Supplementary Material *for* Learnable Human Mesh Triangulation for 3D Human Pose and Shape Estimation

Sungho Chun[1] Sungbum Park[2] Ju Yong Chang[1]
[1]Dept of ECE, Kwangwoon University, Korea [2]NCSOFT, Korea
{asw9161,jychang}@kw.ac.kr, spark0916@ncsoft.com

S1. Introduction

In the supplementary material, we first provide the detailed architecture of the vertex regression module. We then give the additional results of the 54 vertices model, single-view method in the visibility module, and multi-view inconsistency. Also, we present qualitative comparison of surface fitting and joint fitting methods.

S2. Vertex Regression Module

The vertex regression module consists of basic convolution blocks, residual blocks, downsample blocks, upsample blocks, and a $1 \times 1 \times 1$ convolution layer. The basic convolution block consists of a 3D convolution layer, a batch normalization layer, and a ReLU activation function. The residual block contains two 3D convolution layers, two batch normalization layers, two ReLU activation functions, and a residual connection. The downsample block consists of a 3D max pooling layer with a stride of 2. The upsample block consists of a 3D deconvolution layer with a stride of 2, a batch normalization layer, and a ReLU activation function. The vertex regression module is constructed using 3 basic convolution blocks, 20 residual blocks, 5 downsample blocks, 5 upsample blocks, and a $1 \times 1 \times 1$ convolution. Fig. S1 shows the detailed structure of the vertex regression module.

S3. 54 Vertices Model

The 54 vertices model shows worse MPVE and angular distance performance compared to other sub-vertices models because the number of vertices on the arms and hands is relatively small. Too few vertices do not provide enough information to resolve ambiguity in joint rotation and shape estimation. Consequently, the 54 vertices model results in higher wrist rotation errors than the 108 vertices model, which is presented in Table S1. A visualization of the vertex positions of the 54 vertices model and other sub-vertices models is presented in Fig. S2.

S4. Single-view Method for Visibility

In this section, we present justification for the use of I2L-MeshNet [6] in the proposed visibility module. To this end, we construct three visibility modules by combining three state-of-the-art methods for single-view human mesh reconstruction (i.e., I2L-MeshNet, METRO [3], and Graphormer [4]) and a visibility computation algorithm [1]. A detailed procedure for visibility estimation based on single-view mesh reconstruction is presented in Sec. 3.2 of the main paper. Table S2 shows the performance comparison for the cases in which three visibility modules are used in the proposed method. We found that using I2L-MeshNet produces better results than using other methods. Based on this result, we adopt I2L-MeshNet in our visibility module.

S5. Multi-view Inconsistency

This section presents additional examples on multi-view inconsistency in the ablation experiments of the main paper. Fig. S3 gives a scenario where the left hand is invisible due to occlusion. In the second view, it is difficult to determine the exact position of the left hand because the subject's left hand is not visible. However, in the remaining views, the position of the subject's left hand can be easily found. Therefore, in order to reconstruct the left-hand mesh, the model is desirable to have a higher dependence on the features obtained from the remaining views other than the second view. However, the softmax baseline has a relatively high dependence on the features obtained from the second view. As a result, the softmax baseline incorrectly reconstructs the left hand. However, LMT reduces the dependence on the features obtained from the second view and successfully reconstructs the human mesh.

Fig. S4 shows a scenario where the subject's right foot cannot be seen well in the second view. According to the results, the softmax baseline fails to reconstruct the mesh, but LMT reconstructs it successfully. Similar to the case of Fig. S3, it can be seen that the use of visibility reduces the

[1]https://github.com/MPI-IS/mesh

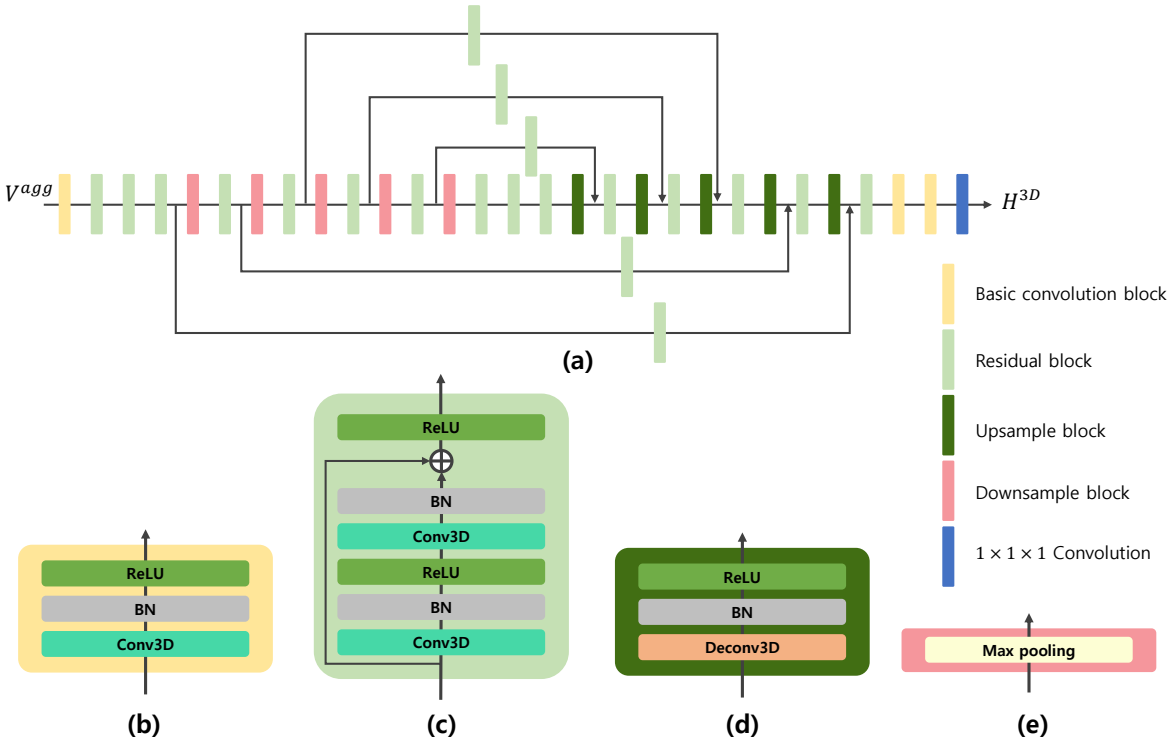

Figure S1: **The architecture of the vertex regression module.** (a) Pipeline of the vertex regression module. (b) Basic convolution block. (c) Residual block. (d) Upsample block. (e) Downsample block.

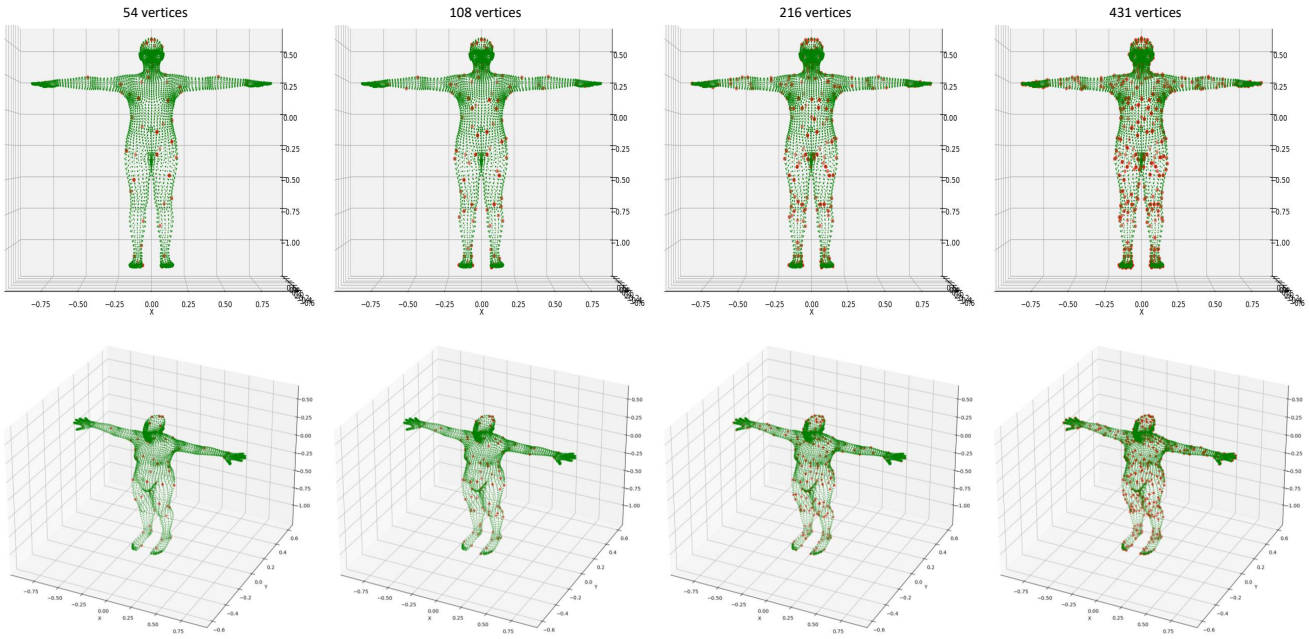

Figure S2: **Visualization of the vertex positions of sub-vertices models.** Green and red point sets denote full-vertices and sub-vertices, respectively.

dependence on the feature obtained from the second view where occlusion occurs.

Angular ↓	pelvis	L-hip	R-hip	torso	L-knee	R-knee	spine	L-ankl	R-ankl	chest	neck	L-thrx	R-thrx	head	L-shld	R-shld	L-elbw	R-elbw	L-wrst	R-wrst
108	5.09	**5.75**	**5.89**	**5.80**	**5.71**	5.75	**5.55**	**8.32**	9.88	**4.59**	13.31	**9.71**	**10.49**	11.11	12.69	**14.66**	13.75	**11.72**	**19.82**	**20.94**
54	**5.04**	6.21	6.30	6.23	6.09	**5.57**	5.85	8.47	**9.69**	4.60	**12.89**	10.17	10.59	**10.70**	**12.10**	15.10	**13.58**	12.57	25.61	24.22

Table S1: **Per-joint rotation error comparison of the 108-vertices and 54-vertices models.** 3D heatmaps with $16 \times 16 \times 16$ resolution are used in both experiments in this table.

Single-view method	MPJPE ↓	MPVE ↓	Angular ↓
I2L-MeshNet [6]	**17.59**	**23.70**	**11.33**
METRO [3]	18.15	23.98	11.55
Graphormer [4]	17.77	24.23	11.52

Table S2: **Ablation results on the single-view mesh reconstruction method in the visibility module.**

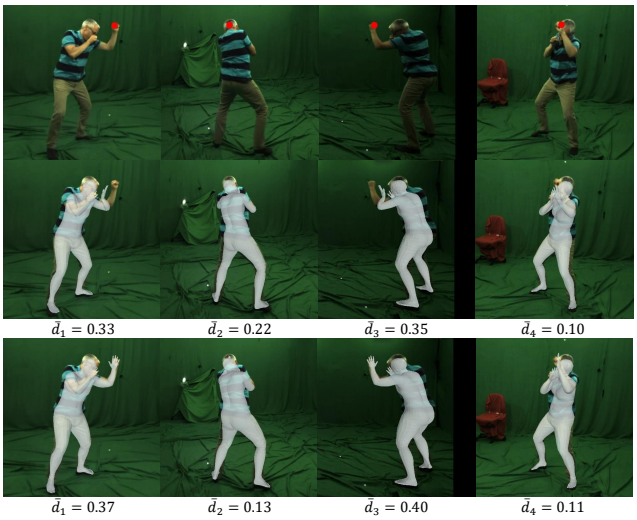

Figure S3: The first row visualizes the input multi-view images. The second and third rows show the reconstructed meshes generated from the softmax baseline and LMT, respectively.

S6. Qualitative Results

This section shows that our surface fitting produces qualitatively better results in terms of joint rotation and shape compared to joint fitting [2, 7]. Fig. S5 shows the human meshes reconstructed by LT-fitting and LMT on the Human3.6M [1] dataset. The second row of Fig. S5 shows that LT-fitting incorrectly predicts the rotations of the left ankle, elbows, and wrists. The fifth row of Fig. S5 shows that LT-fitting incorrectly reconstructs the right knee rotation and human shape. However, in both cases, LMT accurately predicts joint rotation and human shape.

Fig. S6 shows the human meshes reconstructed by LT-fitting and LMT on the MPI-INF-3DHP [5] dataset. The second row of Fig. S6 shows that LT-fitting incorrectly predicts the rotations of the right shoulder, elbows, wrists, knees, and ankles. The fifth row of Fig. S6 shows that LT-fitting incorrectly predicts the rotations of the neck, wrists, elbows, and right knee. However, similar to the results

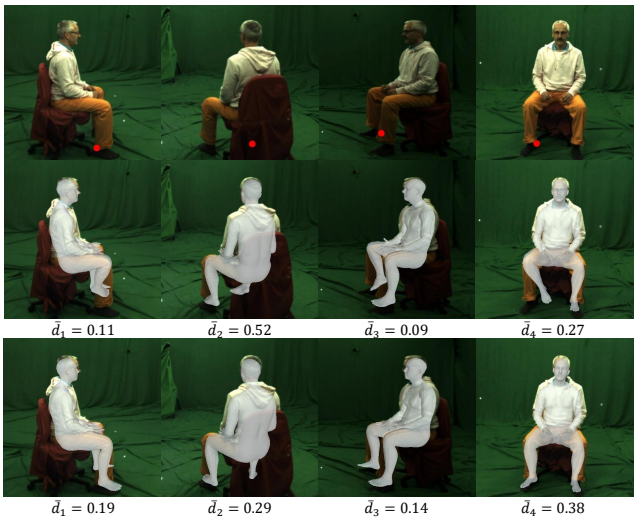

Figure S4: The first row visualizes the input multi-view images. The second and third rows show the reconstructed meshes generated from the softmax baseline and LMT, respectively.

on Human3.6M, LMT accurately predicts joint rotations in both cases. As can be seen from Figs. S5 and S6, the human mesh reconstructed with accurate joint rotation and shape information can explain the human body more naturally, and we qualitatively prove the superiority of surface fitting based on these results.

References

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

[2] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *ICCV*, 2019.

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

[4] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021.

[5] 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 *3DV*, 2017.

[6] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020.

[7] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Light-weight multi-person total capture using sparse multi-view cameras. In *ICCV*, 2021.

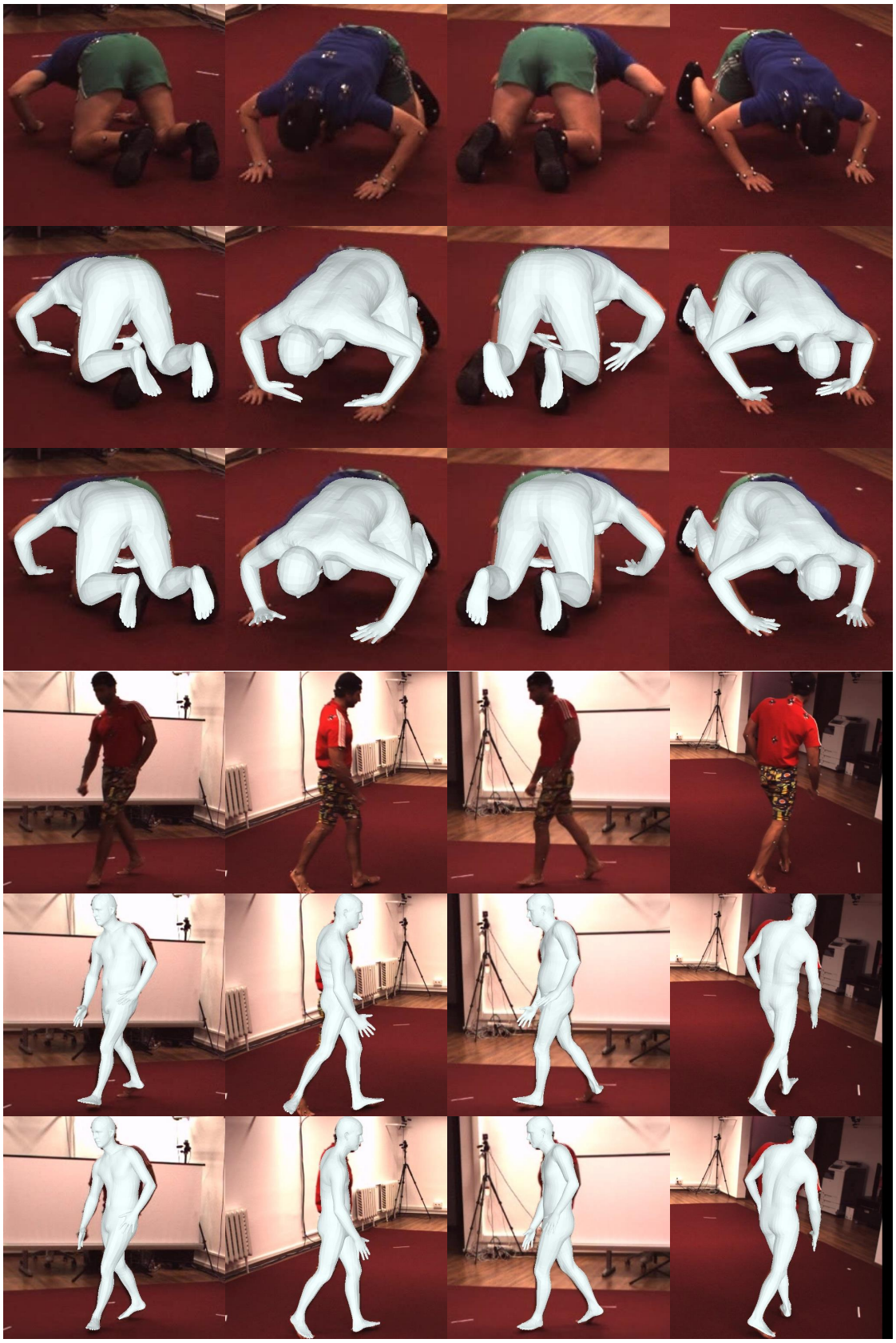

Figure S5: **Qualitative results on Human3.6M.** The first and fourth rows show the input images. The second and fifth rows visualize the human meshes reconstructed by LT-fitting. And the third and sixth rows visualize the human meshes reconstructed by LMT.

Figure S6: **Qualitative results on MPI-INF-3DHP.** The first and fourth rows show the input images. The second and fifth rows visualize the human meshes reconstructed by LT-fitting. And the third and sixth rows visualize the human meshes reconstructed by LMT.