Supplementary Material for "PostureHMR: Posture Transformation for 3D Human Mesh Recovery"

1. More Ablation Experiments

The inference speed. To further understand the performance of our method in terms of inference time, we conducted the following additional comparative experiments. Specifically, we infer the same image multiple times on the same RTX 3090 GPU to get the average speed, and the compared method distribution is HMDiff [3] and VirtualMarker [7]. HMDiff is a human mesh reconstruction algorithm based on the diffusion model, and VirtualMarker is a vertex regression method. As shown in Table 1, our method achieves competitive results against diffusion model-like methods, but is still slower than VirtualMarker, mainly due to multiple iterations.

Method	PostureHMR	HMDiff	VirtualMarker
FPS	19	18	24

Table 1. Inference speed comparison at RTX 3090 GPU.

Impact of the number of different iteration steps. To further illustrate the selection of PostureHMR steps, we conducted additional ablation experiments, changed the total number of iteration steps and gave MPVE test results on the 3DPW and SURREAL data sets. From Table 2, we can observe that when the number of steps increases, the quality of the reconstructed human mesh increases, showing smaller improvement when increasing to 10. Therefore, we set the number of iteration steps to 10 in the experiment.

Dataset	3DPW				SURREAL			
Steps	1	5	10	15	1	5	10	15
MPVE	78.2	76.5	75.4	75.3	44.5	43.0	42.1	42.1

Table 2. Evaluation result of hyperparameter steps k

Fair comparison. During the experiment, different methods may adopt different training strategies, making the experimental comparison potentially unfair. To further verify the effectiveness of the paper's method, we first list some methods for comparison, as shown in Table 3. These methods use the same backbone (HRNet-w48), fine-tuned on 3DPW and approximate data set selection. Experimental results show that our method achieves SOTA results.

In addition, considering the differences in the selection of the comparative experimental data sets mentioned above, we conducted a fairer comparison only using the 3DPW dataset. The experimental results are shown in Table 4.

Method	Dataset	MPVE
HMDiff	COCO and MPII-Pseudo SMPL, 3DHP, UP-3D, H3.6M	82.4
CLIFF [5]	COCO and MPII-Pseudo SMPL, 3DHP, H3.6M	81.2
VirtualMarker	COCO-Pseudo SMPL, 3DHP, UP-3D, H3.6M	77.9
Zolly [9]	COCO, 3DHP, PDHuman, HuMMan, LSPET, H3.6M	76.3
PostureHMR	COCO-Pseudo SMPL, 3DHP, UP-3D, H3.6M	75.4

Table 3. Evaluation results on 3DPW test set.

Method	MPVE	MPJPE	PA-MPJPE
HybrIK [4]	114.5	91.1	67.2
FastMETRO [2]	152.8	136.1	92.5
VirtualMarker	101.32	87.74	54.77
PostureHMR	98.5	84.9	52.2

Table 4. Evaluation results on 3DPW test set.

Impact of feature fusion design. We conducted ablation experiments on the feature fusion method on the SURREAL dataset, and cross-attention [1] fusion of 2D and 3D mesh vertex features. As shown in Table 5, the difference between the two implementation methods is small.

Method	MPVE	MPJPE	PA-MPJPE
Cross-attention	42.4	35.2	27.4
PostureHMR	42.1	35.3	27.4

Table 5. Evaluation results on the SURREAL test set.

Impact of coarse to fine strategy. To further verify the selection of the mesh upsampling strategy, we conducted ablation experiments. As shown in Table 6, the comparison method learns a matrix to achieve interpolation from coarse vertices to fine mesh. Since both are essentially linear layers based on MLP, there is no major difference in the experimental results.

Method	MPVE	MPJPE	PA-MPJPE
Matrix learning	42.3	35.5	27.7
PostureHMR	42.1	35.3	27.4

Table 6. Evaluation results on the SURREAL test set.

More visualization results. Fig. 1 gives a visual comparison of our method with the parametric regression-based CLIFF and vertex regression-based VirtualMarker (VM) methods. Our method outperforms the other two methods

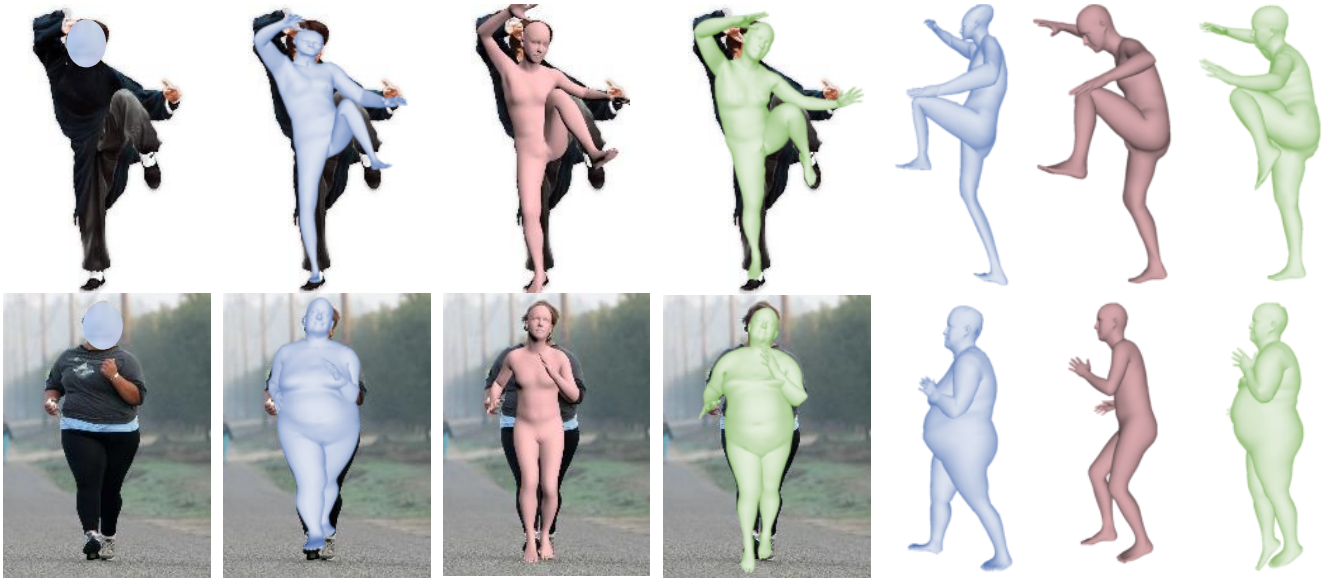


Figure 1. Qualitative comparison on in-the-wild images.

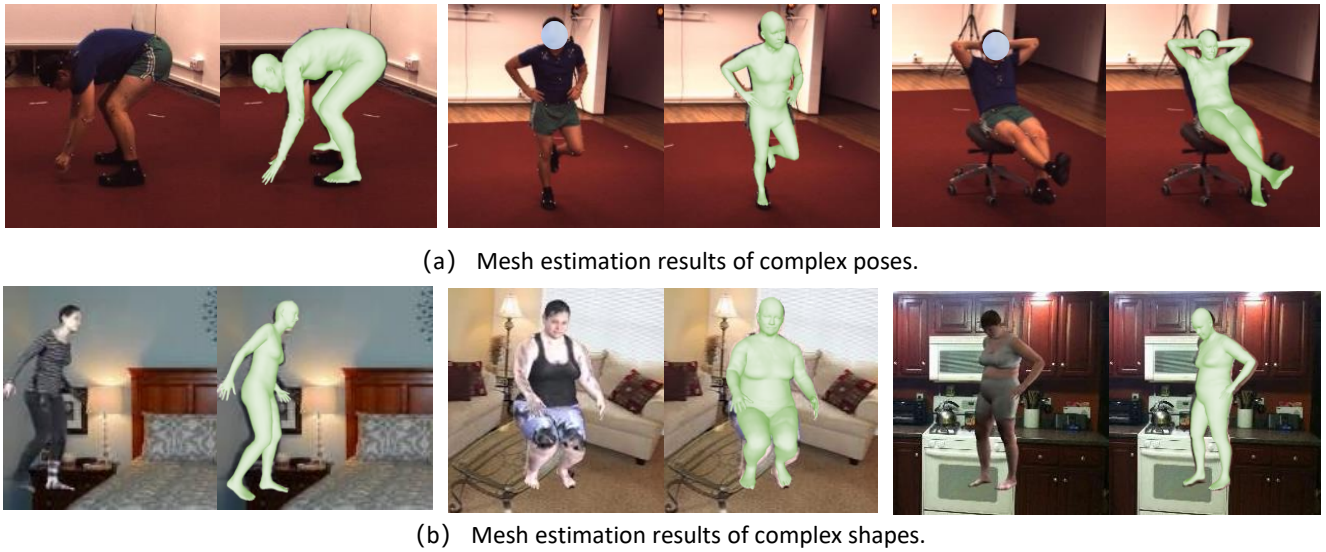


Figure 2. More visualization of human mesh outputs using our method.

under challenging pose and shape inputs. In addition, we give more visualization results in Fig. 2. The first row shows that our algorithm can achieve better modeling results in challenging poses, and the second row gives more complex shape situations.

2. More Discussion of Forward Pass

2.1. Preliminary: SMPL model.

The Skinned Multi-Person Linear model (SMPL) model was proposed in [6], which is a general human body statistical model. Given an initial pose mesh with 6890 vertices, different human bodies can be described through shape parameters (β) and pose parameters (θ). Shape parameters include 10 values describing the shape of the character, such as height, fatness, etc. The pose parameters are defined by

23 joint points and 1 root orientation, with a total of 72 parameters.

For an initial state SMPL model:

$$\mathcal{M}(\beta_0, \theta_0) = T \tag{1}$$

Given the control parameters (β_t, θ_t) input at time t, the new mesh can be summarized as obtained through blend shape and linear blend skinning.

Blend shape. It mainly converts the influence of shape and pose on the initial model into a linear offset. The reason why pose parameters also affect the initial model is because the local expression of the human body is different in different poses. For example, the appearance of the belly is different when sitting and standing. This implementation process is:

$$T_b = T + B_s(\beta_t) + B_p(\theta_t) \tag{2}$$

where $B_s(\cdot)$ and $B_p(\cdot)$ refer to the output offset relative to the vertex by specifying parameters after model training.

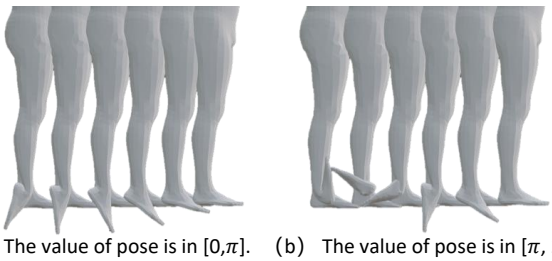
Blend skinning. The previous calculation mainly designed the changes of mesh vertices in the initial posture. Next, skinning calculation is required. The so-called skinning is to drive the vertex changes by moving the key points of the bones. The direct relationship between a vertex and a bone key point becomes a weight. The closer a vertex is to a specific bone point, the stronger its influence will be following changes such as rotation/translation of the bone point. The final mesh output can be achieved through the following:

$$\mathcal{M}(\beta_t, \theta_t) = W(T_b(\beta_t, \theta_t), J(\beta_t), \mathcal{W}) \tag{3}$$

where $J(\cdot)$ is the rest pose joints calculation function and $\mathcal{W}$ is the blend weights. For the specific implementation of blend shape and blend skinning, please refer to [6].

2.2. Limitation of linear interpolation.

Our method's implementation of the forward process mainly performs a simple linear process on the parameters input by SMPL, which may produce non-anthropomorphic postures. However, except for SURREAL, the angles of pose rotation in other datasets are in the range [0, π]. Therefore, most linear interpolation results are anthropomorphic. To intuitively understand this problem, we give two sets of visualization examples in Fig. 3. When the value range exceeds [0, π], the interpolation result will produce the result of the footsteps falling into the body. This problem can be solved using algorithms that reasonably interpolate poses, such as Pose-ndf [8], which is another potential solution that we will explore in future work.



(a) The value of pose is in [0,π]. (b) The value of pose is in [π, 2π].

Figure 3. Visualization of interpolation for different pose values.

References

[1] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *ICCV*, 2021. 1

[2] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *ECCV*, 2022. 1

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

[4] 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. 1

[5] 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. 1

[6] 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. 2, 3

[7] 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

[8] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, et al. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *ECCV*, 2022. 3

[9] 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. 1