PostureHMR: Posture Transformation for 3D Human Mesh Recovery

Yu-Pei Song[1,2], Xiao Wu[1,2]*, Zhaoquan Yuan[1,2], Jian-Jun Qiao[1,2], Qiang Peng[1,2]
[1] Southwest Jiaotong University, Chengdu, China
[2] Engineering Research Center of Sustainable Urban Intelligent Transportation, China
yupei-song@my.swjtu.edu.cn, {wuxiaohk, zqyuan, qpeng}@swjtu.edu.cn, qjjai56@gmail.com

Abstract

Human Mesh Recovery (HMR) aims to estimate the 3D human body from 2D images, which is a challenging task due to inherent ambiguities in translating 2D observations to 3D space. A novel approach called PostureHMR is proposed to leverage a multi-step diffusion-style process, which converts this task into a posture transformation from an SMPL T-pose mesh to the target mesh. To inject the learning process of posture transformation with the physical structure of the human body model, a kinematics-based forward process is proposed to interpolate the intermediate state with pose and shape decomposition. Moreover, a mesh-to-posture (M2P) decoder is designed, by combining the input of 3D and 2D mesh constraints estimated from the image to model the posture changes in the reverse process. It mitigates the difficulties of posture change learning directly from RGB pixels. To overcome the limitation of pixel-level misalignment of modeling results with the input image, a new trimap-based rendering loss is designed to highlight the areas with poor recognition. Experiments conducted on three widely used datasets demonstrate that the proposed approach outperforms the state-of-the-art methods.

1. Introduction

Human Mesh Recovery (HMR) aims to reconstruct a 3D human body mesh from a 2D image. This task has diverse applications, such as virtual reality, human-computer interaction, clothed human reconstruction and posture capture. With the rise of the metaverse, HMR has become a critical technology to create digital humans. Unfortunately, HMR remains a challenging task, due to inherent depth ambiguities, flexible body kinematic structures, diverse visual appearances and ubiquitous part occlusions [18, 20, 42].

Skinned Multi-Person Linear Model (SMPL) [27] is an influential open-source statistical model of the human body, enabling realistic 3D human generation and analysis across a wide variety of applications. With the success of statistical human models, such as SCAPE [1], SMPL [27] and SMPL-X [31], body deformations are factored into identity-dependent and pose-dependent shape deformations, corresponding to shape and pose parameters in SMPL, respectively. In this paper, they are collectively referred as posture. Pioneer methods [7, 17, 29] regress the vertex positions of the mesh, achieving pixel alignment with the image. However, the reconstruction performance is adversely affected due to the lack of depth information, which poses challenges for producing reasonable results from side views. Recently, HMDiff [11] applies the diffusion model to the HMR task, which generates the 3D model from Gaussian noises in a multiple iteration way. It has demonstrated promising performance. However, the overall physical structure is destroyed when adding noises to mesh vertices.

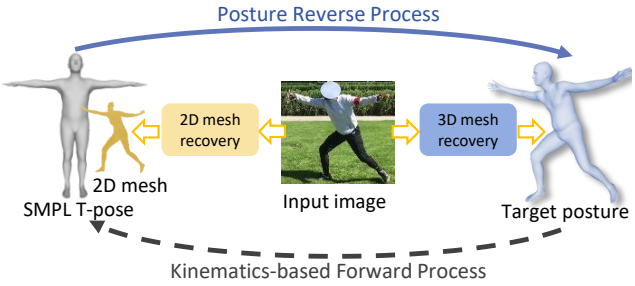

Figure 1. The SMPL T-pose model and the estimated 3D target model from the image are progressively updated through PostureHMR. The 2D mesh estimated by the neural network is combined with the SMPL model of the T-pose, which can gradually transformed to the target posture via a posture reverse process. The learning labels corresponding to the intermediate steps are accurately obtained through a kinematics-based forward process.

To address the limitations of vertex regression methods, incorporating the SMPL T-pose model as input enables the integration of prior knowledge related to depth information. Models utilizing alternative poses may encounter difficulties in preserving accurate structural information at the joint regions. Due to the substantial disparity between the input and output, it becomes challenging to precisely reconstruct

*Corresponding author.

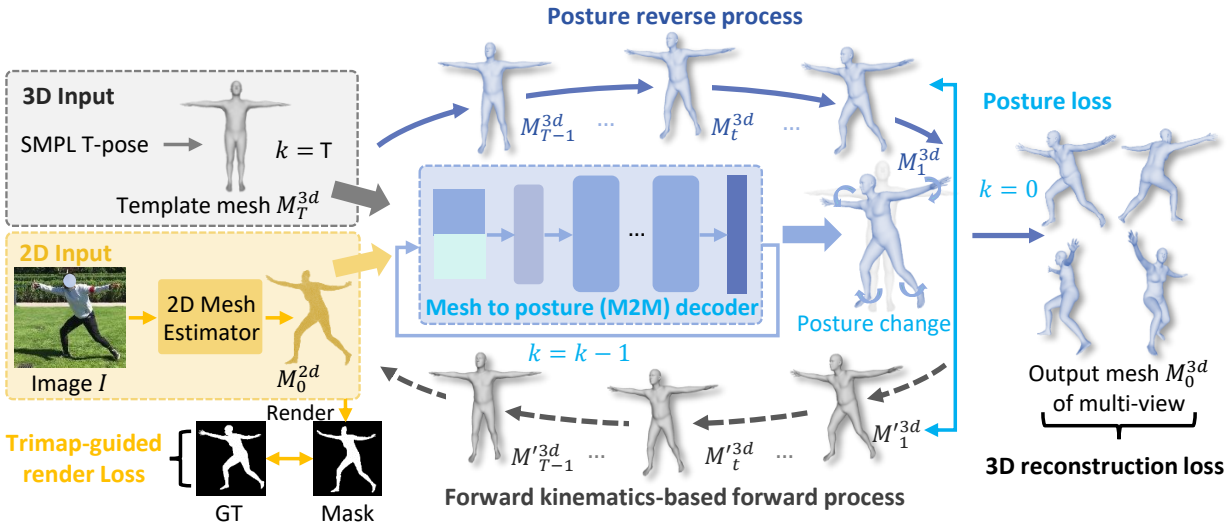

Figure 2. The framework of the proposed PostureHMR, which progressively captures the posture transition from the standard 3D mesh to the target posture. The architecture employs diffusion-style learning with a kinematics-based forward process and posture reverse process. Posture loss, 3D reconstruction loss and trimap-guided rendering loss are comprehensively considered for model training.

the mesh. This process is converted into multiple iterations from the initial posture to the corresponding posture. Acting as the supervisory information, the mesh model generated with the SMPL will guide the learning direction during the intermediate iteration process.

In this paper, we will explore the 3D human mesh recovery from a new perspective. A diffusion-style learning process called PostureHMR is proposed to learn the posture transformations from the standard SMPL T-pose model to the target model, as is illustrated in Fig. 1. It mainly comprises a kinematics-based forward process and a posture reverse process. During the training stage, the SMPL ground truth mesh is gradually transferred to the SMPL template mesh in the forward process. The deep learning model is learned step-by-step from the T-pose to the target posture in the reverse process. During inference, PostureHMR progressively adjusts the SMPL template mesh to align with the posture corresponding to the input image.

The framework of the proposed PostureHMR is illustrated in Fig. 2. To maintain the mesh topology, a forward kinematics (FK) based method is proposed during the forward process, which adopts skeletal animation for pose transformation and linear interpolation within low-dimensional shape space. To mitigate the abstraction of modeling posture from images at the reverse process, a mesh-to-posture (M2P) decoder is designed with 3D mesh and 2D mesh constraints estimated from the input images. This decoder gradually captures non-local and neighborhood interaction information among mesh vertices. Moreover, a trimap-guided rendering loss is provided to address pixel-level misalignment with the input image. Serving as a supervisor, this loss enhances the model attention on areas with poor recognition, facilitating a better alignment. Through the integration of this strategy alongside posture loss and 3D reconstruction loss, the entire model incrementally optimizes the posture transformation, achieving pixel alignment and smooth outcomes.

Different from the classic diffusion models, PostureHMR refrains from utilizing denoising in the posture transformation process. This approach preserves the human body structure throughout the learning process, thereby facilitating detailed modeling. In addition, the inputted SMPL model offers abundant pose, shape and depth information, serving as a robust prior to guide mesh recovery. This compensates for the absence of depth information when inputting a single image, enabling the reconstruction results to yield reasonable modeling outcomes from various viewpoints. The contributions are summarized as follows:

- A posture transformation architecture called PostureHMR is novelly proposed for 3D human mesh recovery. It progressively captures the posture transition from the standard 3D mesh to the target posture, conditioned by the input 2D image. The architecture employs a diffusion-style learning with a kinematics-based forward process and a posture reverse process.
- To model the posture transformation, a forward kinematics-based skeletal animation is adopted to capture the human mesh posture tendency during the forward process. Conditioned by the 2D mesh, a mesh-to-posture (M2P) decoder is designed to learn the posture transition at the reverse process, by guiding the 3D mesh towards the target posture step by step.
- Three factors (posture change, human structure and pixel alignment) are comprehensively considered for model training, corresponding to posture loss, 3D reconstruction loss and trimap-guided rendering loss, respectively.

- Experiments conducted on three public datasets demonstrate that PostureHMR achieves promising performance, which outperforms the state-of-the-art approaches.

2. Related work

2.1. 3D Human Mesh Recovery

Recent research on human mesh recovery from single-image is classified into two categories: model-based methods [4, 15, 16, 23, 41, 47, 55] and model-free ones [6, 17, 24, 25, 29, 49]. Model-based methods regress parameters of the statistical model, which mainly understand the semantic information of the images. However, the transformation from images to abstract parameter space is highly non-linear [10, 39], resulting in a rough alignment of modeling results and image evidence [52]. Some works have introduced different intermediate expressions to reduce the complexity of model learning, such as 2D/3D bone points [8, 22, 28], IUV mapping [52] and markers [51]. Recently, the model-free methods directly regress the vertices of the mesh with superior flexibility, achieving better pixel alignment than model-based methods. However, due to the absence of depth information from a single image, it is difficult to recover reasonable results from the side views [48].

In this work, the challenging problem is explored by leveraging the SMPL template model as an input. The prior knowledge is incorporated to achieve a physically plausible reconstruction. Different from optimization methods [3, 9, 19, 50], posture transformation is deployed in this paper and intermediate models are adopted to guide the iteration.

2.2. Diffusion Models

Diffusion models have gained popularity as generative models, which have proven effective in generating arbitrary high-quality images [34, 35]. These models gradually add Gaussian noises to the original data, subsequently treating the generation process as a denoising task [13]. In addition to image generation, the diffusion models have been applied to other tasks as a new learning framework, such as image classification [32, 36], segmentation [2, 33] and 3D vision [12, 37, 38]. Diff-HMR [5] and EgoHMR [54] generate parameters of SMPL through diffusion models, while HMDiff [11] directly adds noises on vertices of the mesh. However, after introducing noise, the intermediate mesh loses its reasonable physical distribution, thus posing challenges in achieving pixel alignment or obtaining smoothed results.

Unlike traditional diffusion methods, our approach converts it into a posture transformation process, which aligns with the transition of the statistical model from the initial state to the target state. In addition, the prior knowledge of human body structure is preserved during the learning process by introducing skeletal animation.

3. PostureHMR

3.1. Framework

Given a template human mesh M_T^{3d} and an image I, PostureHMR aims to recover the 3D positions of human mesh vertices $M_0^{3d} \in \mathbb{R}^{N\times 3}$ through T reverse steps, where N is the number of vertices. PostureHMR comprises a multi-step diffusion-style process, which consists of a kinematics-based forward process and a posture reverse process. The framework is illustrated in Fig. 2. The forward kinematics-based forward process obtains the posture transformation from the target posture M_0^{3d} to SMPL T-pose M_T^{3d}. The posture reverse process learns the posture change through the mesh-to-posture (M2P) decoder under the condition of 2D mesh M_0^{2d} input, which is estimated from the 2D image.

3D mesh: SMPL is a widely used human body statistical model for realistic human 3D pose and shape estimation. It provides a differentiable function $\mathcal{M}(\theta, \beta)$ that can express a dedicated human body model in any posture. The initial state of SMPL is a T-pose model with 6,890 vertices, and the expressions of different people are obtained by adjusting the shape $\theta \in \mathbb{R}^{23\times 3}$ and the pose $\beta \in \mathbb{R}^{10}$ parameters. In this paper, the template model of SMPL is used as the initial input of the mesh-to-posture (M2P) decoder to model the posture transformation.

2D mesh: A Convolution Neural Network (CNN) backbone network, HRNet [45], is used to extract deep features from the 2D image. A lightweight decode module is then conducted to generate a 2D heatmap. The keypoints $P \in \mathbb{R}^{K\times 2}$ are obtained by calculating the center of mass of the heatmap. To reduce the redundant calculations due to the large number of mesh vertices, a coefficient matrix C is learned. Following [29], the interpolation from sparse points to a complete 2D mesh is implemented by computing $M_0^{2d} = CP$. The number of output points is 431.

3.2. Forward Kinematics-based Forward Process

The forward process of traditional diffusion models introduces Gaussian noises to the ground truth until complete diffusion into the noise distribution. To guarantee that the initial inputs of the model remain a Gaussian distribution throughout both the training and testing phases, it is generally advisable to assign a relatively high value (e.g., 1000) to the iteration number of the diffusion models. Although DDIM [40] speeds up the inference time by reducing the number of iterations during inference, the randomness of the initial sampling can easily lead to the same randomness into the final modeling results [37], when applied to HMR tasks. It is straightforward to apply the original sampling strategy to the vertices by adding noises, which will lose the physical structure of the human body during the intermediate process. Therefore, it is challenging to establish the relationship between 3D modeling derived from random

noises and corresponding input images.

A novel diffusion-style posture transformation method is proposed in this paper. To facilitate the understanding, the HMR task is converted to a transformation, from a model of the initial posture (T-pose) to the target posture of the input image. By using the T-pose model as the end of the forward process, the number of iterations is significantly reduced and multiple samplings are not required during the inference. When the positions of the mesh vertices in the intermediate steps are obtained, the initial and final states are linearly interpolated like noise diffusion. However, this approach fails to capture the physical structure of the human model. Therefore, a new forward process based on forward kinematics is proposed to capture the posture changes through skeleton animation interpolation.

Forward kinematics. Forward kinematics in skeletal animation refers to the use of kinematic equations to calculate the position updates of bone B at different frames. The joint parameters R and template joint positions J are used to calculate the updates:

$$B = \mathrm{FK}(R, J). \tag{1}$$

The keypoint positions of the bones with different interpolation results are obtained by interpolating the rotation information of the bones and combining it with forward kinematics (FK) calculation. The positions of mesh vertices are obtained based on the weight relationship between the bone points and the vertices of SMPL.

FK-based forward process. The forward process provides the supervision guidance for the model learning in the reverse process, which involves a smooth transition from the T-pose to the target posture ($M_T^{3d} \rightarrow M_{T-1}^{3d} \rightarrow ... \rightarrow M_t^{3d} ... \rightarrow M_0^{3d}$). Posture changes are mainly affected by two factors: pose and shape. Therefore, skeletal animation interpolation is adopted for pose, while linear interpolation of the effect of SMPL parameters on vertex offset is used for shape. The shape-and-pose decomposition is illustrated in Fig. 3. Interpolation is decomposed into rotation transformation of pose and linear interpolation of shape. The whole implementation is formulated as:

$$q(M_t^{3d}|M_T^{3d}) = W(R_t^{3d}, J_t, \theta_t, \mathcal{W}), \tag{2}$$

where $W(\cdot)$ is the standard linear blend skinning function. R_t^{3d} represents the vertices of the rest pose after applying parameter blend shape deformation, J_t denotes the joint locations following shape deformation, θ_t signifies the shape parameters at t time step and $\mathcal{W} \in R^{N \times K}$ stands for the blend weights. R_t^{3d} is transformed from the template mesh M_T^{3d} as:

$$R_t^{3d} = M_0^{3d} - \frac{t}{T}(B_S(\beta_0) + B_P(\theta_0)), \tag{3}$$

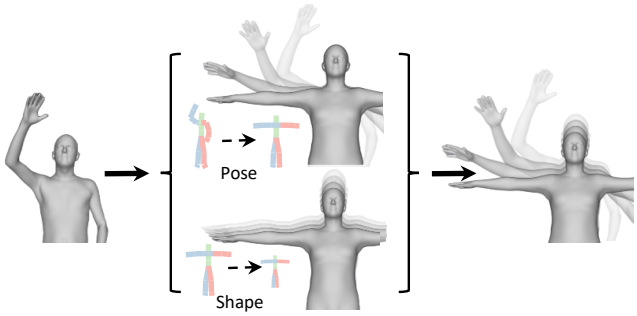

Figure 3. Illustration of the shape-and-pose decomposition. The interpolation process includes pose changes driven by skeletal animation and shape changes, respectively.

where B_S and B_P represent the shape and pose blend shape offsets for the vertex, respectively. θ at step t is obtained from the linear interpolation as:

$$\theta_t = \theta_0 + \frac{t}{T}(\theta_T - \theta_0), \tag{4}$$

where θ_0 and θ_T denote the shape parameters of the initial and final status, respectively.

The proposed method mainly has two advantages over the original noise diffusion process. First, the T-pose is a fixed state that requires fewer steps and reduces the randomness of the network. Second, the intermediate state preserves the physical structure of the human body and imposes more regular constraints on the network during the learning process.

3.3. Posture Reverse Process with 2D Condition

The traditional inverse process is dedicated to gradual denoising and restoring images from Gaussian noise. Meanwhile, the neural network predicts the noise through image/text conditions and steps. For the HMR task, the template model of SMPL replaces Gaussian noises, and in the reverse process, the neural network needs to learn the posture conversion from the T-pose mesh to the target mesh. Therefore, a mesh-to-posture (M2P) decoder is designed to learn posture transitions between different iteration steps. Although it is intuitive to use image features as condition information to reconstruct a specific input image, this reconstruction process is very difficult, since the transformation directly from image input to posture is highly non-linear. Previous works tend to utilize some intermediate expressions as guidance (e.g., 3D pose [22], segmentation map [52]) for HMR task. In this paper, a 2D mesh is employed as a conditioning factor for the modification of 3D mesh. This 2D mesh is derived through image modeling and aims to achieve a align result with the input image. Formally, given the SMPL initial model M_T^{3d}, constraint information 2D mesh M_0^{2d} and the number of iteration steps k, the re-

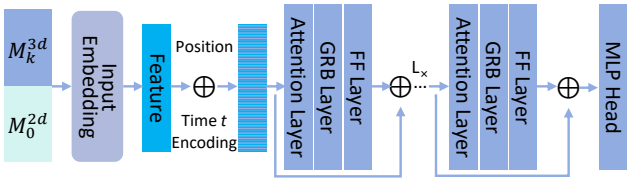

Figure 4. Detailed architecture of the M2P decoder. It combines 3D and 2D mesh as input and outputs body posture transformation.

verse process is formulated as:

$$p_\alpha(M_{0:T}^{3d}) = p(M_T^{3d}) \prod_{k=1}^{T} p_\alpha(M_{k-1}^{3d} | M_k^{3d}, M_0^{2d}), \quad (5)$$

where α is the parameters of M2P decoder.

M2P decoder. Traditional reverse process for image generation generally uses U-net as the decoder F_α. However, in this task, the decoder is built for posture representation. The detailed structure of the network is shown in Fig. 4. To guide the posture learning at the reverse process, 2D mesh M_0^{2d} is used as additional information, which is directly concatenated with 3D mesh M_k^{3d} as inputs. It is embedded via a linear layer. Since different iteration steps share the same network structure, it is necessary to input the current time step information via the sinusoidal methods and add position embedding to maintain the spatial distribution of vertices. These tokens are sent into L transformers with graph blocks, which consist of an attention layer, a graphormer block (GRB) layer and a feedforward (FF) layer. The design of the token is followed by Graphormer [25], which considers non-local interactions among mesh vertices and neighborhood vertex interactions based on mesh topology. Finally, the Multi-Layer Perceptron (MLP) head outputs the posture change from k to $k-1$.

The proposed design has the advantage of modeling the 2D and 3D mesh separately. Image input is more suitable to model the mesh at the 2D level while the M2P decoder learns the prior 3D knowledge of the SMPL model. When the 3D models are constrained with the 2D mesh, the posture changes can be more accurately captured.

3.4. Model Training and Loss Functions

Trimap-based render loss for 2D mesh learning. While model-free methods facilitate the achievement of pixel-aligned with the input image compared to model-based approaches, they often overlook the edge areas of the human body. The trimap regions are illustrated in Fig. 5, which represent the target edge area. This area highly coincides with the region where the reconstruction result error is large. Consequently, this information is employed to assign weights to the modeling process at various locations, thereby enhancing the modeling attention in specific regions. Specifically, differential rendering is performed on the output results M_0^{2d} and ground truth M'^{2d}_0 to obtain the

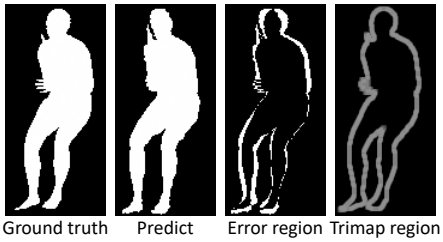

Figure 5. Illustration of the trimap region.

mask. Considering that directly using L1 loss calculation may be influenced by the background, the IoU loss is used instead, which is calculated as follows:

$$\mathcal{L}_{tri} = 2 - \frac{P \cap G}{P \cup G} - \frac{P \cap G \cap C_{tri}}{(P \cup G) \cap C_{tri}}, \quad (6)$$

where P and G represent the model prediction and the ground truth mask, respectively. C_{tri} is the trimap region that is set to the edge l pixel area of the ground truth mask. The 2D reconstruction loss is defined as: $\mathcal{L}_{2d} = \mathcal{L}_{tri} + \lambda_v^{2d}\mathcal{L}_v^{2d}$, where $\mathcal{L}_v^{2d}$ denotes the L1 loss calculation of the 2D mesh.

3D mesh learning. The posture learning in the reverse process is supervised through the labels generated by the forward process. The posture reconstruction loss $\mathcal{L}_{pos}$ is implemented as follows:

$$\mathcal{L}_{pos} = \sum_{k=1}^{T} || (M'^{3d}_{k-1} - M'^{3d}_{k}) - F_\alpha(M_k^{3d}, M_0^{2d}, k) ||_2^2. \quad (7)$$

Following [29], geometry optimization is incorporated in each step, including 3D vertex loss, 3D keypoint loss and surface loss. Please refer to [29] for further details.

Therefore, the loss of 3D reconstruction is expressed as: $\mathcal{L}_{3d} = \mathcal{L}_{pos} + \lambda_v^{3d}\mathcal{L}_v^{3d} + \lambda_j^{3d}\mathcal{L}_j^{3d} + \lambda_s^{3d}\mathcal{L}_s^{3d}$.

4. Experiments

4.1. Datasets and Evaluation Metrics

Human3.6M [14]: It is the largest indoor benchmark dataset for the human pose estimation. The training and testing data are the same as previous works [15, 24, 25]. (S1, S5, S6, S7, S8) are used for training, and (S9, S11) for testing.

3DPW [44]: It is a benchmark dataset for human mesh estimation, which is collected in natural scenes. In the experiment, the model trained by Human3.6M is fine-tuned on its training set to obtain the evaluation results of the test set.

SURREAL [43]: It is a synthetic dataset by combining various SMPL models with arbitrary backgrounds. The partition of training and test sets remains the same as previous works [7, 29].

This work does not involve any human data that raises ethical concerns.

Method	Year	Human3.6M			3DPW		
		MPVE ↓	MPJPE ↓	PA-MPJPE ↓	MPVE ↓	MPJPE ↓	PA-MPJPE ↓
HMR [15]	CVPR'18	96.1	88.0	56.8	152.7	130.0	81.3
PC-HMR [28]	AAAI'21	61.1	47.9	37.3	108.6	87.8	66.9
HybrIK [22]	CVPR'21	65.7	54.4	34.5	86.5	74.1	45.0
ROMP [41]	ICCV'21	-	-	-	108.3	91.3	54.9
PARE [18]	ICCV'21	-	-	-	88.6	74.5	46.5
THUNDR [51]	ICCV'21	-	55.0	39.8	88.0	74.8	51.5
PyMAF [52]	ICCV'21	-	57.7	40.5	110.1	92.8	58.9
ProHMR [20]	ICCV'21	-	-	41.2	-	-	59.8
OCHMR [16]	CVPR'22	-	-	-	107.1	89.7	58.3
3DCrowNet [8]	CVPR'22	-	-	-	98.3	81.7	51.5
CLIFF [23]	ECCV'22	-	47.1	32.7	81.2	69.0	43.0
ProPose [10]	CVPR'23	-	45.7	**29.1**	79.4	68.3	40.6
PLIKS [39]	CVPR'23	-	49.3	34.7	82.6	66.9	42.8
*MeshTransformer [24]	CVPR'21	-	54.0	36.7	88.2	77.1	47.9
*MeshGraphormer [25]	ICCV'21	-	51.2	34.5	87.7	74.7	45.6
*FastMETRO [6]	ECCV'22	-	52.2	33.7	84.1	73.5	44.6
*VisDB [48]	ECCV'22	-	51.0	34.5	85.5	73.5	44.9
*PointHMR [17]	CVPR'23	-	48.3	32.9	84.1	73.9	44.9
*DeFormer [49]	CVPR'23	-	44.8	31.6	82.6	72.9	44.3
*HMDiff [11]	ICCV'23	-	49.3	32.4	82.4	72.7	44.5
*Zolly [46]	ICCV'23	-	49.4	32.3	76.3	65.0	39.8
*VirtualMarker [29]	CVPR'23	58.0	47.3	32.0	77.9	67.5	41.4
*PostureHMR	-	**55.7**	**44.5**	31.0	**75.4**	**64.9**	**39.6**

Table 1. Performance comparison on H3.6M and 3DPW datasets. Model-based and model-free methods (indicated with $*$) cannot be fairly compared since they use different backbone networks and training strategies.

Similar to [17, 22, 25, 39, 52], Mean-PerVertex-Error (**MPVE**), Mean-Per-Joint-Position-Error (**MPJPE**) and Procrustes Analysis MPJPE (**PA-MPJPE**) are used as the performance metrics. These metrics are reported in millimeters (mm) by default. PVE is calculated as the Average Point-to-point Euclidean distance between vertices. Following [22, 29], the metric of MPVE on the H3.6M dataset is also given. MPJPE stands for mean bone key points error. PA-MPJPE calculates MPJPE after aligning the predictions to the ground.

4.2. Implementation Details

The 2D mesh learning process involves cropping every single human region from the input image and uniformly setting it to 256×256. HRNet-W48 is used as the backbone network, which is pre-trained on the COCO [26] 2D pose dataset and the size of the heatmap is 64×64. Similar to [29], the initial output number of the 2D points is 81. Following [11, 23, 24, 29, 52], additional data from MPIINF-3DHP [30], UP-3D [21] and COCO training sets are used for hybrid training to improve the image to 2D mesh reconstruction, and experiments are performed on H3.6M and 3DPW datasets. The 2D mesh estimate model is frozen during diffusion model training. The reverse diffusion steps T is set to 10. The numbers of embedding and transformer channels are kept at 128 and 512, respectively. Similar to previous works [11, 25], the coarse human mesh has 431 vertices and the refined mesh contains 6890 vertices, which is obtained through an MLP layer. The model has trained for 40 epochs in the 2D and 3D learning stages, with the initial learning rate of 0.001 for backbone, $5e^{-4}$ for 2D mesh recovery and $5e^{-5}$ for 3D posture transformation, respectively. It is reduced by half after 30 epochs. The weight of the rendering loss is set to 10 and other losses are the same as [29]. All experiments are carried out on four GeForce RTX 3090 GPUs.

4.3. Comparison with State-of-the-art Methods

Results on H3.6M and 3DPW. The proposed method is compared with the state-of-the-art methods on the H3.6M and 3DPW datasets, which is listed in Table 1. To maintain a fair comparison, the results of PLIKS [39] utilizing additional AGORA data for training are not reported in this paper. Our approach achieves competitive or superior performance among state-of-the-art (SOTA) methods, thus effectively validating the advantages of the posture diffusion learning process. PostureHMR outperforms the vertex regression method (VirtualMarker [29]) and SMPL parametric regression approaches, such as CLIFF [23], PyMAF [52] and HybrIK [22]. HMDiff [11] is similar to PostureHMR as it learns the denoise process to recover the human mesh. However, our approach achieves better results by transform-

Method	Year	MPVE ↓	MPJPE ↓	PA-MPJPE ↓
HMR [15]	CVPR'18	85.1	73.6	55.4
DynaBOA [53]	PAMI'22	70.7	55.2	34.0
*Pose2Mesh [7]	ECCV'20	68.8	56.6	39.6
*PC-HMR [28]	AAAI'21	59.8	51.7	37.9
*VirtualMarker	CVPR'23	44.7	36.9	28.9
*PostureHMR	-	**42.1**	**35.3**	**27.4**

Table 2. Performance comparison on SURREAL. Model-free methods are indicated with ∗.

ing the human mesh reconstruction process into a transition from the T-pose to the target pose, which effectively retains the physical structure of the human body during the learning process. MeshGraphormer [25] also utilizes the T-pose of SMPL as input, which is similar to our method when the number of iterations is set to 1. However, without multiple iterations, this method encounters the problem of substantial differences between the initial posture and the target one.

Results on SURREAL. SURREAL is a simulation dataset that contains more changes in shape. The results are presented in Table 2. Model-based methods are highly nonlinear, such as HMR [15], which model SMPL abstract parameters directly from images. DynaBOA [53] uses 3D pose to improve the difficulty of pose parameters modeling, but it is still limited by the expression of shape. The model-free method represented by VirtualMarker is limited by the ambiguity of depth information. PostureHMR mainly learns the SMPL template transformation of posture and retains more human structure information, which outperforms the state-of-the-art methods.

Qualitative results. PostureHMR is compared with model-based method CLIFF [23] and model-free method VirtualMarker [29], on H3.6M, 3DPW and SURREAL datasets, which are illustrated in Figs. 6, 7 and 8, respectively. From the image perspective, PostureHMR and VirtualMarker have better pixel alignment compared to CLIFF. Thanks to the trimap rendering loss, PostureHMR boosts the performance, since the loss guides the model to focus on non-aligned areas. For the side view, although CLIFF models a satisfactory body shape, CLIFF and VirtualMarker cannot accurately model the pose. The visual appearance generated by VirtualMarker is squashed, which is caused due to the ambiguous depth information. PosetureHMR applies posture transformation to constrain the structure of the human body during the learning phase, ensuring a smooth representation of its form. When coupled with prior depth information, it becomes easier to observe realistic posture representations from a side perspective.

Limitation. Some failure examples are illustrated in Fig. 9. Due to the occlusion or incomplete body shape, our method has incorrect pose estimation. It is notable that the failure area is confined to the invisible regions. As a potential solution, enhancing the number of occlusion samples via data augmentation may be effective.

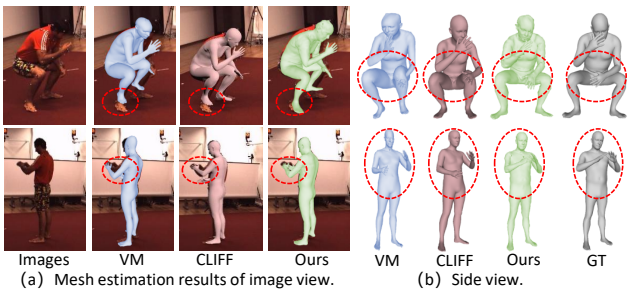

Figure 6. Qualitative comparison with VirtualMarker [29] and CLIFF [23] on H3.6M test set.

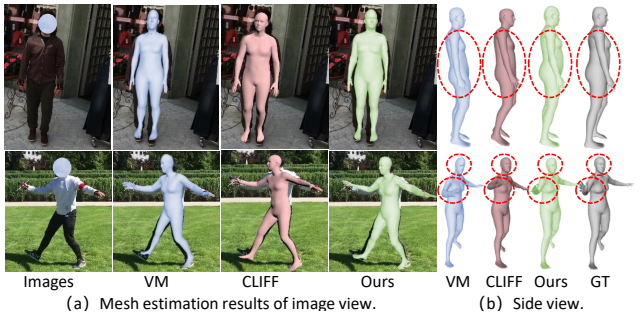

Figure 7. Qualitative comparison on 3DPW test set.

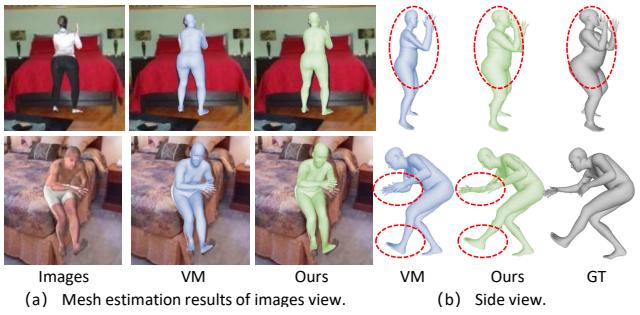

Figure 8. Qualitative comparison on SURREAL test set.

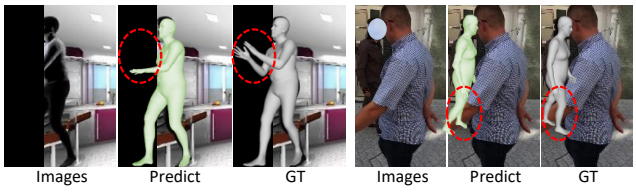

Figure 9. Examples of some failure cases.

4.4. Ablation Study

Impact of posture learning. The effect of posture transformation is first evaluated in Table 3. A baseline model is first constructed, sharing the same backbone and decoder networks as PostureHMR. Its learning is grounded in the denoising technique employed by the diffusion model. Despite the differences in decoder design and the utilization

Methods	MPVE ↓	
	3DPW	SURREAL
Basline	83.2	48.6
Basline + VL	80.3	46.2
Basline + FK	78.5	44.5
Basline + FK + 2D mesh	76.2	42.7
Basline + FK + 2D mesh + render	**75.4**	**42.1**

Table 3. Effect of individual component.

of 2D mesh condition, its fundamental learning principles align with those of HMDiff [11]. The iteration steps are fixed at 1000. In the test phase, DDIM [40] is utilized to expedite the process by reducing the number of iterations to 10. The backbone network learns the condition information to incorporate the global image information, producing a vector with the length of 2048. The vector is concatenated with the 3D mesh vertices. Compared to the baseline, noise diffusion is replaced with vertex linear interpolation (Baseline + VL) between the SMPL T-pose and the ground truth mesh. This approach significantly outperforms the baseline model. However, when forward kinematics (Baseline+FK) is integrated into the posture transformation process, substantial improvement is obtained. Compared to the method (Baseline+VL), this approach effectively preserves the prior knowledge of the physical structure of the human body during the inverse learning process. Fig. 10 depicts the modeling outcomes of FK constraints represented by the green human figure and the results of noise diffusion (w/o FK) illustrated by the blue human figure. Notably, our method performs better on surface details.

Impact of 2D conditions. To validate the effectiveness of the 2D condition design, the initial step involves learning the 2D representation of the mesh (+2D mesh). Following this, this representation is employed as a conditioning factor. When comparing the experimental results with those obtained using image global features as the conditioning factor (Baseline+FK), our 2D mesh conditioning approach demonstrates superior performance. This advantage primarily stems from two factors. Firstly, distinguishing between 2D and 3D information modeling enables the network to effectively mitigate the uncertainty associated with modeling across different dimensions during the learning process. Secondly, compared to image feature information, the 2D mesh offers superior local details, which allows for a more precise constraint on the distribution of 3D mesh reconstruction results within the image's viewing angles.

Impact of Trimap-based Render Loss. The role of trimap-based render loss in PostureHMR is evaluated in Table 3, indicating with (+render). The overall modeling accuracy is further improved, which is mainly due to the optimization of condition input in the 3D mesh modeling process. A refined 2D mesh that aligns closely with the image input aids in achieving precise posture changes in the 3D mesh.

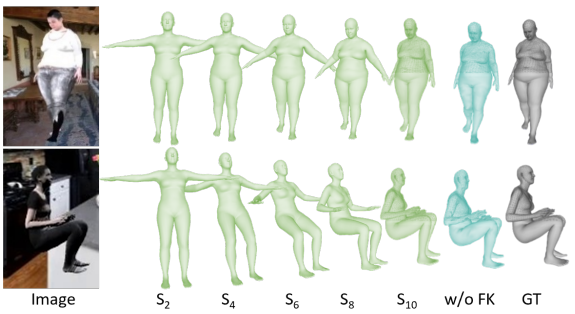

Figure 10. Visualization of the transition from T-pose to the target.

Impact of Diffusion Step k. Consistent with prior research (e.g., [11]) on diffusion models, the performance improvement tends to be slowed down as the number of iterations increases. It usually reaches its climax when k=10. Due to the limited number of diffusion steps and the huge gap between the initial and final postures, it is difficult to accurately model the correlation between them. As the the number of diffusion steps increases, additional intermediate nodes are constructed, so that the gap between the initial and final poses is narrowed, which enhances the overall accuracy. However, with more diffusion steps, it is easier for the network to focus more on the changes between local steps, while ignoring the variations within the entire process. The transition from the T-pose to the target posture is illustrated in Fig. 10, from which we can see how the posture is transferred toward the target status step by step. The pose and shape transformations are simultaneously modeled and constructed throughout the learning process.

5. Conclusion

In this paper, a novel posture transformation framework called PostureHMR is proposed to model a human 3D mesh from a single 2D image. It consists of a kinematics-based forward process and posture reverse process, which progressively convert the SMPL T-pose model to the target model at the inference stage. Moreover, a forward kinematics-based skeletal animation is adopted to capture the posture change, and a mesh-to-posture (M2P) is designed to learn the transformation at the reverse process. A trimap-based rendering loss is proposed to provide better pixel alignment of 2D conditions with input images. PostureHMR outperforms the state-of-the-art methods on three widely used benchmark datasets.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372387, 61802053), Key R&D Program of Guangxi Zhuang Autonomous Region, China (Grant No. AB22080038, AB22080039), Natural Science Foundation of Sichuan (Grant No. 24NSFSC0900).

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, et al. Scape: Shape completion and animation of people. In *SIGGRAPH*, page 408–416, 2005. 1

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022. 3

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, et al. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578, 2016. 3

[4] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *ECCV*, pages 660–677, 2022. 3

[5] Hanbyel Cho and Junmo Kim. Generative approach for probabilistic human mesh recovery using diffusion models. In *ICCVW*, pages 4183–4188, 2023. 3

[6] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *ECCV*, 2022. 3, 6

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 1, 5, 7

[8] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, 2022. 3, 6

[9] Enric Corona, Gerard Pons-Moll, Guillem Alenyà, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *ECCV*, 2022. 3

[10] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, et al. Learning analytical posterior probability for human mesh recovery. In *CVPR*, pages 8781–8791, 2023. 3, 6

[11] Lin Geng Foo, Jia Gong, Hossein Rahmani, and Jun Liu. Distribution-aligned diffusion for human mesh recovery. In *ICCV*, pages 9221–9232, 2023. 1, 3, 6, 8

[12] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, et al. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, pages 13041–13051, 2023. 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NIPS*, pages 6840–6851, 2020. 3

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2014. 5

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3, 5, 6, 7

[16] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, pages 1715–1725, 2022. 3, 6

[17] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, et al. Sampling is matter: Point-guided 3d human mesh reconstruction. In *CVPR*, pages 12880–12889, 2023. 1, 3, 6

[18] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, pages 11127–11137, 2021. 1, 6

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[20] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, pages 11605–11614, 2021. 1, 6

[21] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, et al. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 6

[22] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, et al. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 3, 4, 6

[23] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, pages 590–606, 2022. 3, 6, 7

[24] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 3, 5, 6

[25] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021. 3, 5, 6, 7

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, et al. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6), 2015. 1

[28] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, et al. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI*, pages 2269–2276, 2021. 3, 6, 7

[29] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers. In *CVPR*, pages 534–543, 2023. 1, 3, 5, 6, 7

[30] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, et al. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 6

[31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, et al. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1

[32] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *CVPR*, pages 18434–18443, 2023. 3

[33] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M. Patel. Ambiguous medical image segmentation using diffusion models. In *CVPR*, pages 11536–11546, 2023. 3

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[36] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, pages 8011–8021, 2023. 3

[37] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, et al. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *ICCV*, pages 14761–14771, 2023. 3

[38] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*, pages 702–720, 2022. 3

[39] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, et al. Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation. In *CVPR*, pages 574–584, 2023. 3, 6

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 8

[41] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., et al. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 3, 6

[42] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *IEEE TPAMI*, pages 1–20, 2023. 1

[43] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, et al. Learning from synthetic humans. In *CVPR*, 2017. 5

[44] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 5

[45] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2021. 3

[46] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, et al. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *ICCV*, pages 3925–3935, 2023. 6

[47] Sen Yang, Wen Heng, Gang Liu, Guozhong Luo, Wankou Yang, et al. Capturing the motion of every joint: 3d human pose and shape estimation with independent tokens. In *ICLR*, 2023. 3

[48] Chun-Han Yao, Jimei Yang, Duygu Ceylan, Yi Zhou, Yang Zhou, et al. Learning visibility for robust dense human body estimation. In *ECCV*, pages 412–428, 2022. 3, 6

[49] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *CVPR*, pages 17006–17015, 2023. 3, 6

[50] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *CVPR*, 2018. 3

[51] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, et al. Thundr: Transformer-based 3d human reconstruction with markers. In *ICCV*, pages 12971–12980, 2021. 3, 6

[52] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, et al. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, pages 11446–11456, 2021. 3, 4, 6

[53] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3d human shape and pose from dense body parts. *IEEE TPAMI*, 44(5):2610–2627, 2022. 7

[54] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, et al. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *ICCV*, pages 7989–8000, 2023. 3

[55] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *CVPR*, pages 1611–1620, 2023. 3