμες κατηγορίες. Κάθε δεδομένο εισόδου συνοδεύεται από μια επισημείωση της κατηγορίας στην οποία ανήκει.
- Παλινδρόμηση: Στην προκειμένη περίπτωση σκοπός είναι η εκτίμηση μια συνεχούς τιμής για κάθε δεδομένο εισόδου. Κάθε δεδομένο εισόδου συνοδεύεται από την συνεχή τιμή στην οποία αντιστοιχεί.

Πειραματικά η Παλινδρόμηση αποτελεί ένα δυσκολότερο πρόβλημα καθώς το πεδίο εξόδου εχει μεγαλύτερο πληθάριθμο.

2.1.2 Μη-Επιβλεπόμενη Μάθηση

Σε αντίθεση με την Επιβλεπόμενη μάθηση στο σενάριο της μη επιβλεπόμενης μάθησης η εμπειρία E δίνεται στο μοντέλο με την μορφή δεδομένων χωρίς κάποια επισημείωση. Σκοπός είναι η υλοποίηση μοντέλων ικανών να ανιχνεύουν την λανθάνουσα δομή των δεδομένων εισόδου και να την εκμεταλλευτούν ώστε να επιτύχουν κάποια συγκεκριμένη εργασία. Παραδείγματα κλάσεων εργασιών T αποτελούν τα:

- Μείωση Διάστασης: Στο συγκεκριμένο πρόβλημα στόχος είναι η μείωση της διάστασης των δεδομένων εισόδου χάνοντας όσο το δυνατόν λιγότερη πληροφορία όπως αυτό μετράται σύμφωνα με ένα μέτρο P π.χ. την διατήρηση της διακύμανσης των δεδομένων.



Εικόνα 2.1.1: Παραδείγματα επιβλεπόμενης μάθησης. Στην Ταξινόμηση(αριστερή εικόνα) σκοπός του μοντέλου είναι η κατηγοριοποίηση των δεδομένων σε κατηγορίες. Αυτο συνήθως επιτυγχάνεται βρίσκοντας μια διαχωριστική υπερεπιφάνεια που χωρίζει τον χώρο σε υπερεπίπεδα. Στην παλινδρόμηση σκοπός είναι η μάθηση της πραγματικής διαδικασίας που παράγει κάποια δεδομένα.

- Συσταδοποίηση: Στην προκειμένη περίπτωση σκοπός είναι η δημιουργία συστάδων. Ομοίως το αποτέλεσμα της συσταδοποίησης αποτιμαται με κάποιο μέτρο P οπως είναι το άθροισμα των ευκλείδιων αποστάσεων όλων των σημείων από το κέντρο της συστάδας τους.
- Εκτίμηση Πυκνότητας Πιθανότητας: Σε αυτή τη κλάση προβλημάτων γίνεται η υπόθεση ότι τα δεδομένα εισόδου ακολουθούν μια κατανομή p_data. Στόχος είναι η δημιουργία ενός μοντέλου που αναπαριστά μια κατανομή p_model η οποία προσεγγίζει αυτή την κατανομή. Το μέτρο P σε αυτή την περίπτωση είναι πιθανοφάνεια των δεδομένων εισόδου ως προς την κατανομή p_model

Οπως ήδη αναφέρθηκε στο κεφάλαιο 1 πλέον η πλειοψηφία των μοντέλων μηχανικής μάθησης αποτελούνται από βαθιά νευρωνικά δίκτυα. Στην συνέχεια θα παρουσιάσουμε τις κυριότερες αρχιτεκτονικές βαθιών νευρωνικών δικτύων, τους τρόπους με τους οποίους πραγματοποιείται η μάθηση και τα δημοφιλη παραδείγματα εκπαίδευσης.

2.2 Βαθιά Νευρωνικα Δίκτυα

2.2.1 Νευρωνικά δίκτυα

Το νευρωνικό δίκτυο είναι ένα είδος μοντέλου μηχανικής μάθησης όπου ο βασικός μετασχηματισμός που χρησιμοποιείται είναι το perceptron. Πιο συγκεκριμένα για ένα δείγμα εισόδου $\mathbf{x} \in R^d$ ένα perceptron παράγει ένα πραγματικό αριθμό $y \in R$ ώς

$$y = f(\mathbf{w}^T\mathbf{x} + b)$$

όπου $\mathbf{w} \in R^d$ και $b \in R$ αποτελούν τις παραμέτρους του perceptron και f είναι μια μή γραμμική συνάρτηση όπου ονομάζεται *Συνάρτηση Ενεργοποίησης*. Χρησιμοποιώντας k perceptrons κατασκευάζεται μια στρώση του νευρωνικού δικτύου όπου μετατρέπει ένα διάνυσμα εισόδου $\mathbf{x} \in R^d$ σε ένα διάνυσμα εξόδου $\mathbf{y} \in R^k$. Συχνά στην βιβλιογραφία το σύνολο των k perceptrons αναφέρεται καταχρηστικά ως μια γραμμική στρώση(Linear Layer) ακολουθούμενη από μια συνάρτηση ενεργοποίησης

$$\mathbf{u} = \mathbf{W}^T\mathbf{x} + \mathbf{b}$$
$$\mathbf{y} = \mathbf{f}(\mathbf{u}) = (f(u_1), f(u_2), ..., f(u_k))$$

όπου $\mathbf{W} \in R^{d \times k}$ και $\mathbf{u}, \mathbf{y} \in R^k$.

Βαθύ νευρωνικό δίκτυο ονομάζεται κάθε νευρωνικό δίκτυο με παραπάνω από μια στρώσεις. Εκτός από την τελευταία στρώση που παράγει την τελική εκτίμηση οι ενδιάμεσες στρώσεις ονομάζονται *κρυφές*. Στην βιβλιογραφία επικρατεί η άποψη πως οι κρυφές στρώσεις μέσω της μάθησης παράγουν χρήσιμες αναπαραστάσεις των αρχικών δεδομένων εισόδου τις οποίες χρησιμοποιεί η τελευταία στρώση προκειμένου να παράξει την τελική εκτίμηση. Η επιτυχία των βαθιών νευρωνικών δικτύων βασίζεται στην ιδέα οτι η σύνθεση πολλών απλών μη γραμμικών συναρτήσεων μπορεί να αναπαραστήσει μια πιο περίπλοκη συνάρτηση.

2.2.2 Συναρτήσεις ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης διαδραματίζουν κομβικό ρόλο στην αρχιτεκτονική των βαθιών νευρωνικών δικτύων καθώς χωρίς αυτές το συνολικό μοντέλο θα ήταν συνήθως γραμμικό περιορίζοντας την ικανότητα αναπαράστασης του. Επιπλέον οι διάφορες ιδιότητες τους είναι χρήσιμες ανάλογα με την ποσότητα που μοντελοποιεί το δίκτυο ενώ οι μαθηματικές τους ιδιότητες άλλοτε δυσκολεύουν και άλλοτε διευκολύνουν την εκπαίδευση του δικτύου.

Οι πιο δημοφιλείς συναρτήσεις είναι:

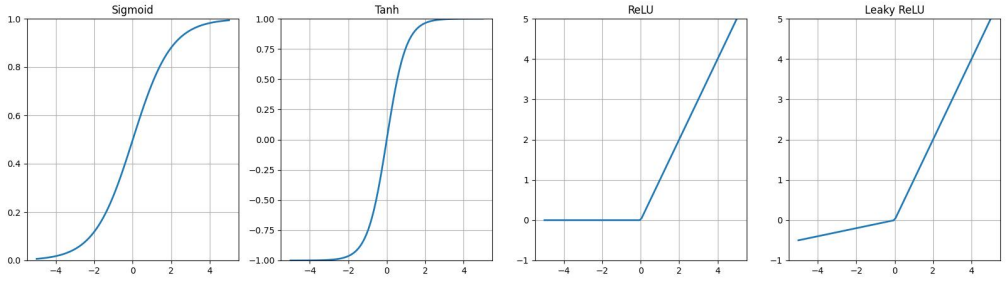
- Η σιγμοειδής συνάρτηση $f(x) = \frac{1}{1+e^{-x}}$: Το σύνολο τιμών της είναι το $[0, 1]$ γεγονός που κάνει χρήσιμη την συνάρτηση για αναπαράσταση πιθανότητας όταν σκοπός είναι η ταξινόμηση σε 2 κατηγορίες. Ωστόσο η μικρή παράγωγος καθώς το x απομακρύνεται από την αρχή των αξόνων ενδέχεται να δυσκολέψει την εκπαίδευση.

- Η υπερβολική εφαπτομένη $f(x) = tanh(x)$: Το σύνολο τιμών της είναι το $[-1, 1]$ γεγονός που την κάνει χρήσιμη όταν το δίκτυο μοντελοποιεί ποσότητες που λαμβάνουν αρνητικές τιμές. Ομοίως με την σιγμοειδή συνάρτηση οι παράγωγοι της μπορεί να δυσκολέψουν την εκπαίδευση.

- Η ανορθωμένη γραμμική συνάρτηση $f(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$: Το σύνολο τιμών της είναι το $[0, \infty]$ γεγονός που την κάνει χρήσιμη σε περιπτώσεις όπου το μοντέλο πρεπει να μπορει προβλέψει μονο θετικές τιμές χωρίς να περιορίζεται από κάποιο πάνω όριο. Πειραματικά έχει βρεθεί πως επιταχύνει την εκπαίδευση. Αν και στο σημείο $x = 0$ δεν υπάρχει η παράγωγος, πρακτικά το πρόβλημα λύνεται θέτοντας αυθαίρετα την παράγωγο ίση με 0.

- Διαρρέουσα ανορθωμένη γραμμική συνάρτηση $f(x) = \begin{cases} x & x > 0 \\ ax & x \leq 0 \end{cases}, a > 0$: Αποτελεί επέκταση της ανορθωμένης γραμμική συνάρτησης και είναι χρήσιμη όταν θέλουμε να συνδυάσουμε τα οφέλη της με την αντιστρεψιμότητα. Συχνά χρησιμοποιείται σε νευρωνικα δίκτυα που θέλουμε να είναι αντιστρέψιμα. Η παράμετρος a επιλέγεται ανάλογα με την περίσταση και συνήθως λαμβάνει τιμές $<< 1$.

Οι συναρτήσεις φαίνονται στην Εικόνα 2.2.1



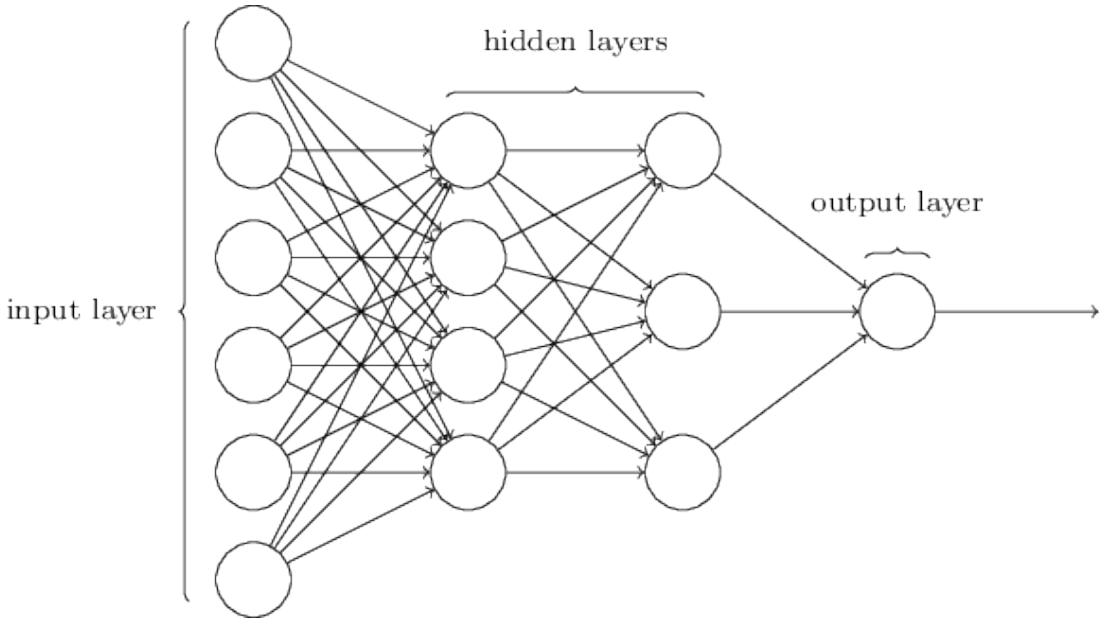
Εικόνα 2.2.1: Οι 4 πιο κοινές συναρτήσεις ενεργοποίησης.

2.2.3 Αρχιτεκτονικές Βαθιών Νευρωνικών δικτύων

Ανάλογα με τον τροπο που συνδέονται μεταξύ τους οι στρωσεις αλλα και τον τρόπο που εφαρμόζονται στα δεδομένα διακρίνουμε τις παρακάτω κατηγορίες νευρωνικών δικτύων.

Πλήρως Συνδεδεμένα Νευρωνικά Δίκτυα

Το συγκεκριμένο δίκτυο αποτελεί την πιο απλή μορφή βαθιών νευρωνικών δικτύων. Αποτελείται από πολλαπλές στρώσεις οπως περιγράφηκαν παραπάνω, όπου κάθε στοιχείο της εξόδου στο i-οστό επίπεδο συνδέεται με όλα τα στοιχεία του προηγούμενο και επόμενου επιπέδου. Στην βιβλιογραφία αναφέρονται και ως Πολλαπλές Στρώσεις Perceptron(Multi Layer Perceptron- MLP). Η αρχιτεκτονική του μοντέλου παρουσιάζεται στην Εικόνα 2.2.2.



Εικόνα 2.2.2: Παράδειγμα αρχιτεκτονικής ενός πλήρως συνδεδεμένου δικτύου(MLP). Εικόνα από [LeD17]

Συνελικτικά Νευρωνικά Δίκτυα

Τα δίκτυα αυτή της κατηγορίας αποτελούνται από συνελικτικές στρώσεις. Η πράξη της συνέλιξης σε αυτό το πεδίο ομοιάζει αρκετά με τον παραδοσιακό ορισμό της καθώς αποτελείται από εναν πυρήνα ο οποίος εφαρμόζεται σε τμήματα του σήματος εισόδου. Πιο συγκεκριμένα για ένα δείγμα εισόδου $\mathbf{x} \in R^d$ και έναν *πυρήνα* $\mathbf{k} \in R^k$ το αποτέλεσμα της πράξης θα είναι ένα διάνυσμα διάστασης $d - k + 1$ το οποίο προκύπτει εφαρμόζοντας σειριακά τον πυρήνα πανω σε τμήματα μήκος k της εισόδου, πολλαπλασιάζοντας στοιχείο με στοιχείο και αθροίζοντας το αποτέλεσμα. Το αποτελεσμα ομοίως με παραπάνω δίνεται ως είσοδος σε μια συνάρτηση ενεργοποίησης. Ο ορισμός μπορεί να επεκταθεί για δισδιάστατα δεδομένα εισόδου με δισδιάστατους πυρήνες και για τρισδιάστατες είσοδους π.χ. τρικαναλική εικόνα με τρισδιάστατους πυρήνες.

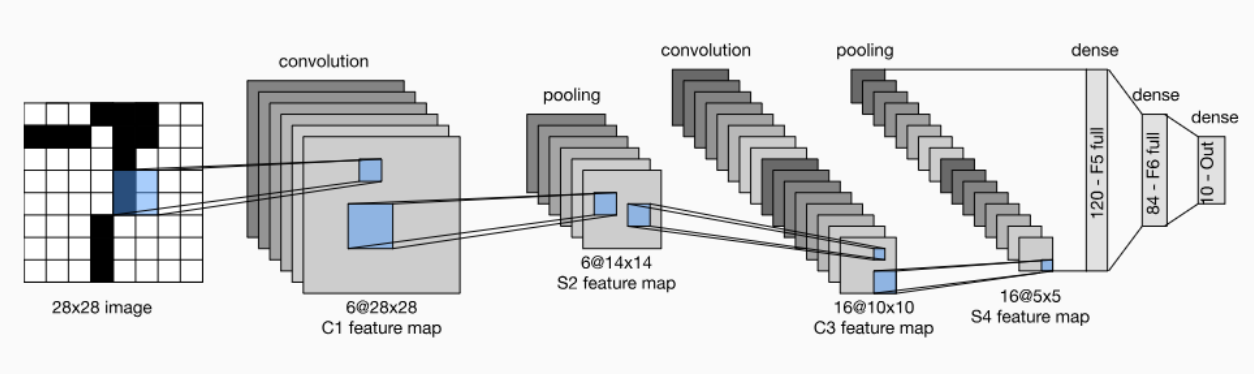
Η αρχιτεκτονική ενός δικτύου προκαταλήπτει το μοντέλο προς ένα συγκεκριμένο τρόπο λειτουργίας (inductive bias). Όσον αφορά στο κίνητρο πίσω από την χρήση της συνέλιξης αποτελείται από δύο παρατηρήσεις:

- Τοπικότητα: Η χρήση συνελικτικών έναντι των πλήρως συνδεδεμένων στρώσεων δικαιολογείται από την παρατήρηση ότι στα σήματα ενδιαφέροντος (π.χ. εικόνα) οι κοντινές περιοχές συσχετίζονται μεταξύ τους περισσότερο από ότι δύο περιοχές σε διαφορετικά σημεία της εικόνας. Επομένως επιλέγοντας συνελικτικές στρώσεις αναγκάζεται το μοντέλο να εκμεταλλευτεί αυτή την ιδιότητα των δεδομένων.
- Στασιμότητα: Το δεύτερο κίνητρο προκύπτει από την παρατήρηση ότι δεν θα έπρεπε να διαδραματίζει σημαντικό ρόλο η τοποθεσία ενός θεμελιώδους μοτίβου στο σήμα. Η χρήση συνελίξεων προκαταβάλλει το δίκτυο προς αυτή την κατεύθυνση.

Τα συνελικτικά δίκτυα χρησιμοποιούν πολλαπλές στρώσεις συνέλιξης ώστε να μετατρέψουν την αρχική είσοδο σε μια χρήσιμη αναπαράσταση.Αφου μειωθεί αρκετά η διάσταση, η τελική αναπαράσταση δίνεται σε ένα πλήρως συνδεδεμένο δίκτυο οπως αυτό που παρουσιάστηκε παραπάνω ώστε να παράξει την τελική εκτίμηση στην απαραίτητη μορφή. Συχνά, προκειμένου να μειωθεί η διάσταση της εισόδου γρηγορότερα χρησιμοποιούνται στρώσεις ομαδοποίησης(Pooling) μεταξύ των στρώσεων συνέλιξης όπου ομαδοποιούν μια γειτονιά από χαρακτηριστικά παράγοντας ένα μεμονωμένο χαρακτηριστικό. Για την υλοποίηση της λειτουργίας μετακινείται ένας πυρήνας πάνω στη είσοδο και ως γειτονιά ορίζεται κάθε φορά το σύνολο των στοιχείων που καλύπτονται από αυτόν. Αυτές οι ομαδοποιήσεις συνήθως ανήκουν σε μια από τις παρακάτω κατηγορίες:

- Ομαδοποίηση Μεγίστου: Στην συγκεκριμένη ομαδοποίηση επιλέγεται το μέγιστο στοιχείο από μια γειτονιά ως το τελικό χαρακτηριστικό αυτής.
- Ομαδοποίηση Μέσο Όρου: Στη συγκεκριμένη περίπτωση εξάγεται ο μέσος όρος των στοιχείων της γειτονιάς ως το τελικό χαρακτηριστικό.

Η συνολική αρχιτεκτονική ενός συνελικτικού δικτύου φαίνεται στην Εικόνα 2.2.3



Εικόνα 2.2.3: Παράδειγμα αρχιτεκτονικής ενός συνελικτικού δικτύου(CNN). Εικόνα από [Zha+23]

Αναδρομικά Νευρωνικά Δίκτυα

Σε αντίθεση με την παραπάνω θεώρηση υπάρχουν σενάρια που το μοντέλο πρέπει να είναι ευαίσθητο ως προς την σειρά με την οποία δέχεται σαν είσοδο τα δεδομένα, όπως π.χ. στην περίπτωση επεξεργασίας χρονοσειρών. Σε αυτή την περίπτωση το δίκτυο λαμβάνει σειριακά ολα τα δεδομένα πρωτού παράξει την τελική εκτίμηση. Τα δίκτυα αυτής της κατηγορίας ονομάζονται Αναδρομικά Νευρωνικά Δίκτυα (ΑΝΔ) καθώς η έξοδος την τρεχουσα χρονική στιγμή t δίνεται σαν επιπλέον είσοδος την επόμενη χρονική στιγμή $t+1$. Κάθε ΑΝΔ απαρτίζεται από ένα κύτταρο το οποιο επαναλαμβάνεται στο πεδιο του χρόνου επομένως στην συνέχεια παρουσιάζουμε το κύτταρο για κάθε μια από τις τρεις πιο δημοφιλείς κατηγορίες ΑΝΔ.

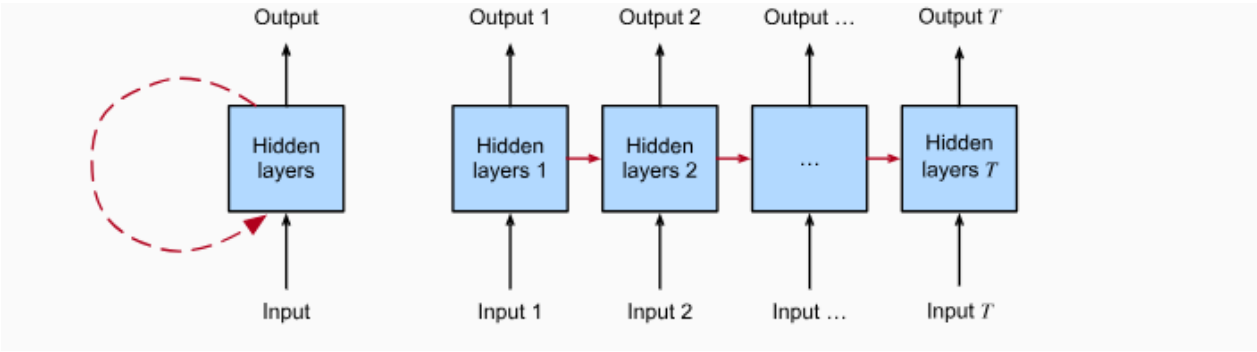
Απλα ΑΝΔ Η πιο απλή μορφή των αναδρομικών νευρωνικών δικτύων είναι τα πλήρως συνδεδεμένα νευρωνικά δίκτυα τα οποία εφαρμόζονται διαδοχικά σε κάθε δείγμα της εισόδου. Πιο συγκεκριμένα το κύτταρο απαρτίζεται από 2 πληρως συνδεδεμένες στρώσεις όπου το αποτέλεσμα της πρώτης στρώσης, το οποίο ονομάζεται και *κρυφή αναπαράσταση* δίνεται σαν επιπλεόν είσοδος την επόμενη χρονική στιγμή ενω το αποτέλεσμα της δεύτερης στρώσης είναι η εκτίμηση του μοντέλου για καθε χρονική στιγμή.

Πιο συγκεκριμένα για μια ακολουθία από δείγματα $\mathbf{x}_i \in R^{d_x}$ το μοντέλο παράγει την κρυφή αναπαράσταση $\mathbf{h}_i \in R^{d_h}$ και την έξοδο $\mathbf{y}_i \in R^{d_y}$ ως εξής:

$$\mathbf{h}_i = f(\mathbf{W}_h^T \mathbf{h}_{i-1} + \mathbf{W}_x^T \mathbf{x}_i + \mathbf{b})$$
$$\mathbf{y}_i = \mathbf{W}_y^T \mathbf{h}_i$$

όπου $\mathbf{W}_h \in R^{d_h \times d_h}$, $\mathbf{W}_x \in R^{d_x \times d_h}$, $\mathbf{W}_y \in R^{d_y \times d_h}$, $\mathbf{b} \in R^{d_h}$ είναι οι παράμετροι του κυττάρου και f είναι η συνάρτηση ενεργοποίησης.

Για την χρονική στιγμή 0 η κρυφή αναπαράσταση της προηγούμενης χρονικής στιγμής θεωρείται το μηδενικό διάνυσμα. Η αρχιτεκτονική ενός απλού Αναδρομικού Δικτύου φαίνεται στην Εικόνα 2.2.4



Εικόνα 2.2.4: Παράδειγμα αρχιτεκτονικής ενός απλού αναδρομικού δικτύου(RNN). Εικονα από [Zha+23]

Ωστόσο αυτά τα δίκτυα υποφέρουν τοσο από προβλήματα εκπαίδευσης,καθώς οι παράγωγοι τους είτε εκτινάσσονται είτε μηδενίζονται(exploding/vanishing gradients problem) οσο και από προβλήματα μοντελοποίησης καθώς πειραμάτικα δείχνεται πως μετά από καποιο μήκος ακολουθίας τα συγκεκριμένα κύτταρα δεν διατηρούν πληροφορία σχετικά με τα αρχικά δείγματα. Προκειμένου να εξαλειφθούν αυτα τα μειονεκτήματα οι ερευνητές είσηγαγαν τα ΑΝΔ με χρήση πυλών τα οποία και παρουσιάζουμε συνοπτικά στις επόμενες δύο παραγράφους.

Αναδρομική Μονάδα με Πύλες Οι Αναδρομικές Μονάδες με Πύλες (Gated Recurrent Unit-GRU) χρησιμοποιούν εναν μηχανισμό με πύλες προκειμένου να ανανεώνουν την κρυφή αναπαράσταση με τέτοιο τρόπο ώστε να είναι εφικτό το μοντέλο να διατηρεί πληροφορία για μεγαλύτερο χρονικό παράθυρο. Η βασική ιδέα των GRU είναι οτι το μοντέλο εκτός από τις κρυφές αναπαραστάσεις εκτιμά και διανύσματα ελέγχου τα οποία δεδομένου οτι προέρχονται από σιγμοειδείς συναρτήσεις έχουν τιμές στο διάστημα $[0, 1]$. Τα διανύσματα ελέγχου λειτουργούν ως μασκες οι οποίες οταν πολλαπλασιαστούν στοιχειο με στοιχείο με ένα διάνυσμα είτε τήνουν να διατηρούν είτε να σβήνουν την αντιστοίχη πληροφορία.

Ομοίως με το παραπάνω απλό κύτταρο, το κύτταρο των GRU διατηρεί για κάθε χρονική στιγμή μια κρυφή αναπαράσταση. Η ειδοποιός διαφορά με το απλό κύτταρο είναι ο τρόπος με τον αυτη η αναπαράσταση ενημερώνεται σε κάθε χρονική στιγμή. Πιο συγκεκριμένα το μοντέλο με βασή την τρέχουσα είσοδο και την κρυφή αναπαράσταση εκτιμά δύο διανύσματα ελέγχου $\mathbf{z}_i \in R^{d_z}$, $\mathbf{r}_i \in R^{d_r}$ τα οποία ονομάζονται διανύσματα *ενημέρωσης* και *επαναφοράς* και υπολογίζονται ως εξής:

$$\mathbf{z}_i = \sigma(\mathbf{W}_z^T \mathbf{x}_i + \mathbf{U}_z^T \mathbf{h}_{i-1} + \mathbf{b}_z)$$
$$\mathbf{r}_i = \sigma(\mathbf{W}_r^T \mathbf{x}_i + \mathbf{U}_r^T \mathbf{h}_{i-1} + \mathbf{b}_r)$$

Στην συνέχεια παράγεται η νέα *υποψήφια* κρυφή αναπαράσταση $\tilde{\mathbf{h}}_i \in R^{d_h}$ χρησιμοποιώντας την τρεχουσα είσοδο και την κρυφή αναπαράσταση της προηγούμενης χρονικής στιγμής αφού πρώτα πολλαπλασιαστεί με το διάνυσμα επαναφοράς ώστε να σβηστεί η θεωρούμενη ως μη χρήσιμη πληροφορία

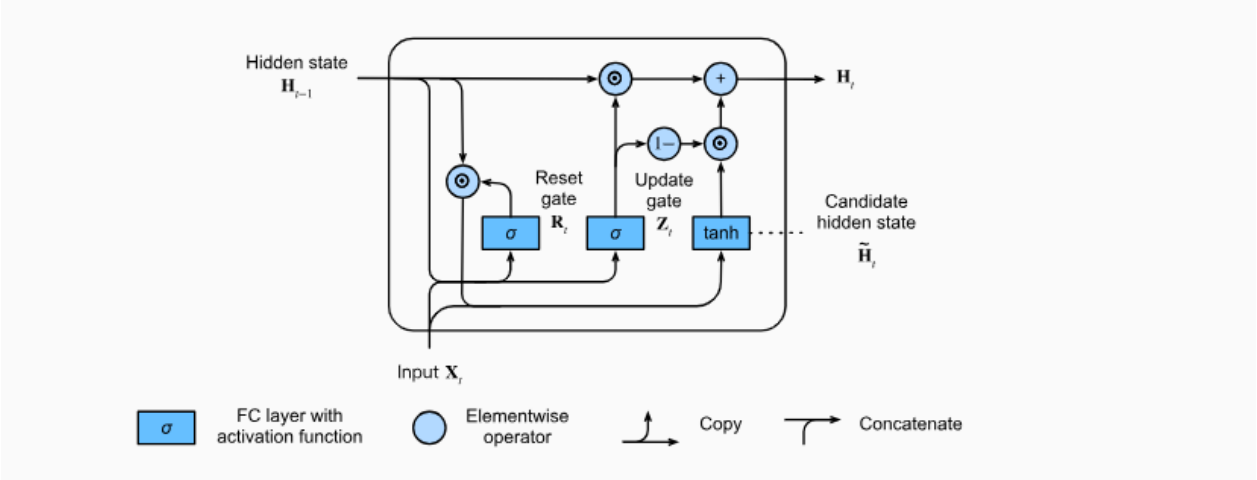
$$\tilde{\mathbf{h}}_i = \phi(\mathbf{W}_h^T \mathbf{x}_i + \mathbf{U}_h^T (\mathbf{h}_{i-1} \odot \mathbf{r}_i) + \mathbf{b}_h)$$

όπου ϕ είναι η υπερβολική εφαπτομένη.

Η τελική κρυφή αναπαράσταση παράγεται ώς ο κυρτός συνδυαμός (στοιχείο με στοιχείο) της τρέχουσας υποψήφιας κρυφής αναπαράστασης και της κρυφής αναπαράστασης της προηγούμενης χρονικής στιγμής χρησιμοποιώντας το διάνυσμα ενημέρωσης

$$\mathbf{h}_i = (1 - \mathbf{z}_i) \odot \mathbf{h}_{i-1} + \mathbf{z}_i \odot \tilde{\mathbf{h}}_i$$

Η αρχιτεκτονιή του κυττάρου των GRU φαίνεται στην Εικόνα 2.2.5.



Εικόνα 2.2.5: Παράδειγμα αρχιτεκτονικής του κυττάρου ενός αναδρομικού δικτύου με πύλες(GRU). Εικονα από [Zha+23]

Δίκτυα Μακρυάς Βραχυπρόθεσμης Μνήμης Ομοίως με τις GRU τα Δίκτυα Μακρυάς Βραχυπρόθεσμης Μνήμης (Long Short Term Memory -LSTM) χρησιμοποιούν διανύσματα ελεγκτές και κρυφές αναπαραστάσεις. Η ειδοποιός διαφόρα από τις GRU είναι οτι τα διανύσματα ελεγκτές δεν υπολογίζονται με βάση την προηγούμενη κρυφη αναπαράσταση, εντούτοις παράγεται ένα ξεχωριστό διάνυσμα εξόδου το οποίο εκτος από το οτι χρησιμοποιείται στον υπόλογισμό των διανυσμάτων ελεγκτών αποτελεί και την έξοδο του μοντέλου κάθε χρονική στιγμή. Πιο συγκεκριμένα το μοντέλο υπολογίζει τα τρια διανύσματα ελεγκτές $\mathbf{f}_i \in R^{d_f}$, $\mathbf{o}_i \in R^{d_o}$, $\mathbf{i}_i \in R^{d_i}$ τα οποία ονομάζονται διανύσματα *λήθης*,*επιλογής εξόδου* και *ενημέρωσης* αντίστοιχα και υπολογίζονται ως

$$\mathbf{f}_i = \sigma(\mathbf{W}_f^T \mathbf{x}_i + \mathbf{U}_f^T \mathbf{h}_{i-1} + \mathbf{b}_f)$$
$$\mathbf{o}_i = \sigma(\mathbf{W}_o^T \mathbf{x}_i + \mathbf{U}_o^T \mathbf{h}_{i-1} + \mathbf{b}_o)$$
$$\mathbf{i}_i = \sigma(\mathbf{W}_i^T \mathbf{x}_i + \mathbf{U}_i^T \mathbf{h}_{i-1} + \mathbf{b}_i)$$

όπου $\mathbf{h}_i$ είναι το διάνυσμα εξόδου της προηγούμενης χρονικής στιγμής και σ η σιγμοειδής συνάρτηση.

Συνοπτικά η λειτουργία τους είναι η εξής:

- $\mathbf{f}_i$: Επιλέγει τα στοιχείο που θα «σβηστούν» από την τρέχουσα κρυφή αναπαράσταση.
- $\mathbf{o}_i$: Επιλέγει τα στοιχεία που θα χρησιμοποιηθούν από την κρυφή αναπαράσταση για τον υπολογισμό του διανύσματος εξόδου.
- $\mathbf{i}_i$: Επιλέγει τα στοιχεία που θα χρησιμοποιηθούν από την υποψήφια κρυφή αναπαράσταση για να ενσωματωθούν στην νέα κρυφή αναπαράσταση.

Στη συνέχεια υπολογίζεται η υποψήφια κρυφή αναπαράσταση $\tilde{\mathbf{c}}_i$ ως

$$\tilde{\mathbf{c}}_i = \phi(\mathbf{W}_c^T \mathbf{x}_i + \mathbf{U}_c^T \mathbf{h}_{i-1} + \mathbf{c}_i)$$

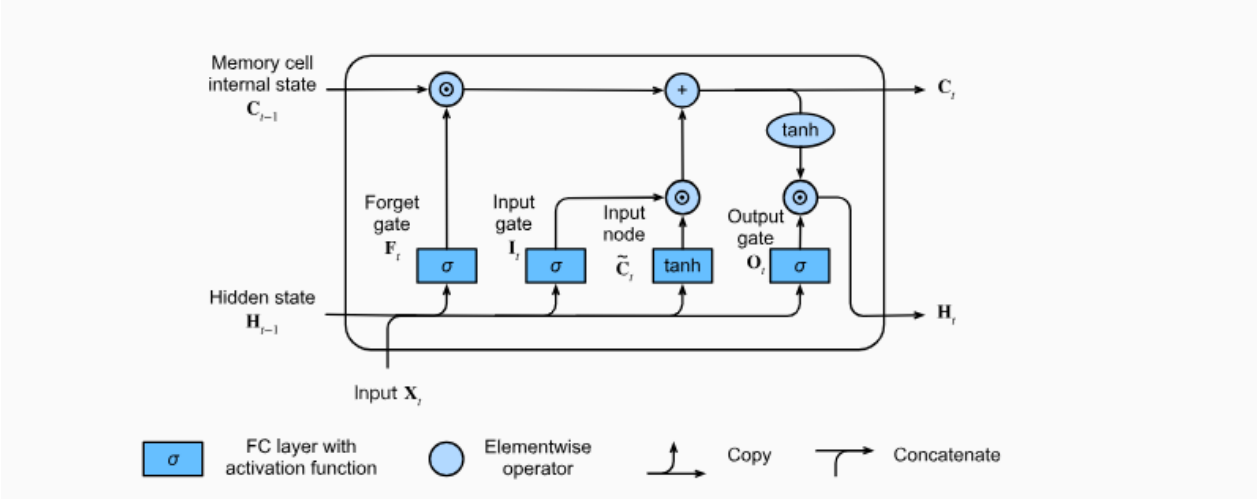
όπου ϕ είναι η υπερβολική εφαπτομένη.

Η νέα κρυφή αναπαράσταση $\mathbf{c}_i$ και το διάνυσμα εξόδου $\mathbf{h}_i$ υπολογίζονται ως

$$\mathbf{c}_i = \mathbf{f}_i \odot \mathbf{c}_{i-1} + \mathbf{i}_i \odot \tilde{\mathbf{c}}_i$$
$$\mathbf{h}_i = \mathbf{o}_i \odot \phi(\mathbf{c}_i)$$

Η αρχιτεκτονική του κυττάρου των LSTM φαίνεται στην Εικόνα 2.2.6.

Οι παραπάνω αρχιτεκτονικές, προκειμένου να αυξηθεί η ικανότητα αναπαράστασης του μοντέλου, επαναλαμβάνονται έτσι ώστε είτε η κρυφή αναπαράσταση είτε το διάνυσμα εξόδου να αποτελεί την είσοδο για το επόμενο



Εικόνα 2.2.6: Παράδειγμα αρχιτεκτονικής του κυττάρου ενός αναδρομικού δικτύου με μακρυά βραχυπρόθεσμη μνήμη(LSTM). Εικονα από [Zha+23]

επίπεδο. Επιπλέον όταν δεν υπάρχει ο περιορισμός της αιτιατότητας τότε τα δίκτυα αποτελούνται από 2 κρυφές αναπαραστάσεις για κάθε χρονική στιγμή, μια που προκύπτει από την εφαρμογή του κυττάρου στην ακολουθία με τη σωστή χρονική σειρά και μια που προκύπτει βλέποντας την ακολουθία με την αντίστροφη χρονική σειρά.

2.2.4 Μετασχηματιστές

Αν και τα παραπάνω δίκτυα αντιμετωπίζουν έως ένα βαθμό τα προβλήματα των απλών αναδρομικών δικτύων, παρουσιάζουν προβλήματα για ακόμα μεγαλύτερες ακολουθίες εισόδου. Επιπλεόν ένα μειονέκτημα των παραπάνω αρχιτεκτονικών είναι η ακολουθιακή διαχείρηση της εισόδου η οποία εισάγει μεγάλες χρονικές καθυστερήσεις και εμποδίζει την παράλληλη επεξεργασία των δεδομένων.

Τις παραπάνω αρχιτεκτονικές ήρθε να ξεπεράσει η αρχιτεκτονική των Μετασχηματιστών [Vas+17]. Το ισχυρό σημείο των παραπάνω αρχιτεκτονικών είναι ο μηχανισμός προσοχής(Attention). Πιο συγκεκριμένα επιτρέπει σε κάθε διάνυσμα της ακολουθίας εισόδου να εστιάσει σε οποιοδήποτε από τα υπόλοιπα διανύσματα εισόδου ώστε να εξάγει την κρυφή του αναπαράσταση.

Αναλυτικότερα έστω $\mathbf{x} = [x1, x_2, ..., x_n] \in R^{n \times d}$ η ακολουθία εισόδου. Χρησιμοποιώντας τρείς πίνακες

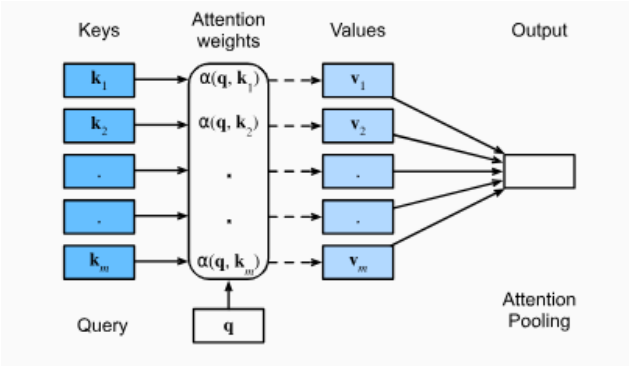
$$W_Q \in R^{d \times d_q}$$
$$W_K \in R^{d \times d_q}$$
$$W_V \in R^{d \times d_v}$$

εξάγονται οι ακολουθίες *ερωτημάτων* $\mathbf{Q} = [q_1, q_2, ..., q_n] \in R^{n \times d_q}$, *κλειδιών* $\mathbf{K} = [k_1, k_2, ..., k_n] \in R^{n \times d_q}$ και *τιμών* $\mathbf{V} = [v_1, v_2, ..., v_n] \in R^{n \times d_v}$.

Για το i-οστό στοιχείο της ακολουθίας εισόδου χρησιμοποιείται το διάνυσμα "ερώτησης" q_i ώστε να συγκριθεί με τα υπόλοιπα διανύσματα της εισόδου μέσω των διανυσμάτων "κλειδιών" των k_j. Αφου βρεθεί το σκόρ του i-οστού στοιχείο μου όλα τα υπόλοιπα και με τον εαυτό του, το διάνυσμα των σκόρ κανονικοποείται με χρήση της συνάρτησης softmax ώστε τα συνολικά σκόρ να αθροίζουν στη μονάδα. Η κρυφή του αναπαράσταση δίνεται από το σταθμισμένο αθροισμα των διανυσμάτων τιμών. Πιο φορμαλιστικά ο τρόπος εξαγωγής της κρυφής αναπαράστασης $\mathbf{h}_i \in R^{d_v}$ περιγράφεται από τις σχέσεις:

$$\mathbf{c} = q_i \mathbf{K}^T$$
$$\tilde{\mathbf{c}} = Softmax(\mathbf{c})$$
$$\mathbf{h}_i = \tilde{\mathbf{c}} \mathbf{V}$$

Ο μηχανισμός προσοχής παρουσιάζεται στην Εικονα 2.2.7



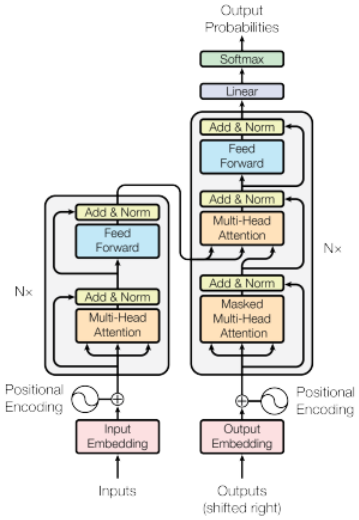
Εικόνα 2.2.7: Οπτική αναπαράσταση του τρόπου λειτουργίας του μηχανισμού προσοχής. Εικονα από [Zha+23]

Οπως φαίνεται από τις παραπάνω σχέσεις οι διαδικασία για κάθε χρονική στιγμή είναι ανεξάρτητη. Επομένως μπορούν να υπολογιστούν οι κρυφές αναπαραστάσεις για όλες τις χρονικές στιγμές παράλληλα. Η σχέση που δίνει τις τελικές κρυφές αναπαραστάσεις για όλες τις χρονικές στιγμές είναι ο διάσημος τύπος του μηχανισμού προσοχής:

$$H = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_q}})\mathbf{V}$$

όπου ο όρος του παρονομαστή λειτουργεί ως σταθερά κανονικοποίησης.

Η αρχιτεκτονική του μετασχηματιστή αποτελείται έναν κωδικοποιητή και εναν αποκωδικοποιητή. Ο κωδικοποιητής(Transformer Encoder) αποτελείται από πολλά επίπεδα μηχανισμού προσοχής όπου μεταξύ τους παρεμβάλονται στρώσεις πλήρως συνδεδεμένων δικτύων και στρώσεις κανονικοποίησης(επεξηγούνται αργότερα). Ο αποκωδικοποιητής ακολουθεί την ίδια αρχιτεκτονική ωστόσο περιέχει και μηχανισμούς Cross Προσοχής όπου οι πίνακες $\mathbf{K}, \mathbf{V}$ εξάγονται από διαφορετική ακολουθία από αυτή που εξάγεται ο πίνακας $\mathbf{Q}$. Η συνολική αρχιτεκτονική φαίνεται στην Εικόνα 2.2.8



Εικόνα 2.2.8: Ο κωδικοποιητής(αριστερά) και αποκωδικοποιητής(δεξιά) της αρχιτεκτονικής του μετασχηματιστή. Εικονα από [Vas+17]

2.2.5 Βοηθητικές Στρώσεις

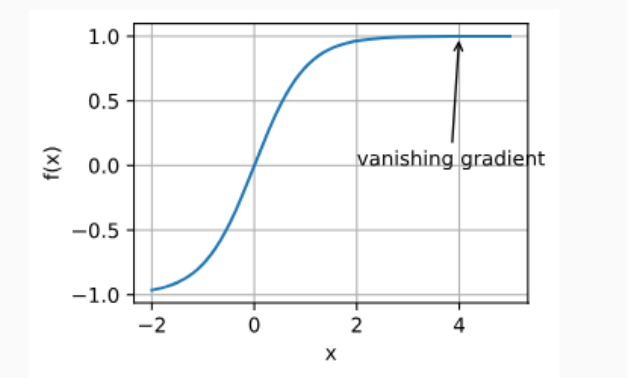
Πειραματικά έχει αποδειχθεί οτι ορισμένες σχεδιαστικές επιλογές στην αρχιτεκτονική του μοντέλου βοηθούν αφενώς το μοντέλο να συγκλίνει γρηγορότερα και αφετέρου να αποφύγει τοπικά ελάχιστα. Σε αυτή την υποενότητα παρουσιάζουμε τις πιο δημοφιλείς από αυτές.

Υπολλειματικές συνδέσεις

Οπως αναφέρθηκε παραπάνω ορισμένες συναρτήσεις ενεργοποίησης μπορει να δυσκολέψουν την διαδιακασία εκπαίδευσης καθώς οπως φαίνεται στην Εικόνα 2.2.9 η παράγωγος στα σημεία μακρυά από την αρχή των αξόνων είναι πολύ μικρή. Με αυτό το κίνητρο οι ερευνητές είσηγαγαν την έννοια των υπολλειματικών συνδέσεων (residual connections) κατα τις οποίες η είσοδος σε ένα υποσύστημα f παρακάμπτει το σύστημα και προστίθεται στην έξοδο του. Η συνολική έξοδος δίνεται από τη σχέση

$$y = f(x) + x$$

και πλεόν η ροή της παραγώγου προς τα πίσω εξασφαλίζεται από τον δεύτερο όρο ακόμα και αν η παράγωγος της f είναι μηδενική στον συγκεκριμένο σημείο. Ένα οπτικό παράδειγμα του προβλήματος και της παραπάνω διαδικασίας φαίνεται στην Εικόνα 2.2.9



Εικόνα 2.2.9: Πρόβλημα μηδενισμού της παραγώγου οταν βρισκόμαστε σε επίπεδο σημείο της συνάρτησης ενεργοποίησης. Εικόνα από [Zha+23]

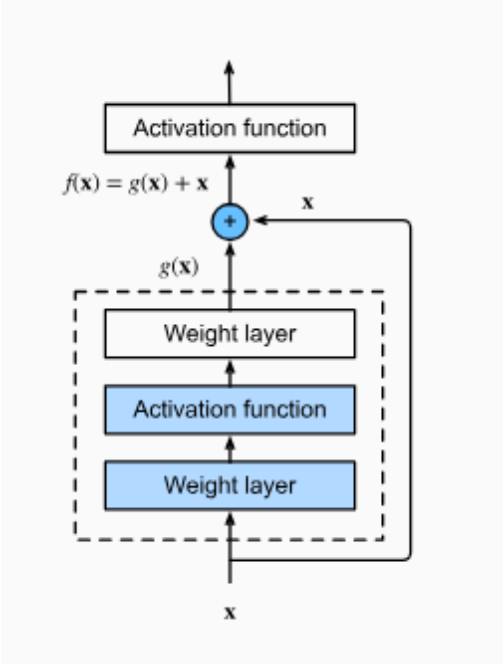
Κανονικοποίηση

Ομοίως με την παραπάνω τενχική έχει πειραματικά αποδειχτεί πως η προσθήκη στρώσεων κανονικοποίησης ανάμεσα στα στις υπόλοιπες στρώσεις των δικτύων επιταχύνει και κάνει πιο σταθερή την εκπαίδευση του δίκτύου. Η στρώση κανονικοποίησης αφού πρώτα κανονικοποιήσει τα δεδομένα ως προς μια από τις διαθέσιμες διαστάσεις εφαρμόζει ένα αφινικό μετασχηματισμό με εκπαιδευόμενες παραμέτρους. Οι τένσορες στα ενδιάμεσα στάδια του δικτύου έχουν τουλάχιστον 2 διαστάσεις, την διάσταση του μικρού συνόλου δεδομένων (batch dimension) και την διάσταση των χαρακτηριστικών (feature dimension). Τα είδη κανονικοποίησης δακρίνονται ανάλογα με τον διάσταση ως προς την οποίο γίνεται η κανονικοποίηση. Τα δύο πιο δημοφιλή είδη είναι:

- Κανονικοποίηση ως προς μικρό σύνολο δεδομένων (Batch Normalization): Στην συγκεκριμένη περίπτωση κανονικοποιούνται οι αναπαραστάσεις ως προς το πλήθος τον δεδομένων το οποία έχουν δωθεί ως είσοδος στο μοντέλο για το συγκεκριμένο βήμα εκπαίδευσης.
- Κανονικοποίηση ως προς την τελευταία διάσταση των χαρακτηριστικών: Στην προκειμένη περίπτωση η κανονικοποίηση γίνεται ως προς το κάθε δείγμα ξεχωριστά.

2.2.6 Εκπαίδευση

Εκπαίδευση ονομάζεται η διαδικασία της μάθησης, κατα την οποία το μοντέλο κάνει χρήση των διαθέσιμων δεδομένων για να μεταβάλει τις παραμέτρους ώστε να αυξηθεί η επίδοση του ως προς το κριτήριο εκπαίδευσης



Εικόνα 2.2.10: Παράδειγμα αρχιτεκτονικής που κανει χρήση υπολλειματικών συνδέσεων. Εικονα από [Zha+23]

L. Στην επιβλεπόμενη μάθηση το κριτήριο εκπαίδευσης L εκφράζει την διαφορά μεταξύ των επισημειώσεων και των εκτιμήσεων του μοντέλου. Ανάλογα με το ειδος επιβλεπόμενης μάθησης το κριτήριο λαμβάνει συνήθως μια από τις δύο παρακάτω μορφές:

- Ταξινόμηση: Το κριτήριο εκπαίδευσης επιλέγεται να είναι η Cross-Εντροπία(Cross Entropy) μεταξύ των πραγματικών επισημειώσεων και των πιθανοτήτων που παράγει το μοντέλο για κάθε κατηγορία. Το λάθος δίνεται από τον τύπο

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \log(\tilde{y}_k^{(i)})$$

  όπου N είναι ο αριθμός των δειγμάτων, K είναι ο αριθμός των κλάσεων, $y^{(i)}$ είναι η one-hot αναπαράσταση των επισημειώσεων και $\log(\tilde{y}^{(i)})$ είναι το διάνυσμα με τις K log πιθανότητες που εκτιμά το μοντέλο. Η ελαχιστοποίηση της Cross Εντροπίας ισοδυναμεί με την μεγιστοποίηση της log πιθανοφάνειας του συνόλου δεδομένων ως προς την κατανομή που εκτιμά το μοντέλο.

- Παλινδρόμηση: Το κριτήριο εκπαίδευσης είναι το Μέσο Τετραγωνικό Λάθος(Mean Squared Error) ανάμεσα στις επισημειώσεις και τις εκτιμήσεις του μοντέλου και δίνεται από τον τύπο:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2$$

  όπου N είναι ο αριθμός των δειγμάτων.

Ομοίως, στο σενάριο της μη επιβλεπόμενης μάθησης επιλέγεται ένα παραγωγίσιμο κριτήριο εκπαίδευσης όπου αξιολογεί την επίδοση του μοντέλου στο συγκεκριμένο πρόβλημα υπολογισμένο στα δεδομένα εκπαίδευσης.

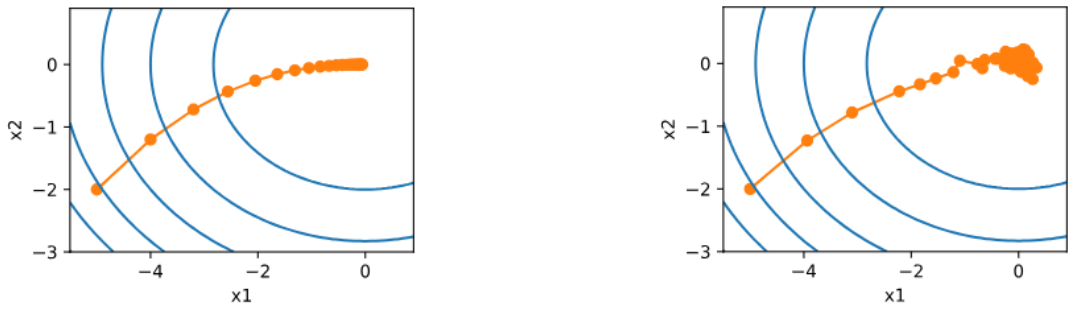
Τοσο τα δίκτυα οσο και τα κριτήρια εκπαίδευσης είναι παραγωγίσιμες συναρτήσεις επομένως μπορούν να υπολογιστούν οι μερικές παράγωγοι ως προς τις παραμέτρους του δικτύου. Ωστόσο στην γενική περίπτωση δεν μπορεί να υπολογιστεί ο κλειστός τύπος που δίνει τα κρίσιμα σημεία λόγως της πολυπλοκότητας των δικτύων. Για αυτό το λόγο η εκπαίδευση του μοντέλου γίνεται με επαναληπτικούς αλγορίθμους που βασίζονται στην παράγωγο ώστε να κάνουν βήματα προς τα ελάχιστα του κριτηρίου εκπαίδευσης.

Ο βασικός αλγόριθμος είναι ο Αλγόριθμος Κατάβασης Κλίσης(Gradient Descent) όπου οι παράμετροι $\mathbf{w}$ του

δικτύου ενημερώνονται σύμφωνα με τον κανόνα:

$$\mathbf{w} = \mathbf{w} - \alpha \frac{\partial L}{\partial \mathbf{w}}$$

όπου α είναι ο ρυθμός μάθησης. Δεδομένου οτι το κριτήριο εκπαίδευσης L υπολογίζεται πάνω σε ένα μικρό σύνολο δεδομένων κάθε φορά ο αλγόριθμος ονομάζεται Στοχαστικός Αλγοριθμός Κατάβασης Κλίσης. Το γεγονός πως οι παράγωγοι κάθε φορά αντιστοιχούν σε ένα διαφορετικό σύνολο δειγμάτων εισάγει θόρυβο στην διαδικασία εκπαίδευσης οπως μπορεί να φανεί στην Εκόνα 2.2.11



(a) Παράδειγμα κατάβασης κλίσης.

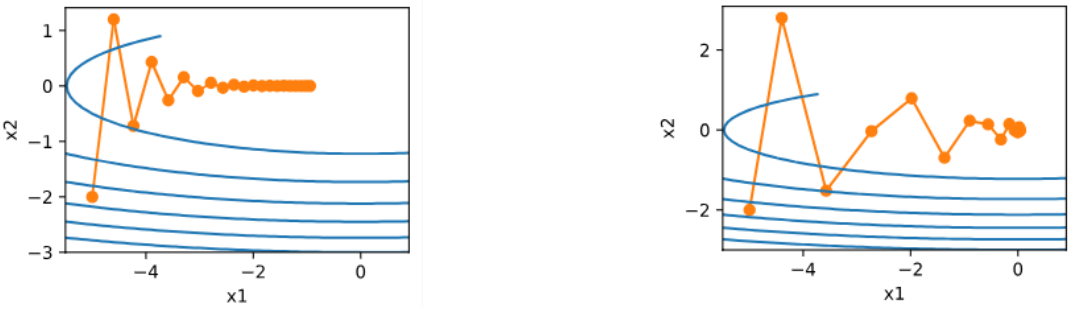
(b) Παράδειγμα στοχαστικής κατάβασης κλίσης.

Εικόνα 2.2.11: Παράδειγμα όπου ο αλγόριθμος στοχαστικής κατάβασης κλίσης παράγει πιο θορυβώδης τροχιές. Εικόνες από [Zha+23]

Η βασική ιδεα του αλγόριθμου είναι η κίνηση προς την κατεύθυνση με την πιο απότομη κλίση. Ωστόσο αυτή η στρατηγική είναι επιρρεπής στο να βρίσκει τοπικά ελάχιστα. Προκειμένου να αντιμετωπιστεί το πρόβλημα οι ερευνητές είσηγαγαν επεκτάσεις στον παραπάνω αλγόριθμο οπως:

- η προσθήκη όρου αδράνειας ώστε να μην αλλάζει απότομα η κατευθυνσή κίνησης.
- η χρήση αυτορυθμιζόμενου ρυθμού μάθησης.
- η χρήση διαφορετικύ ρυθμού μάθησης για διαφοτερικές παραμέτρους του δικτύου.

ένα οπτικό παράδειγμα όπου φαίνεται η επίδραση του όρου αδράνειας φαίνεται στην Εικόνα 2.2.12.



(a) Χωρίς προσθήκη όρου αδράνειας.

(b) Με προσθήκη όρου αδράνειας.

Εικόνα 2.2.12: Παράδειγμα όπου ο αλγόριθμος στοχαστικής κατάβασης κλίσης λόγω διαφορετικής κλίσης στις δύο κατευθύνσεις αδυνατεί να καταλήξει στο σημείο ελαχίστου. Αντιθέτως με την προσθήκη του όρου αδράνειας επιτυχώς καταλήγει στο σημείο ελαχίστου. Εικόνες από [Zha+23]

2.2.7 Γεννετικά Μοντέλα

Σε αντίθεση με τα μοντέλα παλινδρομησης όπου παράγουν σημειακές εκτίμησεις, τα γεννετικά μοντέλα παράγουν μια κατανομή πάνω στον χώρο εξόδου από την οποία συνήθως μπορούμε να δειγματοληπτίσουμε ώστε να πάρουμε δείγματα εξόδου. Οταν η κατανομή που αναπαριστά το μοντέλο $p_{model}(x)$ προσεγγίζει την κατανομή των

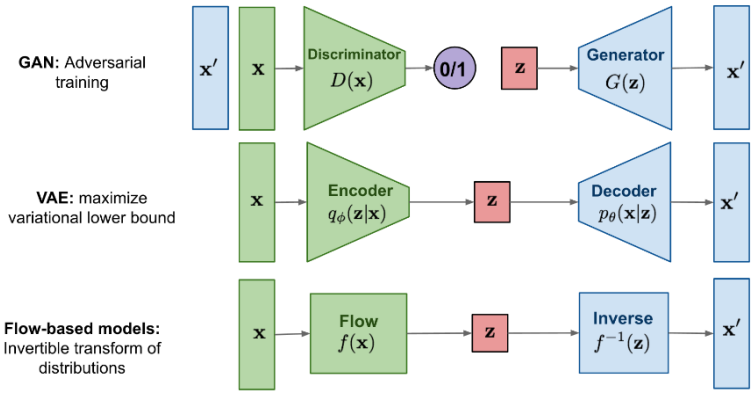
δεδομέων $p_{data}(x)$ τότε το αποτέλεσμα της δειγματοληψίας θα ομοιάζει με τα δείγματα από το σύνολο δεδομένων. Δημοφιλή παραδείγματα τέτοιων μοντέλων αποτελούν τα:

- Variational AutoEncoders: Τα μοντέλα που ακολουθούν αυτό το παράδειγμα κωδικοποιούν τα δεδομένα σε ένα χώρο μικρότερης διάστασης ο οποίος κατέχει εκ κατασκευής του μια δομή. Ομοίως με τους απλούς AutoEncoders, αποτελούνται από εναν Κωδικοποιητή που κωδικοποιεί τα δεδομένα σε αυτο τον *κρυφό χώρο* και έναν Αποκωδικοποιητή που λαμβάνει σαν είσοδο ένα σημείο το κρυφού χώρου και το αποκωδικοποιεί σε ένα σημείο στο χώρο εξόδου. Η ειδοποιός διαφορά μεταξύ των Variational AutoEncoders και των AutoEncoders είναι ο τρόπος εκπαίδευσης, κατα τον οποίο ο Αποκωδικοποιητής θεωρείται ως η κατανομή εξόδου υπο συνθήκη $P(x|z)$ όπου $z \in R^d$ είναι έναν σημείο του κρυφού χώρου ενώ ο Κωδικοποιητής λειτουργεί ως Προσεγγιστική Εκ Των Υστέρων Κατανομή(Approximate Posterior Distribution) $Q(z|x)$ και το μοντέλο εκπαιδεύεται ωστε να μεγιστοποιήσει προσεγγιστικά την πιθανοφάνεια των δειγμάτων του συνόλου δεδομένων.

- Generative Adversarial Networks: Τα μοντέλα που ανήκουν στην συγκεκριμένη κατηγορία αποτελούνται από ένα υπομοντέλο Γεννήτορα,ο οποίος λαμβάνει μια στοχαστική είσοδο και παράγει δείγματα του χώρου εξόδου, και από ένα υπομοντέλο Κριτή ο οποίος αναλαμβάνει να διακρίνει από ένα σύνολο δεδομένων ποια δείγματα από αυτά προέρχεται από την πραγματική κατανομή δεδομένων $p_{data}(x)$ και ποια από την κατανομή του μοντέλου $p_{model}(x)$. Κατα την διάρκεια της εκπαίδευσης, το μοντέλο Κριτής προσπαθεί να διαχωρίσει σωστά τις εισόδους του ενώ το μοντέλο Γεννήτορας προσπαθεί να μειώσει την απόδοση του Κριτή.

- Normalizing Flows: Τα μοντέλα αυτού του παραδείγματος αποτελούνται από ένα αντιστρέψιμο νευρωνικό δίκτυο το οποιο λαμβάνει σαν είσοδο ένα δείγμα από μια βασική κατανομή και παράγει ένα δείγμα από την κατανομή του χώρου εξόδου. Κατα την διάρκεια της εκπαίδευσης μεγιστοποιούμε την πιθανοφάνεια των δεδομένων εκπαίδευσης εκμεταλλευόμενοι την αντιστρεψιμότητα του δικτύου και την ιδιότητα της αλλαγής μεταβλητών.

- AutoRegressive Models: Σε αυτή τη κατηγορία τα μοντέλα δημιουργούν σειριακά το δείγμα εξόδου(υποθέτοντας οτι το δείγμα εξόδου αποτελείται από πολλά στοιχεία) κάνοντας αρχικά μια εκτίμηση για το πρώτο στοιχείο του και στην συνέχεια χρησιμοποιώντας τα ήδη παραγμένα στοιχεία ώστε να παράξουν τα επόμενα. Ομοίως με τα Normalizing Flows κατα την εκπαίδευση μεγιστοποιούμε την πιθανοφάνεια των δεδομένων εκπαίδευσης.



Εικόνα 2.2.13: Παραδείγματα αρχιτεκτονικών παράλληλων γεννετικών δικτύων. Εικονα από [Wen21]

Η επιλογή μεταξύ ενός από τα παραπάνω μοντέλα γίνεται με βάση τις ιδιότητες που ενδιαφέρουν για το εκάστοτε πρόβλημα. Για παράδειγμα τα AutoRegressive μοντέλα είναι ισχύρά ώστοσο η σειριακή διαδικασία δειγματοληψίας τα καθιστά μη αποδοτικά για πρακτικές εφαρμογές. Μεταξύ των υπόλοιπων κλάσσεων που η διαδικασία δειγματοληψίας γίνεται σε ένα βήμα τα VAEs,GAN είναι πιο ισχυρά ωστόσο δεν διαθέτουν την ικανότητα ακριβούς υπολογισμού της πιθανοφάνειας ενός δείγματος. Τέλος οσον αφορά στην σύγκριση μεταξύ των VAEs και GANs, τα GANs παράγουν μεγαλύτερης ποιότητας δείγματα ωστόσο υποφέρουν από θέματα αστάθειας κατα την εκπαίδευση.

Δεδομένου οτι στο επόμενο κεφάλαιο θα παρουσιάσουμε ένα μοντέλο που κάνει χρήση των Normalizing Flows, παρουσιάζουμε παρακάτω συνοπτικά τον τρόπο λειτουργίας τους.

Normalizing flows

Τα normalizing flows ανήκουν στην κατηγορία των γεννετικών μοντέλων. Σε αντίθεση με τα GANs [Goo+14] και τα VAEs [KW13; RMW14] δίνουν την δυνατότητα να υπολογιστεί ή πιθανοφάνεια ενός δείγματος, γεγονός που τα κάνει χρήσιμα για αναπαράσταση πρότερης γνώσης. Βασίζονται στην μέθοδο της αλλαγής μεταβλητών. Πιο συγκεκριμένα αν z τυχαία μεταβλητή που ακολουθεί μια κατανομή p_z και $x = f(z)$ τυχαία μεταβλητή όπου η f είναι μια 1-1 και παραγωγίσιμη συνάρτηση τοτε η x ακολουθεί την κατανομή:

$$p_x(x) = p_z(f^{-1}(x))|det(\frac{\partial f^{-1}(x)}{\partial x})| \tag{2.2.1}$$

Τότε αν επιλέξουμε σαν p_z μια γνωστή κατανομή που γνωρίζουμε να υπολογίζουμε την πιθανοφάνεια ενος δείγματος και επιπλέον μπορούμε να υπολογίσουμε την ορίζουσα της ιακοβιανής τότε μπορούμε να υπολογισουμε την πιθανοφάνεια του δείγματος x. Οι έρευνες σε αυτο το πεδίο εστιάζουν στην παραμετροποίηση των συναρτήσεων f με νευρωνικά δίκτυα σχεδιασμένα έτσι ωστε η f να είναι αντιστρέψιμη και η ορίζουσα της ιακωβιανής να υπολογίζεται αποδοτικά. Από τις πιο διαδεδομένες μεθόδους είναι τα affine coupling layers [DSDB17] και οι επεκτάσεις τους [KD18] όπου η συνάρτηση $y = f(x)$ ορίζεται ως:

$$\begin{aligned} y_{1:d} &= x_{1:d} \\ y_{d+1:D} &= x_{d+1:D} * exp(s(x_{1:d})) + t(x_{1:d}) \end{aligned} \tag{2.2.2}$$

2.3 Παραμετρικά Μοντέλα

Η γεωμετρία ενός τρισδιάστατου αντικειμένου αναπαρίσταται συνήθως στον υπολογιστή ώς ενα πλέγμα από κόμβους(mesh). Προκειμένου να δημιουργηθεί αυτο το σύνολο, στο πεδίο της όρασης υπολογιστών και των γραφικών υπολογιστών χρησιμοποιούνται τα 3D Morphable Models, ένα είδος παραμετρικών μοντέλων, τα οποία λαμβάνουν σαν είσοδο έναν αρκετά μικρότερο αριθμό παραμέτρων και παράγουν σαν έξοδο τις συντεταγμένες κάθε ενός από τους κόμβους του πλέγματος. Τα 3DMM Στην περίπτωση του ανθρωπίνου σχήματος οι ερευνητές Loper et. al [Lop+15] δημιούργησαν το παραγωγίσιμο παραμετρικό μοντέλο SMPL από τρισδιάστατα σκαναρίσματα πολλών διαφορετικών ανθρωπίνων σωμάτων. Πιο συγκεκριμένα το μοντέλο λαμβάνει σαν είσοδο παραμέτρους σχήματος $\boldsymbol{\beta} \in R^{10}$ και τις σχετικές περιστροφές για κάθε μια από τις βασικές 24 αρθρώσεις $\boldsymbol{\vartheta} \in R^{24x3}$ με την μορφή axis-angle(ένα 3Δ διάνυσμα οπου η κατεύθυνση του είναι ο άξονας περιστροφής και η νόρμα του εκφράζει την γωνία περιστροφής), και παράγει σαν έξοδο $N = 6890$ κόμβους που αναπαριστούν το πλέγμα, $M \in R^{3N}$. Παράλληλα παρέχεται και ενας εκτιμητής $\mathbf{J} \in R^{24\times 6890}$ που παράγει τις 24 θέσεις των αρθρώσεων σαν γραμμικό συνδυασμό των κόμβων του πλέγματος. Το μοντέλο αποτελείται από ένα αρχικό πλέγμα $\mathbf{T} \in R^{3N}$ το οποίο ορίζεται να έχει το συγκεκριμένο σχήμα και την συγκεκριμένη πόζα αναφοράς όπως φαίνεται στην εικόνα 2.3.1. Κατα την διάρκεια της εκπαίδευσης του μοντέλου SMPL οι ερευνητές υπολόγισαν τις κύριες διευθύνσεις σχήματος(Principal Components) $\mathbf{S}_i \in R^{3N}$ πάνω σε ένα σύνολο από σκαναρισμένα σώματα διαφορετικού σχήματος τοποθετημένα στην πόζα αναφοράς. Οι παράμετροι σχήματος αποτελούν τους συντελεστές με τους οποίους προστίθενται οι S_i δημιουργώντας τον "διορθωτή σχήματος"

$$B_s(\boldsymbol{\beta}) = \sum_i \beta_i \mathbf{S}_i$$

Ομοίως με τον χώρο του σχήματος, κατα την διάρκεια της εκπαίδευσης του μοντέλου SMPL μαθαίνονται οι διευθύνσεις "διορθωτή πόζας" $P_i \in R^{3N}$ οι οποίες προστίθενται με συντελεστές τα στοιχεία της ανιγμένης πόζας

$$R^*(\boldsymbol{\vartheta}) = R(\boldsymbol{\vartheta}) - R(\boldsymbol{\vartheta}^*)$$

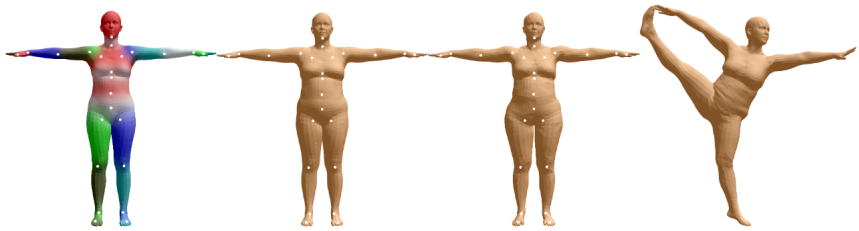
όπου $R(\boldsymbol{\vartheta}) \in R^{23*9}$ είναι το συνολικό διάνυσμα πόζας περιλαμβάνοντας ολα τα στοιχεία από τους 23 πίνακες περιστροφής(εξαιρείται η περιστροφή της πρώτης άρθρωσης που μοντελοποιεί την περιστροφή ως προς το σύστημα αναφοράς) και $\boldsymbol{\vartheta}^*$ είναι η παράμετροι της "ουδέτερης πόζας"(rest post) η οποία υπολογίζεται και αυτή κατα την διαδικασίας εκπαίδευσης του SMPL. Επομένως ο συνολικός διορθωτής πόζας δίνεται από την σχέση

$$B_p(\boldsymbol{\vartheta}) = R^*(\boldsymbol{\vartheta})\mathbf{P}^T$$

με $\mathbf{P} = [P_1, P_2, ...P_{207}] \in R^{3N \times 207}$. Το συνολικό πλέγμα στην ουδετερη πόζα δίνεται πλέον από την σχέση

$$\mathbf{T}_p = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\vartheta})$$

από το πλέγμα $\mathbf{T}_p$ και τον γραμμικό εκτιμητή $\mathbf{J}$ λαμβάνονται οι θέσεις των 24 αρθρώσεων, οι οποίες περιστρέφονται σύμφωνα με τις παραμέτρους πόζας $\boldsymbol{\vartheta}$ και παράγουν τον τελικό σκελετό. Στην συνέχεια ο αλγόριθμος Linear Blend Skinning παίρνει το τελικό πλέγμα στην ουδέτερη πόζα και τον τελικό σκελετό και παράγει το τελικό πλέγμα στην σωστή πόζα μετασχηματίζοντας κάθε κόμβο του πλέγματος με έναν γραμμικό συνδυασμό των μετασχηματισμών των κοντινότερων αρθρώσεων του. Η διαδικασία φαίνεται στην Εικόνα 2.3.1. Το μοντέλο SMPL παρέχεται σε τρεις εκδόσεις: male, female και neutral ανάλογα με το σύνολο στο οποίο έχει εκπαιδευτεί ενώ επίσης εχουν αναπτυχθεί εκδόσεις που ενσωματώνουν παραμετρικά μοντέλα χεριών [RTB17] και προσώπου [Pav+19].



Εικόνα 2.3.1: Η διαδικασία κατα την οποία οι παράμετροι σχήματος και πόζας παράγουν το τελικό πλέγμα. Εικόνα από [Lop+15]

2.4 Μοντέλα Κάμερας

Η σχέση μεταξύ των σημείων στον τρισδιάστατο χώρο και των σημείων στο επίπεδο της εικόνας καθορίζεται από το μοντέλο κάμερας που επιλέγεται. Το μοντέλο της Pinhole κάμερας αποτελεί το πιο διαδεδομένο μοντέλο κάμερας και αποτελεί μια καλή προσέγγιση της πραγματικής διεργασίας που συμβαίνει σε μια φωτογραφική μηχανή. Προσομοιώνει την διαδικασία κατα την οποία σε κάθε σημείο του αισθητήρα της ιδεατής κάμερας καταλήγει μόνο μια ακτίνα φωτός η οποία προέρχεται απο μόνο ένα συγκεκριμένο αντικείμενο της σκηνής. Πιο συγκεκριμένα, οριζουμε το κέντρο προβολής O να είναι η αρχή των αξόνων, το επίπεδο της εικόνας να ειναι το επίπεδο $z = -f$, $f > 0$ οπου f ονομάζεται το εστιακό βάθος και τα αντικείμενα του χώρου να ζούν στον ημίχωρο με θετικά Z. Προκειμένου να βρεθεί σε πιο σημείο του επιπέδου της εικόνας αντιστοιχεί ενα σημείο του χώρου $P = [X, Y, Z]$ προεκτείνεται μια ακτίνα από το σημείο του χώρου στο κέντρο προβολής και συνεχίζεται ώστε να βρεθεί το σημείο τομής $p = [x, y, -f]$ της με το επίπεδο της εικόνας . Οι συντεταγμένες του σημείου τομής δίνονται από την σχέση

$$x = -\frac{f}{Z}X$$
$$y = -\frac{f}{Z}Y$$

Λόγω του τρόπου προβολής, στο πεδίο της εικόνας τα αντικείμενα παρουσιάζονται ανεστραμμένα ως προς τους άξονες x, y. Για αυτό το λόγο αφαιρούμε το αρνητικό πρόσημο από τις παραπάνω σχέσεις το οποίο ισοδυναμεί με το επίπεδο της εικόνας να είναι το $z = f$.

Ένα πιο απλοποιημένο μοντέλο κάμερας είναι η κάμερα ορθογραφικής προβολής οπου τα σημεία προβάλλονται κατευθείαν στο επίπεδο της εικόνας, πρακτικά αγνοώντας την Z συνιστώσα κάθε σημείου. Το συγκεκριμένο μοντέλο αν και απλοικό δίνει ικανοποιητικά αποτελέσματα όταν τα αντικείμενα ενδιαφέροντος βρίσκονται μακρυά από το κέντρο προβολής. Τότε τα ευθυγραμμα τμήματα που συνδέουν τα αντικείμενα με το κέντρο προβολής μπορούν να θεωρηθούν παράλληλα δίνοντας το παραπάνω μοντέλο.

Το μοντέλο που συνδυάζει την απλότητα της ορθογραφικής κάμερας με τα ποιοτικά αποτελέσματα της pinhole κάμερας είναι η κάμερα ασθενούς προοπτικής (weak perspective) η οποία θεωρεί πως ένα σύνολο αντικειμένων είναι αρκετά μακριά από το κέντρο προβολής ώστε οι μεταβολες της Z συνιστώσας να θεωρούνται αμελητέες

και όλα τα αντικείμενα να θεωρείται πως βρίσκονται στο ίδιο βάθος $\bar{Z} = \frac{1}{N}\sum_i Z_i$. Στην συνέχεια εφαρμόζονται οι σχέσεις της pinhole καμερας επομένως οι τελικές θέσεις το πεδίο της εικόνας δίνονται από τις σχέσεις:

$$x = \frac{f}{\bar{Z}}X$$
$$y = \frac{f}{\bar{Z}}Y$$

Η πλειοψηφία των ερευνητικών εργασιών επιλέγει το τελευταίο μοντέλο καθώς οι μεταβολές της συνιστώσας Z των αρθρώσεων του ανθρώπινου σώματος θεωρούνται αμελητέες.

2.5 Κριτήρια αξιολόγησης εκτίμησης 3Δ Πόζας

Ανάλογα με το είδος των επισημειώσεων του σύνολο αξιολόγησης υπάρχει μια πληθώρα απο κριτήρια τα οποία μπορούμε να χρησιμοποιήσουμε ώστε να αξιολογήσουμε το αποτέλεσμα της ανακατασκευής. Στην πιο κοινή περίπτωση όπου οι επισημειώσεις βρίσκονται με την μορφή των θέσεων των αρθρώσεων στον τρισδιάστατο χώρο(Σκελετοί) οι μετρικές που χρησιμοποιούνται είναι οι:

- Mean per Joint Position Error(MPJPE): Προκειμένου να υπολογιστεί το λάθος μεταξύ δύο σκελετών, αρχικά οι δύο σκελετοί ευθυγραμμίζονται μετακινώντας την κεντρική τους άρθρωση, το σημείο ανάμεσα στις δυο αρθρώσεις της λεκάνης, στην αρχή των αξόνων. Στην συνέχεια υπολογίζεται το λάθος ως προς την ευκλείδεια απόσταση για κάθε άρθρωση και εξάγεται ο μέσος όρος.
- Procrustes Aligned Mean per Joint Position Error(PA-MPJPE): Σε αντίθεση με το παραπάνω κριτήριο οπου οι σκελετοί απλά μεταφέρονται ώστε να ευθυγραμμιστούν, στο συγκεκριμένο κριτήριο χρησιμοποιείται η μέθοδος του προκρούστη ώστε να βρεθεί *η περιστροφή* και μετακίνηση που ευθυγραμμίζει καλύτερα τους δυο σκελετούς. Η διαδικασία συνεχίζεται με την έυρεση το μέσου λάθος ως προς τις αρθρώσεις.

Σε περίπτωση όπου οι επισημειώσεις είναι με την μορφή των παραμέτρων SMPL τότε επιπλέον χρησιμοποιείται το κριτήριο:

- Mean Per Vertex Error(MPVE): Προκειμένου να υπολογιστεί το συγκεκριμένο κριτήριο αρχικά τα σύνολα κόμβων ευθυγραμμίζονται μετακινώντας τους ώστε η κεντρική τους αρθρωση,οπως προκύπτει από τις αρθρώσεις, να βρίσκεται στην αρχή των αξόνων. Στη συνέχεια υπολογίζεται το λάθος για κάθε κόμβο του πλέγματος και εξάγεται ο μέσος όρος.

Τέλος στην περίπτωση της εκτίμησης πόζας και σχήματος από βίντεο μας ενδιαφέρει η χρονική συνέπεια μεταξύ των εκτιμήσεων. Σαν κριτήριο αξιολόγησης χρησιμοποιείται η επιτάχυνση των αρθρώσεων.

Κεφάλαιο 3

Σχετική Βιβλιογραφία

3.1 Ιστορική Αναδρομή

Οπως αναφέρθηκε στο κεφάλαιο 1 η ανακατασκευή απο δισδιάστατα δεδομένα της πόζας είναι ένα εξ' ορισμού αμφίσημο πρόβλημα. Οι περισσότερες ερευνητικές εργασίες βασίζονται σε πρότερη γνώση με την μορφή παραμετρικών μοντέλων οπως αυτό που παρουσιάστηκε στο κεφάλαιο 2 προκειμένου να περιορίσουν κάποιους βαθμούς ελευθερίας του προβλήματος και να αποφύγουν μη εφικτές λύσεις. Στο κεφάλαιο αυτό θα παρουσιάσουμε τις μεθόδους που εχουν προταθεί και κάνουν χρήση παραμετρικών μοντέλων και θα επικεντρωθούμε στην κατηγορία των μεθόδων που χρησιμοποιεί νευρωνικά δίκτυα σε συνεργασία με τα παραμετρικά μοντέλα ώστε να αντιμετωπίσει το πρόβλημα.

Ανάλογα με την μορφή των επισημειώσεων, οι αρχικές προτεινόμενες μέθοδοι χρησιμοποιούσαν επαναληπτικούς αλγορίθμους για να μεταβάλλουν τις παραμέτρους του παραμετρικού μοντέλου ώστε να εξηγούνται καλύτερα οι διαθέσιμες παρατηρήσεις. Πιο συγκεκριμένα οι ερευνητές [Bog+16] εισάγουν την μέθοδο SMPLify, οπου χρησιμοποιούν το παραμετρικό μοντέλο SMPL ώστε να ανακατασκευάσουν την 3Δ πόζα και σχήμα με είσοδο μια εικόνα. Αρχικά χρησιμοποιούν ενα προεκπαιδευμένο μοντέλο εκτίμησης δισδιάστατης πόζας ώστε να εντοπίσουν τις θέσεις των αρθρώσεων στην εικόνα εισόδου. Εκκινώντας με τις μέσες παραμέτρους για το μοντέλο SMPL, εξάγουν τις τρισδιάστατες θέσεις των αρθρώσεων και τις προβάλλουν με χρήση μιας pinhole κάμερας στο πεδίο της εικόνας. Ως αντικειμενική συνάρτηση χρησιμοποιούν το μέσο τετραγωνικό σφάλμα ανάμεσα στις επισημειώσεις και τις εκτιμήσεις του μοντέλου. Δεδομένου οτι το παραμετρικό μοντέλο είναι παραγωγίσιμο ο επαναληπτικός αλγόριθμος χρησιμοποιεί τις μερικές παραγώγους του σφάλματος ως προς τις παραμέτρους του μοντέλου SMPL ωστε να τις μεταβάλλει.

Δεδομένου ότι το μοντέλο SMPL δεν εισάγει περιορισμό στις γωνίες περιστροφής των αρθρώσεων η βελτιστοποίηση της παραπάνω αντικειμενικής συνάρτησης οδηγεί συχνά σε μη ρεαλιστικές πόζες που ωστόσο εξηγούν τις επισημειώσεις. Για να αντιμετωπιστεί το πρόβλημα οι ερευνητές εισάγουν επιπλέον όρους στην αντικειμενική συνάρτηση που επιβάλλουν ποινή στο μοντέλο όταν υπάρχει διείσδυση μεταξύ μελών του σώματος, όταν υπάρχουν αρθρώσεις που εκτίνονται πέρα από το φυσικό όριο (π.χ. υπερέκταση αγκώνα) και όταν παράγονται πόζες με χαμηλή πιθανοφάνεια ως προς ενα προεκπαιδευμένο πιθανοτικό μοντέλο.

Συνοπτικά ο αλγόριθμος φαίνεται παρακάτω:

Algorithm 1: Αλγόριθμος SMPLify [Bog+16]

$\theta \leftarrow$ μέση παράμετρος πόζας
$\beta \leftarrow$ μέση παράμετρος σχήματος
repeat
  Υπολογισμός 3Δ θέσεων των αρθρώσεων απο το SMPL μοντέλο $X = J \cdot M(\theta, \beta)$
  Προβολή αρθρώσεων στο πεδίο της εικόνας $\tilde{x} = \Pi(Q)$
  Υπολογισμός σφάλαματος $L = MSE(x, \tilde{x}) + Prior(\theta)$
  Ενημέρωση παραμέτρων. $\theta = \theta - \alpha\frac{\partial L}{\partial \theta}, \quad \beta = \beta - \alpha\frac{\partial L}{\partial \beta}$
until *Σύγκλιση σφάλματος*

Αν και η παραπάνω μέθοδος παράγει ακριβή αποτελέσματα χρειάζεται περίπου 1 λεπτό για κάθε εικόνα το οποίο δεν είναι εφαρμόσιμο στην πράξη. Με την έλευση της βαθιάς μάθησης οι ερευνητές επικεντρώθηκαν στην δημιουργία μεθόδων που θα χρησιμοποιούν νευρωνικά δίκτυα προκειμένου να μάθουν τον μετασχηματισμό από την εκάστοτε είσοδο στο χώρο των παραμέτρων του μοντέλου SMPL. Στις επόμενες ενότητες παρουσιάζουμε τέτοιες μεθόδους τόσο από στατική εικόνα όσο και απο βίντεο.

3.2 Μεθοδοι εκτίμησης από εικόνα.

Σε αυτή τη κατηγορία ανήκουν όλες οι μέθοδοι που χρησιμοποιούν νευρωνικά δίκτυα προκειμένου από την αναπαράσταση μιας τρικαναλικής εικόνας να προβλέψουν τις παραμέτρους του μοντέλου SMPL. Όσον αφορά στο είδος της εισόδου έχουν προταθεί διάφορες επιλογές όπως:

- Θέσεις των αρθρώσεων πάνω στο επίπεδο εικόνας.
- Σιλουέτες στο επίπεδο εικονας.

- Μάσκες Κατάτμησης του Σώματος.
- Ακατέργαστη εικόνα.

Το μοντέλο HMR [Kan+18] είναι από τις πρώτες ερευνητικές εργασίες που χρησιμοποιούν κατευθείαν την εικόνα ως είσοδο προκειμένου να εκτιμήσουν τις παραμέτρους. Η αρχιτεκτονική του φαίνεται στην εικόνα 3.2.1. Πιο συγκεκριμένα αποτελείται από ένα συνελικτικό δίκτυο(ResNet50) το οποίο έχει προεκπαιδευτεί στο σύνολο δεδομένων ImageNet. Οι ενδιάμεσες αναπαραστάσεις από το τελευταίο επίπεδο του ResNet[He+15] δίνονται σαν είσοδοι στην κεφαλή του μοντέλου που εκτιμά τις παραμέτρους σχήματος $\boldsymbol{\beta} \in R^{10}$, πόζας $\boldsymbol{\vartheta} \in R^{72}$ και κάμερας $\mathbf{c} \in R^3$. Οι συγγραφείς υποστηρίζουν πως η εκτίμηση με ένα πέρασμα του νευρωνικού δικτύου είναι δύσκολο πρόβλημα επομένως προχωρούν στις εξής βελτιώσεις:

- Σαν επιπλέον είσοδοι στην κεφαλή του μοντέλου δίνονται οι μέσες τιμές για το σχήμα, την πόζα και την κάμερα και το μοντέλο εκτιμά τις υπολειπόμενες τιμές(residuals) οι οποίες προστίθενται στις μέσες τιμές και παράγουν την τελική εκτίμηση.
- Η διαδικασία εκτίμησης επαναλαμβάνεται πολλαπλές φορές λαμβάνοντας σαν είσοδο κάθε φορά την προηγούμενη εκτίμηση σχήματος,πόζας και κάμερας.

Όσον αφορά στην αναπαράσταση των περιστροφών, στην βιβλιογραφία εκτός από την μορφή axis-angle χρησιμοποιείται και η 6Δ μορφή που προτείνουν οι ερευνητές Zhou et al [Zho+19]. Στην συνέχεια οι παράμετροι σχήματος και πόζας δίνονται σαν είσοδοι στο SMPL μοντέλο το οποίο παράγει το τελικό πλέγμα. Το μοντέλο εκπαιδεύεται με έναν συνδυασμό 2Δ και 3Δ επισημειώσεων. Στα δείγματα για τα οποία παρέχεται η τρισδιάστατη επισημείωση $X^{24\times3}$, με την μορφή 3Δ συντεταγμένων για κάθε μία από τις αρθρώσεις του μοντέλου, οι συγγραφείς ελαχιστοποιούν την L2 νόρμα μεταξύ των επισημειώσεων και των αρθρώσεων του εκτιμώμενου πλέγματος

$$L_{joints} = ||X_{gt} - X_{pred}||_2$$

Στα δείγματα όπου παρέχεται η δισδιάστατη επισημείωση $x^{24\times2}$, οι εκτιμώμενες αρθρώσεις με την χρήση της κάμερας, της οποίας η παράμετροι επίσης έχουν εκτιμηθεί, προβάλλονται στον κανονικοποιημένο χώρο εικόνας και στην συνάρτηση λάθους προστίθεται ένας όρος όπου ελαχιστοποιεί την L2 νόρμα μεταξύ αυτών και των επισημειώσεων(reprojection loss)

$$L_{proj} = ||x_{gt} - x_{proj}||_2$$

Το μοντέλο κάμερας που επιλέγουν είναι η weak perspective κάμερα επομένως αρκεί να εκτιμηθεί η παράμετρος κλίμακας $c_s \in R^+$ και μετατόπισης $c_t \in R^2$. Όπου είναι διαθέσιμες οι επισημειώσεις με την μορφή SMPL παραμέτρων σχήματος και πόζας $\boldsymbol{\Theta} \in R^{82}$ εφαρμόζεται η L2 νόρμα

$$L_{SMPL} = ||\Theta_{gt} - \Theta_{pred}||_2$$

Επιπλέον οι συγγραφείς παρατηρούν ότι στα δείγματα όπου λείπει η τρισδιάστατη επισημείωση, η 2Δ επίβλεψη δεν αρκεί για να παράγει το μοντέλο σωστές πόζες διότι μπορεί να μειώσει το λαθος επαναπροβολής παράγοντας μη αληθοφανείς πόζες που ωστόσο εξηγούν τις επισημειώσεις. Με αυτό το κίνητρο εφαρμόζουν Adversarial εκπαίδευση όπου ενας κριτής $D(x) \in [0, 1]$ εκπαιδευεται να διαχωρίζει μεταξύ αληθινών δεδομένων και δεδομένων που παράγει το μοντέλο. Κατα την εκπαίδευση του κριτή η συνάρτηση λάθους απαιτεί ο κριτής να δίνει χαμηλό σκορ στα εκτιμώμενα δεδομένα και υψηλο σκορ στα αληθινά δεδομένα. Η συνάρτηση λάθους δίνεται από την σχέση:

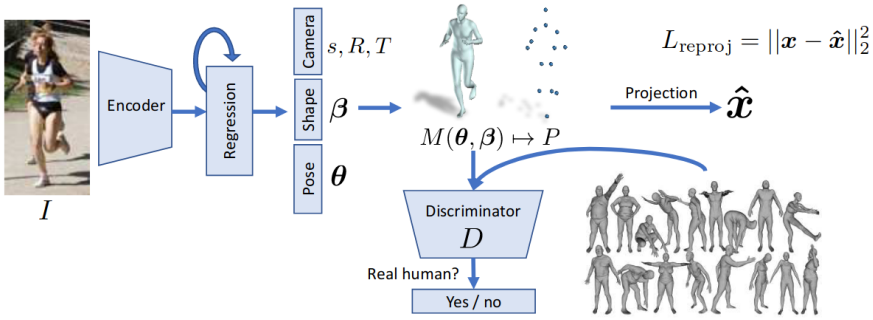
$$L_d = E_{x\sim P_{real}}[1 - D(x)] + E_{x\sim P_{model}}[D(x)]$$

Αντιθέτως κατα την φάση εκπαίδευσης του HMR στην συνάρτηση λάθους προστίθεται ένας όρος που οδηγεί το μοντέλο να παράγει δεδομένα που ο κριτής δίνει υψηλό σκόρ $L_g = E_{x\sim P_{model}}[1 - D(x)]$. Η συνολική συνάρτηση λάθους είναι

$$L = L_{joints} + L_{proj} + L_{SMPL} + L_g \tag{3.2.1}$$

Η συνολική αρχιτεκτονική και το πλαίσιο εκπαίδευσης φαίνονται στην Εικόνα 3.2.1

Παρόμοιες ερευνητικές εργασίες αναπτύχθηκαν ανεξάρτητα περίπου την ίδια περίοδο [Las+17; Pav+18] χρησιμοποιώντας παρόμοιες τεχνικές. Οι ερευνητές [Las+17] εξετάζουν την εκτίμηση χρησιμοποιώντας τεχνητές 2Δ επισημειώσεις αρθρώσεων από 3Δ δεδομένα Πόζας και Σχήματος ενώ οι ερευνητές [Pav+18] χρησιμοποιούν



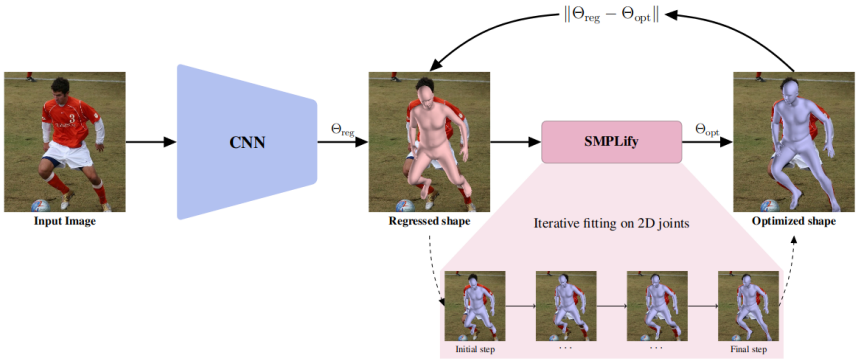
Εικόνα 3.2.1: Η αρχιτεκτονική του μοντέλου HMR. Εικόνα από [Kan+18]

προεκπαιδευμένα δίκτυα ώστε να εξάγουν από εικόνες χαρακτηριστικά σιλουέτας και πόζας. Η επιτυχία της επαναληπτικής μεθόδου του HMR την καθιστά σημείο αναφοράς και μέτρο σύγκρισης στην βιβλιογραφία. Συχνά αποτελεί το βασικό μοντέλο πάνω στο οποίο οι ερευνητές βασίζονται ώστε να παρουσιάσουν πιο αποδοτικούς τρόπους επίβλεψης [Li+22; PKD19; Tri+23]

Σε αντίθεση με τις μεθόδους που κάνουν χρήση επαναληπτικών αλγορίθμων, οι μέθοδοι που εκτιμούν μέσω δικτύων τις παραμέτρους είναι αρκετά ταχύτερες ωστόσο δεν επιτυγχάνουν τα ίδια επίπεδα ακρίβειας. Προκειμένου να εκμεταλλευτούν τα πλεονεκτήματα των δύο μεθόδων οι ερευνητές [Kol+19] προτείνουν ένα συνεργατικό πλαίσιο μεταξύ των 2 μεθόδων. Πιο συγκεκριμένα παρατηρούν ότι η επαναληπτική μέθοδος αργεί να συγκλίνει οταν το σημείο αρχικοποίησης είναι αρκετά διαφορετικό από την τελική πόζα. Η μέθοδος που προτείνουν αποτελείται από 2 στάδια:

- Στάδιο 1: Το μοντέλο HMR με είσοδο την είκονα παράγει την πρώτη εκτίμηση 3Δ πόζας και σχήματος.
- Στάδιο 2: Χρησιμοποιώντας την εκτίμηση του μοντέλου HMR ως σημείο αρχικοποίησης του επαναληπτικού αλγορίθμου εκτελείται ο αλγόριθμος SMPLify ώστε να παράξει μια πιο λεπτομερή εκτίμηση.

Επιπλέον το αποτέλεσμα του σταδίου 2 χρησιμοποιείται στην επίβλεψη του μοντέλου HMR, επομένως και η μέθοδος SMPLify επωφελείται από την αρχική εκτίμηση του μοντέλου HMR αλλα και το μοντέλο HMR επωφελείται από την επίβλεψη μέσω των αποτελεσμάτων της μεθόδου SMPLify. Μια οπτική αναπαράσταση της διαδικασίας φαίνεται στην Εικόνα 3.2.2



Εικόνα 3.2.2: Οπτική αναπαράσταση της διαδικασίας SPIN. Εικόνα από [Kol+19]

Οι επόμενες ερευνητικές εργασίες επικεντρώθηκαν σε συγκεκριμένες πτυχές του προβλήματος όπως:

- Ξεχωριστή εκτίμηση για κάθε μέρος του σώματος.
- Πολλαπλές εκτιμήσεις για μια είσοδο.
- Χρήση καλύτερου μοντέλου κάμερας.
- Πιθανοτική αντιμετώπιση του προβλήματος.

Πιο αναλυτικά οι ερευνητές [Koc+21a] εισάγουν το μοντέλο PARE το οποίο εκτιμά ξεχωριστά τις περιστροφές για κάθε μέρος του σώματος. Σε αντίθεση με την πλειοψηφία των προηγούμενων μοντέλων διατηρεί την χωρική πληροφορία κρατώντας την αναπαράσταση του ResNet50 πριν γινει η συγχώνευση χαρακτηριστικών. Στην συνέχεια χρησιμοποιεί 2 υποδίκτυα προκειμένου να εξάγει μάσκες κατάτμησης και με βάση αυτές να εξάγει ένα χαρακτηριστικό για κάθε μέρος του σώματος. Στην συνέχεια τα χαρακτηριστικά δίνονται σαν είσοδοι σε N ξεχωριστούς εκτιμητές προκειμένου να παράξουν την περιστροφή του συγκεκριμένου σημείου. Πειραματικά δείχνεται πως αυτές οι αρχιτεκτονικές επιλογές κάνουν το μοντέλο πιο ικανό στο σενάριο όπου ο άνθρωπος ενδιαφέροντος εμποδίζεται ή φαίνεται μερικώς στην εικόνα.

Οι ερευνητές [Koc+21b] υποστηρίζουν πως η υπόθεση οτι ο άνθρωπος είναι αρκετά μακριά από την κάμερα ώστε να γίνει χρήση της κάμερας ασθενούς προοπτικής δεν ισχύει πάντα και προτείνουν την χρήση pinhole κάμερας. Επιπλέον, στις μέχρι τώρα προαναφερθείσες εργασίες γίνεται η υπόθεση ότι ισχύουν τα εξής:

- Το υποκείμενο βρίσκεται στην αρχή των αξόνων.
- Η κάμερα έχει σταθερό προσανατολισμό.
- Εκτιμάται η περιστροφή του υποκειμένου ως προς την κάμερα.
- Εκτιμάται η τοποθεσία της κάμερας σε σχέση με την αρχή των αξόνων.

Αντιθέτως στην συγκεκριμένη εργασία οι ερευνητές επιλέγουν

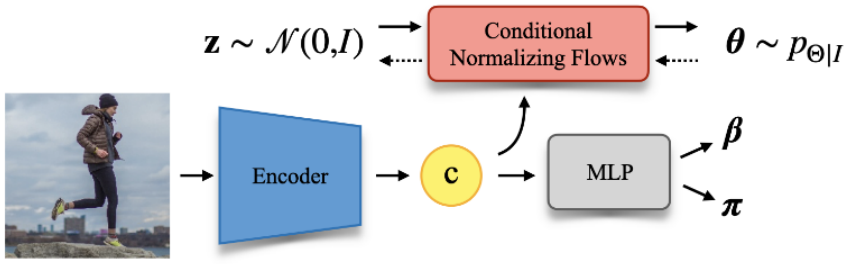
- Να εκτιμάται η περιστροφή της κάμερας ως προς το καθολικό σύστημα αναφοράς.
- Να εκτιμάται η περιστροφή του ανθρώπου ως προς το καθολικό σύστημα αναφοράς.
- Η κάμερα είναι σταθερά τοποθετημένη στην αρχή των αξόνων.
- Να εκτιμάται η τοποθεσία του υποκειμένου ως προς το καθολικό σύστημα αναφοράς.

Ετσι, το υποδίκτυο που είναι υπεύθυνο για την εκτίμηση πόζας και σχήματος εκτιμά επιπλέον την μετατόπιση του ανθρώπου ενω υπάρχει ξεχωριστό υποδίκτυο που εκτιμά την περιστροφή της κάμερας. Πειραματικά αποδεικνύεται οτι το μοντέλο SPEC πετυχαίνει καλύτερη απόδοση σε περιπτώσεις όπου η κάμερα είναι αρκετα κοντά στο αντικείμενο.

Σε περιπτώσεις όπου ο άνθρωπος είτε κρύβεται από κάποιο εμπόδιο είτε είναι μερικώς εμφανής στην εικόνα ιδανικά το μοντέλο θα πρέπει να παράγει πολλαπλές εκτιμήσεις αναπαριστώντας την αβεβαιότητα που υπάρχει στο κρυμμένο μέρος της εικόνας. Οι ερευνητές [Big+20] αντιμετωπίζουν αυτο το πρόβλημα τροποποιώντας το μοντέλο HMR ωστε να παράγει ένα σύνολο από M εκτιμήσεις. Η εκπαίδευση του μοντέλου πραγματοποιείται με το *Καλύτερο-Απο-Μ* κριτήριο σύμφωνα με το οποίο εφαρμόζεται το Μέσο Τετραγωνικό Λάθος μόνο στην καλύτερη από τις M εκτιμήσεις και δίνεται από τον τύπο:

$$L_{best} = \frac{1}{N}\sum_{i=1}^{N} ||X_i - \hat{X}^{m_i^*}||, m_i^* = \text{argmin}_{m=1,\dots,M}||X_i - X^{m_i}||$$

όπου N ο αριθμός των δειγμάτων, X_i η επισημείωση 3Δ αρθρώσεων και X_i^m η m-οστή πρόβλεψη από το σύνολο προβλέψεων για το i-οστό δείγμα.



Εικόνα 3.2.3: Η αρχιτεκτονική του μοντέλου ProHMR. Εικόνα από [Kol+21]

Αν και η παραπάνω μέθοδος παράγει πολλαπλές εκτιμήσεις το γεγονός πως ο αριθμός των εκτιμήσεων εκφράζεται ρητά στη αρχιτεκτονική του μοντέλου συνεπάγεται οτι για διαφορετικό αριθμό εκτιμήσεων πρέπει να εκπαιδευτεί νέο μοντέλο. Αντιθέτως οι ερευνητές [Kol+21] αντιμετωπίζουν το πρόβλημα πιθανοτικά χρησιμοποιώντας γενετικά μοντέλα Normalizing flows προκειμένου να παράγουν δεδομένης μιας εικόνα την υπο συνθήκη κατανομή στον χώρο παραμέτρων SMPL. Το μοντέλο ProHMR εκτός ότι παράγει μεταβλητό αριθμό εκτιμήσεων κατέχει και το γενικότερο πλεονέκτημα των πιθανοτικών μεθόδων που είναι η μοντελοποίηση της αβεβαιότητας μιας εκτίμησης . Η αρχιτεκτονική του μοντέλου φαίνεται στην Εικόνα 3.2.3. Αποτελείται από ένα συνελικτικό δίκτυο ResNet50 προεκπαιδευμένο στο ImageNet. Οι ενδιάμεσες αναπαραστάσεις από το τελευταίο επίπεδο του δικτύου δίνονται σαν είσοδοι σε ενα Glow [KD18] normalizing flow μοντέλο προκειμένου να παράξει μια κατανομή δεδομένης της εικόνας. Παράλληλα δίνονται και σε ενα Feed Forward Net για την εκτίμηση των παραμέτρων σχήματος και κάμερας. Η συνάρτηση f στο μοντέλο Glow αποτελείται από την σύνθεση των συναρτήσεων:

- Conditional Additive Coupling Layer όπου η συνάρτηση $y = f(x, c)$ με $x, y \in R^D$ δίνεται από τις σχέσεις:

$$\begin{aligned} y_{1:d} &= x_{1:d} \\ y_{d+1:D} &= x_{d+1:D} + t(x_{1:d}, c) \end{aligned} \tag{3.2.2}$$

  όπου $\mathbf{c} \in R^K$ είναι η conditioning είσοδος.

- Γενικευμένα permutations της μορφής $y = Wx$ όπου $W \in R^{D \times D}$ αντιστρέψιμος πίνακας.

- Activation Normalization Layer όπου η συνάρτηση $y = f(x)$ είναι της μορφής $y = a \cdot x + b$ με $a, b \in R^D$

Δεδομένου ότι και για τους τρεις τύπους συναρτήσεων η ορίζουσα της ιακωβιανής είναι ανεξάρτητη του x, η κατανομή της εξίσωσης 2.2.1 έχει μεγιστο στον μετασχηματισμό του σημείου όπου έχει μέγιστο η βασική κατανομή, την οποία η συγγραφείς έχουν επιλέξει να μοντελοποιήσουν με την κανονική κατανομή. Επομένως το σημείο $\tilde{x} = f^{-1}(\mathbf{0}, \mathbf{c})$ είναι η πιο πιθανή πόζα.

Η συνάρτηση λάθος του μοντέλου περιλαμβάνει εκτός από τον όρο Negative Log Likelihood(NLL) $L_{nll} = -log(P(x))$ τους όρους της εξίσωσης 3.2.1 χρησιμοποιώντας την πιο πιθανή ποζα $\tilde{x}$ ως την σημειακή εκτίμηση του μοντέλου. Επιπλέον περιλαμβάνει τις αναμενόμενες τιμές των όρων L_{proj} και L_g της εξίσωσης 3.2.1 χρησιμοποιώντας την monte carlo εκτίμηση τους:

$$E_{x \sim p_{model}}[L_{proj} + L_g] = \frac{1}{n} \sum_{x_i} (L_{proj} + L_g)$$

όπου x_i δείγματα από την κατανομή του μοντέλου.

Δεδομένου ότι, οπως έχει ήδη παρουσιαστεί, τα μοντέλα Normalizing Flows παρέχουν την δυνατότητα τόσο του υπολογισμού πιθανοφάνειας όσο και της παραγωγής δειγμάτων οι συγγραφείς προτείνουν δύο τρόπους χρήσης του μοντέλου ProHMR:

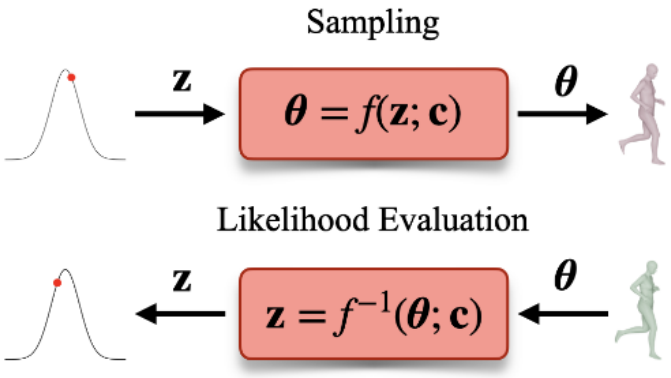
- για εκτίμηση πόζας, όπου το μοντέλο επιστρέφει την πιο πιθανή πόζα $\tilde{x} = f^{-1}(\mathbf{0}, \mathbf{c})$
- για υπολογισμό της πιθανοφάνειας μιας πόζας με χρήση της εξίσωσης 2.2.1.

Μια οπτική αναπαράσταση των παραπάνω διαδικασιών φαίνεται στην Εικόνα 3.2.4

Πρόσφατα έχουν προταθεί ιδέες που ενσωματώνουν την διαδικασία της αντίστροφης κινηματική ανάλυσης. Πιο συγκεκριμένα μια πληθώρα από ερευνητικές εργασιές [Li+21; She+23; Li+23] ισχυρίζονται ότι το πρόβλημα εκτίμησης περιστροφών είναι πιο δύσκολο από την εκτίμηση των θέσεων των αρθρώσεων οπως συμβαίνει στα προβλήματα εκτίμησης τρισδιάστατης πόζας. Για αυτο το λόγο χρησιμοποιούν δίκτυα τα οποια μαθαίνουν την διαδικασία της αντίστροφης κινηματικής ανάλυσης ώστε να μετατρέπουν τις εκτιμήσεις ενός τρισδιάστατου εκτιμητή πόζας σε περιστροφές αρθρώσεων ώστε να είναι συμβατές με το μοντέλο SMPL.

3.3 Μέθοδοι εκτίμησης από βίντεο

Αν και η εκτίμηση από στατικές εικόνες είναι ένα σημαντικό πρόβλημα, η πλειοψηφία των εφαρμογών απαιτεί την εκτίμηση απο μια ακολουθία απο εικόνες. Μια απλή λύση είναι η χρήση στατικών μεθόδων, οπως αυτές που



Εικόνα 3.2.4: Οι δύο τρόποι χρήσης του μοντέλου ProHMR. Εικόνα από [Kol+21]

αναλύθηκαν προηγουμένως, ώστε να γίνεται εκτίμηση της πόζας και σχήματος ξεχωριστά για κάθε Εικόνα της ακολουθίας. Ωστόσο η παραπάνω διαδικασία δεν εκμεταλλεύεται της αλληλοεξαρτήσεις που υπάρχουν μεταξύ των εικόνων ενός βίντεο.

Πιο αναλυτικά, δεδομένου ότι οι κινήσεις ενός ανθρώπου μεταβάλλονται ομαλά μεταξύ δυο κοντινών χρονικών στιγμών η πληροφορία από ένα frame την χρονική στιγμή t περιορίζει τις πιθανές πόζες για την χρονική στιγμή t+1. Επομένως υπάρχει περισσότερη πληροφορία που μπορούν να εκμεταλλευτούν τα μοντέλα ώστε να επιτύχουν καλύτερα αποτελέσματα. Οι μέθοδοι αυτης της παραγράφου μπορούν να χωριστούν σε δύο κατηγορίες ανάλογα με το πότε εισάγονται οι χρονικοί περιορισμοί στην διαδικασία της εκτίμησης. Πιο συγκεκριμένα χωρίζονται σε:

- Μέθοδοι επεξεργασίας εκ των υστέρων(Post processing) οπου αρχικά χρησιμοποιούν ένα στατικό μοντέλο ώστε να παράξουν μια αρχική εκτίμηση για κάθε frame και στην συνέχεια εφαρμόζουν αλγορίθμους όπου τροποποιούν τις εκτιμήσεις ώστε να σέβονται τους παραπάνω περιορισμούς. Ένα τέτοιο παράδειγμα είναι η ερευνητική εργασία των [ADZ19] οπου χρησιμοποιούν το μοντέλο HMR [Kan+18] για να παράξουν αρχικές εκτιμήσεις και στην συνέχεια εφαρμόζουν τον αλγόριθμο bundle adjustment ώστε να παράξουν συνεπείς εκτιμήσεις.

- End-to-End μέθοδοι όπου παράγουν τις εκτιμήσεις με μια ενιαία διαδικασία. Συνήθως αυτές οι μέθοδοι δημιουργούν χρονικά εμπλουτισμένες κρυφές αναπαραστάσεις και στην συνέχεια τις δίνουν σαν είσοδους σε ένα μοντέλο που εκτιμά τις παραμέτρους SMPL χωρίς να ακολουθεί κάποιο επιπλέον βήμα επεξεργασίας.

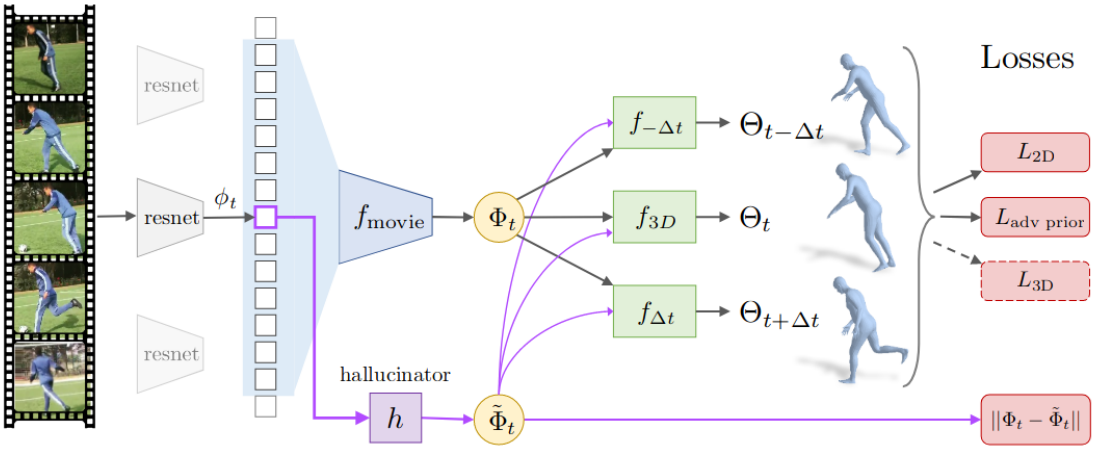
Δεδομένου ότι θα ασχοληθούμε με το δεύτερο είδος σε αυτή την εργασία, θα παρουσιάσουμε ερευνητικές εργασίες αυτής της κατηγορίας.

Η πλειοψηφία των ερευνητικών εργασιών ακολουθεί την ίδια αρχική δομή. Αρχικά ένα προεκπαιδευμένο συνελικτικό δίκτυο ResNet50 χρησιμοποιείται ώστε να εξαχθούν τα χαρακτηριστικά Εικόνας για κάθε frame. Στην συνέχεια οι ερευνητικές εργασίες διαφοροποιούνται ως προς τον τρόπο που συνδυάζεται η πληροφορία από διαφορετικές χρονικές στιγμές.

Οι ερευνητές [Kan+19] χρησιμοποιούν ενα συνελικτικό δίκτυο f_{movie} ώστε να επεξεργαστούν τα χαρακτηριστικά Εικόνας και να παράξουν κρυφές αναπαραστάσεις εμπλουτισμένες με την χρονική πληροφορία. Στην συνέχεια χρησιμοποιούν το μοντέλο HMR ώστε να παράξουν τις παραμέτρους SMPL για κάθε χρονική στιγμή. Επιπλέον, προκειμένου να αναγκάσουν το μοντέλο να μοντελοποιήσει την χρονική πληροφορία χρησιμοποιούν δύο επιπλέον μοντέλα HMR, $f_{\Delta T}$, $f_{-\Delta T}$ τα οποία λαμβάνουν σαν είσοδους την έξοδο του f_{movie} και τις εκτιμήσεις παραμέτρων της χρονικής στιγμή t και εκτιμούν την μεταβολή που πρέπει να προστεθεί στην τρέχουσα πόζα ώστε να παραχθεί η πόζα της επόμενης και προηγούμενης χρονικής στιγμής.

Επιπλέον προκειμένου να είναι δυνατή η χρήση του μοντέλου σε στατικές εικόνες, εκπαιδεύουν επιπλέον ένα μοντέλο "hallucinator" που μαθαίνει να εκτιμά την έξοδο του συνελικτικού δικτύου f_{movie}. Ένα υποπροιόν αυτής της διαδικασίας είναι η ικανότητα του μοντέλου να εκτιμά τις μελλοντικές πόζες απο μια στατική Εικόνα, δίνοντας την εξοδο του "hallucinator" στο μοντέλο $f_{\Delta T}$.

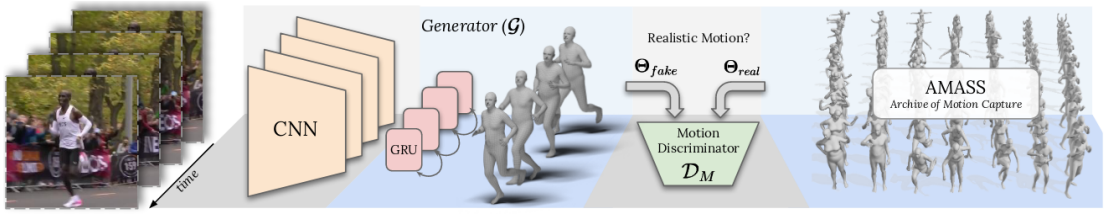
Το μοντέλο εκπαιδεύεται με την ίδια αντικειμενική συνάρτηση που αναλύθηκε για το μοντέλο HMR στην προηγούμενη ενότητα. Η αρχιτεκτονική του μοντέλου φαίνεται στην Εικόνα 3.3.1



Εικόνα 3.3.1: Συνολική αρχτιτεκτονική του μοντέλου HMMR. Εικόνα απο [Kan+19]

Οι ερευνητές [KAB20] επεκτείνουν την μέθοδο [Kan+18] στην χρονική διάσταση. Η ειδοποιός διαφορά της εργασίας τους με την προηγούμενη είναι η αιτιατή αντιμετώπιση αφού για κάθε χρονική στιγμή t το μοντέλο γνωρίζει την πληροφορία μόνο μέχρι την χρονική στιγμή t. Πιο αναλυτικά χρησιμοποιεί ενα αναδρομικό δίκτυο GRU ωστε να παράξει τις χρονικά εμπλουτισμένες κρυφές αναπαραστάσεις. Στην συνέχεια τις δίνει ως εισόδους στο μοντέλο HMR ωστε να εκτιμήσεις τις παραμέτρους SMPL.

Ομοίως κατα την εκπαίδευση χρησιμοποιεί εναν Κριτή, όπου διακρίνει ανάμεσα σε πραγματικές ακολουθίας απο μια βαση δεδομένων και εκτιμώμενες από το μοντέλο ακολουθίες πόζας, και εφαρμόζει Adversarial εκπαίδευση. Η συνολική αρχιτεκτονική και το πλαίσιο εκπαίδευσης φαίνεται στην Εικόνα 3.3.2.



Εικόνα 3.3.2: Η συνολική αρχιτεκτονική και το πλαίσιο εκπαίδευσης του μοντέλου VIBE. Εικόνα απο [KAB20]

Οι ερευνητές [Cho+21] επεκτείνουν την μέθοδο ώστε να χρησιμοποιεί πληροφορία απο τα μελλοντικά frames για την εκτίμηση πιο χρονικά συνεκτικών αποτελεσμάτων ενω οι ερευνητές [Vas+23; Du+23] εφαρμόζουν παρόμοιες τεχνικές στον κρυφό χώρο του μοντέλου ProHMR.

Πρόσφατες εργασίες εκτός από την ανακατασκευή του σχήματος εστιάζουν στην χρήση της πόζας για παρακολουθηση [Raj+22] και εκτίμηση καθολικής τροχιάς [Yua+22; Ye+23].

3.4 Σύνολα Δεδομένων

Τα παραπάνω μοντέλα εκπαιδεύονται πάνω σε ένα πλήθος από σύνολα δεδομένων που περιέχουν είτε 2Δ είτε 3Δ επισημειώσεις. Τα πιο συχνά χρησιμοποιούμενα σύνολα με 2Δ επισημειώσεις είναι τα :

- COCO2014[Lin+14]: αποτελείται από 83 χιλιάδες εικόνες εκπαίδευσης και 41 χιλιάδες εικόνες επαλήθευσης(validation set).

- MPII[And+14]: αποτελείται από 25 χιλιάδες εικόνες εκπαίδευσης.

Όλες οι εικόνες είναι επισημειωμένες με τις θέσεις των φανερών αρθρώσεων στον χώρο της εικόνας(pixel space). Όσον αφορά στα 3Δ σύνολα, χρησιμοποιούνται ευρέως τα:

- Human3.6m [Ion+14]: Αποτελείται από 1.5 εκατομμύρια εικόνες όπου τα 5/7 είναι το σύνολο εκπαίδευσης και οι υπόλοιπες χρησιμοποιούνται στην βιβλιογραφία ως το σύνολο αξιολόγησης(test set).
- MPI-INF-3DHP [Meh+17]: Αποτελείται από 900 χιλιάδες εικόνες εκπαίδευσης.
- 3DPW [Mar+18]: Αποτελείται από περίπου 30 χιλιάδες εικόνες εκπαίδευσης και 30 χιλιάδες εικόνες επαλήθευσης. από την πλειοψηφία της βιβλιογραφίας χρησιμοποιείται μόνο για αξιολόγηση.

Όλες οι παραπάνω εικόνες συνοδεύονται από από 2Δ και 3Δ επισημειώσεις. Επιπλέον στα τρία 3Δ σύνολα οι εικόνες αποτελούν τα frames από βίντεο επομένως χρησιμοποιούνται και από τις μεθόδους της παραγράφου 2.3.

Τέλος για την περίπτωση όπου τα μοντέλα χρειάζονται δεδομένα πόζας και σχήματος χωρίς να ενδιαφέρονται για την αντίστοιχη εικόνα χρησιμοποιείται το σύνολο δεδομένων AMASS[Mah+19].

Κεφάλαιο 4

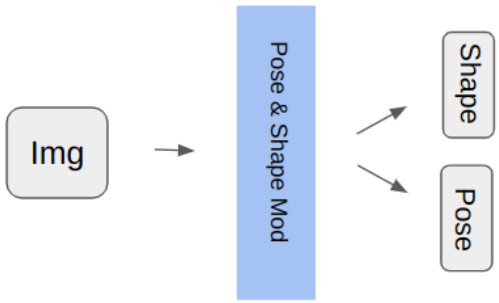
Προτεινόμενη Μεθοδολογία

Αφου έχουμε παρουσιάσει την σχετική βιβλιογραφία, στην συνέχεια προτείνουμε τρόπους επέκτασης των παραπάνω πετυχημένων μοντέλων. Αρχικά εστιάζουμε στο στατικό πρόβλημα και παρουσιάζουμε ένα μοντέλο-επέκταση του μοντέλου HMR που προσπαθεί να αποσυμπλέξει τα δύο προβλήματα. Στην συνέχεια ασχολούμαστε με την δυναμική εκδοχή του προβλήματος. Προτείνουμε 2 τρόπους επέκτασης της αρχιτεκτονικής του μοντέλου VIBE ενώ παράλληλα προτείνουμε και την χρήση ενός βοηθητικού προβλήματος ώστε να επωφεληθούμε από την επιπλέον επίβλεψη.

4.1 Μοντέλα αποσύμπλεξης Πόζας-Σχήματος

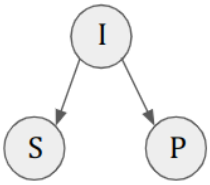
4.1.1 HMR2

Οπως παρουσιάστηκε στο κεφάλαιο 2, οι παράμετροι σχήματος τροποποιούν το αρχικό πλέγμα το οποίο στην συνέχεια επηρεάζει τις θέσεις των αρθρώσεων και τις αποστάσεις μεταξύς τους. Επομένως κατα την διαδικασία εκτίμησης πόζας η πληροφορία αυτή θα πρέπει να είναι διαθέσιμη στο μοντέλο. Ωστόσο παρατηρούμε ότι οι προηγούμενες ερευνητικές εργασίες δεν μοντελοποιούν ρητά αυτή την εξάρτηση αφού εκτιμούν παράλληλα τις παραμέτρους πόζας και σχήματος. Η απλοποιημένη αρχιτεκτονική αυτής της μορφή φαίνεται στην Εικόνα 4.1.1. Με αφορμή αυτή την παρατήρηση τροποποιούμε την αρχιτεκτονική του μοντέλου HMR ώστε να αντικατοπτρίζει την εξάρτηση αυτή.

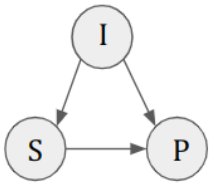


Εικόνα 4.1.1: Ο τρόπος διασύνδεσης των υποσυστημάτων στο μοντέλο HMR.

Πιο συγκεκριμένα τροποποιούμε την αρχιτεκτονική του μοντέλου HMR ωστε να γίνεται πρώτα η εκτίμηση σχήματος με τη χρήση ενός υποσυστήματος και στην συνέχεια, δεδομένων των παραμέτρων σχήματος, η εκτίμηση πόζας από ένα ξεχωριστό υποσύστημα. Ο γράφος εξάρτησης μεταξύ των παραμέτρων του μοντέλου SMPL και της εισόδου φαίνεται στην Εικόνα 4.1.2 και η αρχιτεκτονική του μοντέλου φαίνεται στην Εικόνα 4.1.3.



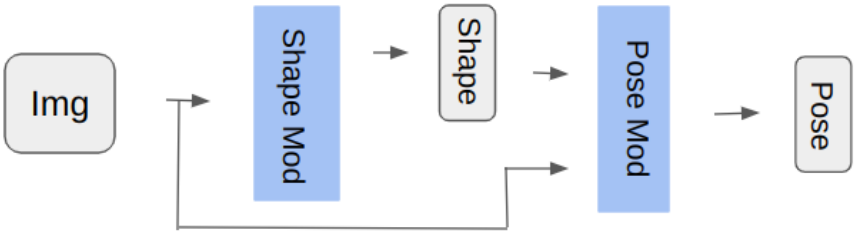
(a) HMR.



(b) HMR2.

Εικόνα 4.1.2: Οι γράφοι εξάρτησης των μοντέλων HMR και HMR2 οπου I,S,P δηλώνουν τις παραμέτρους Εικόνας, Σχήματος και Πόζας. Σε αντίθεση με το μοντέλο HMR, το μοντέλο HMR2 μοντελοποιεί ρητά το γεγονός πως χρειαζόμαστε τα δεδομένα σχήματος προκειμένου να εκτιμήσουμε σωστά την πόζα.

Κατα την πειραματική διαδικασία εκπαιδεύουμε το μοντέλο κάτω από διαφορετικές συνθήκες ώστε καταλάβουμε την επίδραση που έχουν οι αποδόσεις των υποσυστημάτων στην απόδοση του μοντέλου.

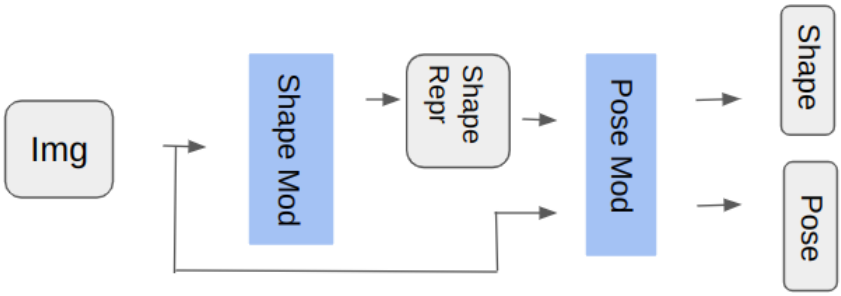


Εικόνα 4.1.3: Ο τρόπος διασύνδεσης των υποσυστημάτων στο μοντέλο HMR2.

4.1.2 HMR3

Προκειμένου να επεκτείνουμε την εφαρμογή του παραπάνω μοντέλου σε σύνολα δεδομένων που δεν διαθέτουν επισημειώσεις SMPL πρέπει να αντικατασταθεί η έξοδος του υποσυστήματος σχήματος από παραμέτρους σχήματος SMPL σε μια πιο γενική αναπαράσταση σχήματος. Επομένως προχωρούμε σε μια επιπλέον τροποποίηση του μοντέλου HMR2.

Πιο συγκεκριμένα, το υποσύστημα σχήματος πλέον παράγει μια γενικευμένη αναπαράσταση σχήματος η οποία δίνεται σαν είσοδος στο υποσύστημα πόζας το οποίο αφενός παράγει την τελική εκτίμηση πόζας, αφετέρου μετασχηματιζει την αναπαράσταση σχήματος στην μορφή των παραμέτρων SMPL έτσι ώστε να δοθούν στο παραμετρικό μοντέλο. Η αρχιτεκτονική του μοντέλου HMR3 φαίνεται στην Εικόνα 4.1.4



Εικόνα 4.1.4: Ο τρόπος διασύνδεσης των υποσυστημάτων στο μοντέλο HMR3.

Η γενικευμένη αναπαράσταση θα πρέπει να έχει τα ακόλουθα χαρακτηριστικά:

- Ανεξάρτητη ως προς το μοντέλο SMPL: Δεν θα πρέπει να εξαρτάται από το μοντέλο SMPL έτσι ώστε να μην χρειάζονται επισημειώσεις SMPL για την εξαγωγή της.
- Αμετάβλητη ως προς την πόζα: Οι αναπαραστάσεις σχήματος εξαγμένες από το ίδιο υποκείμενο σε διαφορετικές πόζες θα πρέπει να ταυτίζονται.

Επομένως επιλέγουμε σαν αναπαράσταση το διάνυσμα διάστασης $n-1$ οπου n είναι ο αριθμός των αρθρώσεων και το στοιχείο στην i-οστή θέση περιέχει την απόσταση της άρθρωσης i ως προς τον γονέα της στο κινηματικό δέντρο. Παρατηρούμε πως αυτή η αναπαράσταση μπορεί να εξαχθεί είτε από τρισδιάστατες επισημειώσεις είτε από επισημειώσεις SMPL αφού πρώτα βρεθούν οι θέσεις των αρθρώσεων και οτι δεν εξαρτώνται από την πόζα του υποκειμένου. Επομένως διαθέτει τα παραπάνω χαρακτηριστικά.

Στο επόμενο κεφάλαιο πραγματοποιούμε μια πληθώρα από πειράματα εκπαίδευσης κάτω από διαφορετικές συνθήκες ώστε να κατανοήσουμε την επίδραση κάθε υποσυστήματος στην επίδοση του μοντέλου.

4.2 Επέκταση της μεθόδου VIBE

Οπως παρουσιάστηκε στο κεφάλαιο 3 οι μέθοδοι τρισδιάστατης ανακατασκευής πόζας και σχήματος από ακολουθία εικόνων(βίντεο) εκμεταλλεύονται τους περιορισμούς που προκύπτουν από το γεγονός οτι η πόζα ενός αν-

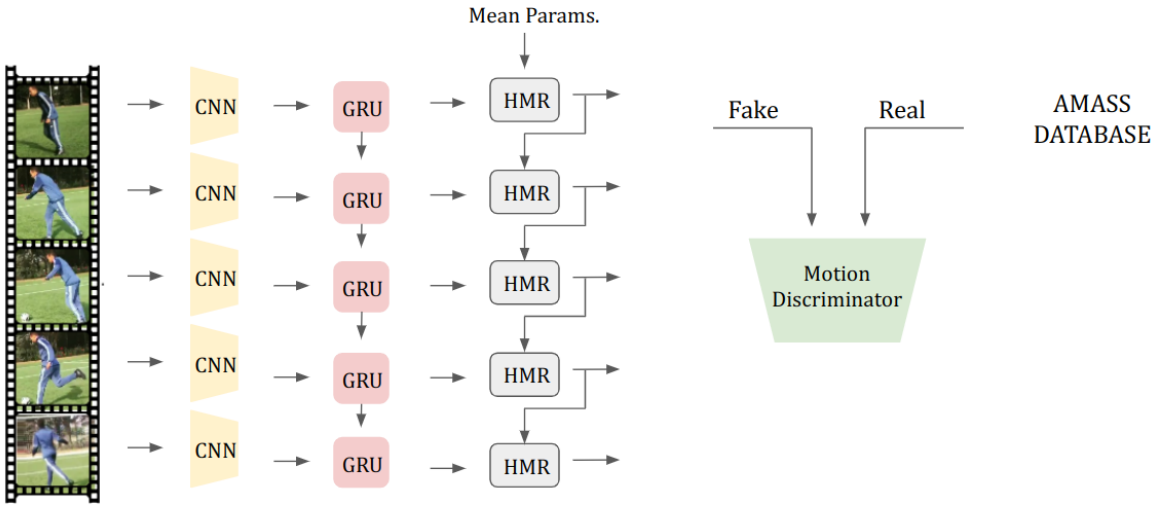
θρώπου μεταβάλλεται ομαλά μεταξύ των χρονικών στιγμών. Σε αυτή την κατηγορία πειραμάτων βασιζόμαστε στο μοντέλο VIBE [KAB20] που παρουσιάστηκε στο κεφάλαιο 3 και το επεκτείνουμε χρησιμοποιώντας παρατηρήσεις ανάλογες με αυτές της προηγούμενης ενότητας.

4.2.1 Autoregressive μοντέλο VIBE

Η πρώτη επέκταση που παρουσιάζουμε προκύπτει από την παρατήρηση ότι η μέθοδος VIBE δεν εκμεταλλεύεται την χρονική εξάρτηση στο πεδίου εξόδου. Πιο συγκεκριμένα, οπως αναφέραμε η μέθοδος VIBE κάνει χρήση ενός χρονικού κωδικοποιητή ώστε να παράξει χαρακτηριστικά εικόνας εμπλουτισμένα με την πληροφορία του χρόνου. Ωστόσο, αφου έχουν παραχθεί αυτά τα χαρακτηριστικά χρησιμοποιεί ένα προεκπαιδευμένο μοντέλο HMR και εκτιμά ανεξάρτητα για κάθε χρονική στιγμή τις παραμέτρους του μοντέλου SMPL.

Αντιθέτως προτείνουμε να τροποποιήσουμε την μέθοδο VIBE ώστε το μοντέλο HMR σε κάθε χρονική στιγμή να λαμβάνει σαν είσοδο εκτός από τα χρονικά εμπλουτισμένα χαρακτηριστικά εικόνας *και* τις παραμέτρους SMPL που έχουν εκτιμηθεί για την προηγούμενη χρονική στιγμή. Αυτο εικάζουμε πως θα βοηθήσει το μοντέλο να παράγει ακολουθίες πόζας με καλύτερη χρονική συνοχή.

Οσον αφορά στην υλοποίηση αυτής της αλλαγής, μια προφανής επιλογή είναι οι εκτιμήσεις της χρονικής στιγμής $t-1$ να δίνονται ως τα σημεία εκκίνησης της διαδικασίας εκτίμησης του μοντέλου HMR για την χρονική στιγμή t. Πιο αναλυτικά, όπως έχει ήδη παρουσιαστεί στο κεφάλαιο 3, το μοντέλο HMR προκειμένου να εκτιμήσει τις παραμέτρους SMPL, λαμβάνει ως είσοδο το χαρακτηριστικό εικόνας και μια αρχική εκτίμηση παραμέτρων και εκτιμά σταδιακά τις διορθώσεις που πρέπει να γίνουν σε αυτές τις παραμέτρους(residuals). Αυτη η αρχική εκτίμηση σε περιπτώσεις που δεν εχουμε κάποια πρότερη γνώση, επιλέγεται να είναι η μέση τιμή των παραμέτρων οπως προκύπτει από το σύνολο δεδομένων εκπαίδευσης. Αντιθέτως στην περίπτωση που μας απασχολεί επιλέγουμε να είναι οι εκτιμήσεις της προηγούμενης χρονική στιγμής.



Εικόνα 4.2.1: Η προτεινόμενη αρχιτεκτονική του μοντέλου VIBE με ακολουθιακή εκτίμηση πόζας.

Η προτεινόμενη αρχιτεκτονική φαίνεται στην Εικόνα 4.2.1. Όπως έχουμε προαναφέρει, οι εκτιμήσεις του μοντέλου HMR περιλαμβάνουν την εκτίμηση πόζας, σχήματος και παραμέτρων κάμερας. Δεδομένου οτι οι παράμετροι αυτές υπόκεινται σε διαφορετικούς περιορισμούς π.χ. η παράμετροι σχήματος οφείλουν να παραμένουν σταθερές για κάθε frame που εικονίζεται οι ίδιος άνθρωπος, αντιθέτως οι παράμετροι πόζας και κάμερας οφείλουν να μεταβάλλονται, εκπαιδεύουμε το μοντέλο κάτω από διαφορετικές συνθήκες όσον αφορά στην μεταφορά παραμέτρων μεταξύ των frames.

4.2.2 VIBE-HMR2

Στην ενότητα με τα πειράματα σχήματος παρουσιάστηκε το μοντέλο HMR2 που μοντελοποιεί ξεχωριστά την εκτίμηση σχήματος και πόζας. Αυτός ο διαχωρισμός εικάζουμε ότι μπορεί να φανεί χρήσιμος στην εκτίμηση πόζας και σχήματος από ακολουθία εικόνων καθώς τα υποσυστήματα σχήματος και πόζας έχουν διαφορετικές ανάγκες.

Για παράδειγμα για μια ακολουθία με το ιδιο υποκείμενο το υποσύστημα σχήματος χρειάζεται να παράγει τις ιδιες παραμέτρους, επομένως μαθαίνει να είναι αμετάβλητο σε χαρακτηριστικά πόζας. Αντιθέτως το υποσύστημα πόζας είναι υπεύθυνο να παράγει την εκτίμηση πόζας δεδομένου του σχήματος και επομένως να εντοπίζει χαρακτηριστικά που είναι χρήσιμα για αυτήν. Το γεγονός ότι τα υποσυστήματα δεν μοιράζονται μέρος της αρχιτεκτονικής, όπως συμβαίνει στο μοντέλο HMR, εικάζουμε οτι θα διευκολύνει την εκπαίδευση και θα βελτιώσει την απόδοση της μεθόδου VIBE.

Δεδομένου οτι το μοντέλο επιτυγχάνει όμοια απόδοση με το HMR σε στατικές εικόνες ενώ παράλληλα κατέχει αυτόν τον διαχωρισμό μας οδηγεί να πειραματιστούμε αντικαθιστώντας στη μέθοδο VIBE το HMR με αυτό. Ομοίως χρησιμοποιούμε ενα προεκπαιδευμένο μοντέλο ResNet50 ώστε να εξάγουμε εκ των προτέρων τα χαρακτηριστικά εικόνας. Επιπλέον τα βάρη του υποσυστήματος πόζας όπου είναι δυνατό εκκινούνται με τις τιμές των βαρών του προεκπαιδευμένου μοντέλου HMR από την μέθοδο [Kol+19].

4.2.3 VIBE με βοηθητική επίβλεψη

Είναι κοινώς αποδεκτό ότι τα κριτήρια Μέσου Τετραγωνικού Λάθους και Μέσου Απόλυτο Λάθους που χρησιμοποιούνται στα προβλήματα Παλινδρόμησης δεν παράγουν το ίδιο καλή επίβλεψη οπως τα κριτήρια που χρησιμοποιούνται στα προβλήματα Ταξινόμησης, με την έννοια οτι επηρεάζονται περισσότερο από τον θόρυβο στα δεδομένα και δεν κατευθείνουν τις παραμέτρους του μοντέλου προς "καλά" τοπικά ελάχιστα.

Με αφορμή την παραπάνω παρατήρηση, στρεφόμαστε σε προβλήματα τα οποία διαδραματίζουν τον ρόλο βοηθητικων προβλημάτων κατα την εκπαίδευση και τα οποία αποτελούν μια δευτερεύουσα πηγή επίβλεψης. Επιλέγουμε το πρόβλημα της κατηγοριοποίησης δράσης από βίντεο οπου μας δίνεται μια ακολουθία από εικόνες στις οποίες ενα υποκείμενο πραγματοποιεί μια δραστηριότητα και σκοπός είναι να ταξινομήσουμε την ακολουθία σε μια από τις διαθέσιμες κλάσεις. Καταλήγουμε σε αυτο το πρόβλημα καθώς αφενός η πόζα αποτελεί ενα χαρακτηριστικό από το οποίο μπορεί να γίνει η ταξινομηση δράσης και αφετέρου είναι ένα πρόβλημα που βασίζεται στην διάσταση του χρόνου.

Όσον αφορά στο σύνολο δεδομένων, επιλέγουμε το Penn Action οπου περιέχει:

- 2326 δείγματα εκπαίδευσης, κάθε δείγμα είναι μια ακολουθία εικόνων.
- 15 διαφορετικούς τύπους δράσης.
- 2Δ επισημειώσεις του σκελετού στο επίπεδο της εικόνας.

.

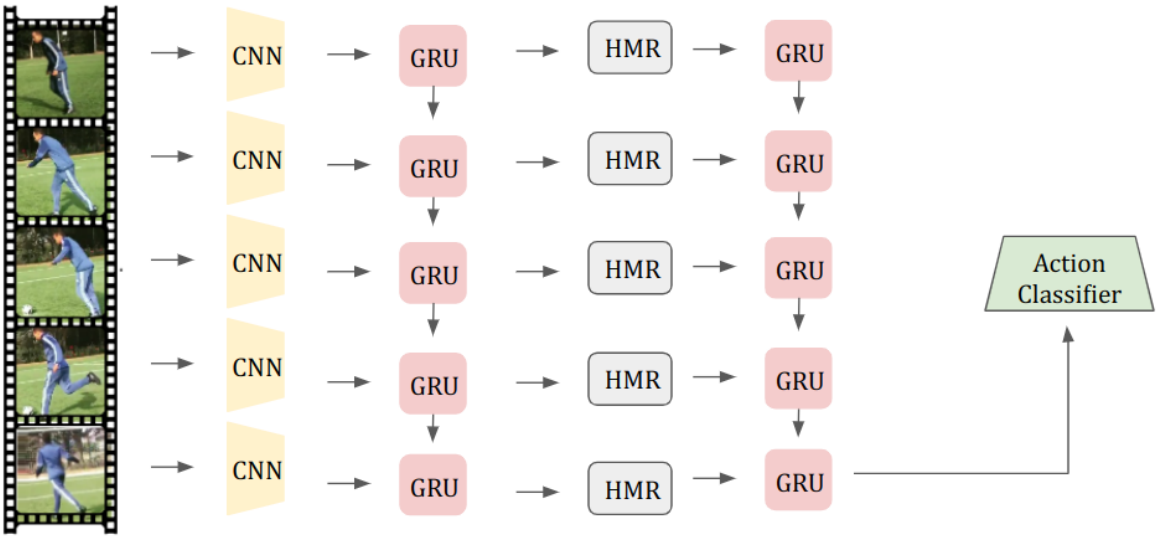
Στην Εικόνα 4.2.2 φαίνονται εικόνες απο τέσσερις δράσεις του συνόλου δεδομένων ενώ στον Πίνακα 4.1 φαίνεται οι κατανομή στις διαφορετικές κλάσεις.

Για την υλοποίηση, επεκτείνουμε την αρχιτεκτονική της μεθόδου VIBE εισάγοντας ένα αναδρομικό νευρωνικό δίκτυο μετά την έξοδο του μοντέλου HMR. Σαν αναδρομικό νευρωνικό δίκτυο επιλέγουμε την αρχιτεκτονική GRU ενός επιπέδου όπου έχει αναλυθεί στο κεφάλαιο 2. Σε κάθε χρονική στιγμή το νευρωνικό δίκτυο λαμβάνει σαν είσοδο τις παραμέτρους πόζας (24 σχετικές περιστροφές για τις 24 αρθρώσεις) που εκτιμά του μοντέλο HMR και παράγει την κρυφή αναπαράσταση για αυτή την χρονική στιγμή. Η κρυφή αναπαράσταση της τελευταίας χρονικής στιγμής δίνεται σαν είσοδος στην κεφαλή ταξινόμησης που παράγει τις πιθανότητες για κάθε κλάση. Η συνολική αρχιτεκτονική φαίνεται στην Εικόνα 4.2.3.

Εικόνα 4.2.2: Δείγματα εικόνων απο το σύνολο δεδομένων PennAction.

Δράση	Δείγματα
Baseball pitch	14460
Baseball swing	7561
Bench press	11576
Bowl	12626
Clean and jerk	23862
Golf swing	9169
Jump rope	3642
Jumping jacks	3362
Pullup	13865
Pushup	10513
Situp	8763
Squat	21351
Strum guitar	4206
Tennis forehand	7203
Tennis serve	11682

Πίνακας 4.1: Αποτελέσματα του μοντέλου HMR3



Εικόνα 4.2.3: Η προτεινόμενη αρχιτεκτονική του μοντέλου VIBE για κατηγοριοποίηση δράσης.

Κεφάλαιο 5

Πειράματα

Model	PA-MPJPE ↓
HMR [JNV21]	**57.5**
HMR Σεν. 1	61.3
HMR Σεν. 2	59.5
HMR Σεν. 3	**58.7**

Πίνακας 5.1: Αποτελέσματα υλοποίησης μοντέλου HMR.

5.1 Υλοποιήσεις Μοντέλων Αναφοράς

Σαν πρώτο βήμα, επιλέγουμε να υλοποιήσουμε δύο από τα πιό δημοφιλή μοντέλα, HMR και ProHMR, τα οποία έχουν παρουσιαστεί αναλυτικά στο κεφάλαιο 3 και το οποία θα λειτουργήσουν και ως τα μοντέλα με την απόδοση αναφοράς στα πειράματα μας. Η υλοποίηση των μοντέλων και η αναπαραγωγή των αποτελεσμάτων που παρουσιάζονται στις εργασίες είναι απαραίτητο βήμα προτού προχωρήσουμε σε τροποποιήσεις καθώς θα επαληθεύσει ότι οποιαδήποτε διαφορά προκύψει στην απόδοση θα οφείλεται σε αυτές τις τροποποιήσεις και όχι σε άλλους παράγοντες της διαδικασίας.

Σε αντίθεση με τον αρχικό τρόπο εκπαίδευσης όπου κάνει χρήση δισδιάστατων και τρισδιάστατων επισημειώσεων, ακολουθούμε την πλειοψηφία της μετέπειτα βιβλιογραφίας και κάνουμε χρήση των επισημειώσεων με την μορφή SMPL παραμέτρων που δημιούργησαν οι ερευνητές [JNV21]. Πιο συγκεκριμένα, για κάθε δεδομένο εισόδου οι ερευνητές βελτιστοποίησαν ως προς το λάθος επαναπροβολής τις παραμέτρους του νευρωνικού δικτύου αντί για τις παραμέτρους του μοντέλου SMPL οπως γίνεται στην μέθοδο SMPLify. Αποτέλεσμα αυτής της διαδικασίας είναι ψευδο-ορθές(pseudo ground truth) επισημειώσεις με την μορφή SMPL για 4 κοινά σύνολα δεδομένων με δισδιάστατες επισημειώσεις:

- COCO: 79 χιλ. επισημειώσεις εκπαίδευσης και 10 χιλ. επισημειώσεις επαλήθευσης.
- MPII: 14.3 χιλ. επισημειώσεις εκπαίδευσης
- LSPet: 3 χιλ. επισημειώσεις εκπαίδευσης και 2.4 χιλ. επισημειώσεις αξιολόγησης.
- OCHuman: 2.5 χιλ επισημειώσεις και 1.7 χιλ. επισημειώσεις αξιολόγησης.

Στα πειράματα μας χρησιμοποιούμε τις επισημειώσεις για το σύνολο COCO ως δείγματα εκπαίδευσης και επαλήθευσης και τις επισημειώσεις για το σύνολο MPII ως δείγματα αξιολόγησης πέρα από τα δείγματα του συνόλου 3DPW. Όπως αναφέρουν οι ερευνητές [JNV21] οι επισημειώσεις με αυτή τη μορφή αναγάγουν το πρόβλημα της ανακατασκευής σε ένα κοινό πρόβλημα παλινδρόμησης το οποίο διευκολύνει την διαδικασία εκπαίδευσης.

5.1.1 Υλοποίηση μοντέλου HMR

Η αρχιτεκτονική και ο τρόπος λειτουργίας του μοντέλου έχουν ήδη παρουσιαστεί στο κεφάλαιο 3. Εδω παρουσιάζουμε τα ποσοτικά και οπτικά αποτελέσματα της υλοποίησης μας. Δεδομένου ότι έχουμε διαφορετικού είδους επισημειώσεις και ότι στην εργασία [JNV21] δεν υπάρχει αναφορά στον τρόπο με τον οποίο συνδυάζονται οι διαφορετικού είδους επισημειώσεις δοκιμάζουμε διαφορετικά σενάρια εκπαίδευσης τα οποία είναι τα εξής:

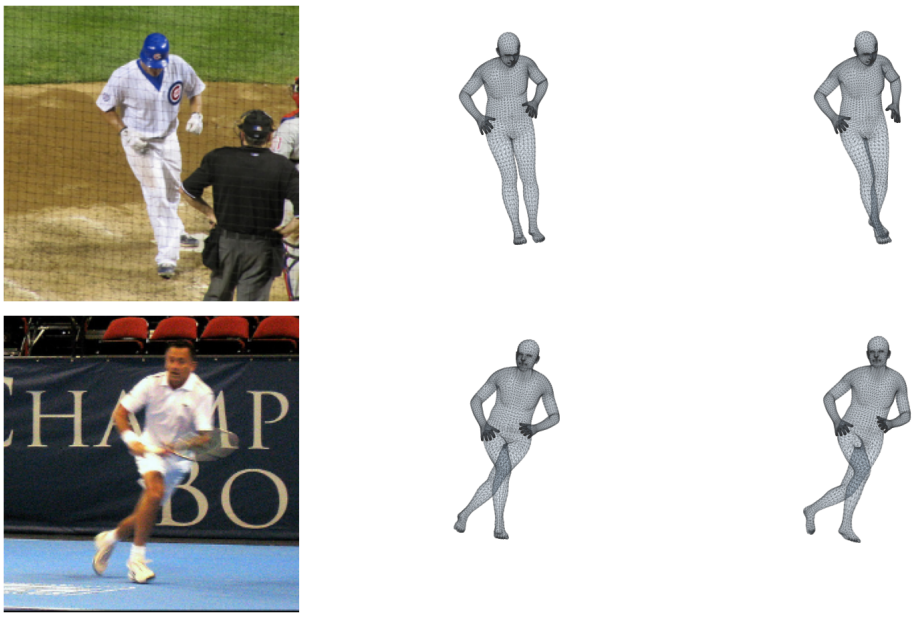
- **Σενάριο 1**: Εκπαίδευση κάνοντας χρήση SMPL και τρισδιάστατων επισημειώσεων.
- **Σενάριο 2**: Εκπαίδευση κάνοντας χρήση SMPL,τρισδιάστατων και δισδιάστατων επισημειώσεων χρησιμοποιώντας σαν κριτήριο το Μέσο Τετραγωνικό Λάθος.
- **Σενάριο 3**: Εκπαίδευση κάνοντας χρήση SMPL,τρισδιάστατων και δισδιάστατων επισημειώσεων χρησιμοποιώντας σαν κριτήριο το Μέσο Απόλυτο Λάθος.

Σημείωση: Οι τρισδιάστατες επισημειώσεις παράγονται από τις SMPL επισημειώσεις. Αν και ο συνδυασμός τους αποτελεί πλεονάζουσα πληροφορία επίβλεψης οι ερευνητές [Kan+18] θεωρούν ότι αυτό βοηθάει την εκπαίδευση του μοντέλου.

Τα αποτελέσματα φαίνονται στον Πίνακα 5.1 για το σύνολο 3DPW. Η μετρική αξιολόγησης είναι η Procrustes Aligned Mean Per Joint Position Error που παρουσιάστηκε στο κεφάλαιο 2 εκφρασμένη σε mm.

Παρατηρούμε πως η προσθήκη του λάθους επαναπροβολής βελτιώνει την απόδοση του μοντέλου ενω η χρήση του μέσου απόλυτου λάθους υπερτερεί έναντι του μέσου τετραγωνικού λάθους. Αυτη η παρατήρηση έρχεται σε συμφωνία με τις παρατηρήσεις των ερευνητών [Pan+22].

Στην Εικόνα 5.1.1 παρουσιάζονται οπτικά αποτελέσματα ανακατασκευής με την χρήση του μοντέλου HMR.



Εικόνα 5.1.1: Παραδείγματα ανακατασκευής με τα μοντέλα HMR(2η στήλη) και ProHMR(3η στήλη)

5.1.2 Υλοποίηση μοντέλου ProHMR

Ομοίως, η αρχιτεκτονική του μοντέλου και ο τρόπος λειτουργίας των Normalizing Flows έχει παρουσιαστεί στο κεφάλαιο 3. Εδώ παρουσιάζουμε τα αποτελέσματα της υλοποίησης μας και κάποια ποιοτικά πειράματα στον *κρυφό χώρο* του μοντέλου.

Δεδομένου ότι ο όρος του λάθος επαναπροβολής όπως φαίνεται από τα παραπάνω αποτελέσματα ωφελεί την εκπαίδευση, εκπαιδεύουμε το συγκεκριμένο μοντέλο κάνοντας χρήση SMPL,τρισδιάστατων και δισδιάστατων επισημειώσεων χρησιμοποιώντας σαν κριτήριο το Μέσο Απόλυτο Λάθος. Οπως παρουσιαστηκε στο κεφάλαιο 3 η συνολική συνάρτηση λαθους δίνεται από την σχέση:

$$L = L_{nll} + L_{joints} + L_{proj} + L_{SMPL} + L_{exp}$$

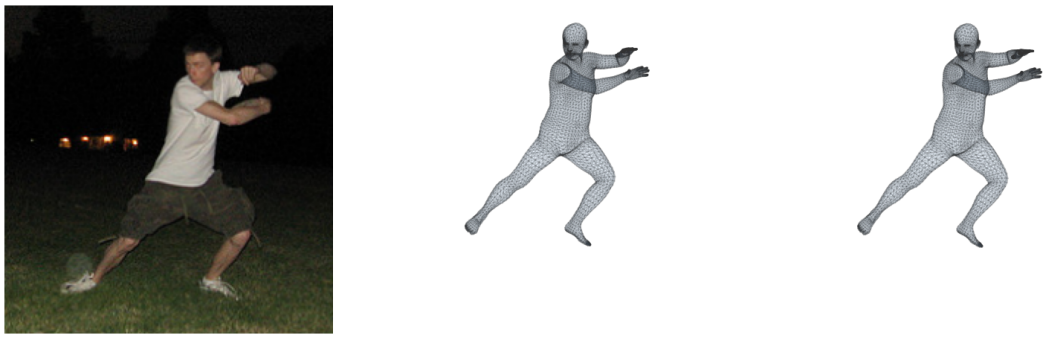
Σε αρχικά πειράματα παρατηρήσαμε ότι η χρήση μόνο του όρου μεγιστοποίησης πιθανοφάνειας οδηγεί ασταθή εκπαίδευση οπου τα βάρη του μοντέλου οδηγούνται σε μεγάλες τιμές οπως παρατηρείται συχνά κατα την εκπαίδευση μοντέλων Normalizing Flows. Εικάζουμε ότι αυτό οφείλεται στον όρο $|det(\frac{\partial f^{-1}(x)}{\partial x})|$ της σχέσης αλλαγής μεταβλητών όπου οδηγεί την εξοδο του μοντέλου να αυξάνεται διαρκώς ώστε να αυξηθεί η πιθανοφάνεια των δειγμάτων. Επομένως ενδιαφέρον παρουσιάζει το γεγονός πως οι όροι ανακατασκευής λειτουργούν και ως περιορισμοί που βοηθούν στην ευστάθεια της εκπαίδευσης.

Τα αποτελέσματα υλοποίησης φαίνονται στον Πίνακα 5.2 για το σύνολο 3DPW. Η μετρική αξιολόγησης ομοίως με πριν είναι η PA-MPJPE εκφρασμένη σε mm.

Παραδείγματα ανακατασκευής φαίνονται στην Εικόνα 5.1.1 και δείγματα από την κατανομή του ProHMR στην Εικόνα 5.1.2. Επιπλέον, προκειμένου να κατανοήσουμε τις ιδιότητες του "κρυφού" χώρου (στην προκειμένη περίπτωση ο χώρος της τυχαίας μεταβλητής που ακολουθεί την βασική κατανομή), επιλέγουμε μια τυχαία κατεύθυνση και καθώς μετακινούμαστε πάνω σε αυτή εισάγουμε τα σημεία στην συνάρτηση f προκειμένου να λάβουμε τα αντίστοιχα σημεία στον χώρο πόζας. Τα δείγματα φαίνονται στην Εικόνα 5.1.3. Ενδιαφέρον

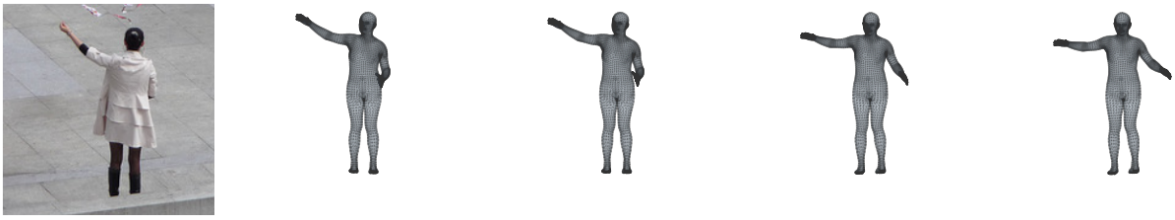
Model	PA-MPJPE ↓
ProHMR [Kol+21]	**59.8**
ProHMR υλοπ.	60.2

Πίνακας 5.2: Αποτελέσματα υλοποίησης μοντέλου ProHMR.



Εικόνα 5.1.2: Δείγματα από την κατανομή του ProHMR για την συγκεκριμένη εικόνα εισόδου.

παρουσιάζει το γεγονός πως καθώς μετακινούμαστε ομαλά στον κρυφό χώρο οι πόζες μεταβάλλονται και αυτές ομαλά.



Εικόνα 5.1.3: Δείγματα από τον κρυφό χώρο του ProHMR καθώς διατρέχουμε ενα περίπατο.

5.2 Πειράματα Σχήματος

5.2.1 HMR2

Προκειμένου να καταλάβουμε την επίδραση του κάθε υποσυστήματος στην επίδοση του μοντέλου πραγματοποιούμε το πείραμα σε δύο διαφορετικά σενάρια:

- **Σενάριο 1**: Κατα την διάρκεια της εκπαίδευσης *και* αξιολόγησης είναι διαθέσιμες οι επισημειώσεις σχήματος επομένως δεν χρειαζόμαστε το υποσύστημα που είναι υπεύθυνο για την εκτίμηση σχήματος. Το υποσύστημα που είναι υπεύθυνο για την εκτίμηση πόζας λαμβάνει τις πραγματικές τιμές των παραμέτρων σχήματος επομένως το μοντέλο είναι υπεύθυνο μονο για την εκτίμηση πόζας. Από αυτό το πείραμα λαμβάνουμε ένα είδους άνω όριο στην επίδοση του μοντέλου.

- **Σενάριο 2**: Κατα την διάρκεια της εκπαίδευσης το μοντέλο εκτιμά τις παραμέτρους σχήματος με το αντίστοιχο υποσύστημα οι οποίες δίνονται σαν είσοδοι στο επόμενο υποσύστημα για την εκτίμηση πόζας. Οι εκτιμώμενες παράμετροι σχήματος επιβλέπονται άμεσα από τις επισημειώσεις.

- **Σενάριο 3**: Κατα την διάρκεια της εκπαίδευσης το μοντέλο εκτιμά τις παραμέτρους σχήματος με το αντίστοιχο υποσύστημα οι οποίες δίνονται σαν είσοδοι στο επόμενο υποσύστημα για την εκτιμηση πόζας. Ωστόσο στο συγκεκριμένο σενάριο οι προβλέψεις σχήματος δεν επιβλέπονται άμεσα μέσω των επισημειώσεων, αντιθέτως η επίβλεψη για το υποσύστημα σχήματος προκύπτει έμμεσα από την επίβλεψη με τις επισημειώσεις πόζας.

Model	MPII ↓	3DPW ↓
HMR baseline	73.9	**58.7**
HMR2 Σενάριο 1	**68.1**	-
HMR2 Σενάριο 2	73	59.0
HMR2 Σενάριο 3	73.1	58.8

Πίνακας 5.3: Αποτελέσματα υλοποίησης μοντέλου HMR2.

Model	6D repr. ↓	Axis-Angle ↓
HMR	0.0716	0.0625
HMR2 Σενάριο 1	**0.0669**	**0.0573**

Πίνακας 5.4: Λάθος στην εκτίμηση περιστροφών οπως εκφράζεται από την L2 νόρμα μεταξύ των εκτιμήσεων και επισημειώσεων.

Με βάση τις παρατηρήσεις μας για την εκπαίδευση του μοντέλου HMR στην προηγούμενη ενότητα, κατα την εκπαίδευση του μοντέλου HMR2 κάνουμε χρήση δισδιάστατων, τρισδιάστατων και SMPL επισημειώσεων και επιλέγουμε ως κριτήριο εκπαίδευσης το Μέσο Απόλυτο Λάθος. Επιπλέον προκειμένου να αξιολογήσουμε το σενάριο 1 χρειαζόμαστε ένα σύνολο με επισημειώσεις SMPL. Για αυτό το λόγο χρησιμοποιούμε τις επισημειώσεις που έχουν προκύψει από την μέθοδο [JNV21] που αναφέρθηκε παραπάνω για το σύνολο MPII. Τα αποτελέσματα ως προς το σύνολο MPII για όλα τα σενάρια και για το σύνολο 3DPW για όλα πλην του σεναρίου 1 φαίνονται στον Πίνακα 5.3 . Η μετρική αξιολόγησης είναι η PA-MPJPE.

Παρατηρούμε ότι το μοντέλο HMR2 στο σενάριο 1 επιτυγχάνει ουσιώδη βελτίωση ως προς το baseline. Προκειμένου να επαληθεύσουμε ότι το κέρδος στην απόδοση προέρχεται από καλύτερη εκτίμηση πόζας και όχι απλά στην καλύτερη θέση των αρθρώσεων υπολογίζουμε το λάθος ως προς τις περιστροφές υπολογίζοντας την L2 νορμα της διαφοράς των εκτιμώμενων περιστροφών με τις επισημειώσεις. Πιο συγκεκριμένα ως αναπαράσταση περιστροφών χρησιμοποιούμε την 6Δ αναπαράσταση των [**ÎÛhou_2018**] και την axis angle αναπαράσταση. Τα αποτελέσματα φαίνονται στον Πίνακα 5.4 και επαληθεύουν πως το κέρδος απόδοσης οφείλεται *και* στην καλύτερη εκτίμηση των περιστροφών.

Στα σενάρια 2 και 3, τα μοντέλα επιτυγχάνουν παρόμοια επίδοση με το baseline. Δεδομένου ότι το σενάριο 3 επιτυγχάνει οριακά καλύτερα αποτελέσματα, η σημαντικότητα της επίβλεψης στο ενδιάμεσο στάδιο φαίνεται πως είναι αμελητέα.

Από τα παραπάνω αποτελέσματα επιβεβαιώνεται η σημασία της πληροφορίας σχήματος κατα την διαδικασία εκτίμησης πόζας αλλα και η ανάγκη για καλύτερη εκτίμηση σχήματος.

5.2.2 HMR3

Ομοίως με την εκπαίδευση του μοντέλου HMR2, εκπαιδεύουμε το μοντέλο HMR3 σε 2 σενάρια:

- **Σενάριο 1**: Κατα την διάρκεια της εκπαίδευσης *και* αξιολόγησης είναι διαθέσιμη η αναπαράσταση σχήματος από τις επισημειώσεις σχήματος επομένως δεν χρειαζόμαστε το υποσύστημα που είναι υπεύθυνο για την εκτίμηση της αναπαράστασης. Ομοίως με παραπάνω από αυτό το πείραμα λαμβάνουμε ένα είδους άνω όριο στην επίδοση του μοντέλου.
- **Σενάριο 2**: Κατα την διάρκεια της εκπαίδευσης το μοντέλο εκτιμά την γενικευμένη αναπαράσταση με το αντίστοιχο υποσύστημα οι οποίες δίνονται σαν είσοδοι στο επόμενο υποσύστημα για την εκτίμηση πόζας και των παραμέτρων σχήματος SMPL. Οι εκτιμώμενες αναπαραστάσεις σχήματος επιβλέπονται άμεσα από τις πραγματικές αναπαραστάσεις σχήματος οπως προκύπτουν από τις επισημειώσεις.

Τα αποτελέσματα φαίνονται στον Πίνακα 5.5 για τα σύνολα δεδομένων MPII και 3DPW ως προς την μετρική PA-MPJPE.

Παρατηρούμε οτι στο σύνολο MPII το μοντέλο κάτω από τις δύο συνθήκες επιτυγχάνει παρόμοια αποτελέσματα ενώ στο σύνολο 3DPW υπάρχει μια μεγαλύτερη απόκλιση. Ωστόσο Θα περιμέναμε το μοντέλο που λαμβάνει τις αληθινές παραμέτρους(Σενάριο 1) να επιτυγχάνει τα ίδια ή καλύτερα αποτελέσματα απο το μοντέλο που τις

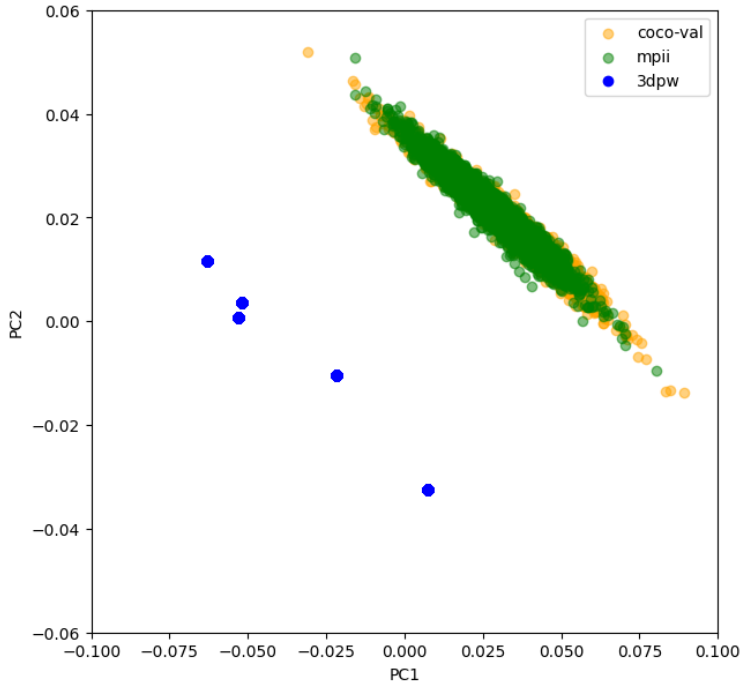
Model	MPII ↓	3DPW ↓
HMR	73.9	**58.7**
HMR3 Σενάριο 1	73.9	59.4
HMR3 Σενάριο 2	**73.6**	58.8

Πίνακας 5.5: Αποτελέσματα του μοντέλου HMR3

εκτιμά(Σενάριο 2). Προκειμένου να διερευνήσουμε την αιτία χρησιμοποιούμε την μέθοδο μείωσης διάστασης PCA που θα μας δώσει την δυνατότητα να οπτικοποιήσουμε ενα υποσύνολο των δεδομένων. Πιο συγκεκριμένα υπολογίζουμε το χαρακτηριστικό σχήματος για τρία σύνολα δεδομένων:

- COCO2014 Validation Set: αποτελεί το σύνολο επαλήθευσης του πειράματος το οποίο δεδομένου ότι είναι υποσύνολο του σύνολου δεδομένων COCO2014, όπως και το σύνολο εκπαίδευσης μας, ακολουθεί την ίδια κατανομή χαρακτηριστικών με το σύνολο εκπαίδευσης.
- MPII: οπως αναλύσαμε παραπάνω, χρησιμοποιούμε αυτό το σύνολο για αξιολόγηση επειδή παρέχει επισημειώσεις SMPL εξαγμένες με την μέθοδο [JNV21].
- 3DWP: το σύνολο αξιολόγησης που παρέχει επισημειώσεις SMPL.

και στην συνέχεια εφαρμόζουμε την μέθοδο PCA. Οπτικοποιούμε το αποτέλεσμα κρατώντας τις δύο πρώτες διαστάσεις για κάθε χαρακτηριστικό. Το αποτέλεσμα της μεθόδου PCA φαίνεται στην Εικόνα 5.2.1



Εικόνα 5.2.1: Το αποτέλεσμα της μεθόδου PCA πάνω στο σύνολο χαρακτηριστικών σχήματος. Παρατηρείται η διαφορά στην κατανομή μεταξύ των συνόλων COCO, MPII και 3DPW.

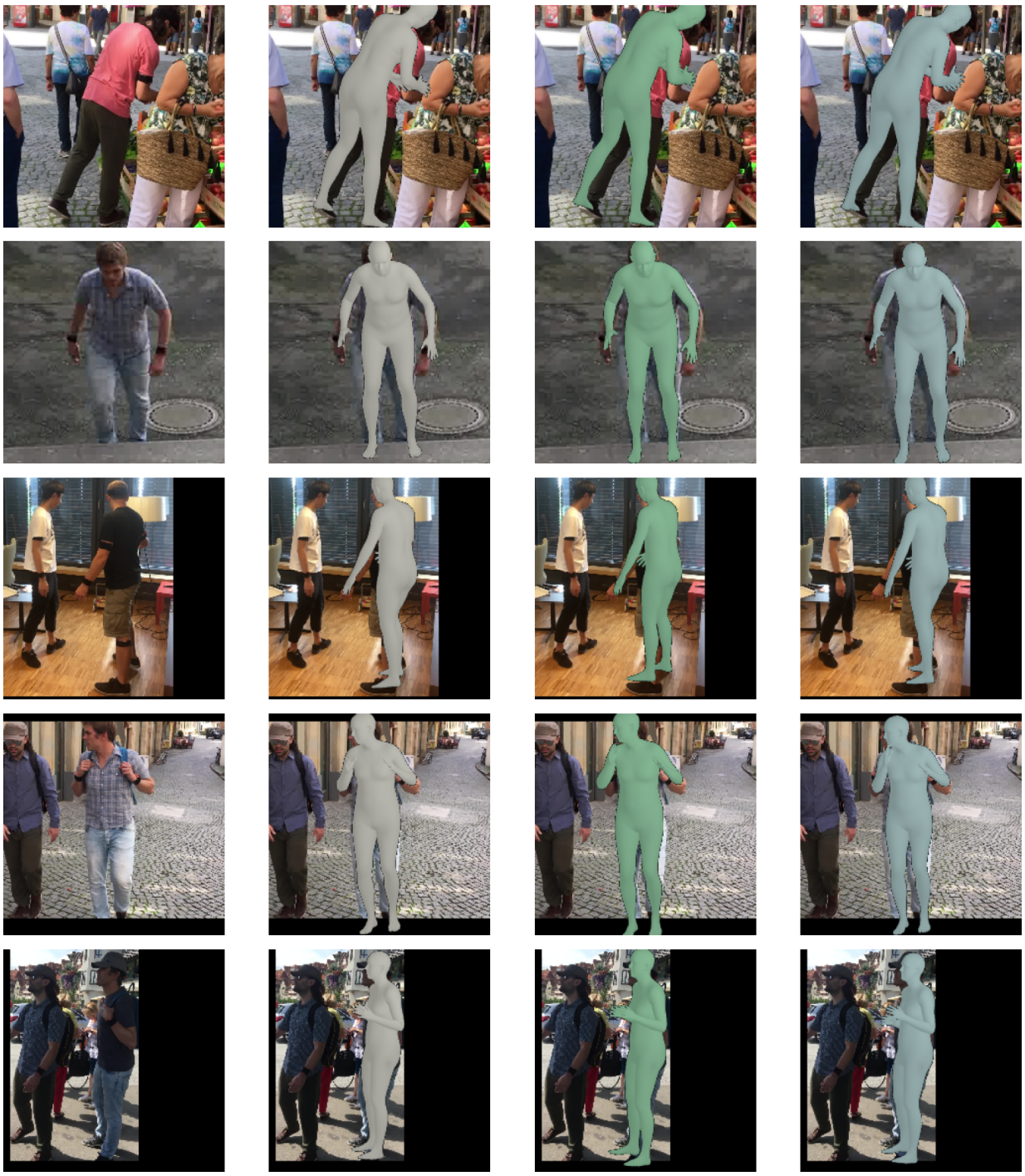
Παρατηρούμε ότ υπάρχει διαφορά ανάμεσα στις κατανομές των συνόλων COCO2014-Val, MPII και 3DPW. Αυτο μπορεί να εξηγήσει το γεγονός πως δυσχεραίνει η απόδοση του μοντέλου στο σύνολο 3DPW καθώς αυτή η μετατόπιση στον χώρο των χαρακτηριστικών είναι μια πληροφορία που το μοντέλο δεν έχει αντιμετωπίσει κατα την διάρκεια της εκπαίδευσης.

Η αιτία αυτής της μετατόπισης δεν είναι αναμενόμενη. Αρχικά σημειώνουμε οτι σε σχέση με τα σύνολα εικόνας COCO2014,MPII το σύνολο βίντεο 3DPW φαίνεται να έχει λιγότερα δείγματα στο σχήμα. Αυτο οφείλεται στο γεγονός πως αν και διαλέξαμε ίσο αριθμό δειγμάτων από τα τρία σύνολα, το σύνολο 3DPW περιέχει

πολλά video με τους ίδιους ανθρώπους, επομένως τα διαφορετικά σύνολα χαρακτηριστικών είναι όσα και οι διαφορετικοί άνθρωποι εμφανίζονται στο βίντεο. Όσον αφορά στην μετατόπιση, εικάζουμε πως οφείλεται στην μέθοδο εξαγωγής [JNV21]. Τα σύνολα COCO2014, MPII που χρησιμοποιούν την ίδια διαδικασία, ακολουθούν την ίδια κατανομή σε αντίθεση με το σύνολο 3DPW που οι παράμετροι σχήματος SMPL έχουν εξαχθεί έπειτα από σκανάρισμα των σωμάτων.

5.2.3 Ποιοτικά αποτελέσματα

Πέρα από την οπτικοποίηση του πλέγματος στον τρισδιάστατο χώρο οπως έγινε στην προηγούμενη ενότητα 5.1 υπάρχει η δυνατότητα να εκμεταλλευτούμε την εκτίμηση παραμέτρων κάμερας και να αξιοποιήσουμε αυτή την εικονική κάμερα για να προσομοιώσουμε την προβολή του πλέγματος(rendering) στο επίπεδο της εικόνας. Στην Εικόνα 5.2.2 φαίνονται τέτοια παραδείγματα για τα μοντέλα HMR, HMR2 και HMR3. Στις περιπτώσεις των μοντέλων HMR2 και HMR3 έχει γίνει χρήση των μοντέλων που εκτιμούν τις αναπαραστάσεις σχήματος. Οπτικά, τα αποτελέσματα είναι παρόμοια και για τα τρία μοντέλα. Παρατηρούμε ότι σε ορισμένες περιπτώσεις όπως στην τρίτη στήλη της δεύτερης εικόνας οτι το μοντέλο HMR2 παράγει πιο αληθοφανείς εκτιμήσεις σχήματος



Εικόνα 5.2.2: Παραδείγματα προσομοίωσης της προβολής του πλέγματος στο επίπεδο της εικόνας κάνοντας χρήση της εκτιμώμενης κάμερας. Παρουσιάζονται αποτελέσματα ανακατασκευής απο τα μοντέλα HMR(2η στήλη) και HMR2(3η στήλη) και HMR3(4η στήλη)

Επιπλέον στην Εικόνα 5.2.3 παρουσιάζουμε και παραδείγματα από τις περιπτώσεις όπου τα μοντέλα λόγω εμποδίων δεν επιτυγχάνουν ικανοποιητικά αποτελέσματα.

Εικόνα 5.2.3: Παραδείγματα οπου εμπόδια μεταξύ της κάμερας και του υποκειμένου οδηγούν τα μοντέλα σε μη ικανοποιητικά αποτελέσματα. Παρουσιάζονται αποτελέσματα ανακατασκευής απο τα μοντέλα HMR(2η στήλη), HMR2(3η στήλη) και HMR3(4η στήλη)

5.3 Πειράματα με είσοδο ακολουθία εικόνων

Αρχικά χρησιμοποιούμε τον κώδικα που παρέχουν οι ερευνητές προκειμένου να αναπαράγουμε τα αποτελέσματα που αναφέρονται στην εργασία. Σαν μια πρώτη επέκταση προτείνουμε την AutoRegressive αντιμετώπιση του προβλήματος χρησιμοποιώντας κάθε φορά την προηγούμενη εκτίμηση ως αρχική εκτίμηση του μοντέλου HMR. Στην συνέχεια προτείνουμε την χρήση του μοντέλου HMR2 της προηγούμενης ενότητας ως αντικατάσταση του μοντέλου HMR ενώ στην συνέχεια εξετάζουμε τρόπους χρήσιμης επίβλεψης χρησιμοποιώντας την εκτιμώμενη πόζα ως μέσω για την κατηγοριοποίηση δράσης.

5.3.1 Αναπαραγωγή Αποτελεσμάτων VIBE

Σαν πρώτο βήμα χρησιμοποιούμε τον κώδικα που παρέχουν οι ερευνητές προκειμένου να αναπαράξουμε τα αποτελέσματα που αναφέρονται στην εργασία. Επιπλέον, δεδομένου ότι κατα την εκπαίδευση του μοντέλου κάθε δείγμα αποτελείται από 16 καρέ ενός βίντεο το οποίο δημιουργεί μεγάλες απαιτήσεις για μνήμη, ακολουθούμε την μέθοδο των συγγραφέων οπου χρησιμοποιούν ένα μοντέλο ResNet50 προεκπαιδευμένο στο αντικείμενο της 3Δ ανακατασκευής πόζας και σχήματος για να εξάγουν εκ των προτέρων τα χαρακτηριστικά εικόνας. Επιπλεόν το μοντέλο HMR που χρησιμοποιείται για να κάνει τις εκτιμήσεις είναι προ εκπαιδευμένο με την μέθοδο [Kol+19]. Τα αποτελέσματα αναπαραγωγής φαίνονται στον Πίνακα 5.6 για το σύνολο δεδομένων 3DPW ως προς την μετρική PA-MPJPE εκφρασμένη σε mm. Αν και οι συγγραφείς χρησιμοποιούν το Μέσο Τετραγωνικό Λάθος, πειραματιζόμαστε και με τη χρήση του Μέσο Απόλυτου Λάθους καθώς όπως φάνηκε στις προηγούμενες ενότητες βοηθάει την εκπαίδευση.

Παρατηρούμε οτι αφενός μπορούμε να αναπαράξουμε τα αποτελέσματα της έρευνας και αφετέρου οτι σε αντίθεση με το μοντέλο HMR, η εκπαίδευση με το Μέσο Απόλυτο Λάθος στο συγκεκριμένο μοντέλο δεν ωφελεί ιδιαίτερα την εκπαιδευση.

5.3.2 AutoRegressive μοντέλο

Τα αποτελέσματα για το αρχικό προτεινόμενο μοντέλο(AR VIBE) φαίνονται στον Πίνακα 5.7 για το σύνολο δεδομένων 3DPW. Προκειμένου να αξιολογήσουμε την χρονική συνοχή των εκτιμήσεων εκτός από την μετρική

Model	3DPW ↓
VIBE	56.5
VIBE αναπαραγ. με MSE	56.5
VIBE αναπαραγ. με MAE	**56.4**

Πίνακας 5.6: Αποτελέσματα αναπαραγωγής του μοντέλου VIBE. Στην πρώτη γραμμή φαίνεται η απόδοση του μοντέλου οπως αναφέρεται στην ερευνητική εργασία. Στις επόμενες δύο γραμμές φαίνονται τα αποτελέσματα των μοντέλων που εκπαιδεύσαμε κάνοντας χρήση του κώδικα που παρείχαν οι συγγραφείς.

Model	PA-MPJPE ↓	ACC ↓
VIBE	56.5	27.1
AR VIBE	56.6	21.8
AR VIBE + More Initial Iters	**56.0**	21.6
AR VIBE + Shape	56.9	22.9
AR VIBE + Single Shape	57.6	**20.2**
AR VIBE + Moving Avg. Shape	56.5	20.9

Πίνακας 5.7: Αποτελέσματα της Auto Regressive μεθόδου.

PA-MPJPE υπολογίζουμε και τη μέση επιτάχυνση των αρθρώσεων.

Παρατηρούμε πως το μοντέλο δεν επιτυγχάνει κάποιο κέρδος στην απόδοση ως προς την μετρική ανακατασκευης ωστόσο επιβεβαιώνεται η παρατήρηση μας περι χρονικής συνοχής καθώς το μοντέλο πετυχαίνει ουσιαστική μείωση της επιτάχυνσης των αρθρώσεων.

Το γεγονός πως δεν βελτιώνεται η απόδοση ανακατασκευής του μοντέλου εικάζουμε ότι οφείλεται στην αδυναμία καλής αρχικής εκτίμησης του μοντέλου. Πιο συγκεκριμένα, το γεγονός οτι μεταφέρουμε τις εκτιμήσεις της προηγούμενης χρονικής στιγμή έχει θετικό και αρνητικό αντίκτυπο. Το θετικό είναι ότι τα αποτελέσματα του μοντέλου έχουν καλύτερη χρονική συνοχή οπως μετράται από την επιτάχυνση των αρθρώσεων. Ωστόσο το αρνητικό είναι οτι μια κακή εκτίμηση την χρονική στιγμή t μεταδίδεται στις επόμενες χρονικές στιγμές δυσχεραίνοντας την διαδικασία των επόμενων προβλέψεων. Προκειμένου να διαπιστώσουμε αν μια καλυτερη εκτίμηση στο αρχικό frame όντως βοηθάει την απόδοση δοκιμάζουμε το εξής πείραμα. Σε αντίθεση με τα υπόλοιπα frames οπου το μοντέλο εκτελεί 3 επαναλήψεις για κάθε εκτίμηση, στο πρώτο frame εκτελούμε 6 επαναλήψεις ώστε να παράξουμε, αν είναι δυνατό, μια καλή εκτίμηση χωρίς να αυξηθεί η υπολογιστική πολυπλοκότητα του μοντέλου. Οπως φαίνεται στον Πίνακα 5.7(AR VIBE + More Initial Iters) οντως αυξάνεται η απόδοση του μοντέλου ως προς την μετρική ανακατασκευής ενισχύοντας την παραπάνω εικασία μας.

Αν και η μετρική ανακατασκευής βελτιώνεται, ποιοτικά παρατηρήσαμε οτι το συγκεκριμένο μοντέλο διαρκώς μεταβάλει τις παραμέτρους σχήματος οδηγώντας σε μη ικανοποιητικά αποτελέσματα. Ένα παράδειγμα αυτής της αστοχίας φαίνεται στην Εικόνα 5.3.1 οπου μεταξύ των frames διακρίνεται οι αλλαγή του σώματος του υποκειμένου.

Εικόνα 5.3.1: Παράδειγμα αστοχίας της μεθόδου AR VIBE οπου οι παράμετροι σχήματος διαρκώς μεταβάλλονται κατα την διάρκεια του βίντεο οδηγώντας σε μη αληθοφανές αποτέλεσμα.

Οι συγγραφείς της μεθόδου VIBE προκειμένου να παράγουν μη μεταβαλλόμενες κατα την διάρκεια του βίντεο παραμέτρους σχήματος, εκτιμούν τις παραμέτρους σχήματος για κάθε χρονική στιγμή και στην συνέχεια υπολογίζουν τον μέσο όρο για το συγκεκριμένο χρονικό παράθυρο. Στην συνέχεια δίνουν αυτές τις παραμέτρους στο μοντέλο SMPL σε κάθε frame.

Model	PA-MPJPE ↓	ACC ↓
VIBE	56.5	**27.1**
VIBE-HMR2	**55.3**	27.3

Πίνακας 5.8: Αποτελέσματα της μεθόδου VIBE με χρήση του μοντέλου HMR2.

Η μέθοδος μας θα μπορούσε να επωφεληθεί από μια τέτοια αντιμετώπιση. Ωστόσο, δεδομένου οτι βάζουμε τον περιορισμό το μοντέλο μας να είναι αιτιατό, δεν μπορεί να εφαρμοστεί ακριβώς στην περίπτωση μας. Για να παρακάμψουμε αυτό το πρόβλημα προτείνουμε 2 τροποποιήσεις:

- Μοντέλο AR VIBE + Moving Avg. Shape: Υπολογισμός κάθε φορά των παραμέτρων σχήματος με τον τύπο του κινούμενου μέσου όρου. Η επιλογή αυτή είναι μια προσέγγιση της αρχική ιδέας και δεν παραβιάζει την αιτιατότητα του προτεινόμενου μοντέλου. Πιο συγκεκριμένα, κάθε χρονική στιγμή εκτιμούμε οπως προηγουμένως τις παραμέτρους σχήματος και πριν τις δώσουμε σαν εισόδους στο μοντέλο SMPL υπολογίζουμε τον κινούμενο μέσο όρο.
- Μοντέλο AR VIB + Single Shape: Χρήση της εκτίμησης σχήματος της πρώτης χρονικής στιγμής για όλες τις εκτιμήσεις κατα την διάρκεια του χρονικού παραθύρου. Προκειμένου να βελτιώσουμε αυτή τη πρώτη εκτίμηση, ομοίως με παραπάνω αυξάνουμε τον αριθμό των επαναλήψεων εκτίμησης σε 6.

Σαν τρίτη επιλογή, για λόγους πληρότητας, δοκιμάζουμε να εκτιμούμε κάθε χρονική στιγμή από την αρχή το σχήμα, χωρίς ωστόσο να επιβάλουμε κάποιον επιπλέον περιορισμό, αναφερόμαστε σε αυτό ως AR VIBE + Shape. Τα αποτελέσματα φαίνονται στον Πίνακα 5.7.

Παρατηρούμε ότι υπάρχει ένα trade off μεταξύ του λάθους ανακατασκευής και της χρονικής συνοχής. Πιο συγκεκριμένα τα μοντέλα AR VIBE + Single Shape και AR VIBE + Moving Avg. Shape επιτυγχάνουν ακόμα μικρότερη επιτάχυνση ωστόσο ανεβάζουν το λάθος ανακατασκευτής συγκριτικά με το μοντέλο AR VIBE + More Inital Iters. Μια επιλογή που φαίνεται να πετυχαίνει εξίσου καλά αποτελέσματα είναι το μοντέλο με τον κινούμενο μέσο όρο.

5.3.3 Χρήση του μοντέλου HMR2

Τα αποτελέσματα φαίνονται στον Πίνακα 5.8 ως προς το σύνολο δεδομένων 3DPW και τις μετρικές PA-MPJPE και Acceleration. Παρατηρούμε οτι το μοντέλο βελτιώνεται ως προς την μετρική ανακατασκευής ενώ παράλληλα δεν μειώνεται η χρονική συνοχή των προβλέψεων του.

Οπως αναφέραμε στην προηγούμενη ενότητα, αυτο εικάζουμε ότι οφείλεται στην διάσπαση του συνολικού μοντέλου σε υποσυστήματα τα οποία μαθαίνουν να εστιάζουν σε διαφορετικά χαρακτηριστικά της εισόδου.

Παρακάτω δείχνουμε ποιοτικά παραδείγματα και σχολιάζουμε τις αδυναμίες αυτής της μεθόδου.

5.3.4 Κατηγοριοποίηση δράσης ως επιπλέον επίβλεψη

Αρχικά εφαρμόζουμε την αρχιτεκτονική που προτάθηκε στην υποενότητα 4.2.3. Ωστόσο παρατηρήσαμε πως από τις 14 αρθρώσεις αξιολόγησης, αυτές με το μεγαλύτερο λάθος ανακατασκευής είναι τα 4 άκρα και οι 4 αμέσως πιο εσωτερικές αρθρώσεις:

- Δεξιός Αστράγαλος
- Δεξί Γόνατο
- Αριστερό Γόνατο
- Αριστερός Αστράγαλος
- Δεξιός Καρπός
- Δεξιός Αγκώνας
- Αριστερός Αγκώνας
- Αριστερός Καρπός

Model	PA-MPJPE ↓
VIBE	56.5
VIBE w/ Action Clf full pose	56.0
VIBE w/ Action Clf partial pose	**55.7**

Πίνακας 5.9: Αποτελέσματα της μεθόδου VIBE με επιπλέον επίβλεψη από το πρόβλημα της κατηγοριοποίησης δράσης.

Joint	VIBE	VIBE w/ Action Clf
R ankle	83.4	**82.5**
R knee	**53.2**	53.3
L knee	**53.9**	54.9
L ankle	79.7	**79.2**
R wrist	91.8	**91.5**
R elbow	49.5	**47.1**
L elbow	53.9	**50.4**
L wrist	**92.9**	94.0

Πίνακας 5.10: Αποτελέσματα ανα άρθρωση της μεθόδου VIBE με επιπλέον επίβλεψη απο το πρόβλημα κατηγοριοποίησης δράσης.

Επομένως σε μια απόπειρα να "στοχεύσουμε" την επίβλεψη σε αυτές τις αρθρώσεις χρησιμοποιούμε τις περιστροφές που αντιστοιχούν μόνο σε αυτές. Τα συνολικά αποτελέσματα φαίνονται στον Πίνακα 5.9 για το σύνολο δεδομένων 3DPW ως προς την μετρική PA-MPJPE. Παρατηρούμε οτι και οι δύο επεκτάσεις βελτιώνουν την απόδοση του αρχικού μοντέλου ως προς το λάθος ανακατασκευής. Επιπλέον παρουσιάζουμε τα επιμέρους λάθη ανακατασκευής σε αυτές τις 8 πιο σημαντικές αρθρώσεις στον Πίνακα 5.10 Παρατηρούμε οτι στις 5 απο τις 8 αρθρώσεις το μοντέλο που χρησιμοποιεί μόνο την πόζα αυτών των 8 αρθρώσεων επιτυγχάνει καλύτερα αποτελέσματα. Ιδίως στους δύο αγκώνες το μοντέλο βελτιώνει την απόδοση κατα τουλαχιστον 2.5mm.

5.3.5 Ποιοτικά αποτελέσματα

Σε αυτή την υποενότητα παρουσιάζουμε αποτελέσματα ανακατασκευής από τα δυναμικά μοντέλα VIBE + Moving Avg. Shape, AR VIBE + Single Shape και VIBE-HMR2. Για λόγους σύγκρισης παρουσιάζουμε και τα αποτελέσματα από το μοντέλο της μεθόδου VIBE με το προεκπαιδευμένο μοντέλο που παρέχουν οι ερευνητές.

Αρχικά συγκρίνουμε τις αιτιατές υλοποιήσεις με το βασικό μοντέλο VIBE. Τα αποτελέσματα παρουσιάζονται στην Εικόνα 5.3.2.

Παρατηρούμε οτι οι διαφορές στην μετρική ανακατασκευής δεν είναι εύκολα φανερές καθώς ολα τα μοντέλα δίνουν ικανοποιητικά αποτελέσματα. Επιπλέον παρατηρούμε ότι το μοντέλο VIBE + Moving Avg. Shape έχει σταματήσει να μεταβάλλει δραστικά το σχήμα του υποκειμένου μεταξύ των frames. Απο τα ποιοτικά πειράματα που διεξήγαμε, καταλήξαμε πως απο τις υλοποιήσεις μας, καλύτερα δείγματα δίνουν τα μοντέλα VIBE + Moving Avg. Shape και AR VIBE + Single Shape.

Στην συνέχεια συγκρίνουμε το μοντέλο VIBE-HMR2 με το βασικό μοντέλο VIBE. Τα αποτελέσματα παρουσιάζονται στην εικόνα 5.3.3. Ομοίως με παραπάνω οι διαφορές στην μετρική ανακατασκευής δεν είναι εύκολα αντιληπτές. Ωστόσο αν εστιάσουμε στην άρθρωση του αριστερού καρπού φαίνεται πως το μοντέλο VIBE-HMR2 παράγει καλύτερη εκτίμηση.

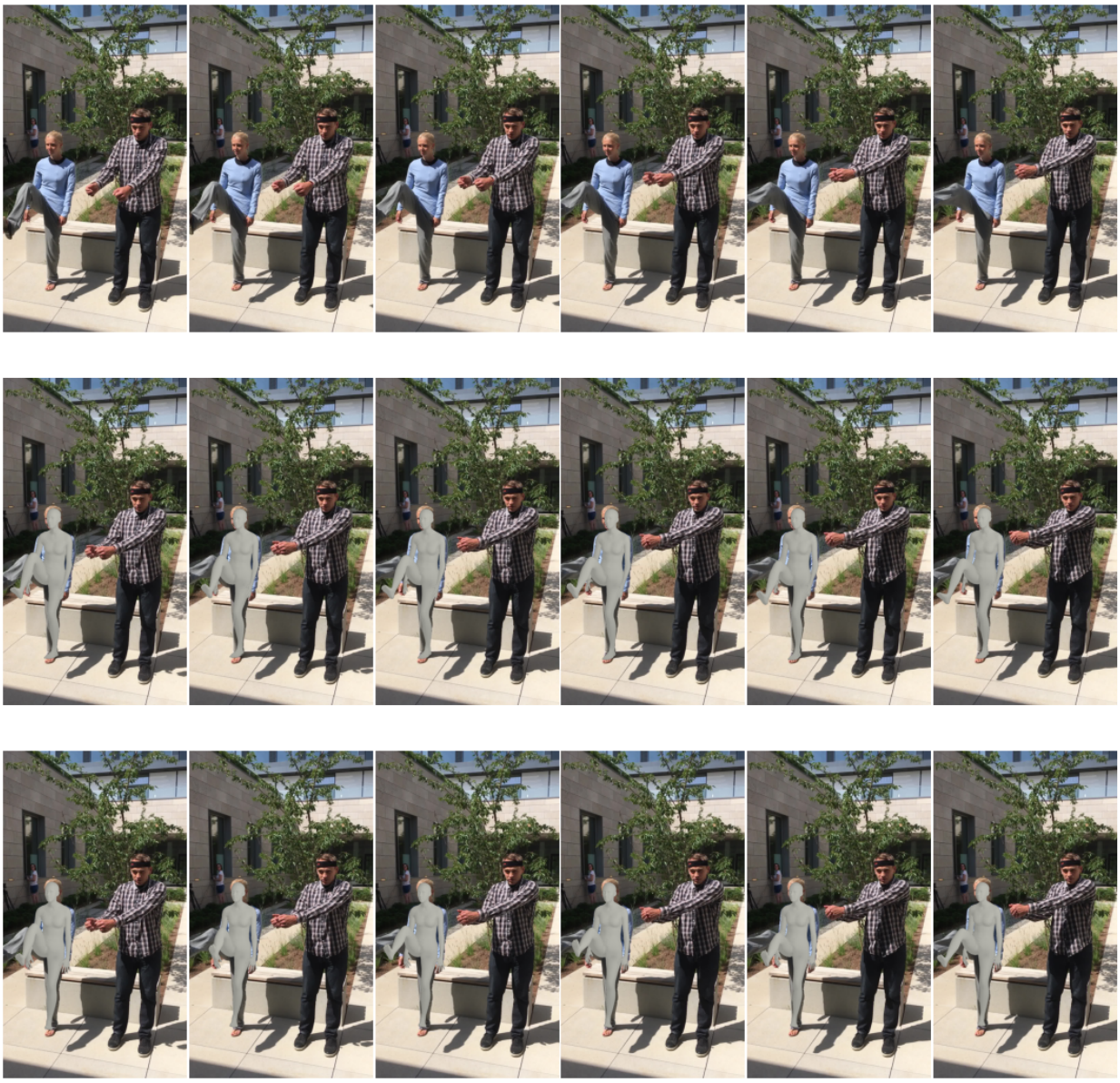
Τέλος παρουσιάζουμε ποιοτικά αποτελέσματα μεταξύ του προτεινόμενου αιτιατού μοντέλου AR VIBE + Moving Avg. Shape και του μοντέλου VIBE-HMR2 ώστε να συγκρίνουμε την χρονική συνοχή των εκτιμήσεων. Τα αποτελέσματα φαίνονται στην Εικόνα 5.3.4.

Πιο συγκεκριμένα έχουμε επιλέξει frames οπου το υποκείμενο κινεί το δεξί πόδι το οποίο μετα απο το πρώτο frame βρίσκεται εκτός του οπτικού πεδίου της κάμερας. Παρατηρούμε οτι το αιτιατό μοντέλο(2η γραμμή) μεταξύ των frames 2-3(στήλες 2 και 3) και 5-6(στήλες 5 και 6) παράγει πιο συνεκτικές εκτιμήσεις για την άρθρωση του γονάτου καθώς η θέση της δεν μεταβάλλεται απότομα μεταξύ των frames. Σε αντίθεση, είναι φανερή η

Εικόνα 5.3.2: Παραδείγματα ανακατασκευής απο τα μοντέλα VIBE(2η στήλη), AR VIBE + Single Shape(3η στήλη), AR VIBE + Moving Avg. Shape(4η στήλη).

Εικόνα 5.3.3: Παραδείγματα ανακατασκευής από τα μοντέλα VIBE(2η στήλη) και VIBE-HMR2(3η στήλη).

αντίστοιχη μετατόπιση στα αποτελέσματα του μοντέλου VIBE-HMR2. Παράλληλα είναι εμφανής η αδυναμία του αιτιατού μοντέλου να ανιχνεύσει επιτυχώς την άρθρωση του αστραγάλου. Αντιθέτως το μοντέλο VIBE-HMR2, σε μερικά frames(3 και 4) ανιχνεύει καλύτερα την άρθρωση.



Εικόνα 5.3.4: Παραδείγματα ανακατασκευής απο τα μοντέλα AR VIBE + Moving Avg. Shape(2η γραμμή) και VIBE-HMR2(3η γραμμή).

Κεφάλαιο 6

Επίλογος

6.1 Σύνοψη και Συμπεράσματα

Στην παρούσα εργασία μελετήσαμε το πρόβλημα της τρισδιάστατης ανακατασκευής πόζας και σχήματος τόσο από στατικές εικόνας όσο και από βίντεο. Και στα δύο προβλήματα ο τρόπος σκέψης ήταν να εκμεταλλευτούμε την φυσική εξάρτηση των δεδομένων πόζας και σχήματος αλλα και την σχέση των δεδομένων πόζας μεταξύ συνεχόμενων χρονικών στιγμών και να μοντελοποιήσουμε ρητά αυτή την εξάρτηση στην αρχιτεκτονική.

Στην περίπτωση των στατικών εικόνων προσπαθήσαμε να δούμε το πρόβλημα από την οπτική γωνία της εκτίμησης πόζας δεδομένου του σχήματος, γεγονός που μας οδήγησε να χωρίσουμε το συνολικό πρόβλημα σε 2 υποπροβλήματα:

- Εκτίμηση σχήματος δεδομένης της εικόνας.
- Εκτίμηση πόζας δεδομένης της εικόνας και του σχήματος.

Πραγματοποιήσαμε έναν αριθμό από πειράματα εκπαίδευσης όπου απομονώναμε τα υποσυστήματα ώστε να μελετήσουμε τον τρόπο που επηρεάζει καθένα από αυτά την απόδοση του μοντέλου. Τα τελικά μοντέλα επιτυγχάνουν απόδοση όμοια με αυτή του βασικού μοντέλου.

Όσον αφορά στο δυναμικό πρόβλημα της εκτίμησης τρισδιάστατης πόζας και σχήματος από βίντεο προσπαθήσαμε να μοντελοποιήσουμε καλύτερα την χρονική εξάρτηση μεταξύ δεδομένων πόζας τροποποιώντας το βασικό μοντέλο ώστε να εκτιμά ακολουθιακά την πόζα για κάθε frame. Το μοντέλο, με αυτό τον τρόπο, παρέμεινε αιτιατό ενώ παράλληλα παρήγαγε χρονικά πιο συνεκτικές εκτιμήσεις.

Στην συνέχεια τροποιήσαμε το βασικό μοντέλο VIBE ώστε να χρησιμοποιεί το στατικό μοντέλο HMR2 που προτείναμε παραπάνω. Η τροποποίηση αυτή βελτίωσε την μετρική ανακατασκευής κρατώντας τις υπόλοιπες μετρικές αμετάβλητες. Εικάζουμε ότι αυτή η βελτίωση οφείλεται στην διάσπαση του μοντέλου σε υποσυστήματα που λόγω της εργασίας που εκτελούν χρειάζεται να είναι αμετάβλητα σε διαφορετικά χαρακτηριστικά.

Τέλος προτείναμε μια εναλλακτική μορφή εκπαίδευσης κάνοντας χρήση της ιδέας μάθησης πολλαπλών εργασιών(Multitask Learning). Πιο συγκεκριμένα χρησιμοποιήσαμε το αποτέλεσμα της εκτίμησης πόζας προκειμένου να κατηγοριοποιήσουμε το βίντεο εισόδου σε μια από τις διαθέσιμες κατηγορίες. Η επιπλέον επίβλεψη που προέκυψε από την βοηθητική εργασία βοήθησε το μοντέλο να μειώσει την μετρική ανακατασκευής.

Συμπερασματικά, δείξαμε πειραματικά ότι απλές αλλαγές στην αρχιτεκτονική ώστε να μοντελοποιούνται ρητά εξαρτήσεις μεταξύ των δεδομένων μπορούν να ωφελήσουν το μοντέλο ώστε:

- να επιτυγχάνει μικρότερο λάθος ανακατασκευής.
- να παράγει πιο συνεκτικές εκτιμήσεις μεταξύ των χρονικών στιγμών.
- να είναι πιο κατανοητή η λειτουργία των υποσυστημάτων του.

6.2 Μελλοντικές Επεκτάσεις

Η παραπάνω εργασία δέχεται πολλές φυσικές επεκτάσεις. Παρακάτω παρουσιάζουμε τις πιο ενδιαφέρουσες από αυτές.

Όσον αφορά στο στατικό μοντέλο θα είχε νόημα να εξεταστεί η περεταίρω διάσπαση του υποσυστήματος πόζας σε διαφορετικά υποσυστήματα για κάθε μέρος του σώματος ή για ομάδες μερών. Μια παρόμοια ιδέα έχει εφαρμοστεί από τους ερευνητές [Geo+20] χωρίς ωστόσο να μοντελοποιούν ρητά την εξάρτηση των μερών από τις παραμέτρους σχήματος.

Όσον αφορά στο δυναμικό μοντέλο ενδιαφέρον παρουσιάζει η επέκταση του μοντέλου της ενότητας 5.3.2 χρησιμοποιώντας την μέθοδο SMPLify [Bog+16]. Πιο αναλυτικά αναφέραμε πως στην συγκεκριμένη μέθοδο, δεδομένου ότι οι εκτιμήσεις παράγονται ακολουθιακά, διαδραματίζει σημαντικό ρόλο η αρχική εκτίμηση. Δεδομένου ότι η μέθοδος SMPLify, οπου βελτιστοποιεί τις παραμέτρους του μοντέλου SMPL ώστε να ταιριάζει στην εικόνα, παράγει εκτιμήσεις μεγαλύτερης ακρίβειας [Bog+16; Kol+19], μπορεί να χρησιμοποιηθεί μόνο στο πρώτο frame ώστε να παράξει ενα καλό σημείο εκκίνησης και στην συνέχεια να χρησιμοποιηθεί η μέθοδος της ενότητας 5.3.2 κανονικά.

Δεδομένου ότι το αιτιατό μοντέλο AR VIBE παράγει καλύτερα αποτελέσματα ως προς την επιτάχυνση ενω διατηρεί την επίδοση ανακατασκευής του βασικού μοντέλου και δεδομένου ότι το μοντέλο VIBE-HMR2 παράγει καλύτερα αποτελέσματα ως προς την μετρική ανακατασκευής ενώ διατηρεί την ίδια χρονική συνοχή με το baseline θα ήταν λογικό να δοκιμαστεί ο συνδυασμός τους. Πιο συγκεκριμένα, μια άμεση επέκταση της παραπάνω εργασίας είναι η τροποποίηση του μοντέλου AR VIBE ώστε να χρησιμοποιεί το μοντέλο HMR2. Αρχικά πειράματα που εκτελέσαμε έδειξαν ότι η εκπαίδευση αυτού το μοντέλου είναι πιο ασταθής πιθανών λόγω της αποσύμπλεξης των δύο υποσυστημάτων. Ωστόσο αξίζει να διερευνηθεί περαιτέρω.

Τέλος, ενδιαφέρον παρουσιάζει η ιδέα του συνδυασμού της μεθόδου 5.3.2 και ProHMR [Kol+21]. Πιο συγκεκριμένα στα πειράματα υλοποίησης παρατηρήσαμε πως ο κρυφός χώρος του μοντέλου ProHMR κατέχει μια δομή: κοντινά σημεία του κρυφού χώρου μετασχηματίζονται σε όμοιες πόζες. Επομένως μια άμεση επέκταση θα ήταν η χρήση του Χρονικού Κωδικοποιητή(Temporal Encoder) του μοντέλου VIBE [KAB20] ώστε να προβλέπει ακολουθιακά σημεία στον κρυφό χώρο του ProHMR οπου με την χρήση των Normalizing Flows θα μετασχηματίζονται σε ακολουθίες πόζας.

Εν κατακλείδι, η παραπάνω εργασία προσφέρεται ως αρχικό σημείο προκειμένου να υλοποιηθούν πιο περίπλοκες τεχνικές που δανείζονται ιδέες από στατικά μοντέλα.

Βιβλιογραφία

[And+14] Andriluka, M. et al. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[ADZ19] Arnab, A., Doersch, C., and Zisserman, A. "Exploiting temporal context for 3D human pose estimation in the wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3395–3404.

[Big+20] Biggs, B. et al. "3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20496–20507.

[Bog+16] Bogo, F. et al. "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer. 2016, pp. 561–578.

[Cho+21] Choi, H. et al. "Beyond static features for temporally consistent 3d human pose and shape from a video". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 1964–1973.

[DSDB17] Dinh, L., Sohl-Dickstein, J., and Bengio, S. *Density estimation using Real NVP*. 2017. arXiv: 1605.08803 [cs.LG].

[Du+23] Du, Y. et al. "Flowpose: Conditional Normalizing Flows for 3D Human Pose and Shape Estimation from Monocular Videos". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097101.

[Geo+20] Georgakis, G. et al. "Hierarchical kinematic human mesh recovery". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer. 2020, pp. 768–784.

[Goo+14] Goodfellow, I. et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[He+15] He, K. et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385.

[Ion+14] Ionescu, C. et al. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339.

[JNV21] Joo, H., Neverova, N., and Vedaldi, A. "Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation". In: *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, pp. 42–52.

[Kan+18] Kanazawa, A. et al. "End-to-End Recovery of Human Shape and Pose". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7122–7131. DOI: 10.1109/CVPR.2018.00744.

[Kan+19] Kanazawa, A. et al. "Learning 3d human dynamics from video". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5614–5623.

[KW13] Kingma, D. P. and Welling, M. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[KD18] Kingma, D. P. and Dhariwal, P. "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in neural information processing systems* 31 (2018).

[KAB20] Kocabas, M., Athanasiou, N., and Black, M. J. "Vibe: Video inference for human body pose and shape estimation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5253–5263.

[Koc+21a] Kocabas, M. et al. "PARE: Part attention regressor for 3D human body estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11127–11137.

[Koc+21b] Kocabas, M. et al. "SPEC: Seeing people in the wild with an estimated camera". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11035–11045.

[Kol+19] Kolotouros, N. et al. "Learning to reconstruct 3D human pose and shape via model-fitting in the loop". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2252–2261.

[Kol+21] Kolotouros, N. et al. "Probabilistic modeling for human mesh recovery". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11605–11614.

[KSH12] Krizhevsky, A., Sutskever, I., and Hinton, G. E. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012.

[Las+17] Lassner, C. et al. "Unite the people: Closing the loop between 3d and 2d human representations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6050–6059.

[LeD17] LeDell, E. *Statistical Learning Data Mining IV*. 2017.

[Li+21] Li, J. et al. "HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 3382–3392. DOI: 10.1109/CVPR46437.2021.00339.

[Li+23] Li, J. et al. "NIKI: Neural Inverse Kinematics With Invertible Neural Networks for 3D Human Pose and Shape Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 12933–12942.

[Li+22] Li, Z. et al. "Cliff: Carrying location information in full frames into human pose and shape estimation". In: *European Conference on Computer Vision*. Springer. 2022, pp. 590–606.

[Lin+14] Lin, T. et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312.

[Lop+15] Loper, M. et al. "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Trans. Graph.* 34.6 (2015). ISSN: 0730-0301. DOI: 10.1145/2816795.2818013.

[Mah+19] Mahmood, N. et al. "AMASS: Archive of Motion Capture as Surface Shapes". In: *International Conference on Computer Vision*. Oct. 2019, pp. 5442–5451.

[Mar+18] Marcard, T. von et al. "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera". In: *European Conference on Computer Vision (ECCV)*. 2018.

[Meh+17] Mehta, D. et al. "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision". In: *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE. 2017. DOI: 10.1109/3dv.2017.00064.

[Pan+22] Pang, H. E. et al. "Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26034–26051.

[PKD19] Pavlakos, G., Kolotouros, N., and Daniilidis, K. "Texturepose: Supervising human mesh estimation with texture consistency". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 803–812.

[Pav+18] Pavlakos, G. et al. "Learning to estimate 3D human pose and shape from a single color image". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 459–468.

[Pav+19] Pavlakos, G. et al. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10975–10985.

[Raj+22] Rajasegaran, J. et al. "Tracking people by predicting 3D appearance, location and pose". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 2740–2749.

[RMW14] Rezende, D. J., Mohamed, S., and Wierstra, D. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on*

Machine Learning. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, 2014, pp. 1278–1286.

[RTB17] Romero, J., Tzionas, D., and Black, M. J. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 36.6 (Nov. 2017).

[She+23] Shetty, K. et al. "PLIKS: A Pseudo-Linear Inverse Kinematic Solver for 3D Human Body Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 574–584.

[Tri+23] Tripathi, S. et al. "3D human pose estimation via intuitive physics". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 4713–4725.

[Vas+23] Vasilikopoulos, N. et al. "TAPE: Temporal Attention-Based Probabilistic Human Pose and Shape Estimation". In: *Image Analysis*. Springer Nature Switzerland, 2023, pp. 418–431. DOI: 10.1007/978-3-031-31438-4_28.

[Vas+17] Vaswani, A. et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[Wen21] Weng, L. "What are diffusion models?" In: *lilianweng.github.io* (2021).

[Ye+23] Ye, V. et al. "Decoupling human and camera motion from videos in the wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 21222–21232.

[Yua+22] Yuan, Y. et al. "GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11038–11049.

[Zha+23] Zhang, A. et al. *Dive into Deep Learning*. Cambridge University Press, 2023.

[Zho+19] Zhou, Y. et al. "On the continuity of rotation representations in neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5745–5753.