SkelFormer: Markerless 3D Pose and Shape Estimation using Skeletal Transformers

Vandad Davoodnia[1,2], Saeed Ghorbani[2], Alexandre Messier[2], and Ali Etemad[1]

[1] Queen's University, Canada
[2] Ubisoft LaForge, Canada
{vandad.davoodnia,ali.etemad}@queensu.ca
{saeed.ghorbani,alexandre.messier}@ubisoft.com

Abstract. We introduce SkelFormer, a novel markerless motion capture pipeline for multi-view human pose and shape estimation. Our method first uses off-the-shelf 2D keypoint estimators, pre-trained on large-scale in-the-wild data, to obtain 3D joint positions. Next, we design a regression-based inverse-kinematic skeletal transformer that maps the joint positions to pose and shape representations from heavily noisy observations. This module integrates prior knowledge about pose space and infers the full pose state at runtime. Separating the 3D keypoint detection and inverse-kinematic problems, along with the expressive representations learned by our skeletal transformer, enhance the generalization of our method to unseen noisy data. We evaluate our method on three public datasets in both in-distribution and out-of-distribution settings and observe better performance than prior works. Moreover, ablation experiments demonstrate the impact of each of the modules of our architecture. Finally, we study the performance of our method in dealing with noise and heavy occlusions and find considerable robustness with respect to other solutions. Supplementary materials and demonstration video available at `https://vdavoodnia.github.io/projects/skelformer/`.

Keywords: Markerless Motion Capture · Multi-view Human Pose Estimation · Inverse-kinematics · Skeletal Transformers

1 Introduction

Motion capture is an active field of research with applications in sports, entertainment, health, and human-computer interaction. Currently, optical motion capture technology offers the most reliable and accurate solution by using a large number of cameras that detect markers attached to the actor's body. As a result, optical motion capture is costly and has a time-consuming setup process, preventing its practical application in low-budget or outdoor settings. In contrast, markerless optical motion capture offers a more convenient and portable solution for capturing the pose, generally at the cost of accuracy. Therefore, a significant amount of research has been dedicated to improving markerless motion capture in recent years, delivering high-quality animations by using only a few RGB cameras [1,2]. Yet, markerless approaches either take a long time to process, *e.g.*, 6 minutes for a 30-second video [3], or they struggle to perform well in in-the-wild environments [4]. Optimization-based solutions that fit a parametric model to the detected keypoints often exhibit long run-times, while regression models trained on controlled and in-studio datasets lack generalization due to low-diversity backgrounds, appearance, and lighting conditions [5].

In this paper, we propose a novel pipeline for markerless motion capture, which we name SkelFormer. At a high level, SkelFormer consists of two main modules: a 3D keypoint estimator and a skeletal transformer. First, to simplify 3D keypoint detection while maintaining generalizability to a wider distribution of scenarios, our method uses a Direct-Linear-Transformation (DLT) [6] triangulation method on the output of off-the-shelf 2D keypoint estimators trained on in-the-wild data. Next, we propose a skeletal transformer motivated by Inverse-Kinematics (IK) approaches to generate body pose and shape parameters rather than relying on the commonly used optimization methods. This module significantly reduces computational overhead while exhibiting more accurate performance. However, the misalignment between the estimated 3D keypoints and the body joint configuration of motion capture data makes the integration of keypoint estimators and our IK module challenging. To address this, we propose a simple joint regressor, trained on a small set of synthetic, and use it to generate synthetic keypoints from motion capture data that are aligned with 2D keypoint estimators. Lastly, we apply several augmentations on the acquired keypoints and train our IK component using the noisy data.

To rigorously test the performance of our method, we evaluate SkelFormer in both In-Distribution (InD) and Out-of-Distribution (OoD) settings against prior works while noting that most previous works have been tested in InD settings. Next, detailed ablation experiments demonstrate the impact of each component. Finally, we study the performance of our method and examine its robustness to highly noisy and occluded data.

In summary, we propose SkelFormer, a regression-based IK solution that converts 3D body keypoint positions to a full-body pose and shape. Our model bridges the gap between the most accurate 3D keypoint detection algorithms and human pose and body mesh estimation with negligible performance degradation. SkelFormer achieves robust results and outperforms others in InD scenarios. Next, we find strong cross-dataset generalizability through OoD evaluation on two unseen datasets, achieving competitive performance to InD multi-view solutions. Additionally, our method exhibits high robustness (less than half of the error of optimization-based solutions) in severely noisy and occluded scenarios.

2 Related Work

2.1 Keypoint Detection

The 2D keypoint estimation field has seen substantial progress in recent years. Generally, 2D keypoint estimation models are categorized into top-down and bottom-up approaches, each with their trade-offs in speed and accuracy [7–9]. Due to the availability of large-scale datasets, such as COCO WholeBody [10,11] and Halpe [12], 2D estimators have expanded into whole-body keypoints, potentially impacting 3D human pose and shape estimation. Previous works have proposed several strategies to infer the 3D keypoints of a subject, including semi-supervised learning [13,14], temporal [15,16], and multi-view [17,18] modelling. PoseBert [15] and volumetric Learnable Triangulation (LT) [17] are notable examples of temporal and multi-view methods, respectively, reporting 3D keypoint estimation with an error of below pixel-level accuracy. Our method leverages the advances in 3D keypoint estimation by using off-the-shelf models.

2.2 Pose and Shape Estimation

Regression-based methods generally predict the parameters of a body model, *e.g.*, SMPL [19,20] represented by body shape and pose components. The research on body pose and shape regression can be categorized into single-view and multi-view problems. Single-view approaches generally suffer from the inherent 2D image to 3D pose ambiguities, resulting in worse performance compared to multi-view methods. For instance, Pose2Mesh [21] uses a GraphCNN, consisting of a mesh coarsening encoder-decoder architecture, to regress the human body and shape from a single image. Similarly, GTRS [22] proposes a lightweight graph-based transformer network to uplift 2D keypoints to 3D pose and shape parameters. PyMAF [23] also explores visual encoder feature maps for pose regression. We design a multi-view markerless motion capture pipeline by taking inspiration from the advances in single-view research [1,22].

In the context of markerless motion capture via multiple views, the majority of methods are supervised, utilizing strategies such as collaborative learning [24], volumetric feature aggregation [25], multi-view feature fusion via attention [2], and pixel-aligned feedback fusion [4]. Since regression models typically rely heavily on the availability of annotated data and the diversity of postures in the training data, they tend to be limited to in-studio quality and do not perform well on OoD evaluations with background and appearance shifts [4]. Although a common solution to this problem is to pre-train the network on in-the-wild datasets, it does not guarantee better generalizability as the neural networks are susceptible to over-fitting and catastrophic forgetting. This is also evident by the best estimation error of 93 *mm* on the in-the-wild Ski-Pose [26] dataset reported in a recent work [2], which is three times bigger than their 33 *mm* error on the Human3.6m [5] dataset. We address the OoD generalization limitation in multi-view setups by incorporating prior knowledge of human pose into the solution in an IK solver by training on a large set of motion capture data.

Optimization-based approaches fit the parameter of the SMPL model to features extracted from an image, such as 2D/3D keypoints and silhouettes [3]. Simplify-x [27] introduced VPoser, a human variational pose prior trained on a large collection of motion capture data, to reduce the complexity of the optimization space. Subsequently, the majority of recent works rely on VPoser to fit the SMPL model to 3D predictions, which is a time-consuming process that can take up to a day to process a one-hour-long video [3, 28]. Furthermore, variational Gaussian models are prone to mean collapse due to the prior distribution assumption. Additionally, as the optimization process is sensitive to initialization, their potential to fit noisy data accurately is hindered [29]. To address this challenge, several works like ProHMR [30] have proposed using regression-based models to initialize the optimization variables. Although this approach can speed up and improve accuracy, optimization algorithms remain susceptible to reaching undesirable local minima, especially on noisy, exotic, or unseen poses as discussed in [29]. In summary, though capable of modelling complex movements and interactions, optimization-based models often require careful hyper-parameter tuning for each recording sequence, making them difficult to use in the face of multiple constraints and impractical for fast-paced production. We show that compared to optimization solutions, our skeletal transformer performs more accurately with better noise and occlusion robustness.

2.3 Inverse-kinematics Models

IK is the task of obtaining body joint rotations given several pose constraints, with applications in robotics [31] and animation [32]. In the context of human pose and shape estimation, HybrIK [1] proposed a hybrid analytical-neural IK solution that obtains the SMPL body rotations given the 3D keypoints estimated from monocular images. They designed their model to disambiguate the 2D image to 3D pose estimation by considering the shape of the human body and breaking the joint rotations to their swing and twist components. However, to our knowledge, IK applications of neural networks have not been explored for multi-view pose and shape estimation. This may be due to the superior performance of optimization methods, such as VPoser [27], yet at a high computational cost.

2.4 Skeletal Neural Networks

Previous research has reported superior performance for human motion modelling [33], 3D keypoint refinement [34], and 2D to 3D uplifting [15] using skeletal neural networks. In these approaches, the human body structure is modelled as a skeletal graph, and the neural networks exploit the graph structure by learning local and global pose features. Similar techniques have been proposed for hand pose estimation, where the margins for error are much lower than human body pose estimation [35]. Motivated by the recent success of transformers in several fields, such as natural language processing [36] and computer vision [37], they have been used for motion inbetweening and completion [38,39] of human poses, achieving high-quality results. We design a transformer model to capture the full pose state via contextualized latent representations from 3D joint positions.

3 Methodology

Overview. As illustrated in Fig. 1, our pipeline starts with a 3D keypoint estimator consisting of different sub-modules for tracking, 2D keypoint estimation,

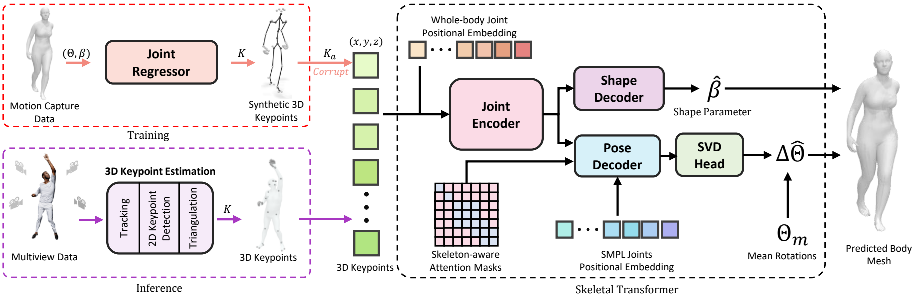


Fig. 1: An overview of the proposed skeletal transformer pipeline is demonstrated. During training, noisy 3D keypoints are generated using our joint regressor, while during inference, 3D keypoint are provided by off-the-shelf models. Then, our proposed skeletal transformer maps the keypoints onto the SMPL pose and shape parameters.

and triangulation, for which we use off-the-shelf models. Our proposed skeletal transformer then maps the estimated 3D keypoints onto the SMPL pose and shape parameters. The details of each part are given below.

3.1 3D Keypoint Estimation

Human Tracking. We employ Faster R-CNN [40], a well-established object detection model, to track the subjects in the input frames. Although more advanced methods, such as 3D skeleton tracking modules for crowded scenes [41], can be used, we did not observe any misidentification during single-person experiments.
2D Pose Estimation. To estimate the 2D joints, we employ HRNet-W48+Dark [7, 9] trained on COCO WholeBody dataset [11]. Since this model is trained on in-the-wild datasets, it helps with the generalizability of our pipeline.
Triangulation. We choose a simple triangulation method by employing DLT [6] on 2D keypoints given the extrinsic camera parameters. For this purpose, we consider 2D detection scores for assigning point occlusions.

3.2 Skeletal Transformer

Traditional IK solvers often assume noise-free constraints, which is not always the case for observations, *e.g.*, 3D keypoints in markerless motion capture. As a result, iterative IK solvers generally perform better than regression models in reaching the local minima given the ground-truth joints, at the cost of additional computations. However, noise and occlusions can cause iterative IK solvers to reach sub-optimal solutions. To address this issue and speed up the process, we introduce an end-to-end learnable pose reconstruction model as illustrated in Fig. 1. Following, we present the detailed components of our skeletal transformer.
Joint Encoder. As depicted in Fig. 1, our network consists of a 3D joint encoder followed by pose and shape decoders. To account for the order of the joints

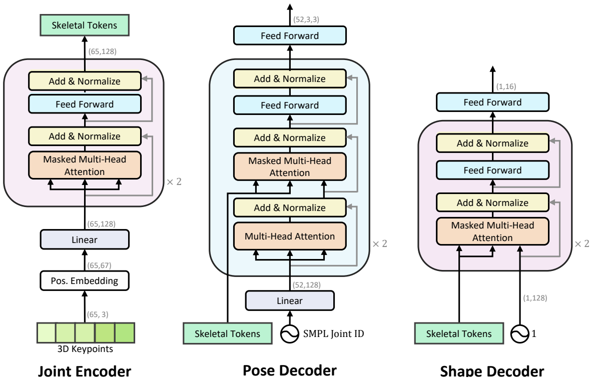


Fig. 2: Detailed architectures of our modules are presented.

in a skeleton, we first inject information about their ordering by concatenating a positional embedding (*i.e.*, joint ID embedding) vector to each joint input [42]. The results are then passed through an embedding layer to match the dimensionality of the transformer's hidden layers. Next, we pass the embedded data through a series of transformer encoder blocks consisting of self-attention and feed-forward layers, followed by layer normalization, to obtain the skeletal tokens (see Fig. 2). To model joint occlusions, we mask out corresponding connections, encouraging the network to use contextual information embedded within the rest of the joints. The result is a latent representation of the joint positions, which is passed to the following decoders.

Pose Decoder. Our pose decoder estimates the body pose given the latent representations of joint positions. A positional embedding is fed into the decoder to relay information about all 52 SMPL+H skeleton joints. Then, the skeletal tokens and occlusion masks are passed through the multi-head attention layers in the decoder. Specifically, the decoder consists of several blocks of self-attention, multi-head attention, and feed-forward layers followed by layer normalization (see Fig. 2). Additionally, to block the unwanted correlations between far-away joints (*e.g.*, left and right hands) that existed in the training set, we set the attention weights such that each joint only attends to the other joints that are in a distance of less than 4 nodes in the kinematic tree.

Symmetric Orthogonalization. The next step is to obtain the joint rotations from the decoder's output. Although previous works [27, 30, 43] have suggested using 6-DoF representations of rotation matrices, we find that SVD symmetric orthogonalization proposed in a recent work [44] yields more accurate results and converges faster. Therefore, the output of the pose decoder is passed through a fully connected residual layer that outputs a square matrix $M_{3\times3}$ with SVD of $U\Sigma V^T$. Then, its symmetric orthogonalization $\Theta \in SO(3)$ is obtained by:

$$\Theta = U\Sigma_o V^T, where\ \Sigma_o = diag(1, 1, det(UV^T)). \tag{1}$$

Shape Decoder. To obtain the shape parameters, we design a shape decoder with a similar architecture to our pose decoder. Since the order of joints does not affect the shape decoder, we remove the self-attention layers and set the sequence length of the shape decoder to one, effectively reducing it to a feed-forward network with multi-head attention (see Fig. 2).

3.3 Joint Regressor

Our goal is to train our skeletal transformer on a large collection of motion capture data consisting of samples represented by SMPL shape and pose parameters. However, the skeleton configuration extracted from whole-body keypoint detectors is not aligned with the SMPL joints. To solve the alignment issue, we use a joint regressor to convert SMPL representations to our desired skeletal configurations. Currently available joint regressors have been reported to be either inaccurate [45], or simply obtained by selecting vertices on the *surface* of the body, which is not biomechanically correct nor in accordance with existing body models. Therefore, we propose a novel joint regressor and training scheme using a small amount of synthetic data to align the SMPL with the 3D keypoints.

The joint regressor, defined as a linear layer $K = \mathcal{J}V$, is trained to map the body mesh vertices $V \in \mathbb{R}^{6890 \times 3}$ to 3D keypoints $K \in \mathbb{R}^{J \times 3}$, where J is the number of joints in the 3D keypoint configuration. In order to train the joint regressor, we randomly take 10,000 SMPL body samples from the AMASS dataset [46] and render them from four orthogonal views after removing the root translation and adding random root rotations augmentation. Next, following the process of Sec. 3, we use HRNet-W48+Dark [7, 9] to estimate whole-body keypoints followed by DLT triangulation for obtaining 3D keypoints.

To encourage sparsity and avoid out-of-body predictions, previous works [45] have suggested using an L_2 regularization on the joint regressor with the goal of (a) enforcing a sum of 1 for vertex weights of each 3D joint; and (b) encouraging all of the weights of the joint regressor to lie between 0 and 1. However, doing so creates a trade-off between regularization and accuracy. To solve this issue, we apply a temperature-scaled Softmax function over the trainable parameters of the joint regressor ϕ, thus automatically satisfying both constraints. Doing so also gives us control over the vertex sparsity for each of the 3D joints. The joint regressor's weights for the i_{th} keypoint are computed as:

$$\mathcal{J}_i(\phi) = \frac{e^{\phi_i / T}}{\sum_{j=1}^{J} e^{\phi_j / T}}, \tag{2}$$

where T controls the sharpness of the distribution of the vertex weights. We use an L-BFGS optimizer [47] to increase the training efficiency. We empirically set the temperature to $T = 10$ so that 3 to 10 vertices contribute to each joint.

3.4 Data Preparation

We extract pairs of 3D keypoints K (using our joint regressor), body joint rotation matrices $\Theta \in \mathbb{R}^{52 \times 3 \times 3}$, and shape parameters $\beta \in \mathbb{R}^{16}$ from human motion capture data. Next, we measure the average joint rotations across the dataset to normalize Θ. As some layers in the skeletal transformer share weights across all the joints, the symmetry of the left and right rotations is important. Yet, the mean pose Θ_m provided and used in previous works [30,43,48] is not symmetrical. To address this, we add the mediolateral mirrored samples to the dataset and use quaternion averaging as described in [49] to measure Θ_m, subsequently normalizing the pose by $\Delta\Theta = \Theta_m^{-1}\Theta$. During training, we apply a range of online augmentations to diversify the data and enhance our model's robustness. The augmentations include joint occlusions, body rotation, keypoint noise addition, shape augmentation, outlier addition, and mediolateral mirroring.

3.5 Training

To train our skeletal transformer, we feed the model with augmented data K_a to obtain pose and shape parameters, which are then passed through the SMPL Forward Kinematics (FK) layer. We then use a combination of rotational, positional, and shape losses to leverage different scopes and granularities to help with training [50]. Our final loss is calculated as the sum of the following losses.
Rotation Loss. Our rotational loss is the sum of the global and local geodesic distances between the predicted $\hat{\Theta}$ and ground-truth rotations Θ of every joint:

$$L_R(\Theta, \hat{\Theta}) = \arccos\left(\frac{\mathrm{tr}(\Theta\hat{\Theta}^{\intercal}) - 1}{2}\right). \tag{3}$$

Position Loss. Our model uses the SMPL FK layer to obtain the keypoints $\hat{K}$ and vertices $\hat{V}$. We then define a positional loss using $L_P = L_{1;s}(\hat{K}, K) + L_{1;s}(\hat{V}, V)$, where the $L_{1;s}$ represents a smoothed L_1 loss [51]. This loss can be seen as a combination of L_1 and L_2 distances, which is less susceptible to outliers than L_2, and it has less near-zero penalty than the L_1 distance.
Shape Loss. Finally, we minimize the $L_S = L_2(\hat{\beta}, \beta)$ distance between the estimated $\hat{\beta}$ and ground-truth shape parameters β to train our shape decoder.

4 Experiments

4.1 Datasets

AMASS. The Archive of Motion Capture as Surface Shapes (AMASS) [46] is a collection of 3D human pose and shape information collected from multiple motion capture databases. It contains over 40 hours of motion capture data from more than 300 subjects and spans over 11,000 actions. We follow the standard train, test, and validation splits used in prior works [3,27]. This dataset is used solely to train the skeletal transformer IK solver.
Human3.6m. Human3.6m [5] is the standard benchmark for evaluating 3D human pose, shape, and body estimation in multi-view and single-view approaches. Following previous works [52], we perform our evaluations using Protocol-I,

where the root-centred MPJPE and PA-MPJPE on subjects 9 and 11 are measured. We use this dataset for ***InD*** evaluation of our method.
RICH. Real scenes, Interaction, Contact and Humans (RICH) dataset [53] is a recently published dataset of multi-view videos with accurate markerless motion-captured bodies and scenes. The test set contains one withheld scene and 7 unseen subjects in 52 scenarios, captured using four cameras. We use this dataset for ***OoD*** evaluation of our method in outdoor settings.
MPI-INF-3DHP. The Max Planck Institute for Informatics 3D Human Pose dataset (MPI-INF-3DHP) [54] is a collection of over 1.5 million frames captured from eight angles, featuring eight actors performing various actions like sitting, dancing, and exercising. We follow the previous works by evaluating our model on subject 8 of the training set [25,55]. We use this dataset for ***OoD*** evaluations.

4.2 Evaluation Metrics

To evaluate the performance of our method, we employ standard evaluation metrics in 3D pose estimation literature. Mean-Per-Joint-Position-Error (**MPJPE**) measures the Euclidean distance between the estimated joint positions and the ground-truth joint positions, averaged over all joints in the skeleton. **PA-MPJPE** is an extension of MPJPE, where a rigid alignment between the estimated and ground-truth keypoints is applied prior to error measurement. This metric shows how well the skeleton is estimated, regardless of scaling and rotation. Additionally, we report **MPVPE** and **PA-MPVPE** to show the error between ground-truth and predicted vertices of body mesh. Additionally, we report **AUC** and **PCK** at a threshold of 150 *mm* according to MPI-INF-3DHP [54] evaluation criteria. Finally, **Rotation Error** is measured by the geodesic distance between the ground-truth and predicted poses.

4.3 Implementation Details

Hyper-parameters. We select 65 joints from the 133 joints of the whole-body skeleton configuration, excluding most facial landmarks while keeping the eyes, ears, and nose. We choose two transformer blocks for the encoder and each of the decoders. As illustrated in Fig. 2, We use a positional embedding of size 64 in the encoder and decoder while setting hidden layer dimensions to 128 within all layers. The shape and pose decoder heads use 1024-dimensional residual layers.
Optimization. We train our model using AdamW optimizer [56] with a batch size of 1024 on an NVIDIA A4000 GPU. We chose a learning rate of 1e-3, which is warmed up with a factor of 1e-4 for the first 2000 iterations and gradually reduced with a cosine annealing scheduler over 50000 iterations until it reaches 1e-7. The training of the network takes less than 18 hours to complete.
Mirror Test. Similar to the flipping test performed in 2D keypoint estimation methods [7,37], we perform a mirroring test during inference to reduce the model's biases towards left and right body parts.
Computation Cost. Our model contains 6.631 M parameters with 159.482 M Floating Point Operations (FLOPs) for a single input. As a result, our skeletal

Table 1: The comparison of our method in InD settings against prior multi-view works on the full test set of the Human3.6m dataset. * denotes the results from using ground-truth 3D keypoints as input

Method	**MPJPE↓**	**PA-MPJPE↓**	**Output**
CPN+DLT [57]	32.1	27.8	Joint Pos. Only
LT [17]	20.7	17.0	Joint Pos. Only
Pose2Mesh [21]*	29.0	23.0	Mesh Only
Huang *et al.* [58]	58.2	47.1	Joint Rot.+Mesh
Shin and Halilaj (SPIN^{4,cal}) [25]	49.8	35.4	Joint Rot.+Mesh
Shin and Halilaj (main) [25]	46.9	32.5	Joint Rot.+Mesh
Gong *et al.* [59]	53.8	42.4	Joint Rot.+Mesh
Jiang *et al.* [60]	50.2	37.3	Joint Rot.+Mesh
Jia *et al.* [4]	33.0	26.9	Joint Rot.+Mesh
SMPLify-X (LT) [27]	26.3	21.2	Joint Rot.+Mesh
SkelFormer (CPN)	33.5	27.8	Joint Rot.+Mesh
SkelFormer (LT)	**25.2**	**20.6**	Joint Rot.+Mesh

Table 2: Comparison of our method in OoD settings against prior works.

Method	**MPVPE↓**	**MPJPE↓**	**PA-MPJPE↓**	**PCK↑**	**AUC↑**	**OoD**
			RICH Dataset			
METRO [61]	134.5	129.6	-	-	-	✓
METRO [61]	107.9	98.8	-	-	-	✗
SA-HMR [62]	103.0	93.9	-	-	-	✗
IPMAN-R [63]	89.9	79.0	47.6	-	-	✗
SPIN [43]	129.5	112.2	71.5	-	-	✓
PARE [64]	125.0	107.0	73.1	-	-	✓
CLIFF [65]	122.3	107.0	67.2	-	-	✓
SkelFormer (HRNet)	**39.9**	**44.2**	**35.6**	-	-	✓
			MPI-INF-3DHP Dataset			
Liang and Lin [55]	-	-	59.0	95.0	65.0	✗
Shin and Halilaj [25]	-	-	50.2	97.4	65.5	✗
Jia *et al.* [4]	-	-	**48.4**	**98.6**	67.3	✗
SkelFormer (HRNet)	-	-	54.8	97.5	**67.4**	✓

transformer can solve the IK problem in 66 *ms* for a batch size of 512 using approximately 4G of GPU memory. Consequently, our SkelFormer pipeline takes 274 *ms* to predict body pose and shape parameters from a four-camera frame.

4.4 Results

InD Testing. To evaluate SkelFormer, we first compare its performance to prior works on the Human3.6m [5] dataset. In this experiment, most benchmarks pre-train their models on multiple datasets and fine-tune them on the Human3.6m training set. Accordingly, we report the performance of our model using keypoint estimators, *i.e.*, CPN [57] and LT [17], trained on Human3.6m dataset (InD). To this end, we use 3D predictions from CPN (followed by DLT triangulation [6]) and LT, both of which are trained following the standard evaluation protocol [5]. Moreover, we train our skeletal transformer on the AMASS dataset [46] using a 17-joint configuration as per Human3.6M. Delving deep into the results in Tab. 1, we observe that our method outperforms the best regression model [4]. Additionally, SkelFormer outperforms optimization solutions from multi-view 2D keypoints, 3D keypoint [58], and SMPLify-X (LT) [27]. A notable solution is

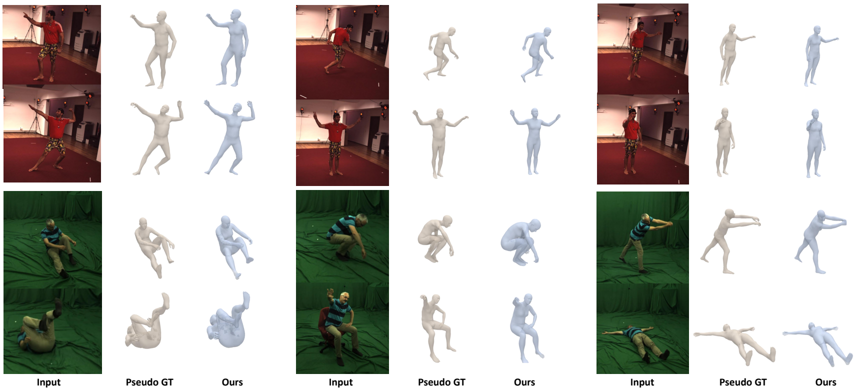


Fig. 3: A visual comparison with the pseudo-ground-truth from [52] is provided, presenting the realism and accuracy of our SkelFormer.

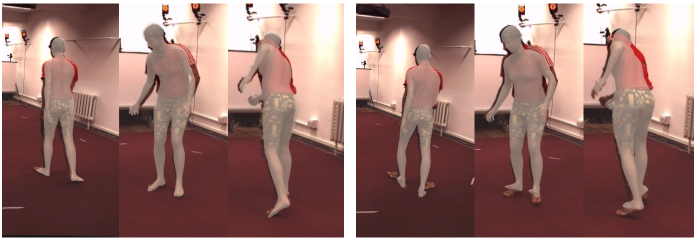
Pose2Mesh [21], another regression-based approach that exploits the 3D ground truth as input. However, SkelFormer outperform Pose2Mesh despite not using the ground-truth information. Finally, although not directly comparable, our method achieves the closest performance to solutions that only predict joint positions, namely, LT [17] and CPN+DLT [57].

OoD Testing. Next, to test our method in OoD settings, we use HRNet-W48+Dark [7, 9] as the keypoint extractor, which has not been trained on the RICH [53] or the MPI-INF-3DHP [54] datasets. Moreover, we keep our entire pipeline frozen and do not fine-tune any of its components on any portion of these datasets. The results are presented in Tab. 2. It should be noted that while prior works on the RICH dataset are monocular pose estimation approaches, all prior works on MPI-INF-3DHP are multi-view solutions and use all available views. On the RICH dataset, Tab. 2 shows that other solutions, such as SPIN [43] and CLIFF [65], suffer greatly in OoD setups (obtaining 172% and 127% additional error w.r.t. their InD performance on Human3.6m dataset). In contrast, our method shows competitive results compared to InD solutions by outperforming prior works on RICH while showing relatively minor degradation on the MPI-INF-3DHP dataset. Finally, we demonstrate the fitting quality of our model compared to the pseudo ground truth [52] on the Human3.6m and the MPI-INF-3DHP datasets in Fig. 3, showing improvements in the feet and hands.

Ablation Study. We test the importance of different network components and report the results in Tab. 3. Given our goal of increasing generalizability in the presence of noise and occlusions, we conduct experiments on motion capture data from the AMASS [46] testing set at the presence of 20% occlusion and additive Gaussian noise with $\sigma = 20$ *mm*. First, we demonstrate the effectiveness of symmetric orthogonalization by replacing our SVD operation with the commonly used 6-DoF representation, showing a drop of 0.8 *mm* and 0.19° of MPVPE and rotational error when SVD is removed. Next, we experiment with different combinations of local and global (after FK) rotational loss functions to train

Table 3: Ablation Study results in the presence of 20% occlusion and a Gaussian noise of $\sigma = 20$ *mm*.

Experiments	MPVPE↓	PA-MPVPE↓	Rot. Error↓
SVD Symmetric Orthogonalization	**18.6**	**12.1**	**2.94**
6-DoF Symmetric Orthogonalization	19.2	12.6	3.13
Local Rot. Loss	19.9	13.3	3.01
Global Rot. Loss	19.3	13.2	3.63
Global+Local Rot. Loss	**18.6**	**12.1**	**2.94**
Part-based Decoder Att. Weights	18.8	12.1	2.96
$d = 1$ Decoder Att. Weights	19.6	12.8	2.97
$d = 2$ Decoder Att. Weights	19.1	12.4	2.93
$d = 3$ Decoder Att. Weights	19.3	12.6	2.96
$d = 4$ **Decoder Att. Weights**	**18.6**	**12.1**	**2.94**
$d = 5$ Decoder Att. Weights	18.9	12.6	2.97
w/o Decoder Att. Weights	19.7	12.8	3.04
w/o Rotation Normalization	19.5	13.1	3.01
w/o Mirror Test	27.7	19.7	3.56
w/o Shape Aug.	19.8	12.8	2.92



(a) VPoser on our joint regressor

(b) VPoser on Moon et al. 2022 joint regressor

Fig. 4: The fitting performance of VPoser is demonstrated while using (a) our proposed joint regressor; and (b) the joint regressor from [52].

our model. We observe that combining global and local rotation losses results in a better performance. Next, we showcase the effectiveness of our decoder masking strategy using the attention weights in three experiments: *i*) part-based experiment, where we restrict the attention within upper right, upper left, lower right, lower left, and center regions of the body; *ii*) node distance experiments, where we restrict the attention based on the node distance d in the skeleton kinematic tree with values between 1 and 5; and *iii*) without skeleton-aware attention masks, where we allow each joint to attend to any other joints in the decoder. We observe that the best results are obtained for $d = 4$, improving over the vanilla transformer decoder by 1.1 *mm* MPVPE and 0.2° rotation error. Lastly, we show the impact of rotation normalization, mirror testing, and shape augmentation, where a significant drop in performance is seen in the individual absence of each of these components, highlighting their significance. Finally, we perform a qualitative experiment on our joint regressor. To this end, we use VPoser [27] to fit onto 3D keypoints using our joint regressor and the one provided in prior works [27] on Human3.6m dataset [5]. In Fig. 4, we show the visual fidelity of the effect of our joint regressor in comparison to [52], specifically, in better fitting to the feet regions.

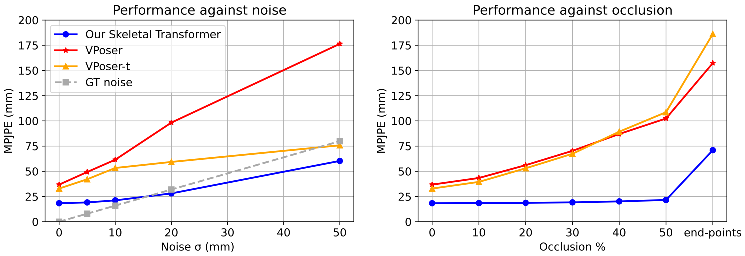


Fig. 5: Robustness of our skeletal transformer is highlighted in the presence of different levels of noise and occlusion by comparing it against VPoser and VPoser-t.

Robustness to Noise and Occlusions. To evaluate the performance of our model on motion capture data, we perform experiments by simulating noise and occlusion. For these experiments, we compare our skeletal transformer with VPoser [27] and its temporal version, VPoser-t [3], which tries to maximize temporal consistency during optimization. Figure 5 demonstrates the performance of different models on the AMASS [46] testing set. In the first experiment, we introduce varying noise levels to the input data and evaluate the robustness of our method. More specifically, Gaussian noise with varying standard deviations up to 50 *mm* is added to the input, effectively increasing the MPJPE of ground truth up to 80 *mm* (referred to as GT noise). However, our skeletal transformer predicts body pose and shape parameters, which result in lower error after $\sigma = 15$ *mm*. Additionally, its performance only degrades by 19.7 *mm* at maximum noise level, while VPoser fails to predict less noisy poses. Lastly, by comparing our method to a temporal model (VPoser-t), we demonstrate the model's robustness to noisy scenarios. Next, we report the performance of our model in the presence of occlusions, where each joint is randomly masked with varying occlusion amounts of up to 50%. Figure 5 shows that the performance of our skeletal transformer barely changes between 0 and 50% occlusion, thus showing the model's capability to exploit local and global joint information. In contrast, VPoser's performance drops by more than 20 *mm* in the presence of only 20% occlusion. We finally report the performance in an extreme occlusion scenario, where only 7 end-point keypoints are provided. Our model outperforms other solutions while maintaining a reasonable accuracy.

Qualitative Assessments. Figure 6 demonstrates the visual quality of SkelFormer on RICH [53] dataset in equally-sampled frames from the testing set. Our method yields the correct pose and shape with high overlaps with the subject.

5 Conclusion

In this paper, we presented SkelFormer, a novel multi-stage pipeline consisting of keypoint estimators and a skeletal transformer for markerless human motion capture. Our method leverages large amounts of motion capture data to address the poor generalization of multi-view human shape and pose estimation

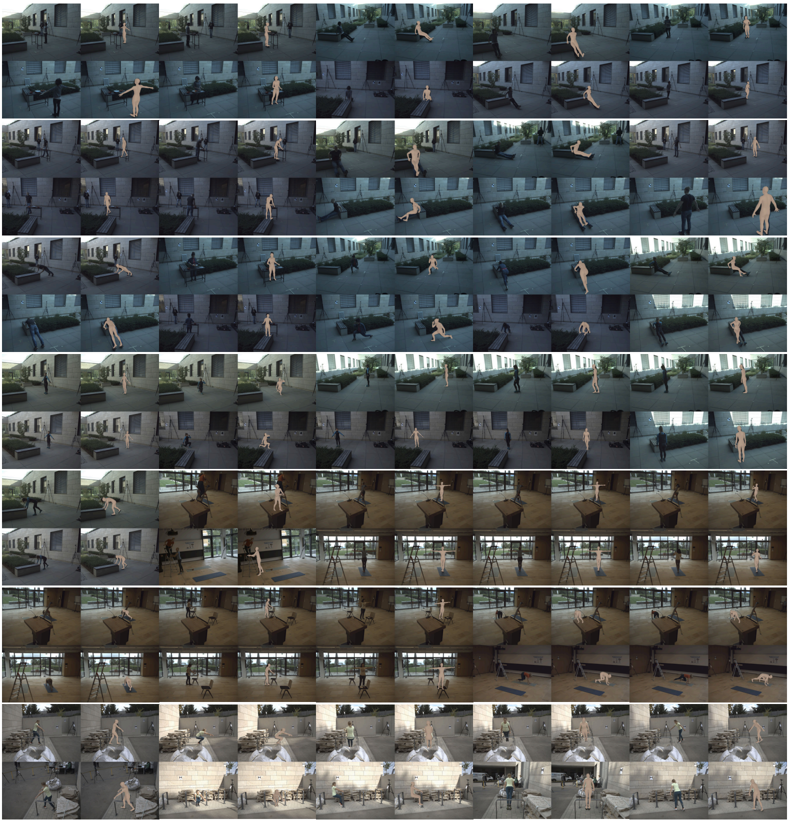


Fig. 6: Sample results on the RICH dataset are presented.

approaches while outperforming optimization approaches in accuracy. Through extensive experiments, we demonstrated the effectiveness of SkelFormer in several challenging conditions, including InD and OoD settings. Specifically, We achieve the best results in InD experiments among prior multi-view approaches and show competitive OoD performance, demonstrating the generalization of our pipeline. Furthermore, we show the effectiveness of the proposed elements in our skeletal transformer through our ablation study. Finally, our single-frame skeletal transformer exhibits higher noise and occlusion robustness than optimization approaches that rely on temporal data.

Future Work. The major limitation of this work is the accumulation of errors in the multi-stage pipeline. Although our network mitigates jitters and occlusions better than other approaches, it is still prone to artifacts caused by mediolateral mix-ups and incorrect tracking. More accurate human trackers [40, 41], 2D keypoint estimators, and triangulation techniques may be explored to remedy such issues. Another promising direction is to use SkelFormer to initialize generative models such as DMMR [28] and HuMoR [3] to further refine the motions.

Acknowledgment

This work was partially funded by Mitacs through the Accelerate program.

References

1. J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, “Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3383–3393.
2. Z. Yu, L. Zhang, Y. Xu, C. Tang, L. Tran, C. Keskin, and H. S. Park, “Multiview human body reconstruction from uncalibrated cameras,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
3. D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, “Humor: 3d human motion model for robust pose estimation,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 488–11 499.
4. K. Jia, H. Zhang, L. An, and Y. Liu, “Delving deep into pixel alignment feature for accurate multi-view human mesh recovery,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 989–997.
5. C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.
6. R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge University Press, 2004.
7. K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5693–5703.
8. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.
9. F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, “Distribution-aware coordinate representation for human pose estimation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7093–7102.
10. S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, “Whole-body human pose estimation in the wild,” in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 196–214.
11. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.
12. H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, “Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
13. R. Mitra, N. B. Gundavarapu, A. Sharma, and A. Jain, “Multiview-consistent semi-supervised learning for 3d human pose estimation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6907–6916.
14. V. Davoodnia, S. Ghorbani, and A. Etemad, “Estimating pose from pressure data for smart beds with deep image-based pose estimators,” *Applied Intelligence*, vol. 52, no. 2, pp. 2119–2133, 2022.

15. W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 15 085–15 099.
16. V. Davoodnia and A. Etemad, "Human pose estimation from ambiguous pressure recordings with spatio-temporal masked transformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
17. K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7718–7727.
18. Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng, "Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild," *International Journal of Computer Vision*, vol. 129, pp. 703–718, 2021.
19. M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–16, 2015.
20. J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1–17, 2017.
21. H. Choi, G. Moon, and K. M. Lee, "Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 769–787.
22. C. Zheng, M. Mendieta, P. Wang, A. Lu, and C. Chen, "A lightweight graph transformer network for human mesh reconstruction from 2d human pose," in *ACM International Conference on Multimedia (ACMMM)*, 2022, pp. 5496–5507.
23. H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 446–11 456.
24. Z. Li, M. Oskarsson, and A. Heyden, "3d human pose and shape estimation through collaborative learning and multi-view model-fitting," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1888–1897.
25. S. Shin and E. Halilaj, "Multi-view human pose and shape estimation using learnable volumetric aggregation," *arXiv preprint arXiv:2011.13427*, 2020.
26. J. Spörri, "Reasearch dedicated to sports injury prevention-the'sequence of prevention'on the example of alpine ski racing," *Habilitation with Venia Docendi in Biomechanics*, vol. 1, no. 2, p. 7, 2016.
27. G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 975–10 985.
28. B. Huang, Y. Shu, T. Zhang, and Y. Wang, "Dynamic multi-person mesh recovery from uncalibrated multi-view cameras," in *IEEE International Conference on 3D Vision (3DV)*, 2021, pp. 710–720.
29. L. Metz, C. D. Freeman, S. S. Schoenholz, and T. Kachman, "Gradients are not all you need," *arXiv preprint arXiv:2111.05803*, 2021.
30. N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 605–11 614.
31. A. Csiszar, J. Eilers, and A. Verl, "On solving the inverse kinematics problem using neural networks," in *IEEE International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, 2017, pp. 1–6.

32. R. Villegas, J. Yang, D. Ceylan, and H. Lee, “Neural kinematic networks for unsupervised motion retargetting,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8639–8648.
33. S. Raab, I. Leibovitch, P. Li, K. Aberman, O. Sorkine-Hornung, and D. Cohen-Or, “Modi: Unconditional motion synthesis from diverse data,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 873–13 883.
34. T. Jiang, N. C. Camgoz, and R. Bowden, “Skeletor: Skeletal transformers for robust body-pose estimation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3394–3402.
35. A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu, and Q. Xu, “Learning skeletal graph neural networks for hard 3d pose estimation,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 436–11 445.
36. L. Floridi and M. Chiriatti, “Gpt-3: Its nature, scope, limits, and consequences,” *Minds and Machines*, vol. 30, pp. 681–694, 2020.
37. Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “Vitpose: Simple vision transformer baselines for human pose estimation,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 38 571–38 584, 2022.
38. Y. Duan, Y. Lin, Z. Zou, Y. Yuan, Z. Qian, and B. Zhang, “A unified framework for real time motion completion,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 4459–4467.
39. J. Qin, Y. Zheng, and K. Zhou, “Motion in-betweening via two-stage transformers,” *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 1–16, 2022.
40. S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.
41. L. Bridgeman, M. Volino, J.-Y. Guillemaut, and A. Hilton, “Multi-person 3d pose estimation and tracking in sports,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, 2019.
42. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
43. N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, “Learning to reconstruct 3d human pose and shape via model-fitting in the loop,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2252–2261.
44. J. Levinson, C. Esteves, K. Chen, N. Snavely, A. Kanazawa, A. Rostamizadeh, and A. Makadia, “An analysis of svd for deep rotation estimation,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 22 554–22 565, 2020.
45. E. Hedlin, H. Rhodin, and K. M. Yi, “A simple method to boost human pose estimation accuracy by correcting the joint regressor for the human3. 6m dataset,” in *IEEE Conference on Robots and Vision (CRV)*, 2022, pp. 1–7.
46. N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, “Amass: Archive of motion capture as surface shapes,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5442–5451.
47. D. C. Liu and J. Nocedal, “On the limited memory bfgs method for large scale optimization,” *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.
48. A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7122–7131.
49. F. L. Markley, Y. Cheng, J. L. Crassidis, and Y. Oshman, “Averaging quaternions,” *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 4, pp. 1193–1197, 2007.

50. A. Sengupta, I. Budvytis, and R. Cipolla, "Synthetic training for accurate 3d human pose and shape estimation in the wild," *arXiv preprint arXiv:2009.10013*, 2020.
51. R. Girshick, "Fast r-cnn," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
52. G. Moon, H. Choi, and K. M. Lee, "Neuralannot: Neural annotator for 3d human mesh training sets," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2299–2307.
53. C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black, "Capturing and inferring dense full-body human-scene contact," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 274–13 285.
54. D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *IEEE International Conference on 3D Vision (3DV)*, 2017, pp. 506–516.
55. J. Liang and M. C. Lin, "Shape-aware human pose and shape reconstruction using multi-view images," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4352–4362.
56. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2018.
57. Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7103–7112.
58. Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black, "Towards accurate marker-less human shape and pose estimation over time," in *IEEE International Conference on 3D Vision (3DV)*, 2017, pp. 421–430.
59. X. Gong, L. Song, M. Zheng, B. Planche, T. Chen, J. Yuan, D. Doermann, and Z. Wu, "Progressive multi-view human mesh recovery with self-supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 676–684.
60. X. Jiang, X. Nie, Z. Wang, L. Liu, and S. Liu, "Multi-view human body mesh translator," *arXiv preprint arXiv:2210.01886*, 2022.
61. K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1954–1963.
62. Z. Shen, Z. Cen, S. Peng, Q. Shuai, H. Bao, and X. Zhou, "Learning human mesh recovery in 3d scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 038–17 047.
63. S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas, "3d human pose estimation via intuitive physics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4713–4725.
64. M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3d human body estimation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 127–11 137.
65. Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," in *European Conference on Computer Vision (ECCV).* Springer, 2022, pp. 590–606.