
NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group


# Generalizing 3D Human Shape and Pose Estimation for Diverse Non-Adult Populations

DIPLOMA THESIS

of


**Georgios V. Chatzichristodoulou**

**Supervisor**: Petros Maragos
Professor Emeritus, NTUA

**Co-Supervisors**: Georgios Pavlakos
Assistant Professor, UT Austin

Niki Efthymiou
Postdoctoral Researcher, NTUA



Athens, October 2025

**NATIONAL TECHNICAL UNIVERSITY OF ATHENS**
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group

# Generalizing 3D Human Shape and Pose Estimation for Diverse Non-Adult Populations

DIPLOMA THESIS

of

**Georgios V. Chatzichristodoulou**

**Supervisor**: Petros Maragos
Professor Emeritus, NTUA

**Co-Supervisors**: Georgios Pavlakos
Assistant Professor, UT Austin

Niki Efthymiou
Postdoctoral Researcher, NTUA

Approved by the examining committee on 17 October 2025.

| ..... | ..... | ..... |
|---|---|---|
| Petros Maragos | Athanasios Rontogiannis | Ioannis Kordonis |
| Professor Emeritus, NTUA | Associate Professor, NTUA | Assistant Professor, NTUA |

Athens, October 2025

.....

**Γεώργιος Β. Χατζηχριστοδούλου**
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

# Abstract

The accurate three-dimensional (3D) shape and pose estimation of humans from a single image constitutes a fundamental and complex problem in Computer Vision, with critical applications spanning health, biomechanics, Virtual Reality (VR), and animation. Despite significant advancements for the adult population, the majority of current methods fail to generalize effectively to children and infants due to their unique anthropometric proportions and the scarcity of specialized datasets required for model training. This diploma thesis addresses this challenge by introducing a comprehensive framework designed to bridge this domain gap. We propose an optimization-based method that extends a top-performing model by incorporating the SMPL-A body model, enabling the concurrent and accurate modeling of adults, children, and infants. Leveraging this approach, we generated pseudo-ground-truth annotations for publicly available databases of child and infant images. Utilizing this new training data, we then developed and trained a specialized transformer-based Deep Learning model capable of real-time 3D human reconstruction. Furthermore, we introduce the BabyRobot dataset, which contains the 3D reconstructions produced by our method from videos of children interacting with robots with many actions, gestures and movements in the environment. Our methods contribute to the anonymization of sensitive data, like that of children and infants, since the 3D reconstructions provide information about the body and the motion of humans, but not their identity. Our results demonstrate a substantial improvement in the quality of shape and pose estimation for child and infant images, while simultaneously maintaining high performance across the adult population.

**Keywords:** 3D Computer Vision, Human Mesh Recovery, 3D Shape and Pose Estimation, SMPL-A, Vision Transformer, Pediatric Population

# Περίληψη

Η ακριβής τρισδιάστατη (3Δ) εκτίμηση του σχήματος και της πόζας των ανθρώπων από μία μόνο εικόνα αποτελεί ένα θεμελιώδες πρόβλημα στην Όραση Υπολογιστών με ευρείες εφαρμογές σε τομείς όπως η εικονική πραγματικότητα και το animation, αλλά και στο χώρο της υγείας και της εμβιομηχανικής. Ενώ έχει σημειωθεί σημαντική πρόοδος για τον ενήλικο πληθυσμό, η πλειοψηφία των υφιστάμενων μεθόδων αποτυγχάνει να γενικεύσει με ακρίβεια σε παιδιά και βρέφη λόγω των ιδιαίτερων ανθρωπομετρικών τους αναλογιών και της έλλειψης μεγάλων, εξειδικευμένων συνόλων δεδομένων για την εκπαίδευση σχετικών μοντέλων. Η παρούσα διπλωματική εργασία αντιμετωπίζει αυτήν την πρόκληση εισάγοντας ένα ολοκληρωμένο πλαίσιο που γεφυρώνει αυτό το χάσμα. Προτείνουμε μια μέθοδο βελτιστοποίησης που επεκτείνει ένα σύγχρονο επιτυχημένο μοντέλο, υιοθετώντας το ενιαίο SMPL-A μοντέλο ανθρώπινου σώματος για την ταυτόχρονη μοντελοποίηση ενηλίκων, παιδιών και βρεφών. Χρησιμοποιώντας αυτή τη μέθοδο, δημιουργήσαμε ψευδο-επισημειώσεις (pseudo-annotations) για δημόσιες βάσεις δεδομένων παιδικών εικόνων. Με αυτό το νέο σύνολο εκπαίδευσης, αναπτύξαμε και εκπαιδεύσαμε ένα μοντέλο Βαθιάς Μάθησης ικανό για 3Δ ανακατασκευή ανθρώπων σε πραγματικό χρόνο. Οι μέθοδοι μας μπορούν να βοηθήσουν στην ανωνυμοποίηση ευαίσθητων πληροφοριών, όπως αυτές των παιδιών και των βρεφών, αφού θα παρέχεται η πληροφορία για το σώμα και την κίνηση του ανθρώπου, αλλά όχι της ταυτότητάς του. Σε αυτό το πλαίσιο, παρουσιάζουμε το σύνολο δεδομένων BabyRobot, το οποίο περιέχει τις 3Δ ανακατασκευές από βίντεο παιδιών που αλληλεπιδρούν με ρομπότ με πλούσιες δράσεις, χειρονομίες και κίνηση στο χώρο. Τα αποτελέσματά μας καταδεικνύουν ουσιαστική βελτίωση στην ποιότητα εκτίμησης σχήματος και πόζας σε παιδικές και βρεφικές εικόνες, διατηρώντας παράλληλα την υψηλή απόδοση στον ενήλικο πληθυσμό.

**Λέξεις-Κλειδιά:** 3Δ Όραση Υπολογιστών, Human Mesh Recovery, Εκτίμηση 3Δ Σχήματος και Πόζας, SMPL-A, Vision Transformer, Παιδιατρικός Πληθυσμός

# Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνονται οι σπουδές μου στη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου και θα ήθελα να ευχαριστήσω όσους συνέβαλαν σε αυτό. Στον επιβλέποντα Καθηγητή Πέτρο Μαραγκό που με εισήγαγε στον κόσμο της Όρασης Υπολογιστών μέσα από τα μαθήματά του και μου έδωσε την ευκαιρία να εκπονήσω την διπλωματική μου εργασία υπό την επίβλεψή του. Στους συνεπιβλέποντές μου, τον Επίκουρο Καθηγητή Γιώργο Παυλάκο για την καθοδήγηση και τις ιδέες του σε αυτό το ιδιαίτερο πεδίο έρευνας και τη Δρ. Νίκη Ευθυμίου για τα σχόλια και την πολύτιμη βοήθειά της κατά τη διάρκεια της διπλωματικής. Η συστηματική υποστήριξή τους υπήρξε καθοριστική για την ομαλή και έγκαιρη ολοκλήρωση της εργασίας. Επίσης, θα ήθελα να ευχαριστήσω όλους όσους συμπλήρωσαν το ερωτηματολόγιο που διενεργήθηκε στα πλαίσια της εργασίας αυτής. Ένα μεγάλο ευχαριστώ οφείλω και στους φίλους μου, για τις στιγμές χαλάρωσης και τις συζητήσεις που λειτούργησαν ως πολύτιμη πηγή ισορροπίας και δύναμης κατά τη διάρκεια αυτού του ταξιδιού. Ιδιαίτερη μνεία αξίζει και στο πιάνο μου, το οποίο προσέφερε την απαραίτητη δημιουργική διέξοδο. Τέλος, ευχαριστώ θερμά τους γονείς και την αδερφή μου για την ξεχωριστή βοήθεια τους καθ'όλη τη διάρκεια των σπουδών μου, όπως και τους παππούδες και τις γιαγιάδες μου για τις γνώσεις και την ενθάρρυνσή τους.

# Contents

# List of Figures

# List of Tables

# Κεφάλαιο 1

# Εκτεταμένη Περίληψη στα Ελληνικά

## 1.1 Εισαγωγή

Η τρισδιάστατη (3Δ) εκτίμηση του σχήματος και της πόζας (3D shape and pose estimation) των ανθρώπων αποτελεί ένα θεμελιώδες πρόβλημα του πεδίου της Όρασης Υπολογιστών. Σκοπός είναι η ανακατασκευή μίας 3Δ αναπαράστασης του ανθρώπινου σώματος από τα δεδομένα εισόδου, τα οποία μπορεί να είναι μία δισδιάστατη (2Δ) εικόνα, ένα βίντεο ή και συνδυασμός εικόνων από διαφορετικές όψεις. Η φύση του προβλήματος είναι εγγενώς δύσκολη καθώς κατά την προβολή μίας πραγματικής σκηνής του 3Δ χώρου στο 2Δ επίπεδο της εικόνας χάνεται πληροφορία για το βάθος της σκηνής. Αυτές οι πληροφορίες είναι απαραίτητες, καθώς κατά την εκτίμηση του 3Δ σχήματος και πόζας από μία 2Δ εικόνα γίνεται η αντίστροφη διαδικασία, δηλαδή από το 2Δ επίπεδο πηγαίνουμε στο 3Δ. Το πρόβλημα γίνεται ακόμα δυσκολότερο όταν στην εικόνα υπάρχουν περισσότερα άτομα που αλληλεπιδρούν ή αντικείμενα που κρύβουν μέρη του σώματος των ανθρώπων.

Αρκετές μέθοδοι έχουν προταθεί στη βιβλιογραφία για την αντιμετώπιση των δυσκολιών που περιγράφηκαν παραπάνω. Ειδικότερα, με τον ραγδαία αναπτυσσόμενο τομέα της Βαθιάς Μάθησης και τη χρήση δεδομένων μεγάλης κλίμακας συνεχώς προτείνονται όλο και πιο αποτελεσματικές μέθοδοι. Αρχιτεκτονικές όπως τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks) (CNN) ή πιο πρόσφατα οι Transformers με τον μηχανισμό της Προσοχής (Attention Mechanism) [66] έχουν φέρει την επανάσταση στον χώρο της Όρασης Υπολογιστών. Αυτές οι τεχνολογίες μπορούν να μάθουν τις χωρικές σχέσεις και το περιεχόμενο μιας εικόνας, διευκολύνοντας τη δημιουργία αλγορίθμων που λύνουν προβλήματα στα οποία η κλασική Όραση Υπολογιστών δεν

είχε ανάλογη επιτυχία.

Οι κύριες κατηγορίες μεθόδων για την εκτίμηση του 3Δ σχήματος και της πόζας των ανθρώπων είναι:

- **Μέθοδοι βασισμένες στη Βελτιστοποίηση**: Οι πιο αποτελεσματικές μέθοδοι, οι οποίες βασίζονται κατά κανόνα στην ελαχιστοποίηση μίας αντικειμενικής συνάρτησης. Συνήθως, προσπαθούν να ελαχιστοποιήσουν την απόσταση μεταξύ της προβολής του 3Δ πλέγματος (3D mesh) που δημιουργείται για τον άνθρωπο στο επίπεδο της εικόνας και κάποιων σημείων ενδιαφέροντος (keypoints), κυρίως συνδέσμων του σώματος, που έχουν εκτιμηθεί προηγουμένως από κάποιο άλλο μοντέλο. Παρά την αποτελεσματικότητά τους, αυτές οι μέθοδοι είναι αρκετά χρονοβόρες και υπολογιστικά ακριβές, αφού χρειάζεται σημαντικός χρόνος μέχρι να λυθεί το πρόβλημα βελτιστοποίησης.

- **Μέθοδοι βασισμένες στην Παλινδρόμηση (Regression) - Μέθοδοι Βαθιάς Μάθησης**: Οι πιο σύγχρονες τεχνικές, όπου βασίζονται στην εκπαίδευση ενός προβλέπτη (νευρωνικού δικτύου), ο οποίος έχει μάθει να εκτιμάει αμέσως τις παραμέτρους ενός μοντέλου που περιγράφει το ανθρώπινο σώμα, μόλις του δοθεί η είσοδος. Για την πρόβλεψη το μοντέλο χρησιμοποιεί, συνήθως, ένα νευρωνικό δίκτυο το οποίο εκπαιδεύεται σε αρκετά μεγάλο πλήθος δεδομένων. Συνεπώς, αν και πολύ πιο γρήγορες αυτές οι μέθοδοι, εξαρτώνται από την ποιότητα και το πλήθος των δεδομένων εκπαίδευσης.

### 1.1.1 Προκλήσεις κατά την εκτίμηση 3Δ Σχήματος και Πόζας

Οι περισσότερες από τις μεθόδους που έχουν αναπτυχθεί συμβάλλουν στην σωστή εκτίμηση του 3Δ σχήματος και της πόζας των ανθρώπων από μία εικόνα. Η πλειοψηφία τους έχει σχεδιαστεί ώστε να λειτουργεί για ενήλικο πληθυσμό. Όταν οι μέθοδοι αυτές, ωστόσο, εφαρμοστούν σε παιδιά ή βρέφη, η ποιότητα των αποτελεσμάτων υστερεί σημαντικά από αυτή των ενηλίκων. Για την αποτυχία των τεχνικών αυτών ευθύνονται τόσο οι γενικότερες προκλήσεις κατά την εκτίμηση του 3Δ σχήματος και της πόζας ενός ανθρώπου, όσο και συγκεκριμένα προβλήματα που αφορούν τόσο τα διαθέσιμα δεδομένα όσο και ηθικά και νομικά ζητήματα γύρω από τα δεδομένα παιδιών και βρεφών.

Πιο συγκεκριμένα, ένα βασικό ζήτημα κατά τη μελέτη του σχήματος και της πόζας ενός ανθρώπου είναι η αβεβαιότητα που υπάρχει κατά την εκτίμηση τους, κυρίως, όταν υπάρχουν μέρη του σώματος που αποκρύπτονται από φυσικά εμπόδια, τον ίδιο ή άλλο άνθρωπο, ή δεν είναι στο πεδίο της εικόνας. Όταν

κάποιο μέρος του σώματος δεν είναι ορατό, πολλές διαφορετικές εκδοχές σχήματος και πόζας μπορούν να θεωρηθούν σωστές, δεδομένης της ασάφειας που υπάρχει.

Η δυσκολία στη γενίκευση σε παιδιά και βρέφη των μοντέλων που έχουν αναπτυχθεί για ενήλικες κυρίως, δεν οφείλεται μόνο στην αδυναμία του υπολογιστή να εφαρμόσει σωστά τη μέθοδο ή το μοντέλο στις εικόνες, αλλά και στην φυσιολογία του ανθρώπου. Το ανθρώπινο σώμα από την βρεφική ηλικία μέχρι την ενηλικίωση και κατά τη διάρκεια της ζωής μας αλλάζει σημαντικά. Αυτό σημαίνει ότι αλλάζουν και οι αναλογίες των διαφόρων μερών του σώματος. Για παράδειγμα, το κεφάλι ενός βρέφους είναι δυσανάλογα μεγάλο σε σχέση με το υπόλοιπο σώμα του, ενώ τα χέρια του είναι αρκετά μικρά σε σχέση με τον κορμό του. Αντίθετα, αυτές οι αναλογίες αλλάζουν στους ενήλικες ή στους εφήβους. Συνεπώς, αυτές οι διαφορές στις αναλογίες και στο σώμα, δυσκολεύουν την δημιουργία ενός καθολικού μοντέλου για την περιγραφή του ανθρώπινου σώματος για άτομα κάθε ηλικίας.

Ένα ακόμα σημαντικό ζήτημα σχετικά με τα παιδιά είναι η ευαίσθητη φύση των δεδομένων τους. Αρκετά ηθικά ζητήματα αλλά και οι νομοθεσίες των εκάστοτε κρατών περιορίζουν την δημιουργία μεγάλων συνόλων δεδομένων με φωτογραφίες παιδιών τα οποία είναι απαραίτητα για την εκπαίδευση μοντέλων εξειδικευμένα στην εκτίμηση πόζας και σχήματος παιδιών.

### 1.1.2 Εφαρμογές

Η μελέτη του 3Δ σχήματος και της πόζας ενός ανθρώπου απαριθμεί πλήθος εφαρμογών σε διαφορετικά πεδία και επιστημονικές κοινότητες. Στον χώρο της υγείας, αποτελεί έναν μη επεμβατικό τρόπο για τη μελέτη της κίνησης και της ανάπτυξης ενός ανθρώπου. Από τη μελέτη της ανάπτυξης ενός βρέφους και τον εντοπισμό νευρολογικών ή κινησιολογικών ασθενειών, έως την μελέτη της φυσικής κατάστασης ενός αθλητή, ο τομέας αυτός της Όρασης Υπολογιστών προσφέρει πολυάριθμες λύσεις. Με τη δημιουργία 3Δ ανακατασκευών των ανθρώπων αποκρύπτονται οι πληροφορίες για την ταυτότητα των ανθρώπων, χωρίς να χάνεται η πληροφορία για το σώμα, τις κινήσεις και τις αλληλεπιδράσεις τους. Με αυτόν τον τρόπο μπορούν να δημιουργηθούν μεγάλες ανοιχτές βάσεις δεδομένων με πλήρως ανώνυμα δεδομένα. Επίσης, μια σημαντική εφαρμογή της μελέτης του 3Δ σχήματος και πόζας είναι ότι μπορεί να αποτελέσει ένα σημαντικό δείκτη για την πρώιμη διάγνωση του αυτισμού σε παιδιά. Τέλος, μέσω των μοντέλων περιγραφής του ανθρώπινου σώματος, η εκτίμηση της 3Δ πόζας και σχήματος βρίσκει εφαρμογές στην εικονική και στην επαυξημένη πραγματικότητα, μέσω, για παράδειγμα, των ρεαλιστικών άβαταρς, όπως και στο πεδίο της παρακολούθησης ανθρώπων (tracking) σε κινούμενο βίντεο και στην αναγνώριση δράσεων.

### 1.1.3 Συνεισφορές

Με αφορμή το γεγονός ότι οι περισσότερες μέθοδοι για την εκτίμηση του 3Δ σχήματος και πόζας των ανθρώπων εστιάζουν σε ενήλικο πληθυσμό, στόχος της διπλωματικής εργασίας είναι να βελτιώσει την ποιότητα των αποτελεσμάτων στην 3Δ μοντελοποίηση παιδιών και βρεφών. Γι'αυτό το λόγο οι κύριες συνεισφορές μας συνοψίζονται παρακάτω:

- Προτείνουμε μία μέθοδο βελτιστοποίησης, επέκταση μίας σύγχρονης μεθόδου, η οποία χρησιμοποιεί το ενιαίο μοντέλο SMPL-A και είναι ικανή να μοντελοποιήσει τόσο ενήλικες όσο και μωρά και παιδιά κάθε ηλικίας.

- Χρησιμοποιούμε τη μέθοδο βελτιστοποίησης για να παράξουμε επισημειώσεις για εικόνες παιδιών και βρεφών από δημόσιες βάσεις δεδομένων ώστε να δημιουργήσουμε ένα σύνολο δεδομένων που μπορεί να χρησιμοποιηθεί για την εκπαίδευση μοντέλων βαθιάς μάθησης. Με αυτά τα δεδομένα εκπαιδεύουμε ένα τέτοιο μοντέλο το οποίο είναι ικανό από μία εικόνα σε πραγματικό χρόνο να κάνει 3Δ ανακατασκευή των ανθρώπων που βρίσκονται σε αυτή.

- Οι μέθοδοι που προτείνουμε μπορούν να χρησιμοποιηθούν για την ανωνυμοποίηση και απόκρυψη ευαίσθητων δεδομένων μέσω της δημιουργίας 3Δ ανακατασκευών ανθρώπων κάθε ηλικίας. Οι 3Δ ανακατασκευές του ανθρώπινου σώματος μπορούν να χρησιμοποιηθούν στη δημιουργία datasets σε δεδομένα που σε διαφορετική περίπτωση θα ήταν αδύνατο να δημοσιευτούν λόγω της ευαισθησίας τους. Σε αυτό το πλαίσιο, παρουσιάζουμε και το σύνολο δεδομένων BabyRobot με 3Δ ανακατασκευές από παιδιά που αλληλεπιδρούν με ρομπότ με πλούσιες δράσεις, χειρονομίες και κινήσεις στο χώρο.

## 1.2 Θεωρητικό Υπόβαθρο

Παρακάτω παρουσιάζονται συνοπτικά κάποια θεμελιώδη στοιχεία θεωρητικού υποβάθρου για την κατανόηση της μεθοδολογίας που αναπτύχθηκε. Ένα πιο εκτεταμένο θεωρητικό υπόβαθρο υπάρχει στο αγγλικό κείμενο στο Κεφάλαιο 3.

### 1.2.1 Το Skinned Multi-Person Linear (SMPL) Μοντέλο

Τις περισσότερες φορές η δημιουργία 3Δ ανακατασκευών για το ανθρώπινο σώμα απαιτεί ένα μοντέλο που να το περιγράφει. Σε αυτό το πλαίσιο έχουν

αναπτυχθεί διαφορετικά μοντέλα. Το Skinned Multi-Person Linear (SMPL) [36] είναι ένα από τα πιο ρεαλιστικά και δυνατά μοντέλα που μπορούν να χρησιμοποιηθούν για αυτόν το σκοπό. Έχει εκπαιδευτεί σε ένα σύνολο δεδομένων με ποικιλία στο σχήμα και στην πόζα των ανθρώπων, ώστε να μπορεί να γενικεύσει όσο το δυνατόν καλύτερα.

Το SMPL μοντελοποιεί το σχήμα των ανθρώπων ως ένα συνδυασμό ενός σχήματος που εξαρτάται από την ταυτότητα του ανθρώπου και ενός σχήματος που εξαρτάται από την εκάστοτε πόζα που έχει. Χρησιμοποιώντας vertex-based skinning με corrective blend shapes παράγει το τελικό 3Δ πλέγμα (mesh) του ανθρώπου.

Η μοντελοποίηση ξεκινάει με ένα mesh template, το μέσο ανθρώπινο σώμα, που αποτελείται από $N = 6890$ κορυφές και $K = 23$ αρθρώσεις. Οι βασικές παράμετροι που χρησιμοποιεί το SMPL για την «παραμόρφωση» αυτού του mesh template είναι οι παράμετροι σχήματος $\boldsymbol{\beta} \in \mathbb{R}^n$ (συνήθως $n = 10$) και οι παράμετροι πόζας $\boldsymbol{\theta}$. Οι παράμετροι $\boldsymbol{\theta}$ είναι συνολικά $3 \cdot K + 3 = 72$, καθώς κάθε σύνδεσμος χρειάζεται 3 παραμέτρους για να περιγραφεί η περιστροφή του, και επιπλέον 3 παράμετροι συνολικά για τον κεντρικό προσανατολισμό του σώματος.

Τελικά, το SMPL είναι ένα μοντέλο $M(\boldsymbol{\beta}, \boldsymbol{\theta}; \Phi) : \mathbb{R}^{|\boldsymbol{\theta}| \times |\boldsymbol{\beta}|} \to \mathbb{R}^{3N}$ το οποίο κάνει μία απεικόνιση από τις παραμέτρους σχήματος και πόζας σε κορυφές ενός 3Δ mesh. Με $\Phi$ συμβολίζουμε το πλήρες σύνολο παραμέτρων του SMPL, δηλαδή του template mesh, των blend weights, του πίνακα κυρίων συνιστωσών που χρησιμοποιείται για τα blend shapes του σχήματος, του πίνακα που χρησιμοποιείται για να βρεθεί η θέση των joints και, τέλος, του πίνακα που περιέχει το σύνολο των blend shapes για την πόζα.

### 1.2.2 Το Skinned Multi-Infant Linear (SMIL) Μοντέλο

Το SMPL είναι ένα πολύ δυνατό και περιγραφικό μοντέλο για το ανθρώπινο σώμα. Ωστόσο, το γεγονός ότι έχει εκπαιδευτεί σε 3Δ mesh μόνο ενηλίκων το αποτρέπει από την επιτυχή μοντελοποίηση βρεφών. Για το λόγο αυτό, στηριζόμενο στο SMPL, αναπτύχθηκε το Skinned Multi-Infant Linear (SMIL) [19] με στόχο την καλύτερη μοντελοποίηση των βρεφών. Το SMIL είναι ένα μοντέλο που έχει εκπαιδευτεί σε χαμηλής ποιότητας RGB-D δεδομένα βρεφών που κινούνται ελεύθερα στο χώρο ώστε να μπορεί να εφαρμοστεί στη συνέχεια σε πραγματικές συνθήκες. Είναι ένα μοντέλο ικανό να παράξει ένα ρεαλιστικό σώμα βρεφών με τις σωστές αναλογίες λόγω του διαφορετικού template mesh, ειδικά σχεδιασμένο για βρέφη, που χρησιμοποιεί όπως και χάρις στην στοχευμένη σε βρεφικές αναλογίες βελτιστοποίηση των παραμέτρων σχήματος και

πόζας.

### 1.2.3 SMPL-A

Με την ανάπτυξη του SMPL και του SMIL δημιουργήθηκαν δύο μοντέλα που μπορούν να περιγράψουν πολύ αποτελεσματικά ενήλικες και βρέφη, αντίστοιχα. Ωστόσο, η αδυναμία του SMPL να μοντελοποιήσει παιδιά και βρέφη και, αντίστοιχα, του SMIL ενήλικες δημιουργεί ένα κενό στην καθολική μοντελοποίηση όλων των ηλικιών. Το SMPL-A [47] είναι ένα μοντέλο που προσπαθεί να καλύψει αυτό το κενό. Χρησιμοποιεί κατά βάση τη σχεδίαση του SMPL με τη διαφορά να έγκειται στο template για το σώμα που χρησιμοποιεί. Συγκεκριμένα, το SMPL-A παρεμβάλει ένα SMPL template σώματος ενήλικα $T_A$ και ένα SMIL template ενός παιδιού $T_C$ για να προκύψει το τελικό template $T_F$ σύμφωνα με την παρακάτω σχέση:

$$T_F = \alpha T_C + (1 - \alpha) T_A$$

όπου $\alpha$ είναι μία νέα παράμετρος που εισάγεται, το βάρος παρεμβολής (interpolation weight). Η παράμετρος αυτή παίρνει τιμές στο διάστημα $\alpha \in [0, 1]$, όπου $\alpha = 0$ όταν έχουμε template ενήλικα, και $\alpha = 1$ template παιδιού. Το shape space για το SMPL-A παραμένει το ίδιο με του SMPL, ωστόσο εφαρμογές έχουν δείξει ότι ακόμα και με αυτή την αλλαγή στο body template, η κοινή μοντελοποίηση ενηλίκων και παιδιών είναι δυνατή.

### 1.2.4 Vision Transformer

Ο Vision Transformer (ViT) [12] είναι μία αρχιτεκτονική βαθιάς μάθησης για επεξεργασία εικόνων που βασίζεται στον Transformer [66], ο οποίος αρχικά είχε αναπτυχθεί για εφαρμογές φυσικής επεξεργασίας γλώσσας (Natural Language Processing) (NLP). Σε αντίθεση με προηγούμενες αρχιτεκτονικές, όπως τα CNNs, ο ViT χρησιμοποιεί τον self-attention μηχανισμό [66], ο οποίος του επιτρέπει να μαθαίνει εξαρτήσεις από όλη την εικόνα, και όχι τοπικά χαρακτηριστικά.

Ο ViT, αντί να επεξεργάζεται μια εικόνα ως ένα σύνολο από pixels, τη χωρίζει σε μικρά μη επικαλυπτόμενα τμήματα (patches) και τα αντιμετωπίζει σαν να ήταν λέξεις σε μία πρόταση, όπως δηλαδή ο Transformer που είχε αρχικά αναπτυχθεί για εφαρμογές NLP. Κάθε ένα από αυτά τα τμήματα προβάλλεται γραμμικά σε ένα διάνυσμα, το embedding. Τα patch embeddings ενώνονται με embeddings που δίνουν πληροφορίες για τη θέση του patch στην αρχική εικόνα και το συνολικό embedding περνάει από έναν κλασικό Transformer encoder με multi-head self-attention επίπεδα, layer normalization και υπολειμματικές συνδέσεις (residual connections).

Ένας ViT χρησιμοποιεί δύο είδη μηχανισμού attention:

- **Self-attention**: Υπολογίζει τη σχέση μεταξύ διαφορετικών patches της εικόνας, ωθώντας το μοντέλο να εστιάσει στις πιο σημαντικές περιοχές της εικόνας, ανεξαρτήτως της χωρικής απόστασης τους.
- **Cross-attention**: Σε πολυτροπικά συστήματα, όταν χειρίζονται περισσότερα από ένα modalities, για παράδειγμα το κείμενο, ο ViT χρησιμοποιεί cross-attention για να διαχειριστεί ταυτόχρονα τα διαφορετικά είδη πληροφορίας.

Ο ViT μπορεί να είναι πολύ πιο αποτελεσματικός από ένα CNN, καθώς αντιμετωπίζει επιτυχώς προβλήματα όπως το locality και το translation invariance. Ωστόσο, για να γίνει αυτό χρειάζεται να εκπαιδευτεί σε μεγάλο πλήθος δεδομένων καλής ποιότητας ή να χρησιμοποιηθούν αποδοτικές τεχνικές επαύξησης δεδομένων (data augmentation techniques).

## 1.3 Σχετική Βιβλιογραφία

Η εκτίμηση του 3Δ σχήματος και της πόζας των ανθρώπων από μία εικόνα ή ένα βίντεο έχει μελετηθεί εκτενώς από την ερευνητική κοινότητα τα τελευταία χρόνια. Από κλασικές μεθόδους της Όρασης Υπολογιστών μέχρι την χρήση σύγχρονων τεχνικών Βαθιάς Μάθησης, κάθε μέθοδος προσπαθεί να συμβάλλει στη βελτίωση της ποιότητας των αποτελεσμάτων. Παρακάτω παρουσιάζονται συνοπτικά κάποιες από τις πιο αποτελεσματικές μεθόδους στο πεδίο καθώς και μέθοδοι που αντιμετωπίζουν σχετικά προβλήματα όπως η εκτίμηση της 2Δ πόζας των ανθρώπων.

### 1.3.1 Εκτίμηση 2Δ πόζας

Ένα θεμελιώδες πρόβλημα της Όρασης Υπολογιστών αποτελεί η εκτίμηση της 2Δ πόζας των ανθρώπων. Αυτό το πρόβλημα είναι πολύ σημαντικό καθώς μπορεί να οδηγήσει στην κατασκευή ενός σκελετού για τον κάθε άνθρωπο, το οποίο έχει αρκετές εφαρμογές σε διαφορετικά πεδία, όπως η μελέτη της κίνησης ή η αξιολόγηση και βελτίωση της τεχνικής των αθλητών.

Επιπρόσθετα, οι περισσότερες μέθοδοι βελτιστοποίησης που έχουν αναπτυχθεί για την εκτίμηση της 3Δ πόζας των ανθρώπων χρησιμοποιούν την 2Δ πόζα, είτε ως έναν τρόπο ανύψωσης από το 2Δ επίπεδο στο 3Δ είτε ως σφάλμα προβολής της 3Δ θέσης των αρθρώσεων σε σχέση με τη 2Δ θέση τους. Συνεπώς, μία αποτελεσματική μέθοδος για την αντιμετώπιση αυτού του

προβλήματος πρέπει να στηρίζεται σε μία εξίσου αποτελεσματική μέθοδο για την εκτίμηση της 2Δ πόζας.

Το πρόβλημα έχει μελετηθεί και λύσεις έχουν προταθεί που βασίζονται τόσο σε παραδοσιακές μεθόδους [58, 67] όσο και σε πιο πρόσφατες αρχιτεκτονικές Βαθιάς Μάθησης [35, 61, 65]. Οι πρώτες μέθοδοι που αναπτύχθηκαν χρησιμοποιούν παραδοσιακές τεχνικές της Όρασης Υπολογιστών. Μία συνηθισμένη τεχνική είναι τα part-based μοντέλα, τα οποία χωρίζουν το σώμα σε μέρη εντοπίζοντας τα ξεχωριστά, και στη συνέχεια μοντελοποιούν τις μεταξύ τους αποστάσεις ώστε να εκτιμήσουν την πόζα. Επίσης, πολλές φορές χρησιμοποιούνται χαρακτηριστικά του σώματος που εξάγονται από descriptors, όπως τα Histograms of Oriented Gradients (HOG) [10], καθώς και πιθανοτικές μέθοδοι [59] οι οποίες έχουν ως στόχο την ελαχιστοποίηση μίας αντικειμενικής συνάρτησης που ελέγχει την πιθανότητα οι θέσεις των αρθρώσεων να είναι τέτοιες ώστε η πόζα να είναι εφικτή ανατομικά. Αν και όλες αυτές οι μέθοδοι έθεσαν τα θεμέλια για την μελέτη του προβλήματος της εκτίμησης της 2Δ πόζας του ανθρώπου, είναι αρκετά κοστοβόρες υπολογιστικά, ενώ η ακρίβεια στην εκτίμηση της πόζας είναι αρκετά περιορισμένη, ειδικά σε πολύπλοκες πόζες.

Το ταχέως αναπτυσσόμενο και υποσχόμενο πεδίο της Βαθιάς Μάθησης έχει φέρει σημαντική πρόοδο στο χώρο με σύγχρονες μεθόδους οι οποίες βασίζονται κυρίως στα CNNs και στους ViTs. Αυτά τα μοντέλα μπορούν να εκπαιδευτούν ώστε να μάθουν εύρωστα χαρακτηριστικά σε πολλά επίπεδα απευθείας από τα δεδομένα. Η εποχή των μεγάλων δεδομένων (Big Data) επιτρέπει την εκπαίδευση τέτοιων μοντέλων σε μεγάλο πλήθος δεδομένων εκπαίδευσης υψηλής ποιότητας και αρκετά μεγάλης διακύμανσης, ώστε το μοντέλο να μάθει να γενικεύει καλύτερα. Το OpenPose [7] και το ViTPose [71] αποτελούν δύο από τις πιο χαρακτηριστικές και αποτελεσματικές μεθόδους που έχουν αναπτυχθεί για την εκτίμηση της 2Δ πόζας.

Το OpenPose αποτελεί μία μέθοδο που επιτρέπει την εκτίμηση σε πραγματικό χρόνο της 2Δ πόζας για περισσότερα από ένα άτομα σε μία σκηνή. Βασίζεται σε ένα CNN το οποίο έχει εκπαιδευτεί σε ένα dataset μεγάλης κλίμακας το οποίο περιέχει και σκηνές με αρκετούς ανθρώπους οπότε να μπορεί να αντιμετωπίσει επιτυχώς αντίστοιχες περιπτώσεις. Οι δυνατότητες του, ωστόσο, περιορίζονται όταν υπάρχουν αρκετά μη ορατά μέρη του σώματος, σε πολύπλοκες πόζες ή σε χαμηλής ποιότητας εικόνες.

Το ViTPose στηρίζεται σε μία απλή αρχιτεκτονική, η οποία χρησιμοποιεί, όπως φανερώνει και το όνομά του, έναν ViT. Ο self-attention μηχανισμός επιτρέπει στο ViT να μαθαίνει καλύτερα τα χαρακτηριστικά σε μία εικόνα, μελετώντας τις σχέσεις μεταξύ των σημείων όλης της εικόνας, σε αντίθεση με τα CNNs, τα οποία περιορίζονται σε πιο τοπικά χαρακτηριστικά. Συγκριτικά με το OpenPose, το ViTPose μπορεί να εκτιμήσει με μεγαλύτερη επιτυχία δύσκολες πόζες σε μεγαλύτερο ηλικιακό φάσμα. Γι'αυτό το λόγο, στις μεθό-

δους μας χρησιμοποιούμε αυτό το μοντέλο για την εκτίμηση της θέσης των 2Δ αρθρώσεων.

### 1.3.2 Εκτίμηση 3Δ σχήματος και πόζας

Η ανάγκη για καλύτερη μοντελοποίηση του ανθρώπινου σώματος και η αξιοποίηση αυτής σε ποικίλες εφαρμογές οδήγησε στην ανάπτυξη μεθόδων για την εκτίμηση του 3Δ σχήματος και της πόζας των ανθρώπων. Η εκτίμηση της 3Δ πόζας έχει μελετηθεί και ως ξεχωριστό πρόβλημα, καθώς δεν επηρεάζεται σε μεγάλο βαθμό από το σχήμα. Αντίθετα, οι περισσότερες μέθοδοι που κάνουν εκτίμηση του 3Δ σχήματος, κάνουν ταυτόχρονα και εκτίμηση της 3Δ πόζας, αφού το σχήμα εξαρτάται από την πόζα του ανθρώπου.

Οι πρώτες μέθοδοι που αναπτύχθηκαν για την εκτίμηση της 3Δ πόζας των ανθρώπων βασίζονταν στην μετατροπή της 2Δ πόζας σε 3Δ [8, 41]. Με την εκτίμηση της 2Δ θέσης των αρθρώσεων, συνήθως με ένα CNN, χρησιμοποιούσαν ένα νευρωνικό δίκτυο ή έναν regressor για την εκτίμηση της 3Δ θέσης των αρθρώσεων. Πιο σύγχρονες μέθοδοι έχουν αναπτυχθεί για την εκτίμηση της 3Δ πόζας οι οποίες δεν χρειάζονται το ενδιάμεσο στάδιο της εκτίμησης της 2Δ πόζας. Αυτές οι μέθοδοι χρησιμοποιούν κυρίως τεχνικές βαθιάς μάθησης [33, 60] αλλά και πιο καινοτόμες ιδέες, όπως σήματα WiFi [72].

Μία από τις βασικότερες προκλήσεις στην μελέτη του 3Δ κόσμου από 2Δ σήματα είναι η αβεβαιότητα του βάθους. Για να αντιμετωπιστεί αυτό το πρόβλημα στην εκτίμηση του 3Δ σχήματος και της πόζας, πολλές μέθοδοι χρησιμοποιούν κάμερες που καταγράφουν διαφορετικές όψεις του κόσμου (multi-view cameras) [11, 20] ή βίντεο [52, 73]. Εφόσον το σχήμα ενός ανθρώπου δεν αλλάζει κατά τη διάρκεια ενός βίντεο και δεν εξαρτάται από την οπτική που θα το δει κάποιος, οι μορφές αυτές βοηθούν στην καλύτερη κατανόηση και εκτίμηση του σχήματος των ανθρώπων. Λύσεις στο πρόβλημα προσφέρουν και οι κάμερες RGB-D, οι οποίες δίνουν κάθε στιγμή πληροφορίες για το βάθος της σκηνής [5].

Σε περιπτώσεις μελέτης της κίνησης, για την αποφυγή παράλογης κίνησης έχουν προταθεί εμβιομηχανικά (biomechanics) μοντέλα της κίνησης και του ανθρώπινου σώματος [28, 70]. Φαινόμενα όπως το «skating» των ποδιών, δηλαδή όταν τα πόδια του ανθρώπου γλιστρούν με αφύσικο τρόπο, θα μπορούσαν να αποφευχθούν λαμβάνοντας υπόψη εμβιομηχανικούς περιορισμούς.

Στην περίπτωση όπου η είσοδος είναι μία απλή RGB εικόνα, η μελέτη γίνεται δυσκολότερη. Ωστόσο, αρκετές μέθοδοι έχουν αναπτυχθεί με στόχο την επίλυση του προβλήματος. Όπως αναφέρθηκε, αυτές οι μέθοδοι είναι κυρίως βασισμένες στην βελτιστοποίηση [14, 48], στην παλινδρόμηση με τεχνικές βαθιάς μάθησης [24, 75] καθώς και σε συνδυασμό βελτιστοποίησης και παλινδρόμησης [29].

Στη συνέχεια περιγράφονται συνοπτικά δύο αποτελεσματικές μέθοδοι, μία βελτιστοποίησης και μία παλινδρόμησης, που χρησιμοποιήσαμε στα αρχικά πειράματά μας.

### SMPLify-X

Το SMPLify-X [48] αποτελεί μία από τις πιο αποτελεσματικές μεθόδους βελτιστοποίησης για την εκτίμηση του 3Δ σχήματος και της πόζας των ανθρώπων από μία εικόνα RGB. Στηρίζεται σε μία παλαιότερη μέθοδο, το SMPLify [6], την οποία επεκτείνει και βελτιώνει. Σκοπός είναι η ελαχιστοποίηση μίας αντικειμενικής συνάρτησης μέσα από την οποία θα προκύψουν οι παράμετροι του μοντέλου SMPL-X [48]. Όπως και στις περισσότερες παρόμοιες μεθόδους, έτσι και σε αυτή, η αντικειμενική συνάρτηση βασίζεται στην ελαχιστοποίηση ενός σφάλματος προβολής της θέσης των 3Δ αρθρώσεων με την 2Δ αντίστοιχη θέση που έχει ανιχνεύσει το OpenPose. Η αντικειμενική συνάρτηση είναι η παρακάτω:

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C}$$

Οι παράμετροι $\theta_b, \theta_f$ και $m_h$ είναι τα διανύσματα για την πόζα του σώματος, του προσώπου και των χεριών, αντίστοιχα, ενώ οι όροι $E_{m_h}(m_h), E_{\theta_f}(\theta_f)$ και $E_\mathcal{E}(\psi)$ είναι L2 priors για την πόζα των χεριών, του προσώπου και τις εκφράσεις του προσώπου αντίστοιχα. Ο όρος $E_\beta(\beta)$ εκφράζει την απόσταση Mahalanobis μεταξύ των παραμέτρων του σχήματος κατά τη βελτιστοποίηση και της κατανομής των τιμών του σχήματος στο σύνολο εκπαίδευσης του SMPL-X. Τέλος, ο όρος $E_\alpha(\theta_b)$ είναι ένας prior για τους αγκώνες και τα γόνατα, ενώ ο όρος $E_J$ είναι ο όρος σφάλματος για την απόσταση μεταξύ των 2Δ keypoints που ανιχνεύθηκαν από το OpenPose και της προβολής των 3Δ σε 2Δ keypoints.

Το SMPLify-X εισάγει τον Variational Human Body Pose Prior (VPoser) που εκφράζεται με τον όρο $E_{\theta_b}(\theta_b)$ στην αντικειμενική συνάρτηση και σκοπός του είναι η αποτροπή ανέφικτων στάσεων σώματος. Επίσης, εισάγεται ένα νέο interpenetration loss ($E_C(\theta)$ στην αντικειμενική συνάρτηση) για την αποφυγή «συγκρούσεων» και «διείσδυσης» ενός μέρους του σώματος με ένα άλλο τα οποία δεν είναι φυσικά ερμηνεύσιμα. Τέλος, έχουν αναπτυχθεί μοντέλα στοχευμένα για κάθε φύλο, και γι'αυτό το λόγο χρησιμοποιείται ένας Deep Gender Classifier για την αναγνώριση του φύλου του κάθε ανθρώπου. Σε περίπτωση όπου δεν ανιχνευθεί με μεγάλη σιγουριά κάποιο φύλο, χρησιμοποιείται ένα ουδέτερο μοντέλο.

### BEV

Το BEV [62] είναι μία μέθοδος που χρησιμοποιεί τεχνικές βαθιάς μάθησης και παλινδρόμηση για την πρόβλεψη της 3Δ πόζας και του σχήματος των ανθρώπων σε μία εικόνα ή ένα βίντεο. Έχει αναπτυχθεί για περιπτώσεις, όπου σε μία εικόνα συνυπάρχουν πολλοί άνθρωποι, καθώς εισάγει μία απεικόνιση «bird's-eye-view» για τον εντοπισμό των κέντρων των σωμάτων των ανθρώπων. Ταυτόχρονα, η χρήση ενός τροποποιημένου SMPL-A μοντέλου την καθιστά μία αποτελεσματική στην εκτίμηση του 3Δ σχήματος παιδιών μέθοδο. Βέβαια, η εκτίμηση της πόζας παραμένει ένα ζήτημα αφού αρκετές φορές είναι λανθασμένη.

Το BEV χρησιμοποιεί ένα νευρωνικό δίκτυο για την εξαγωγή χαρτών του κέντρου των σωμάτων, του localization offset από την μπροστινή όψη και δύο χαρτών για το bird's eye view. Αυτοί οι τέσσερις χάρτες συνδυάζονται για να παράξουν χάρτες για την 3Δ θέση των κέντρων και του offset. Στη συνέχεια, αυτοί οι νέοι χάρτες προβλέπουν την 3Δ μετατόπιση των ανθρώπων, η οποία τελικά σε συνδυασμό με έναν χάρτη χαρακτηριστικών για το 3Δ πλέγμα του ανθρώπου κάνει regress τις SMPL-A παραμέτρους.

Όπως προαναφέρθηκε, το BEV χρησιμοποιεί ένα μερικώς τροποποιημένο SMPL-A μοντέλο για την καλύτερη περιγραφή των βρεφών. Συγκεκριμένα, δεδομένου ότι το SMPL-A χρησιμοποιεί μόνο το shape space του SMPL τόσο για ενήλικες όσο και για βρέφη, επειδή το SMPL δεν έχει μοντελοποιηθεί κατάλληλα για παιδιά και βρέφη, το σχήμα που προκύπτει με το SMPL shape space δεν είναι πολλές φορές ρεαλιστικό. Έτσι, όταν το βάρος παρεμβολής $\alpha$ είναι μεγαλύτερο από ένα κατώφλι $t_\alpha$, δηλαδή μοιάζει περισσότερο για βρέφος, χρησιμοποιείται το μοντέλο SMIL. Αντίθετα, όταν $\alpha \leq t_\alpha$, χρησιμοποιείται το SMPL μοντέλο. Η αλλαγή αυτή βοηθάει στην καλύτερη μοντελοποίηση του σχήματος των ανθρώπων γενικότερα, αλλά και ειδικά των βρεφών.

## 1.4 Μεθοδολογία

Στην παρούσα εργασία προτείνουμε δύο νέες μεθόδους για την εκτίμηση 3Δ πόζας και σχήματος των ανθρώπων. Για να το πετύχουμε αυτό, επεκτείνουμε υπάρχουσες αποτελεσματικές για ενήλικες μεθόδους χρησιμοποιώντας σύνολα δεδομένων με εικόνες που περιέχουν παιδιά και βρέφη, ώστε να μπορεί να γίνει καλύτερη μοντελοποίηση αυτών των ηλικιών. Στόχος μας είναι η δημιουργία μεθόδων που εκτιμούν την 3Δ πόζα και σχήμα το ίδιο καλά ανεξάρτητα από την ηλικία του εκάστοτε ανθρώπου. Αμφότερες οι μέθοδοι λειτουργούν με είσοδο τόσο εικόνα όσο και βίντεο, όπως και για εικόνες με ένα ή πολλούς ανθρώπους.

Η πρώτη μέθοδος είναι μια μέθοδος βασισμένη σε βελτιστοποίηση. Είναι

αρχικά υλοποιημένη για να λειτουργεί σε βίντεο, ωστόσο εμείς την επεκτείνουμε και σε εικόνες RGB. Σκοπός είναι η ελαχιστοποίηση μίας αντικειμενικής συνάρτησης η οποία αποτελείται από αρκετούς όρους σφάλματος και κανονικοποίησης με σημαντικότερο αυτό του σφάλματος προβολής των 3Δ σημείων στα ψευδο-πραγματικά 2Δ αντίστοιχα σημεία. Τελικά, η μέθοδος εφαρμόζει το μοντέλο SMPL-A σε κάθε άνθρωπο των δεδομένων εισόδου.

Βασικός στόχος μας είναι η εκπαίδευση ενός μοντέλου που μπορεί να κάνει απευθείας πρόβλεψη των παραμέτρων πόζας και σχήματος βάσει ενός μοντέλου, όπως το SMPL-A, για κάθε άνθρωπο σε μία εικόνα. Αυτό πετυχαίνουμε με τη δεύτερη μέθοδο. Η πολύ καλή ποιότητα των αποτελεσμάτων της πρώτης μεθόδου βοήθησε στη δημιουργία ψευδο-πραγματικών επισημειώσεων σε εικόνες παιδιών και βρεφών από ανοιχτά σύνολα δεδομένων της ερευνητικής κοινότητας, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση ενός νευρωνικού δικτύου που κάνει εκτίμηση των παραμέτρων σχήματος και πόζας βάσει του μοντέλου SMPL-A. Αυτή η μέθοδος αποτελεί και μία αρκετά πιο γρήγορη μέθοδο συγκριτικά με την πρώτη καθώς δεν υπάρχει η φάση της βελτιστοποίησης.

### 1.4.1 Μέθοδος βασισμένη στη Βελτιστοποίηση

Ένα βίντεο είναι μια ακολουθία από εικόνες όπου συνήθως υπάρχει κάποια αλλαγή στη σκηνή μεταξύ τους. Ωστόσο, είναι δυνατό αυτές οι εικόνες να είναι και οι ίδιες. Με αυτή τη θεώρηση μπορούμε να ανάγουμε το πρόβλημα της εκτίμησης 3Δ σχήματος και πόζας των ανθρώπων από μία εικόνα στο αντίστοιχο πρόβλημα με είσοδο ένα βίντεο. Πάνω σε αυτήν την παρατήρηση και σε πειράματα που έγιναν σε σχετικές μεθόδους και δείχνουν ότι όταν η είσοδος είναι ένα βίντεο τα αποτελέσματα είναι καλύτερα συγκριτικά με είσοδο μία εικόνα βασίζεται η πρώτη μέθοδος μας.

Συγκεκριμένα, βασιζόμαστε στο SLAHMR (Simultaneous Localization and Human Mesh Recovery) [73], μία αποτελεσματική μέθοδο βελτιστοποίησης για την 3Δ εκτίμηση σχήματος και πόζας των ανθρώπων σε ένα βίντεο, και την τροποποιούμε κατάλληλα ώστε να δουλεύει το ίδιο καλά για εικόνες, και για ανθρώπους ανεξαρτήτως ηλικίας. Ο τρόπος λειτουργίας του SLAHMR απεικονίζεται στην Εικόνα 1.1.

Το SLAHMR έχει σχεδιαστεί ώστε να λειτουργεί για βίντεο που έχουν ληφθεί υπό πραγματικές συνθήκες, και όχι σε ελεγχόμενα πειραματικά περιβάλλοντα, το οποίο εισάγει δυσκολίες όπως τις απότομες κινήσεις και το φυσικό background. Χρησιμοποιεί προχωρημένες τεχνικές για το tracking τον ανθρώπων στο βίντεο, ενώ πέρα από την ανακατασκευή του σώματος των ανθρώπων, ανακτά και τις τροχιές τους, όπως και την κίνηση της κάμερας σε ένα κοινό σύστημα συντεταγμένων του κόσμου. Σε ένα βίντεο που έχει ληφθεί υπό πραγματικές συνθήκες η κίνηση της κάμερας είναι μία σημαντική παράμετρος, η

οποία πρέπει και αυτή να μοντελοποιηθεί ώστε η ανακατασκευή των ανθρώπων να έχει γίνει σωστά.

Όπως αναφέρθηκε, το σύστημά μας λαμβάνει ως είσοδο ένα βίντεο $T$ καρέ (frames) το οποίο περιέχει $N$ ανθρώπους. Κάθε άνθρωπος $i$ τη χρονική στιγμή $t$ αναπαρίσταται ως:

$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}$$

όπου $\Phi_t^i \in \mathbb{R}^3$ είναι ο ολικός προσανατολισμός (global orientation), $\Theta_t^i \in \mathbb{R}^{22\times 3}$ εκφράζει την πόζα από 22 αρθρώσεις, $\beta^i \in \mathbb{R}^{11}$ οι παράμετροι για το σχήμα σε όλες τις χρονικές στιγμές $t$, με την $11^\eta$ τιμή να αντιστοιχεί στο βάρος παρεμβολής $\alpha$, και $\Gamma_t^i \in \mathbb{R}^3$ η μετατόπιση του root (root translation).

Το πρώτο βήμα είναι η κατά καρέ εκτίμηση της στάσης για κάθε άνθρωπο, κάνοντας 3D tracking σε όλα τα καρέ χρησιμοποιώντας το 4DHumans [16] tracking σύστημα.

Μία σωστή μελέτη του προβλήματος οφείλει να κάνει σωστή μοντελοποίηση και της κίνησης της κάμερας, αφού σε ένα βίντεο η κίνηση ενός ανθρώπου στο σύστημα συντεταμένων της κάμερας είναι μία συνάρτηση της κίνησης τόσο του ανθρώπου όσο και της κάμερας στο σύστημα συντεταγμένων του κόσμου.

Γι'αυτό το λόγο χρησιμοποιείται το DROID-SLAM [64], ένα σύστημα SLAM για να εκτιμηθεί ο μετασχηματισμός $\{\hat{R}_t, \hat{T}_t\}$ από τον κόσμο στην κάμερα για κάθε χρονική στιγμή $t$. Για την εκτίμηση της κλίμακας της κάμερας $\alpha_c$ και των τροχιών των ανθρώπων χρησιμοποιείται ένας prior για την κίνηση των ανθρώπων στον κόσμο.

Ξεκινάμε με την αρχικοποίηση του ολικού προσανατολισμού και του root translation στο σύστημα συντεταγμένων του κόσμου, ενώ η κλίμακα της κάμερας αρχικοποιείται στην τιμή $\alpha_c = 1$:

$$\begin{aligned} {}^w\Phi_t^i &= R_t^{-1c}\hat{\Phi}_t^i, & {}^w\Gamma_t^i &= R_t^{-1c}\hat{\Gamma}_t^i - \alpha_c R_t^{-1} T_t, \\ \beta_i &= \hat{\beta}_i, & \Theta_t^i &= \hat{\Theta}_t^i, \end{aligned}$$

Οι αρθρώσεις στο σύστημα του κόσμου εκφράζονται ως:

$${}^w\mathbf{J}_t^i = \mathcal{M}({}^w\mathbf{\Phi}_t^i, \mathbf{\Theta}_t^i, \beta^i) + {}^w\,\mathbf{\Gamma}_t^i$$

όπου $\mathcal{M}$ είναι η συνάρτηση που χρησιμοιεί το SMPL-A για να δημιουργήσει τις αρθρώσεις και τα σημεία της $3\Delta$ αναπαράστασης του ανθρώπου.

Η αντικειμενική συνάρτηση κατά τη βελτιστοποίηση βασίζεται, όπως και οι περισσότερες σχετικές μέθοδοι, στο σφάλμα προβολής των $3\Delta$ αρθρώσεων στο $2\Delta$ επίπεδο βάσει των $2\Delta$ αρθρώσεων $x_t^i$ που έχουν έχουν ανιχνευθεί:

$$E_{\text{data}} = \sum_{i=1}^{N} \sum_{t=1}^{T} \psi_t^i \rho(\Pi_K((R_t \cdot^w \mathbf{J_t^i} + \alpha \mathbf{T_t}) - \mathbf{x_t^i})$$

όπου $\Pi_K(\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T) = K \begin{bmatrix} \frac{x_1}{x_3} & \frac{x_2}{x_3} & 1 \end{bmatrix}^T$ είναι η προοπτική προβολή με intrinsic πίνακα της κάμερας $K \in \mathbb{R}^{2\times 3}$, $\rho$ η Geman-McClure συνάρτηση [4] και $\psi_t^i$ το confidence score των ανιχνευμένων $2\Delta$ αρθρώσεων.

Σε αυτό το στάδιο, η βελτιστοποίηση γίνεται αποκλειστικά στον ολικό προσανατολισμό και στο root translation ${}^w\Phi_t^i, {}^w\Gamma_t^i$:

$$\min_{\{\{{}^w\Phi_t^i, {}^w\Gamma_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}}$$

Στο δεύτερο στάδιο γίνεται ομαλοποίηση της μετάβασης μεταξύ των στάσεων των ανθρώπων στον κόσμο. Ο prior για την ομαλοποίηση των αρθρώσεων ορίζεται ως:

$$E_{\text{smooth}} = \sum_{i}^{N} \sum_{t}^{T} \|\mathbf{J_t^i} - \mathbf{J_{t+1}^i}\|^2$$

Priors όροι χρησιμοποιούνται, επίσης, για την βελτιστοποίηση της κλίμακας της κάμερας $\alpha_c$, τις παραμέτρους του σχήματος $\beta_i$ και πόζας $\Theta_t^i$ των ανθρώπων όπως και για την κίνηση των ανθρώπων στον κόσμο.

Η συνάρτηση βελτισοποίησης γίνεται τελικά σε αυτό το στάδιο:

$$\min_{\alpha, \{\{{}^w\mathbf{P_t^i}\}_{\mathbf{t=1}}^{\mathbf{T}}\}_{\mathbf{i=1}}^{\mathbf{N}}} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + \lambda_{\text{smooth}} E_{\text{smooth}}$$

όπου $E_{\text{pose}} = \sum_{i=2}^{N} \sum_{t=1}^{T} \|\zeta_t^i\|^2$ και $E_\beta = \sum_i^N \|\beta^i\|^2$ είναι priors για τη πόζα και το σχήμα, αντίστοιχα, με $\zeta_t^i \in \mathbb{R}^{32}$ να αναπαριστά τις παραμέτρους της πόζας $\Theta_t^i$ στον latent χώρο του μοντέλου VPoser. Σε αυτό το στάδιο γίνεται βελτιστοποίηση και στις παραμέτερους σχήματος και πόζας των ανθρώπων, όπως και στην κλίμακα της κάμερας.

Για την φυσικότητα της ανθρώπινης κίνησης χρησιμοποιείται ένας εκπαιδεύσιμος prior, βασισμένος στο HuMoR [54], ο οποίος χρησιμοποιεί έναν Conditional Variational Autoencoder (CVAE). Ο CVAE μαθαίνει την κατανομή πιθανότητας των φυσικών κινήσεων του σώματος, επιτρέποντας στο μοντέλο να παράγει ρεαλιστικές ακολουθίες κίνησης και να τις χρησιμοποιεί ως περιορισμό κατά την βελτιστοποίηση, εξασφαλίζοντας έτσι την φυσικότητα του τελικού αποτελέσματος.

Εκτός από τους κύριους όρους σφάλματος με τους prior για την κίνηση, δύο επιπλέον όροι κανονικοποίησης εισάγονται κατά τη βελτιστοποίηση για

να εξασφαλισθεί ότι η κίνηση είναι φυσικά πραγματοποιήσιμη και εύλογη. Ο πρώτος όρος αφορά ένα σφάλμα σταθερότητας ($E_{\text{stab}}$) ο οποίος κανονικοποιεί την ταχύτητα και τις θέσεις των αρθρώσεων όπως αυτά έχουν προβλεφθεί. Έτσι, ο όρος σφάλματος γίνεται:

$$E_{\text{prior}} = \lambda_{\text{CVAE}} E_{\text{CVAE}} + \lambda_{\text{stab}} E_{\text{stab}}$$

Ο δεύτερος όρος κανονικοποίησης αποτρέπει το συχνό φαινόμενο σε αυτές τις μεθόδους, όπου ο άνθρωπος φαίνεται να κάνει skating. Για την κανονικοποίηση ελέγχεται η ταχύτητα εκείνων των αρθρώσεων που είναι πιθανότερο να βρίσκονται σε επαφή με το έδαφος της σκηνής:

$$E_{\text{skate}} = \sum_{i}^{N} \sum_{t}^{T} \sum_{j}^{J} c_t^i(j) \| J_t^i(j) - J_{t+1}^i(j) \|$$

όπου $c_t^i(j)$ η πιθανότητα επαφής ενός συνδέσμου $j$ με το έδαφος, για τον άνθρωπο $i$ τη χρονική στιγμή $t$, και $J_t^i(j)$ η θέση του.

Τέλος, ο όρος $E_{\text{con}}$ προσπαθεί να εξασφαλίσει ότι αυτά τα σημεία επαφής θα παραμείνουν κοντά στο έδαφος:

$$E_{\text{con}} = \sum_{i}^{N} \sum_{t}^{T} \sum_{j}^{J} c_t^i(j) \max(d(J_t^i(j), g) - \delta, 0).$$

όπου $d(\mathbf{p}, \mathbf{g})$ είναι η απόσταση ενός σημείου $\mathbf{p} \in \mathbb{R}^{\mathbf{3}}$ από το έδαφος $g \in \mathbb{R}^3$, και $\delta$ μία μικρή σταθερά. Συνδυάζοντας όλους αυτούς τους όρους σφάλματος εξασφαλίζουμε την ομαλότητα της κίνησης, αλλά και την ρεαλιστικότητά της. Να επισημανθεί ότι το έδαφος $g$ βελτιστοποιείται ως μία ελεύθερη μεταβλητή, κοινή για όλους τους ανθρώπους σε όλες τις χρονικές στιγμές.

Το τελικό πρόβλημα βελτιστοποίησης προσθέτοντας όλους τους όρους σφάλματος περιγράφεται ως:

$$\min_{\alpha_c, g, \{s_0^i\}_{i=1}^N, \{\{\mathbf{z_t^i}\}_{\mathbf{t=1}}^{\mathbf{T}}\}_{\mathbf{i=1}}^{\mathbf{N}}} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + E_{\text{prior}} + E_{\text{env}}$$

όπου $E_{\text{env}} = \lambda_{\text{skate}} E_{\text{skate}} + \lambda_{\text{con}} E_{\text{con}}$. Στο τελικό αυτό στάδιο πραγματοποιείται βελτιστοποίηση στις μεταβλητές για την ομαλότητα της κίνησης.

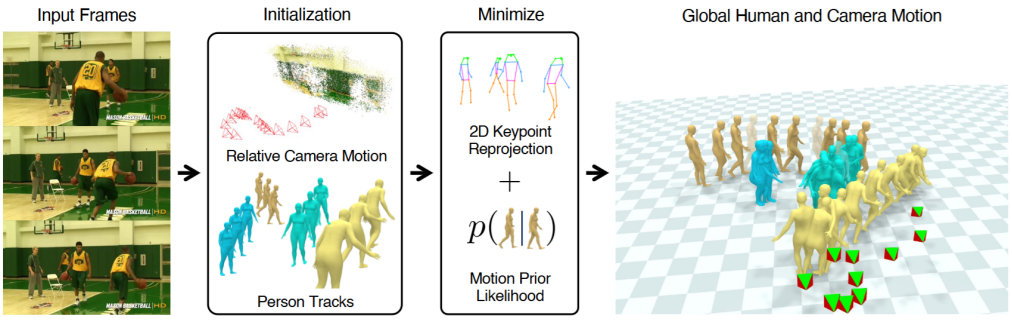



Figure 1.1: SLAHMR Pipeline. Εικόνα από [73]

### 1.4.2 Μέθοδος βασισμένη στην Παλινδρόμηση (Μέθοδος Βαθιάς Μάθησης)

Η πρώτη μέθοδος που περιγράφηκε αποτελεί μία εξαιρετικά αποτελεσματική μέθοδο βελτιστοποίησης για την εκτίμηση της 3Δ στάσης και σχήματος των ανθρώπων σε μία σκηνή. Ωστόσο, το γεγονός ότι βασίζεται στη βελτιστοποίηση έχει ως βασικό μειονέκτημα τον μεγάλο χρόνο εκτέλεσης μέχρι να λυθεί το πρόβλημα βελτιστοποίησης. Έτσι, δεν μπορεί να εφαρμοστεί σε πραγματικό χρόνο, ενώ απαιτεί και σημαντικά μεγάλη υπολογιστική ισχύ.

Τα προβλήματα αυτά προσπαθεί να λύσει η δεύτερη μέθοδος που προτείνεται, η οποία βασίζεται σε αρχιτεκτονική βαθιάς μάθησης, με χρήση παλινδρόμησης. Στηρίζεται στο HMR2.0 [16], μία μέθοδο όπου δεδομένης μίας εικόνας, χρησιμοποιεί έναν Vision Transformer για την πρόβλεψη των παραμέτρων SMPL για κάθε άνθρωπο της εικόνας. Είναι μία πολύ αποτελεσματική και γρήγορη μέθοδος, με απλή αρχιτεκτονική η οποία παρουσιάζεται στην Εικόνα 1.2.

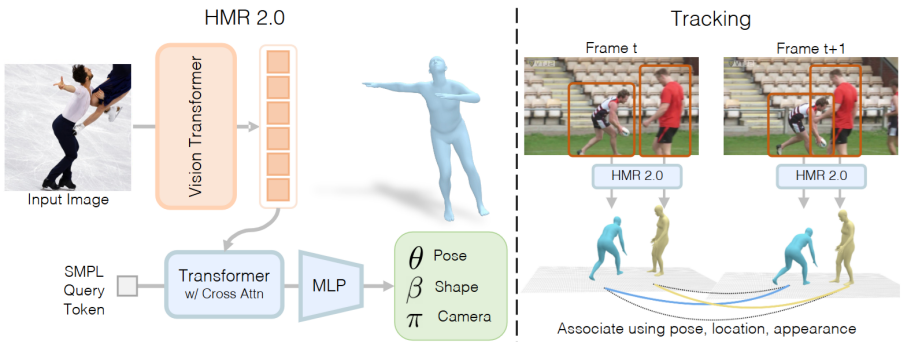



Figure 1.2: HMR2.0 Overview. Εικόνα από [16]

Η ιδέα βασίζεται στο Human Mesh Recovery (HMR) [24], με βασική δι-

αφορά την αντικατάσταση του CNN με έναν ViT. Το HMR2.0 ακολουθεί μία απλή end-to-end transformer αρχιτεκτονική, η οποία ωστόσο του επιτρέπει να επιτυγχάνει πολύ καλά αποτελέσματα στη 3Δ ανακατασκευή. Με την επέκταση μας στη μέθοδο και στην αρχιτεκτονική, με τη χρήση δεδομένων παιδιών και μωρών και με την αντικατάσταση του μοντέλου περιγραφή του ανθρώπινου σώματος από το SMPL στο SMPL-A, καταφέρνουμε να έχουμε πολύ καλά αποτελέσματα στην εκτίμηση του 3Δ σχήματος και πόζας για άτομα κάθε ηλικίας.

Η αρχιτεκτονική του HMR2.0 αποτελείται από έναν ViT ο οποίος «κόβει» την εικόνα σε κομμάτια, στο καθένα από τα οποία εξάγει tokens. Αυτά επεξεργάζονται από έναν αποκωδικοποιητή Transformer (Transformer decoder) με multi-head self-attention και καταλήγουν σε ένα MLP που κάνει πρόβλεψη για τις SMPL-A παραμέτρους σχήματος $\boldsymbol{\beta}$ και πόζας $\boldsymbol{\theta}$ καθώς και την μετατόπιση της κάμερας $\pi$. Το βάρος παρεμβολής $\alpha$ υπολογίζεται ως η $11^{\eta}$ τιμή της παραμέτρου $\boldsymbol{\beta}$.

Το μοντέλο εκπαιδεύεται σε ένα συνδυασμό συνόλων δεδομένων τα οποία έχουν διαφορετικές επισημειώσεις, και έτσι χρησιμοποιείται ένας συνδυασμός $2\Delta$ και $3\Delta$ όρων σφάλματος καθώς και ένας discriminator.

Συγκεκριμένα, με είσοδο μία εικόνα $I$, η πρόβλεψη του μοντέλου είναι $\Theta = [\boldsymbol{\theta}, \boldsymbol{\beta}, \pi] = f(I)$. Ο πρώτος όρος σφάλματος που χρησιμοποιείται όταν υπάρχουν επισημειώσεις για τις πραγματικές τιμές της πόζας $\boldsymbol{\theta}^*$ και του σχήματος $\boldsymbol{\beta}^*$ είναι ένα MSE σφάλμα στις προβλέψεις:

$$\mathcal{L}_{\text{smpl}} = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$$

Στην περίπτωση που υπάρχουν επισημειώσεις για τα $3\Delta$ σημεία ενδιαφέροντος $X^*$, προστίθεται ένας L1 όρος σφάλματος για την απόσταση από τα σημεία ενδιαφέροντος που έχουν προβλεφθεί $X$:

$$\mathcal{L}_{\text{kp3D}} = \|X - X^*\|_1$$

Με παρόμοιο τρόπο, όταν υπάρχουν επισημειώσεις για τα αντίστοιχα $2\Delta$ σημεία $x^*$, προσθέτουμε ένα L1 σφάλμα με την προβολή των $3\Delta$ σημείων $\pi(X)$:

$$\mathcal{L}_{\text{kp2D}} = \|\pi(X) - x^*\|_1$$

Τέλος, όπως αναφέρθηκε, χρησιμοποιείται ένας discriminator $D_k$, για να αποτρέπει μη φυσικές πόζες, που έχει εκπαιδευθεί για κάθε παράμετρο του μοντέλου περιγραφής του ανθρώπινου σώματος, δηλαδή τις παραμέτρους πόζας $\theta_b$, του σχήματος $\boldsymbol{\beta}$ και τις κατά μέλος σχετικές περιστροφές $\theta_i$. Ο όρος σφάλματος για τον generator είναι ο παρακάτω:

$$\mathcal{L}_{\text{adv}} = \sum_{k} (D_k(\theta_b, \boldsymbol{\beta}) - 1)^2$$

### Λεπτομέρειες για την εκπαίδευση και την αξιολόγηση

Για την εκπαίδευση του μοντέλου χρησιμοποιούμε το dataset το οποίο δημιουργήσαμε από εικόνες από το SyRIP [21] και το Relative Human [62] με επισημειώσεις από την πρώτη μέθοδο μας καθώς και το συνδυασμό των datasets του HMR2.0 [16].

Για την αξιολόγηση, συγκρίναμε το μοντέλο μας με τρεις μεθόδους της βιβλιογραφίας σε 4 datasets που περιέχουν είτε αποκλειστικά, είτε εν μέρει εικόνες παιδιών ή βρεφών. Συγκεκριμένα, χρησιμοποιούμε τα SyRIP [21] και Relative Human [62], όπου περιέχουν επισημειώσεις για τα 2Δ keypoints, ενώ με τη μέθοδο βελτιστοποίησής μας εξάγουμε παραμέτρους SMPL-A και κάμερας. Επίσης, χρησιμοποιούμε το ChildPlay [63] στο οποίο παράγουμε επισημειώσεις τόσο για τις παραμέτρους SMPL-A και κάμερας από τη μέθοδο βελτιστοποίησης μας, όσο και τις θέσεις των 2Δ keypoints από το ViTPose.

**BabyRobot Dataset** Κατά την αξιολόγηση των μοντέλων χρησιμοποιούμε, ακόμα, ένα καινούριο σύνολο δεδομένων, το BabyRobot. Αποτελείται από εικόνες παιδιών ηλικίας 6-10 ετών που αλληλεπιδρούν με ρομπότ και κινούνται ελεύθερα στο εργαστήριο. Υπάρχουν τρεις διαφορετικές κάμερες για κάθε παιδί και κάθε σκηνή, μία μπροστά από το παιδί (δίπλα στο ρομπότ), μία αριστερά και μία δεξιά του ρομπότ. Κατά την αξιολόγηση χρησιμοποιούμε εικόνες από όλες τις κάμερες για κάθε παιδί. Η έντονη κίνηση στο χώρο καθώς και τα αντικείμενα που υπάρχουν τοποθετημένα εισάγουν occlusions και δυσκολίες στην εκτίμηση της πόζας των παιδιών. Δεδομένου ότι δεν έχουμε επισημειώσεις για τις εικόνες, χρησιμοποιούμε τη μέθοδο βελτιστοποίησής μας για να δημιουργήσουμε τις παραμέτρους SMPL-A και κάμερας, και το ViTPose για τα 2Δ keypoints.

**Μετρικές αξιολόγησης** Ως μετρικές, χρησιμοποιούμε για το 3Δ σχήμα τη μέση διαφορά ύψους από το mesh που προκύπτει από την εκάστοτε μέθοδο και του ψευδο-πραγματικού από τις επισημειώσεις των datasets (AHD), καθώς και την ποσοστιαία αντίστοιχη διαφορά (APHD). Για τα datasets που περιέχουν μόνο εικόνες παιδιών ή μόνο βρεφών εξάγουμε και το μέσο ύψος που έχει προβλεφθεί.

Για την 3Δ πόζα χρησιμοποιούμε την μετρική Mean Per Joint Position Error (MPJPE), δηλαδή το μέσο σφάλμα της 3Δ θέσης των αρθρώσεων.

Για την 2Δ πόζα, χρησιμοποιούμε το Percentage of Correct Keypoints (PCK), μία μετρική σχετικά με το ποσοστό των προβολών των keypoints στο 2Δ επίπεδο που βρίσκονται σε απόσταση από την πραγματική θέση των keypoints μικρότερη από ένα κατώφλι. Σημειώνουμε ότι σε όλα τα datasets για το πραγματικό ύψος όπως και τη 3Δ θέση των keypoints χρησιμοποιούμε τις προβλέψεις της πρώτης μεθόδου μας, ενώ για τα 2Δ keypoints, το SyRIP και το Relative Human περιέχουν στις επισημειώσεις τις θέσεις τους, και στα ChildPlay και BabyRobot τις εξάγουμε από το ViTPose.

**Ποιοτική αξιολόγηση** Τέλος, στον τομέα της 3Δ Όρασης Υπολογιστών πολύ σημαντική είναι η ποιοτική οπτική αξιολόγηση των αποτελεσμάτων, καθώς οι μετρικές και ο υπολογιστής δεν είναι ικανοί να τα αξιολογήσει πλήρως. Για το λόγο αυτό διεξάγαμε μία έρευνα χρηστών υποκειμενικής αξιολόγησης των αποτελεσμάτων μας, συγκριτικά με το HMR2.0 και το BEV. Έτσι, μπορούμε να έχουμε ποιοτική αξιολόγηση των αποτελεσμάτων από ένα μεγαλύτερο δείγμα ανθρώπων, το οποίο οδηγεί σε ασφαλέστερα συμπεράσματα για την αποτελεσματικότητα του μοντέλου μας.

## 1.5 Πειράματα

Κατά τη φάση της ανάπτυξης των μεθόδων, έγιναν διαφορετικά πειράματα μέχρι η ποιότητα των αποτελεσμάτων να είναι η επιθυμητή. Πειράματα έχουν γίνει τόσο κατά την ανάπτυξη της μεθόδου που βασίζεται στη βελτιστοποίηση, όσο και κατά τη διάρκεια της ανάπτυξης της δεύτερης μεθόδου και συγκεκριμένα κατά την εκπαίδευση του μοντέλου. Τα πρώτα χρησιμοποιούν τροποποιημένες προσεγγίσεις των μεθόδων SMPLify-X [48] και SLAHMR [73] ενώ στη δεύτερη φάση όλα τα πειράματα βασίζονται στο μοντέλο HMR2.0 [16]. Στη συνέχεια παραθέτουμε επιγραμματικά τα πειράματα που έχουν γίνει. Η πλειοψηφία αυτών των πειραμάτων χαρακτηρίστηκε ως ανεπιτυχής με βάση τα αποτελέσματά τους. Για περισσότερες λεπτομέρειες παραπέμπουμε τον αναγνώστη στο Κεφάλαιο 6.

### Πειράματα για τη μέθοδο βελτιστοποίησης

Τα πειράματα βελτιστοποίησης βασίστηκαν σε δύο μεθόδους, το SMPLify-X και το SLAHMR, ενώ χρησιμοποιήθηκε εν μέρει και το BEV.

- Πειράματα βασισμένα στο SMPLify-X
  - Εφαρμογή της αρχικής μεθόδου SMPLify-X σε δεδομένα παιδιών και βρεφών.

  - Αντικατάσταση του SMPL-X μοντέλου με το SMPL-A που μοντελοποιεί καλύτερα τα παιδιά.
  - Χρήση των παραμέτρων σχήματος από το BEV, οι οποίες είναι πιο ακριβείς από αυτές του SMPLify-X και βελτιστοποίηση SMPLify-X για τις παραμέτρους της πόζας.
  - Αντικατάσταση του OpenPose με το ViTPose για τα $2\Delta$ keypoints, καθώς αυτά είναι μεγαλύτερης ακρίβειας και εφαρμόζονται και σε δυσκολότερες πόζες.
  - Grid Search για να βρούμε την βέλτιστη τιμή του βάρους $\alpha$ επιλέγοντας αυτή που δίνει το μικρότερο fitting loss.

- Πειράματα βασισμένα στο SLAHMR

  - Εφαρμογή της αρχικής μεθόδου SLAHMR σε δεδομένα παιδιών και βρεφών με τη χρήση του μοντέλου SMPL-A αντί για το SMPL.
  - Grid search στην τιμή του βάρους παρεμβολής $\alpha$ και επιλογή αυτής με το μικρότερο άθροισμα σφαλμάτων.
  - Πάγωμα των παραμέτρων σχήματος $\boldsymbol{\beta}$ είτε σε μηδενική τιμή (εκτός από το $\alpha$) είτε στις τιμές του BEV και βελτιστοποίηση για τις υπόλοιπες παραμέτρους.
  - Βελτιστοποίηση συγκεκριμένα για το SyRIP dataset [21], όπου περιέχει εικόνες βρεφών και μέσω πειραμάτων grid search διαπιστώθηκε ότι υπάρχει μία σχεδόν βέλτιστη τιμή $\alpha \approx 0.9$ για τη μοντελοποίηση αυτών των περιπτώσεων.
  - **Τελικό μοντέλο που χρησιμοποιήσαμε:** Αρχικοποίηση του $\alpha$ στην τιμή 1, ώστε να ξεκινήσει από ένα σημείο η βελτιστοποίηση που πιθανώς να μοντελοποιήσει καλύτερα παιδιά και βρέφη. Όπως και φάνηκε τελικά αυτό βοήθησε τη βελτιστοποίηση και με αυτό τον τρόπο λάβαμε τα καλύτερα αποτελέσματα χωρίς να χρειαστεί grid search για την τιμή του $\alpha$.

### Πειράματα για τη μέθοδο βαθιάς μάθησης

Τα πειράματα για τη μέθοδο βαθιάς μάθησης στηρίχθηκαν στο HMR2.0, με τις διαφορές να εντοπίζονται στον τρόπο εκπαίδευσης. Έγιναν πειράματα τόσο εκπαίδευσης όσο και fine-tuning.

Εφόσον παρέχεται ένα προεκπαιδευμένο μοντέλο HMR2.0, καθώς και όλα τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση, ο πιο εύκολος τρόπος το μοντέλο να «μάθει» και τα δικά μας δεδομένα είναι το fine-tuning στα δικά

μας δεδομένα. Το μοντέλο HMR2.0 έχει τρεις τύπους παραμέτρων, ViT backbone, SMPL-A head και discriminator. Τα πειράματα fine-tuning διεξάγονται σε δύο τομείς, στα δεδομένα που χρησιμοποιούνται και στις παραμέτρους που ανανεώνονται κάθε φορά.

Ως προς τα δεδομένα, αρχικά κάναμε fine-tuning χρησιμοποιώντας μόνο τα δικά μας δεδομένα. Τα δικά μας δεδομένα σε σύγκριση με τα δεδομένα του HMR2.0 περιέχουν αρκετά μεγάλο πλήθος περιπτώσεων παιδιών διαφόρων ηλικιών και βρεφών. Στα επόμενα πειράματα χρησιμοποιήσαμε ένα μείγμα των δικών μας δεδομένων και των δεδομένων του HMR2.0, δίνοντας μεγαλύτερη πιθανότητα να δοθεί δεδομένο εκπαίδευσης από τα δικά μας δεδομένα, ώστε να μπορέσει το μοντέλο να μάθει καλύτερα τα παιδιά και τα βρέφη.

Ως προς τα βάρη που ανανεώνονται, αρχικά έγινε fine-tuning σε όλα τα βάρη ταυτόχρονα. Επειδή δεν δούλεψε όπως περιμέναμε, κάναμε fine-tuning διαδοχικά σε κάθε τύπο παραμέτρων, κρατώντας παγωμένα τα υπόλοιπα βάρη. Τέλος, χρησιμοποιήσαμε και παγώσαμε το pre-trained ViT backbone πριν την εκπαίδευση από το HMR2.0, κάνοντας fine-tuning το SMPL-A head και τον discriminator.

Εφόσον το fine-tuning δεν μπόρεσε να δουλέψει, αποφασίσαμε να κάνουμε εκπαίδευση ενός νέου μοντέλου. Για να είμαστε σίγουροι ότι το μοντέλο εκπαιδεύεται κάναμε μία πρώτη εκπαίδευση χρησιμοποιώντας μόνο τα δικά μας δεδομένα. Δεδομένου ότι το μοντέλο έμαθε τα δεδομένα αυτά προχωρήσαμε σε εκπαίδευση χρησιμοποιώντας όλα τα δεδομένα, δηλαδή τα δικά μας και αυτά που χρησιμοποίησαν οι συγγραφείς του HMR2.0, για να έχουμε ένα μοντέλο που να δουλεύει το ίδιο καλά για ενήλικες αλλά και για παιδιά και βρέφη. Ομοίως με το fine-tuning, επειδή η εκτίμηση της πόζας παρουσίαζε κάποιες ανακρίβειες, χρησιμοποιήσαμε το pre-trained backbone και εκπαιδεύσαμε τα βάρη από τους άλλους δύο τύπους παραμέτρων, χωρίς ωστόσο και αυτό να έχει πολύ καλά αποτελέσματα.

Με βάση την αξιολόγηση των προηγούμενων πειραμάτων, υιοθετήσαμε μια υβριδική προσέγγιση εκπαίδευσης για το τελικό μας μοντέλο. Η διαδικασία που ακολουθήθηκε περιλαμβάνει τα ακόλουθα βήματα:

- **Αρχική Εκπαίδευση (Training from Scratch):** Για αρχή, εκπαιδεύσαμε ένα μοντέλο με την δικιά μας τροποποιημένη αρχιτεκτονική του HMR2.0 από την αρχή, χρησιμοποιώντας συνδυασμό των αρχικών δεδομένων εκπαίδευσης του HMR2.0 και των δικών μας συνόλων δεδομένων που περιέχουν περισσότερες εικόνες παιδιών και βρεφών.

- **Υβριδική Μοντελοποίηση και Fine-Tuning:** Παρατηρήθηκε ότι το μοντέλο που προέκυψε από την αρχική εκπαίδευση παρουσίαζε καλή εκτίμηση του $3\Delta$ σχήματος, αλλά υπολειπόταν στην ακρίβεια της πόζας.

Συγκριτική δοκιμή με το αρχικό HMR2.0 checkpoint σε εικόνες παιδιών επιβεβαίωσε την υπεροχή του HMR2.0 στην εκτίμηση της πόζας. Δεδομένου ότι το ViT backbone του HMR2.0 ήταν προεκπαιδευμένο στη 2Δ εκτίμηση keypoints, θεωρήσαμε ότι η χρήση του, σε συνδυασμό με το SMPL-A head και τον discriminator του δικού μας εκπαιδευμένου μοντέλου, θα βελτίωνε συνολικά την απόδοση. Η υπόθεσή μας επιβεβαιώθηκε. Για την τελική βέλτιστη ευθυγράμμιση των διαφορετικών παραμέτρων του υβριδικού μοντέλου, εφαρμόσαμε fine-tuning για μερικές επιπλέον εποχές, επιτυγχάνοντας τελικά τα βέλτιστα αποτελέσματα.

## 1.6 Αποτελέσματα

Στον Πίνακα 1.1 παρουσιάζονται τα αποτελέσματα της αξιολόγησης του μοντέλου μας στα datasets που χρησιμοποιήσαμε σε σύγκριση με 3 μοντέλα από τη βιβλιογραφία για το 3Δ σχήμα. Όπως παρατηρούμε **σχεδόν σε όλα τα datasets το μοντέλο μας επιτυγχάνει την καλύτερη επίδοση**. Συγκεκριμένα, στο SyRIP όπου περιέχει μόνο εικόνες μωρών παρατηρούμε ότι η διαφορά είναι πολύ μεγαλύτερη, κάτι που σημαίνει ότι το μοντέλο μας έχει βελτιώσει σημαντικά τα αποτελέσματα στις συγκεκριμένες ηλικίες. Στα Relative Human και ChildPlay επειδή υπάρχουν άτομα διαφορετικών ηλικιών αλλά και πολλά occlusions και truncations η διαφορά δεν είναι τόσο μεγάλη, είναι ωστόσο ενδεικτική της βελτίωσης των αποτελεσμάτων σε άτομα μικρών ηλικιών, αφού εκεί είναι η κύρια διαφορά των μοντέλων.

Table 1.1: Αξιολόγηση του μοντέλου μας με τις μετρικές AHD (m) και APHD (%). Μικρότερη απόλυτη τιμή δηλώνει καλύτερα αποτελέσματα.

| **Μέθοδος** | **SyRIP** | | **Relative Human** | | **ChildPlay** | | **BabyRobot** | |
|---|---|---|---|---|---|---|---|---|
| | **AHD ↓ (m)** | **APHD ↓ (%)** | **AHD ↓ (m)** | **APHD ↓ (%)** | **AHD ↓ (m)** | **APHD ↓ (%)** | **AHD ↓ (m)** | **APHD ↓ (%)** |
| ProHMR [30] | -1.011 | -161.76 | -0.098 | -17.73 | -0.477 | -92.01 | -0.562 | -56.69 |
| HMR2.0b [16] | -0.980 | -157.62 | -0.075 | -16.18 | -0.468 | -90.80 | -0.534 | -54.27 |
| BEV [62] | -0.528 | -91.8 | **-0.009** | -12.5 | **-0.067** | **-42.8** | -0.359 | -38.92 |
| **Το μοντέλο μας** | **-0.118** | **-25.97** | 0.088 | **-4.56** | -0.190 | -60.23 | **-0.354** | **-36.83** |

Στο BabyRobot η μέθοδος μας είναι και πάλι καλύτερη συγκριτικά με τις υπόλοιπες μεθόδους. Εδώ παρατηρείται μεγάλη ομοιότητα με τα αποτελέσματα του BEV, αφού η χρήση του SMPL-A μοντέλου βοηθάει το μοντέλο να εκτιμήσει καλύτερα το σχήμα των παιδιών.

Επίσης, όπως παρατηρούμε στον Πίνακα 1.2 **η μέθοδος μας παράγει το πιο λογικό μέσο ύψος**, με τις υπόλοιπες μεθόδους να δίνουν κατά βάση σώμα ενήλικα.

| **Μέθοδος** | **SyRIP** | **BabyRobot** |
|---|---|---|
| ProHMR | 1.712 | 1.718 |
| HMR2.0b | 1.705 | 1.717 |
| BEV | 1.249 | 1.528 |
| **Το μοντέλο μας** | **0.848** | **1.524** |

Table 1.2: Μέσο ύψος ανθρώπων σε μέτρα.

Στον Πίνακα 1.3 συγκρίνονται τα μοντέλα ως προς την 3Δ πόζα με τη μετρική MPJPE. Παρατηρούμε ότι **η μέθοδος μας είναι καλύτερη από το BEV**, δηλαδή τη μέθοδο που παράγει το πιο καλό 3Δ σχήμα για παιδιά και βρέφη. Επίσης, με εξαίρεση το SyRIP, **το μοντέλο μας έχει παρόμοια τιμή της μετρικής με το μοντέλο HMR2.0b**, αποδεικνύοντας την πολύ καλή ποιότητα των αποτελεσμάτων στην εκτίμηση της 3Δ πόζας.

Table 1.3: Αξιολόγηση 3Δ πόζας. Αξιολόγηση του μοντέλου με τη μετρική MPJPE (MPJPE σε mm). Μικρότερη τιμή ↓ υποδεικνύει καλύτερο μοντέλο.

| **Μέθοδος** | **SyRIP** | **ChildPlay** | **BabyRobot** |
|---|---|---|---|
| ProHMR [30] | 515.24 | 494.82 | 505.56 |
| BEV [62] | 452.73 | 424.41 | 380 |
| HMR2.0b [16] | **55.47** | **314.92** | **252.08** |
| **Το μοντέλο μας** | 287.84 | 318.26 | 258.51 |

Επιπρόσθετα, στον Πίνακα 1.4 παρουσιάζονται τα αποτελέσματα για την εκτίμηση της 2Δ πόζας, όπου **παρατηρείται καλύτερη εκτίμηση σε όλα τα dataset εκτός από το Relative Human για το μοντέλο μας συγκριτικά με το BEV**, τη μέθοδο, δηλαδή, που μπορεί να μοντελοποιήσει καλύτερα παιδιά και βρέφη μαζί με τους ενήλικες. Τα αποτελέσματά μας σε σύγκριση με το HMR2.0b υπολείπονται, ωστόσο βρίσκονται σε αρκετά υψηλά επίπεδα δείχνοντας την ικανότητα του μοντέλου μας και στη εκτίμηση της 2Δ πόζας.

Τέλος, για να πάρουμε μία γενικότερη ποιοτική αξιολόγηση των αποτελεσμάτων, διεξάγουμε μία έρευνα σε μορφή ερωτηματολογίου, όπου συγκρίνουμε το μοντέλο μας με το HMR2.0 και το BEV. Οι συμμετέχοντες καλούνται να απαντήσουν για 25 ζεύγη ανακατασκευών (10 συγκρίσεις του μοντέλου μας με το BEV, 10 συγκρίσεις του μοντέλου μας με το HMR2.0 και 5 για το BEV με το HMR2.0) τρεις ερωτήσεις, οι οποίες αφορούν την καλύτερη εκτίμηση 3Δ σχήματος, την καλύτερη εκτίμηση 3Δ πόζας και το καλύτερο συνολικό

Table 1.4: Αξιολόγηση 2Δ πόζας. PCK σκορ των 2Δ προβολών των keypoints σε διαφορετικές τιμές κατωφλίου. Υψηλότερο σκορ ↑ υποδεικνύει καλύτερο μοντέλο.

| **Μέθοδος** | **SyRIP** | | **Relative Human** | | **ChildPlay** | | **BabyRobot** | |
|---|---|---|---|---|---|---|---|---|
| | **@0.05** | **@0.1** | **@0.05** | **@0.1** | **@0.05** | **@0.1** | **@0.05** | **@0.1** |
| BEV [62] | 0.34 | 0.57 | 0.32 | 0.55 | 0.43 | 0.73 | 0.61 | 0.86 |
| HMR2.0b [16] | **0.79** | **0.98** | **0.48** | **0.62** | **0.76** | **0.94** | **0.97** | **0.99** |
| **Το μοντέλο μας** | 0.63 | 0.88 | 0.30 | 0.51 | 0.51 | 0.81 | 0.89 | 0.97 |

αποτέλεσμα που λαμβάνουν. Συνολικά έχουμε 39 εικόνες με τις ανακατασκευές τους, οπότε κάθε συμμετέχων λαμβάνει ένα τυχαίο σύνολο από 25 ζεύγη.

Στην έρευνα συμμετείχαν συνολικά 30 άνθρωποι με διαφορετικό ακαδημαϊκό υπόβαθρο και διαφορετικές ηλικίες ώστε να έχουμε μία πιο ολοκληρωμένη άποψη για την οπτική ποιοτική αξιολόγηση από τον ανθρώπινο παράγοντα. Τα αποτελέσματα δείχνουν **ξεκάθαρη υπεροχή του μοντέλου μας** συγκριτικά με τα άλλα δύο μοντέλα συνολικά, αλλά και σε κάθε κατηγορία ξεχωριστά. Συγκεκριμένα, **περίπου το 75% των απαντήσεων συμφώνησε ότι το μοντέλο μας παράγει καλύτερα αποτελέσματα συγκριτικά με το BEV για κάθε κατηγορία**. Στη σύγκριση με το HMR2.0 η διαφορά είναι μικρότερη, αφού λαμβάνουμε λίγο περισσότερο από το 50% των ψήφων, με τη διαφορά ότι εδώ υπάρχουν περισσότεροι αναποφάσιστοι. Αυτός είναι ένας ακόμα δείκτης της πολύ καλής επίδοσης του μοντέλου μας, αφού το HMR2.0 θεωρείται ένα από τα καλύτερα μοντέλα στην εκτίμηση του 3Δ σχήματος και της πόζας. Συνολικά, **το μοντέλο μας λαμβάνει περισσότερο από το 60% των ψήφων αν συγκεντρώσουμε τις ψήφους από κάθε κατηγορία**, δείχνοντας ότι η αξιολόγηση από τους ανθρώπους θεωρεί πλειοψηφικά το μοντέλο μας πιο αξιόπιστο στην εκτίμηση του 3Δ σχήματος και της πόζας για παιδιατρικό πληθυσμό. Τα αναλυτικά αποτελέσματα για κάθε κατηγορία σύγκρισης παρουσιάζονται στο αγγλικό κείμενο στους Πίνακες 7.5 και 7.6.

Μερικά παραδείγματα από τα αποτελέσματα των μεθόδων παρουσιάζονται στις εικόνες 1.3, για τη μέθοδο βελτιστοποίησης, και 1.4, για το προτεινόμενο μοντέλο. Συγκεκριμένα για το προτεινόμενο μοντέλο γίνεται και μία σύγκριση των ποιοτικών αποτελεσμάτων του με τα αντίστοιχα αποτελέσματα των HMR2.0 και BEV. Όπως παρατηρείται, η μέθοδος βελτιστοποίησης παράγει ρεαλιστικές και υψηλής ακρίβειας 3Δ ανακατασκευές, τόσο ως προς το σχήμα όσο και ως προς τη πόζα, ενώ τα αποτελέσματα του προτεινόμενου μοντέλου μας είναι τα πιο ρεαλιστικά σε σύγκριση με το HMR2.0 και το BEV, αφού το σώμα που παράγεται όταν η εικόνα περιέχει βρέφος είναι πράγματι βρέφους, με ακριβή 3Δ πόζα. Για περισσότερα ποιοτικά αποτελέσματα σε βρέφη, παιδιά

Figure 1.3: Παραδείγματα της μεθόδου βελτιστοποίησής μας σε άτομα διαφορετικής ηλικίας.

μεγαλύτερης ηλικίας αλλά και ενήλικες παραπέμπουμε τον αναγνώστη στο αγγλικό κείμενο στο Κεφάλαιο 7.

## 1.7 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Η παρούσα διπλωματική εργασία αντιμετωπίζει το θεμελιώδες πρόβλημα της 3Δ εκτίμησης του σχήματος και της πόζας του ανθρώπου, εστιάζοντας στη γενίκευση των μεθόδων για μη-ενήλικους πληθυσμούς. Οι σύγχρονες τεχνικές δυσκολεύονται να γενικεύσουν σε εικόνες παιδιών και βρεφών, λόγω της εξάρτησής τους από γεωμετρικά πρότυπα βασισμένα σε ενήλικες. Για να γεφυρώσουμε αυτό το χάσμα, προτείνουμε μία καινοτόμο προσέγγιση. Συγκεκριμένα, προτείνουμε δύο μεθόδους για την εκτίμηση του 3Δ σχήματος και

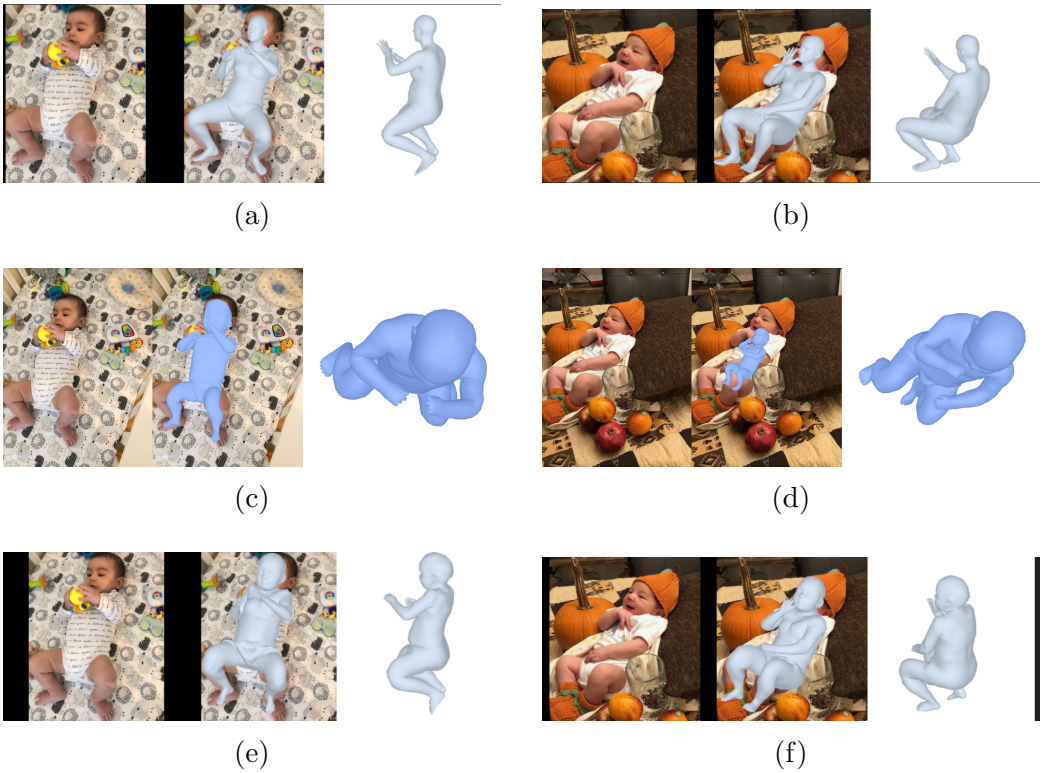


Figure 1.4: Παραδείγματα του προτεινόμενου μοντέλου ((e),(f)) σε σύγκριση με το HMR2.0b ((a),(b)) και το BEV ((c),(d)).

πόζας ανθρώπων από μία εικόνα για ανθρώπους κάθε ηλικίας. Επεκτείνουμε σύγχρονες μεθόδους που λειτουργούν αποτελεσματικά για ενήλικα άτομα ώστε να είναι περισσότερο αποτελεσματικά και για άτομα μικρότερης ηλικίας, δηλαδή παιδιά και βρέφη. Η πρώτη μέθοδος βασίζεται σε βελτιστοποίηση, ενώ η δεύτερη είναι μέθοδος βαθιάς μάθησης. Λόγω της έλλειψης δεδομένων για την εκπαίδευση του μοντέλου μας, ώστε να μάθει να κάνει εκτίμηση για το σχήμα και την πόζα σε εικόνες παιδιών, χρησιμοποιούμε τη μέθοδο βελτιστοποίησης για να παράξουμε ψευδο-πραγματικές επισημειώσεις σε εικόνες από δημοσίως διαθέσιμα σχετικά σύνολα δεδομένων. Οι δύο αυτές μέθοδοι που αναπτύσσουμε είναι ικανές να μοντελοποιήσουν άτομα κάθε ηλικίας επιτυχώς. Επίσης, οι μέθοδοι μας μπορούν να χρησιμοποιηθούν για την ανωνυμοποίηση ευαίσθητων δεδομένων, αφού οι ακριβείς 3Δ ανακατασκευές που δημιουργούν μπορούν να αντικαταστήσουν τον άνθρωπο και να αποκρύψουν στοιχεία για την ταυτότητά του. Μια εφαρμογή αυτού, είναι στο BabyRobot dataset, όπου με τη χρήση των 3Δ ανακατασκευών μπορεί πλέον να μετατραπεί από μια κλειστή πλούσια βάση με αλληλεπιδράσεις παιδιών με ρομπότ σε μια ανοιχτή βάση ανωνυμοποιη-

μένη ως προς τα φυσιολογικά χαρακτηριστικά του σώματος ενός παιδιού που αλληλεπιδρά με ρομπότ μέσω δράσεων και κινήσεων.

Οι μέθοδοι μας αν και αποτελεσματικές στην εκτίμηση του 3Δ σχήματος και πόζας των ανθρώπων χρήζουν περαιτέρω βελτίωσης. Αρχικά, τα αντικειμενικά προβλήματα της αβεβαιότητας της πόζας λόγω φυσικών ή ανθρώπινων εμποδίων και του βάθους της σκηνής παραμένουν, και από μία μόνο εικόνα συνήθως αυτά δεν είναι πλήρως αντιμετωπίσιμα. Εικόνες από διαφορετικές όψεις, βίντεο ή κάμερες βάθους RGB-D μπορούν να συντελέσουν στην επίλυση αυτών των προβλημάτων. Επίσης, η ικανότητα γενίκευσης του μοντέλου SMPL-A παραμένει περιορισμένη σε άτομα μικρής ηλικίας. Αυτό οφείλεται στις φυσικές διαφορές ενός σώματος ενήλικα και ενός μωρού ή παιδιού οι οποίες μεταφέρονται και στα τεχνητά μοντέλα και δεν είναι εύκολο να μοντελοποιηθούν με ένα μοντέλο. Συγκεκριμένα, το SMPL-A χρησιμοποιεί για όλα τα άτομα ένα ενιαίο shape space, αυτό του ενήλικα, το οποίο έχει ως αποτέλεσμα μερικές φορές η ανακατασκευή να είναι παράλογη. Η ανάπτυξη πιο δυνατών μοντέλων περιγραφής του ανθρώπου θα μπορούσε να λύσει τέτοια προβλήματα, όπως και η μη παραμετρική μοντελοποίηση, η οποία, ωστόσο, προϋποθέτει την ύπαρξη σχετικών δεδομένων εκπαίδευσης. Τα δεδομένα εκπαίδευσης, ειδικά για παιδιά, παραμένουν ίσως το μεγαλύτερο εμπόδιο στην πρόοδο του πεδίου. Η ευαισθησία αυτών των δεδομένων απαιτεί αυστηρή νομοθεσία και δεοντολογία, καθιστώντας την συλλογή τους μια διαρκή πρόκληση.

# Chapter 2

# Introduction

Three-dimensional (3D) pose and shape estimation is a fundamental challenge in computer vision. It aims to reconstruct a person's detailed 3D body model, including both their posture and shape, from various inputs like a single 2D image, a video, or data from multi-view cameras. This task is inherently complex due to its ill-posed nature, as a significant amount of depth information is lost when a 3D scene is projected onto a 2D plane. Despite this difficulty, remarkable progress has been made, particularly for adult subjects, by leveraging large-scale datasets and sophisticated deep learning architectures. These advancements have enabled a wide range of applications, including augmented reality, human-computer interaction, and medical motion analysis.

Modern approaches to 3D pose and shape estimation generally fall into two primary categories:

- **Optimization-based methods:** These techniques iteratively refine the parameters of a body model to minimize the discrepancy between the projected 3D model and observable features in the 2D image, such as joint locations or silhouettes. While powerful and robust, these methods can be computationally expensive.

- **Regression-based (Deep Learning) methods:** These models, typically neural networks, are trained to directly predict a body model's parameters from an input image. They offer fast inference speeds, but their effectiveness is heavily dependent on the availability of large-scale datasets with 3D ground truth for training.

The recent surge in performance within computer vision is largely attributable to the advancements in deep learning. Architectures such as Convolutional Neural Networks (CNNs) have proven exceptionally effective at

extracting complex features from images. More recently, the integration of attention mechanisms and transformer-based models [66], originally developed for natural language processing, has further enhanced the ability of these systems to understand spatial relationships and context within an image. This progress, however, is fundamentally driven by data; the power of these models lies in their ability to learn from millions of labeled examples, a prerequisite that becomes the central challenge when addressing sensitive and difficult-to-acquire data for specific populations.

## 2.1 Challenges in 3D Shape and Pose Estimation

While these methods have been highly successful for adults, extending them to younger populations, such as children and babies, remains a significant and largely unresolved challenge. The core obstacles are both ethical and technical, creating a unique research gap.

**Data Scarcity and Ethical Constraints:** The most significant obstacle is the fundamental scarcity of high-quality data. Existing human body models, like SMPL [36], are built from thousands of 3D scans of adults. The strict legislation protecting minors, coupled with the ethical complexities of acquiring such data, makes it exceptionally difficult to obtain the necessary datasets for a child-specific body model. This data sparsity creates a considerable "domain gap" that renders adult-based models fundamentally ill-suited for accurate pose and shape estimation in children.

**Anthropometric Discrepancy:** The proportions of the human body change dramatically from infancy through childhood to adulthood. For instance, an infant's head is disproportionately large, and their limbs are shorter relative to their torso. These anthropometric variations, which are crucial for accurate modeling, are not captured by models trained exclusively on adult data.

**Occlusions and Pose Ambiguity:** Occlusions are a crucial problem in 3D shape and pose estimation. Natural obstacles or strange poses can hide parts of the human body, creating ambiguities. When some parts of the body are missing, a computer, like a human, may perceive the shape and pose in multiple ways. This problem is particularly challenging in children and infants due to their unpredictable movements, which makes the collection of high-quality, occlusion-free data almost impossible.

**Privacy and Security Concerns:** Reconstructing a child's 3D shape and pose, especially from images of their face or identifiable features, raises

significant privacy and security concerns. Future solutions must not only be accurate but also incorporate robust methods for de-identification and data protection to safeguard the subjects.

These unique challenges underscore the need for a dedicated approach to pediatric 3D shape and pose estimation.

## 2.2 Motivation and Applications

Despite these challenges, the ability to accurately model the 3D shape and pose of children has critical and transformative applications in pediatrics and healthcare. This technology could serve as a non-invasive, objective tool for developmental assessments, tracking motor skills and physical growth. For example, clinicians could use it to quantify a child's gait, a key indicator of neurological and musculoskeletal health, without needing intrusive equipment. This technology could also facilitate the early detection of musculoskeletal disorders like scoliosis and assist in diagnosing conditions like autism by analyzing a child's posture and movement patterns [15, 27]. These applications underscore the profound clinical impact and potential of this research to improve children's health outcomes globally.

For the generic applications of 3D shape and pose estimation of humans, it can help with the task of tracking humans as well as the task of action recognition. For example, in sports analysis, coaches can use 3D pose estimation to meticulously analyze an athlete's form and technique, providing objective data to optimize performance and prevent injuries. This capability extends to fields like virtual reality (VR) and augmented reality (AR), where realistic human avatars are created and animated in real-time [56].

## 2.3 Contributions

The contributions of this thesis address the critical data and methodological limitations in 3D shape and pose estimation for non-adult subjects. Our primary contributions are summarized in the following points:

- **Specialized Optimization Technique:** We successfully adapt and optimize an existing optimization-based technique for 3D shape and pose estimation, specifically to **enable robust and accurate reconstruction for infants and young children.**
- **Novel Deep Learning Model for Non-Adults:** We introduce a specialized, highly accurate HMR-like deep learning model for 3D shape

and pose estimation. This model is engineered by successfully integrating a pre-trained backbone with a customized SMPL-A head, achieving superior performance on child and infant imagery.

- **Public Dataset Release and Data Anonymization:** We release the BabyRobot dataset, a new resource comprising the 3D reconstructions of children interacting with robots. This release demonstrates a methodology for the ethical handling and anonymization of sensitive child imagery, as the shared 3D body representation replaces potentially identifying photographic data.

## 2.4 Thesis Structure

The remainder of this thesis is structured as follows: Chapter 3 analyzes the theoretical background necessary to explain the proposed methodology. Chapter 4 provides a comprehensive review of existing literature. Chapter 5 details the proposed methodology. Chapter 6 presents and analyzes the experiments that were performed, while Chapter 7 shows the results of the evaluation. Finally, Chapter 8 concludes the work and discusses potential avenues for future research.

# Chapter 3

# Theoretical Background

## 3.1 Fundamentals of 3D Representation

The world that we live in, as well as most of the objects around us, exists in three dimensions. The same applies for humans. Thus, in order to describe the world and the objects it contains, we need a three-dimensional (3D) coordinate system.

The most common coordinate system in the world is the **3D Cartesian coordinate system**, where each point $\mathbf{p} \in \mathbb{R}^3$ is defined by a tuple of three numbers $(x, y, z)$. The three planes of the coordinate system are pairwise perpendicular.

However, Cartesian coordinates are not the only option. Other coordinate systems are often used depending on the application:

- **Cylindrical coordinates** $(r, \phi, z)$: Used when there is rotational symmetry around an axis. A point is defined by its distance $r$ from the axis, the angle $\phi$ around the axis, and the height $z$.

- **Spherical coordinates** $(\rho, \theta, \phi)$: Useful for representing points on spheres or in 3D space with radial symmetry. A point is defined by its distance $\rho$ from the origin, the inclination angle $\theta$, and the azimuthal angle $\phi$. Figure 3.1 illustrates the way Cartesian, Cylindrical, and Spherical Coordinates are calculated.

- **Barycentric coordinates**: Often used in computer graphics for representing positions inside triangles or tetrahedra, useful for interpolation.

- **Homogeneous coordinates**: Homogeneous coordinates introduce an extra dimension to represent a point $\mathbf{X} = [X, Y, Z]^T$ that belongs in

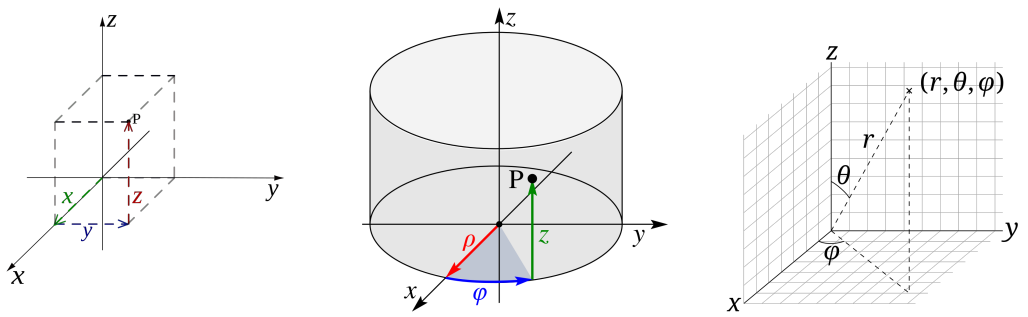
an Euclidean 3D space as $\tilde{\mathbf{X}} = (X_1, X_2, X_3, X_4)$ in homogeneous coordinates, where $X_4 \neq 0$ is an arbitrary normalization parameter. To transform a vector from 4D to 3D we use the following equation:

$$(X, Y, Z) = \left(\frac{X_1}{X_4}, \frac{X_2}{X_4}, \frac{X_3}{X_4}\right)$$

In an analogous way a point in the 2D image plane $\mathbf{x} = [x, y]^T$ can be expressed as $\tilde{\mathbf{x}} = [x_1, x_2, x_3]^T$ with $x_3 \neq 0$ and

$$(x, y) = \left(\frac{x_1}{x_3}, \frac{x_2}{x_3}\right)$$

Typically, $X_4 = 1$ for points in space and $X_4 = 0$ for points at infinity. This representation is widely used in computer graphics and computer vision because it allows translations, rotations, scaling, and perspective projections to be expressed as matrix multiplications.




(a) Cartesian Coordinates (b) Cylindrical Coordinates (c) Spherical Coordinates

Figure 3.1: Three Main Coordinate Systems. Figures from [68]

Each coordinate system has its advantages depending on whether the problem involves symmetry, rotations, projections, or transformations.

While cylindrical, spherical, and similar systems describe alternative parameterizations of space, in computer vision we mainly use Cartesian coordinates organized in different reference frames, such as the World and Camera coordinate systems.

**World Coordinate System** ($\mathcal{C}_W$) The coordinate system that defines the position and the orientation of an object in the real world is called the *World Coordinate System*. It is an arbitrary global reference frame and is typically denoted by axes $(X_W, Y_W, Z_W)$ and a center $O_W$ where the three axes intersect.

**Camera Coordinate System ($\mathcal{C}_C$)** The *Camera Coordinate System* is a local reference frame attached to the camera itself. It is typically used to describe the position of 3D points relative to the camera after applying the extrinsic transformation (rotation and translation) from the world coordinate system.

In this system, the origin $O_C$ is located at the optical center of the camera (also called the pinhole), and the axes are defined as follows:

- The $Z_C$-axis points forward along the camera's optical axis (towards the scene).
- The $X_C$-axis is horizontal and usually points to the right in the image plane.
- The $Y_C$-axis is vertical and points downward in the image plane (following the image coordinate convention).

Any 3D point expressed in the world coordinate system $\mathbf{P}_W = (X_W, Y_W, Z_W)^T$ can be transformed to the camera coordinate system $\mathbf{P}_C = (X_C, Y_C, Z_C)^T$ using an extrinsic transformation:

$$\mathbf{P}_C = \mathbf{R}\,\mathbf{P}_W + \mathbf{t}$$

where $\mathbf{R} \in SO(3)$ is a $3 \times 3$ rotation matrix that aligns the world axes to the camera axes, and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector that represents the camera's position in the world.

**Image Coordinate System ($\mathcal{C}_I$) and Perspective Projection** The points of the 3D Camera Coordinate System, when they are projected onto a 2D plane of the camera, belong in a 2D coordinate system that is called *Image Coordinate System* and is usually normal to the z-axis of the camera coordinate system. The projection from the 3D $\mathcal{C}_\mathcal{C}$ system to the $\mathcal{C}_I$ is commonly the perspective projection, and is described with the pinhole model. The $X$ and $Y$ coordinates of the camera system are projected into the 2D plane, which is at a distance $f$ (focal length) from the camera coordinate system. The $x$ and $y$ coordinates of the projection in the image plane can be found using the law of similar triangles as follows:

$$x = f\frac{X_C}{Z_C} \quad , \quad y = f\frac{Y_C}{Z_C} \tag{3.1}$$

Figure 3.2 shows an example of how a point of a surface is projected to a point in the image plane using the perspective projection.

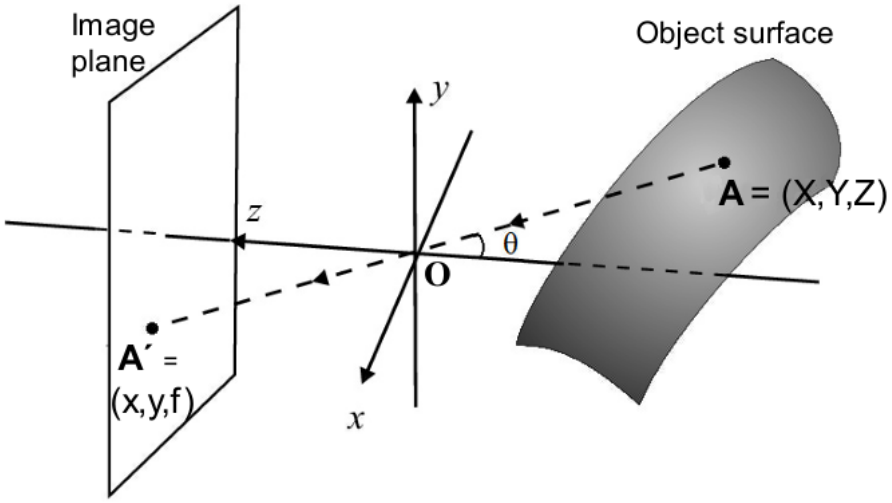



Figure 3.2: Perspective Projection: A point **A** in the camera coordinate system is projected using the pinhole camera model to the point $\mathbf{A}'$ in the image plane. The image plane is at a distance $f$ (focal length) from the camera coordinate system. Figure from [40]

The **pinhole camera model** is a simplified, idealized representation of a camera. It is a mathematical model based on the concept of a camera obscura, a light-proof box with a tiny hole, or pinhole, on one side. All light from a scene passes through this single point, projecting an inverted image onto a sensor plane on the opposite side. The model assumes an infinitely small pinhole, which prevents any blurring and gives the image an infinite depth of field. Because all light rays from a 3D point converge at the pinhole before hitting the image plane, we can use simple geometric principles, like the law of similar triangles, to derive the relationship between a 3D point in the camera's coordinate system and its corresponding 2D projection on the image plane. An illustration of the pinhole camera model is shown in Figure 3.3.

The equation 3.1 is non-linear between the world and image coordinates, something that makes it difficult to analyze. However, using homogeneous coordinates, these equations can be written in a linear form. The perspective projection of 3D world points $\mathbf{x}_W = \begin{bmatrix} X, Y, Z \end{bmatrix}$ to 2D image points $\mathbf{x}_I = \begin{bmatrix} x, y \end{bmatrix}$ can be expressed as a transformation from 4D homogeneous world coordinates with $X_4 = 1$ to 3D homogeneous image coordinates. More formally, using matrices:

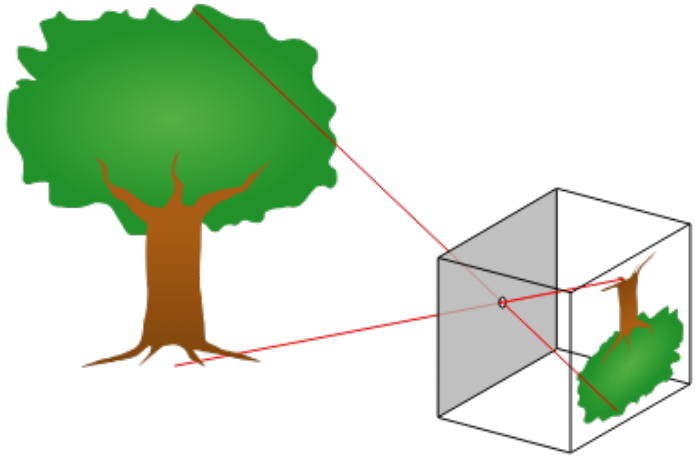


Figure 3.3: The Pinhole Camera Model. Figure from [9]

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \mapsto \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

where $\mathcal{P} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ is the projection matrix.

Solving the above system in terms of $(x_1, x_2, x_3)$ and changing the homogeneous coordinates to Cartesian we have:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{x_1}{x_3} \\ \frac{x_2}{x_3} \end{bmatrix} = \begin{bmatrix} f\frac{X}{Z} \\ f\frac{Y}{Z} \end{bmatrix}$$

## 3.2 Machine Learning

Machine learning, a subfield of artificial intelligence, addresses the challenge of enabling computers to learn from data without being explicitly programmed for every possible scenario. The core objective is to develop algorithms that train computational models to identify underlying relationships and patterns within a given dataset. Once trained, these models can make predictions, classify new data, or make decisions. As a result, machine learning has become a foundational technology across numerous domains, from

natural language processing and computer vision to healthcare and finance.

### 3.2.1 Types of Machine Learning

Machine learning algorithms are typically divided into two main categories:

- **Supervised Learning**: In this type of learning, the data used have labels, meaning we know the correct output for each input. The model's goal is to learn a mapping from inputs to outputs. The most common types of problems are *classification*, which outputs a class (e.g., spam or not spam), and *regression*, which outputs a continuous value (e.g., a house price).

- **Unsupervised Learning**: Unsupervised learning handles data that do not have labels. The algorithm's objective is to discover hidden structures or patterns within the data. The most common unsupervised learning problem is *clustering*, where the algorithm groups similar data points into "clusters".

### 3.2.2 Data Subsets

As mentioned previously, the training of a machine learning model requires data that are similar to those that the model will be used on after its training. These data are divided into three essential subsets to ensure accurate evaluation and to prevent overfitting:

- **Training Set**: This is the largest subset of the dataset and is used to train the model, allowing it to learn the underlying patterns and relationships in the data.

- **Validation Set**: This subset is used during the training process to tune hyperparameters and evaluate the model's performance on unseen data. It helps in preventing overfitting and provides a basis for making adjustments to the model's architecture.

- **Test Set**: This set is used only after the training is complete to provide an unbiased final evaluation of the model's performance and its ability to generalize to new, unseen data. It is crucial that the model has not seen any data from the test set during its training or validation phases.

## 3.3 Deep Learning

In recent years, the field of artificial intelligence has been revolutionized by deep learning, a powerful class of machine learning methods that enables computational models to discover intricate representations within vast amounts of data. Unlike traditional machine learning approaches that often rely on manual feature engineering, deep learning models autonomously learn hierarchical features from raw data, such as images or text. This paradigm shift has been made possible by a convergence of factors, including the availability of large-scale datasets, significant advancements in computational power (particularly with GPUs), and the development of sophisticated algorithms. This has led to breakthrough performance in a wide array of domains, from natural language processing to computer vision.

The field of Deep Learning is built upon computational models that are designed to learn and represent data through multiple levels of abstraction. The cornerstone of these models is the artificial neural network (ANN), a concept that has revolutionized computer vision and is central to the approach taken in this thesis.

### 3.3.1 Neural Networks

A neural network, or artificial neural network (ANN), is a computational model inspired by the structure and function of the human brain. At its core, an ANN consists of a collection of interconnected processing units called neurons, which are organized into distinct layers: an input layer, one or more hidden layers, and an output layer. Each neuron in a given layer is connected to all neurons in the subsequent layer, with each connection assigned a numerical value known as a weight ($w$). Data is fed into the input layer and is processed sequentially as it passes through the network. The output of each neuron is determined by a weighted sum of its inputs, which is then passed through a non-linear activation function ($f$) before being transmitted to the next layer. The network's ability to learn complex patterns and relationships from data is achieved through an iterative training process, where the weights ($w_{ij}$) are systematically adjusted to minimize the difference between the network's predicted output and the true output, a process often guided by a loss function ($\mathcal{L}$) and an optimization algorithm such as backpropagation [57].

### 3.3.2 Feedforward Neural Networks (FNN)

A **Feedforward Neural Network (FNN)** is a foundational type of artificial neural network where information flows exclusively in one direction,

from the input layer, through one or more hidden layers, and to the output layer. The name "feedforward" reflects this unidirectional flow, as there are no loops or feedback connections. The network learns by adjusting the **weights** and **biases** associated with the connections between neurons.

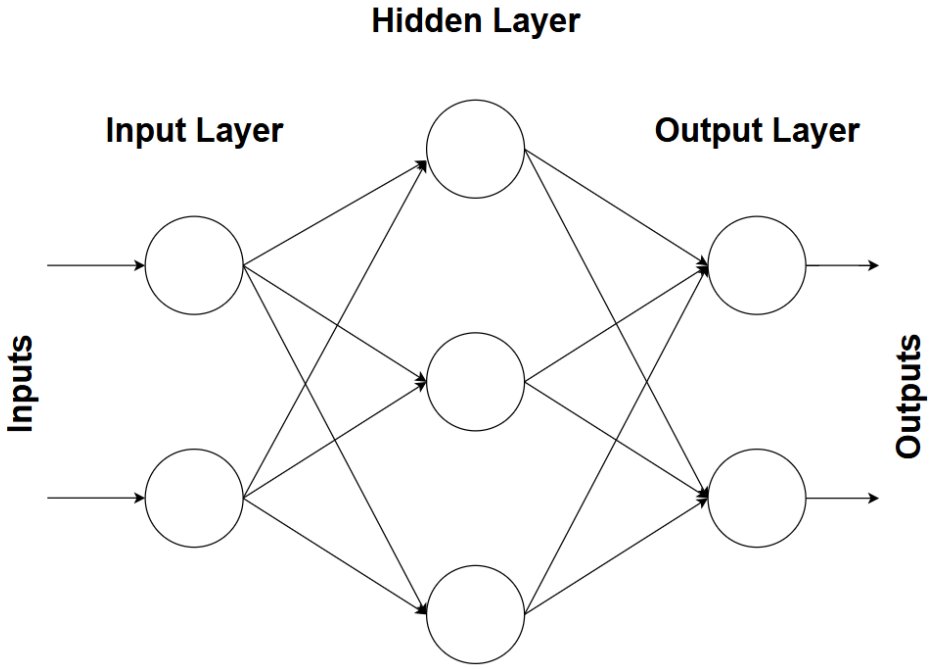



Figure 3.4: A simple Feedforward Neural Network with an input layer with 2 inputs, 1 hidden layer with 3 hidden states and an output layer with 2 outputs.

### The structure of an FNN

The structure of an FNN, as shown in Figure 3.4, is straightforward, containing three distinct types of layers:

- **The Input Layer:** This layer receives the raw input data. Each neuron in this layer typically corresponds to a single feature of the input data.
- **The Hidden Layers:** Positioned between the input and output layers, these layers are responsible for learning complex patterns and relationships within the data. The number of hidden layers and neurons within them can vary depending on the complexity of the problem.

- **The Output Layer:** This layer produces the final result of the network. The number of neurons here is determined by the task; for example, it equals the number of classes in a classification problem or the number of outputs in a regression problem.

### Training an FNN

Training an **FNN** is the process of teaching it to map specific inputs to desired outputs by adjusting its internal parameters—the **weights** and **biases**—to minimize the error between the network's output and the expected output. This error is quantified by a **loss function**. The training process is an iterative cycle consisting of three main steps:

1. **Forward Propagation**: The input data is passed through the network, layer by layer, to produce an output.

2. **Loss Calculation**: The loss function measures the discrepancy between the network's prediction and the ground-truth data.

3. **Backpropagation and Parameter Update**: The error is propagated backward through the network to calculate the gradient of the loss with respect to each weight and bias. An **optimizer** (e.g., Stochastic Gradient Descent) then uses these gradients to update the parameters, gradually reducing the loss.

This cycle is repeated for many iterations, or **epochs**, until the network's performance is satisfactory.

### Activation Functions

**Activation functions** are mathematical operations applied to the output of each neuron. Their primary role is to introduce **non-linearity** into the network, enabling it to learn complex, non-linear relationships in the data. Without them, a neural network, no matter its depth, would only be capable of learning linear functions. Common activation functions include:

- **Sigmoid**: Squashes values between 0 and 1, often used for binary classification output layers.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- **ReLU (Rectified Linear Unit)**: Outputs the input for positive values and zero otherwise. It is widely used in hidden layers due to its computational efficiency.

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

- **Softmax**: Converts a vector of numbers into a vector of probabilities, used in the output layer for multi-class classification.

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

  where $z_i$ is the logit (the output of the previous layer in the network) for the $i$-th class and $K$ is the number of classes.

### Challenges

The training of a neural network presents several challenges, one of the most significant being the training duration. A model trained for a duration greater or less than the optimal number of epochs may face problems with overfitting or underfitting, respectively.

**Overfitting** Overfitting occurs when a model is trained for too long. This leads to poor generalization, which is the ability to make accurate predictions on unseen data. The primary effect of overfitting is that the model learns the training data too well, memorizing noise and specific examples rather than learning the underlying patterns. As a result, its performance on new, unknown data is poor.

**Underfitting** Conversely, underfitting is the opposite problem, occurring when a model has not been trained for a sufficient number of epochs. In this case, the model fails to learn the fundamental patterns in the training data, as it hasn't seen enough examples. Consequently, the model cannot generalize well to either the training data or unseen data.

A simple example of the overfitting and underfitting problems in a classification task is illustrated in Figure 3.5. The two classes are the "blue" and the "red" points.

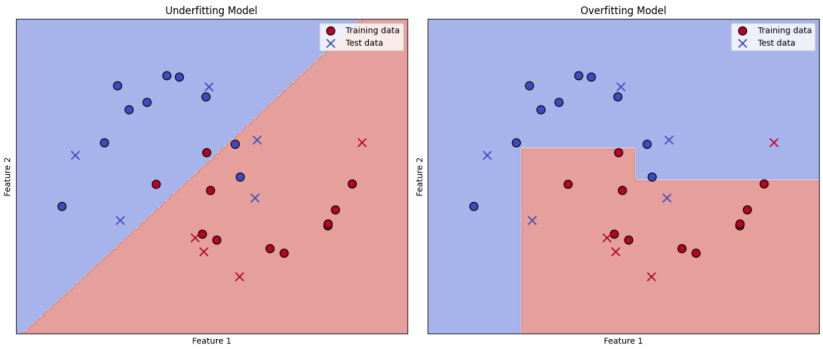



Figure 3.5: A simple illustration of overfitting and underfitting in a classification task.

### 3.3.3 Convolutional Neural Networks (CNN)

A **convolutional neural network (CNN)** is a type of deep neural network that has revolutionized the field of computer vision [31]. Unlike traditional neural networks, its architecture is based on the convolution operation, which allows it to automatically and efficiently learn spatial hierarchies of features from data. While predominantly used for image analysis, CNNs are versatile and can be applied to other types of data, such as text, audio, and video.

A CNN's architecture is built upon a series of specialized hidden layers that process input data. These layers include one or more **convolutional layers**, each of which uses a learnable kernel (or filter). This kernel slides over the input data, performing a dot product to produce a new representation called a feature map. This process allows the network to detect features such as edges, textures, and patterns.

After each convolutional layer, a non-linear activation function (most commonly **ReLU**, or **sigmoid**) is applied to the feature map to introduce non-linearity. The hidden layers can also include pooling layers, which reduce the spatial dimensions of the feature maps, thereby decreasing the computational load and making the network more robust to minor variations in feature location. Other common layers include normalization layers and fully connected layers. Before the data is processed by the final fully connected layers, a flattening layer is applied to transform the multi-dimensional feature map into a one-dimensional vector, as required by the input of a fully connected network [39].

More formally, the convolution operation involves a kernel $K$ with dimensions $f \times f$ that slides over the input image $I$. The feature map $F$, is calculated as the sum of the element-wise products of the kernel and the

corresponding portion of the input image.

For a 2D input image $I$ and a 2D kernel $K$, the convolution operation is formally given by:

$$(I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n)$$

In practice, a cross-correlation operation is typically used for computational efficiency:

$$F(i, j) = \sum_{m=0}^{f-1} \sum_{n=0}^{f-1} I(i + m, j + n)K(m, n)$$

The spatial dimensions of the output feature map, $F_{\text{size}}$, are determined by the input size $I_{\text{size}}$, kernel size $K_{\text{size}}$, stride $S$ (the number of pixels the kernel shifts), and padding $P$ (adding zeros around the input border), as follows:

$$F_{\text{size}} = \frac{I_{\text{size}} - K_{\text{size}} + 2P}{S} + 1$$

Because **Max Pooling** is the most common type of pooling in a CNN, we present the operation of this pooling layer in more detail. It selects the maximum value from a small window (e.g., $2 \times 2$) of the feature map to represent that entire region. The operation is defined by:

$$F_{\text{pooled}}(i, j) = \max_{m,n \in W} F(iS + m, jS + n)$$

where $W$ is the pooling window size and $S$ is the stride. The operations for average pooling and min pooling are analogous, simply replacing the max function with average or min, respectively.

### 3.3.4 Transformer Network and Attention Mechanism

The **Transformer architecture**, introduced by Vaswani et al. [66], marked a significant paradigm shift in deep learning, particularly within the field of Natural Language Processing (NLP). Unlike previous sequential models like CNNs, the Transformer relies entirely on the attention mechanism to capture long-range dependencies within a sequence. This parallelizable design allows the transformer to model complex relationships between words in a highly efficient manner [23].

The core innovation of the Transformer is the **self-attention mechanism**, which enables the model to weigh the importance of all other words in

a sentence when encoding a specific word. This mechanism is defined by three learned linear projections of the input embedding $\mathbf{x}$: a **query** ($\mathbf{q}$), a **key** ($\mathbf{k}$), and a **value** ($\mathbf{v}$). These vectors are derived from the input embedding $\mathbf{x}_i$ by multiplying it with learned weight matrices $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$:

$$\mathbf{q}_i = \mathbf{x}_i\mathbf{W}^Q, \quad \mathbf{k}_i = \mathbf{x}_i\mathbf{W}^K, \quad \mathbf{v}_i = \mathbf{x}_i\mathbf{W}^V$$

The attention score between token $i$ and token $j$ is calculated as the dot product of the query vector of token $i$ and the key vector of token $j$, scaled by the square root of the dimension of the key vectors, $d_k$, to stabilize gradients:

$$\text{score}(\mathbf{q}_i, \mathbf{k}_j) = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}}$$

These scores are then normalized using a softmax function to obtain the attention weights, $\alpha_{ij}$, which represent the probability distribution of attention from token $i$ to all other tokens in the sequence:

$$\alpha_{ij} = \text{softmax}(\text{score}(\mathbf{q}_i, \mathbf{k}_j)) = \frac{\exp(\text{score}(\mathbf{q}_i, \mathbf{k}_j))}{\sum_{k=1}^{N} \exp(\text{score}(\mathbf{q}_i, \mathbf{k}_k))}$$

Finally, the output of the self-attention layer for token $i$, $\mathbf{a}_i$, is a weighted sum of all the value vectors, where the weights are the computed attention scores:

$$\mathbf{a}_i = \sum_{j=1}^{N} \alpha_{ij}\mathbf{v}_j$$

It is worth noting that for tasks like language generation, the summation is over the preceding tokens ($j \leq i$), which is known as **masked attention**.

### Multi-head Attention

To exploit a wider range of information, the Transformer architecture uses **multi-head attention**. This mechanism performs the self-attention calculation in parallel $h$ times, with each "head" using its own set of learned Q, K, and V weight matrices. Each head can learn a distinct set of relationships, allowing the model to attend to information from different representation subspaces jointly. The outputs of these parallel heads are then concatenated and projected back to the original dimension.

#### Transformer Blocks

The complete Transformer model is built from stacked layers called **Transformer blocks**. Beyond the self-attention mechanism, a standard Transformer block includes a fully-connected (FFN), **residual connections** [18] around both the self-attention and FFN sub-layers, and **layer normalization** [3]. The residual connections and layer normalization are crucial for enabling the training of deep networks and preventing vanishing gradients, ensuring information flows effectively through the model. The FFN, which is applied independently to each position, provides the model with additional capacity to process the output of the attention mechanism.

### 3.3.5 Vision Transformer (ViT)

The Vision Transformer (ViT) [12] is a deep learning architecture for image classification tasks that applies the Transformer model to sequences of image patches. Unlike traditional CNNs, ViT replaces convolutional layers with self-attention mechanisms, enabling the model to capture long-range dependencies across the entire image.

An image is first split into fixed-size non-overlapping patches (e.g., $16 \times 16$ pixels), and each patch is linearly projected into an embedding vector. These patch embeddings are then concatenated with positional encodings that provide spatial information about the location of each patch in the original image.

The resulting sequence of embeddings is passed through a standard Transformer encoder composed of multi-head self-attention layers, layer normalization, and residual connections. A learnable [CLS] token[1] is often prepended to the input sequence and is used for classification tasks after the Transformer layers. These procedures are illustrated briefly in Figure 3.6.

ViT primarily uses:

- **Self-attention:** Computes the relevance between patches, allowing the model to focus on important regions of the image irrespective of their spatial distance.

- **Cross-attention (in multimodal setups):** When used in conjunction with other modalities (e.g., text), ViT may incorporate cross-

[1]Used primarily in Transformer-based models, the [CLS] token is a special token that gets added to the start of a sentence. Its purpose is to act as a single, condensed representation of the entire text, which the model can then use to perform tasks like sentence classification.

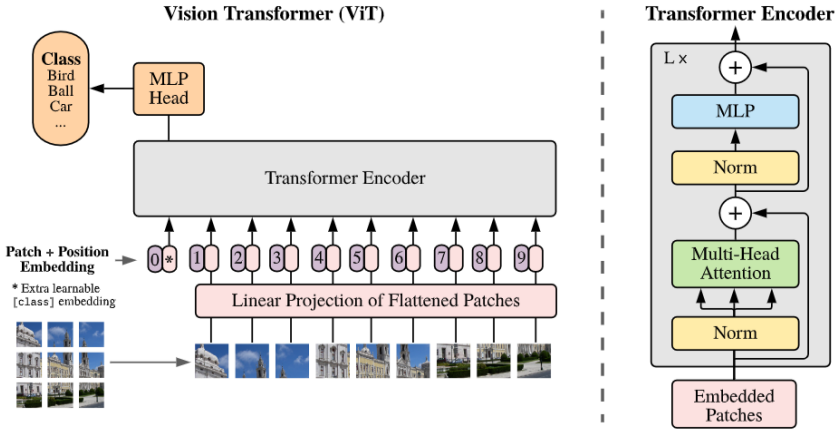



Figure 3.6: Vision Transformer Overview. Figure from [12]

attention mechanisms to align information from both sources. This is common in models like CLIP [51].

Compared to CNNs, ViT demonstrates competitive or superior performance when trained on large datasets. However, it requires significantly more data or strong data augmentation techniques due to the lack of inductive biases like locality and translation invariance found in CNNs.

## 3.4 Human Body Models

In order to correctly estimate the 3D shape and pose of a human, we need a strong model that can express them. Different models have been proposed to describe the human body. SMPL [36] is one of the most powerful and realistic models that can be used for this scope. Based on SMPL, there has been developed a family of models on it, like SMPL-X [48], SMIL [19] and SMPL-A [47]. SMPL has also been the inspiration for SMAL [76], a model that applies the SMPL concept to animals. Except for these SMPL-based models, others have also been proposed. For example, STAR [44] uses a different formulation for its blend shapes than SMPL, which are corrective shapes that deform the mesh to look more natural. SKEL [26] combines the SMPL surface mesh with a biomechanically accurate skeleton. SCAPE [2] is one of the predecessors of SMPL, which also uses a template mesh and a set of parameters to model body deformations.

### 3.4.1 Skinned Multi-Person Linear (SMPL) Model

SMPL [36] is a learned human body model that has been trained on a large dataset of various human shapes and poses. This training enables it to represent a wide variety of human body types accurately.

The SMPL model separates the body shape into two components, as shown in Figure 3.7: identity-dependent shape and non-rigid pose-dependent shape. SMPL utilizes a vertex-based skinning approach with corrective blend shapes.

SMPL modeling starts with a mesh template $\overline{\mathbf{T}} \in \mathbb{R}^{3N}$, the average body shape and pose, which consists of $N = 6890$ vertices and $K = 23$ joints. SMPL uses two parameters to deform the mesh template: shape parameter $\boldsymbol{\beta}$ and pose parameter $\boldsymbol{\theta}$.

The $\boldsymbol{\beta} \in \mathbb{R}^n$ (commonly $n = 10$) parameters are derived from the $n$ principal components of the PCA on a dataset of thousands of 3D body scans. $\boldsymbol{\beta}$ shape parameters control the identity-dependent shape of the person, like the height and weight, the muscle tone, and, in general, the proportions of the body (like longer legs or arms).

For the pose, SMPL uses a skeleton of $K = 23$ joints. For each joint, three parameters are needed to describe the rotation, using axis-angle rotation. Three more parameters are also essential for the global orientation of the body. Thus, the total number of pose parameters is $3 \cdot K + 3 = 72$.

More formally, SMPL is a model $M(\boldsymbol{\beta}, \boldsymbol{\theta}; \Phi) : \mathbb{R}^{|\boldsymbol{\theta}| \times |\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{3N}$ that maps the shape and pose parameters to vertices to create a 3D mesh. $\Phi = \{\overline{\mathbf{T}}, \mathcal{W}, \mathcal{S}, \mathcal{J}, \mathcal{P}\}$ is the full set of parameters of the SMPL model. $\mathcal{W} \in \mathbb{R}^{N \times K}$ is a set of blend weights. $\mathcal{S} = \left[\mathbf{S_1}, \ldots, \mathbf{S}_{|\boldsymbol{\beta}|}\right] \in \mathbb{R}^{3N \times |\boldsymbol{\beta}|}$ is the matrix of the orthonormal principal components of shape displacements and are used in the shape blend shapes.

$$B_S(\boldsymbol{\beta}; \mathcal{S}) = \sum_{n=1}^{|\boldsymbol{\beta}|} \beta_n \mathbf{S_n}$$

where $\boldsymbol{\beta} = \left[\beta_1, \ldots, \beta_{|\boldsymbol{\beta}|}\right]^T$. The linear function $B_S(\boldsymbol{\beta}; \mathcal{S})$ is fully defined by the matrix $\mathcal{S}$, which is learned from training meshes.

$\mathcal{J}$ represents the learned matrix that transforms rest vertices into rest joints. The 3D location of a joint is determined by the body shape $\boldsymbol{\beta}$, so the joints are a function of $\boldsymbol{\beta}$.

$$J(\boldsymbol{\beta}; \mathcal{J}, \overline{\mathbf{T}}, \mathcal{S}) = \mathcal{J}(\overline{\mathbf{T}} + B_S(\boldsymbol{\beta}; \mathcal{S}))$$

$\mathcal{P} = \left[\mathbf{P_1}, \ldots, \mathbf{P_{9K}}\right] \in \mathbb{R}^{3N \times 9K}$ is a matrix of all pose blend shapes, where $\mathbf{P} \in \mathbb{R}^{\mathbf{3N}}$ are vectors of vertex displacemetns. The matrix $\mathcal{P}$ fully defines the

pose blend shape function $B_P(\boldsymbol{\theta}; \mathcal{P})$ from which $\mathcal{P}$ is learned.

$$B_P(\boldsymbol{\theta}; \mathcal{P}) = \sum_{n=1}^{9K} (R_n(\boldsymbol{\theta}) - R_n(\boldsymbol{\theta}^*))\mathbf{P_n}$$

where $\boldsymbol{\theta}^*$ is the rest pose and $R_n(\boldsymbol{\theta})$ denotes the $n^{th}$ element of $R(\boldsymbol{\theta})$, a function $R : \mathbb{R}^{|\boldsymbol{\theta}|} \rightarrow \mathbb{R}^{9K}$ that maps a pose vector $\boldsymbol{\theta}$ to a vector of concatenated part relative rotation matrices. Thus, the total number of pose blend shapes is $23 \times 9 = 207$.

Applying a standard blend skinning function $W(\cdot)$ (dual-quaternion or linear) to rotate the vertices around the estimated joint centers with smoothing defined by the blend weights, the SMPL model is finally defined as:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}; \Phi) = W\left(T_P(\boldsymbol{\beta}, \boldsymbol{\theta}; \overline{\mathbf{T}}, \mathcal{S}, \mathcal{P}), J(\boldsymbol{\beta}; \mathcal{J}, \overline{\mathbf{T}}, \mathcal{S}), \boldsymbol{\theta}, \mathcal{W}\right)$$

and each vertex is transformed as:

$$\mathbf{t}'_\mathbf{i} = \sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{w_{k,i}} \mathbf{G}'_\mathbf{k}(\boldsymbol{\theta}, \mathbf{J}(\boldsymbol{\beta}; \mathcal{J}, \overline{\mathbf{T}}, \mathcal{S}))\mathbf{t_{P,i}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \overline{\mathbf{T}}, \mathcal{S}, \mathcal{P})$$

where

$$\mathbf{t}_{P,i}(\boldsymbol{\beta}, \boldsymbol{\theta}; \overline{\mathbf{T}}, \mathcal{S}, \mathcal{P}) = \bar{t}_i + \sum_{m=1}^{|\boldsymbol{\beta}|} \beta_m s_{m,i} + \sum_{n=1}^{9K} \left(R_n(\boldsymbol{\theta}) - R_n(\boldsymbol{\theta}^*)\right) \mathbf{p}_{n,i}$$

is the vertex $i$ after the blend shapes and $\mathbf{s}_{m,i}, \mathbf{p}_{n,i} \in \mathbb{R}^3$ the elements of the shape and pose blend shapes corresponding to template vertex $\bar{\mathbf{t}}_i$.

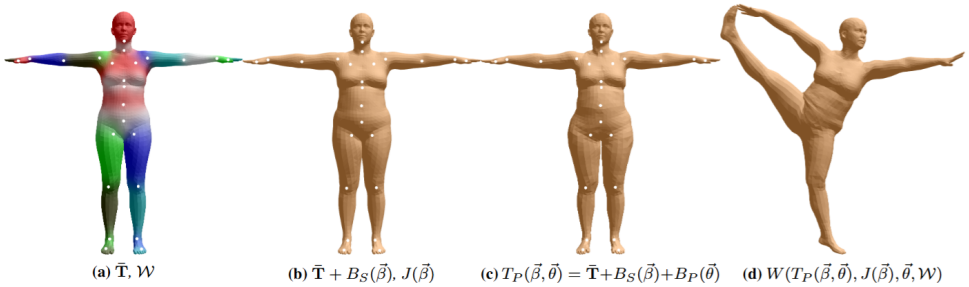



Figure 3.7: SMPL Model: (a): The template mesh $\overline{\mathbf{T}}$ with blend weights $\mathcal{W}$ indicated with the different colors and the joints with the white points. (b): The mesh when we add the identity-specific blend shape. (c): The mesh with the addition of the blend shape specific to the pose. (d): Final mesh with the desired pose. Figure from [36]

### 3.4.2 Skinned Multi-Infant Linear (SMIL) Model

While SMPL is a widely used model for human body reconstruction, it fails to produce accurate results for infants due to the lack of training on infant scans. To address this limitation, the SMIL model [19] was introduced as a specialized, learned model for infant human body reconstruction. The development of SMIL was a crucial step in creating a tool tailored for this specific population, whose body proportions and movements are significantly different from those of adults. By focusing on infants, SMIL overcomes the inherent biases of models trained on adult data.

The SMIL model was trained on a unique dataset consisting of relatively low-quality RGB-D data captured from freely moving infants. This training approach is particularly noteworthy because it accounts for the real-world challenges of collecting data from babies, who are not cooperative subjects. The model's ability to learn from this less-than-ideal data demonstrates its robustness and its practical utility. This method ensures that SMIL can accurately handle the dynamic, unconstrained movements and squirming that are characteristic of infants.

SMIL's inspiration from the SMPL framework suggests a familiar underlying structure, likely involving a template mesh that is then deformed by shape and pose parameters. However, unlike SMPL, these parameters and the template itself are specifically optimized for infant anatomy, capturing the unique bone structure and fat distribution of babies. This specialized approach makes SMIL a valuable tool for applications in pediatric healthcare, developmental research, and realistic infant animation in computer graphics, providing a reliable method for generating accurate 3D representations of infant bodies.

### 3.4.3 The SMPL Model Family

The success of the SMPL model has inspired the creation of a family of related models that extend its capabilities in various ways. These extensions build upon the core principles of SMPL while adding new features for more detailed human body representation. They all leverage a learned, low-dimensional parameterization to generate a wide range of realistic human shapes and poses.

#### SMPL+H

**SMPL+H** [55] is an extension of SMPL that allows for modeling of both the body and the hands. It achieves this by integrating **MANO** (Hand

Model with Articulated and Non-rigid defOrmations), a separate, learned model designed specifically for reconstructing human hands. By attaching the MANO model to the SMPL body, SMPL+H provides a fully articulated model of the body, including the hands, with a combined set of parameters for shape and pose. This allows for applications in virtual reality, robotics, and haptic feedback systems where detailed hand pose is essential for interaction. The model provides a unified framework for capturing and animating the entire upper body, from the torso to the fingertips.

#### SMPL-X

**SMPL-X** (SMPL eXpressive) [48] is a more comprehensive extension that models the face, hands, and body within a single framework. Its mesh contains $N = 10,475$ vertices and $K = 54$ joints, a significant increase over the original SMPL model. This additional complexity allows SMPL-X to include joints for the neck, jaw, eyeballs, and fingers. As a result, SMPL-X can effectively model facial expressions as well as the detailed position of the fingers, making it a powerful tool for capturing nuanced human movements and expressions. The model includes an additional set of "expression" parameters to control facial animations. This makes it particularly useful for creating realistic avatars in computer graphics, virtual humans for simulations, and for analyzing non-verbal communication from video data.

### 3.4.4 SMPL-A

While SMPL can model a range of adult bodies, the absence of children's 3D scans in its training data hinders its ability to generalize and effectively model children's bodies. Thus, Patel et al. introduce the SMPL-A [47] model, which models effectively both child and adult bodies.

The key difference between SMPL and SMPL-A is the body template. SMPL-A uses as body template an interpolation of an adult $T_A$ and a child $T_C$ template. $T_A$ is the SMPL adult template, while $T_C$ is the SMIL infant tempalte. This interpolation is controlled by a shape interpolation parameter-weight $\alpha \in [0, 1]$ according to the following equation:

$$T_F = \alpha T_C + (1 - \alpha) T_A$$

SMPL-A adapts SMPL's adult shape space, so the only difference between SMPL and SMPL-A is the body template. Experiments have shown that this change enables the creation of more accurate body shapes for children. For the aforementioned reasons, most of our experiments and our final method utilize SMPL-A.

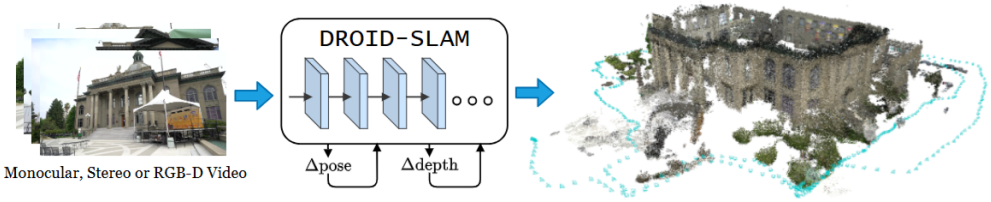



Figure 3.8: Simultaneous Localization and Mapping (SLAM). Figure from [64]

## 3.5 Simultaneous Localization and Mapping

Many computer vision applications, such as autonomous vehicles and robotics, require an accurate 3D representation of the surrounding environment. Humans naturally excel at localizing themselves and building a mental map of unknown places while moving through them. Simultaneous Localization and Mapping (SLAM) aims to replicate this capability: it constructs a map of an unknown environment while simultaneously estimating the agent's position within it. Common approaches to SLAM include particle filters and Kalman filters. An example of a SLAM system is illustrated in Figure 3.8.

Although SLAM might seem unrelated to estimating the 3D shape and pose of humans, it can serve as a crucial first step—especially in video scenarios—by providing spatial consistency and enabling robust detection and tracking of individuals across frames. More specifically, in Section 5.1 we will examine a way of how SLAM can help to get more accurate 3D reconstructions of humans.

# Chapter 4

# Literature Review

The estimation of 3D human shape and pose from a single image or a video has been a focal point of research in the field for several years. Many attempts have been proposed that try to solve this problem. The two primary categories of such methods are optimization-based and regression-based. In this chapter, we will give a brief overview of such methods, and we will examine more extensively some of the most effective and frequently used. Furthermore, we will present an overview of methods in similar problems, like 2D pose estimation from images.

## 4.1 2D Pose Estimation

Accurate 2D joint estimation is a prerequisite for most methods that estimate the 3D shape and pose of humans. Consequently, a reliable 3D shape and pose estimator must be built upon a solid 2D joint estimator. This problem has been studied for many years, with solutions evolving from traditional computer vision methods [58, 67] to highly effective deep learning approaches [7, 35, 61, 65, 71].

### 4.1.1 Traditional Methods

Early computer vision methods for 2D human joint estimation often relied on traditional techniques. These approaches typically utilize **part-based models**, which break down the human body into individual parts (e.g., head, torso, arms) and detect them independently before modeling their spatial relationships to estimate a full pose. Other methods used **hand-crafted features** from predefined descriptors like Histograms of Oriented Gradients (HOG) [10] to represent the visual characteristics of the body. **Probabilistic**

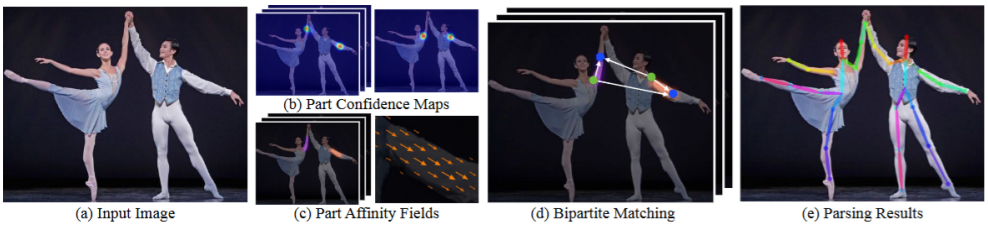



Figure 4.1: OpenPose Overview. Figure from [7]

**methods** [59] also played a role, aiming to minimize a cost function that balances the likelihood of a joint's location with the anatomical plausibility of the overall pose. While foundational, these methods were computationally intensive and lacked the ability to generalize well to complex poses, variations in lighting, and occlusions.

### 4.1.2 Deep Learning Methods

The advent of deep neural networks led to significant advancements in 2D pose estimation. Most modern methods leverage **CNNs** and **ViTs** as their basic components, as they are capable of learning robust, high-level features directly from data.

#### OpenPose

**OpenPose** [7] is a pioneering real-time, multi-person method that was once considered the state-of-the-art. It follows a bottom-up approach, first detecting all body keypoints in the image and then associating them with specific individuals. Its core is a CNN trained on large-scale pose datasets, making it particularly effective in crowded scenes due to its ability to handle multiple people simultaneously. However, its performance can degrade significantly in cases of heavy occlusion, extreme poses, or low-resolution imagery. An overview of OpenPose is illustrated in Figure 4.1. Our first experiments utilize the 25 keypoints (the names are shown in Figure 4.2) estimated by OpenPose.

#### ViTPose

**ViTPose** [71] is the current state-of-the-art 2D pose estimator. As its name suggests, it employs a **Vision Transformer (ViT)** as its core architecture, which is illustrated in Figure 4.3. Unlike CNN-based approaches, which primarily capture local features, ViTPose leverages the **self-attention**

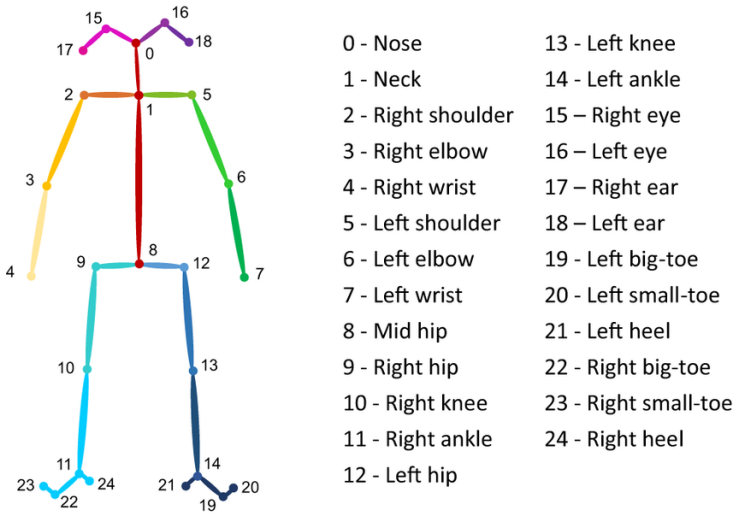



Figure 4.2: Names of Keypoints that OpenPose detects. Figure from [74]

**mechanism** to capture long-range dependencies between keypoints across the entire image. This global awareness enables more robust pose estimation in complex and challenging scenes. The model's simple, scalable, and flexible nature makes it highly transferable across different datasets and tasks. In our experiments, we use ViTPose to provide more accurate and consistent pose estimates, especially in challenging conditions where OpenPose struggles.

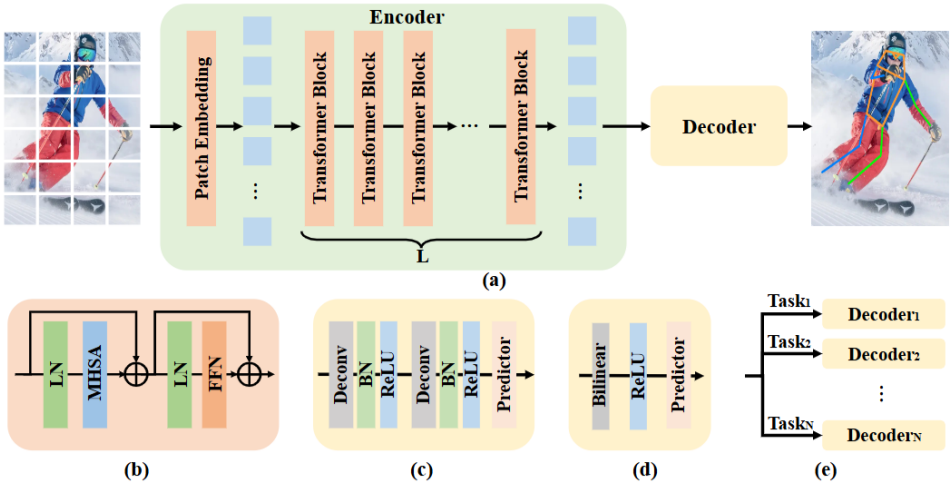



Figure 4.3: (a) The framework of ViTPose. (b) The transformer block. (c) The classic decoder. (d) The simple decoder. (e) The decoders for multiple datasets. Figure from [71]

## 4.2 3D Shape and Pose Estimation of Human Body

A significant body of research has been dedicated to addressing the problem of 3D shape and pose estimation. While the estimation of 3D human pose has been studied with considerable success as a standalone problem, the reconstruction of a detailed 3D body shape inherently requires pose information. Consequently, most contemporary methods jointly estimate both 3D pose and shape to accurately represent the human silhouette.

Early approaches to 3D pose estimation primarily involved a lifting process from 2D to 3D. These methods first estimated 2D joint locations and pose, which were then used to infer the corresponding 3D joint coordinates [8, 41, 43, 45]. Most of these methods use a CNN to estimate the 2D joints and then, using a neural network or a regressor, they estimate the 3D location of the joints. End-to-end methods have been proposed to solve the problem, bypassing the intermediate step of estimating the 2D pose. These methods use modern deep learning methods [33, 60] or more innovative ways, like the WiFi Signals [72].

To address the challenge of depth and shape ambiguity, many state-of-the-art methods utilize multi-view cameras [11, 20, 50] or videos as input [52, 73]. By observing a subject from multiple viewpoints, either spatially (multi-view) or temporally (video), these models can effectively resolve ambiguities. Since a person's shape parameters remain constant within a short time frame, analyzing different views from a video sequence helps the model better understand the body's structure and produce more accurate shape estimates. A similar principle is applied in multi-shot reconstruction, where multiple still images from different angles are used to reconstruct a static human subject [49]. Another standard method to deal with the depth ambiguity is the use of RGB-D cameras, where except for the image, we also get the depth [5].

More recently, modern methods have integrated the biomechanics of the human body and motion [28, 70]. By taking biomechanical constraints into account, these models can generate more plausible and smoother motion, effectively rejecting physically impossible poses and improving the robustness of the estimation.

The primary focus of this thesis is the estimation of three-dimensional shape and pose from a single RGB image. A single RGB image represents the most challenging form of input data for this problem, as it obscures depth information and provides only a view of a human or humans in a multi-person scenario. Despite the challenging nature of the problem, the research community has developed a plethora of methods that attempt to solve it, with

great results. These methods, as explained in Chapter 2 can be optimization-based [6, 14, 48], regression-based with the use of deep learning methods [13, 16, 24, 62, 75] or a combination of optimization and regression [29].

The following subsections delve into a selection of effective methods used in our initial experiments, with which our results are compared.

### 4.2.1 SMPLify

In 2016, Bogo et al. [6] introduced SMPLify, the first method that can estimate the 3D shape and pose of a human body from a single unconstrained image. SMPLify pipeline starts with an estimation of the 2D joints of the human in the picture. This is done with the DeepCUT CNN [53], which finds the 2D location of the joints along with a confidence score $w_i$. Then, SMPLify tries to fit a SMPL model onto these 2D joints. The optimization process minimizes the reprojection loss between the projected 2D keypoints of the predicted 3D mesh and the estimated 2D joints from DeepCUT. Furthermore, SMPLify adds an interpenetration term in the loss function that is differentiable to both body shape and pose, and helps to prevent implausible poses. SMPLify prevents interpenetration by approximating the body surface as a set of capsules, where each capsule has a specific radius and axis length. These steps are illustrated briefly in Figure 4.4.

More formally, let $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \gamma)$ be the SMPL body model with shape parameters $\boldsymbol{\beta}$, pose parameters $\boldsymbol{\theta}$ and translation $\gamma$. Let $J(\boldsymbol{\beta})$ be the function that predicts 3D skeleton joint locations from body shape. For each joint $i$, the posed 3D joint is denoted as $R_{\boldsymbol{\theta}}(J(\boldsymbol{\beta})_i)$ where $R_{\boldsymbol{\theta}}$ is the global rigid transformation induced by pose $\boldsymbol{\theta}$.

The objective function that SMPLify tries to minimize during optimization is a sum of five error terms:

$$E(\boldsymbol{\beta}, \boldsymbol{\theta}) = E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; K, J_{est}) + \lambda_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \lambda_\alpha E_\alpha(\boldsymbol{\theta}) + \lambda_{sp} E_{sp}(\boldsymbol{\theta}; \boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}}(\boldsymbol{\beta})$$

where $K$ are the camera parameters and $\lambda_\theta, \lambda_\alpha, \lambda_{sp}, \lambda_\beta$ weights.

The first term $E_J$ penalizes the 2D distance between the estimated 2D joints and the SMPL projected joints based on the following equation:

$$E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; K, J_{est}) = \sum_{\text{joint i}} w_i \rho(\Pi_K(R_\theta(J(\beta)_i)) - J_{est,i})$$

where $\Pi_K$ is the 3D to 2D projection of the camera with parameters $K$ and $w_i$ are the confidence scores of the estimation of joints from DeepCUT.

$E_\alpha(\boldsymbol{\theta})$ is a pose prior that penalizes elbows and knees that have unnatural bending:

$$E_\alpha(\theta) = \sum_i \exp(\boldsymbol{\theta}_i)$$

where $i$ is a counter over knees and elbows pose parameters.

The $E_\theta(\boldsymbol{\theta})$ term is used for the pose prior to eliminate implausible poses. It is built by fitting SMPL to the CMU marker data using MoSh [37] and by an approximation of a sum of a mixture of Gaussians as described in the following equation:

$$\begin{aligned} E_{\boldsymbol{\theta}}(\boldsymbol{\theta}) &\equiv -\log \sum_j (g_j \mathcal{N}(\boldsymbol{\theta}; \mu_{\theta,j}, \Sigma_{\theta,j})) \\ &\approx -\log \left( \max_j (c g_j \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu_{\theta,j}}, \boldsymbol{\Sigma_{\theta,j}})) \right) \\ &= \min_j (-\log(c g_j \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\theta,j}, \Sigma_{\theta,j}))) \end{aligned}$$

where $g_j$ are the mixture model weights of 8 Gaussians, and $c$ a positive constant.

For the interpenetration error term, SMPLify separates the human body into "capsules" and checks for intersection between capsules that cannot be intersected in natural poses. For simplicity, these capsules are simplified into spheres with centers $C(\boldsymbol{\theta}, \boldsymbol{\beta})$ and radii $r(\beta)$. For the penalty term, a 3D isotropic Gaussian with $\sigma(\boldsymbol{\beta}) = \frac{r(\boldsymbol{\beta})}{3}$ is is used to describe each sphere. The error term is a mixture of 3D Gaussians as described in the following equation:

$$E_{sp}(\boldsymbol{\theta}; \boldsymbol{\beta}) = \sum_i \sum_{j \in I(i)} \exp \left( \frac{\|C_i(\boldsymbol{\theta}, \boldsymbol{\beta}) - C_j(\boldsymbol{\theta}, \boldsymbol{\beta})\|^2}{\sigma_i^2(\boldsymbol{\beta}) + \sigma_j^2(\boldsymbol{\beta})} \right)$$

Finally, a shape prior is used, which is defined as:

$$E_\beta(\boldsymbol{\beta}) = \boldsymbol{\beta} \Sigma_\beta^{-1} \boldsymbol{\beta}$$

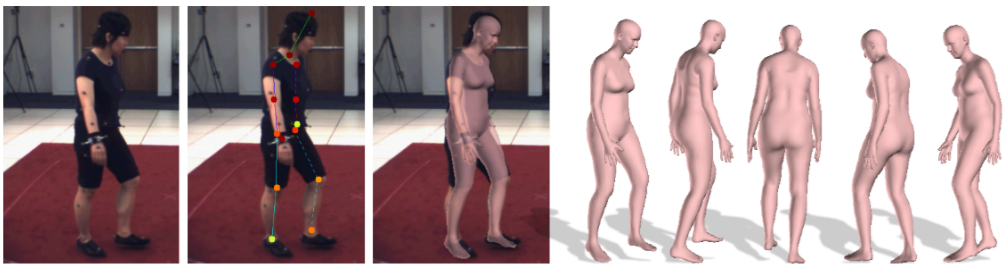


Figure 4.4: SMPLify Overview. Figure from [6]

### 4.2.2 SMPLify-X

Pavlakos et al. [48] introduced SMPLify-X, a method that builds upon the foundational principles of SMPLify but incorporates significant improvements. SMPLify-X leverages the more expressive SMPL-X model, which is capable of representing the human body, hands, and face. The pipeline is designed to be compatible with the entire SMPL family of models.

Figure 4.5: SMPLify-X examples. Figure from [48]

Similar to the original SMPLify, SMPLify-X begins by using 2D image keypoints to fit the 3D body model. These keypoints, including those for the body, hands, and face, are extracted jointly using OpenPose. Then, fitting SMPL-X to the image is an optimization problem whose goal is to minimize the objective function:

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C}$$

where $\theta_b, \theta_f$ and $m_h$ are the pose vectors for the body, face and the two hands, respectively. The notation and the error terms are similar to the SMPLify ones. Specifically, $\theta$ is the pose parameters, $\beta$ are the shape parameters, while $E_{m_h}(m_h), E_{\theta_f}(\theta_f)$, and $E_{\mathcal{E}}(\psi)$ are $L_2$ priors for the hand pose, facial pose, and facial expressions. $E_\beta(\beta) = \|\beta\|^2$ is the Mahalanobis distance between the shape parameters of the optimization and the shape distribution in the SMPL-X training dataset. $E_\alpha(\theta_b) = \sum_{i \in (\text{elbows, knees})} \exp(\theta_i)$ is a prior for elbows and knees.

Similar to SMPLify, SMPLify-X has a data term that penalizes the distance between the 2D keypoints detected by OpenPose and the 3D to 2D projected keypoints. Specifically, the data term is

$$E_J(\beta, \theta, K, J_{est}) = \sum_{\text{joint } i} \gamma_i \omega_i \rho(\Pi_K(R_\theta(J(\beta)_i)) - J_{est,i})$$

where $\Pi_K$ is the 3D to 2D projection with intrinsic camera parameters $K$, $\omega_i$ is the detection confidence score for each joint $i$ and $\gamma_i$ are the weights per joint for annealed optimization, as empirically it was found that an annealing scheme for these weights helps optimization of the objective function to deal with ambiguities and local optima.

Another contribution of SMPLify-X is the Variational Human Body Pose Prior (VPoser) that penalizes impossible poses. The error term $E_{\theta_b}(\theta_b)$ in the objective function describes this pose prior. VPoser is trained and tested on a set of approximately 1M and 65k poses, respectively, making it a very effective pose prior.

SMPLify-X introduces a new interpenetration loss to avoid self-collisions and penetrations of body parts that are physically impossible. This is done by detecting colliding triangles $\mathcal{C}$ on the mesh using Bounding Volume Hierarchies (BVH). Then, for each pair of colliding triangles $f_s$ and $f_t$, the algorithm computes a local conic 3D distance field $\Psi$, i.e., a function that gives the signed distance from any point in space to the surface of the triangle. The sign of the distance indicates whether the point is inside or outside the mesh. Finally, the loss term penalizes the depth of intrusion by taking the vertices of one colliding triangle and evaluating their position within the distance field of the other colliding triangle. The collision term in the objective function is defined as:

$$E_C(\theta) = \sum_{(f_s(\theta), f_t(\theta)) \in \mathcal{C}} \left\{ \sum_{v_s \in f_s} \| - \Psi_{f_t}(v_s) n_s \|^2 + \sum_{v_t \in f_t} \| - \Psi_{f_s}(v_t) n_t \|^2 \right\}$$

Finally, SMPLify-X uses a Deep Gender Classifier to detect the gender of humans in the images. The proportions and the shape of men's and women's bodies are different, and, therefore, knowing the gender of each human in the image can improve the quality of the fitting by using a gender-specific model. If the detected gender probability is below a threshold, a gender-neutral body model is fitted. Some reconstructions produced from SMPLify-X are shown in Figure 4.5.

### 4.2.3 ProHMR

The ProHMR [30] (Probabilistic Human Mesh Recovery) method proposes a probabilistic way to solve the problem of 3D human reconstruction from 2D data, which may be an image or keypoints. To this end, given the input, this method learns a mapping from the input to a distribution of plausible 3D poses. This distribution is regressed using Normalizing Flows, which are used for the representation of complex distributions as a series of invertible transformations of a simple base distribution (typically a standard multivariate Gaussian).

The architecture of ProHMR comprises a CNN that encodes the input image to get a context vector $\mathbf{c}$ that is used as the conditioning input to the Normalizing Flow model to get the distribution of SMPL pose parameters $\theta$. For the task of 3D pose regression, the authors decide to select the mode of the distribution as the most appropriate choice. The same vector $\mathbf{c}$ is the input to an MLP, which outputs the SMPL shape $\beta$ and camera parameters $\pi$, as they do not depend on the pose.

For the task of body fitting, the logic is similar to SMPLify, with the use of the following objective function:

$$\lambda_J E_J - \ln p_{\Theta|I}(\theta|\mathbf{c}) + \lambda_\beta E_\beta$$

where $E_J$ is the term for the 2D keypoint reprojection loss, $E_\beta$ a quadratic penalty on the shape coefficients, and $E_{\theta|I} = -\ln p_{\Theta|I}(\theta|\mathbf{c})$ a pose prior that models the likelihood of a given pose conditioned on the image evidence. This prior is used instead of the standard 3D priors about the 3D pose and the unnatural rotations of elbows and knees. The architecture of ProHMR is presented in Figure 4.6.

### 4.2.4 BEV

The Bird's-Eye-View (BEV) [62] method is particularly effective at producing accurate 3D shapes for children and babies. This success is primarily

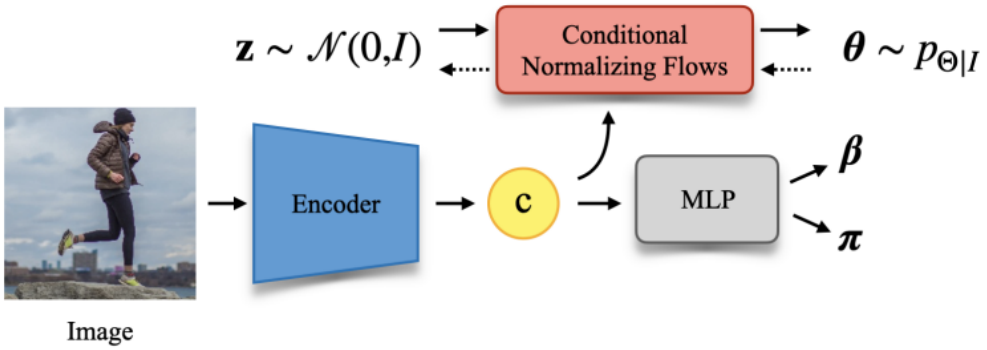



Figure 4.6: ProHMR Architecture. Figure from [30]

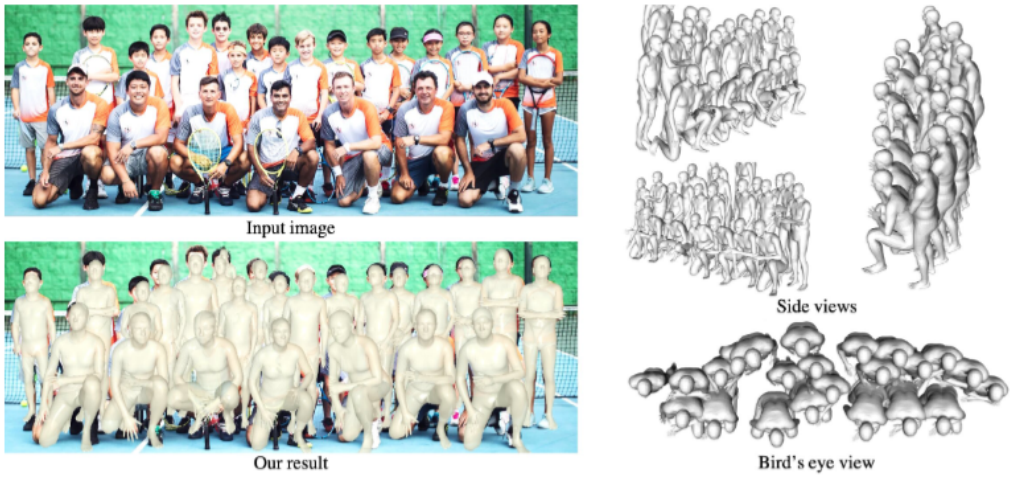



Figure 4.7: Bird's-eye-view example. Figure from [62]

attributed to its use of a modified SMPL-A model to represent human figures. As explained in 3.4.4, SMPL-A is a reliable model for representing both adults and children within a single framework. However, a challenge with this approach is that pose estimation remains problematic in many cases.

BEV addresses the multi-person problem by estimating the depth of each individual using a "bird's-eye-view" map, shown in Figure 4.7, which provides a top-down estimate of body centers. This is a regression-based method that operates as follows:

1. Given an input image, a network first generates four feature maps: two for the body center and localization offset in the frontal view, and two for the same in the bird's-eye view.

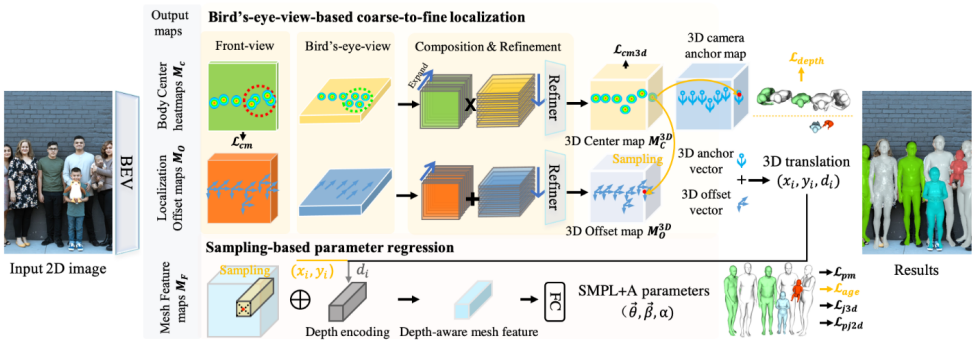



Figure 4.8: BEV Overview. Figure from [62]

2. These four maps are then combined to produce 3D Center and Offset maps.

3. These maps are subsequently used to predict the 3D translation of each person $(x_i, y_i, d_i)$.

4. Finally, the 3D translation and a mesh feature map are used to regress the SMPL-A parameters.

An overview of the BEV method is illustrated in Figure 4.8.

To address the issue of modeling adults and children with a single model, the authors of BEV utilize a modified SMPL-A model. Specifically, instead of using only the SMPL model, they also employ the SMIL model when the interpolation parameter $\alpha$ is above a threshold $t_\alpha$. As explained in 3.4.4, $\alpha \in [0, 1]$, with $\alpha = 0$ corresponding to a purely adult template and $\alpha = 1$ representing a child template. Conversely, when $\alpha \leq t_\alpha$, the SMPL model is used, along with the extra $\alpha$ parameter. This is based on the intuition that the shape space of infants is distinct from that of children and adults, and consequently, using the SMPL shape space for infants may produce incorrect results. Therefore, when the subject appears to be an infant, the SMIL model is employed for its specialized infant shape space.

# Chapter 5

# Methodology

We leverage existing approaches and propose two complementary methods for estimating the 3D shape and pose of humans from single images or videos. Both methods support single and multi-person scenarios and are designed to better model children and babies. Our goal is to develop effective methods that can be universally applied to humans of all ages.

The first method is an optimization-based approach that fits the SMPL-A human body model to images or videos by minimizing a reprojection-based objective function. This objective combines alignment between 3D keypoints and pseudo-ground-truth 2D detections with additional regularization terms and improved strategies for detection and tracking.

The second method is a learning-based approach that predicts 3D shape and pose from a single image. While this method offers much faster inference than the optimization-based approach, it requires large-scale annotated datasets for effective training. To address the lack of child-specific annotations, we generate high-quality pseudo-ground-truth labels using the first method, thereby augmenting existing datasets.

## 5.1 Optimization-based algorithm

In many scenarios, the goal extends beyond reconstructing 3D shape and pose from single images to handling full video sequences. The single-image case can be regarded as a special case of video-based reconstruction, where the sequence consists of a single repeated frame. Preliminary experiments have shown that when the input of a 3D shape and pose estimation system is a video, the results are better compared to the single-image case. Therefore, we use SLAHMR (Simultaneous Localization and Human Mesh Recovery) [73] as a baseline for our approach, with modifications to adjust it to our problem.

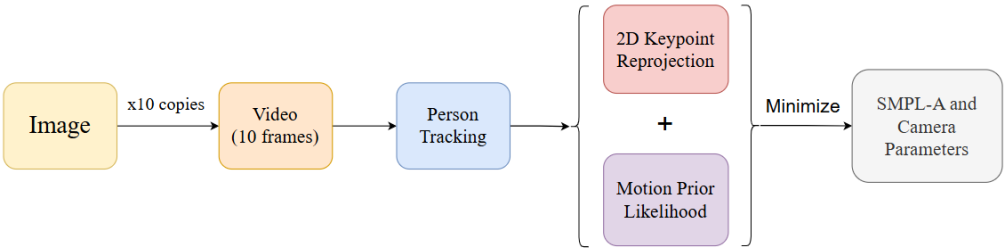



Figure 5.1: Pipeline of our optimization method

SLAHMR is an optimization-based approach designed for this problem. It operates on videos in the wild, supports multiple people per frame, and jointly recovers both human trajectories and camera motion in a common world coordinate system. The key insight is that in unconstrained videos, camera motion is often significant, and accurate reconstruction requires explicitly modeling this motion.

SLAHMR was originally developed for the SMPL+H model, but it is possible to adapt its pipeline to other members of the SMPL family. Our first method builds upon SLAHMR. We use the SMPL-A model instead of SMPL+H and we modify the pipeline to be able to handle both images and videos. Originally, SLAHMR works with videos, but our extension, when the input is an image, creates a small video of 10 frames of the same image. Figure 5.1 shows a brief overview of the pipeline of our optimization method.

For a video of $T$ frames containing $N$ people, each person $i$ at time step $t$ is represented as:

$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}$$

where $\Phi_t^i \in \mathbb{R}^3$ is the global orientation, $\Theta_t^i \in \mathbb{R}^{22\times 3}$ the body pose with 22 joint angles, $\beta^i \in \mathbb{R}^{11}$ the shape over all time steps $t$, where the $11^{th}$ value is the $\alpha$ interpolation weight, and $\Gamma_t^i \in \mathbb{R}^3$ the root translation.

The first step is to estimate each person's per-frame pose $\hat{\mathbf{P}}_t^i$ and compute their unique identity track associations over all frames using a 3D tracking system, PHALP [52]. Recently, the method had been compatible with 4DHumans tracking, which is analyzed in Section 5.2. In our method we use 4DHumans tracking system, as a more modern and effective method than PHALP.

In a video, the net motion, *i.e.*, a person's motion in the camera coordinates, depends both on the human's and camera's motion in the world frame. Therefore, the camera motion should also be modeled in a correct way. Let ${}^c\mathbf{P}_t^i = \{{}^c\Phi_t^i, \Theta_t^i, \beta^i, {}^c\Gamma_t^i\}$ the pose in the camera frame and

${}^{w}\mathbf{P}_t^i = \{{}^{w}\Phi_t^i, \Theta_t^i, \beta^i, {}^{w}\Gamma_t^i\}$ the pose in the world with the same local pose $\hat{\Theta}_t^i$ and shape $\hat{\beta}^i$ parameters.

SLAHMR uses DROID-SLAM [64], a SLAM system, to estimate the world-to-camera transform at each time $t$, $\{\hat{R}_t, \hat{T}_t\}$. This is essential to compute the relative camera motion between video frames. A human motion in the world prior is used to determine the camera scale $\alpha_c$ and people's global trajectories. The camera scale $\alpha_c$ is important to be estimated correctly to place the people in the world, so the human bodies and motion are plausible.

First, the global orientation and root translation in the world coordinate frame using the estimated camera transforms and camera-frame parameters are initialized. Camera scale is initialized in the value of $\alpha_c = 1$.

$$
\begin{aligned}
{}^{w}\Phi_t^i &= R_t^{-1c}\hat{\Phi}_t^i, & {}^{w}\Gamma_t^i &= R_t^{-1c}\hat{\Gamma}_t^i - \alpha_c R_t^{-1} T_t, \\
\beta_i &= \hat{\beta}_i, & \Theta_t^i &= \hat{\Theta}_t^i,
\end{aligned}
$$

The world frame joints are expressed as:

$$
{}^{w}\mathbf{J}_t^i = \mathcal{M}({}^{w}\Phi_t^i, \Theta_t^i, \beta^i) + {}^{w}\Gamma_t^i
$$

where $\mathcal{M}$ is the differentiable function that SMPL model uses to generate the mesh vertices and joints.

Similar to SMPLify, SLAHMR defines a 2D joint reprojection loss to align the projected 3D to 2D joints with the detected from ViTPose 2D keypoints $x_t^i$:

$$
E_{\text{data}} = \sum_{i=1}^{N}\sum_{t=1}^{T} \psi_t^i \rho(\Pi_K(R_t \cdot {}^{w}\mathbf{J}_\mathbf{t}^\mathbf{i} + \alpha \mathbf{T}_\mathbf{t}) - \mathbf{x}_\mathbf{t}^\mathbf{i})
$$

where $\Pi_K(\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T) = K \begin{bmatrix} \frac{x_1}{x_3} & \frac{x_2}{x_3} & 1 \end{bmatrix}^T$ is perspective camera projection with camera intrinsics matrix $K \in \mathbb{R}^{2\times 3}$, $\rho$ is the robust Geman-McClure function [4] and $\psi_t^i$ are the confidence scores of the detected 2D keypoints.

At this stage of the optimization, due to the under-constrained reprojection loss, the optimization is being held only to the global orientation and root translation ${}^{w}\Phi_t^i, {}^{w}\Gamma_t^i$ of the human pose parameters:

$$
\min_{\{\{{}^{w}\Phi_t^i, {}^{w}\Gamma_t^i\}_{t=1}^{T}\}_{i=1}^{N}} \lambda_{\text{data}} E_{\text{data}}
$$

The optimization lasts 30 iterations with $\lambda_{\text{data}} = 0.001$.

For the camera scale $\alpha_c$, human shape $\beta_i$ and body pose $\Theta_t^i$ optimization, additional priors about human movement in the world are used. This

optimization stage smooths the transitions between poses in the world trajectories so that the displacements of the people are plausible. The prior of joint smoothness is defined as:

$$E_{\text{smooth}} = \sum_{i}^{N} \sum_{t}^{T} \|\mathbf{J_t^i} - \mathbf{J_{t+1}^i}\|^{\mathbf{2}}$$

The other priors concern the pose $E_{\text{pose}} = \sum_{i=2}^{N} \sum_{t=1}^{T} \|\zeta_t^i\|^2$ and the shape $E_\beta = \sum_i^N \|\beta^i\|^2$, where $\zeta_t^i \in \mathbb{R}^{32}$ represent the body pose parameters $\Theta_t^i$ in the latent space of the VPoser model. The updated objective function to be minimized is the following:

$$\min_{\alpha, \{\{{}^{w}\mathbf{P_t^i}\}_{\mathbf{t=1}}^{\mathbf{T}}\}_{\mathbf{i=1}}^{\mathbf{N}}} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + \lambda_{\text{smooth}} E_{\text{smooth}}$$

The optimization is performed over 60 iterations using $\lambda_{\text{smooth}} = 5, \lambda_\beta = 0.05$ and $\lambda_{\text{pose}} = 0.04$.

To ensure the temporal consistency and naturalness of reconstructed human motion, SLAHMR introduces a method that incorporates a learned motion prior. The approach models the likelihood of a human trajectory using a Conditional Variational Autoencoder (**CVAE**) to regularize the output. This approach uses the transition-based motion prior HuMoR [54].

The CVAE is trained to predict the distribution of the next pose parameters $\mathbf{P_t}$, velocity and joint locations ($s_t$) given the previous ones ($s_{t-1}$), leveraging a latent variable $\mathbf{z}_t \in \mathbb{R}^{48}$ to capture motion complexity. The transition likelihood is modeled as:

$$p_\theta(s_t|s_{t-1}) = \int_{z_t} p_\theta(\mathbf{z_t}|\mathbf{s_{t-1}}) \mathbf{p}_\theta(\mathbf{s_t}|\mathbf{z_t}, \mathbf{s_{t-1}})$$

The conditional prior $p_\theta(\mathbf{z_t}|\mathbf{s_{t-1}})$ is a Gaussian distribution with mean $\mu_\theta(s_{t-1})$ and covariance $\sigma_\theta(s_{t-1})$. This learned prior is then used in an energy term on the latents $\mathbf{z_t^i}$:

$$E_{\text{CVAE}} = -\sum_{i}^{N} \sum_{t}^{T} \log \mathcal{N}(\mathbf{z_t^i}; \mu_\theta(\mathbf{s_{t-1}^i}), \sigma_\theta(\mathbf{s_{t-1}^i}))$$

The method performs a global optimization over a sequence of states to recover complete human motion trajectories for multiple individuals. The process begins by initializing the motion transition latent variables, $\mathbf{z}_t^i$, for each person $i$ at each time step $t$. These latent variables are derived from the previous state, $s_{t-1}^i$, using a pre-trained HuMoR encoder ($\mu_\phi$). The

next state, $s_t^i$, is then recursively generated from the previous state and the latent variable via a HuMoR decoder ($\Delta_\theta$). This autoregressive process is represented by the equations:

$$\mathbf{z}_t^i = \mu_\phi(s_t^i, s_{t-1}^i), \quad s_t^i = s_{t-1}^i + \Delta_\theta(z_t^i, s_{t-1}^i)$$

In addition to the primary motion prior losses, the optimization includes two extra regularization terms to ensure physical plausibility. The first, a stability loss ($E_{\text{stab}}$), regularizes the predicted velocity and joint locations, making the motion more consistent. Thus, the prior optimization terms are:

$$E_{\text{prior}} = \lambda_{\text{CVAE}} E_{\text{CVAE}} + \lambda_{\text{stab}} E_{\text{stab}}$$

The second, a foot-skate loss ($E_{\text{skate}}$), explicitly prevents unrealistic sliding by penalizing the velocity of joints that are likely to be in contact with the ground plane of the scene $g \in \mathbb{R}^3$. This is calculated as:

$$E_{\text{skate}} = \sum_i^N \sum_t^T \sum_j^J c_t^i(j) \| J_t^i(j) - J_{t+1}^i(j) \|$$

where $c_t^i(j)$ is the ground contact probability for joint $j$ of person $i$ at time $t$, and $J_t^i(j)$ is its position. A final term, $E_{\text{con}}$, encourages these same contact points to stay close to the ground, calculated as:

$$E_{\text{con}} = \sum_i^N \sum_t^T \sum_j^J c_t^i(j) \max(d(J_t^i(j), g) - \delta, 0).$$

Here, $d(\mathbf{p}, \mathbf{g})$ is the distance from a point $\mathbf{p} \in \mathbb{R}^\mathbf{3}$ to the ground plane $g$, and $\delta$ is a small threshold. These combined losses ensure that the final reconstructed trajectories are not only temporally smooth but also physically realistic. Plane $g$ is optimized as a free variable shared across all people and timestamps.

Combining all the losses of this stage, the final optimization problem is:

$$\min_{\alpha_c, g, \{s_0^i\}_{i=1}^N, \{\{\mathbf{z_t^i}\}_{\mathbf{t=1}}^{\mathbf{T}}\}_{\mathbf{i=1}}^{\mathbf{N}}} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + E_{\text{prior}} + E_{\text{env}}$$

where $E_{\text{env}} = \lambda_{\text{skate}} E_{\text{skate}} + \lambda_{\text{con}} E_{\text{con}}$, $\lambda_{\text{CVAE}} = 0.075, \lambda_{\text{skate}} = 100$ and $\lambda_{\text{con}} = 10$.

The optimization at all stages is performed with the L-BFGS algorithm and a learning rate of 1.

Figure 5.2 illustrates the overall SLAHMR pipeline, providing a visual summary of the key methodological steps described in the preceding sections.

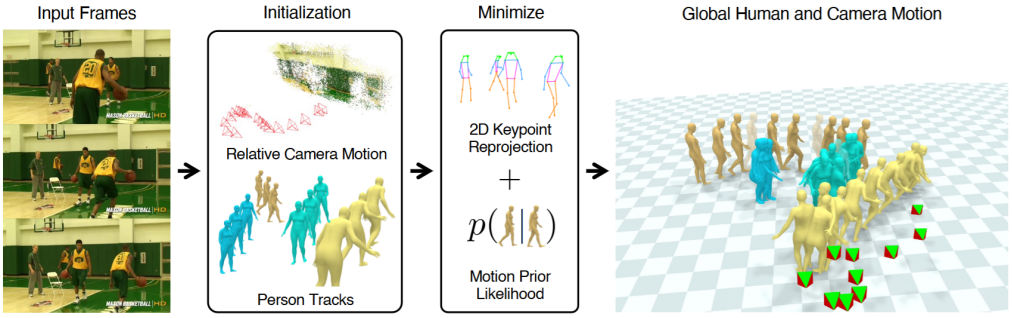



Figure 5.2: SLAHMR Pipeline. Figure from [73]

## 5.2 Learning-based method

Although the first method can reconstruct humans in a photo or a video very effectively, its optimization-based nature has some significant drawbacks. The most notable issue is that optimization requires a considerable amount of time to run, which makes these methods unable to operate in real-time. Additionally, the first method is highly sensitive to initialization. If the starting parameters are not chosen carefully, the optimization process may get stuck in a local minimum, leading to an inaccurate or anatomically implausible reconstruction.

These issues are being addressed by deep learning methods. Methods like Human Mesh Recovery (HMR) [24] use regression to predict the pose, shape, and camera parameters from an input image. Using these parameters and a model like SMPL, they can reconstruct the 3D shape and pose of all the humans in an image.

HMR2.0 [16] is an extension of HMR that uses a ViT instead of a CNN. The architecture is very simple, since HMR2.0 has an end-to-end transformer architecture with two main components. Our learning-based method uses as a starting point the HMR2.0 and leverages it to handle both adults and children by extending the training set with child and baby data. Our two core changes are the use of SMPL-A model instead of SMPL and the use of training data that contains children and babies.

A **ViT** is used to patchify the image and extract the image tokens to get processed by a **Transformer decoder** with multi-head self-attention that ends with an MLP that predicts the SMPL-A shape $\beta$ and pose $\theta$ parameters as well as the camera translation $\pi$. The interpolation weight $\alpha$ is predicted

as the $11^{th}$ $\beta$ parameter. These components are also illustrated in Figure 5.3, providing a visual summary of the HMR2.0 architecture.

HMR2.0 predictor $f$ is trained on a large mixture of datasets with a combination of 2D and 3D losses and a discriminator. Different datasets used had different annotations, so the losses are used if there are ground-truth annotations or, in some cases, pseudo-ground-truth annotations. More specifically, let $I$ be the input image and $\Theta = \left[\theta, \beta, \pi\right] = f(I)$ the model predictions on image I. If ground-truth SMPL-A shape parameters $\beta^*$ and pose parameters $\theta^*$ are available, an MSE loss is used for the model predictions:

$$\mathcal{L}_{\texttt{smpl}} = \|\theta - \theta^*\|_2^2 + \|\beta - \beta^*\|_2^2$$

When the dataset provides accurate ground-truth 3D keypoints $X^*$, a L1 loss is added to penalize the distance from the predicted 3D keypoints $X$:

$$\mathcal{L}_{\texttt{kp3D}} = \|X - X^*\|_1$$

Similarly, if there are 2D keypoints annotations $x^*$, an L1 loss is used to penalize the projection of the predicted 3D keypoints $\pi(X)$:

$$\mathcal{L}_{\texttt{kp2D}} = \|\pi(X) - x^*\|_1$$

Finally, to get plausible 3D poses, a discriminator $D_k$ is trained for each factor of the body model, *i.e.*, the body pose parameters $\theta_b$, the shape parameters $\beta$ and the per-part relative rotations $\theta_i$ with the generator loss expressed as:

$$\mathcal{L}_{\texttt{adv}} = \sum_k (D_k(\theta_b, \beta) - 1)^2$$

As stated in Section 6.1.2 SLAHMR last version uses HMR2.0 tracking system, named 4DHumans. The core idea of 4DHumans is the same as the PHALP, which detects people in individual frames and "lifts" them to 3D, predicting the 3D pose, the location in 3D space and the 3D appearance from the texture map. At each recursion step, these three parameters per person are predicted for the next frame and then the best matches between the top-down predictions and the bottom-up detections of people in that frame after lifting them in 3D are found. The current state of each tracked object is updated with new observations, and this process is repeated.

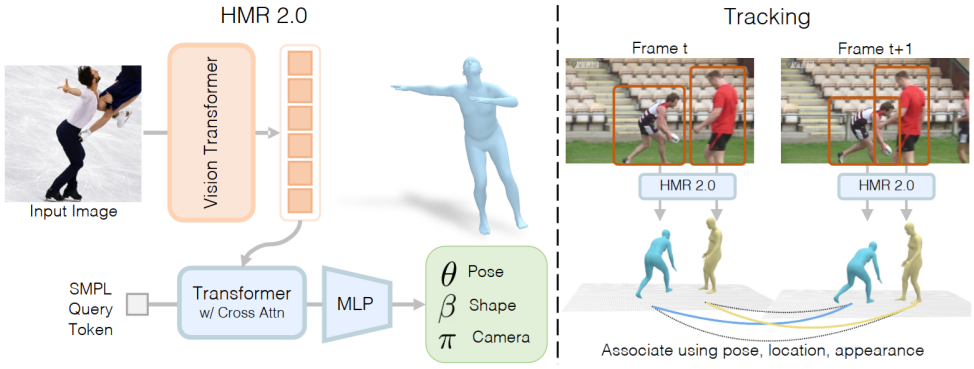



Figure 5.3: HMR2.0 Architecture. Figure from [16]

## 5.3 Architecture Details

The model that we train consists of one ViT image encoder and a transformer decoder. The ViT encoder is taken from the ViTPose model, which was pre-trained for the 2D joint detection task. It takes as input a $256 \times 192$ image and consists of 50 transformer layers. The encoder outputs $16 \times 12$ image tokens, each of dimension 1280. These tokens serve as the encoded representation of the input image for the decoder.

The transformer decoder has 6 layers, each with multi-head self-attention, multi-head cross-attention, and feed-forward blocks with layer normalization. It has a hidden dimension of 2048. Both the self-attention and cross-attention blocks use 8 heads, each with a dimension of 64. The feed-forward MLP has a hidden dimension of 1024.

For the SMPL-A parameters prediction, a 2048-dimensional learnable SMPL-A query token is fed into the transformer decoder. The decoder uses cross-attention to attend to the $16 \times 12$ image tokens from the ViT encoder. The output of the decoder is then passed through a linear layer to predict the final parameters. The output of the network consists of the pose ($\theta$), the shape ($\beta$), and the camera ($\pi$) parameters.

Table 5.1 shows the number of trainable parameters for the backbone, the SMPL-A head and the discriminator, for a total of 671M parameters.

| Name | Type | Number of Trainable Parameters |
|---|---|---|
| backbone | ViT | 630 M |
| smpla_head | SMPL-A Transformer Decoder Head | 39.5 M |
| discriminator | Discriminator | 1.8 M |

Table 5.1: Trainable Parameters for Model Components

## 5.4 Experimental Setup

### 5.4.1 Implementation Details

For the code of the methods we use **PyTorch** [46]. For both the training and fine-tuning of the model, we use the same configuration. The batch size is set to 16, we use **AdamW optimizer** [38] with a learning rate of $\mathbf{4 \cdot 10^{-5}}$, $\beta_1 = 0.9, \beta_2 = 0.999$ and a weight decay of $\mathbf{10^{-4}}$. The first training phase lasts for **2.5M steps**, while the fine-tuning phase lasts **1.5M steps**. For the different weights used during training we set the values to $\mathcal{L}_{\text{kp3D}} = \mathbf{0.05}, \mathcal{L}_{\text{kp2D}} = \mathbf{0.05}, \mathcal{L}_{\text{adv}} = \mathbf{0.0005}$ and the terms within $\mathcal{L}_{\text{smpl}}$ weigh **0.0015** and **0.001** for the $\beta$ and $\theta$, respectively.

### 5.4.2 Training Datasets and Annotations Preparation

A significant challenge in the field of 3D child shape and pose estimation stems from the paucity of training data, largely attributable to the sensitivity of these data. The datasets that contain child and baby images are very few, meaning that the datasets with annotations for 3D shape and pose estimation are even fewer.

For the training, we use the mixture of datasets used in the HMR2.0 model training, *i.e.*, Human3.6M [22], MPI-INF-3DHP [42], COCO [34] and MPII [1]. These datasets contain annotations for the bounding boxes and the 3D or 2D location of keypoints that are used for training. Three more datasets, InstaVariety [25], AVA [17] and AI Challenger [69], are used in the training of HMR2.0 with pseudo-ground truth annotations. Specifically, a detector [32] is used for the bounding boxes, then ViTPose [71] for the corresponding 2D keypoints and ProHMR [30] to get pseudo-ground truth SMPL parameters for the pose $\theta^*$ and the shape $\beta^*$ with camera $\pi^*$. Additionally, we use SyRIP [21] and Relative Human [62] datasets with pseudo-ground truth SMPL parameters from our optimization method and 2D keypoints from their ground-truth annotations. We observed that some of their annotations are erroneous in these two datasets. In such cases, we use the 2D keypoints from ViTPose. While this approach was generally effective, cer-

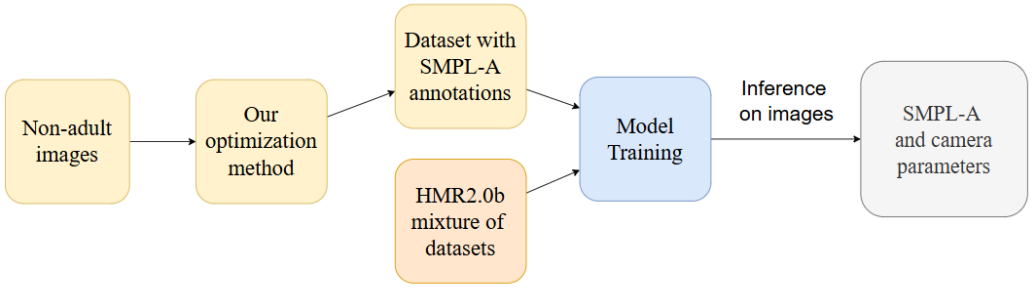



Figure 5.4: The pipeline of our proposed model. We begin from images of non-adult population, we extract annotations and train a model to regress the SMPL-A and camera parameters from a single image.

tain limitations, particularly with extreme poses and unclear facial features, prevented us from using the entire datasets for training. The training of HMR-like models requires high-quality, accurately annotated data to ensure robust generalization in complex scenarios. Figure 5.4 concludes how, from raw children images, we train our proposed model to estimate the 3D shape and pose from a single image.

Our method occasionally lacked quality when the human's face was not clear or visible. This is a critical issue because facial proportions are a key determinant in distinguishing between an infant and an adult. To mitigate this problem, we decided to constrain our training set to only include data where the ViTPose detections for face keypoints had a confidence score greater than 0.7.

This constraint offered a dual benefit and a clear trade-off:

- **Improved Data Quality:** It ensured that our model was trained on data where infants were correctly identified as infants, preventing them from being misclassified as adults.
- **Reduced Dataset Size:** This filter significantly reduced our training data by more than 50%, as less than half of the total images in the datasets met this strict criterion.

Since the $\beta$ parameter in our dataset has 11 components (including the SMPL-A interpolation weight $\alpha$), while the other training datasets utilize the standard SMPL model with 10 shape components, we implemented a data harmonization step. During the loading of samples from SMPL-based datasets, we automatically reshape the $\beta$ parameters to a consistent size of $\beta \in \mathbb{R}^{11}$ by setting the SMPL-A interpolation weight $\alpha$ to zero. This ensures

a uniform input shape for the model training. Finally, for the validation, we use a mixture of our validation dataset and COCO.

In Table 5.2 we show the weights for each dataset, *i.e.*, the probability that samples from this dataset are used for the first training phase. We assign a greater probability in our dataset since our main focus is on child and baby data. A smaller weight would diminish the ability of our model to learn these data since it would see only adult ones. In the fine-tuning phase, where the model has learned the 3D shape of children, we reduce the weight for our dataset so it sees more diverse examples to capture the 3D pose better. The weights are shown in Table 5.3.

In both phases, the validation datasets have the same weight.

Table 5.2: Training Dataset Configuration

| **Dataset Type** | **Dataset Name** | **Weight** |
|---|---|---|
| **TRAIN DATASETS** | | |
| H36M | H36M-TRAIN-WMASK | 0.05 |
| MPII | MPII-TRAIN-WMASK | 0.05 |
| COCO | COCO-TRAIN-2014-WMASK-PRUNED | 0.05 |
| MPI-INF-3DHP | MPI-INF-TRAIN-PRUNED | 0.05 |
| AVA | AVA-TRAIN-MIDFRAMES-1FPS-WMASK | 0.10 |
| AIC | AIC-TRAIN-WMASK | 0.10 |
| INSTA | INSTA-TRAIN-WMASK | 0.10 |
| Ours (from SyRIP and Relative Human) | **TRAIN Subset** | **0.50** |
| **VALIDATION DATASETS** | | |
| Ours (from SyRIP and Relative Human) | **VAL Subset** | **0.50** |
| COCO | COCO-VAL | 0.50 |

Table 5.3: Fine-Tuning Dataset Configuration

| **Dataset Type** | **Dataset Name** | **Weight** |
|---|---|---|
| **TRAIN DATASETS** | | |
| H36M | H36M-TRAIN-WMASK | 0.0875 |
| MPII | MPII-TRAIN-WMASK | 0.0875 |
| COCO | COCO-TRAIN-2014-WMASK-PRUNED | 0.0875 |
| COCO (ViTPose) | COCO-TRAIN-2014-VITPOSE-REPLICATE-PRUNED12 | 0.0875 |
| MPI-INF-3DHP | MPI-INF-TRAIN-PRUNED | 0.0875 |
| AVA | AVA-TRAIN-MIDFRAMES-1FPS-WMASK | 0.0875 |
| AIC | AIC-TRAIN-WMASK | 0.0875 |
| INSTA | INSTA-TRAIN-WMASK | 0.0875 |
| Ours (from SyRIP and Relative Human) | **TRAIN Subset** | **0.3000** |
| **VALIDATION DATASETS** | | |
| Ours (from SyRIP and Relative Human) | **VAL Subset** | **0.50** |
| COCO | COCO-VAL | 0.50 |

## 5.5 Evaluation

### 5.5.1 Evaluation Baselines

First, we examine whether our approach is better than the HMR2.0 checkpoint in child modeling, as it is the starting point of our approach. HMR2.0 has excellent performance in estimating the 3D shape and pose of humans from a single image. We also compare our method with ProHMR [30], a similar probabilistic method for human mesh recovery that uses a distribution of plausible 3D poses to mitigate the reconstruction ambiguity. Finally, we compare our method with BEV [62], a model that uses a modified SMPL-A model and demonstrates efficacy in the reconstruction of children.

### 5.5.2 Evaluation Datasets

Given that our primary objective is 3D shape and pose estimation for children and infants, the evaluation necessitates datasets within this domain. Unfortunately, as with the training phase, the scarcity of publicly available, annotated data persists. Therefore, our evaluation utilizes both test subsets from our test pool and an external dataset with challenging characteristics.

#### SyRIP and Relative Human

We use the test subsets of SyRIP and Relative Human. The SyRIP test set specifically contains 100 infant images. The Relative Human test set provides a broader challenge, consisting of 1836 multi-person images that include subjects of various ages. Both datasets contain ground-truth 2D keypoint annotations. The necessary 3D keypoints and shape parameters ($\beta$) for the evaluation split are estimated using our optimization method, thus providing the pseudo-ground-truth data for these datasets.

#### ChildPlay

Furthermore, we incorporate the ChildPlay dataset [63] as an external validation source. This dataset, which features videos of children interacting with adults, is characterized by frequent and significant occlusions and truncations. While the inherent pose ambiguity within these images makes the dataset unsuitable for inclusion in our primary training pool, it serves as a robust test of model generalization. For evaluation purposes, we randomly sample 1000 distinct frames from the ChildPlay videos to form a dedicated test set. ChildPlay dataset does not provide any annotations for

our evaluation. Therefore, we generate pseudo-ground-truth annotations for 2D keypoints using ViTPose, and for the 3D keypoints and shape parameters $\beta$ using our optimization method.

#### BabyRobot

Finally, we evaluate our method on the BabyRobot dataset. This dataset contains 1000 images of children aged 6 to 10 years old, where each child interacts with a robot and moves freely in the environment. It utilizes 3 cameras: one placed in front of the child and robot, one to the left, and one to the right. The dataset includes images from all the cameras for every child. To use these images for our evaluation, we generate pseudo-ground-truth annotations for the SMPL parameters using our optimization method and for the 2D keypoints using ViTPose. Some examples of images from the BabyRobot dataset are shown in Figure 5.5.

(a) Left camera (b) Front camera (c) Right camera

Figure 5.5: Example images from the 3 camera views of the BabyRobot dataset.

### 5.5.3 Evaluation Metrics

The performance of our models is rigorously evaluated using a combination of standard quantitative metrics and specialized domain-specific measurements.

The height of the human subjects offers a reliable metric for evaluating 3D shape estimation quality. Unlike adults, children and infants possess a specific, age-dependent range of possible heights. Since some of our evaluation datasets contain data exclusively from children or infants, it is possible to assess the reconstructed human height range. Furthermore, the dataset annotations enable a direct comparison between the pseudo-ground-truth height and the predicted height for each subject.

More specifically, we introduce two quantitative metrics based on height: the **Average Height Difference (AHD)** between the pseudo-ground-truth

and predicted height, and the **Average Percentage Height Difference (APHD)**, calculated as a percentage of the ground truth.

Let $N$ denote the total number of subjects, $H_i^*$ be the pseudo-ground-truth height (in centimeters), and $H_i$ be the predicted height for subject $i$. We calculate AHD and APHD as follows:

$$AHD = \frac{1}{N}\sum_{i=1}^{N}(H_i^* - H_i) \quad \text{(in cm)}$$

$$APHD = \frac{100\%}{N} \cdot \sum_{i=1}^{N}\frac{H_i^* - H_i}{H_i^*}$$

Absolute values are omitted to assess systemic bias; this reveals whether the model consistently over-predicts (positive difference) or under-predicts (negative difference) the height, rather than just the magnitude of the error.

To find both the predicted and the pseudo-ground-truth height of a person, we follow a simple procedure: we calculate the 3D mesh vertices from the SMPL (or SMPL-A) model using the predicted or pseudo-ground-truth shape parameters ($\beta$) and a zero pose ($\theta = \vec{0}$) and global orientation. The resulting mesh has no joint rotations, allowing the height to be calculated simply as the difference between the vertices with the maximum and minimum $y$-axis values. For all the datasets, the pseudo-ground-truth shape parameters are provided by our optimization method.

For the evaluation of our 3D pose and shape estimation model, we primarily use a common metric in the field, the **Mean Per Joint Position Error (MPJPE)**. MPJPE quantifies the average error between the predicted 3D joint locations and their corresponding ground-truth locations, specifically measured using the $L_2$ norm (Euclidean distance). The formula is given by:

$$\text{MPJPE} = \frac{1}{K}\sum_{i=1}^{K}\|J_i - J_i^*\|_2$$

where $K$ is the number of joints, $J_i$ is the predicted 3D location of joint $i$, and $J_i^*$ is the ground-truth location of the same joint. Since no dataset provides the ground-truth 3D location of the joints, we use our optimization method to get their pseudo-ground-truth location.

For the 2D pose estimation, we use **PCK** metric (**Percentage of Correct Keypoints**) with two different thresholds, 0.05 and 0.1. It is a metric for the 2D pose that uses the reprojected 2D keypoints of the generated 3D mesh. The PCK formula for a threshold $\tau$ is the following:

$$\text{PCK}_\tau = \frac{\sum_{i=1}^{N} \mathbb{I}\left(\frac{\|\mathbf{p}_i - \mathbf{p}_i^*\|_2}{D} \leq \tau\right)}{\sum_{i=1}^{N} \mathbb{I}(\mathbf{v}_i > 0)}$$

where $N$ is the number of keypoints, $\mathbb{I}$ an indicator function, $p_i$ and $p_i^*$ the predicted and ground-truth location of the keypoint $i$, $D$ a normalization factor equal to the length of the diagonal of the image and $\mathbf{v}_i$ the confidence score of the keypoint $i$ detection. The SyRIP and Relative Human datasets initially provide ground-truth 2D keypoint annotations. Conversely, the ChildPlay dataset only contains images, so we employ ViTPose to extract the necessary 2D keypoints along with their confidence scores. The same keypoint extraction process was also applied to the BabyRobot dataset.

In 3D vision research, it is also essential to assess models and their results qualitatively. Some subtle information, such as the plausibility of body proportions, interpenetration of limbs, or smoothness of the reconstructed mesh, is not fully captured by quantitative metrics; thus, visual evaluation remains a critical approach to comparing model quality.

### Subjective Evaluation

To provide a **quantitative assessment of qualitative performance**, we conducted a comprehensive subjective study. The primary objective was to compare the efficacy of our proposed method against two baselines HMR2.0 and BEV.

We developed a dedicated user questionnaire utilizing a carefully **compiled test set of 39 images**, sampled proportionally from the four primary evaluation datasets. For each image, we generated and included the corresponding 3D reconstructions produced by all three methods.

During the evaluation, each participant was presented with a **randomized pair of reconstructions** (stimuli) and was instructed to address three distinct evaluative criteria:

1. **Shape Fidelity:** Which reconstruction exhibits the most accurate body shape and topology?
2. **Pose Accuracy:** Which reconstruction demonstrates the most faithful estimation of the subject's 3D pose?
3. **Overall Efficacy:** Which reconstruction delivers the most effective and visually convincing result?

If the participant could not determine the best reconstruction between the two options, we provided a third possible answer: "Cannot determine".

The questionnaire provided a total of **25 paired comparisons** per participant. This comparative matrix was structured as follows:

- **Our Method vs. HMR2.0:** 10 comparisons
- **Our Method vs. BEV:** 10 comparisons
- **Baselines Comparison (HMR2.0 vs. BEV):** 5 comparisons

The comparisons between HMR2.0 and BEV served to establish an internal performance benchmark for the existing methods and a way to prevent participants from learning the identity of each method.

Based on the aggregated results from the survey, we derived comprehensive statistics concerning the preferred shape and pose estimation for each method and for every image. This analysis allows us to not only compare the overall efficacy of the proposed method against the baselines but also to investigate performance variance across different datasets and demographic groups (e.g., subject age).

# Chapter 6

# Experiments

## 6.1 Optimization-Based Experiments

Before selecting our optimization-based method as the primary approach for generating training annotations, we conducted several experiments using SMPLify-X and SLAHMR, applying multiple modifications to assess their performance on our target domain. Below, we summarize these experiments and their outcomes.

### 6.1.1 SMPLify-X

Using SMPLify-X as a baseline pipeline, we performed five variations to evaluate its applicability for children and babies.

#### Simple Optimization

As a baseline, we tested the standard SMPLify-X pipeline on our dataset. While the method performs well for older children, it fails for younger children and babies, with errors in both shape and pose. This limitation arises partly from inaccurate 2D keypoint detections due to the lack of child-specific examples in OpenPose training data. Since SMPLify-X minimizes 2D reprojection error, poor 2D detections propagate to incorrect 3D reconstructions.

#### SMPL-A Instead of SMPL

A major limitation of SMPLify-X for children is its reliance on adult-oriented models such as SMPL or SMPL-X. To address this, we replaced the model with SMPL-A, which introduces an additional $\alpha$ parameter for inter-

polating between adult and child mesh templates. Despite this modification, improvements were modest, and significant errors persisted.

### BEV for Shape Parameters

Since previous attempts failed to recover accurate shapes, we incorporated BEV, a method reported to provide superior shape estimates. We initialized SMPLify-X with BEV's shape parameters and kept them fixed during optimization. This approach yielded the best results among SMPLify-X variations for older children but remained unreliable for babies, often predicting adult-like proportions.

As we see in Figures 6.1 and 6.2, the shape of BEV is not accurate even in cases where the face keypoints are clear in the image. Also, the estimated pose is incorrect with fewer or more problems, in Figure 6.1 and Figure 6.2, respectively.

(a) (b)

Figure 6.1: Example of SMPL-A with BEV predicted shape parameters. The BEV predicted shape is not accurate, and the pose has been estimated incorrectly (the right arm).

### ViTPose for 2D Keypoint Detection

To mitigate errors from inaccurate 2D keypoints, we replaced OpenPose with ViTPose. This change improved pose estimation for challenging cases but did not fully resolve 3D shape inaccuracies. Figure 6.3 illustrates a direct comparison of SMPLify-X estimations when initialized using OpenPose versus ViTPose 2D keypoint predictions. The figure clearly demonstrates a marked improvement in the quality of the resulting 3D pose reconstruction when utilizing ViTPose.

(a) (b)

Figure 6.2: Example of SMPL-A with BEV predicted shape parameters. Inaccurate pose estimation from SMPLify-X.

(a) (b)

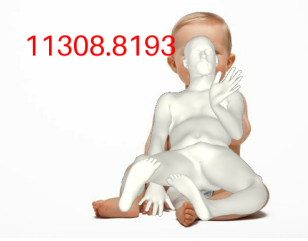
Figure 6.3: Comparison of SMPLify-X with OpenPose 2D Keypoints (a) and ViTPose 2D Keypoints (b). The predictions of ViTPose are more accurate, leading to a better 3D Pose Estimation.

### Grid Search on $\alpha$

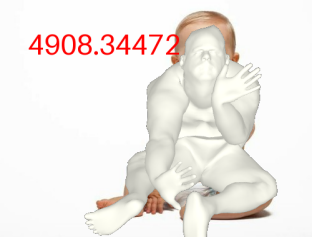
We explored a grid search over the SMPL-A interpolation parameter $\alpha$ within $[0, 1]$. For each value $\alpha \in \{0, 0.1, \ldots, 1\}$, we performed SMPLify-X optimization without adjusting $\alpha$. We then selected the reconstruction with the lowest fitting loss. Results showed minimal improvement compared to standard optimization. As shown in Figure 6.4, the optimal value for $\alpha$ is 0.9, which also yields the best qualitative results. For smaller $\alpha$ values, the model struggles to adapt to the infant's proportions, resulting in unrealistic reconstructions that resemble an adult body. As $\alpha$ increases, the reconstruction more accurately captures the infant's body shape. However, at the extreme value representing a pure infant template ($\alpha = 1$), the optimization fails, producing a 3D mesh that is not anatomically plausible. This highlights a critical balance in the model: too little or too much emphasis on the infant template can lead to poor results, underscoring that the optimal solution lies in a carefully tuned blend.


(a) $\alpha = 0$

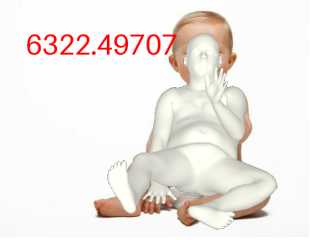

(b) $\alpha = 0.1$




(c) $\alpha = 0.2$




(d) $\alpha = 0.3$




(e) $\alpha = 0.4$


(f) $\alpha = 0.5$


(g) $\alpha = 0.6$


(h) $\alpha = 0.7$


(i) $\alpha = 0.8$


(j) $\alpha = 0.9$


(k) $\alpha = 1$


(l) $\alpha = 0.59$, from BEV shape

Figure 6.4: SMPLify-X results of the grid search on $\alpha$. The figures show the performance of the model with different parameter settings. The red value is the fitting loss for each case.

#### Extra experiments

Finally, we conducted two more experiments that produced poor results. In the first one, we used the SMIL model to test the pipeline. As we already knew, SMIL is an infant-specific model and not a universal model. Hence, the modeling of adults and children with SMIL using SMPLify-X pipeline is unsuccessful. The main problems are the proportions of the torso and the head, which are thinner in adults than in infants.

The second experiment was inspired by an observation from the grid search on the $\alpha$ experiment. Specifically, it appears that children of similar age are expressed with a similar $\alpha$ value. Knowing the age of each person in an image could give us some statistics on the distribution of the $\alpha$ value and the ages. Using this distribution, we would then select the $\alpha$ value based on it, either without performing a grid search or by performing a grid search on a smaller $\alpha$ value interval. The main issue with this experiment is the age estimation model. Unfortunately, these models work only with the face and are trained mostly on adult data, being unreliable in estimating the age of children and babies.

### 6.1.2 SLAHMR

We also evaluated SLAHMR adapting the pipeline to work with the SMPL-A model and conducted several experiments to improve performance for children and babies.

#### Simple Optimization with SMPL-A

First, we tested the SLAHMR pipeline using the SMPL-A model. Consistent with preliminary observations, SLAHMR outperformed SMPLify-X in most cases. However, for babies, shape estimation remained problematic, particularly when 2D facial keypoints were inaccurate or unclear. In such cases, incorrect face proportions led to suboptimal $\alpha$ values and unrealistic body shapes.

In Figure 6.5, we can see the improvement of SLAHMR with SMPL-A compared to SMPLify-X with SMPL-A. In the SLAHMR case, the body shape corresponds to an infant with a correct 3D pose, while in the SMPLify-X case, both the estimated 3D pose and shape are incorrect.

#### Grid Search on $\alpha$

To address sensitivity to $\alpha$, we performed a grid search similar to the SMPLify-X experiments. Specifically, SLAHMR was run 11 times with $\alpha \in$


Figure 6.5: Comparison of SLAHMR with SMPL-A (left) and SMPLify-X (right).

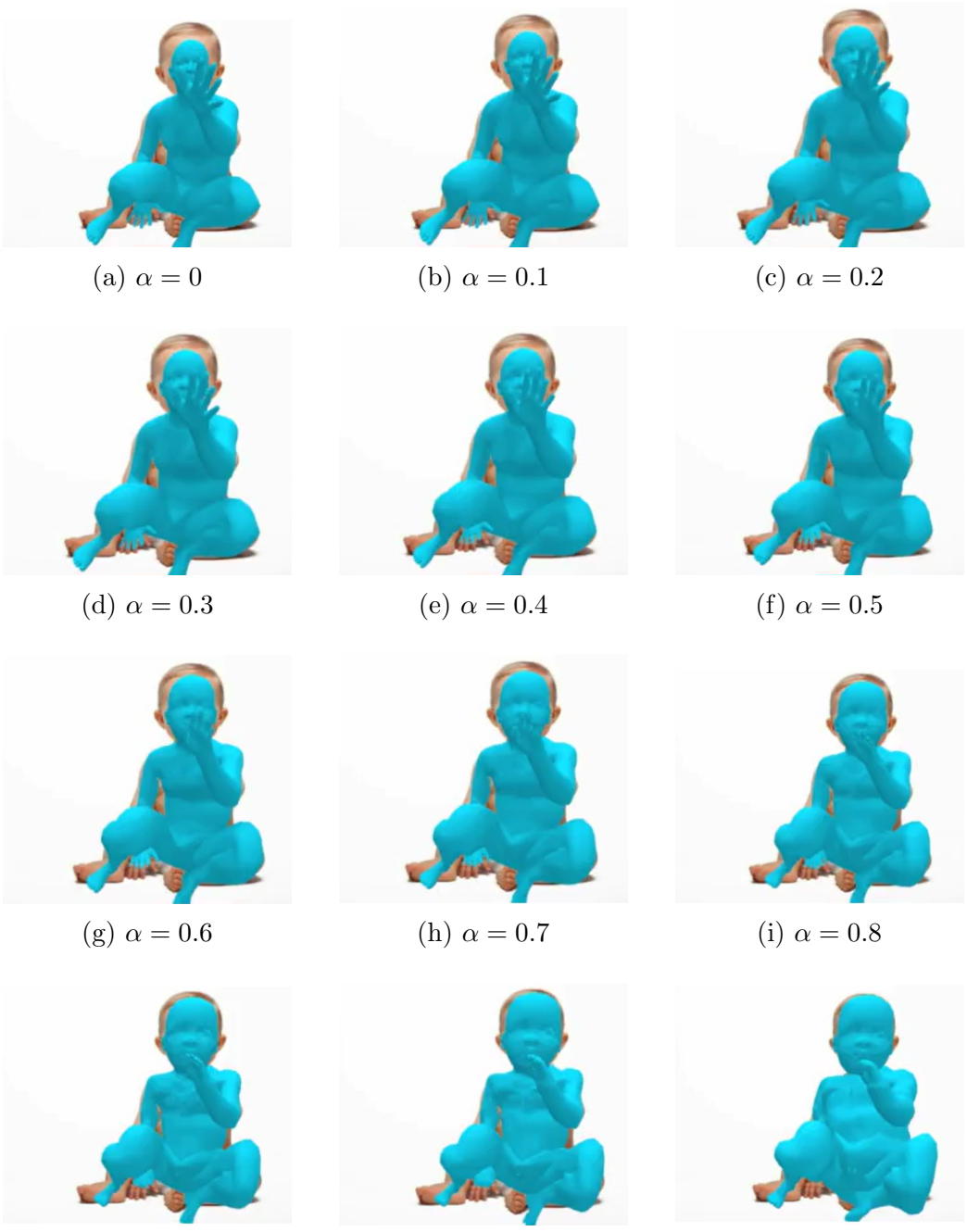
$\{0, 0.1, 0.2, \ldots, 1\}$, and the solution with the lowest total loss was selected. Although this approach occasionally improved results, it was computationally expensive and impractical for large-scale annotation. An example of the 11 results of the grid search is illustrated in Figure 6.6. Figures corresponding to small $\alpha$ values demonstrate a fundamental limitation of the model: they fail to capture the child's unique body morphology, instead regressing to a generic adult shape within the infant image. Conversely, as the $\alpha$ values increase, the model successfully approximates an infant's body shape. However, this accuracy comes at a cost, as these reconstructions display significant and unrealistic deformations in peripheral areas, most notably the feet, due to the complex pose of the infant. This suggests a trade-off between overall body shape accuracy and the preservation of fine details.

### Freezing Shape Parameters

We explored freezing the shape parameters $\boldsymbol{\beta}$ while optimizing the remaining parameters. Two strategies were tested: (1) setting all *betas* to zero and optimizing only $\alpha$, and (2) initializing shape with BEV predictions and keeping them fixed. Both strategies yielded similar results to those of previous experiments, without significant improvements in quality. An example is shown in subfigure (l) of Figure 6.6.

### Final configuration - Initialization of $\alpha$ at 1

As discussed in Section 3.4.4, $\alpha = 1$ corresponds to the child-specific template. Since our primary focus is on children and babies, we initialized $\alpha$ at 1 instead of 0, allowing the optimization to proceed from a child-centered



(a) $\alpha = 0$ (b) $\alpha = 0.1$ (c) $\alpha = 0.2$

(d) $\alpha = 0.3$ (e) $\alpha = 0.4$ (f) $\alpha = 0.5$

(g) $\alpha = 0.6$ (h) $\alpha = 0.7$ (i) $\alpha = 0.8$

(j) $\alpha = 0.9$ (k) $\alpha = 1$ (l) $\alpha = 1$, $\beta = \vec{0}$

Figure 6.6: SLAHMR results of the grid search on $\alpha$. Figure (l) shows the results when all the shape parameters are equal to 0, and the interpolation weight is equal to 1.

starting point. This modification consistently improved the quality of the reconstruction compared to previous configurations. This setup represents the finalized optimization-based method used to generate the pseudo-ground-truth annotations for the subsequent deep learning model training.

#### SyRIP-Specific Optimization

The previous experiments suggested that for baby images, $\alpha$ values near 0.9 often yield the best results. The SyRIP dataset contains many such cases where facial keypoints are unreliable, leading to poor shape estimation. To address this, we ran SLAHMR with $\alpha$ fixed at 0.9 while optimizing all other parameters. Results were comparable to those of our final configuration, with slight improvements for unclear faces. However, SLAHMR still struggled to produce plausible shapes for some baby images, even with this constraint.

## 6.2 Training Experiments

With all the data prepared, we conducted a series of experiments to optimize our model's performance. Our starting point was the HMR2.0b model, as the authors provided both the model checkpoint and the datasets. The model consists of three trainable parameter types: the Vision Transformer (ViT) backbone, the SMPL-A head, and the discriminator. The following sections briefly describe the different training experiments.

### 6.2.1 Fine-Tuning HMR2.0

The most direct approach was to fine-tune the HMR2.0 model checkpoint on our new dataset, which mostly contains data on infants and children. This method presented challenges due to the demographic differences between the original HMR2.0 training data (predominantly adults) and our dataset, as well as the disparity in dataset sizes. Our experiments involved modifying two factors: the training data and the model weights being optimized.

For the training data, we fine-tuned models using only our new dataset in some experiments, while in others, we used a mixture of our dataset and the original HMR2.0 datasets. We assigned a greater weight to our dataset in the mixture, aiming to extend the model's knowledge to children and babies, which our dataset primarily contains.

Regarding the model's parameters, we explored three configurations:

1. **All weights fine-tuned:** Initially, we fine-tuned all the weights of the HMR2.0 model, including the ViT backbone, the SMPL-A head, and

Figure 6.7: Example from fine-tuning. The model cannot learn the training data and fits an adult body to infants.

the discriminator. The results showed that the model failed to learn from our training data. An example is shown in Figure 6.7, where while the 3D pose is accurate, the body shape refers to an adult.

2. **Sequential fine-tuning:** In a second approach, we fine-tuned the ViT backbone separately, followed by the SMPL-A head and the discriminator. Despite this different training flow, the results were consistent with the case of training the entire model.

3. **Frozen backbone:** Finally, we froze the pre-trained ViT backbone and fine-tuned only the SMPL-A head and the discriminator. This approach also did not lead to an improvement in results.

We should note that since the HMR2.0 original checkpoint had a slightly different architecture (it uses SMPL instead of SMPL-A), we made a "model surgery" for the fine-tuning. More specifically, we expanded those parameters relative to the shape from 10 dimensions to 11. The initialization of these new weights and biases was either zero or a small random value in other experiments.

### 6.2.2 Training from Scratch

Given the unsuccessful fine-tuning attempts, we decided to evaluate the model's capacity to learn from the children and baby data by training it from scratch. Indeed, the model was able to be trained correctly and learned the training data. Therefore, a different training practice was followed.

Using the mixture of datasets as shown in Table 5.2, we trained a new model from scratch. Although the 3D shape was predicted with great results, the 3D pose exhibited issues. Since the ViT backbone is primarily responsible

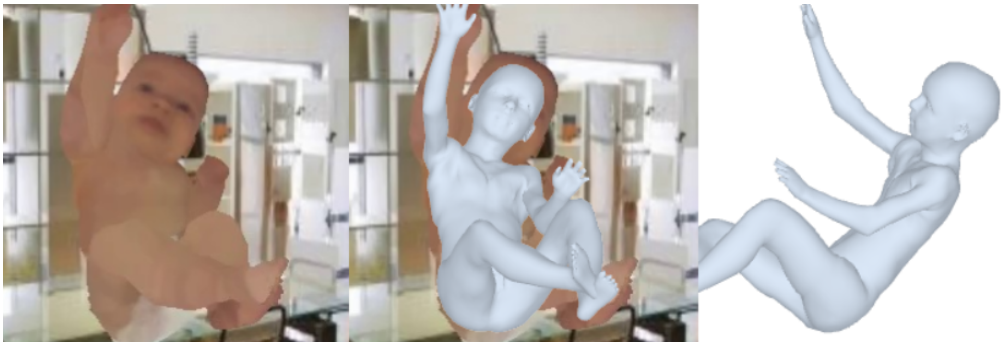


Figure 6.8: Example during the training of the model from scratch. The model learns the training data and correctly predicts the 3D shape and pose of the infant in the image.

for the 3D pose, we used the pre-trained backbone and trained the other two types of parameters from scratch. The results indicated that while the model could accurately estimate the pose of adults, it was unable to correctly fit to an infant, with notable inaccuracies in the head region.

### 6.2.3 Final Training Configuration

Due to the unsatisfactory results obtained from more straightforward training and fine-tuning approaches, we adopted a combined training and hybrid model strategy. Initially, we trained a model from scratch, utilizing both the original HMR2.0 training datasets and our custom datasets. While this model learned to estimate the 3D shape successfully, we observed persistent issues with pose estimation. Testing the original HMR2.0 checkpoint on the same images revealed its superior pose accuracy. Consequently, we devised a hybrid model: we combined the SMPL-A head and the discriminator from our newly trained model with the ViT backbone derived from the original HMR2.0 checkpoint (which was pre-trained on 2D keypoint estimation). Despite the immediate improvement in results, this new hybrid configuration requires further fine-tuning for a few epochs to ensure optimal adaptation and alignment between the distinct components. Following this fine-tuning, the model achieved the best results across all previous experiments, demonstrating high accuracy in both 3D shape and pose estimation. Figure 6.8 shows an example of a reconstruction during the training process, where it is clear that the model learns both the shape and the pose.

# Chapter 7

# Results and Discussion

## 7.1 Optimization-based Method

Finally, our first method seems to provide high-quality results in most cases. It can model all ages and correctly identifies whether there is a baby, a children or an adult. We provide some examples of how it works in Figure 7.1 for the ChildPlay, in Figure 7.2 for the SyRIP, in Figure 7.3 for the Relative Human, and in Figure 7.4 for the BabyRobot dataset. In all cases, the results are very satisfactory for every age group the method tries to model. It can capture difficult poses, multi-person scenarios, and natural or not obstacles and occlusions. The quality of these results led us to use this method for the annotations for our custom training dataset to train the deep learning model.

Figure 7.1: Examples of the optimization-based method from the Childplay dataset. The method can handle difficult poses and humans of every age.

Figure 7.2: Examples of the optimization-based method from the SyRIP dataset. The babies are modeled correctly, even in difficult poses.

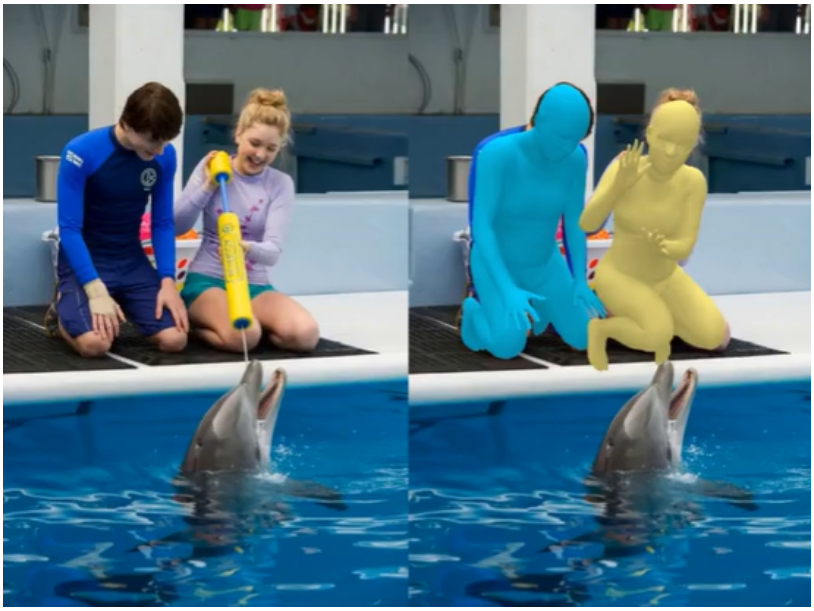


Figure 7.3: Examples of the optimization-based method from the Relative Human dataset.

Figure 7.4: Examples of the optimization-based method from the BabyRobot dataset.

## 7.2 Evaluation Results and Discussion

### Quantitative Evaluation

To assess our model's performance, we present a comprehensive evaluation across several key metrics and datasets.

**Height** Table 7.1 shows the general 3D evaluation results across all datasets using the AHD (Average Height Difference) and APHD (Average Per-Joint Height Difference) metrics.

Table 7.1: Model Evaluation using AHD (m) and APHD (%) metrics. Lower absolute value is better.

| Method | SyRIP | | Relative Human | | ChildPlay | | BabyRobot | |
|---|---|---|---|---|---|---|---|---|
| | AHD ↓ (m) | APHD ↓ (%) | AHD ↓ (m) | APHD ↓ (%) | AHD ↓ (m) | APHD ↓ (%) | AHD ↓ (m) | APHD ↓ (%) |
| ProHMR [30] | -1.011 | -161.76 | -0.098 | -17.73 | -0.477 | -92.01 | -0.562 | -56.69 |
| HMR2.0b [16] | -0.980 | -157.62 | -0.075 | -16.18 | -0.468 | -90.80 | -0.534 | -54.27 |
| BEV [62] | -0.528 | -91.8 | **-0.009** | -12.5 | **-0.067** | **-42.8** | -0.359 | -38.92 |
| **Our Model** | **-0.118** | **-25.97** | 0.088 | **-4.56** | -0.190 | -60.23 | **-0.354** | **-36.83** |

Based on these results, our proposed model demonstrates very good performance, particularly when considering subjects with non-adult anthropometry. **The overall lowest AHD and APHD confirm the superiority of our approach in handling diverse human scales**.

The results on the SyRIP dataset, which is dedicated exclusively to infants and toddlers, most clearly highlight this achievement. Here, **our model achieves a significantly lower APHD than all competing methods**, confirming the effectiveness of incorporating the SMPL-A model and specialized training strategies for accurately estimating highly non-adult body shapes. In Figure 7.5, we provide some examples of the fitting of our model compared to the HMR2.0b checkpoint and BEV. It is obvious that our model fits a baby's body to the images, while the original HMR2.0 provides an adult body, leading to an unrealistic shape and pose. The BEV results show an improved body shape for infants, but the estimated pose and the overall fitting quality are generally unreliable.

In the Relative Human and ChildPlay datasets, which contain a mix of children and adults, the results are more balanced across all methods. This moderation is primarily due to two factors: 1) the presence of severe occlusions and truncations, which limit the performance gains of all techniques, and 2) the high estimation accuracy of existing baseline methods on the adult subjects, given their extensive training on standard adult datasets. Despite these challenges, **our model maintains a clear performance advantage on these datasets except the ChildPlay dataset and BEV method, which shows a better APHD over our method**. This happens mainly for two reasons. The first one is that BEV in about 30% of the images did not detect any human and could not make its estimations. These subjects are not considered in the metric and have affected the results. Secondly, BEV

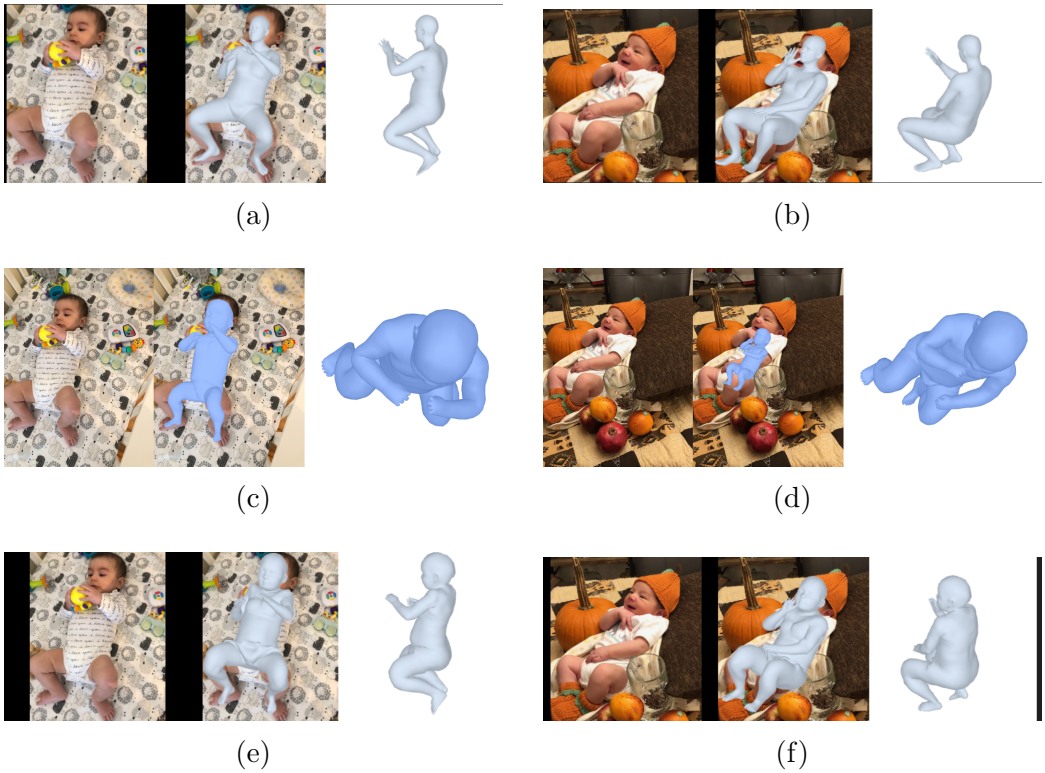


Figure 7.5: Examples of the proposed model ((e),(f)) compared to the HMR2.0b model ((a),(b)) and BEV ((c),(d)) in images from the SyRIP dataset.

is actually an effective method in child shape modeling and it can produce better results than existing methods. In Figures 7.6 and 7.7 we illustrate some results in Relative Human and ChildPlay images. Our model provides as good as the original's HMR2.0 results for adults and significantly better for children. Once again, the BEV results reveal the recurring issues with pose estimation, despite achieving high accuracy in estimating the children's shape.

The BabyRobot evaluation provides a compelling case study regarding the influence of age on model generalization since the model's performance remains comparable to the best baselines. This phenomenon is largely attributed to the older average age and resulting body proportions of the children within this dataset. As their height and limb-to-torso ratios begin to approximate adult metrics, the geometric priors embedded in general adult body models become surprisingly effective, enabling baseline methods to achieve results that nearly match our specialized approach. Moreover, the

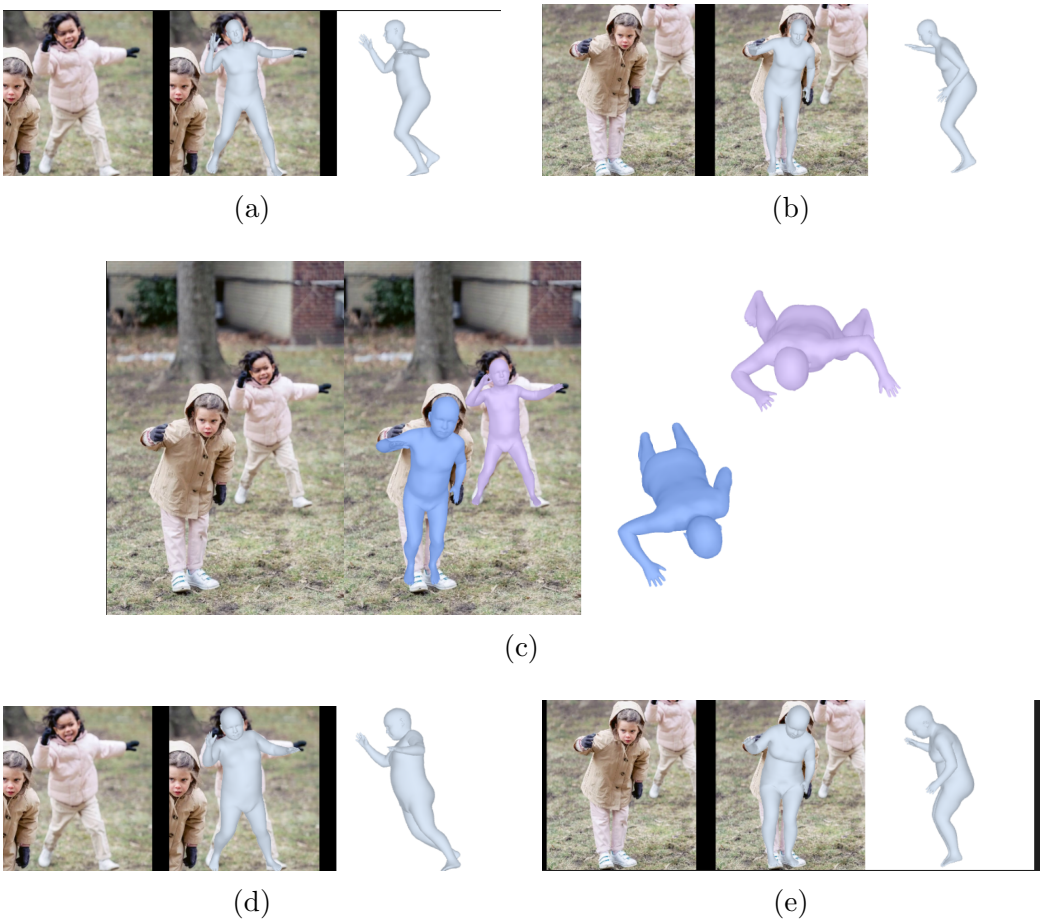
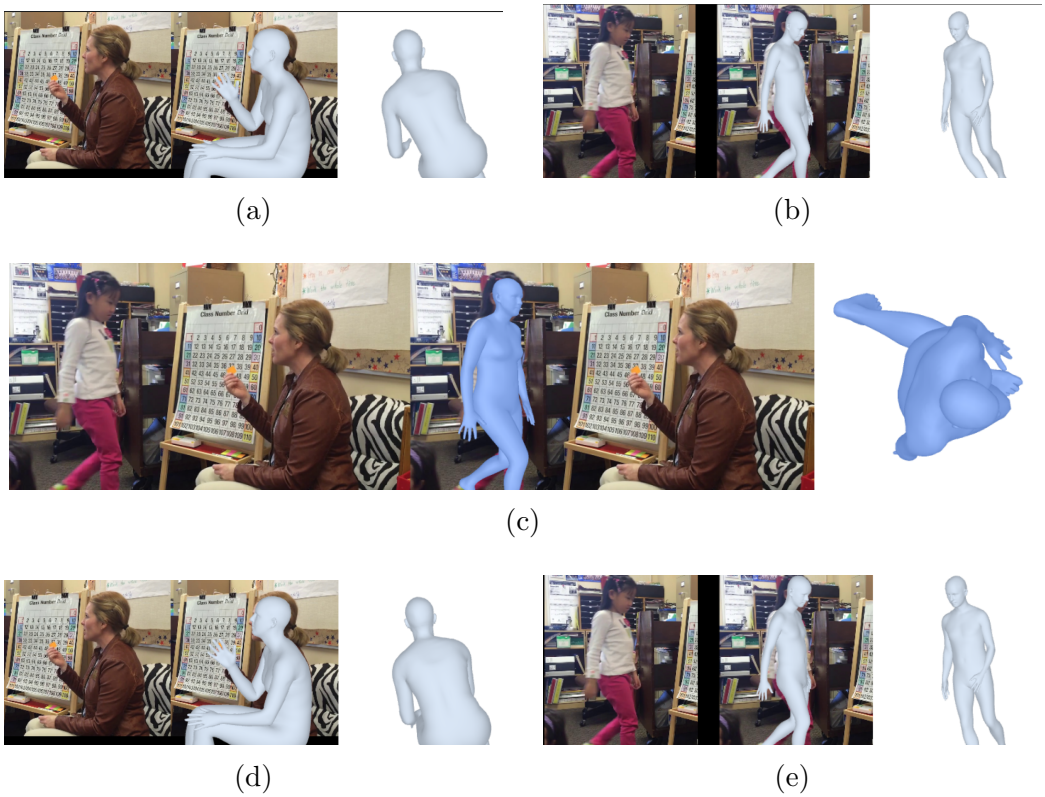


Figure 7.6: Examples of the proposed model ((d),(e)) compared to the HMR2.0b model ((a),(b)) and BEV ((c)) in images from the Relative Human dataset.

small amount of training data for these ages contributes to this result. Again, BEV results are very close to ours due to the use of SMPL-A, but our model is slightly better. Two visual examples comparing our method to the HMR2.0b checkpoint and BEV are illustrated in Figure 7.8. Crucially, **our method consistently captures the child's shape more accurately**. This leads to a more realistic pose, as the imposed adult body proportions of HMR2.0b often fail to map precisely to the child's specific poses. More specifically, in both presented cases, the legs of the children appear erroneously bent when modeled with HMR2.0b, while our model yields a visually more accurate and realistic pose. The illustration of the BEV proves the quantitative results in BabyRobot.

In Table 7.2, we present the average predicted height (in meters) from



(a) (b) (c) (d) (e)

Figure 7.7: Examples of the proposed model ((d),(e)) compared to the HMR2.0b model ((a),(b)) and BEV (c) in images from the ChildPlay dataset.

our method compared to the evaluation baselines. This specific analysis is conducted only on the SyRIP and BabyRobot datasets because these are the only two evaluation sets comprised exclusively of infants and children, respectively. Including other datasets with subjects from all age groups would lead to misleading conclusions regarding the model's accuracy in the non-adult domain.

| **Method** | **SyRIP** | **BabyRobot** |
|---|---|---|
| ProHMR | 1.712 | 1.718 |
| HMR2.0b | 1.705 | 1.717 |
| BEV | 1.249 | 1.528 |
| **Our Model** | **0.848** | **1.524** |

Table 7.2: Average predicted height in m.

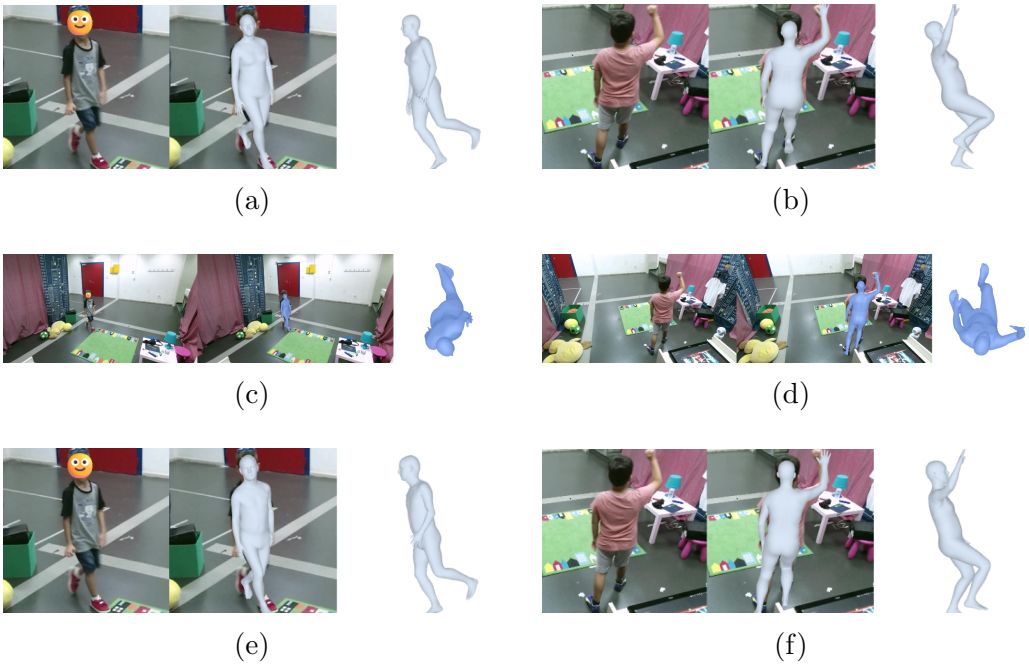


Figure 7.8: Examples of the proposed model ((c),(d)) compared to the HMR2.0b model ((a),(b)) in images from the BabyRobot dataset.

Based on the results, the superiority of our method is clear. The SyRIP dataset contains only infants, yet the predicted average heights reported by the baselines are significantly overestimated and highly unrealistic for this age group, suggesting a reliance on adult geometric priors. In contrast, our method, which is trained using specialized data and the SMPL-A model, yields **an average predicted height of 0.848m, providing a more realistic and biologically plausible result for infants**. Similarly, for the BabyRobot dataset, our approach provides a more realistic average height estimation, further validating the necessity of a child-specific body modeling strategy.

Table 7.3: 3D Pose Estimation. Model Evaluation using MPJPE (MPJPE in mm). Lower ↓ is better.

| **Method** | **SyRIP** | **ChildPlay** | **BabyRobot** |
|---|---|---|---|
| ProHMR [30] | 515.24 | 494.82 | 505.56 |
| BEV [62] | 452.73 | 424.41 | 380 |
| HMR2.0b [16] | **55.47** | **314.92** | **252.08** |
| **Our Model** | 287.84 | 318.26 | 258.51 |

**3D Pose** For 3D pose accuracy, we present the evaluation results in Table 7.3 using the primary metric, MPJPE (Mean Per Joint Position Error), comparing our model with the baselines across the three key datasets.

The evaluation of the 3D pose shows that **our model surpasses BEV**, the method that can estimate the 3D shape and pose of children and infants more accurately, in all datasets. A surprising result is that while HMR2.0 estimates infants in SyRIP with an average height of 1.70m, it achieves the lowest MPJPE in this dataset. This can be explained by the very good fitting of the SMPL model to the 2D keypoints from HMR2.0. Our model achieves **almost the same MPJPE with HMR2.0 in ChildPlay and BabyRobot,** showing its effectiveness in 3D pose estimation.

Table 7.4: 2D Pose Evaluation. PCK scores of projected keypoints at different thresholds. Higher ↑ is better.

| **Method** | **SyRIP** | | **Relative Human** | | **ChildPlay** | | **BabyRobot** | |
|---|---|---|---|---|---|---|---|---|
| | **@0.05** | **@0.1** | **@0.05** | **@0.1** | **@0.05** | **@0.1** | **@0.05** | **@0.1** |
| BEV [62] | 0.34 | 0.57 | 0.32 | 0.55 | 0.43 | 0.73 | 0.61 | 0.86 |
| HMR2.0b [16] | **0.79** | **0.98** | **0.48** | **0.62** | **0.76** | **0.94** | **0.97** | **0.99** |
| **Our Model** | 0.63 | 0.88 | 0.30 | 0.51 | 0.51 | 0.81 | 0.89 | 0.97 |

**2D Pose** Finally, in Table 7.4, we present the results of the 2D pose evaluation using the PCK (Percentage of Correct Keypoints) metric with two different thresholds (0.05 and 0.1). For this comparison, all available evaluation datasets are used to assess the robustness of the methods' 2D keypoint prediction.

Based on the results, **our model effectively estimates 2D pose across all datasets, outperforming the BEV method,** which is designed to model children, on every dataset except Relative Human. The superior performance of BEV on the Relative Human dataset is attributed to its inclusion of this dataset in its training. The HMR2.0b model consistently achieves the highest PCK score for 2D pose estimation across all datasets. While our model was trained on the same datasets as HMR2.0b, the difference in results can be attributed to architectural variations and the incorporation of the SMPL-A body model, in contrast to HMR2.0b's use of the standard SMPL model.

### Subjective Study

The subjective study was conducted with a sample of 30 anonymous participants from different academic backgrounds to achieve a more diverse

representation. Table 7.5 shows the cumulative results of the pairwise results, Our model vs BEV and Our model vs HMR2.0. In Table 7.6, we further analyze these numbers for pairwise comparison by question category, *i.e.*, shape, pose, and overall.

Table 7.5: Subjective Study Results: Cumulative Pairwise Comparison (Overall Win Rate). The total number of comparisons is 900 per pair.

| **Comparison Pair** | **Method** | **Win Rate (%)** |
|---|---|---|
| **Our Model vs. BEV** | **Our Model** | **75.89** |
| | BEV | 17.22 |
| | Cannot Determine | 6.89 |
| **Our Model vs. HMR2.0** | **Our Model** | **50.22** |
| | HMR2.0 | 32.44 |
| | Cannot Determine | 17.34 |

The results of the subjective user study show a clear improvement of our method over the baselines, HMR2.0 and BEV, in modeling non-adult populations. **Our model was preferred across every category tested: shape fidelity, pose accuracy, and overall quality, when compared head-to-head with both BEV and HMR2.0.**

More specifically, **our model was selected in approximately 75% of the comparison votes against BEV in all categories**, demonstrating a significantly higher perceived quality of reconstruction by human evaluators. When compared against the HMR2.0 model, our model still outperforms the baseline, though the margin is smaller, accompanied by a greater percentage of indecisive votes. The fact that our model maintains an advantage, particularly in the pose category, is encouraging given HMR2.0's established performance in 3D pose estimation. For completeness, we note that in the direct comparison between BEV and HMR2.0, the participants overwhelmingly preferred the HMR2.0 reconstructions, confirming its established performance and highlighting the need for improvement from BEV.

Aggregating the results across all categories, we calculated the cumulative win rate. When considering only the comparisons involving our model, *i.e.*, eliminating the BEV vs. HMR2.0 comparisons to avoid misleading totals, **our method was preferred in more than 60% of the total answers, unequivocally providing the best overall reconstructions**. These aggregated results also show that participants perceived our shape estimation as more accurate than our pose estimation—a finding directly confirmed by

Table 7.6: Subjective Study Results: Pairwise Comparison Win Rates by Category. The total number of comparisons for each pair is 300 per category for all Our Model vs. baseline pairs (Our Model vs. BEV and Our Model vs. HMR2.0) and 150 per category for the HMR2.0 vs. BEV pair.

| Category | Comparison Pair | Method | Win Rate (%) |
|---|---|---|---|
| **Shape** | BEV vs. Our Model | **Our Model** | **76.33** |
| | | BEV | 16.33 |
| | | Cannot Determine | 7.34 |
| | HMR2.0 vs. Our Model | **Our Model** | **52.33** |
| | | HMR2.0 | 29.67 |
| | | Cannot Determine | 18.00 |
| | HMR2.0 vs. BEV | **HMR2.0** | **69.33** |
| | | BEV | 22.00 |
| | | Cannot Determine | 8.67 |
| **Pose** | BEV vs. Our Model | **Our Model** | **76.67** |
| | | BEV | 20.00 |
| | | Cannot Determine | 3.33 |
| | HMR2.0 vs. Our Model | **Our Model** | **49.00** |
| | | HMR2.0 | 35.33 |
| | | Cannot Determine | 15.67 |
| | HMR2.0 vs. BEV | **HMR2.0** | **76.00** |
| | | BEV | 19.33 |
| | | Cannot Determine | 4.67 |
| **Overall** | BEV vs. Our Model | **Our Model** | **74.67** |
| | | BEV | 15.33 |
| | | Cannot Determine | 10.00 |
| | HMR2.0 vs. Our Model | **Our Model** | **49.33** |
| | | HMR2.0 | 32.33 |
| | | Cannot Determine | 18.34 |
| | HMR2.0 vs. BEV | **HMR2.0** | **72.00** |
| | | BEV | 19.33 |
| | | Cannot Determine | 8.67 |

our quantitative evaluation metrics. The final results lead to the strong conclusion that **the proposed model establishes a clear subjective performance advantage over both baselines, with a substantial preference gap of over 10% in win rate for the total comparison votes**.

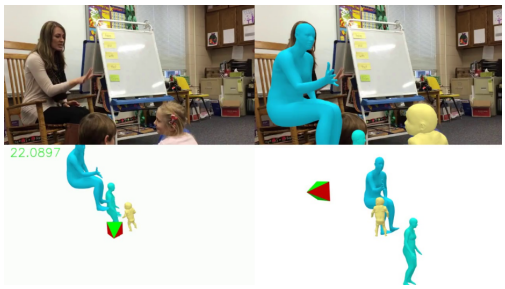
## 7.3 Limitations

Despite achieving promising results in the estimation of 3D shape and pose, the visualization and quantitative evaluation of our model reveal several inherent limitations within the current methodology, particularly when applied to challenging imagery.

First, our methods inevitably contend with the depth and pose ambiguity inherent to 3D reconstruction from a single 2D image. This ambiguity is severely exacerbated when significant occlusion (whether self-occlusion or scene occlusion) or truncation is present. Such failure examples are illustrated in Figure 7.9.

As the evaluation results confirm, 3D pose estimation remains a significant challenge in datasets featuring both occlusions and non-adult subjects. Specifically, in these challenging evaluation subsets, the MPJPE metric reached values of approximately 10cm. An average error of 10cm between the predicted and ground-truth joint locations represents a substantial error, highlighting a critical area for future improvement.


(a) Failure in 3D pose estimation due to natural occlusions from the infant's pose.



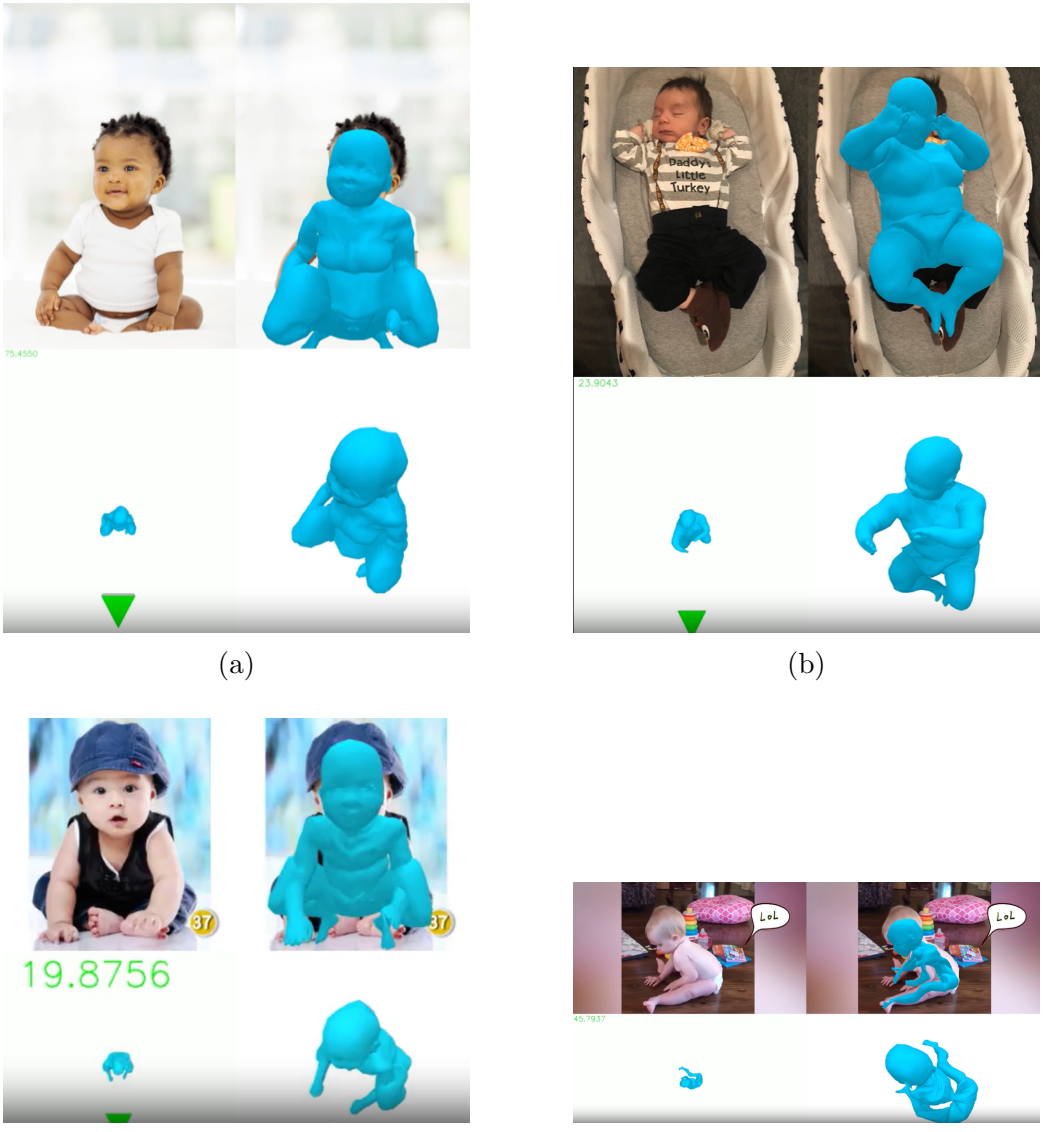

(b) Failure of 3D Pose Estimation due to Severe Truncation. The model estimates a SMPL body in a standing pose because the image is cropped to only show the child's face, while a human understands that the children are sitting.

Figure 7.9: Failures of our models in 3D pose estimation.

Furthermore, the model's capacity for accurate child modeling is significantly hampered by the scarcity of child-specific training data and the disproportionately large volume of adult examples in the current dataset. As previously explained, the sensitivity and ethical constraints surrounding children's data hinder the development of large-scale, child-specific datasets. Decreasing the size of the adult training sets would, however, deteriorate the generalization quality of the results due to a reduction in observed exam-

ples. Consequently, there is an unavoidable trade-off between generalization quality and child-specific accuracy.

It should be mentioned that despite the general good performance of our methods, the ability of the SMPL-A model to accurately describe an infant can be problematic in combination with the optimization process. These limitations are visualized in Figure 7.10, where complex poses and occlusions lead to anatomically implausible 3D shape estimations for the infants.




(a) (b)

(c) (d)

Figure 7.10: Visualizing the optimization method's limitations on infant data. Challenging input conditions, including severe physical occlusions and complex poses, frequently result in anatomically implausible 3D body shape estimations.

# Chapter 8

# Conclusion and Future Work

## 8.1 Summary of Contributions

In this thesis, we address the critical challenge of 3D human shape and pose estimation with a specific focus on the non-adult population. Current state-of-the-art methods, primarily designed and trained on adult bodies, exhibit significant performance degradation when evaluated on images of children and infants due to the reliance on adult-centric geometric priors (e.g., the standard SMPL model). Motivated by this gap, we propose a novel, two-stage methodological approach that leverages and significantly improves existing deep learning systems for robust application across all human age groups.

We introduce a specialized optimization-based shape and pose estimation method, adapted specifically for use with the SMPL-A body model, which is highly effective across diverse adult and non-adult proportions. This method is crucially employed to generate high-quality pseudo-ground-truth annotations for existing public datasets that feature children and infants, directly addressing the critical scarcity of annotated non-adult training data. We then utilize this newly augmented dataset, combined with a diverse mixture of predominantly adult datasets, to train a novel, specialized HMR-like transformer-based neural network. This network estimates 3D shape and pose via single-image regression, successfully integrating a robust pre-trained ViT backbone with a customized prediction head tailored for the 11-parameter SMPL-A body representation.

Our approach quantitatively and qualitatively demonstrates excellent performance, establishing a new benchmark against modern, similar works in the challenging domain of non-adult 3D estimation. Furthermore, the subjective user study validates the perceived quality and anatomical plausibility of our

reconstructions among a diverse sample of participants. Finally, we establish a viable methodology for ethically releasing sensitive data by sharing 3D human reconstructions instead of raw imagery. This process inherently anonymizes the identity of the subjects (children and infants) while providing accurate 3D body and motion information essential for subsequent motion analysis and action recognition tasks. The BabyRobot dataset exemplifies this approach, offering 3D reconstructions of children interacting with robots, complete with diverse actions, gestures, and spatial movements.

## 8.2 Future Work

Our current methodology utilizes the SMPL-A model, which is limited to modeling the human body's shape and pose. To advance this work and achieve a more comprehensive representation of human form, several key areas will be addressed in future research.

Future work will incorporate more sophisticated and holistic models, such as SMPL+H and SMPL-X, to enable the modeling of articulated hands and facial expression, respectively. This will provide a more complete, and arguably socially relevant, representation of the human body, capturing crucial details involved in interaction and expression.

However, this expansion introduces significant ethical and technical challenges, particularly when extending to sensitive populations like children and babies. Modeling features like facial expression and identity necessitates meticulous attention to data privacy, consent, and subject well-being. Research in this domain must prioritize these factors above purely technical performance.

To overcome the inherent limitations of adult-centric body models, the development of child-specific body models is critical. Since children and babies exhibit distinct anthropometric proportions and a unique shape-space manifold compared to adults, generic adult models like SMPL-A often introduce unrealistic deformations or anatomical inaccuracies. While the creation of such a model promises a significant improvement in result quality, it is contingent upon securing the high-quality, dense 3D scan training data that remains scarce in this sensitive domain.

Last but not least, the use of different types of input information can drastically improve the quality of 3D reconstructions. Specifically, leveraging temporal data (video), since a person's shape is constant across multiple frames, can constrain the optimization for better shape estimation. Furthermore, the use of multiple views of the same scene (multi-view cameras) is beneficial for both pose and shape estimation, as different spatial perspec-

tives reduce 3D ambiguity and yield more accurate reconstructions. These advanced methodologies can then be applied to existing datasets, such as BabyRobot, which contains videos from multiple cameras. This ultimately allows us to create highly accurate 3D meshes of people that can be used as high-fidelity ground-truth annotations in future 3D shape and pose estimation tasks.

# Bibliography


[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," *ACM Transactions on Graphics (TOG)*, vol. 24, p. 408–416, 2005.

[3] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[4] J. T. Barron, "A general and adaptive robust loss function," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] R. Bashirov, A. Ianina, K. Iskakov, Y. Kononenko, V. Strizhkova, V. Lempitsky, and A. Vakhitov, "Real-time rgbd-based extended body pose estimation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *European Conference on Computer Vision (ECCV)*, 2016.

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 43, p. 172–186, 2019.

[8] C.-H. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] W. contributors, "Pinhole camera model — Wikipedia, the free encyclopedia," 2025, [Online; accessed 19-September-2025]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pinhole_camera_model&oldid=1286010356

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[11] Z. Dong, J. Song, X. Chen, C. Guo, and O. Hilliges, "Shape-aware multi-person pose estimation from multi-view images," in *International Conference on Computer Vision (ICCV)*, 2021.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[13] S. K. Dwivedi, Y. Sun, P. Patel, Y. Feng, and M. J. Black, "TokenHMR: Advancing human mesh recovery with a tokenized pose representation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] T. Fan, K. V. Alwala, D. Xiang, W. Xu, T. Murphey, and M. Mukadam, "Revitalizing optimization for 3d human pose and shape estimation: A sparse constrained formulation," in *International Conference on Computer Vision (ICCV)*, 2021.

[15] U. J. Ganai, A. Ratne, B. Bhushan, and K. Venkatesh, "Early detection of autism spectrum disorder: gait deviations and machine learning," *Scientific Reports*, vol. 15, p. 873, 2025.

[16] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," in *International Conference on Computer Vision (ICCV)*, 2023.

[17] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Müller-Felber, and A. Sebastian Schroeder, "Learning an infant body model from rgb-d data for accurate full body motion analysis," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018.

[20] M. Hofmann and D. M. Gavrila, "Multi-view 3d human pose estimation in complex environment," *International Journal of Computer Vision*, vol. 96, pp. 103–124, 2012.

[21] X. Huang, N. Fu, S. Liu, and S. Ostadabbas, "Invariant representation learning for infant pose estimation with small data," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021.

[22] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, pp. 1325–1339, 2013.

[23] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed., 2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[24] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] M. Keller, K. Werling, S. Shin, S. Delp, S. Pujades, C. K. Liu, and M. J. Black, "From skin to skeleton: Towards biomechanically accurate 3D digital humans," *ACM Transactions on Graphics (TOG)*, vol. 42, pp. 1–12, 2023.

[27] N. Kojovic, S. Natraj, S. P. Mohanty, T. Maillart, and M. Schaer, "Using 2d video-based pose estimation for automated prediction of autism

spectrum disorders in young children," *Scientific Reports*, vol. 11, p. 15069, 2021.

[28] F. Koleini, M. U. Saleem, P. Wang, H. Xue, A. Helmy, and A. Fenwick, "Biopose: Biomechanically-accurate 3d pose estimation from monocular videos," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[29] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *International Conference on Computer Vision (ICCV)*, 2019.

[30] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *International Conference on Computer Vision (ICCV)*, 2021.

[31] P. Koutras, G. Retsinas, and P. Maragos, *Deep Learning for Computer Vision Applications*. Kallipos, Open Academic Editions, 2025, [Chapter]. [Online]. Available: https://hdl.handle.net/11419/15132

[32] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *European Conference on Computer Vision (ECCV)*, 2022.

[33] Z. Liao, J. Zhu, C. Wang, H. Hu, and S. L. Waslander, "Multiple view geometry transformers for 3d human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[34] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[35] Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang, and X. Wang, "Deep dual consecutive network for human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, pp. 248:1–248:16, 2015.

[37] M. M. Loper, N. Mahmood, and M. J. Black, "Mosh: Motion and shape capture from sparse markers," *ACM Transactions on Graphics (TOG)*, vol. 33, pp. 220–1, 2014.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2017.

[39] P. Maragos, *Topics in Computer Vision and Machine Learning*. Kallipos, Open Academic Editions, 2025, [Postgraduate textbook].

[40] ——, *Image Analysis and Computer Vision*, 2025. [Online]. Available: http://cvsp.cs.ntua.gr/courses/vision/material.shtm

[41] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *International Conference on Computer Vision (ICCV)*, 2017.

[42] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *International Conference on 3D Vision (3DV)*, 2017.

[43] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[44] A. A. A. Osman, T. Bolkart, and M. J. Black, "STAR: A sparse trained articulated human body regressor," in *European Conference on Computer Vision (ECCV)*, 2020.

[45] S. Park, J. Hwang, and N. Kwak, "3d human pose estimation using convolutional neural networks with 2d pose information," in *European Conference on Computer Vision (ECCV)*, 2016.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[47] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "AGORA: Avatars in geography optimized for regression analysis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[48] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face,

and body from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[49] G. Pavlakos, J. Malik, and A. Kanazawa, "Human mesh recovery from multiple shots," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Harvesting multiple views for marker-less 3d human pose annotations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[52] J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik, "Tracking people by predicting 3d appearance, location and pose," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[53] M. Rajchl, M. C. H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 36, pp. 674–683, 2017.

[54] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, "Humor: 3d human motion model for robust pose estimation," in *International Conference on Computer Vision (ICCV)*, 2021.

[55] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics (ToG)*, vol. 36, 2017.

[56] A. Roussos and P. Maragos, *Three-Dimensional Modeling of Deformable Objects*. Kallipos, Open Academic Editions, 2025, [Chapter]. [Online]. Available: https://hdl.handle.net/11419/15135

[57] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[58] B. Sapp, A. Toshev, and B. Taskar, "Cascaded models for articulated pose estimation," in *European Conference on Computer Vision (ECCV)*, 2010.

[59] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A joint model for 2d and 3d pose estimation from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[60] B. T. Soroush Mehraban, Vida Adeli, "Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[61] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[62] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3D people in depth," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[63] S. Tafasca*, A. Gupta*, and J.-M. Odobez, "Childplay: A new benchmark for understanding children's gaze behaviour," in *International Conference on Computer Vision (ICCV)*, 2023.

[64] Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[65] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[67] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *European Conference on Computer Vision (ECCV)*, 2008.

[68] Wikipedia contributors, "Coordinate system — Wikipedia, the free encyclopedia," 2025, [Online; accessed 19-September-2025]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Coordinate_system&oldid=1296595106

[69] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, and Y. Wang, "Large-scale datasets for going deeper in image understanding," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2019.

[70] Y. Xia, X. Zhou, E. Vouga, Q. Huang, and G. Pavlakos, "Reconstructing humans with a biomechanically accurate skeleton," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[71] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[72] K. Yan, F. Wang, B. Qian, H. Ding, J. Han, and X. Wei, "Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[73] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa, "Decoupling human and camera motion from videos in the wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[74] M. Zhang, Y. Zhou, X. Xu, Z. Ren, Y. Zhang, S. Liu, and W. Luo, "Multi-view emotional expressions dataset using 2d pose estimation," *Scientific Data*, p. 649, 2023.

[75] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3d human reconstruction from a single image," in *International Conference on Computer Vision (ICCV)*, 2019.

[76] S. Zuffi, A. Kanazawa, D. Jacobs, and M. J. Black, "3D menagerie: Modeling the 3D shape and pose of animals," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.