Supplementary Materials

1. Notation Table

The symbols used in this paper are summarized in Table 1.

Symbol	Description
I	The input image
f_{enc}	The extracted feature from CNN and transformer encoder
f_{2D}	The feature for generating 2D hypotheses
f_{3D}	The feature for generating 3D hypotheses
f_{hmr}	The feature for generating SMPL parameters
z_{2D}	The sampled latent feature in 2D HGM
z_{3D}	The sampled latent feature in 3D HGM
μ_{2D}	Predicted mean for the 2D pose hypotheses distribution
σ_{2D}	Predicted standard deviation for the 2D pose hypotheses distribution
μ_{3D}	Predicted mean for the 3D pose hypotheses distribution
σ_{3D}	Predicted standard deviation for the 3D pose hypotheses distribution
H_{2D_heat}	2D pose hypotheses in the form of 2D heatmap
H_{3D}	3D pose hypotheses in the form of 3D coordinates
$H_{3D->2D}$	Projected 2D coordinates from H_{3D}
H'_{2D_heat}	2D heatmap after making Gaussian blobs on $H_{3D->2D}$
$f_{2D\_Sampled}$	The sampled feature by H_{2D_heat}
$f_{3D\_Sdampled}$	The sampled feature by H'_{2D_heat}
θ	Pose parameters of SMPL [4]
β	Shape parameters of SMPL [4]
M	The final mesh
J_{3D}	The body joints regressed from M
K	Number of hypotheses
N_J	Number of body joints

Table 1. A list of the symbols used in the main manuscript.

2. Hyperparameters

The hyperparameters adopted for training our framework are summarized in Table 2.

Table 2. A summary of the hyperparameters used in our framework.

Hyperparameter Settings	
Learning Rate	10^{-4}
Weight Decay Factor	10^{-4}
Betas	$(0.9, 0.999)$
Batch Size	64
Image Crop Size	224×224
Epochs	60
N_J	24
k	81
λ_{hmr}	60
λ_{pose}	5
λ_{smpl}	1
λ_{3D}	300
λ_{2D}	200
λ_{reg}	10
Hardware Settings	
CPU	AMD Ryzen™ 9 5950X
Main Memory	128GB
GPU	NVIDIA RTX 3090
GPU Memory	24GB

3. Visualization of Hypotheses

In this section, we present additional visualizations of 2D and 3D pose hypotheses, as depicted in Fig. 1. For the 2D poses, we visualize the keypoints for the Right Ankle, Right Knee, Right Hip, Left Hip, Left Knee, and Left Ankle. For the 3D poses, we focus on the left and right wrists and the left and right ankles.

4. Visualization of Additional Qualitative Results

Fig. 2 presents the additional qualitative results of both FastMETRO [1] and our method on the Human3.6M and 3DPW datasets.

5. A Guideline for Reproduction

Our implementation follows the previous Transformer-based methods [1–3]. For detailed information and code, please refer to our anonymous repository: https://anonymous.4open.science/r/Progressive-Hypothesis-Transformer-for-3D-Human-Mesh-Recovery.

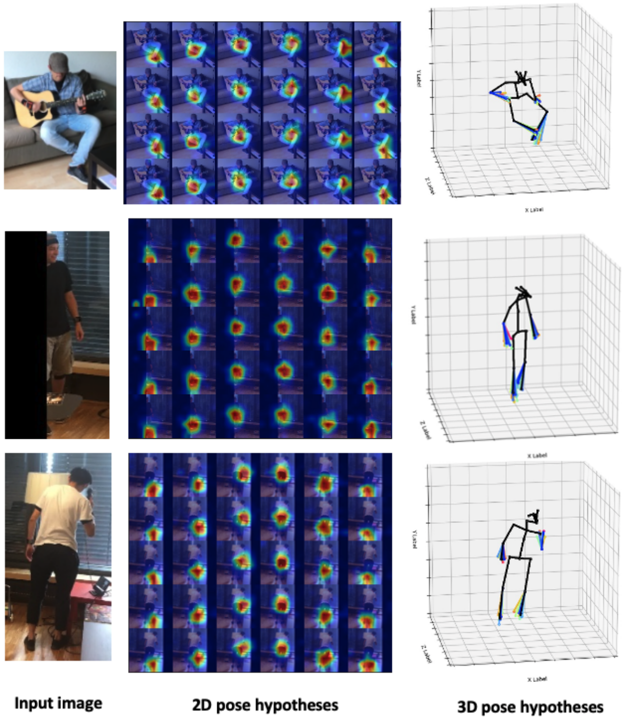

Figure 1. Visualization of our 2D/3D hypotheses. For 2D poses, we visualize the hypotheses of Right Ankle, Right Knee, Right Hip, Left Hip, Left Knee, and Left Ankle (from left to right on the image). For 3D poses, we visualize left and right wrists, as well as left and right ankles. Different colors represent different hypotheses.

References

[1] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 342–359, 2022. 1, 2

[2] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 1

[3] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 12939–12948, 2021. 1

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

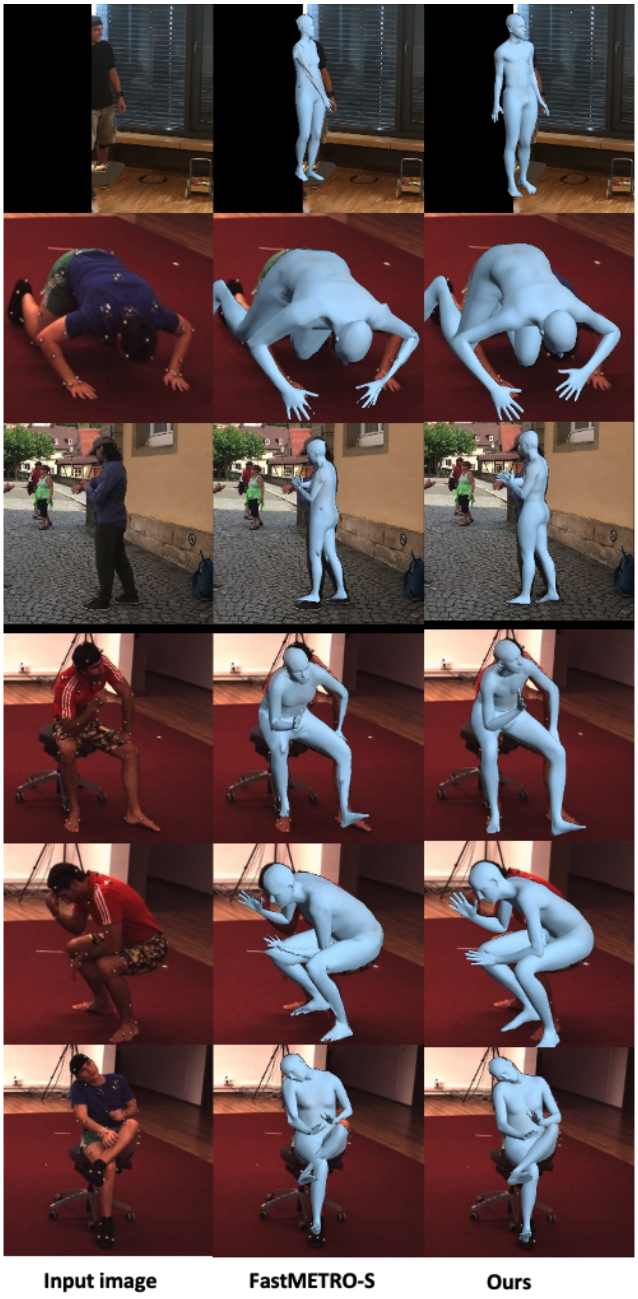

Figure 2. Qualitative comparisons of FastMETRO [1] and the proposed method on the Human3.6M and 3DPW datasets. Please note that our model size is comparable to that of FastMETRO-S.