Supplementary Material: Learning Human Mesh Recovery in 3D Scenes

Zehong Shen Zhi Cen Sida Peng Qing Shuai Hujun Bao Xiaowei Zhou†
State Key Lab of CAD&CG, Zhejiang University

1. Implementation Details

1.1. Network Architecture

Root and Contact Module. We use HRNet [1] as the CNN backbone, which is initialized with METRO [2] pretrained weights. Specifically, the backbone extracts stage-4 features $f \in \mathbb{R}^{56\times56\times64}$ from an input image $I \in \mathbb{R}^{224\times224\times3}$. We use a fully-connected layer to map f to initial 2D root heatmap $f_1 \in \mathbb{R}^{56\times56}$, normalized depth map $f_2 \in \mathbb{R}^{56\times56}$, and image feature map $f_3 \in \mathbb{R}^{56\times56\times32}$. During training, we use an extra fully-connected layer to map f to a 2D contact segmentation map $f_4 \in \mathbb{R}^{56\times56\times8}$.

We follow SMAP [3] to estimate the initial human root from f_1 and f_2. Using the initial root estimation, we select interest nearby scene points, voxelize these points, and construct 3D features $g \in \mathbb{R}^{N\times35}$, as described in the Sec.3.2 of the main paper. For $g \in \mathbb{R}^{N\times35}$, N indicates the voxel numbers, and feature dimension 35 consists of 3 for the vector representation and 32 for the unprojected f_3.

The Sparse 3D CNN builds upon the publicly available architecture[1] of SPVCNN [4], which consists of downsample and upsample modules with residual connections and point-wise transforms. We refer readers to [4] for more details. We add one fully-connected layer at the input to map g from 35 dimensions to 32. We modify the output fully-connected layer to output a feature dimension of 12, which consists of 1 for raw confidence c_i^{raw} and 3 for refined offset$_i$, and 8 for contact segmentation results.

We experimentally find a soft confidence weighting over all voxels gives the best performance,

$$c = Softmax(\sigma(c^{raw})) \tag{1}$$

where σ is sigmoid.

Following PROX [5], we use seven contact categories as segmentation labels, which is illustrated in Fig. 1.

Mesh Recovery Module. As illustrated in Fig. 2, we elaborate on "Figure 4" of the main paper with more details of the residual connection and positional embedding. The main module is a stack of three submodules that share the same structure. The output feature dimensions for each submodule are 512, 128, and 3. In Fig. 3, we give the definition of self-attention layer and cross-attention layer used in the parallel scene network. Specifically, self-attention is used to process scene point features. The cross-attention is used to fuse scene point features to vertex features.

[1] https://github.com/mit-han-lab/spvnas

Figure 1. **Definition of Contact Categories.** We use the definition from PROX [5]. The color of different contact regions aligns with the figures of the main paper.

1.2. Loss

We elaborate on the definition of the loss mentioned in Sec.3.4 of the main paper. The notation with $\bar{\cdot}$ indicates the groundtruth.

Root and Contact. L_{R2D} is the L2 of 2D root heatmap. We create the ground truth heatmap $\bar{f}_1$ by projecting the ground truth root 3D to the image plane and using a 3×3 gaussian kernel to smooth the target point.

$$L_{\text{R2D}} = \sum \|f_1 - \bar{f}_1\|_2 \tag{2}$$

L_{RZ} is the L1 of depth prediction. We use ground truth point $(\bar{x}, \bar{y})$ to pick the depth value from the predicted depth map and compute loss with the ground truth depth, which is

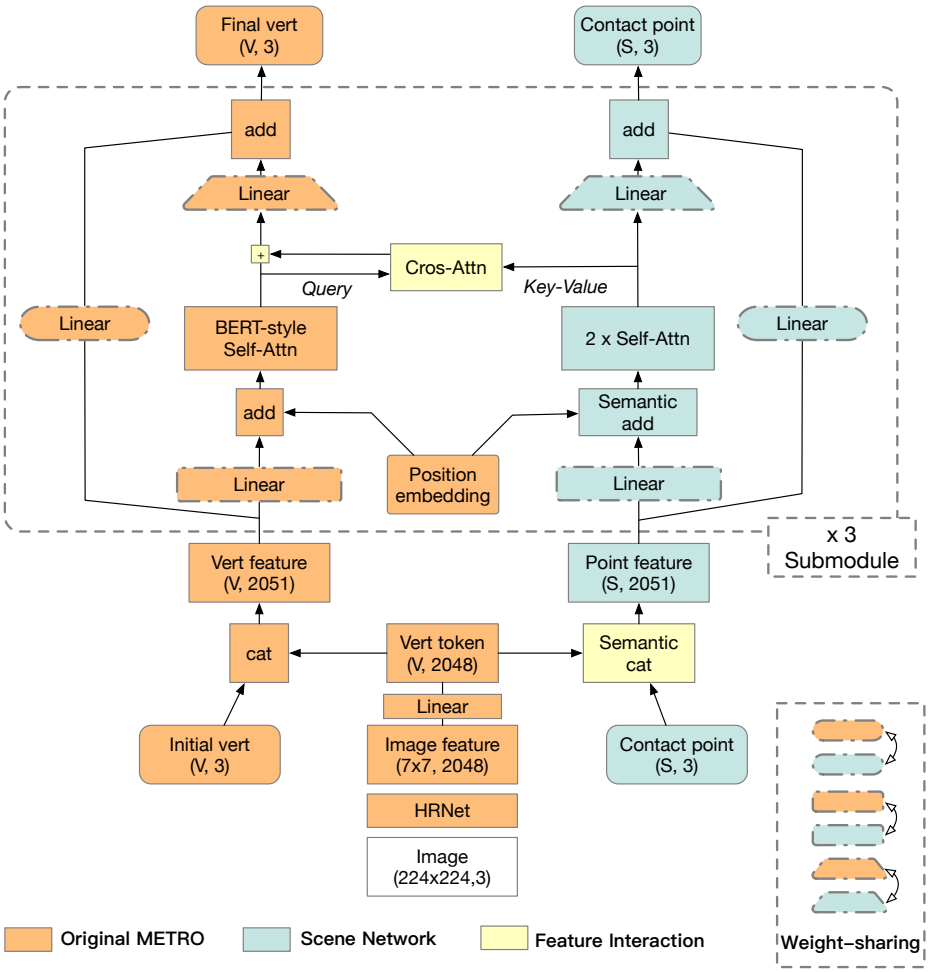

Figure 2. **Mesh Recovery Module.** We give the network design details, including residual connection and positional embedding. The semantic-add works similarly to the semantic-cat, as described in the main paper.

normalized by focal length and image size.

$$L_{\text{RZ}} = \sum \|f_2(\bar{x}, \bar{y}) - \bar{Z}\|_1 \tag{3}$$

L_{ROV} is the L1 of offset vector difference. We compute ground truth offset vector $\bar{o}_i$ for each voxel point.

$$L_{\text{ROV}} = \sum_i \|o_i^* - \bar{o}_i\|_1 \tag{4}$$

L_{R3D} is the L1 of refined root position. Since the refined root r^* is a confidence-based weighted average of all prediction, this loss, together with L_{ROV}, encourage the network to learn confidence.

$$L_{\text{R3D}} = \|r^* - \bar{r}_{3D}\|_1 \tag{5}$$

L_{C} is the cross-entropy loss for N voxel points and the MSE loss for 2D contact map, which is an auxiliary training task,

$$L_{\text{C}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} log\frac{exp(x_{n,c})}{\sum_{i=1}^{C} exp(x_{n,i})} y_{n,c} + \sum \|f_4 - \bar{f}_4\|_2 \tag{6}$$

Human Mesh Recovery. L_{V} is the mean L1 of translation-aligned vertex error.

$$L_{\text{V}} = \frac{1}{V}\sum_i^V \|v_i - \bar{v}_i\|_1 \tag{7}$$

L_{J} is the mean L1 of translation-aligned joint 3D error. We use 14 joints of H36M following METRO [2].

$$L_{\text{J}} = \frac{1}{J}\sum_i^J \|j_i - \bar{j}_i\|_1 \tag{8}$$

L_{CP} is the mean L1 of the input scene points p and out-

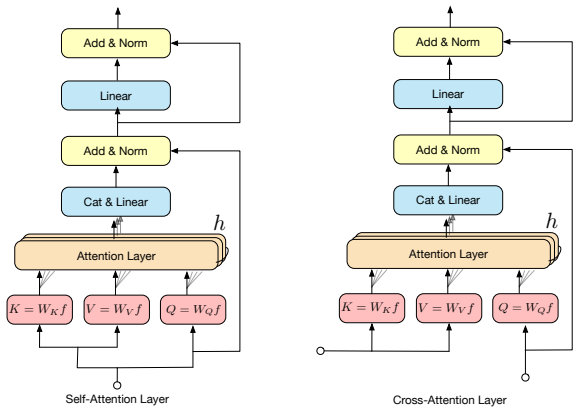

Figure 3. **Self and Cross Attention Layers** share the same core network architecture. The difference is at the input, where self-attention fuses one input feature, and cross-attention fuses two input features.

method	Penetration↓	Reprojection↓
PROX (RGBD) [5]	7.35	19.36
PROX+HuMoR (RGBD) [6]	3.15	10.89

Table 1. **Comparison of the pseudo ground-truth in PROX qualitative.** We re-generate the pseudo ground-truth for the PROX [5] qualitative dataset with PROX+HuMoR [6] using the RGBD input. We report the penetration and reprojection error compared to the originally released PROX results.

put scene points $\tilde{p}$.

$$L_{CP} = \frac{1}{N}\sum_{i}^{N} \|p_i - \tilde{p}_i\|_1 \tag{9}$$

L_{GV} is the mean L1 of vertex error in global coordinates, which is the camera coordinates in our implementation.

$$L_{GV} = \frac{1}{V}\sum_{i}^{V} \|v_i^g - \bar{v}_i^g\|_1 \tag{10}$$

1.3. Dataset Preparation

We combine HuMoR [6] and PROX [5] to re-generate the pseudo ground-truth. As shown in the 1, the penetration and reprojection error are all reduced. The penetration is the average over all penetrated vertices. The reprojection error is calculated on the 2D joints from openpose [7] with confidence higher than 0.75.

1.4. More ablations

Number of contact categories. Tab. 2 shows the ablation results for the predicted number of contact categories. The root and contact module classifies scene contacts as 1/432/8 categories, where the human mesh recovery module is adapted accordingly. The 432 stands for treating each downsampled vertex of the human mesh as an individual class.

Method	G-MPJPE↓	G-MPVE↓	PenE↓	ConFE↓	MPJPE↓	MPVE↓
Dataset GT [8]	/	/	9.8	10.8	/	/
METRO [2]†	511.7	509.7	33.6	37.6	98.8	107.9
SA-HMR-1	278.3	286.7	16.8	23.2	98.1	107.9
SA-HMR-432	275.6	283.4	15.8	22.7	96.2	105.5
SA-HMR-8	**264.6**	**272.7**	**14.9**	**19.0**	**93.9**	**103.0**

Table 2. **Ablation of number of contact categories on RICH dataset.**

2. Concluding Remarks

2.1. Limitation and Future Works

The current datasets [5, 8, 9] that contain images and scene-scans are limited in capacity, and capturing ground truth data is challenging due to occlusion. Additionally, the current data does not account for object deformation, and the groundtruth training labels still exhibit human-scene penetration, making it difficult for SA-HMR to effectively learn to estimate scene contacts. Synthetic data appears to be a promising direction to address these limitations.

In situations where there are inconsistencies between the pre-scanned scene and the image, our model tends to predict results that align with the pre-scanned scene, which produces better visualization but lower accuracy. In the future, we also plan to enhance the system by incorporating dynamic object detection as additional 3D cues.

2.2. Social Impact

Accurate and efficient mesh reconstruction in the scene has mostly positive use cases in VR/AR, games, and Human-Computer Interaction. However, we also see a possibility that our results could be used for fake video production with recent advances in SMPL-based neural body rendering. Being aware of this, we will make our models available only for research purposes.

References

[1] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *T-PAMI*, 2020. 1

[2] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1, 2, 3

[3] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, 2020. 1

[4] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*, 2020. 1

[5] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 1, 3

[6] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 3

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[8] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. 3

[9] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape, motion and social interactions from head-mounted devices. In *ECCV*, 2022. 3