ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Image Processing And Computer Vision

3D HUMAN SHAPE AND POSE ESTIMATION FROM MULTI-VIEW IMAGES

CANDIDATE
Alessandro Folloni

SUPERVISOR
Prof. Samuele Salti

CO-SUPERVISOR
Matteo Fabbri, PhD

Academic year 2023-2024

Session 1st

Contents

List of Figures

List of Tables

Abstract

3D human shape and pose estimation is at the core of this thesis, which presents an end-to-end pipeline for reconstructing accurate body models from multi-view images of fitness exercises.

The process begins with videos from four cameras, from which a pre-trained model extracts 2D keypoints for each single frame. These keypoints are then fused by a specialized neural network to reconstruct precise 3D positions, overcoming the limitations of monocular methods. In a subsequent step, a regression model computes SMPL-X parameters to provide a detailed representation that captures both the overall pose and intricate body features.

Overall, integrating 2D detection, multi-view 3D reconstruction, and SMPL-X modeling establishes a solid foundation for innovative systems in human motion analysis and optimization, with promising applications in virtual coaching and fitness.

Chapter 1

Introduction

1.1 Context and Motivation

Over the last decade, computer vision and artificial intelligence (AI) have significantly advanced the field of human motion analysis. Deep learning methods have enabled highly accurate activity recognition, action classification, and pose estimation by extracting complex features from large volumes of visual data. In this context, understanding human behavior and accurately estimating poses are central tasks with applications in sports, medical analysis and more.

A growing trend in these fields is the shift from two-dimensional (2D) keypoints detection to three-dimensional (3D) representations of the human body. Although 2D methods are mature and perform well in many cases, they cannot capture the full depth and complexity of real-world motions—especially when accurate measurements of joint angles or limb orientations are desired. For instance, in fitness training or physical therapy, subtle postural deviations may not be evident from a single-camera view or a 2D projection. Accurate 3D pose estimation addresses these limitations by providing a robust assessment of human movement in free space.

Parametric body models such as SMPL [20] and SMPL-X [27] have

gained traction in this transition to 3D. Unlike simple skeleton-based approaches, these models integrate both the shape and pose of the body into a continuous representation, enabling the creation of realistic and personalized body models. Despite their promise, applying SMPL or SMPL-X in real-world settings presents many challenges.

1.2 Problem Definition

The primary objective of this thesis is to develop a pipeline that accurately retrieves SMPL-X parameters from multi-view video sequences depicting fitness exercises. In our typical setup, we have access to the recordings of multiple cameras that simultaneously recorded a subject performing various movements. From this raw footage, our goal is not only to obtain a 3D skeleton but also to derive a detailed parametric model that captures the subject's body shape and pose at specific instants.

To achieve this, our pipeline is structured in distinct stages:

1. **2D Keypoints Extraction:** In the first phase, an object detection model (e.g., YOLOv8s) is employed to extract 2D keypoints from each camera view. These keypoints, which represent crucial joint locations, form the foundational data for subsequent processing.

2. **3D Pose Estimation:** The extracted 2D keypoints are integrated into our dataset and then employed to get the corresponding 3D keypoints. This is made through a specialized neural network, which converts these spatial coordinates into a coherent 3D skeleton. This stage leverages geometric information across multiple views to resolve depth ambiguities.

3. **SMPL-X Parameter Regression:** With a similar process, a Transformer-based model processes the 3D keypoints and regresses the detailed SMPL-X parameters. This transformation converts the sparse

3D joint data into a continuous high-dimensional representation that encapsulates both body shape and pose.

1.3 Challenges

This objective introduces several technical and methodological challenges relevant to our work:

- **Data Synchronization:** With multiple cameras, precise synchronization of video streams is essential. Even minor temporal misalignment can lead to significant errors in the 3D reconstruction. Fortunately, our controlled fitness dataset helps mitigate this issue.

- **Robust 2D Keypoints Extraction:** The reliability of the entire pipeline depends on the accurate detection of 2D keypoints. Although frameworks like YOLO deliver high accuracy, adverse conditions, such as low-light environments, can degrade performance. However, in our controlled setting, these challenges are minimized.

- **Effective Conversion from 2D to 3D and to SMPL-X:** Transitioning from 2D keypoints to a 3D skeleton, and then to a detailed parametric model, is inherently complex. Each stage requires specialized models to capture the necessary spatial relationships and ensure that the final SMPL-X parameters are anatomically consistent. This sequential conversion is critical to overcome the limitations of monocular methods and to achieve a detailed understanding of human motion.

- **Computational Complexity:** Processing high-resolution multi-view videos, extracting keypoints, reconstructing 3D poses, and fitting a detailed parametric model are all computationally intensive tasks. Our approach leverages optimized models along with the substantial computational power to balance speed and accuracy.

Overall, our work demonstrates that by sequentially extracting 2D keypoints, reconstructing a robust 3D pose, and accurately regressing SMPL-X parameters, it is possible to develop an effective pipeline for fitness applications and, more generally, estimating the human model's parameters.

1.4 GoatAI

For this thesis, I completed my internship at GoatAI, an innovative startup based in Modena, Italy, and born inside the University of Modena and Reggio Emilia. The company focuses on developing advanced solutions in artificial intelligence and computer vision, specializing in human behavior analysis and 3D pose estimation for fitness applications. My work at GoatAI involved collaborating with a multidisciplinary team to address practical challenges in real-world data processing and model deployment, which directly informed the development of the pipeline presented in this thesis.

Chapter 2

Related Works

2.1 Human Behavior Understanding in Fitness Scenarios

Human Behavior Understanding (HBU) broadly refers to AI-driven tasks aimed at analyzing, interpreting, and predicting human actions in diverse environments [33, 30]. Historically, HBU found early applications in surveillance, social robotics, and retail analytics, yet its relevance has expanded in recent years to areas like healthcare, sports, and fitness. The ability to automatically perceive and interpret how people move, interact, or perform activities holds particular promise in fitness-related contexts, where precise motion tracking and feedback can significantly enhance training outcomes.

In Figure 2.1, we see a conceptual flow [25] where raw video data is fed into an action recognition algorithm, which can then supply detailed information about the subject.

2.1.1 Growth of Virtual Coaching and At-Home Fitness

A convergence of technological and societal factors—such as improved internet connectivity, affordable cameras, and an increased emphasis on personal wellness—has propelled the adoption of virtual coaching platforms [7].

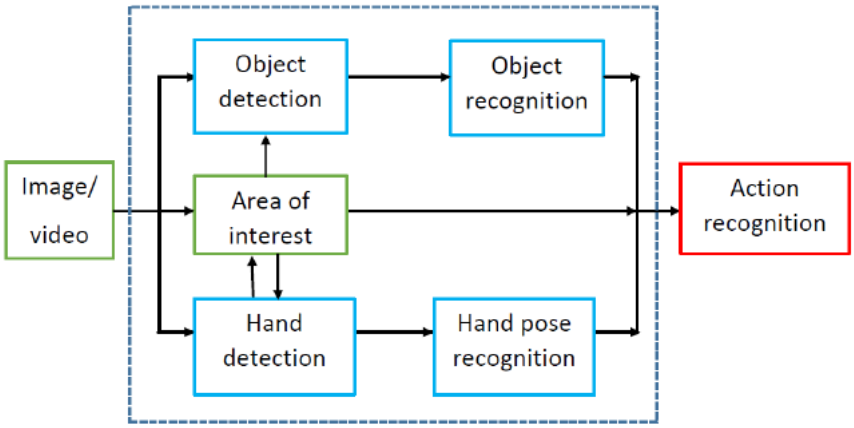


Figure 2.1: High-level illustration of a typical HBU pipeline. The system detects the objects involved, interprets actions, and provides a feedback, in this case a simple action recognition.

Alongside professional instructors hosting live or recorded sessions, many users now rely on AI-driven solutions to provide immediate guidance on exercise form, workout progress, and potential risk of injury. This trend gained further momentum with global shifts toward remote and home-based exercise programs [35]. As a result, the demand for robust, real-time analysis of human movement has never been higher, underscoring the critical role of HBU in modern fitness ecosystems.

Beyond convenience, these AI-assisted solutions can enable large-scale data collection on user performance, potentially giving rise to more personalized workout plans. For example, advanced platforms can factor in a user's historical motion data, injury history, and even daily readiness metrics from wearable sensors, thus delivering context-aware coaching suggestions. This level of personalization is poised to revolutionize the traditional "one-size-fits-all" approach to fitness.

2.1.2 Motion Tracking as a Core Component of HBU

A crucial aspect of HBU in fitness is the accurate reconstruction of human motion. While early methods often employed wearable sensors (e.g., accelerometers or inertial measurement units) to track body movements [40, 10],

computer-vision-based approaches have steadily gained popularity due to their non-intrusive nature and ability to capture entire body kinematics. The ability to analyze workouts from a simple webcam or smartphone camera opens the door to wide accessibility, especially in at-home settings, where users might not possess specialized wearable devices.

However, reliance on standard cameras also presents challenges. Inconsistencies in camera placement, lighting, and background clutter can degrade the quality of extracted motion signals, making robust algorithmic pipelines essential. In particular, 2D and 3D pose estimation techniques—discussed in detail in Sections 2.2 and beyond—form the computational backbone of these systems, converting raw video frames into interpretable joint positions and skeletal configurations.

2.1.3 Challenges in Human Behavior Understanding for Fitness

Despite its potential, HBU in fitness applications poses several unique challenges:

- **High Inter-Subject Variability:** Individuals differ significantly in body shape, joint flexibility, and movement patterns [30]. This diversity complicates the task of creating generalized models that can accurately recognize correct form or detect subtle errors for all users, from novices to advanced athletes.

- **Environmental Constraints:** Lighting variations, cluttered backgrounds, and non-ideal camera angles (common in home environments) can degrade pose estimation accuracy [34]. Moreover, partial occlusions—caused by furniture or even the user's own limbs—can introduce significant noise in detected keypoints.

- **Real-Time Feedback Requirements:** The utility of HBU in fitness

often hinges on delivering immediate feedback or interactive coaching [41, 19]. Achieving near real-time performance while maintaining high accuracy demands efficient algorithms that can handle complex full-body movements at sufficient frame rates.

- **Complex Motions and Multi-Action Workouts:** Exercises like burpees, snatches, or dance routines involve compound, fast, and often non-repetitive motions. Systems must be robust to dynamic posture changes, rapid transitions, and multiple exercise phases [37].

- **Data Privacy and Ethical Considerations:** Large-scale video recordings can pose risks related to user privacy, security, and data misuse. Strategies such as anonymization, secure storage, and user consent protocols are vital to maintaining trust in HBU-based fitness platforms.

Meeting these challenges requires a fusion of advanced computer vision, deep learning, and domain knowledge in biomechanics. Virtual coaching tools must effectively map visual data onto meaningful feedback—e.g., calling out a user's improper knee alignment during a squat or suggesting a change in hand placement during a push-up to avoid shoulder strain.

2.1.4 Advances Driven by 2D and 3D Pose Estimation

Recent progress in pose estimation—a sub-domain of HBU—has significantly bolstered fitness applications. High-performing 2D pose estimators like OpenPose, HRNet, and AlphaPose [6, 36, 39] form the foundation for systems that detect anatomical keypoints in real time. However, 2D estimations alone provide limited insight into depth or 3D skeletal orientation—elements crucial for advanced motion analysis. Consequently, many researchers have explored 3D pose estimation strategies, either through monocular camera setups refined by deep learning [22, 14] or multi-view configurations exploiting triangulation [28, 23].

In the fitness domain, 3D reconstructions of the human skeleton can yield actionable metrics such as joint angles, velocity profiles, or limb trajectories—facilitating more nuanced feedback (e.g., "Keep your back straight," or "Extend your hips further"). Studies have shown that detailed 3D tracking can reduce the incidence of exercise-related injuries by highlighting incorrect movement patterns [3, 17].

2.1.5 Personalization and SMPL-based Modeling

Moving beyond joint-based representations, parametric models like SMPL or SMPL-X add shape information to the pose data [20, 27]. These models capture body morphology, allowing systems to account for individual anthropometry—an especially relevant factor in fitness. For instance, two users performing the same squat may present drastically different angles or stances due to variance in limb proportions. By fitting SMPL-based models to user-specific shapes, HBU pipelines can deliver personalized guidance (e.g., adjusting squat depth recommendations based on torso-to-leg ratios).

Furthermore, SMPL-based representations lend themselves well to graphical feedback systems. Rendering a user's digital avatar from multiple perspectives or overlaying corrective markers can offer a more intuitive understanding of proper exercise form, further enhancing the coaching experience [29, 42]. This approach can also assist in advanced analyses of muscle load or joint stress, informing both athletic training and rehabilitation protocols.

2.1.6 Summary of HBU in Fitness Contexts

In summary, Human Behavior Understanding in fitness scenarios aims to transform raw sensor or video data into actionable feedback, assisting users in achieving better technique, safety, and results. This objective fuels the need for robust pose estimation—2D, 3D, and shape-aware parametric models—and real-time analytic capabilities that can thrive in less-controlled settings.

By reliably detecting deviations in form or identifying areas for improvement, HBU-based systems hold the potential to democratize high-quality coaching experiences, bridging the gap between professional trainers and at-home exercisers.

Building on the discussions in this section, the subsequent parts of this chapter delve deeper into the specific methodologies that make HBU possible, starting with 2D pose estimation as the foundational element and progressively moving towards advanced 3D reconstructions and SMPL-based modeling. Each stage addresses different facets of the challenges outlined above, collectively forging a comprehensive approach to human-centric analysis in fitness and beyond.

2.2 2D Pose Estimation

2D pose estimation is a critical component in our pipeline, serving as the foundation for subsequent 3D reconstruction and SMPL/SMPL-X fitting. Over the years, methods for 2D pose estimation have evolved from traditional, handcrafted feature-based approaches to sophisticated deep learning models. In the context of analyzing images of a person during training, these methods must accurately detect keypoints and robustly handle challenging conditions such as rapid motion, occlusions, and varying lighting.

2.2.1 Early Approaches and Their Limitations

Traditional methods relied on handcrafted feature extraction techniques such as Histograms of Oriented Gradients (HOG) [9] and Scale-Invariant Feature Transform (SIFT) [21], combined with graphical models like Pictorial Structures [11] to model the spatial relationships between body parts. Although pioneering, these methods struggled with occlusions, high intra-class pose variations, and were computationally expensive for dynamic environments like fitness sessions.

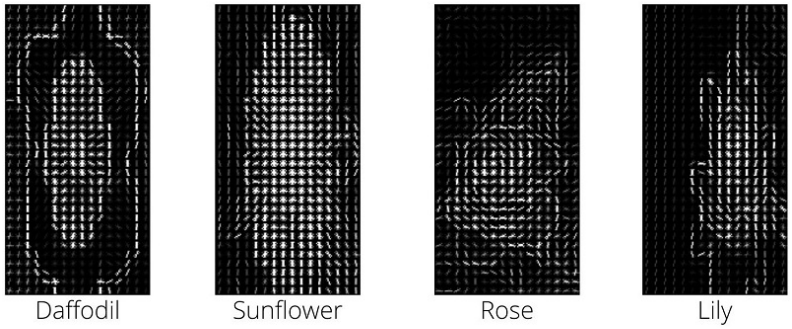


Figure 2.2: Example of feature detection using HOG (Adapted from Dalal and Triggs, 2005).

2.2.2 Deep Learning Revolution in 2D Pose Estimation

The advent of deep convolutional neural networks marked a significant improvement. Early deep-learning models, such as Convolutional Pose Machines and Hourglass Networks [38, 24], demonstrated that hierarchical feature extraction greatly improves keypoint detection, even in crowded or variable lighting conditions. OpenPose [6] further advanced the field by using part affinity fields to simultaneously detect body parts and assign them to specific individuals, while HRNet [36] maintained high-resolution feature representations for improved accuracy. However, these state-of-the-art methods can be computationally demanding.

2.2.3 The YOLO Family and the Choice of YOLOv8s

The YOLO (You Only Look Once) family of detectors has undergone significant evolution over the past several years, transforming from a groundbreaking real-time object detection framework into a versatile tool that now extends to human pose estimation. This real-time capability made YOLO highly attractive for applications requiring rapid decision-making and low latency.

In our application, where the goal is to support fitness coaching through real-time analysis of human motion, the trade-off between speed and accuracy

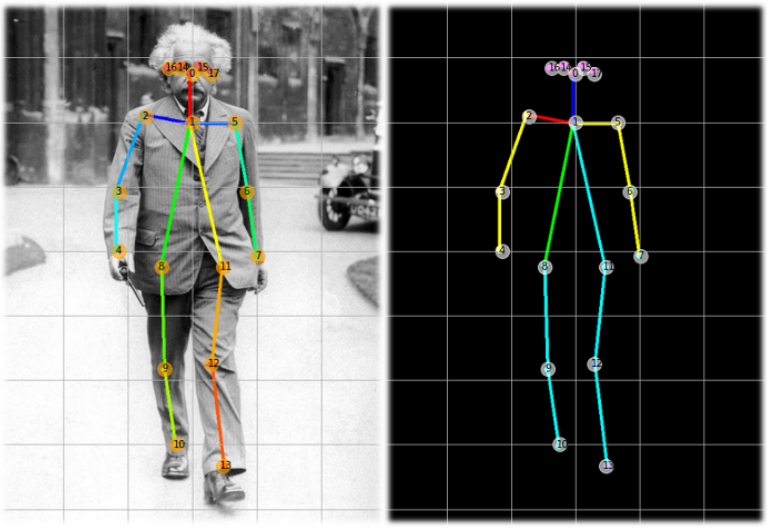


Figure 2.3: Example of OpenPose usage on a 2D image representing Albert Einstein. Source: Towards Data Science.

is critical. While there exist more advanced and computationally intensive models capable of achieving very high accuracy, they often require specialized hardware and longer inference times that are not suitable for live feedback. After evaluating several versions of YOLO and other deep learning-based approaches, we chose YOLOv8s—a lightweight and efficient variant that offers competitive accuracy with the advantage of fast inference speeds. This model is particularly adept at handling high-speed movements and partial occlusions, which are common in dynamic fitness environments.

YOLOv8s benefits from an optimized architecture that minimizes computational overhead while retaining a robust ability to detect keypoints reliably. Its design includes high performing feature extractors and multi-scale detection mechanisms, which ensure that even small or partially obscured joints are accurately localized. Furthermore, the availability of pre-trained weights allows for rapid deployment and fine-tuning on our dataset without the need for extensive retraining from scratch. This capability not only reduces the development time but also enhances the overall performance of our system.

In summary, YOLOv8s was selected for its ability to deliver real-time keypoint detection performance that meets the demanding requirements of fitness

applications. Its integration into our pipeline enables efficient extraction of 2D keypoints from multiview video feeds, which subsequently serve as the foundation for our 3D reconstruction and parametric modeling stages.

Figure 2.4: Illustration of YOLOv8s keypoint detection in an exercise scenario. Source: Adapted from [32].

2.2.4 Integration with Multi-View Imaging

While 2D pose estimation typically processes single images, incorporating multiple camera views greatly improves robustness. Each camera in a multiview setup offers a unique perspective, helping to alleviate occlusions or distortions in other views. In our pipeline, YOLOv8s is applied independently to each camera feed, and the resulting 2D keypoints are later fused into a coherent 3D reconstruction. This strategy is especially relevant in fitness scenarios, where dynamic movements often lead to self-occlusions.

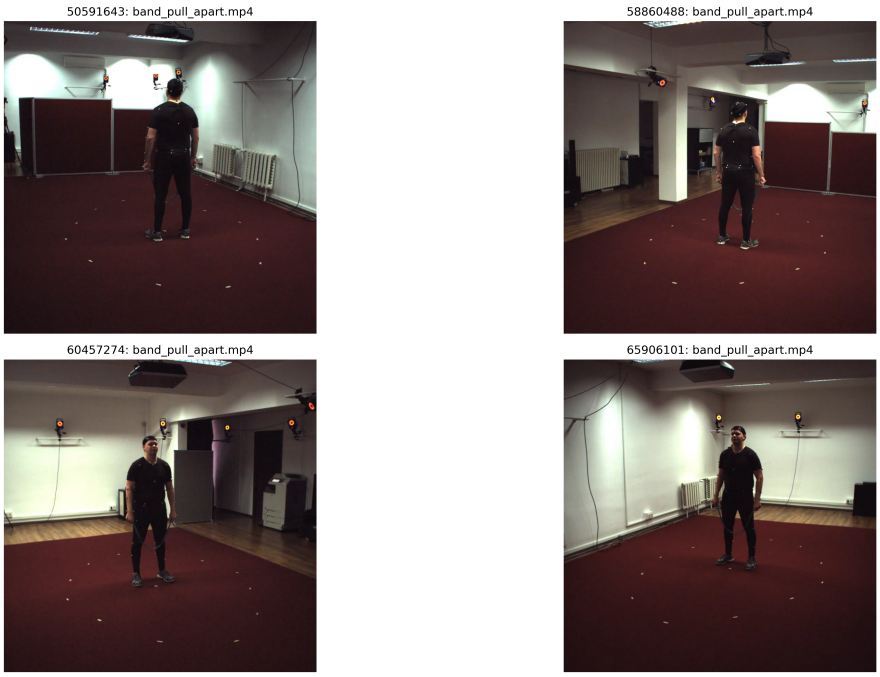


Figure 2.5: Schematic representation of our multi-view imaging setup.

2.3 3D Pose Estimation from Single and Multiple Images

Accurate 3D human pose estimation is essential for applications ranging from action recognition and biomechanics to virtual coaching. While 2D pose estimation provides valuable information on joint locations, it cannot capture the depth and complex spatial relationships necessary for detailed motion analysis. Therefore, methods have evolved to infer 3D poses from 2D data using both monocular and multiview strategies.

2.3.1 Monocular 3D Pose Estimation

Monocular approaches derive depth information from a single video feed or image sequence, making them attractive for low-cost, easy-to-deploy systems. Early methods relied on heuristic constraints, manual annotations, or

silhouette-based techniques to approximate 3D structure [31, 43]. More recent deep learning models directly regress 3D joint positions from 2D keypoints or raw RGB images [22, 44]. However, these methods often suffer from depth ambiguities and occlusions, particularly in dynamic environments such as fitness training.

2.3.2 Multiview 3D Pose Estimation

Multiview methods address the limitations of monocular setups by leveraging multiple synchronized cameras. By triangulating corresponding keypoints from different views using epipolar geometry [4, 23, 28], these approaches achieve more robust and accurate 3D reconstructions. Although multiview systems require additional hardware and synchronization, they provide the precision needed for high-accuracy applications.

2.4 Historical Evolution of 3D Body Modeling

The development of 3D body modeling has evolved from simple geometric representations to advanced data-driven models. In the 1970s, researchers approximated body parts using basic shapes (e.g., cylinders, cones) connected by simple rotational joints [2]. The 1980s and 1990s introduced kinematic chains, inverse kinematics, and early physics-based constraints, yet these methods were limited by manual intervention and rigid representations. A breakthrough occurred in the early 2000s with data-driven models like SCAPE [1], which learned a low-dimensional shape space from 3D scans, paving the way for modern parametric models such as SMPL [20] and SMPL-X [27].

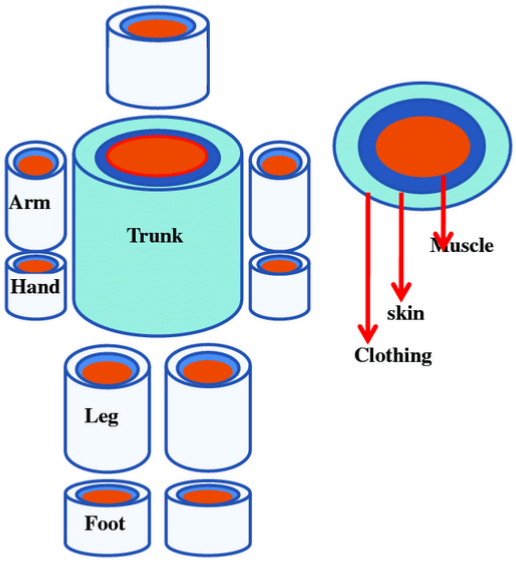


Figure 2.6: An early 3D human body model represented using primitive shapes (e.g., cylinders). Adapted from [18].

2.4.1 Deep Learning Approaches for 3D Pose Estimation

Modern 3D pose estimation increasingly relies on deep learning techniques that integrate geometric and learned priors. Approaches such as those proposed in [22, 44] directly regress 3D joint positions from 2D keypoints, while our pipeline employs a Transformer-based model to fuse multiview 2D detections (obtained via YOLOv8s) into a coherent 3D skeleton. This deep learning approach is critical for handling noise, occlusions, and the inherent uncertainties of single-view predictions.

2.4.2 Challenges and Future Research Directions

Despite significant progress, several challenges remain. Temporal synchronization is crucial, as precise time alignment between multiple cameras is necessary to avoid reconstruction errors. Occlusions and complex motions present another difficulty, as dynamic movements in fitness routines often lead to partial occlusions, requiring robust strategies to track disappearing and reappearing joints. Additionally, domain adaptation remains a challenge since many models are trained on controlled datasets and may not generalize well to

varied real-world conditions. Techniques such as synthetic data augmentation can help bridge this gap [29]. Finally, computational efficiency is a concern, as high-resolution multiview video processing and deep learning-based reconstruction demand substantial computational resources.

2.4.3 Insights and Relevance to Modern Applications

The evolution from primitive, heuristic-based models to advanced, data-driven methods highlights key lessons: high-quality 3D data, sophisticated deformation techniques, and efficient computational methods are vital for realistic human body modeling. Our pipeline—combining YOLOv8-based 2D detection, Transformer-based 3D reconstruction, and SMPLX parameter regression—demonstrates that integrating geometric cues with deep learning yields robust 3D representations. These advancements are particularly relevant for virtual coaching, where precise, real-time feedback on human motion can lead to improved performance and injury prevention.

In summary, while monocular methods offer simplicity, multiview approaches coupled with deep learning techniques provide the accuracy needed for complex, real-world applications. This section sets the stage for the subsequent detailed discussion on SMPL and related parametric models.

2.5 SMPL and Similar Models

Parametric body models have become increasingly popular as a method for representing the human body in a low-dimensional yet anatomically precise manner [20]. Unlike traditional skeleton-only representations, these models capture both pose (i.e., joint rotations) and shape (i.e., body proportions), allowing for a more realistic and personalized depiction of an individual. Such models are indispensable in applications ranging from computer graphics and virtual reality to advanced motion analysis in various domains, as they provide a unified framework for fitting a human mesh to observed data from images,

videos, or 3D scans.

2.5.1 SMPL Basics

SMPL (Skinned Multi-Person Linear) is one of the most widely adopted parametric models for human body representation. It represents the human body using two distinct sets of parameters:

- **Shape Parameters β:** This vector encodes variations in body morphology across individuals, such as differences in height, weight, and limb proportions. Typically defined in a low-dimensional space (often 10–20 dimensions) via Principal Component Analysis (PCA) on a large dataset of 3D body scans, β enables the generation of a diverse range of body shapes with only a few coefficients.

- **Pose Parameters θ:** This vector describes the rotation of the joints, commonly using an axis-angle representation for 24 body parts. The pose parameters specify how each joint deviates from a canonical or rest pose, thereby capturing the articulation necessary for realistic body motion.

The SMPL model is defined by a function $M(\beta, \theta)$ that outputs a high-resolution mesh of the human body. The process of generating this mesh involves several key steps, explained in the following sections.

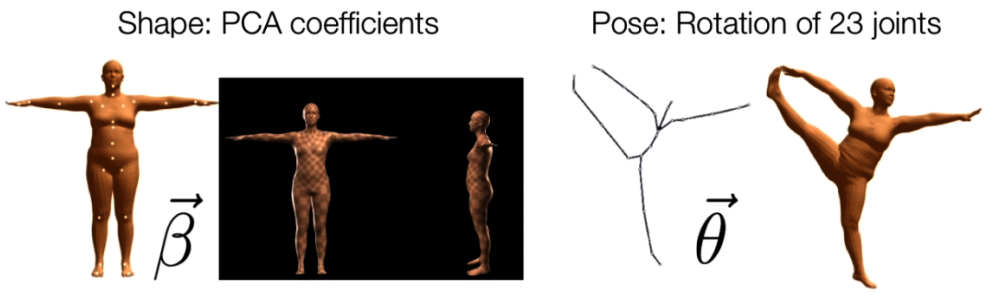


Figure 2.7: Contribution of β and θ inside the SMPL body model. Adapted from [13].

2.5.2 Template Mesh

SMPL begins with a learned template mesh that represents an average human body in a neutral pose. This mesh serves as the base structure, which is later deformed according to individual-specific shape and pose variations.

2.5.3 Shape Blend Shapes

The shape vector $\boldsymbol{\beta}$ modulates the template mesh via shape blend shapes—precomputed deformation vectors derived from PCA on 3D scans. By linearly combining these vectors with weights provided by $\boldsymbol{\beta}$, the model adapts the template to accurately reflect the target body shape.

2.5.4 Pose Blend Shapes

Rotating the template mesh using joint transformations does not fully capture the complex deformations that occur during articulation (such as muscle bulging or skin stretching). SMPL addresses this issue by incorporating pose blend shapes, which are learned corrective deformations that adjust the mesh after the initial joint rotations. These corrections help ensure that the final output exhibits natural, non-linear deformations corresponding to realistic human motion.

2.5.5 Linear Blend Skinning (LBS)

Once the blend shapes are applied, SMPL uses Linear Blend Skinning (LBS) to deform the mesh according to the pose parameters $\boldsymbol{\theta}$. In LBS, each vertex is influenced by nearby joints through precomputed skinning weights. Formally, the transformed position $\hat{v}_i$ of a vertex v_i is computed as:

$$\hat{v}_i = \sum_{j=1}^{J} w_{ij}\, T_j(\boldsymbol{\theta}) \Big(v_i + b_i^{\text{shape}}(\boldsymbol{\beta}) + b_i^{\text{pose}}(\boldsymbol{\theta}) \Big),$$

where w_{ij} are the skinning weights, $T_j(\boldsymbol{\theta})$ are the transformation matrices corresponding to each joint, and b_i^{shape} and b_i^{pose} denote the shape and pose blend shapes for vertex i.

2.5.6 Differentiability and Fitting

A key strength of SMPL is its full differentiability, which allows the model to be embedded into optimization frameworks or end-to-end deep learning pipelines. This differentiability enables the simultaneous optimization of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ by minimizing an error function (e.g., the difference between the projected mesh vertices and detected landmarks), thereby facilitating robust model fitting.

2.5.7 SMPL-X and Other Extensions

While SMPL provides a robust foundation for modeling human body shape and pose, it does not capture fine-grained details such as hand articulations and facial expressions. SMPL-X [27] extends SMPL to address these limitations, and its enhancements include:

- **Detailed Hand Modeling:** SMPL-X introduces separate hand models that capture the intricate degrees of freedom in finger and wrist motions. This extension is particularly useful for applications requiring fine motor analysis, such as sign language recognition or analyzing grip strength and technique.

- **Enhanced Facial and Head Features:** In addition to the body, SMPL-X integrates a more expressive facial model that captures subtle expressions and head movements. This allows for a more holistic understanding of human behavior and can be beneficial in applications such as emotion recognition or detailed motion capture in virtual environments.

- **Unified Parameter Space:** By combining body, hand, and face modeling into a single, coherent framework, SMPL-X offers a unified representation that can be used in a variety of applications without requiring separate models for each body part.

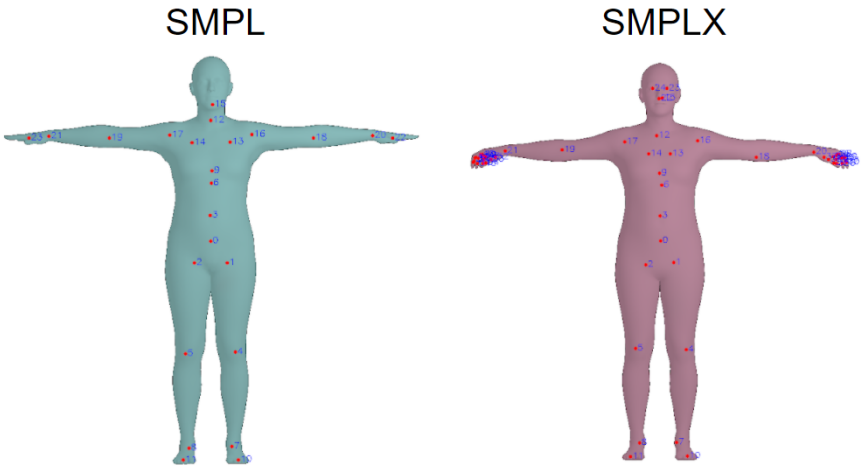


Figure 2.8: Comparison of SMPL and SMPL-X: SMPL-X extends the basic SMPL model by incorporating detailed hand and facial features.

2.5.8 Challenges in Fitting Parametric Models

Fitting parametric models like SMPL or SMPL-X to real-world data presents several challenges. Initialization sensitivity is a key issue, as optimization-based methods such as iterative closest point or gradient descent require an accurate initial estimate. On the other hand, a poor initialization can lead to convergence at local minima, resulting in anatomically implausible meshes. Another challenge is data sparsity and noise, especially in uncontrolled environments like home gyms or outdoor settings, where 3D keypoints from pose estimation are often sparse and noisy, degrading the quality of the fitted model. The scarcity of high-quality datasets with comprehensive ground-truth annotations for both body shape and pose further limits the training of deep learning models for direct SMPL parameter regression. Additionally, variability in apparel and occlusions complicates the process, as clothing can obscure body

contours and self-occlusions, such as crossed arms, make precise alignment more difficult.

While deep learning approaches have been developed to regress SMPL parameters directly from visual inputs, they require large amounts of training data and careful regularization to ensure anatomically consistent reconstructions.

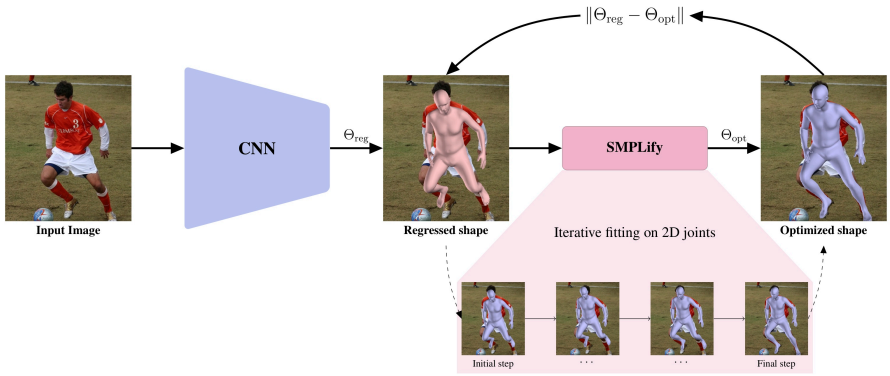


Figure 2.9: Optimization Flowchart for Fitting Parametric Models: from a single RGB image, a CNN regresses an initial 3D shape, and an iterative fitting process (e.g., SMPLify) refines that shape using 2D joint constraints.

2.5.9 Relevance to Various Applications

While our primary focus is on fitness and virtual coaching, parametric models like SMPL and SMPL-X offer benefits across multiple domains. Their ability to capture both shape and pose enables the creation of personalized digital avatars, essential for tailored fitness feedback, virtual try-on, gaming, and medical diagnostics. These models also enhance visualization by generating detailed 3D meshes from multiple viewpoints, allowing the overlay of additional data such as muscle activation maps or joint stress indicators, improving interpretability and user engagement. Moreover, their anatomically consistent structure supports biomechanical analysis, enabling the estimation of forces, torques, and muscle loads, which is crucial for injury prevention and performance optimization in sports and rehabilitation. Additionally, the differentiability of these models facilitates seamless integration into AI pipelines,

allowing end-to-end optimization from image capture to detailed mesh reconstruction, thereby improving accuracy and robustness in various applications.

2.6 Alternative Approaches to Human Body Modeling

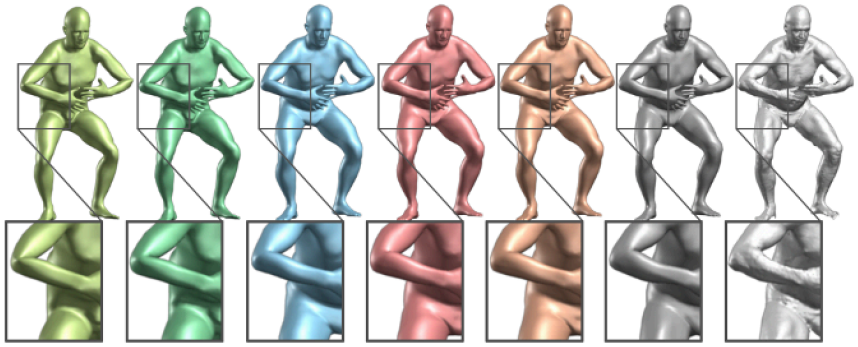


Figure 2.10: Overview of Alternative Approaches to Human Body Modeling: A comparative chart of LBS, Dual-quaternion blend skinning (DBQS), BlendSCAPE, SMPL-LBS, SMPL-DQBS, SMPL-X and the 3D scan, highlighting their key features and differences.

In addition to SMPL and its extensions, several alternative models and fitting approaches have been proposed over the years to address the challenges of capturing human shape and pose from visual data. These methods differ in their underlying assumptions, model complexity, and optimization strategies. In this section, we provide an overview of some of the prominent approaches, including SCAPE, SMPLify, SMPLify-X, STAR, and SPIN.

2.6.1 SCAPE

SCAPE (Shape Completion and Animation for People) was one of the first models to address the complex deformations of the human body in a data-driven manner [1]. Unlike SMPL, which uses a linear formulation, SCAPE models non-linear deformations by combining per-triangle deformation matrices with a global shape space. This approach allows SCAPE to capture more

detailed variations in body shape and motion. However, its high computational cost and the difficulty of fitting the model to noisy data have limited its widespread adoption in real-time applications.

2.6.2 SMPLify

SMPLify [5] represents a significant step forward by proposing an optimization-based method to fit the SMPL model directly to 2D keypoints extracted from images. By minimizing the re-projection error between the SMPL model and detected 2D landmarks, SMPLify bridges the gap between 2D pose estimation and 3D model fitting. This approach demonstrated that even a simple optimization framework could produce accurate 3D reconstructions from single images, although it often requires a good initialization and strong pose priors to avoid local minima.

2.6.3 SMPLify-X

Building on the ideas of SMPLify, SMPLify-X [16] extends the fitting framework to the SMPL-X model, which incorporates additional detail in the hands and face. SMPLify-X adapts the optimization procedure to account for these extra parameters, enabling more detailed reconstructions that capture fine-grained upper-body movements. This enhanced fitting process is particularly useful for applications that require detailed analysis of hand gestures or facial expressions, though it also introduces additional complexity and computational load.

2.6.4 STAR

STAR (Sparse Trained Articulated Human Body Regressor) [26] is another notable approach that seeks to reduce the dimensionality of the human body representation. By employing a sparse regression framework, STAR achieves competitive performance with fewer parameters than traditional models. This

Figure 2.11: Example of SMPLify-X body model with a successful application.

reduction in complexity can be particularly advantageous when computational resources are limited, or when the application demands faster inference times. STAR's specialized shape space also allows it to generalize well across diverse populations, addressing one of the key challenges in body modeling.

2.6.5 SPIN

SPIN (Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop) [15] takes a hybrid approach by integrating deep learning with traditional model-based optimization. In SPIN, a neural network regressor produces an initial estimate of the SMPL parameters, and this estimate is then refined through an iterative optimization process that incorporates the SMPL model in the loop. This end-to-end approach helps improve the accuracy and robustness of the reconstruction, especially in the presence of noisy or incomplete input data. SPIN's ability to iteratively refine the model parameters makes it particularly attractive for applications that require precise 3D reconstructions.

2.6.6 ExPose

ExPose [12] integrates deep learning with model-based optimization to estimate 3D human pose and shape in real time. By leveraging both 2D keypoints and parametric models, ExPose achieves robust performance even under challenging conditions such as occlusions and diverse poses. Its innovative approach of incorporating model-fitting within a deep learning framework offers improved accuracy and generalization compared to purely optimization-based methods.

Chapter 3

Methodology

This chapter details the end-to-end methodology developed in this thesis to extract detailed SMPL-X parameters from multi-view video sequences of fitness exercises. Our approach transforms raw video data into a robust 3D representation of human motion and subsequently generates a detailed parametric model of the human body. The proposed pipeline is made up of several interlinked stages, which are described below.

First, we get access to four synchronized video recordings capturing the same fitness scene from different angles. These multi-view videos provide diverse perspectives essential for accurate 3D reconstruction.

Next, the YOLOv8s-pre-trained model is applied to each camera feed to extract 2D keypoints corresponding to critical body joints. These keypoints serve as the foundational data for subsequent processing.

After the keypoints extraction, a preprocessing step filters the data by selecting only the frames that exhibit significant movement changes. This reduces redundancy and enhances data quality. The filtered 2D keypoints, together with their corresponding ground-truth 3D joint positions, are aggregated into a comprehensive JSON structure (named `mega_dict_2d3d`). An analogous process constructs `mega_dict3dSMPLX`, which links the 3D joint data with the corresponding SMPL-X parameters.

Using the information stored in these dictionaries, our pipeline then employs a first neural network (experimenting with various architectures) to fuse the multi-view 2D keypoints into a coherent 3D skeleton, effectively mitigating occlusions and depth ambiguities. Building upon the reconstructed 3D skeleton, a second model regresses high-dimensional SMPL-X parameters, converting the sparse 3D joint data into a detailed and anatomically plausible parametric representation of the human body.

Finally, the predicted SMPL-X parameters are used to render detailed 3D meshes on specific graphics software, enabling both quantitative evaluation and qualitative visual inspection of the reconstructed models.

In the following sections, we provide a detailed account of each stage of the pipeline, including the architectures of our deep learning models, the preprocessing techniques used to build the dictionaries, and the training protocols implemented.

3.1 Model Architectures and Implementation

This section provides a detailed description of the deep learning architectures and implementation strategies employed in our pipeline. Our objective is to convert raw, multi-view 2D keypoints data into accurate 3D representations and ultimately into a detailed SMPL-X parametrization of the human body. We explored various architectures during our research; however, our experiments revealed that Transformer-based models consistently outperformed other alternatives for both the 2D-to-3D and the 3D-to-SMPLX tasks.

3.1.1 2D-to-3D Pose Reconstruction Module

The first stage of our pipeline reconstructs a coherent 3D skeleton from multi-view 2D keypoints. This process begins by extracting 2D keypoints from four synchronized video streams, which are captured from different viewpoints of the subject. We employ the YOLOv8s detector, a lightweight and efficient

version of YOLO, specifically fine-tuned for keypoint detection in human poses. The YOLOv8s model operates on each individual frame of the video, detecting the locations of the keypoints corresponding to the human body. In our case, each frame contains 17 keypoints, which represent key body joints such as the wrists, elbows, shoulders, hips, knees, and ankles, as well as the neck and head.

For every detected keypoint, the model outputs two parameters: the x and y coordinates of the keypoint in the 2D image plane. These coordinates are recorded for each frame of the video, for every subject in the dataset. Since we are working with four synchronized cameras, this results in four sets of 17 keypoints per frame for each subject.

To ensure consistent quality and mitigate the effects of varying lighting conditions, camera angles, and other environmental factors, the extracted keypoints are normalized. This normalization process scales the coordinates to a common reference system, reducing any distortions that might arise due to differences in camera settings or perspectives. Despite this normalization, some level of uncertainty persists, as the predicted keypoints are not always perfectly accurate due to limitations in the detector, occlusions, or ambiguities in pose estimation. These uncertainties can introduce small errors in the 2D keypoints, which will be addressed in later stages of the pipeline.

Once the keypoints are extracted and normalized, they are structured into a unified representation that provides a comprehensive view of the human pose across all video streams, enabling the subsequent stages of the pipeline to perform 3D reconstruction. Each frame's set of keypoints forms the basis for the 3D model reconstruction, with each set being processed sequentially to construct the 3D skeleton of the subject for that frame.

Transformer-Based Approach In our best-performing approach, a linear projection first maps the low-dimensional 2D keypoints into a higher-dimensional space, enriching the representation to capture complex spatial

relationships. The Transformer encoder then applies self-attention to fuse multi-view information, mitigating occlusions and misalignment. The encoder's output is reshaped into a 25×3 matrix, where each row represents a joint's (x, y, z) coordinates, forming the robust 3D skeleton for subsequent SMPL-X regression.

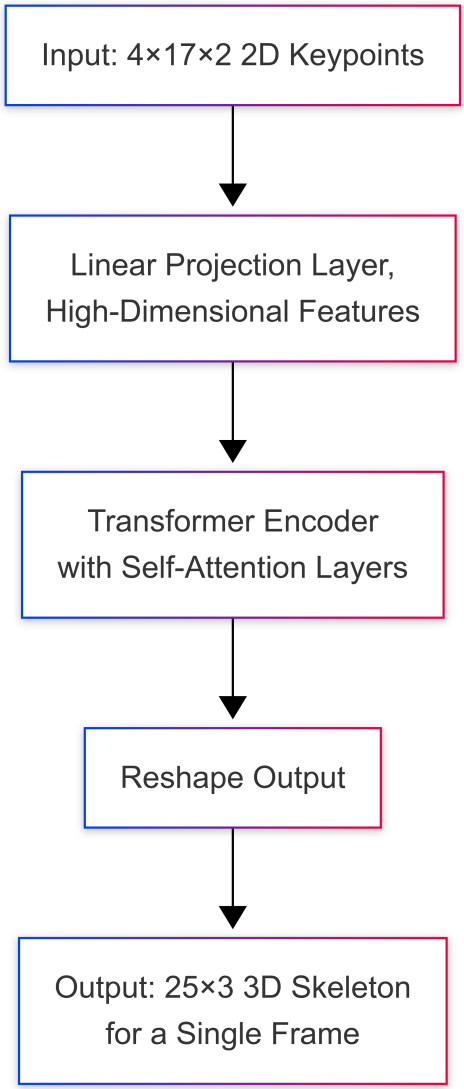


Figure 3.1: Schematic diagram of the 2D-to-3D Pose Reconstruction Module for the Transformer model. The diagram illustrates the flow from multi-view 2D keypoints (17×2 per camera) through a linear projection and a Transformer encoder to produce a 3D skeleton (25×3).

Alternative Architectures In addition to the Transformer model, we also evaluated three alternative architectures: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Fully Connected Neural Networks (FCNNs). Each of these models employs a distinct strategy for processing multi-view 2D keypoints into a 3D skeleton:

- **CNN-Based Approach:** This model treats the 2D keypoints as an image-like structure, applying convolutional layers to extract spatial features before regressing the 3D coordinates. While CNNs excel at capturing local spatial patterns, their fixed receptive fields limit their ability to model long-range dependencies between keypoints.
- **RNN-Based Approach:** The RNN model sequentially processes keypoints over time, leveraging recurrent layers to encode temporal dependencies. This architecture is particularly effective for sequences but may be less optimal when reconstructing single-frame poses due to its reliance on past observations.
- **FCNN-Based Approach:** A fully connected network directly maps the concatenated 2D keypoints from all views to their corresponding 3D coordinates. While computationally efficient, this model struggles with occlusions, as it lacks an explicit mechanism for reasoning about missing keypoints.

Overall, each model demonstrates strengths and limitations depending on the input characteristics and the nature of occlusions present in the scene. The Transformer-based approach, however, consistently achieves the best performance due to its ability to capture long-range dependencies and dynamically integrate multi-view information, making it the most effective choice for robust 3D pose reconstruction.

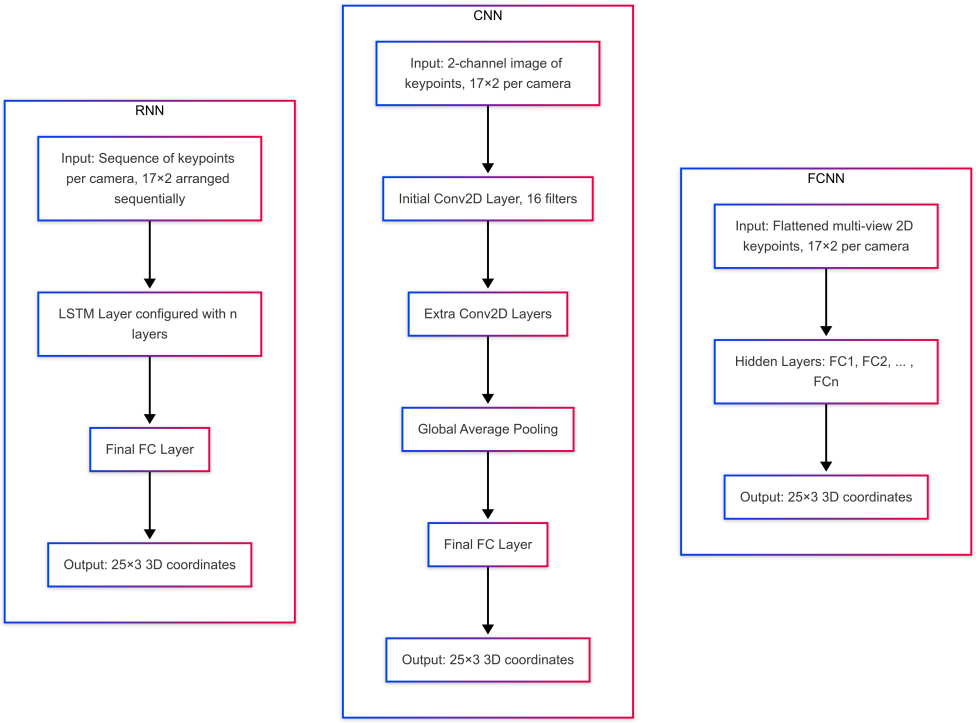


Figure 3.2: Diagram of the different 2D-to-3D pose reconstruction architectures.

3.1.2 3D-to-SMPLX Parameter Regression Module

In this stage, the aim is to convert the 3D skeleton obtained in the previous step into a detailed SMPL-X parameter vector. The input consists of a 3D skeleton with dimensions 25×3, where each of the 25 joints is defined by three coordinates (x, y, z) representing its position in 3D space. These joint positions are derived from the previous stage, where 2D keypoints were first extracted and then triangulated to obtain the 3D positions.

Once the 3D skeleton is available, the next step involves mapping these joint positions to the SMPL-X model. The SMPL-X model is a parametric model designed to represent the human body in a highly detailed manner, capturing both the global pose and body shape, as well as finer details such as facial expressions, hand movements, and eye orientation.

The final result of this stage is a 188-dimensional vector. This vector encapsulates various aspects of the body, such as body pose, overall body shape,

and detailed expressions like hand gestures and facial motions. These parameters are encoded in the vector, allowing the SMPL-X model to accurately represent the person in the 3D space.

We adopted a Transformer-based architecture for this regression task because of its ability to capture long-range dependencies and global context. The key components are as follows:

1. **Input Preprocessing:** The 3D skeleton is reshaped into a sequential format that preserves spatial relationships among joints.

2. **Linear Projection:** A linear layer projects the input into a higher-dimensional feature space (of size *d_model*), enriching the representation to capture subtle joint interactions.

3. **Transformer Encoder Layers:** A stack of encoder layers employs self-attention to dynamically weigh each joint's contribution, integrating global contextual information to address occlusions and misalignment. The number of layers is set by *transformer_num_layers*.

4. **Flattening and Regression:** The encoder output is flattened and passed through a fully connected layer that regresses the 188-dimensional SMPL-X parameter vector.

This module effectively captures the complex mapping between the 3D skeleton and SMPL-X parameters, yielding anatomically plausible and detailed reconstructions.

3.1.3 Implementation Details

Our implementation is built on PyTorch, with NumPy and SciPy handling numerical and rotation computations, while experiment tracking is managed using wandb. The system follows a modular pipeline with two main stages: 2D-to-3D pose reconstruction and 3D-to-SMPLX parameter regression, ensuring a unified data structure for aligning multi-view inputs with 3D outputs.

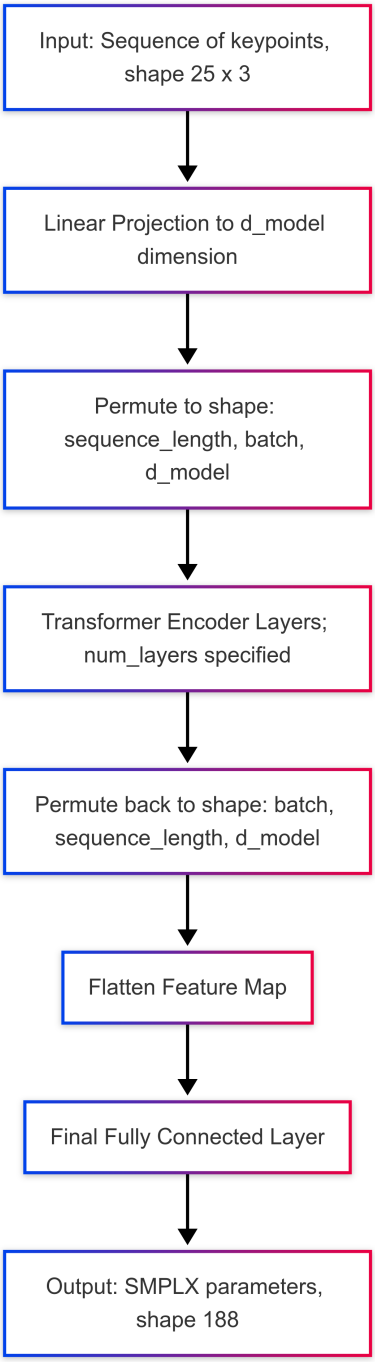


Figure 3.3: Schematic diagram of the 3D-to-SMPLX Regression Module using the Transformer architecture.

Training configurations, including batch size, learning rate, and dropout, are defined in configuration files, with the Adam optimizer and MSE loss guiding model updates. The framework is designed for scalability, leveraging GPU clusters and allowing seamless integration of additional data modalities or model improvements.

3.2 Data Processing and Preprocessing

Our preprocessing pipeline converts raw multiview videos into structured, high-quality data for training. This stage minimizes noise, ensures temporal consistency, and selects the most informative frames for 3D reconstruction and SMPLX regression.

3.2.1 2D Keypoint Detection and Normalization

Frames from each of the four camera views are processed with a YOLOv8s-based detector to extract 17 keypoints per frame. The raw coordinates are normalized using a dataset-wide mean and standard deviation, ensuring consistency despite varying lighting and camera conditions. Frames with missing or inconsistent keypoints are trimmed or zero-padded so that only frames with valid detections across all views are retained. Figure 3.4 outlines the process from frame extraction to the construction of a unified keypoint representation.

3.2.2 Custom Dataset Construction

To streamline downstream processing, we aggregate the data into two efficient and structured dictionaries:

1. `mega_dict_2d3d`: Aligns normalized 2D keypoints from all views with corresponding 3D joint positions, filtering out frames with negligible motion.

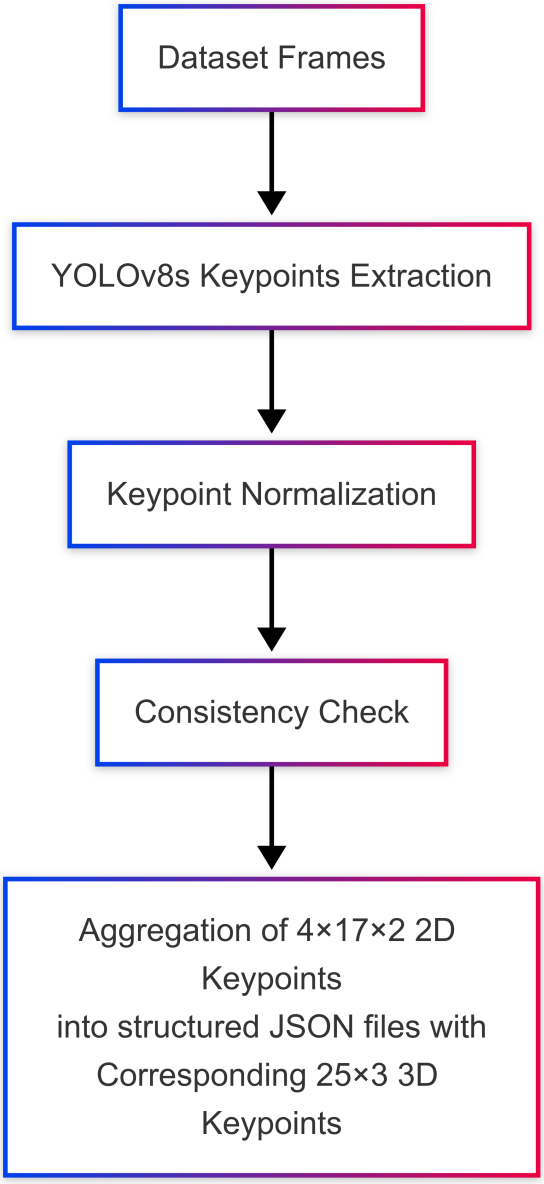


Figure 3.4: Flowchart of 2D keypoint extraction, normalization, and aggregation.

2. `mega_dict3dSMPLX:` Associates filtered 3D joint data with corresponding SMPLX parameters, always filtering out redundant frames.

These dictionaries are built by loading JSON data, filtering frames based on movement thresholds, and ensuring consistency across camera views.

3.2.3 Error Handling and Quality Assurance

Robust checks ensure consistency in frame counts across cameras, verify that keypoint data have the correct dimensions and are non-empty, and confirm the presence of all necessary annotations, including 3D joints and SMPL-X parameters. Frames that fail these checks are either discarded or corrected, maintaining high-quality training data that accurately represents real-world human motion.

3.3 Training Protocols

This section outlines our training strategy, covering hyper-parameter selection, loss functions, optimizers, and validation procedures for both the 2D-to-3D reconstruction and SMPL-X regression modules. Our goal is to achieve robust convergence and generalization.

3.3.1 Overview

We developed two distinct models—one for 2D-to-3D reconstruction and one for SMPL-X parameter regression—each trained separately. Detailed quantitative results and sample outputs for each module are presented in the following chapter. An end-to-end qualitative demo, which combines the outputs of both modules, is also provided to visually validate the complete pipeline.

3.3.2 Model Choice and Hyper-parameters

Training is configured via a JSON file that specifies key hyper-parameters, including:

- **Model Type:** Specifies the architecture used (e.g., Transformer, CNN, RNN, FCNN) for the 2D-to-3D reconstruction module; for the 3D-to-SMPLX part, as already mentioned, we only use Transformer-based architectures.
- **Model Details:** Architecture-specific parameters such as number of layers and hidden sizes.
- **Batch Size, Learning Rate, Epochs, and Dropout:** Default values are tuned based on model requirements.
- **Optimizer Settings:** We use the Adam optimizer, leveraging its adaptive learning rate.

A global random seed is set to ensure reproducibility.

3.3.3 Loss Functions

For our regression tasks, we use Mean Squared Error (MSE) as the primary loss function. MSE penalizes larger errors more heavily and provides smooth gradients for optimization. Additional metrics are monitored for comprehensive performance evaluation.

3.3.4 Optimizer and Learning Rate Scheduling

The Adam optimizer is initialized with a fixed learning rate (as specified in the configuration). Although we do not use advanced scheduling in our baseline experiments, variations such as decay or cyclic learning rates are potential avenues for future improvements.

3.3.5 Experiments' Selection

We systematically explored a wide range of hyper-parameter configurations for each model to identify the optimal settings in terms of performance. In the following chapter, we present the quantitative results along with detailed information on the hyper-parameters used for each model.

Chapter 4

Experimental Results

4.1 Introduction

In this chapter, we present a comprehensive evaluation of our proposed pipeline for 3D human pose estimation and SMPL/SMPL-X parameter regression. Our experiments demonstrate that the integration of multi-view 2D keypoints detection, robust 3D reconstruction, and advanced parametric modeling yields results that are both quantitatively accurate and visually convincing for fitness applications.

We begin by describing the datasets and evaluation metrics used to assess the performance of each module. Next, we detail the quantitative results obtained from our experiments for both models. We then present qualitative visualizations of the reconstructed 3D poses and generated SMPL-X meshes, highlighting their anatomical plausibility and visual coherence.

The performance of each model is evaluated independently, both quantitatively and qualitatively, before we conclude with an end-to-end qualitative demonstration of the complete pipeline.

4.2 Dataset Overview

Our experiments leverage the FIT3D dataset [8], a comprehensive collection of multi-view video recordings specifically designed to capture human motion in fitness environments. This dataset has been curated to record subjects performing a range of gym exercises (e.g., squats, lunges, bench presses, deadlifts) from multiple perspectives. The rich and diverse data provided by FIT3D makes it ideally suited for evaluating advanced computer vision methods for 3D pose estimation and parametric body modeling.

4.2.1 Dataset Organization and Frame Extraction

The FIT3D dataset is organized in a clear hierarchical structure that facilitates both data management and efficient processing. At the highest level, the dataset is stored in a root directory that contains a separate folder for each subject. Within each subject folder, the data are further subdivided into several key components:

- **Camera Parameters:** A folder containing calibration files for each of the four cameras.
- **2D Keypoints:** A folder with normalized 2D keypoints files. These files are organized in sub-folders, one per camera. They are extracted using YOLOv8s, as already discussed in the previous sections.
- **Videos:** Raw video recordings from all four synchronized cameras are stored in a dedicated folder, with sub-folders corresponding to each camera.
- **3D Joints:** A single JSON file per exercise, containing the ground truth 3D joint positions. Ground truth 3D coordinates are provided on a frame-by-frame basis and serve as the benchmark for evaluating our 3D reconstruction methods.

- **SMPL-X Parameters:** A single JSON file per exercise, providing the detailed SMPL-X parametrization. They are detailed parametric representations capturing both body shape and pose—including finer details such as hand gestures and facial expressions— provided for each exercise.

Raw videos are decomposed into individual frames, with strict temporal synchronization maintained across all camera views. This alignment is critical because it ensures that the same moment in time is captured from multiple perspectives. The extracted frames are then processed to extract keypoints, and consistency is verified across the four cameras. Finally, the processed data are aggregated into the pre-mentioned unified JSON structures which consolidate the multi-view 2D keypoints and the corresponding 3D annotations for each exercise, as better explained in the following sections.
Figure 4.1 shows a simplified diagram of the dataset hierarchy.

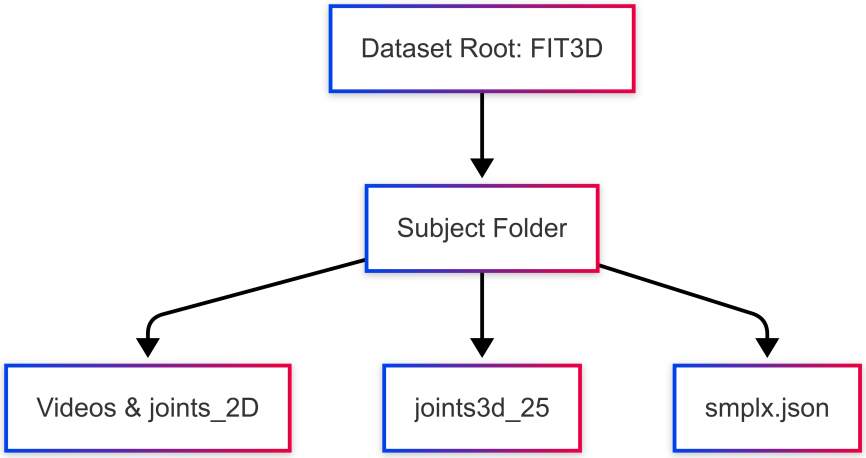


Figure 4.1: Simplified hierarchical organization of the FIT3D dataset. Videos & joints_2D contain extra folders for each camera while joints_3D and smplx contain unique json files for each exercise.

4.2.2 Dictionary Construction

To streamline training, the raw annotations are processed into the unified JSON structures, that are essentially custom dictionaries with the corresponding data. During the creation of these dictionaries, a filtering procedure is applied to retain only those frames that exhibit significant motion—determined by a predefined threshold based on the difference in 3D joint positions. This filtering process reduces noise and focuses the training data on dynamic, informative frames.

These thresholds are experimented qualitatively in order to get a good amount of data that could be representative enough for the entire dataset, excluding redundant information; we tried with different values until getting the desired output.

4.2.3 Summary of Key Attributes

Table 4.1 summarizes the primary attributes of the FIT3D dataset, highlighting its multiview configuration, high resolution, frame rate, and multiple annotation layers.

Attribute	Description
Number of Subjects	8
Number of Cameras	4 (multiview setup)
Resolution	High resolution (e.g., 1920×1080)
Frame Rate	30 fps
2D Keypoints	Extracted using YOLOv8, normalized with global statistics
3D Joint Positions	Ground truth for each exercise on each subject (on a frame basis)
SMPL/SMPL-X Parameters	Detailed parametric representations of body shape and pose for each exercise on each subject (on a frame basis)
Mega Dictionaries	Consolidated JSON structures (`mega_dict_2d3d` and `mega_dict3dSMPLX`)

Table 4.1: Summary of key attributes of the FIT3D dataset.

4.2.4 Training, Validation and Test sets

Since our dataset contained only eight subjects with all the necessary data, we had to carefully decide on a reasonable division into training, validation, and test sets. These splits are fixed to ensure comparable and accurate results across experiments. Specifically, we allocated 70% of the data to the training set, 20% to the validation set, and 10% to the test set for both our architectures.

Regarding sampling, the split is not stratified by subject. Instead, all frames from all subjects are pooled and then partitioned according to the specified percentages. This approach allows each subject's data to appear in all splits, although it may introduce partial overlap in subject identity across training, validation, and test sets.

4.3 Evaluation Metrics

To quantitatively assess the performance of our 3D pose estimation and SMPL-X parameter regression pipelines, we employ a suite of evaluation metrics. These metrics enable a comprehensive analysis of the errors in predicted joint positions and model parameters. In the following subsections, we describe each metric in detail, including its mathematical formulation, its interpretation, and its relevance to our task.

4.3.1 Mean Per Joint Position Error (MPJPE)

MPJPE is one of the most commonly used metrics in 3D human pose estimation. It quantifies the average Euclidean distance between the predicted 3D joint positions and the corresponding ground truth positions. Formally, if $\mathbf{p}_i^{\text{pred}}$ and $\mathbf{p}_i^{\text{gt}}$ denote the predicted and ground truth coordinates for the i-th joint, respectively, and N is the total number of joints, then MPJPE is defined as:

$$\text{MPJPE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i^{\text{pred}} - \mathbf{p}_i^{\text{gt}}\|_2$$

This metric is particularly intuitive as it directly measures the spatial error in units. A lower MPJPE indicates that the predicted 3D pose is closer to the ground truth, which is critical in applications such as virtual coaching where even small deviations can impact performance feedback.

4.3.2 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is another widely used metric that measures the average absolute difference between the predicted and ground truth values. For our regression tasks (both for 3D joint positions and SMPL/SMPL-X parameters), MAE is defined as:

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^{M} |y_i^{\text{pred}} - y_i^{\text{gt}}|$$

where M is the total number of scalar predictions. MAE is particularly useful because it provides an interpretable error magnitude without overly penalizing larger errors, unlike squared errors. This is important when the error distribution contains outliers that we do not want to dominate the overall performance measure.

4.3.3 Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

The Mean Squared Error (MSE) measures the average of the squares of the differences between the predicted and actual values:

$$\text{MSE} = \frac{1}{M} \sum_{i=1}^{M} (y_i^{\text{pred}} - y_i^{\text{gt}})^2$$

Squaring the differences ensures that larger errors are penalized more significantly. However, because the error is squared, the result is not in the same units as the original measurements.

To address this, the Root Mean Squared Error (RMSE) is computed by taking the square root of the MSE:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

RMSE is particularly valuable because it expresses error in the same units as the original data, providing a clear and intuitive sense of the average error magnitude. In our context, MSE serves as the models' loss function, offering a baseline measure of variance around the ground truth. RMSE further enhances interpretability by revealing how far, on average, our predictions deviate from the ground truth in a directly comparable scale.

4.3.4 Groupwise Mean Absolute Error

For a more detailed analysis, particularly in the context of SMPL/SMPL-X parameter regression, we compute the Groupwise MAE. This metric evaluates the MAE for distinct groups of parameters (e.g., translations, global orientation, body pose, hand poses, facial expressions) separately. This approach allows us to identify specific areas where the model may be underperforming.

For instance, let $\mathbf{y}_g^{\text{pred}}$ and $\mathbf{y}_g^{\text{gt}}$ denote the predicted and ground truth values for group g, respectively. Then, the Groupwise MAE for group g is defined as:

$$\text{Groupwise MAE}_g = \frac{1}{M_g} \sum_{i=1}^{M_g} \left| y_{g,i}^{\text{pred}} - y_{g,i}^{\text{gt}} \right|$$

where M_g is the number of parameters in group g. This granular analysis is particularly useful in applications like virtual coaching, where understanding errors in specific joints or regions (e.g., the arms or the torso) can provide actionable insights for improving performance.

4.3.5 Note about the Units

Accurately assessing the performance of our model requires well-defined metrics, each associated with a specific unit of measurement, such as centimeters. However, in our experiments, FIT3D does not provide explicit unit references, meaning we lack a definitive scale for absolute measurements. To address this and ensure consistency across the dataset, we applied normalization to all the involved data (e.g. 2D keypoints, 3D keypoints and SMPL-X parameters). This preprocessing step allows us to analyze relative performance trends effectively, without the need for explicit unit conversions.

Normalization is performed by computing the mean and standard deviation over the entire dataset (using the training set data) for each coordinate. Each value is then standardized by subtracting this mean and dividing by the standard deviation. This approach removes the influence of the original scale and ensures that our evaluation metrics reflect relative differences across samples.

4.4 Quantitative Results

This section presents a comprehensive quantitative evaluation of our proposed pipeline. Using the already discussed metrics, we assess two main modules: the 2D-to-3D reconstruction module and the SMPL-X parameter regression module.

4.4.1 2D-to-3D Reconstruction Performance

In the first stage of our pipeline, multi-view 2D keypoints are fused into coherent 3D joint positions. Our evaluation is detailed in Tables 4.2, 4.3, 4.4, and 4.5, which report, on a separate table for each architecture experimented, the configuration number (ID), model specifics, and test results (loss and MPJPE); training and validation metrics are also reported in Appendix A.1.

Notably, Transformer-based setups consistently yield the lowest MPJPE, highlighting their superior capability in modeling spatial-temporal relationships. RNN models perform competitively, while FCNN and CNN approaches exhibit higher errors. Several experiments were run with few epochs to assess early convergence and control training time.

Table 4.2: Evaluation of 2D-to-3D Reconstruction Experiments (Transformer). Columns: **ID** is the experiment identifier; **BS** is the batch size; **d** is the embedding dimension; **n** is the number of attention heads; **L** is the number of layers; **dp** is the dropout rate; **Ep** is the number of training epochs; **LR** is the learning rate; **Test Loss** is the loss on the test set (MSE); **Test MPJPE** is the test mean per joint position error.

ID	BS	d	n	L	dp	Ep	LR	Test Loss	Test MPJPE
1	64	128	2	2	0.2	25	5×10^{-5}	4.29×10^{-6}	1.93×10^{-4}
2	64	128	4	10	0.5	100	0.001	3.83×10^{-4}	0.00170
3	64	256	8	12	0.2	200	0.0005	3.85×10^{-4}	0.00170
4	64	128	2	2	0.4	15	0.0005	4.19×10^{-6}	1.74×10^{-4}
5	64	128	2	2	0.5	10	0.0005	6.85×10^{-6}	2.25×10^{-4}
6	64	128	8	6	0.5	100	0.0005	3.29×10^{-6}	1.54×10^{-4}
7	32	256	4	4	0.2	25	1×10^{-5}	2.16×10^{-6}	8.98×10^{-5}
8	32	256	4	4	0.3	20	1×10^{-5}	2.73×10^{-6}	1.03×10^{-4}
9	32	256	4	12	0.3	100	1×10^{-5}	8.87×10^{-7}	5.87×10^{-5}
10	32	512	8	10	0.4	50	0.001	1.67×10^{-4}	7.96×10^{-4}
11	32	256	4	3	0.5	10	0.001	3.27×10^{-5}	3.83×10^{-4}
12	32	256	4	3	0.3	10	0.001	2.29×10^{-6}	9.66×10^{-5}
13	32	512	8	8	0.3	50	0.0001	8.16×10^{-7}	5.53×10^{-5}
14	32	512	4	10	0.4	50	0.0001	7.90×10^{-7}	5.40×10^{-5}
15	32	256	4	4	0.4	15	0.0001	5.70×10^{-6}	1.54×10^{-4}
16	32	256	4	3	0.4	15	0.0001	7.02×10^{-6}	1.76×10^{-4}
17	16	256	2	12	0.2	200	5×10^{-5}	1.98×10^{-7}	1.71×10^{-5}
18	16	128	2	10	0.5	100	5×10^{-5}	2.54×10^{-6}	6.60×10^{-5}
19	16	128	2	2	0.3	20	1×10^{-5}	3.07×10^{-6}	7.69×10^{-5}

Continued on next page

Table 4.2 (continued)

ID	BS	d	n	L	dp	Ep	LR	Test Loss	Test MPJPE
20	16	512	2	8	0.5	50	0.001	7.81×10^{-5}	3.89×10^{-4}
21	16	512	8	6	0.2	25	0.001	8.15×10^{-5}	4.03×10^{-4}
22	16	512	8	6	0.5	10	0.001	7.80×10^{-5}	3.85×10^{-4}
23	16	512	8	6	0.5	20	0.0001	5.45×10^{-7}	3.37×10^{-5}
24	16	512	8	6	0.4	15	0.0001	6.92×10^{-7}	3.79×10^{-5}

Table 4.3: Evaluation of 2D-to-3D Reconstruction Experiments (RNN). Columns: **ID** is the experiment identifier; **BS** is the batch size; **h** is the hidden size; **L** is the number of layers; **dp** is the dropout rate; **Ep** is the number of epochs; **LR** is the learning rate; **Test Loss** is the loss on the test set; **Test MPJPE** is the mean per joint position error.

ID	BS	h	L	dp	Ep	LR	Test Loss	Test MPJPE
25	64	512	2	0.3	20	5×10^{-5}	1.25×10^{-5}	3.06×10^{-4}
26	64	512	3	0.4	15	0.0005	1.79×10^{-6}	1.19×10^{-4}
27	64	512	3	0.5	10	0.0005	3.39×10^{-6}	1.63×10^{-4}
28	32	128	2	0.2	25	1×10^{-5}	3.46×10^{-6}	1.16×10^{-4}
29	32	256	2	0.3	20	1×10^{-5}	3.36×10^{-6}	1.14×10^{-4}
30	32	128	1	0.3	10	0.001	1.10×10^{-6}	6.61×10^{-5}
31	32	256	3	0.4	15	0.0001	6.39×10^{-6}	1.64×10^{-4}
32	16	128	1	0.2	25	0.001	6.41×10^{-7}	3.76×10^{-5}
33	16	256	1	0.5	10	0.001	6.82×10^{-7}	3.60×10^{-5}
34	16	256	2	0.5	20	0.0001	5.67×10^{-7}	3.45×10^{-5}

Table 4.4: Evaluation of 2D-to-3D Reconstruction Experiments (FCNN). Columns: **ID** is the experiment identifier; **BS** is the batch size; **Hidden** represents the concatenated sizes of the hidden layers; **Ep** is the number of training epochs; **LR** is the learning rate; **Test Loss** is the loss on the test set; **Test MPJPE** is the mean per joint position error.

ID	BS	Hidden	Ep	LR	Test Loss	Test MPJPE
35	64	256,128	20	5×10^{-5}	3.31×10^{-5}	5.17×10^{-4}

Continued on next page

Table 4.4 (continued)

ID	BS	Hidden	Ep	LR	Test Loss	Test MPJPE
36	64	256,128,64	15	0.0005	2.44×10^{-5}	4.47×10^{-4}
37	32	512,256	25	1×10^{-5}	2.61×10^{-6}	1.02×10^{-4}
38	32	512,256,128	10	0.001	7.77×10^{-6}	1.74×10^{-4}
39	32	512,256,128,64	15	0.0001	1.27×10^{-5}	2.26×10^{-4}
40	16	1024,512,256,128,64	25	0.001	5.52×10^{-6}	1.09×10^{-4}
41	16	1024,512	10	0.001	2.80×10^{-6}	7.70×10^{-5}
42	16	1024,512,256	20	0.0001	2.12×10^{-6}	6.92×10^{-5}

Table 4.5: Evaluation of 2D-to-3D Reconstruction Experiments (CNN). Columns: **ID** is the experiment identifier; **BS** is the batch size; **ch1** to **ch5** represent the number of filters in each convolutional layer (if a row has fewer than 5 layers, the remaining columns are left blank); **Ep** is the number of epochs; **LR** is the learning rate; **Test Loss** is the loss on the test set; **Test MPJPE** is the mean per joint position error.

ID	BS	ch1	ch2	ch3	ch4	ch5	Ep	LR	Test Loss	Test MPJPE
43	64	256	512	1024			20	5×10^{-5}	1.96×10^{-4}	1.41×10^{-3}
44	64	64	128	256			15	0.0005	2.29×10^{-5}	3.92×10^{-4}
45	64	128	256	512			10	0.0005	4.54×10^{-5}	5.99×10^{-4}
46	32	64	128				20	1×10^{-5}	5.63×10^{-6}	1.54×10^{-4}
47	32	128	256	512	1024		15	0.0001	1.89×10^{-4}	9.26×10^{-4}
48	16	64	128	256	512	1024	25	0.001	2.58×10^{-6}	7.48×10^{-5}
49	16	64	128	256	512		10	0.001	5.70×10^{-6}	1.06×10^{-4}
50	16	128	256				20	0.0001	5.86×10^{-6}	9.86×10^{-5}

4.4.2 Numerical Evaluation for 2D-to-3D

Our experiments demonstrate that the proposed multi-view 2D-to-3D reconstruction module achieves excellent performance across different architectures with very low error rates, as summarized in Table 4.6. In particular, the Transformer-based model (namely configuration ID=17) achieved the lowest MPJPE of 1.71×10^{-5}, outperforming the other approaches. The superior

performance of the Transformer can be attributed to its self-attention mechanism, which efficiently captures global spatial dependencies and integrates information from multiple views. This allows the model to learn robust feature representations and subtle inter-joint relationships, leading to very good results across different hyper-parameters' configurations.

Also, the other architectures achieved interesting results. In particular, the RNN-based models—which obtained the third lowest MPJPE—leveraged sequential processing to capture temporal dependencies among keypoints, enabling efficient performance. In contrast, the CNN-based models, although effective at extracting local spatial features via convolutional filters, yielded higher error rates due to their limited ability to capture global context. Similarly, the FCNN-based model, relying solely on fully-connected layers, did not exploit spatial or temporal relationships effectively, which resulted in relatively poorer performance. Overall, these findings underscore the importance of combining both local and global feature modeling to achieve accurate 3D reconstruction. Further experiments could leverage extra training with increasing number of epochs on every architecture and additional testing on the various specifics.

Model and ID	**Test Loss**	**Test MPJPE**
Transformer 17	1.98×10^{-7}	1.71×10^{-5}
Transformer 23	5.45×10^{-7}	3.37×10^{-5}
RNN 33	6.82×10^{-7}	3.60×10^{-5}

Table 4.6: 3 best performing models for 2D-to-3D reconstruction.

4.4.3 SMPL/SMPL-X Parameter Regression Performance

In the second module of our pipeline, 3D keypoints are used to regress SMPL-X parameters. Our evaluation (see Table 4.7) compares different configurations. As before, each configuration is identified by an ID followed by a concise description of the model details and the test metrics that capture various aspects of accuracy. The proposed setups achieve low errors in this

high-dimensional regression task, highlighting their effectiveness in modeling complex inter-parameter relationships.

Table 4.7: Main evaluation of 3D-to-SMPLX reconstruction experiments. Columns: **ID** is the experiment identifier; **BS** is the batch size; **d** is the embedding dimension; **n** is the number of attention heads; **L** is the number of layers; **dp** is the dropout rate; **Ep** is the number of training epochs; **LR** is the learning rate; **Test Loss** is the loss on the test set (MSE); **Test MAE** is the mean absolute error; **Test RMSE** is the root mean square error.

ID	BS	d	n	L	dp	Ep	LR	Test Loss	Test MAE	Test RMSE
1	16	128	8	28	0.5	80	3.5×10^{-4}	8.67×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
2	64	512	8	27	0.5	75	3.4×10^{-4}	8.68×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
3	48	384	6	26	0.45	70	3.3×10^{-4}	8.68×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
4	32	256	8	25	0.4	65	3.2×10^{-4}	8.68×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
5	24	128	4	24	0.35	60	3.1×10^{-4}	8.69×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
6	16	512	8	23	0.3	55	3.0×10^{-4}	8.67×10^{-2}	1.54×10^{-1}	2.94×10^{-1}
7	64	384	6	22	0.5	80	2.9×10^{-4}	8.68×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
8	48	128	8	21	0.5	75	2.8×10^{-4}	8.67×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
9	32	512	8	20	0.5	90	2.7×10^{-4}	8.69×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
10	24	384	6	19	0.45	85	2.6×10^{-4}	8.70×10^{-2}	1.54×10^{-1}	2.94×10^{-1}
11	16	256	8	18	0.45	80	2.5×10^{-4}	1.27×10^{-1}	1.54×10^{-1}	2.95×10^{-1}
12	64	128	8	17	0.4	75	2.4×10^{-4}	8.73×10^{-2}	1.54×10^{-1}	2.94×10^{-1}
13	48	512	8	16	0.35	70	2.3×10^{-4}	8.72×10^{-2}	1.54×10^{-1}	2.94×10^{-1}
14	32	384	6	15	0.3	65	2.2×10^{-4}	8.72×10^{-2}	1.54×10^{-1}	2.94×10^{-1}
15	24	128	4	14	0.25	60	2.1×10^{-4}	1.65×10^{-2}	3.06×10^{-2}	5.97×10^{-2}
16	16	512	8	13	0.2	55	2.0×10^{-4}	8.67×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
17	64	384	6	12	0.5	70	2.0×10^{-4}	5.48×10^{-3}	3.77×10^{-2}	7.40×10^{-2}
18	48	256	8	11	0.5	85	1.8×10^{-4}	2.22×10^{-2}	3.72×10^{-2}	7.30×10^{-2}
19	32	128	8	10	0.5	100	1.7×10^{-4}	1.33×10^{-2}	6.35×10^{-3}	6.35×10^{-2}
20	24	512	8	9	0.45	95	1.6×10^{-4}	3.34×10^{-2}	6.72×10^{-3}	8.19×10^{-2}
21	16	384	6	8	0.4	90	1.5×10^{-4}	2.68×10^{-2}	6.48×10^{-3}	8.05×10^{-2}
22	64	256	8	7	0.35	85	1.5×10^{-4}	1.10×10^{-2}	3.11×10^{-2}	5.57×10^{-2}
23	48	128	4	6	0.3	80	1.5×10^{-4}	1.03×10^{-2}	3.20×10^{-2}	5.66×10^{-2}

Continued on next page

Table 4.7 (continued)

ID	BS	d	n	L	dp	Ep	LR	Test Loss	Test MAE	Test RMSE
24	32	512	8	5	0.25	75	1.0×10^{-4}	1.15×10^{-2}	2.77×10^{-2}	5.37×10^{-2}
25	24	384	6	4	0.2	65	1.0×10^{-4}	8.72×10^{-3}	2.72×10^{-2}	5.24×10^{-2}
26	16	256	8	3	0.1	60	1.0×10^{-4}	7.93×10^{-3}	2.54×10^{-2}	4.86×10^{-2}
27	64	128	2	10	0.5	70	9.0×10^{-4}	8.67×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
28	48	512	8	9	0.5	80	8.0×10^{-4}	8.68×10^{-2}	1.55×10^{-1}	2.95×10^{-1}
29	32	384	6	8	0.45	100	7.0×10^{-4}	8.67×10^{-2}	1.54×10^{-1}	2.95×10^{-1}
30	24	256	4	5	0.4	90	6.0×10^{-4}	1.32×10^{-2}	6.09×10^{-2}	9.00×10^{-2}
31	16	256	4	7	0.35	100	5.0×10^{-4}	1.30×10^{-2}	6.08×10^{-2}	9.00×10^{-2}
32	16	128	2	6	0.3	90	4.0×10^{-4}	1.57×10^{-2}	4.46×10^{-3}	3.52×10^{-2}
33	16	512	8	5	0.25	80	3.0×10^{-4}	1.88×10^{-2}	4.01×10^{-3}	3.30×10^{-2}
34	32	384	6	4	0.2	70	2.0×10^{-4}	1.16×10^{-2}	3.15×10^{-3}	2.96×10^{-2}
35	24	256	8	3	0.15	65	1.0×10^{-4}	7.72×10^{-3}	2.67×10^{-2}	2.73×10^{-2}
36	32	128	4	3	0.15	120	4.0×10^{-4}	1.78×10^{-2}	4.27×10^{-3}	6.54×10^{-2}
37	24	256	2	10	0.5	300	3.0×10^{0}	1.33×10^{-1}	1.56×10^{-1}	1.56×10^{-1}
38	16	512	8	8	0.4	250	2.0×10^{0}	1.31×10^{-1}	1.55×10^{-1}	1.55×10^{-1}
39	64	384	6	6	0.3	200	2.0×10^{-4}	2.28×10^{-2}	5.22×10^{-3}	3.77×10^{-2}
40	48	256	4	4	0.2	150	1.0×10^{-4}	8.00×10^{-3}	2.59×10^{-3}	5.09×10^{-2}
41	16	128	2	2	0.1	100	1.0×10^{-5}	5.39×10^{-3}	5.86×10^{-3}	7.66×10^{-2}

4.4.4 Discussion and Best Configurations

Our evaluation of the SMPL-X regression module considered three key error metrics: RMSE, MSE (test loss), and MAE. Based on the experimental results summarized in Table 4.7, we identify the following top-performing configurations for each metric:

Top Three by RMSE:

- **Configuration 35:** $d = 256,\ n = 8,\ L = 3,\ dp = 0.15$ with RMSE = 2.73×10^{-2}

- **Configuration 34:** $d = 384,\ n = 6,\ L = 4,\ dp = 0.2$ with RMSE = 2.96×10^{-2}
- **Configuration 33:** $d = 512,\ n = 8,\ L = 5,\ dp = 0.25$ with RMSE = 3.30×10^{-2}

Top Three by MSE (Test Loss):

- **Configuration 41:** $d = 128,\ n = 2,\ L = 2,\ dp = 0.1$ with MSE = 5.39×10^{-3}
- **Configuration 17:** $d = 384,\ n = 6,\ L = 12,\ dp = 0.5$ with MSE = 5.48×10^{-3}
- **Configuration 35:** $d = 256,\ n = 8,\ L = 3,\ dp = 0.15$ with MSE = 7.72×10^{-3}

Top Three by MAE:

- **Configuration 34:** $d = 384,\ n = 6,\ L = 4,\ dp = 0.2$ with MAE = 3.15×10^{-3}
- **Configuration 33:** $d = 512,\ n = 8,\ L = 5,\ dp = 0.25$ with MAE = 4.01×10^{-3}
- **Configuration 32:** $d = 128,\ n = 2,\ L = 6,\ dp = 0.3$ with MAE = 4.46×10^{-3}

Overall, although configuration 35 delivers the best RMSE and configuration 41 the best MSE, configuration 34 offers a compelling balance by achieving the lowest MAE while maintaining a competitive RMSE. Given that MAE robustly reflects the average prediction error, configuration 34 emerges as the best overall model for SMPL-X parameter regression.

4.4.5 Group-wise Error Analysis

To further analyze the performance of the regression module, we computed the MAE for the SMPL-X parameters. This group-wise evaluation was useful to guarantee effective and additional indications about our approach but they were not used as key metrics for assessing the performance, nonetheless some groups—such as hand poses and facial expressions—exhibit slightly higher error rates compared to others, indicating more difficulties on some body parts, in particular hand poses. For simplicity and brevity, we report the table with these results in the Appendix A.2.

The regression metrics indicate that our approach is capable of accurately capturing the complex, high-dimensional space of SMPL/SMPL-X parameters. Variations in hyper-parameters such as dropout, Transformer depth, and learning rate have a noticeable impact on performance. The group-wise analysis, in particular, suggests that further improvements might be possible by tailoring the network architecture or loss functions to better handle challenging parameter groups, such as hand poses.

4.5 Qualitative Results

Building on our quantitative findings, this section provides a detailed visual evaluation of the pipeline's outputs. We assess the reconstructed 3D poses and the corresponding SMPL/SMPL-X models, emphasizing their visual quality, anatomical plausibility, and consistency across different viewpoints. In our analysis, we present results from multiple models, and then focus on a specific configuration that exemplifies the best trade-off between accuracy and visual coherence. This qualitative evaluation not only confirms the effectiveness of our approach but also highlights areas where further refinements could enhance performance.

4.5.1 Visual Examples of 3D Pose Reconstructions

The first stage of our pipeline converts multi-view 2D keypoints detections into coherent 3D joint reconstructions. Given the extensive experiments conducted, we present visual results for the three best performing models to highlight their capabilities in handling various scenarios. Specifically, for each model we showcase on Figure 4.2 two representative exercises: one depicting a relatively easy situation and another featuring a more challenging pose.

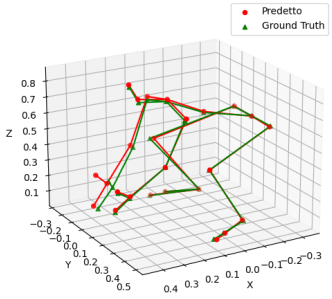
Additionally, we show some extra visual results in Appendix B.1.

4.5.2 2D-to-3D Visual Analysis

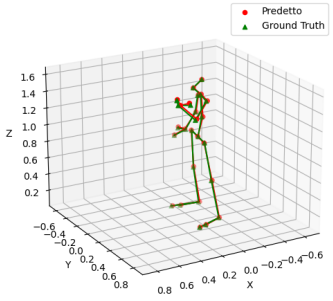
The visual results demonstrate that our 2D-to-3D reconstruction models are effective in converting multi-view 2D keypoints detections into coherent 3D poses. Although the reconstructions are not perfect, the images clearly show that some models perform with higher precision than others. In particular, the first Transformer-based model (ID 17) has delivered excellent results across a variety of scenarios, effectively handling occlusions and dynamic environments. Its outputs exhibit strong anatomical plausibility and maintain consistent joint localization even when some views are partially obscured or when the subject is performing rapid movements. These observations validate the robustness of our approach, while also indicating that there is still room for further refinement to address occasional inaccuracies.

4.5.3 SMPL-X Model Visualizations

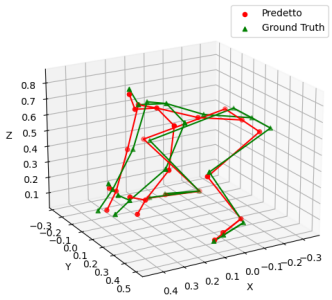
Following the 3D reconstruction, our pipeline regresses SMPL/SMPL-X parameters to generate detailed parametric body models. The generation of these meshes is slightly trickier compared to the 2D-to-3D part, due to the inner complexity of representing the SMPL-X model with traditional tools inside the programming languages. Nonetheless, we present in Figure 4.3 some examples obtained using Blender, a software for graphics, where we visualize



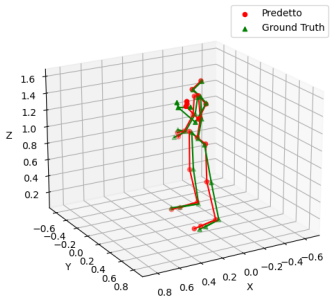
(a) Transformer (ID 17) - Subject 11, Warmup 18



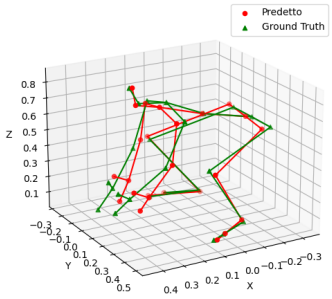
(b) Transformer (ID 17) - Subject 5, Dumbbell Curl



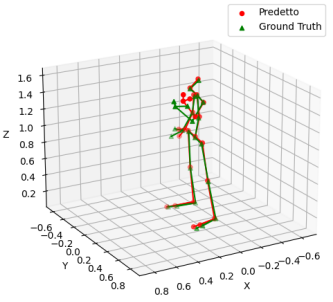
(c) Transformer (ID 23) - Subject 11, Warmup 18



(d) Transformer (ID 23) - Subject 5, Dumbbell Curl



(e) RNN (ID 33) - Subject 11, Warmup 18



(f) RNN (ID 33) - Subject 5, Dumbbell Curl

Figure 4.2: Visual examples of 3D pose reconstructions for the three best performing models. Each model is evaluated on the same two exercises.

some ground truth meshes together with the predictions made by the 3 best models, as discussed in the quantitative section. In every image, we present first of all (on the left) the ground truth SMPL-X mesh, followed by the predictions made by the Transformer-based model at configuration/ID 34, 35 and 41, in order.

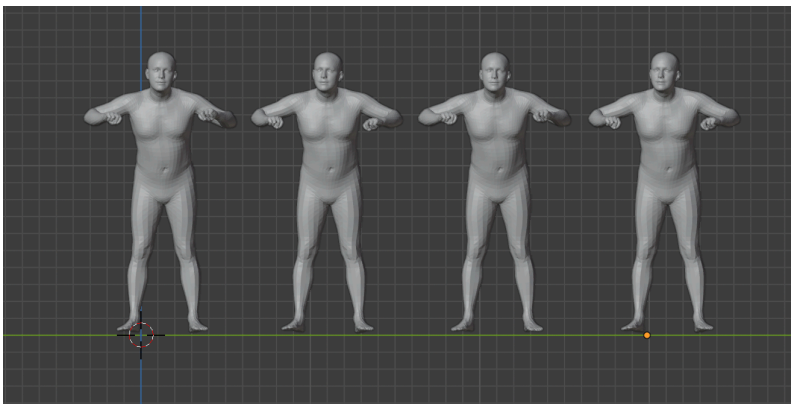
These visualizations not only capture the overall body pose but also reveal subtle anatomical details—such as precise hand articulations and facial expressions. The generated models are anatomically plausible, with smooth deformations and minimal limb interpenetration, underscoring the accuracy of our regression module. Although minor discrepancies are noticeable compared to the ground truth (for example, slight differences in the subject's height or facial orientation), the overall reconstruction is highly efficient and faithful.

4.5.4 Discussion of Visual Quality and Anatomical Plausibility

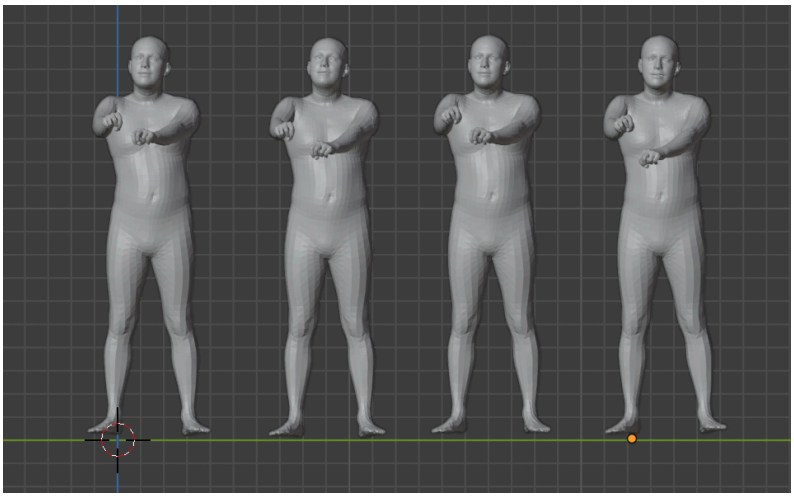
The qualitative results validate the overall effectiveness of our pipeline in reconstructing accurate 3D poses and generating detailed SMPL-X models from multi-view videos. The reconstructed poses closely align with the ground truth, and the resulting models display realistic body proportions with smooth deformations. Nevertheless, some limitations persist:

- **Preservation of Fine Details:** While the reconstructed body shape and pose are generally accurate, subtle anatomical details—such as facial expressions and the precise dimensions of certain body parts—may exhibit slight discrepancies from the ground truth. Although these variations are minor, refining the model's ability to capture intricate details could further enhance realism.

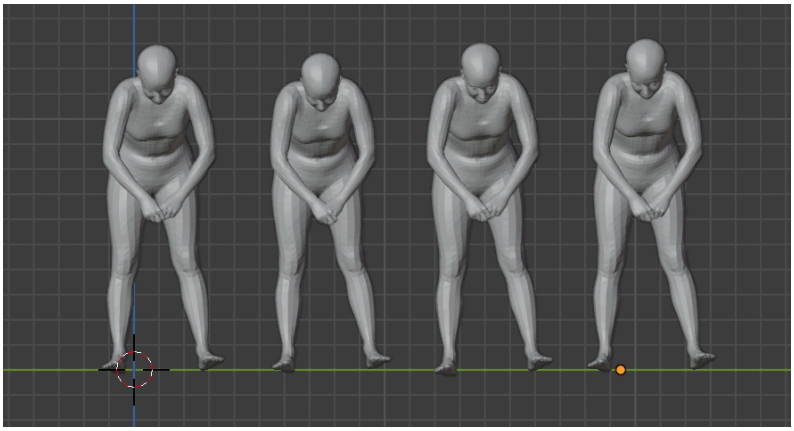
- **Robustness to Extreme Poses:** In highly dynamic or extreme motion scenarios, minor deviations in joint positioning can occur. Introducing additional temporal smoothing techniques or incorporating more



(a) Sample 0: Comparison for a slightly folded position with open arms.



(b) Sample 1: Comparison for a standing position with arms upfront.



(c) Sample 2: Comparison for a squat position.

Figure 4.3: Visual examples of 3D pose reconstructions for the three best performing models, arranged vertically. On the left, the ground truth, followed by the predictions of the three best performing models (configurations 34, 35 and 41).

advanced motion priors could help improve accuracy, particularly in challenging motion sequences.

- **Consistency Across Views:** The multi-view fusion strategy effectively reduces occlusions; however, occasional inconsistencies still arise in cluttered environments. Further optimizing the fusion algorithm could enhance coherence across viewpoints, ensuring a more seamless reconstruction.

Overall, these observations confirm the robustness of our approach while also highlighting specific areas for refinement. The insights gained from this analysis provide a clear path for future improvements to further enhance both the anatomical plausibility and consistency of the reconstructed models.

4.5.5 End-to-End Pipeline

In this section, we present the complete pipeline in action. Four multi-view videos from our dataset are processed by extracting 2D keypoints from one frame per view and then feeding them into our approach to generate the final SMPL-X body model. The results demonstrate the end-to-end effectiveness of our system in producing anatomically plausible and detailed 3D reconstructions.

In the given example, we randomly selected subject s05 performing the clean and press exercise at frame 814. The overall pipeline is summarized in Figure 4.4. Figure 4.5 shows the multi-view images, while Figure 4.6 compares the predicted 3D joints with the ground truth. The SMPL-X reconstruction is illustrated in Figure 4.7.

In an end-to-end pipeline, errors from individual stages tend to accumulate. For example, small inaccuracies during 2D keypoints extraction are magnified in the 3D reconstruction stage, resulting in noticeable misalignment in the 3D joint estimates. When these 3D joints are subsequently used to regress

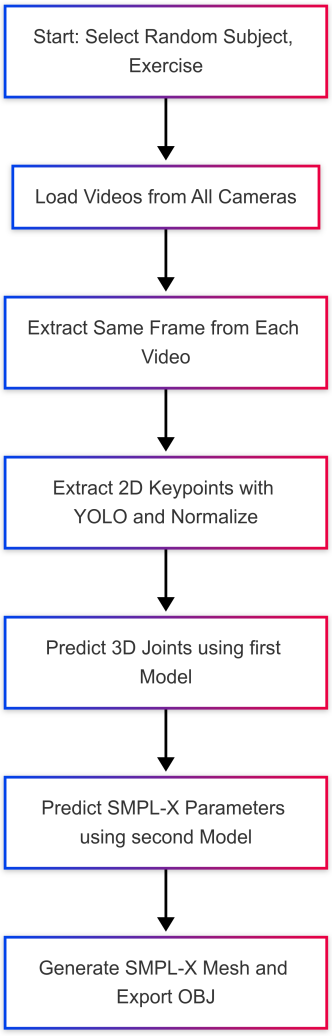
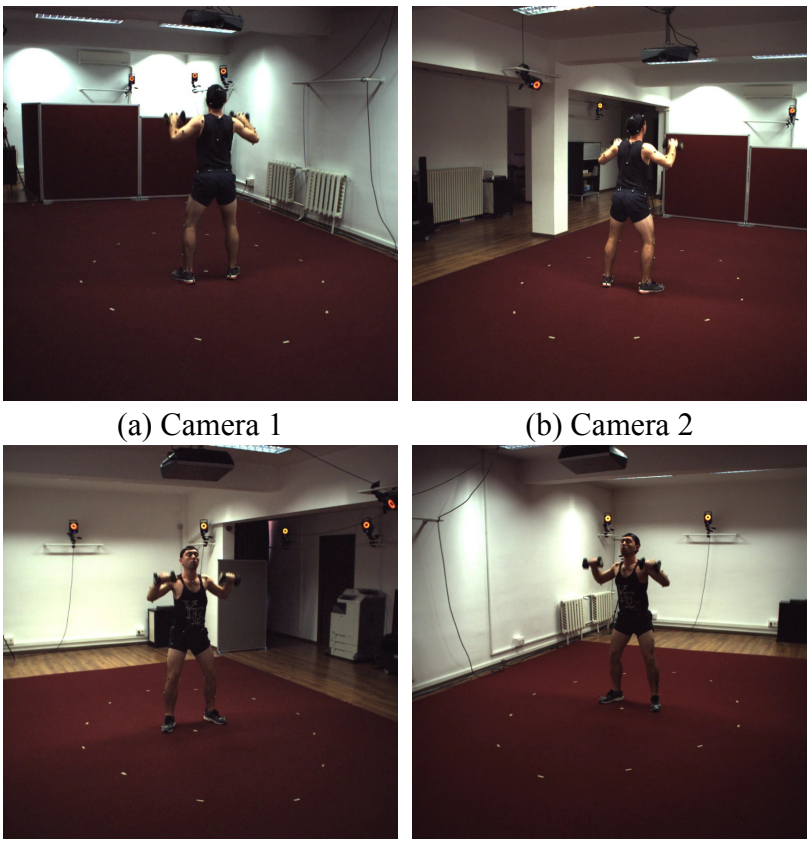


Figure 4.4: Diagram of the complete pipeline, from multi-view video processing to 3D SMPL-X body model reconstruction.

the SMPL–X parameters, the compounded errors can lead to certain body parts – such as the face – appearing less accurate. Despite these challenges, the overall pipeline demonstrates robust performance and efficiency in generating anatomically plausible 3D reconstructions. The system successfully integrates multiple processing steps into a cohesive workflow, and while there is room for further refinement to minimize error propagation, the results validate the practical viability of the approach.



(a) Camera 1 (b) Camera 2

(c) Camera 3 (d) Camera 4

Figure 4.5: Multi-view images captured from the four cameras.

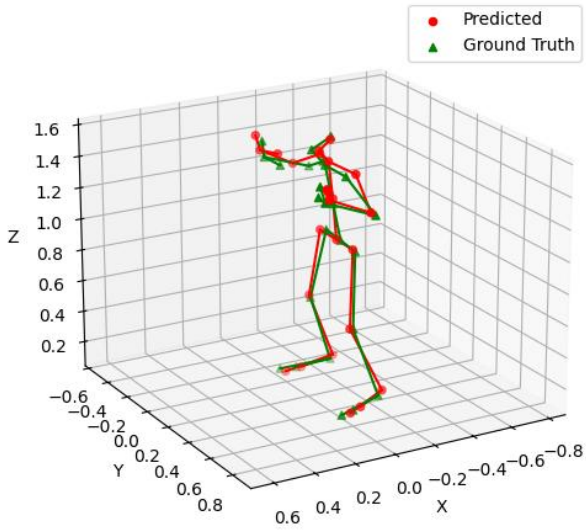


Figure 4.6: Comparison between predicted joints3D and ground truth.

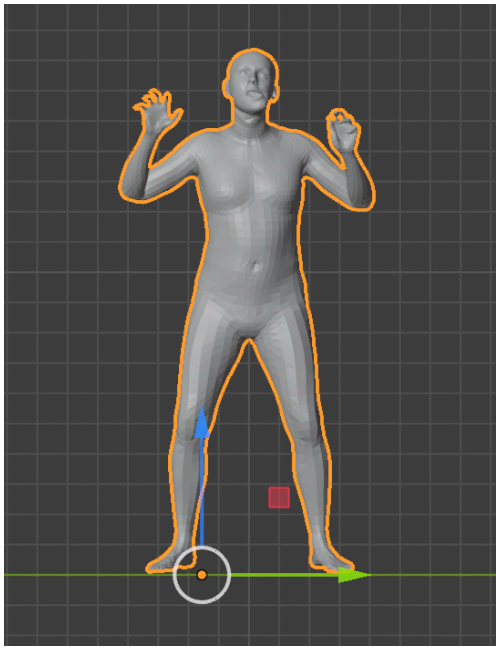


Figure 4.7: Reconstructed SMPL-X mesh.

4.6 Discussion

Our work introduces an innovative and modular pipeline that integrates multi-view 2D-to-3D reconstruction with SMPL/SMPL-X parameter regression in a unified, end-to-end framework. This approach is designed to efficiently capture both the global spatial relationships and the detailed anatomical features of the human body, thereby producing highly accurate 3D representations. The modular design allows each component to be optimized independently while maintaining seamless integration, making our pipeline versatile and adaptable to various scenarios.

The experimental results provide strong evidence for the effectiveness of our method. In the 2D-to-3D reconstruction module, we observed very low MPJPE values, indicating that the network is able to accurately infer 3D joint positions from multi-view 2D keypoints. In particular, Transformer-based architectures excel in this task by leveraging self-attention mechanisms to model global spatial dependencies. The quantitative metrics are corroborated by

qualitative assessments: reconstructed 3D poses consistently align well with the ground truth, and visual overlays of the predicted SMPL/SMPL-X models reveal that even subtle anatomical details, such as hand gestures and facial expressions, are captured with high fidelity.

In the SMPL/SMPL-X parameter regression module, our network successfully maps the 3D keypoints to a high-dimensional parameter space. Although different configurations yield slight variations in metrics such as MSE, MAE, and RMSE, the overall performance is robust. For instance, some configurations excel in reducing average errors, while others are more effective at mitigating larger deviations. This trade-off highlights the capacity of our model to balance different aspects of error minimization, ensuring that both average performance and the handling of outlier cases are addressed effectively.

Furthermore, our group-wise error analysis provides additional insights into the performance nuances of our regression module. While the overall parameter predictions are accurate, certain groups—particularly those related to hand poses and facial expressions—exhibit relatively higher errors. This suggests that these components, which capture more intricate and variable details, present a greater challenge. Nevertheless, the accuracy achieved across all parameter groups underscores the robustness of our approach and its capability to capture complex, interdependent features of human pose and shape.

Overall, the results obtained from our pipeline demonstrate that a modular design combined with advanced deep learning architectures can lead to significant improvements in 3D human pose estimation and parameter regression. The quantitative metrics validate the high accuracy of our reconstruction and regression processes, while the qualitative evaluations reinforce the visual and anatomical plausibility of the output. Our discussion confirms that this innovative approach is a promising solution for detailed 3D human modeling.

Chapter 5

Conclusions and Future Work

This thesis has presented a comprehensive pipeline for extracting detailed SMPL-X parameters from multi-view video sequences of fitness exercises. By integrating optimal 2D keypoints detection, robust 3D reconstruction, and advanced Transformer-based models for parametric regression, our work demonstrates that it is possible to derive anatomically plausible and accurate representations of the human body from raw multi-view videos using a modular approach.

Our pipeline employs standard preprocessing techniques to consolidate the most informative frames into unified dictionaries. These dictionaries serve as the basis for both 2D-to-3D pose estimation and SMPL-X parameter regression.

5.1 Summary of Contributions

The primary contributions of this thesis can be summarized as follows:

- **Multi-view Approach:** We developed a robust multi-view pipeline that uses four synchronized video streams and employed YOLOv8s detector for reliable 2D keypoints extraction. The preprocessing procedures then filter and consolidate the data into dictionaries, ensuring that only the most informative frames are used for training.

- **2D-to-3D Reconstruction Module:** We implemented and evaluated various architectures for 3D pose reconstruction, with Transformer-based models demonstrating particularly strong performance in capturing global spatial dependencies and achieving low MPJPE.
- **SMPL-X Parameter Regression:** We proposed a novel regression module based on Transformer architectures to accurately predict high-dimensional SMPL-X parameters from 3D keypoints, achieving detailed and anatomically coherent models.
- **Extensive Experimental Validation:** Through rigorous evaluations on the FIT3D dataset, our approach has been validated both quantitatively and qualitatively, highlighting its effectiveness in capturing the complexity of human motion in fitness environments.

5.2 Future Work

While the results of this thesis are promising, several challenges and limitations remain, suggesting avenues for future research:

5.2.1 Real-Time Processing

One of the key challenges in this research is improving the computational efficiency of the entire pipeline to enable real-time performance. Current methods may not be fast enough for applications that require immediate feedback, such as virtual coaching or live analysis during training sessions. Future work could focus on optimizing the algorithms, possibly through hardware acceleration or model compression techniques, to allow seamless real-time performance without sacrificing accuracy or robustness.

5.2.2 Robustness to Occlusions and Extreme Poses

While the current system performs well in controlled environments, it struggles with occlusions or extreme poses where key parts of the subject's body are obscured or moved outside the normal range. These scenarios are common in dynamic environments like sports or fitness activities, where rapid movements and multiple people can cause partial occlusion. Future research could involve the integration of additional sensor modalities, such as depth sensors or multi-view cameras, or advanced machine learning techniques like adversarial training, to help the model better handle occlusions and extreme poses.

5.2.3 Broader Generalization

The current model is trained and tested on a relatively narrow dataset, which may limit its generalization to a wider range of subjects, environments, and exercise types. For instance, people with different body types, ages, or levels of athletic ability may not be well-represented in the current dataset, and certain types of exercises may be underrepresented. Expanding the dataset to include more diverse subjects and activities would help improve the robustness and generalization of the model, making it more versatile for use in various applications.

5.2.4 Quantitative Comparisons with Alternative Approaches

Although the demo provides a functional proof of concept, a more comprehensive evaluation is necessary to quantify its performance. Future work should involve comparing the proposed approach with other emerging techniques in the field, using standardized benchmarks and metrics. This will help to identify the strengths and weaknesses of different models and allow for a more informed decision on the best approach for specific use cases. This comparison could include metrics such as accuracy, processing speed, robustness to

noise, and scalability.

5.2.5 Interactive and Personalized Applications

An exciting avenue for future work lies in the development of interactive systems that leverage the reconstructed 3D models for personalized performance analysis. By integrating the 3D models with real-time tracking, it would be possible to provide users with personalized coaching feedback, such as posture correction, motion efficiency, and injury prevention recommendations. Moreover, such systems could be tailored to specific exercises, providing users with insights into their performance and areas for improvement based on their unique movements and body mechanics.

5.2.6 Dynamic Tracking and Integration

The current system treats the reconstructed 3D model as a static representation of the subject, but future work could focus on integrating the 3D model with the subject's motion in real time. This dynamic tracking would allow the model to adapt and move in sync with the subject throughout the video. Such an enhancement would improve applications in augmented reality (AR) and motion analysis, where the model's ability to interact with real-world environments or to track complex movements is crucial. By integrating motion tracking technologies like IMUs or advanced computer vision techniques, the reconstructed model could become a dynamic representation of the subject's ongoing movements, enabling more accurate and interactive applications.

5.3 Final Remarks

In conclusion, this thesis lays a solid foundation for advanced AI-driven systems in human motion analysis. The integration of multi-view data acquisition, deep learning-based 3D reconstruction, and precise parametric regression

has proven effective in generating detailed and accurate 3D models. Future work will focus on addressing current limitations, extending practical applicability, and conducting comprehensive comparisons with similar emerging technologies.

Appendix A

A.1 2D-to-3D Train and Validation

We report the mentioned table containing the results of the experimented configurations also during training and validation. These results are not crucial for our considerations but were essential in order to guarantee efficient training and guidelines during the study.

Table A.1: Supplementary training and validation metrics by configuration.

ID	Train Loss	Train MPJPE	Val Loss	Val MPJPE
1	1.20×10^{-3}	4.94×10^{-2}	2.75×10^{-6}	1.06×10^{-4}
2	4.55×10^{-3}	9.30×10^{-2}	1.19×10^{-5}	1.94×10^{-4}
3	8.10×10^{-2}	3.89×10^{-1}	1.35×10^{-4}	7.11×10^{-4}
4	8.10×10^{-2}	3.89×10^{-1}	1.35×10^{-4}	7.11×10^{-4}
5	1.25×10^{-3}	4.80×10^{-2}	2.87×10^{-6}	9.89×10^{-5}
6	1.91×10^{-3}	5.88×10^{-2}	3.84×10^{-6}	1.19×10^{-4}
7	9.80×10^{-4}	4.26×10^{-2}	3.21×10^{-6}	9.19×10^{-5}
8	1.16×10^{-3}	4.69×10^{-2}	1.13×10^{-6}	4.43×10^{-5}
9	1.63×10^{-3}	5.53×10^{-2}	4.08×10^{-6}	6.18×10^{-5}
10	6.10×10^{-4}	3.44×10^{-2}	4.64×10^{-7}	3.09×10^{-5}
11	8.13×10^{-2}	3.90×10^{-1}	7.28×10^{-5}	3.77×10^{-4}
12	3.11×10^{-3}	7.56×10^{-2}	1.12×10^{-5}	1.59×10^{-4}
13	1.07×10^{-3}	4.58×10^{-2}	1.01×10^{-6}	4.68×10^{-5}

Continued on next page

Table A.1 (continued)

ID	Train Loss	Train MPJPE	Val Loss	Val MPJPE
14	4.90×10^{-4}	3.08×10^{-2}	4.25×10^{-7}	2.80×10^{-5}
15	4.80×10^{-4}	3.06×10^{-2}	6.20×10^{-7}	3.13×10^{-5}
16	3.67×10^{-3}	8.38×10^{-2}	3.07×10^{-6}	7.64×10^{-5}
17	4.57×10^{-3}	9.28×10^{-2}	3.74×10^{-6}	8.53×10^{-5}
18	1.50×10^{-4}	1.72×10^{-2}	7.43×10^{-8}	8.31×10^{-6}
19	2.20×10^{-3}	6.46×10^{-2}	1.27×10^{-6}	3.45×10^{-5}
20	8.15×10^{-2}	3.91×10^{-1}	3.17×10^{-5}	1.78×10^{-4}
21	8.17×10^{-2}	3.92×10^{-1}	3.42×10^{-5}	1.87×10^{-4}
22	8.20×10^{-2}	3.93×10^{-1}	3.22×10^{-5}	1.77×10^{-4}
23	8.60×10^{-4}	4.09×10^{-2}	5.15×10^{-7}	2.15×10^{-5}
24	1.02×10^{-3}	4.53×10^{-2}	5.73×10^{-7}	2.17×10^{-5}
25	4.15×10^{-3}	8.97×10^{-2}	7.92×10^{-6}	1.69×10^{-4}
26	7.00×10^{-4}	3.83×10^{-2}	1.33×10^{-6}	6.55×10^{-5}
27	1.36×10^{-3}	5.34×10^{-2}	2.32×10^{-6}	9.22×10^{-5}
28	2.45×10^{-3}	7.07×10^{-2}	1.89×10^{-6}	5.79×10^{-5}
29	2.50×10^{-3}	7.02×10^{-2}	2.62×10^{-6}	6.35×10^{-5}
30	5.90×10^{-4}	3.42×10^{-2}	7.79×10^{-7}	3.75×10^{-5}
31	5.08×10^{-3}	1.01×10^{-1}	3.85×10^{-6}	8.41×10^{-5}
32	6.50×10^{-4}	3.62×10^{-2}	4.19×10^{-7}	1.99×10^{-5}
33	7.10×10^{-4}	3.65×10^{-2}	4.28×10^{-7}	2.04×10^{-5}
34	1.01×10^{-3}	4.59×10^{-2}	4.55×10^{-7}	1.97×10^{-5}
35	1.53×10^{-2}	1.77×10^{-1}	2.10×10^{-5}	2.75×10^{-4}
36	1.38×10^{-2}	1.67×10^{-1}	1.26×10^{-5}	2.27×10^{-4}
37	3.69×10^{-3}	8.91×10^{-2}	1.69×10^{-6}	5.52×10^{-5}
38	8.74×10^{-3}	1.33×10^{-1}	3.78×10^{-6}	8.89×10^{-5}
39	1.45×10^{-2}	1.72×10^{-1}	6.52×10^{-6}	1.16×10^{-4}
40	1.08×10^{-2}	1.47×10^{-1}	2.81×10^{-6}	5.50×10^{-5}
41	8.09×10^{-3}	1.21×10^{-1}	2.05×10^{-6}	4.60×10^{-5}
42	7.19×10^{-3}	1.21×10^{-1}	1.53×10^{-6}	3.79×10^{-5}

Continued on next page

Table A.1 (continued)

ID	Train Loss	Train MPJPE	Val Loss	Val MPJPE
43	5.20×10^{-3}	1.01×10^{-1}	7.61×10^{-5}	6.31×10^{-4}
44	2.66×10^{-3}	7.41×10^{-2}	7.36×10^{-6}	1.63×10^{-4}
45	4.11×10^{-3}	9.14×10^{-2}	1.46×10^{-5}	2.46×10^{-4}
46	3.13×10^{-3}	8.03×10^{-2}	3.51×10^{-6}	8.12×10^{-5}
47	7.00×10^{-3}	1.17×10^{-1}	5.88×10^{-5}	3.59×10^{-4}
48	4.66×10^{-3}	9.73×10^{-2}	1.49×10^{-6}	4.16×10^{-5}
49	1.08×10^{-2}	1.47×10^{-1}	3.27×10^{-6}	5.55×10^{-5}
50	3.06×10^{-3}	7.90×10^{-2}	1.38×10^{-6}	3.83×10^{-5}

A.2 3D-to-SMPLX Group-wise Testing results

The table below presents our group-wise quantitative evaluation of the 3D-to-SMPLX reconstruction results. Each column corresponds to the mean absolute error (MAE) for different model components (e.g., Betas, Body, Expression, Global Orientation, Jaw, Left Hand, Left Eye, Right Eye, Translational error). While these metrics were not used to drive the main decision-making process in our study, they offer valuable insights into how the model performs across various body parts. We believe these detailed results are useful for guiding future research aimed at refining and optimizing reconstruction quality for specific regions.

Table A.2: Supplementary group-wise MAE results by configuration.

ID	Betas	Body	Expr	Glob	Jaw	LHand	LEye	REye	Transl
1	0.73945	0.18825	0.12922	0.23862	0.01731	0.07871	0.03530	0.02492	0.12782
2	0.73920	0.18889	0.12910	0.23472	0.01496	0.07894	0.03660	0.02732	0.12810
3	0.73856	0.18978	0.12870	0.23794	0.01601	0.07900	0.03498	0.02498	0.12934
4	0.74110	0.18796	0.12904	0.24010	0.01474	0.07879	0.03511	0.02515	0.12871
5	0.74092	0.18869	0.12958	0.24182	0.01466	0.07939	0.03506	0.02553	0.13035

Continued on next page

Table A.2 (continued)

ID	Betas	Body	Expr	Glob	Jaw	LHand	LEye	REye	Transl
6	0.73906	0.18873	0.12908	0.23878	0.01577	0.07858	0.03538	0.02489	0.12951
7	0.73925	0.18849	0.12874	0.24222	0.01502	0.07904	0.03522	0.02485	0.12997
8	0.73966	0.18872	0.12902	0.23856	0.01494	0.07874	0.03536	0.02539	0.12893
9	0.73994	0.18845	0.12860	0.24571	0.01739	0.07912	0.03533	0.02514	0.12930
10	0.73919	0.18889	0.12887	0.23661	0.01505	0.07859	0.03504	0.02484	0.12867
11	0.73937	0.18923	0.12895	0.23734	0.01461	0.07865	0.03503	0.02506	0.12727
12	0.73947	0.18880	0.12892	0.23972	0.01497	0.07889	0.03499	0.02485	0.12970
13	0.73959	0.18877	0.12889	0.23782	0.01629	0.07875	0.03528	0.02576	0.12812
14	0.73855	0.18920	0.12935	0.24173	0.01464	0.07871	0.03507	0.02521	0.12787
15	0.09995	0.03184	0.02951	0.05366	0.00359	0.02251	0.00640	0.00721	0.01655
16	0.73952	0.18898	0.12880	0.23948	0.01538	0.07867	0.03523	0.02509	0.12819
17	0.13895	0.04195	0.03435	0.05721	0.00464	0.02481	0.00810	0.00889	0.02858
18	0.13947	0.03979	0.03508	0.05389	0.00421	0.02528	0.00748	0.00841	0.02937
19	0.14688	0.04094	0.03875	0.08068	0.00436	0.02893	0.00808	0.00860	0.03330
20	0.15369	0.04649	0.03933	0.07274	0.00467	0.02876	0.00884	0.00953	0.03245
21	0.15423	0.04664	0.03981	0.06712	0.00443	0.02916	0.00819	0.00925	0.03263
22	0.10783	0.03179	0.02976	0.05022	0.00394	0.02282	0.00618	0.00725	0.02669
23	0.11380	0.03135	0.03089	0.05118	0.00439	0.02407	0.00652	0.00743	0.02812
24	0.09275	0.02805	0.02683	0.04802	0.00362	0.02045	0.00616	0.00686	0.02417
25	0.09188	0.02645	0.02627	0.04259	0.00382	0.02083	0.00565	0.00661	0.02503
26	0.08082	0.02436	0.02515	0.04101	0.00378	0.01988	0.00573	0.00629	0.02413
27	0.73902	0.18845	0.12865	0.23685	0.01536	0.07861	0.03518	0.02495	0.12834
28	0.73903	0.18959	0.12973	0.23984	0.01515	0.07918	0.03617	0.02770	0.12897
29	0.73895	0.18851	0.12879	0.23856	0.01708	0.07887	0.03526	0.02581	0.12795
30	0.24178	0.06631	0.05802	0.10731	0.00641	0.03981	0.01177	0.01265	0.04316
31	0.23957	0.06700	0.05668	0.10512	0.00662	0.03833	0.01215	0.01261	0.04367
32	0.12383	0.03622	0.03310	0.05810	0.00530	0.02553	0.00679	0.00787	0.02961
33	0.11330	0.03488	0.03158	0.05450	0.00367	0.02359	0.00675	0.00771	0.02716
34	0.10366	0.03021	0.02812	0.04616	0.00387	0.02178	0.00621	0.00687	0.02551

Continued on next page

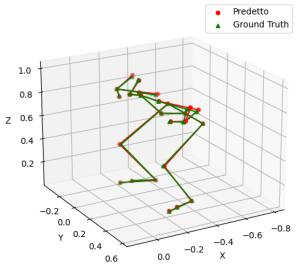
Table A.2 (continued)

ID	Betas	Body	Expr	Glob	Jaw	LHand	LEye	REye	Transl
35	0.09099	0.02606	0.02700	0.04219	0.00350	0.02113	0.00577	0.00668	0.02585
36	0.73903	0.19125	0.12914	0.23970	0.02430	0.08037	0.04799	0.02591	0.13347
37	0.74032	0.18862	0.12936	0.24437	0.01649	0.07991	0.03694	0.02925	0.13135
38	0.13593	0.04081	0.03452	0.06603	0.00444	0.02597	0.00718	0.00837	0.02944
39	0.08641	0.02523	0.02499	0.04122	0.00339	0.01998	0.00521	0.00622	0.02334
40	0.14882	0.03788	0.03893	0.06936	0.00601	0.03129	0.00682	0.00836	0.03645
41	0.14882	0.03788	0.03893	0.06936	0.00601	0.03129	0.00682	0.00836	0.00539

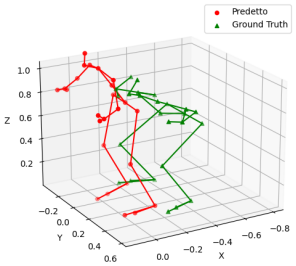
Appendix B

B.1 2D-to-3D Extra Visual Samples

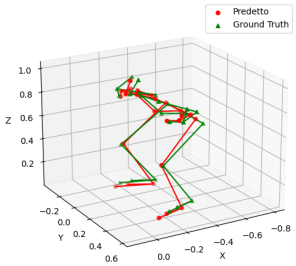
In this section, we present additional visual samples from our 2D-to-3D reconstruction experiments. These extra images offer a qualitative perspective on the model outputs, complementing our evaluations. Notably, the best three configurations demonstrate remarkable accuracy, as their reconstructed poses align closely with the ground truth. In contrast, models with lower performance exhibit more visible discrepancies, which highlight areas for potential improvement. While these visual samples were not directly used to make the final decisions in our study, they serve as an important resource for understanding the subtle differences in reconstruction quality across various architectures, and they may guide future refinements in our approach. We show some plots comparing the best models with some worse configurations: in the first spot we see the prediction by the best model, namely Transformer with ID=17, followed by the other models.



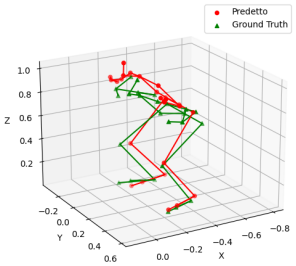
(a) Subject 8 Squat: Transformer Configuration 17



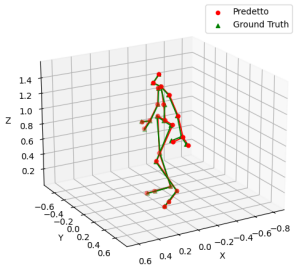
(b) Subject 8 Squat: Transformer Configuration 21



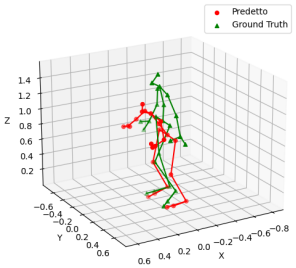
(c) Subject 8 Squat: CNN Configuration 46



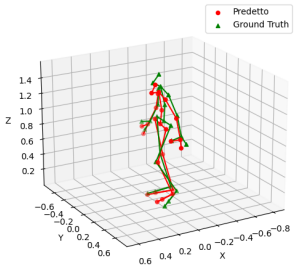
(d) Subject 8 Squat: FCNN Configuration 40



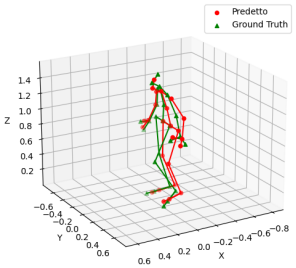
(e) Subject 7 Warmup 15: Transformer Configuration 17



(f) Subject 7 Warmup 15: Transformer Configuration 21



(g) Subject 7 Warmup 15: CNN Configuration 46

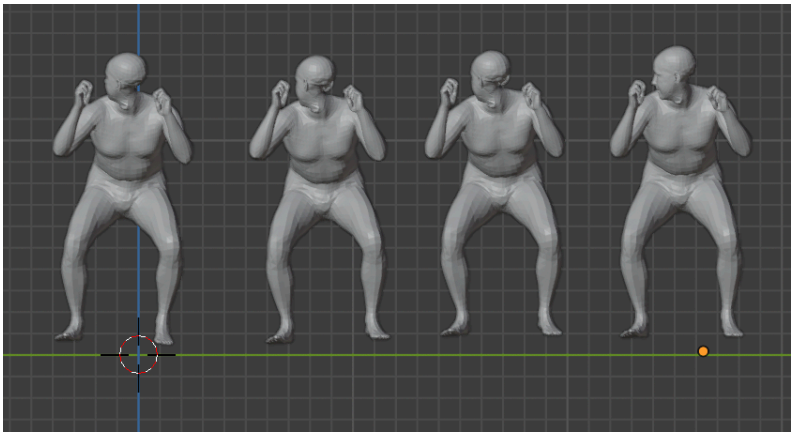


(h) Subject 7 Warmup 15: FCNN Configuration 40

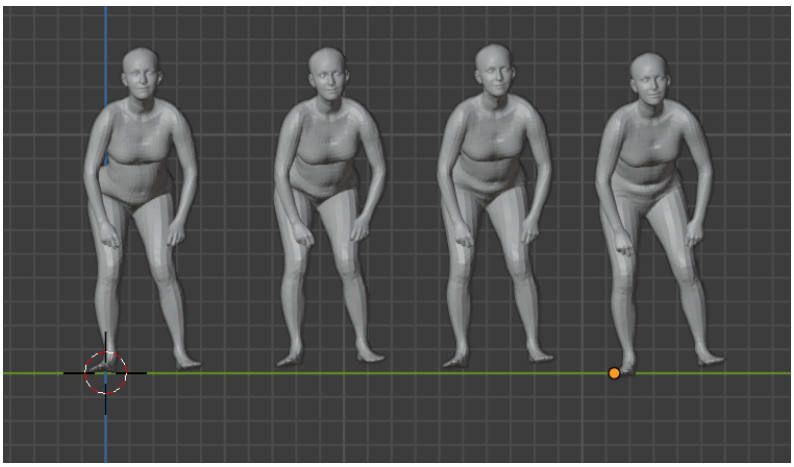
Figure B.1: Additional visual examples in Appendix B.

B.2 3D-to-SMPLX Extra Visual Samples

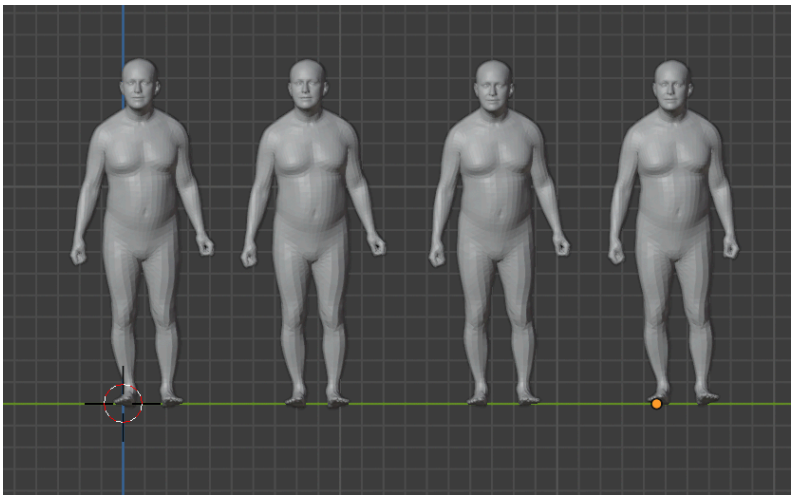
In this section, we present additional visual samples from our 3D-to-SMPLX reconstruction experiments, focusing again exclusively on the three best-performing configurations. While other models achieved competitive evaluation scores, their reconstructions exhibited noticeable inaccuracies, particularly in fine details and body proportions. In contrast, the selected configurations consistently produced convincing results, accurately capturing body shape, pose, and facial expressions. These qualitative examples further illustrate their reliability, particularly in regions such as the hands and face. Although these samples were not used to determine final rankings, they provide valuable insights that can guide future improvements in reconstructing intricate anatomical details. Inside each image on Figure B.2, in order, we find the ground truth, Transformer Configuration 34, Configuration 35 and 41.



(a) Sample 4: Comparison on a folded pose with arms up.



(b) Sample 5: Comparison on a static pose with back folded.



(c) Sample 6: Comparison on a perfectly standing pose.

Figure B.2: Additional visual examples of SMPL-X reconstruction.

Bibliography

[1] D. Anguelov, P. Srinivasan, D. Koller, L. Pishchulin, S. Savarese, and J. Davis. Scape: shape completion and animation for people. In *ACM Transactions on Graphics (TOG)*, volume 24 of number 3, pages 408–416. ACM, 2005.

[2] U. Author. Early approaches to 3d human modeling using geometric primitives. Unpublished Report, 1975. Details on using cylinders and cones for body modeling.

[3] B. A. Ballester, J. D. Pedrera-Zamorano, et al. Analysis of lower limb movement during step exercise through video-based pose estimation. *Gait & Posture*, 55:229–235, 2017.

[4] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. 3d pictorial structures revisited: multiple human pose estimation. In *2014 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 467–475. IEEE, 2014.

[5] C. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black. Smplify: 3d human pose and shape estimation from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–579, 2016.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

[7] J.-K. Chen, P. Kwong, C. Chang, V. Luk, R. Bajcsy, J.-C. Chen, and Q. Cheng. Wearable sensors for reliable fall detection. *IEEE Engineering in Medicine and Biology Society (EMBC)*:3712–3715, 2016.

[8] N. Cognome, N. Altro, and N. Altro. Fit3d: a comprehensive multi-view dataset for 3d human pose estimation in fitness environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1243, 2020.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893. IEEE, 2005.

[10] S. Doherty, H. Boulton, et al. Wearable technology for automated fall detection: evaluating approaches for effectiveness and acceptability. In *IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 1–3. IEEE, 2017.

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[12] R. C. Guler, M. Umar, and D. P. Siewiorek. Expose: exemplar-based 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1243, 2020. DOI: `10.1109/CVPR42600.2020.01234`.

[13] K. Ha. Smpl model introduction. `https://khanhha.github.io/posts/SMPL-model-introduction/`, 2021. Accessed: 2025-02-26.

[14] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018.

[15] N. Kolotouros, G. Pavlakos, and M. J. Black. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 101–110, 2019.

[16] C. Lassner, G. Pons-Moll, and M. J. Black. Smplify-x: 3d human pose and shape estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10766–10775, 2017.

[17] W. Li, F. Zhang, and S. Wang. Pose-based multi-stream network for fine-grained action recognition. *Neurocomputing*, 421:153–165, 2021.

[18] J. Lim, H. Choi, E. Roh, H. Yoo, and E. Kim. Assessment of airflow and microclimate for the running wear jacket with slits using cfd simulation. *Fashion and Textiles*, 2, December 2015. DOI: 10.1186/s40691-014-0025-2.

[19] L. Liu, B. Chen, and Z. Yang. Attention mechanism exploits single-frame cues for gesture recognition. *Neurocomputing*, 398:72–82, 2020.

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[22] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2017.

[23] D. Mehta, S. Sridhar, O. Sotnychenko, H. R. Xchang, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: real-time 3d human pose estimation with a single rgb camera. In *ACM Transactions on Graphics (ToG)*, volume 36 of number 4, page 44. ACM, 2017.

[24] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016.

[25] T. Nguyen, J.-C. Nebel, and F. Flórez-Revuelta. Recognition of activities of daily living with egocentric vision: a review. *Sensors*, 16:72, January 2016. DOI: 10.3390/s16010072.

[26] A. A. A. Osman, D. Tzionas, G. Pavlakos, K. Schindler, S. Tang, and M. J. Black. STAR: a sparse trained articulated human body regressor. https://arxiv.org/abs/2008.08622, 2021. Accessed: 2023-XX-XX.

[27] G. Pavlakos, G. Georgakis, V. Muhr, et al. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[28] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6988–6997, 2017.

[29] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7753–7762, 2019.

[30] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[31] V. Ramakrishna, D. Munoz, M. Hebert, R. Andrew, and T. Kanade. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision (ECCV)*, pages 573–586. Springer, 2012.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[33] A. A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli. Challenges in human behavior understanding: bridging the gap between perception and modeling. In *Human Behavior Understanding*, pages 1–12. Springer, 2010.

[34] M. Shah, A. Kale, W. Tavanapong, M. Abidi, and B. Abidi. Understanding human behavior from motion imagery. In *SoutheastCon, 2003. Proceedings. IEEE*, pages 214–219. IEEE, 2003.

[35] W. Shi, X. Wang, et al. Exercise gesture recognition and feedback with wearable sensors and deep learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–14. ACM, 2019.

[36] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*:5693–5703, 2019.

[37] P.-S. Wang et al. Global attention pooling networks for point cloud understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8632–8641, 2020.

[38] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.

[39] Y. Xiu, J. Li, H. Wang, H.-S. Fang, and C. Lu. PoseFlow: efficient online pose tracking. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[40] C.-H. Yang, Y.-C. Hsu, et al. Assessing physical activity: an automated classification of body postures and movements using wearable sensors. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5172–5175. IEEE, 2008.

[41] D. Yoo, M.-C. Kim, G. Jeong, M. Kang, and S.-W. Lee. Deep pose estimation for workout feedback. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1689–1694. IEEE, 2018.

[42] A. Zanfir, E. Oneata, I. Georgescu, A.-I. Popa, M. Leordeanu, and C. Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, pages 465–481. Springer, 2020.

[43] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*:4966–4975, 2015.

[44] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 398–407, 2017.