Supplementary Material for Sampling is Matter: Point-guided 3D Human Mesh Reconstruction

Jeonghwan Kim[1*] Mi-Gyeong Gwon[1*] Hyunwoo Park[1]
Hyukmin Kwon[2] Gi-Mun Um[2] Wonjun Kim[1†]
[1]Konkuk University [2]Electronics and Telecommunications Research Institute
{jhkim0759,kmk3942,pzls,wonjkim}@konkuk.ac.kr {hmkwon,gmum}@etri.re.kr

1. Introduction

This supplementary material provides more descriptions of the proposed method as follows. First, the detailed architecture of the proposed method is presented in Section 2. In the following, implementation details of the proposed method are explained in Section 3. Finally, additional qualitative results and visual materials with the analysis follow in Section 4.

2. Architecture Details

The proposed method consists of two main parts: feature sampling and mesh regression. We clarify the configuration of hyper-parameters for each part in the following subsections.

2.1. Feature sampling

To sample vertex-relevant features in the embedding space, we adopt HRNet-W32 [19] as the backbone network, which is trained based on three 2D pose datasets [9, 12, 20] and one human detection dataset [16], by following [17]. The backbone feature X_b is decoded through heatmap, feature, and grid feature decoders, respectively (see Fig. 1). Specifically, heatmap and feature decoders take $X_b^1 \in \mathbb{R}^{C\times H\times W}$ (where $C = 32$ and $H = W = 56$), which is encoded through the top path of HRNet as their inputs as shown in Fig. 1. On the other hand, the input of the grid feature decoder is extracted from the bottom path of the backbone as $X_b^2 \in \mathbb{R}^{8C\times H/8\times W/8}$ to consider the global context of the input image. The detailed architecture of each decoder is provided in Table 1.

2.2. Mesh regression

Our sampled features are reshaped as the vertex token, i.e., $\hat{V} \in \mathbb{R}^{N\times D}$ where N and D are set to 431 and 512, respectively. Similarly, grid features are also input to the transformer encoder as the grid token $\hat{G} \in \mathbb{R}^{Z\times D}$ where Z is set to 49($= 7 \times 7$). On the other hand, the joint token $\hat{J} \in \mathbb{R}^{K\times D}$ is randomly initialized and optimized regardless of the feature sampling process. Here K is set to 14, which indicates the number of keypoints of the human body. Moreover, we leverage the vertex token to generate the camera token $\hat{T} \in \mathbb{R}^{1\times D}$ by a single linear layer. The camera token is updated to estimate the camera parameter which is utilized for projecting 3D joint positions onto the

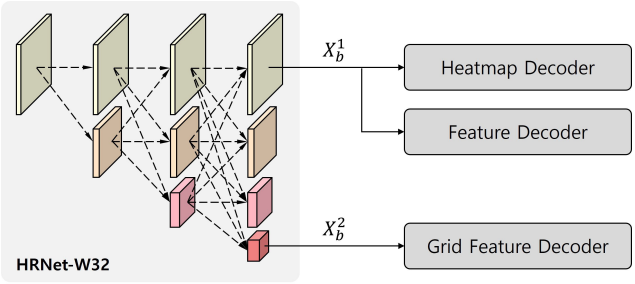

Figure 1. The detailed process by which the backbone features are input into each decoder. Note that we present a simplified version of the HRNet backbone in this figure for compact representation.

Decoder architecture			
Module	Layer type	Input dim.	Output dim.
Heatmap decoder	1×1 Conv.	32	431
	ResBlock [4]	431	431
	ResBlock [4]	431	431
Feature decoder	1×1 Conv.	32	512
	ResBlock [4]	512	512
	ResBlock [4]	512	512
Grid feature decoder	1×1 Conv.	256	512
	ResBlock [4]	512	512

Table 1. The detailed architecture of heatmap, feature, and grid feature decoders.

*equal contribution
†corresponding author

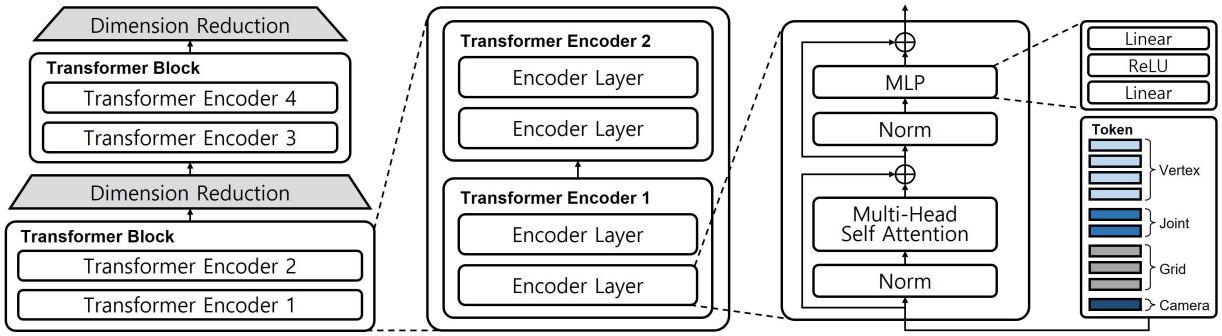

Figure 2. The detailed architecture of the sequence of transformer encoders in the proposed method.

2D space when calculating the 2D joint loss. Note that the camera token is implicitly optimized by using projected 2D joints. Such four types of tokens are updated through the sequence of transformer encoders to reconstruct the 3D human mesh. The detailed architecture is illustrated in Fig 2. Specifically, two transformer blocks, which consist of two transformer encoders respectively, are stacked with linear projectors for reducing the dimension of tokens from 512 to 128, and from 128 to 3 (corresponding to 3D coordinates). Each transformer encoder conducts multi-head self-attention where our progressive attention masking scheme is applied.

3. Implementation Details

In this Section, we provide detailed explanations about benchmark datasets that we adopted in this work, the optimization process of the vertex estimation, the computational costs of our proposed network, and the ablation study according to different settings of our masking scheme.

3.1. Datasets

Human3.6M [5] is a multi-view 3D pose dataset collected by using the motion capture system in the indoor environment. This dataset consists of videos taken with five female and six male subjects from 17 different scenarios. Our proposed method is trained using subjects S1, S5, S6, S7, and S8, and evaluated with subjects S9 and S11.

3DPW [18] is the most widely employed dataset for the 3D human pose estimation since it contains various real-world images with 3D pose and shape annotations. This dataset is composed of 60 video sequences recorded with seven individual subjects.

MuCo-3DHP [14] is constructed based on multiple-human scenes, which are synthesized by using the MPI-INF-3DHP [13] dataset, to consider strong inter-person occlusions in complex backgrounds.

UP-3D [8] consists of various outdoor images with 2D joint annotations and pseudo labels of the 3D human mesh, which are obtained by using [2].

Methods	# of Params	FPS	MPJPE	PA-MPJPE
METRO [10]	230.4M	19.55	54.0	36.7
MeshGraphormer [11]	226.5M	18.67	51.2	34.5
FastMETRO [3]	153.0M	21.88	52.2	33.7
Ours	59.1M	21.83	48.3	32.9

Table 2. Comparison of the computational costs with the performance for previous model-free methods based on the Human3.6M dataset. Note that FPS of each method is recomputed on our experimental environment for the fair comparison.

COCO [12] contains images taken under the real-life context, which are labeled with 2D joint annotations. We also employed the pseudo labels provided by [6] for the ground truth of the 3D human mesh.

MPII [1] is collected from YouTube videos. This dataset includes large amounts of images that capture diverse human activities and corresponding 2D joint annotations.

FreiHAND [21] is the first large-scale real-world dataset for 3D hand pose and shape estimation, which is annotated by high-quality labels with 21 hand keypoints. This dataset is used to demonstrate the generalization ability of our proposed method. Note that we use test-time augmentation for the performance evaluation of the proposed method.

3.2. Vertex Optimization

Since the SMPL model has been most widely adopted for 3D human representation, the total number of vertices constituting a full body mesh generally follows that of SMPL, i.e., 6,890. However, estimating such a large number of vertices at once probably causes the problem of redundancy in prediction due to the spatial locality of vertices as mentioned in [7]. Therefore, we first estimate sparse vertices and then expand them into dense vertices for efficient training just like previous model-free methods. More concretely, the ground truth of the vertex is compressed twice with a factor of 4 (i.e., 6,890$\rightarrow$1,732$\rightarrow$431), based on the downsampling technique introduced in [15]. After such 431 vertices are estimated from the sequence of transformer encoders, upsampling is performed in a reverse way (i.e., 431$\rightarrow$1,732$\rightarrow$6,890) to make a full body structure by using

Figure 3. More results of 3D human mesh reconstruction by the proposed method on the 3DPW dataset.

Figure 4. More results of 3D human mesh reconstruction by the proposed method on the COCO dataset.

Distance threshold for defining vertex connection (M)				MPJPE	PA-MPJPE
Enc1	Enc2	Enc3	Enc4		
✗	✗	✗	✗	50.9	33.3
✗	✗	3	1	50.7	33.8
✗	5	3	1	49.9	33.7
7	5	3	1	48.3	32.9

Table 3. Performance analysis of the proposed method according to the change of progressive attention masking based on the Human3.6M dataset. Note that Enc($\cdot$) indicates each transformer encoder shown in Fig. 2. The number in the left side (i.e., 1, 3, 5, and 7) denotes the distance threshold for computing self-attentions.

the pre-defined matrix U [15]. These three types of vertex estimation are respectively guided by each ground truth, which is generated during the downsampling process, via their own vertex loss. By doing this, the computational cost in terms of memory and time for training the network can be reduced as well.

3.3. Computational Costs

By giving the vertex-relevant information to the sequence of transformer encoders, the proposed method can efficiently learn to reconstruct the 3D human mesh even with low computational requirements. Table 2 shows the computational costs in terms of the number of parameters and the run-time speed (fps). As can be seen, a notably small number of trainable parameters is required for the proposed network compared to previous methods, which leads to reduction of the overall training time. For example, it takes five days to train the proposed network for 50 epochs with two NVIDIA RTX A6000 GPUs. Moreover, the proposed method shows the competitive processing speed while outperforming previous methods.

3.4. Ablation Study

To investigate the advantage of the proposed progressive attention masking scheme, we conduct more comparative experiments and corresponding results are shown in Table 3. As can be seen, the performance of MPJPE is improved as more masks are used in a progressive manner, while PA-MPJPE is slightly dropped when not all the encoders use masks, compared to the baseline model. It is noteworthy that the proposed method, i.e., the case when progressive attention masking is applied to all the transformer encoders, shows the best performance as shown in the bottom row of Table 3. Based on this result, it is thought that our progressive attention masking is helpful for improving the performance of 3D human mesh reconstruction.

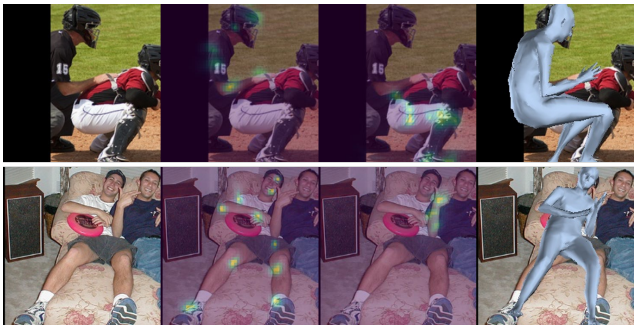

Figure 5. Some failure cases of the proposed method on the COCO dataset. From left to right: input images, visualization of the activated positions on the left person in the heatmap, visualization of the activated positions on the right person in the heatmap, and results of 3D human mesh reconstruction.

4. Additional Visual Materials

4.1. Qualitative Results

In this subsection, we provide the analysis of qualitative results by the proposed method and then discuss about the failure cases.

Analysis of qualitative results. Several examples of the 3D human mesh reconstruction for 3DPW [18] and COCO [12] are shown in Figs. 3 and 4, respectively. As can be seen, 3D meshes are accurately fitted to the target body region under real-world environments. In particular, the proposed method is robust to self-occlusions occurring in extreme poses, e.g., the first rows of Fig. 3 and 4, by effectively focusing on features sampled at each vertex point. Furthermore, since the proposed method also considers the local relation between vertices by progressive attention masking, it yields the reliable result of 3D human mesh reconstruction even in object-occluded situations as shown in the second rows of Fig. 3 and 4.

Discussion of failure cases. Some failure cases of the proposed method are shown in Fig. 5. As can be seen, the proposed method often suffers from ambiguities driven by severe inter-person occlusions. Since our method relies on the sampled feature, the incorrectly predicted heatmap (which is utilized for point-guided feature sampling) may adversely affect the result of 3D human mesh reconstruction. For example, we can see that the activated positions of the heatmap appear on the region of two different persons in a single frame, leading to generation of the invalid output mesh in Fig. 5. To cope with this limitation, our future work is to guide the network to distinguish individual persons in multi-person situations with extreme occlusion.

Figure 6. Several examples of predicted heatmaps and the corresponding results of 3D human mesh reconstruction by the proposed method on the 3DPW dataset. Note that the heatmaps are normalized before visualization. 1st column: input images. 2nd to 5th columns: visualization of predicted heatmaps. 6th column: ground truth. 7th column: results of 3D human mesh reconstruction.

4.2. Heatmap Visualization

Figure 6 shows several examples of predicted heatmaps and corresponding output meshes. Note that heatmaps for the same vertex are presented in the same columns. It can be seen that the heatmap decoder in our proposed method successfully estimates positions of projected vertices. In particular, the heatmap for the right hand is accurately predicted even with the severe occlusion (see the first example of the fifth column in Fig. 6). From these examples, we can see that such estimated heatmaps play a significant role to generate the 3D human body fitted to the target body region accurately as shown in the seventh column of Fig. 6.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3686–3693, 2014. 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Eur. Conf. Comput. Vis.*, pages 561–578, 2016. 2

[3] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *Eur. Conf. Comput. Vis.*, pages 342–359, 2022. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 1

[5] 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 Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 2

[6] 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 *Int. Conf. Comput. Vis.*, pages 2252–2261, 2019. 2

[7] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4501–4510, 2019. 2

[8] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6050–6059, 2017. 2

[9] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10863–10872, 2019. 1

[10] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1954–1963, 2021. 2

[11] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Int. Conf. Comput. Vis.*, pages 12939–12948, 2021. 2

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 1, 2, 4

[13] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Int. Conf. 3D Vis.*, pages 506–516, 2017. 2

[14] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *Int. Conf. 3D Vis.*, pages 120–130, 2018. 2

[15] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *Eur. Conf. Comput. Vis.*, pages 704–720, 2018. 2, 4

[16] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. CrowdHuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1

[17] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Int. Conf. Comput. Vis.*, pages 11179–11188, 2021. 1

[18] 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 *Eur. Conf. Comput. Vis.*, pages 601–617, 2018. 2, 4

[19] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, 2021. 1

[20] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *Int. Conf. Multimedia and Expo*, pages 1480–1485, 2019. 1

[21] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 813–822, 2019. 2