GoatAI

Alma Mater Studiorum
Università di Bologna

3D Human Shape and Pose Estimation from Multi-view Images

Master's Degree in Artificial Intelligence

Candidate

Alessandro Folloni

Supervisor

Prof. Samuele Salti

Co – Supervisor

Matteo Fabbri PhD.

GoatAI

Introduction

Fitness is a rapidly growing industry, and AI can play a key role by enabling personalized training, optimizing performance, and providing smarter health insights.

GoatAI

Motivations and Objectives

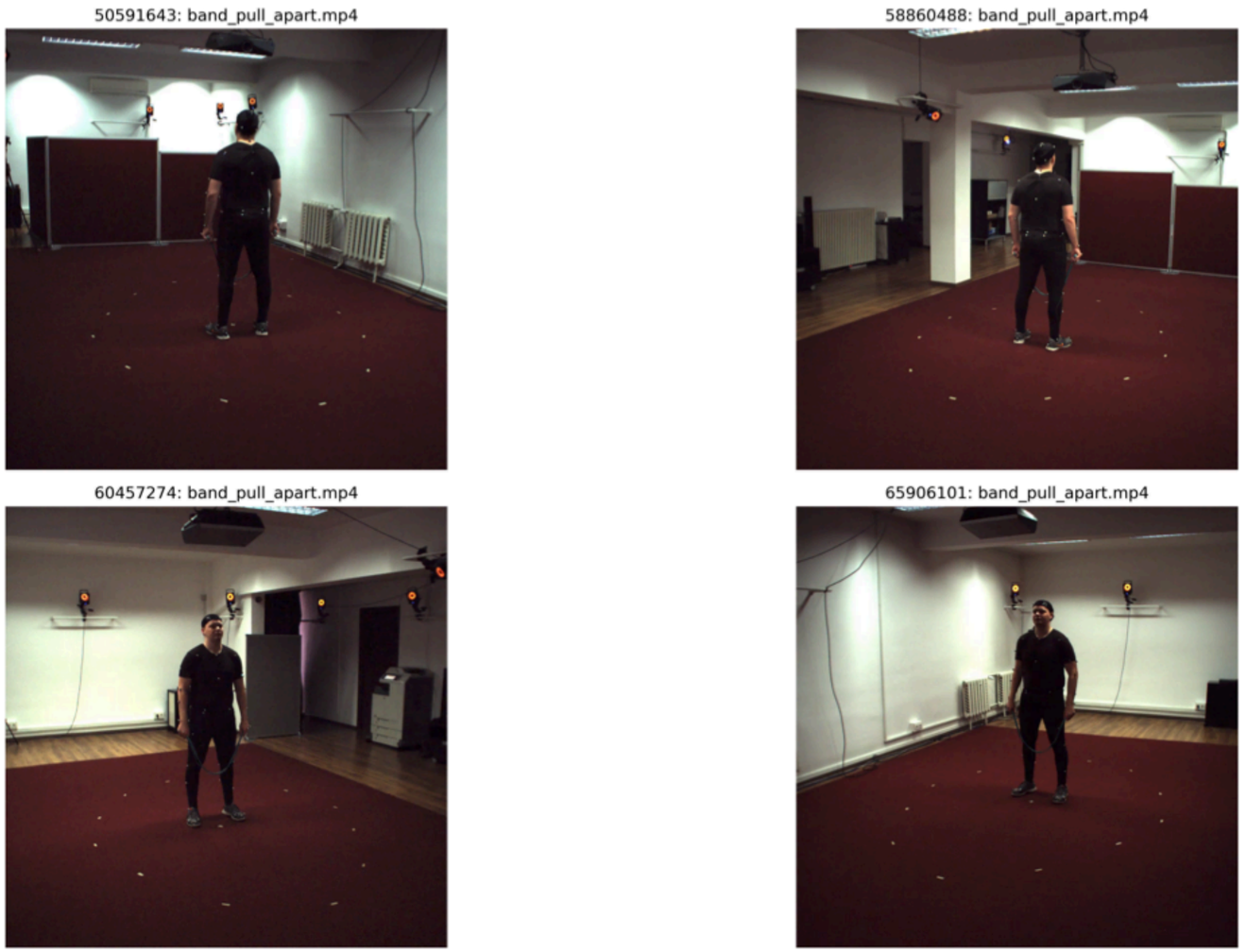
The final goal is to develop an AI Coach that observes your gym training and provides real-time guidance.

- We aim to create accurate 3D human reconstructions using multi-view images.
- Traditional methods often miss subtle details like slight movements, expressions, and hand gestures.
- Our pipeline fuses multi-view data to reconstruct 3D joints and estimate detailed SMPL-X parameters.

GoatAI

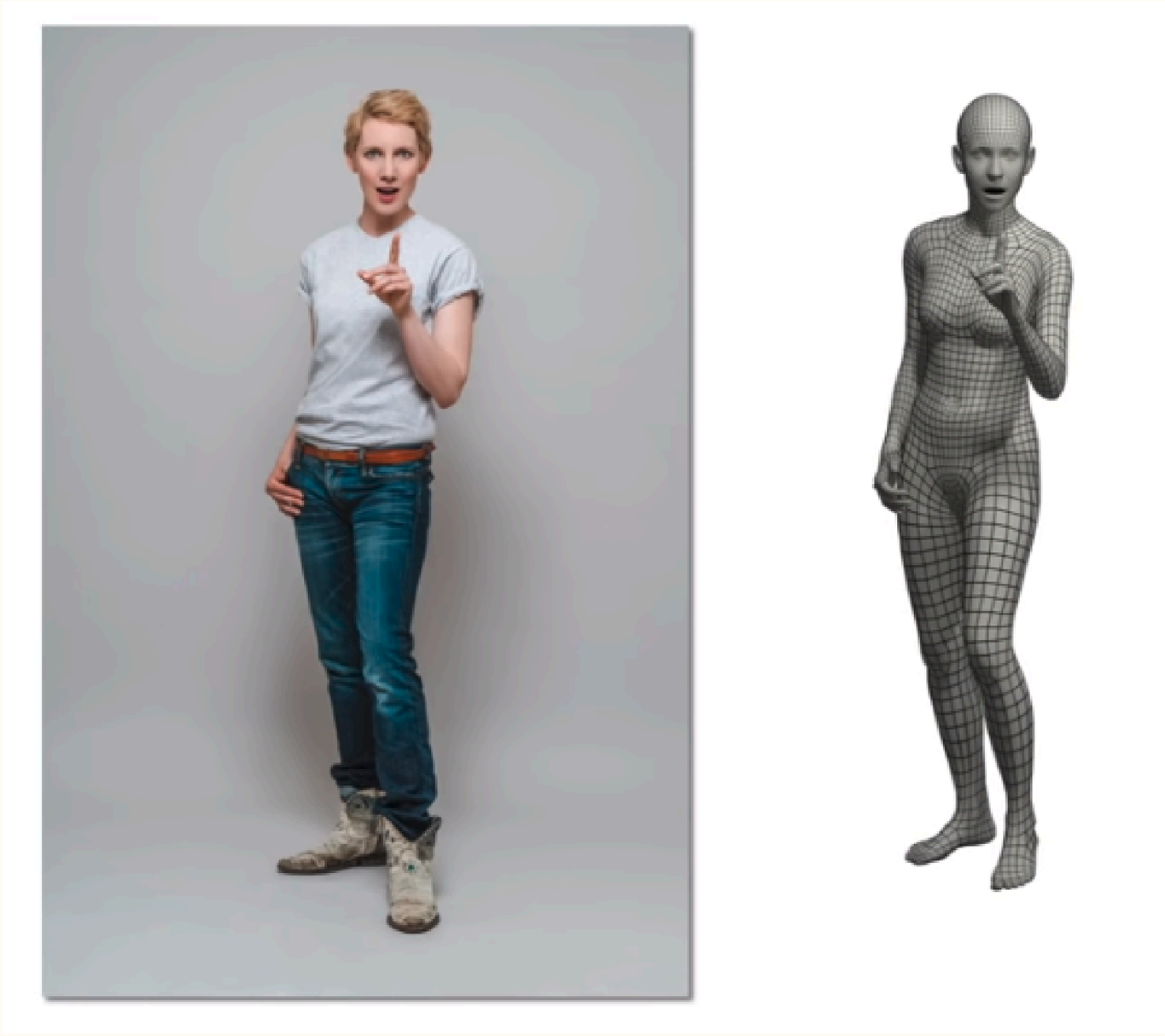
Setup

Our approach relies on multi-view imaging—capturing subjects from multiple angles to integrate diverse spatial information. This ensures robust and accurate 3D reconstructions.



GoatAI

High-level algorithm

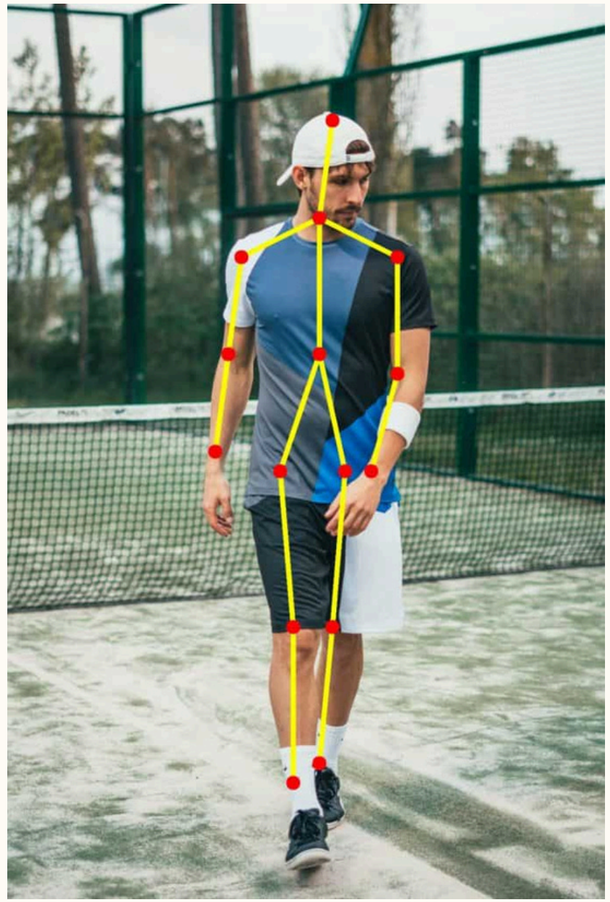


GoatAI

Some definitions

2D Keypoints: (x,y) coordinates of body landmarks.
2D Joints: those keypoints that mark body articulations.

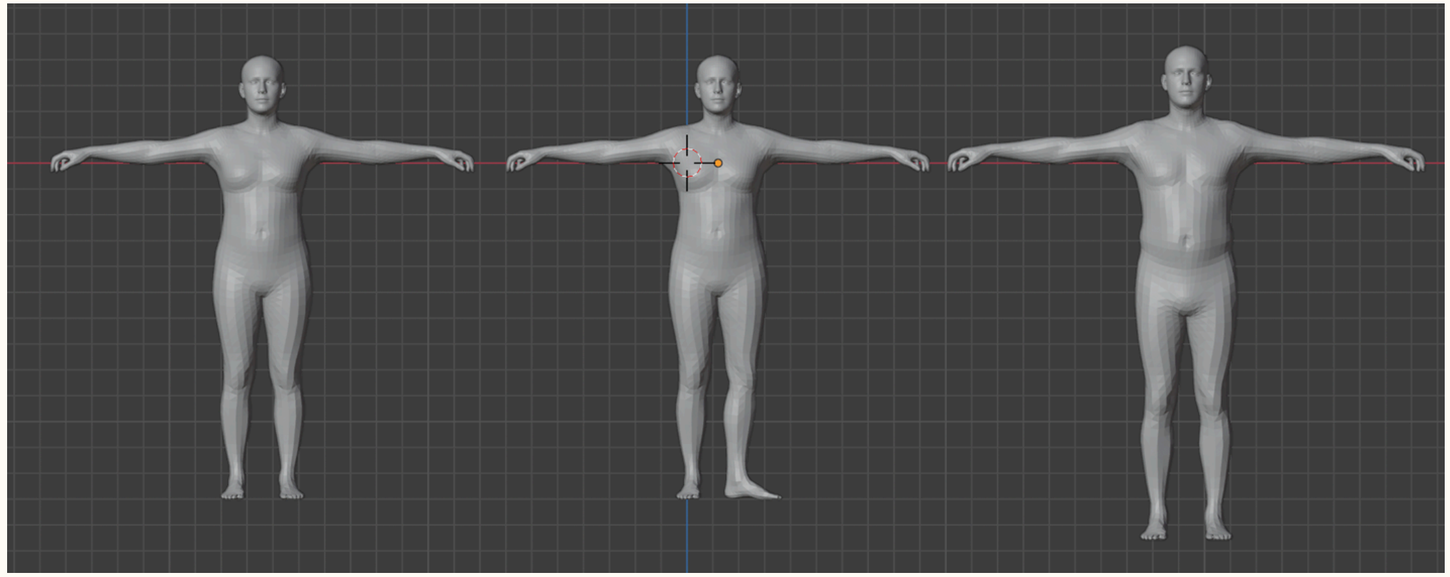
3D Keypoints: (x,y,z) coordinates of landmarks via multi-view triangulation.
3D Joints: those keypoints organized as a 3D skeletal structure.



GoatAI

Some definitions

SMPL-X models are parametric representations of the human body that capture shape, pose, and fine details like facial expressions and hand gestures, enabling realistic 3D reconstructions.

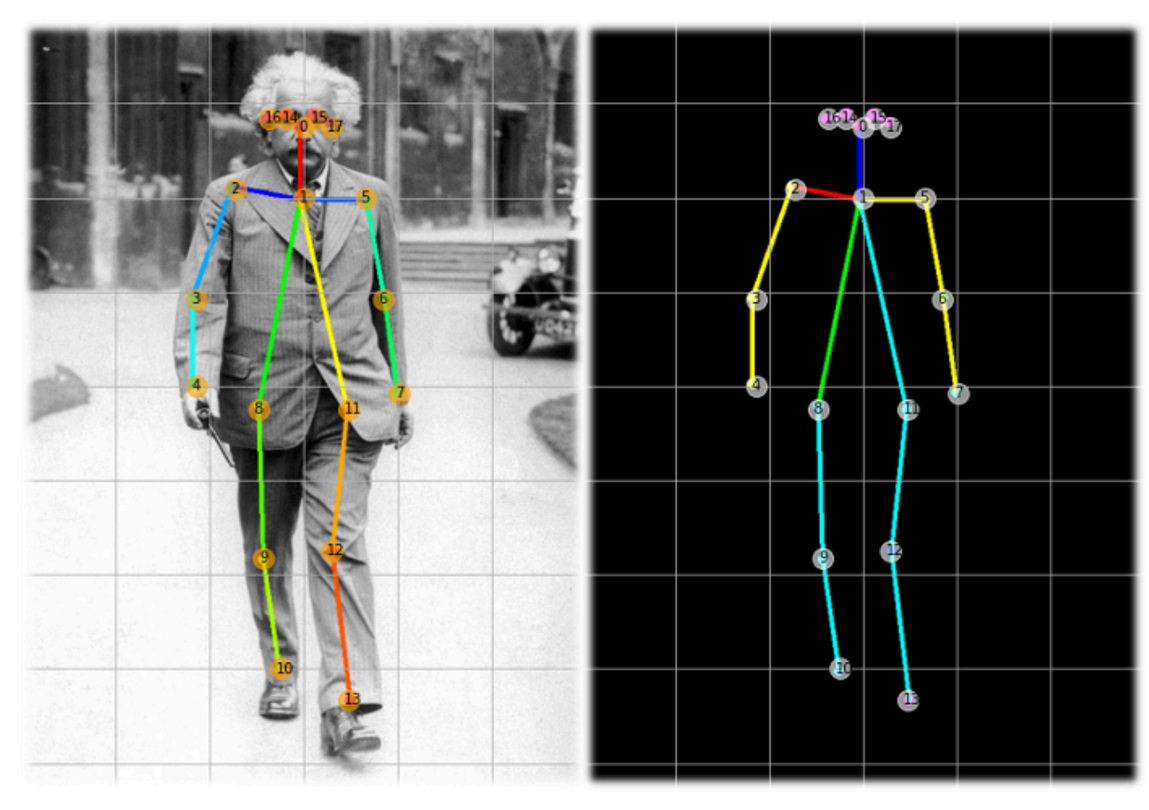


SMPLX mesh comparison: neutral body, pose variation (θ) and shape variation (β).
Bogo, F., El-Nouby, A., Black, M. J., Romero, J. “SMPL-X: A New 3D Human Body Model with Expressive Hands and Face.” arXiv:2006.09018 (2020).

GoatAI

Low-Level Process

2D Keypoint Extraction: We use YOLOv8 to extract 2D keypoints from each camera view. YOLOv8 processes each image in real-time, detecting and localizing essential body landmarks with high accuracy. These 2D keypoints serve as the foundational input for our subsequent 3D reconstruction stages.



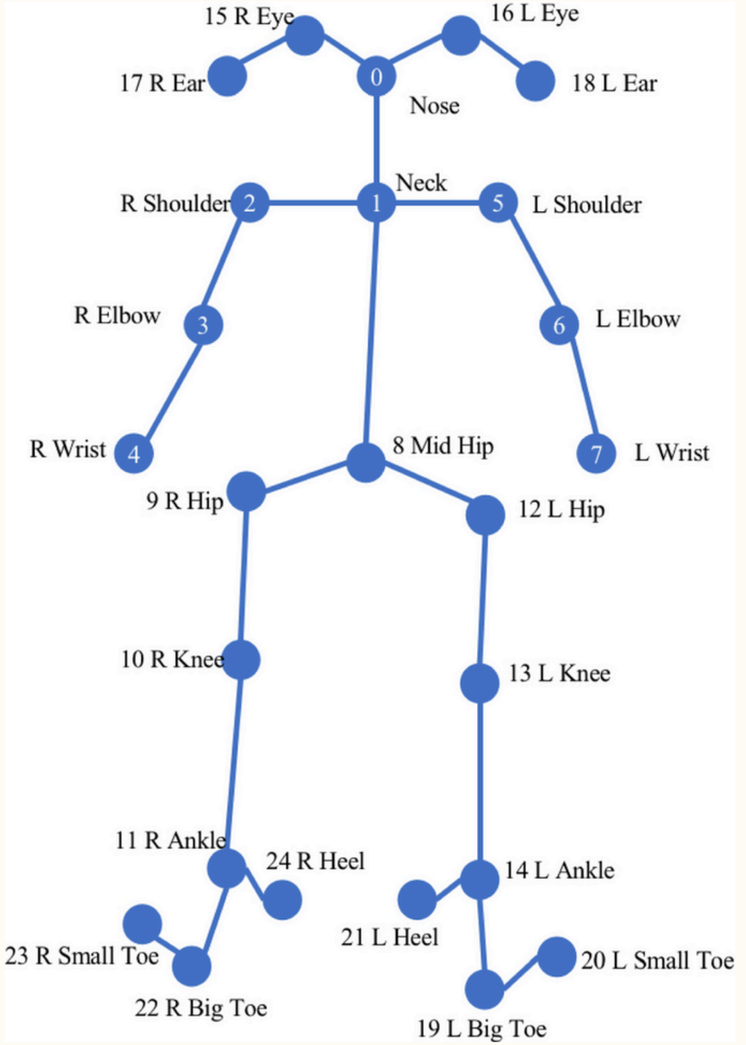
Input: single images/frames
Output: 17 x 2 parameters

GoatAI

Low-Level Process: 2D to 3D

In our 3D keypoint extraction stage, we tested multiple architectures—Transformers, RNNs, FCNNs, and CNNs—while tuning key hyperparameters such as model depth, number of layers, dropout, batch size, and learning rate.

Input: 4 x 17 x 2 parameters
Output: 25 x 3 parameters

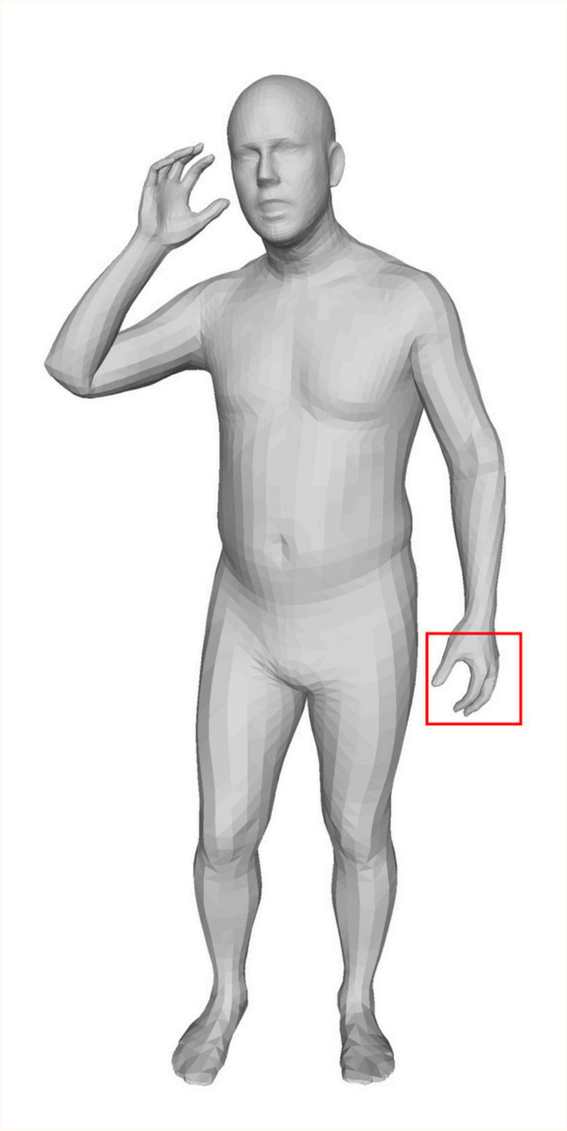


GoatAI

Low-Level Process: 3D to SMPL-X

The 3D pose is refined into detailed SMPL-X parameters using a Transformer-based model that leverages self-attention to capture global dependencies. This process yields accurate estimates of body shape, pose, facial expressions, and hand gestures.

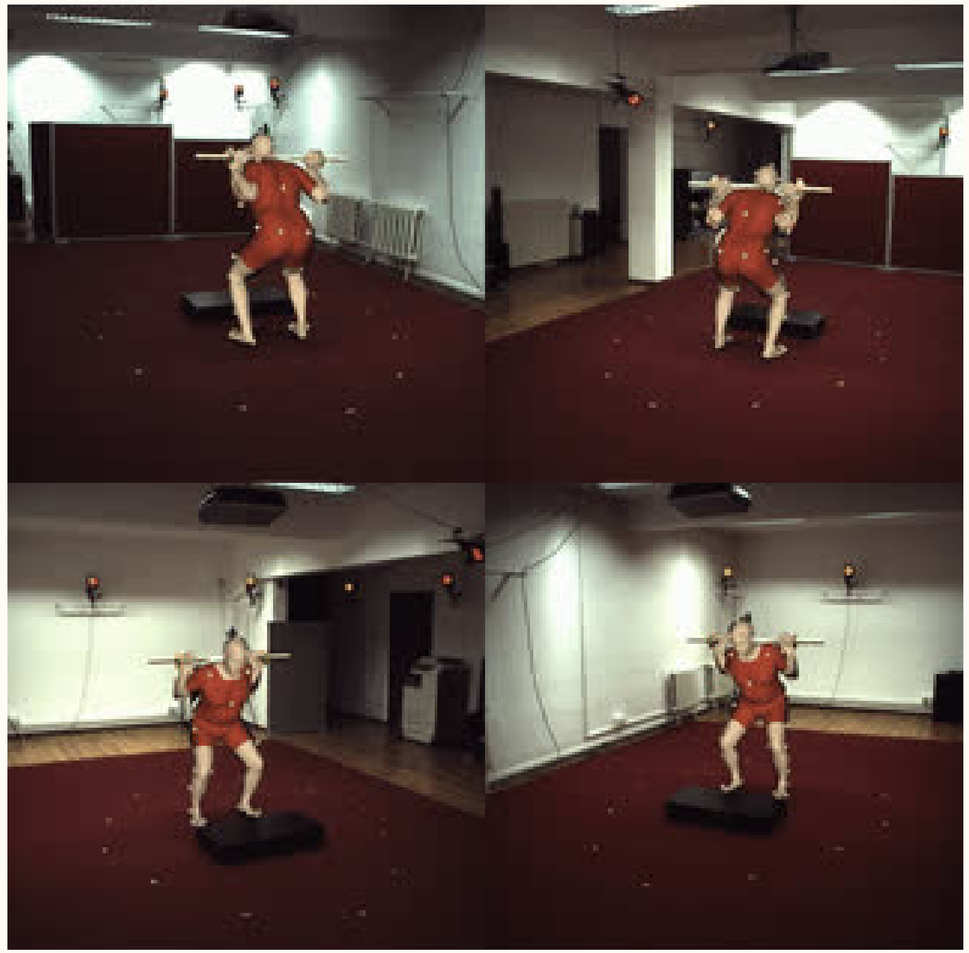
Input: 25 x 3 parameters
Output: 188 parameters



GoatAI

Dataset

We use the FIT3D dataset, a multi-view collection capturing gym exercises (e.g., squats, bench presses) from multiple angles.
This diverse data is crucial for training and evaluating our 3D reconstruction and SMPL-X models.



GoatAI

Quantitative Results: 2D to 3D

We evaluated several architectures (Transformers, RNNs, FCNNs, CNNs).

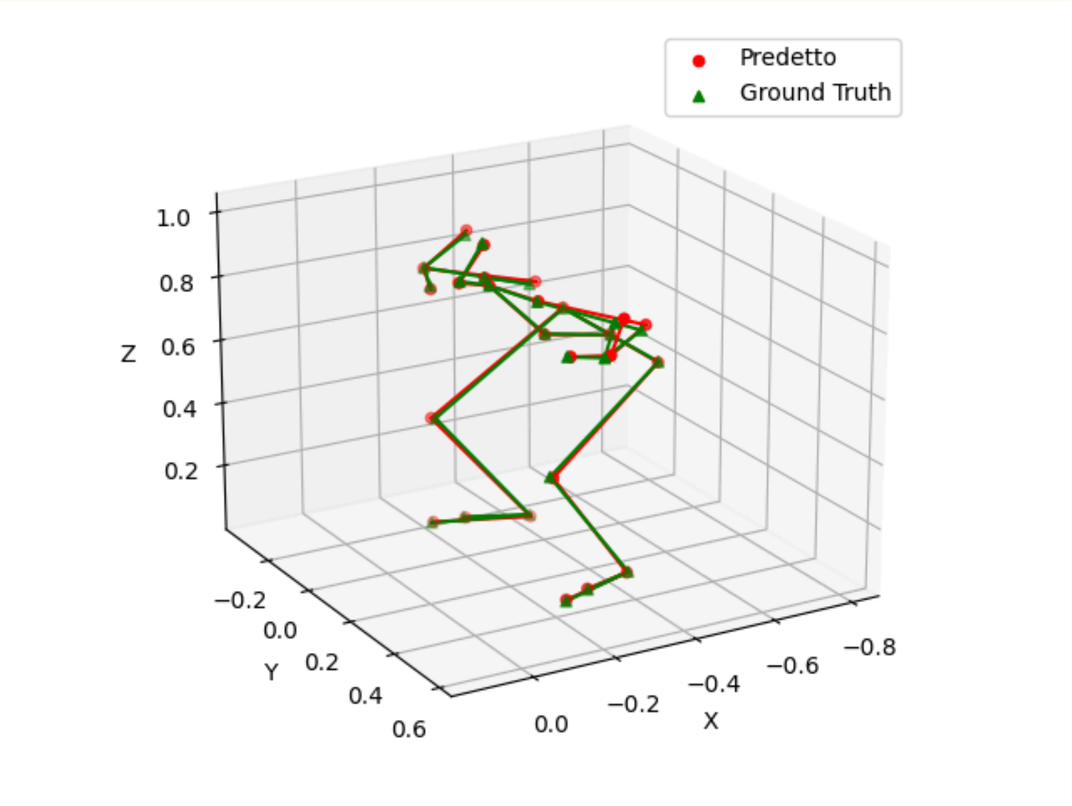
- Loss: MSE
- Metrics: MPJPE (average distance between predicted and ground truth joints)

ID	Architecture	BS	d/h	n	L	dp	Ep	LR	Test Loss	Test MPJPE
17	Transformer	16	256	2	12	0.2	200	5×10^{-5}	1.98×10^{-7}	$\mathbf{1.71 \times 10^{-5}}$
23	Transformer	16	512	8	6	0.5	20	1×10^{-4}	5.45×10^{-7}	3.37×10^{-5}
34	RNN	16	256	—	2	0.5	20	1×10^{-4}	5.67×10^{-7}	3.45×10^{-5}
33	RNN	16	256	—	1	0.5	10	1×10^{-3}	6.82×10^{-7}	3.60×10^{-5}
32	RNN	16	128	—	1	0.2	25	1×10^{-3}	6.41×10^{-7}	3.76×10^{-5}
24	Transformer	16	512	8	6	0.4	15	1×10^{-4}	6.92×10^{-7}	3.79×10^{-5}
14	Transformer	32	512	4	10	0.4	50	1×10^{-4}	7.90×10^{-7}	5.40×10^{-5}
13	Transformer	32	512	8	8	0.3	50	1×10^{-4}	8.16×10^{-7}	5.53×10^{-5}
9	Transformer	32	256	4	12	0.3	100	1×10^{-5}	8.87×10^{-7}	5.87×10^{-5}
18	Transformer	16	128	2	10	0.5	100	5×10^{-5}	2.54×10^{-6}	6.60×10^{-5}

GoatAI

Qualitative Results: 2D to 3D

Among the architectures we evaluated, the Transformer-based model stood out by achieving the lowest MPJPE.
Its self-attention mechanism effectively captures both spatial and temporal dependencies in the multi-view data, enabling a highly accurate reconstruction of 3D human poses.



GoatAI

Quantitative Results: 3D to SMPL-X

We evaluated a Transformer-based model for 3D-to-SMPLX regression.

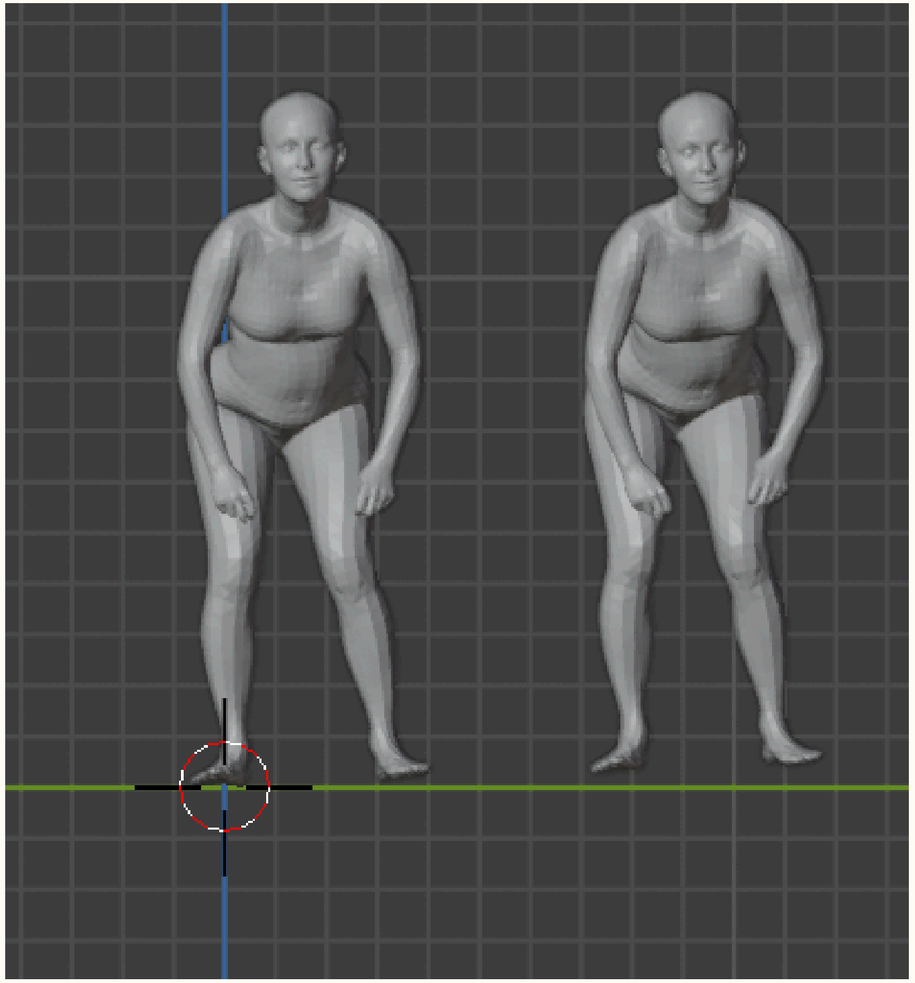
- Loss: MSE
- Metrics: MAE and RMSE

ID	BS	d	n	L	dp	Ep	LR	Test Loss	Test MAE	Test RMSE
41	16	128	2	2	0.1	100	1.0×10^{-5}	$\mathbf{5.39 \times 10^{-3}}$	5.86×10^{-3}	7.66×10^{-2}
35	24	256	8	3	0.15	65	1.0×10^{-4}	7.72×10^{-3}	2.67×10^{-2}	$\mathbf{2.73 \times 10^{-2}}$
26	16	256	8	3	0.1	60	1.0×10^{-4}	7.93×10^{-3}	2.54×10^{-2}	4.86×10^{-2}
40	48	256	4	4	0.2	150	1.0×10^{-4}	8.00×10^{-3}	$\mathbf{2.59 \times 10^{-3}}$	5.09×10^{-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}
23	48	128	4	6	0.3	80	1.5×10^{-4}	1.03×10^{-2}	3.20×10^{-2}	5.66×10^{-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}
24	32	512	8	5	0.25	75	1.0×10^{-4}	1.15×10^{-2}	2.77×10^{-2}	5.37×10^{-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}
31	16	256	4	7	0.35	100	5.0×10^{-4}	1.30×10^{-2}	6.08×10^{-2}	9.00×10^{-2}

GoatAI

Qualitative Results: 3D to SMPL-X

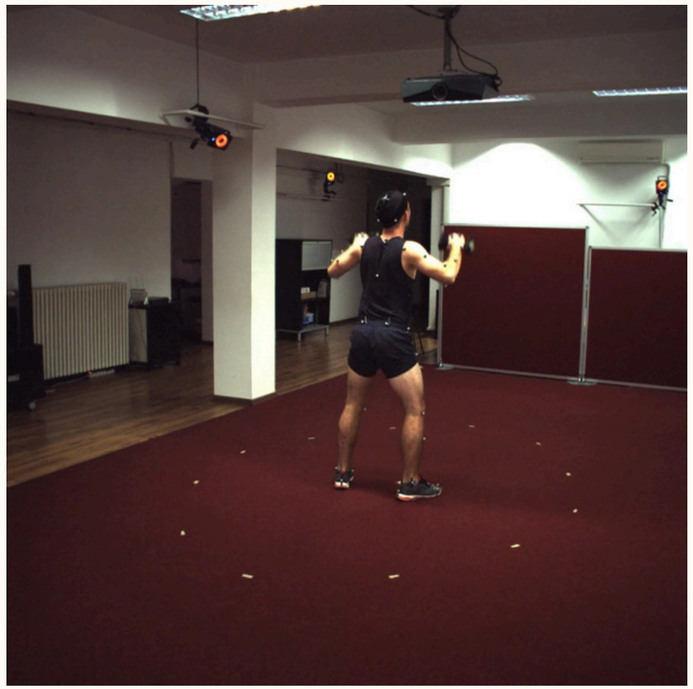
Among the evaluated configurations, one stood out by achieving the best overall performance. This configuration recorded the second lowest MAE while maintaining competitive RMSE, demonstrating its ability to precisely capture the complex, high-dimensional relationships among SMPL-X parameters.

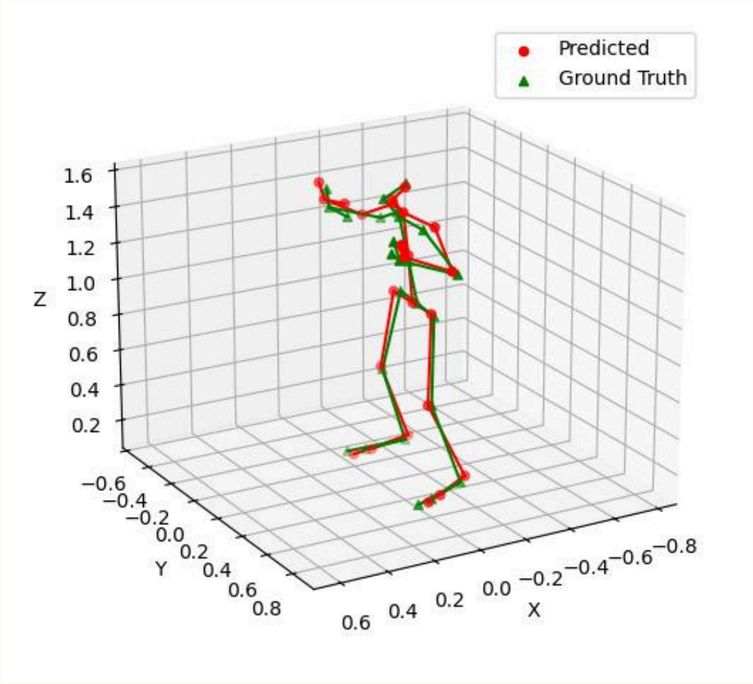


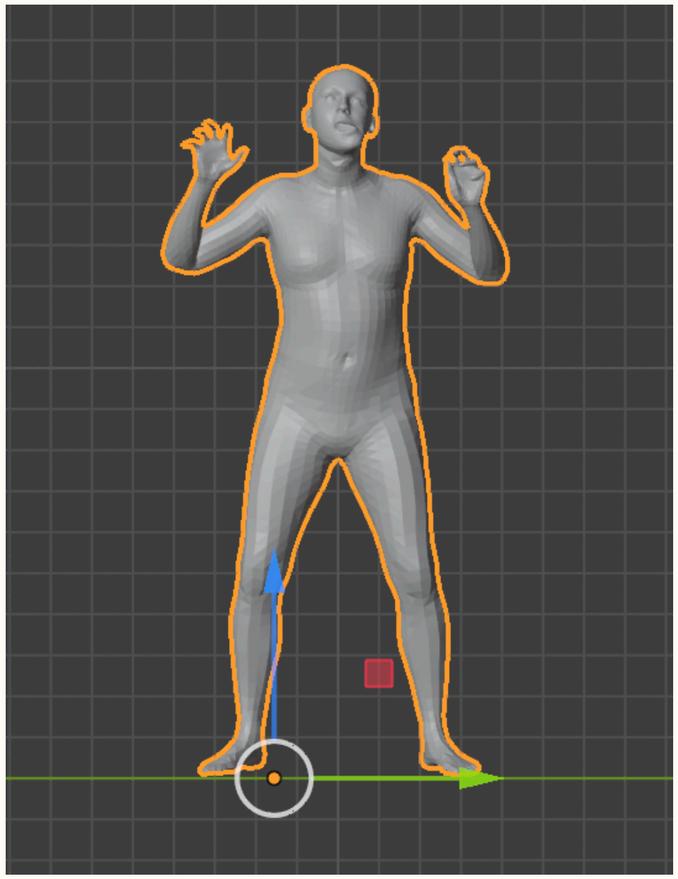
Left: ground-truth mesh; Right: corresponding model prediction.

GoatAI

Results: end-to-end pipeline







Despite some error accumulation across stages, the overall results remain highly promising.

GoatAI

Conclusions and Future Work

Our pipeline effectively integrates multi-view 2D keypoints extraction, 3D keypoints prediction, and SMPL-X parameter estimation, with the Transformer-based model demonstrating superior performance in both cases.
Although error accumulation between stages is evident, the overall results are promising.

Future work will focus on refining hand-specific parameters alongside broader model parameters, extracting a larger set of keypoints to enable higher-fidelity reconstruction, and integrating monocular approaches to deliver more generalizable, robust, and real-time capable results.

GoatAI

Alma Mater Studiorum
Università di Bologna

Thank you for your attention