Supplementary Material of "Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose"

Hongsuk Choi*, Gyeongsik Moon*, and Kyoung Mu Lee

ECE & ASRI, Seoul National University, Korea
{redarknight,mks0601,kyoungmu}@snu.ac.kr

In this supplementary material, we present more experimental results that could not be included in the main manuscript due to the lack of space.

1 Qualitative results

1.1 Shape recovery

We trained and tested Pose2Mesh on SURREAL dataset [20], which have various samples in terms of the body shape, to verify the capability of shape recovery. As shown in Figure 1, Pose2Mesh can recover a 3D body shape corresponding to an input image, though not perfectly. The shape features of individuals, such as the bone length ratio and fatness, are expressed in the outputs of Pose2Mesh. This implies that the information embedded in joint locations (*e.g.* the distance between hip joints) carries a certain amount of shape cue.

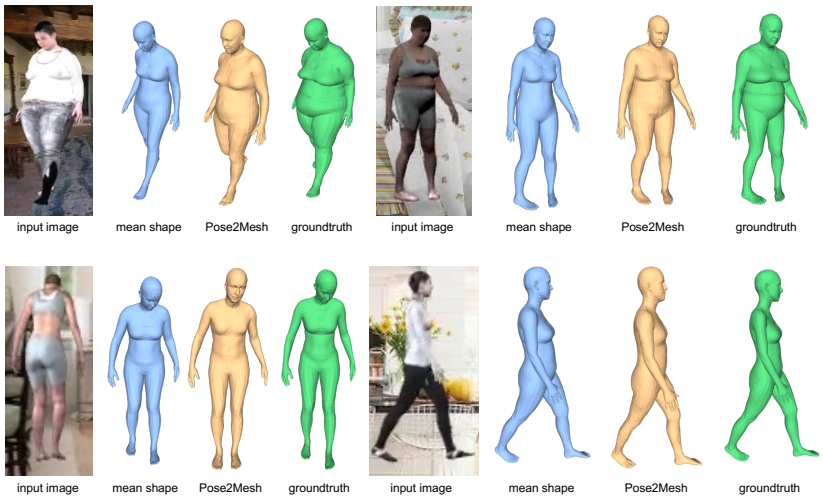

Fig. 1. The Pose2Mesh predictions compared with the groundtruth mesh, and the mesh decoded from groundtruth pose parameters and the mean shape parameters.

*equal contribution

1.2 Additional results

Here, we present more qualitative results on COCO [7] validation set and FreiHAND [21] test set in Figure 2. The images at the fourth row show some of the failure cases. Although the people on the first and second images appear to be overweight, the predicted meshes seem to be closer to the average shape. The right arm pose of the mesh in the third column is bent, though it appears straight.

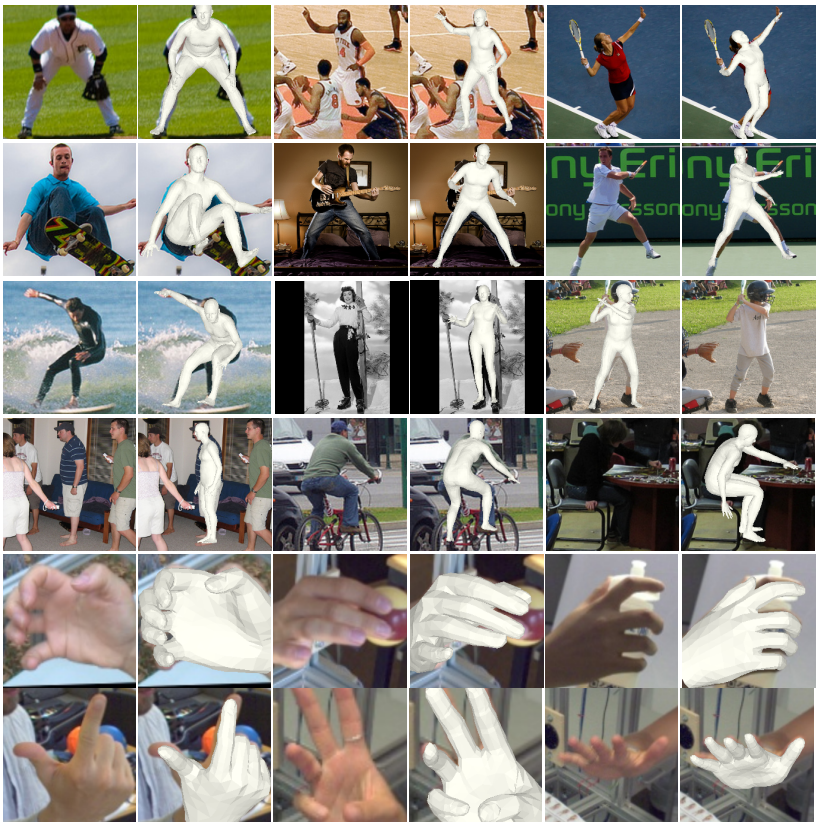

Fig. 2. Additional qualitative results on COCO and FreiHAND.

1.3 Comparison with the state-of-the-art

We present the qualitative comparison between our Pose2Mesh and GraphCMR [6] in Figure 3. We regard GraphCMR as a suitable comparison target, since it is also the model-free method and regresses coordinates of human mesh defined by SMPL [9] using GraphCNN like ours. As the figure shows, our Pose2Mesh provides much more visually pleasant mesh results than GraphCMR. Based on the loss function analysis in Section 7 and the visual results of GraphCMR, we

conjecture that the surface losses such as the normal loss and the edge loss are the reason for the difference.

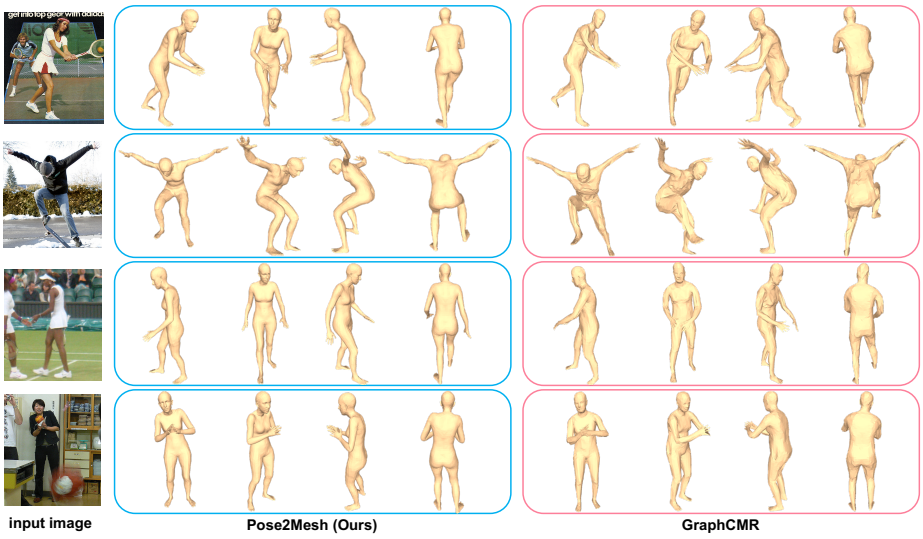

Fig. 3. The mesh quality comparison between our Pose2Mesh and GraphCMR [6].

2 Details of PoseNet

2.1 Network architecture

Figure 4 shows the detailed network architecture of PoseNet. First, the normalized input 2D pose vector is converted to a 4096-dimensional feature vector by a fully-connected layer. Then, it is fed to the two residual blocks, where each block consists of a fully connected layer, 1D batch normalization, ReLU activation, and the dropout. The dimension of the feature map in the residual block is 4096, and the dropout probability is set to 0.5. Finally, the output from the residual block is converted to $(3J)$-dimensional vector, the 3D pose vector, by a fully-connected layer. The 3D pose vector represents the root-relative 3D pose coordinates.

2.2 Accuracy of PoseNet

We present MPJPE and PA-MPJPE of PoseNet on the benchmarks in Table 1. For the Human3.6M benchmark [2], 14 common joints out of 17 Human3.6M defined joints are evaluated following [4–6, 16]. For the 3DPW benchmark [11], COCO defined 17 joints are evaluated and $\mathcal{J}\mathbf{M}$ from the groundtruth SMPL

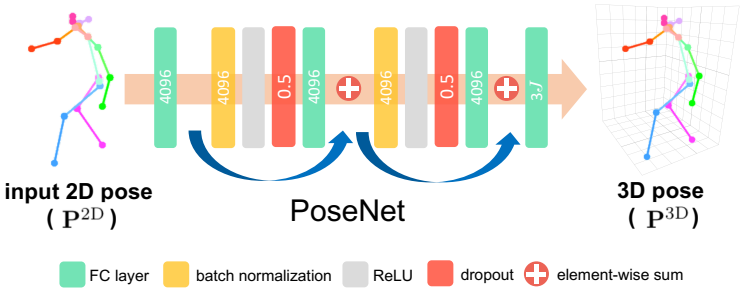

Fig. 4. The detailed network architecture of PoseNet.

meshes are used as groundtruth. The 2D pose outputs from [19] and [18] are taken as test inputs on Human3.6M and 3DPW respectively. For the FreiHAND benchmark, only FreiHAND train set is used during training, and 21 MANO [17] hand joints are evaluated by the official evaluation website. The 2D pose outputs from [18] are taken as test inputs.

Table 1. The MPJPE and PA-MPJPE of PoseNet on each benchmark.

train set	Human3.6M		Human3.6M + COCO	
benchmark	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
Human3.6M	65.1	48.4	66.7	48.9
3DPW	105.0	62.9	99.2	61.0

benchmark	PA-MPJPE
FreiHAND	8.56

3 Pre-defined joint sets and graph structures

We use different pre-defined joint sets and graph structures for Human3.6M, 3DPW, and FreiHAND benchmarks, as shown in Figure 5. To be specific, we employ Human3.6 body joints, COCO body joints, MANO hand joints for Human3.6M, 3DPW, FreiHAND benchmarks, respectively, in both training and testing stages. For the COCO joint set, we additionally define pelvis and neck joints that connect the upper body and lower body. The pelvis and neck coordinates are calculated as the middle point of right-left hips and right-left shoulders, respectively.

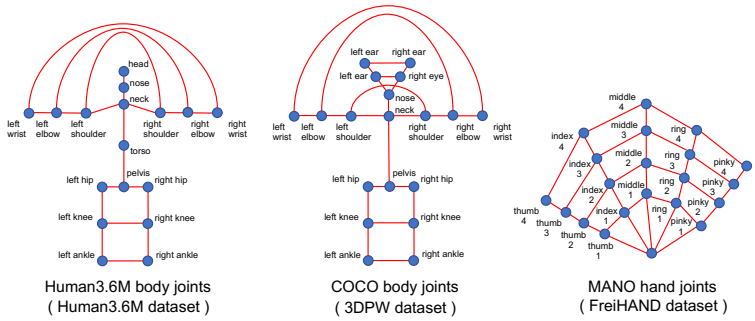

Fig. 5. The joint sets and graph structures of each dataset that are used in Pose2Mesh.

4 Pseudo-groundtruth SMPL parameters of Human3.6M dataset

Mosh [8] method can compute SMPL parameters from the marker data in Human3.6M dataset. Since Human3.6M dataset does not provide 3D mesh annotations, most of the previous 3D pose and mesh estimation papers [4–6,16] used the SMPL parameters obtained by Mosh method as the groundtruth for the supervision. However, due to the license issue, the SMPL parameters are not currently available. Furthermore, the source code of Mosh is not publicly released.

For the 3D mesh supervision, we alternatively obtain groundtruth SMPL parameters by applying SMPLify-X [15] on the groundtruth 3D joint coordinates of Human3.6M dataset. Although the obtained SMPL parameters are not perfectly aligned to the groundtruth 3D joint coordinates, we confirmed that the error of the SMPLify-X is much less than those of current state-of-the-art 3D human pose estimation methods, as shown in Table 2. Thus, we believe using SMPL parameters obtained by SMPLify-X as groundtruth is reasonable. For the fair comparison, all the previous works and our system are trained on our SMPL parameters from SMPLify-X.

During the fitting process of SMPLify-X, we adopted a neutral gender SMPL body model. However, we empirically found that the fitting process produces gender-specific body shapes, which correspond to each subject. As a result, since most of the subjects in the training set of Human3.6M dataset are female, our Pose2Mesh trained on Human3.6M dataset tends to produce female body shape meshes. We tried to fix the identity code of the SMPL body model obtained from the T-pose; however, it produces higher errors. Thus, we did not fix the identity code for each subject.

5 Synthetic data from AMASS

We leverage additional synthetic data from AMASS [10] to boost the performance of Pose2Mesh. AMASS is a new database that unifies 15 different optical marker-based mocap datasets within a common framework. It created SMPL

Table 2. The MPJPE comparison between SMPLify-X fitting results and state-of-the-art 3D human pose estimation methods. "*" takes multi-view RGB images as inputs.

methods	MPJPE
Moon et al. [12]	53.3
Sun et al. [19]	49.6
Iskakov et al. [3]*	20.8
SMPLify-X from GT 3D pose	**13.1**

parameters from mocap data by a method named Mosh++. We used CMU dataset [1] from the database in training.

To be specific, we generated paired 2D pose-3D mesh data by projecting a 3D pose obtained from a mesh to the image plane, using camera parameters from Human3.6M. As shown in Table 3, when AMASS is added, both the joint error and surface error decrease. Exploiting AMASS data in this fashion is not possible for [4], [6], and [5], since they need pairs of image and 2D/3D annotations.

Table 3. The MPJPE and MPVPE of our Pose2Mesh on 3DPW with accumulative training datasets. The 2D pose outputs from [18] are used for input to Pose2Mesh.

train sets	MPJPE	MPVPE
Human3.6M+COCO	91.4	109.3
Human3.6M+COCO+AMASS	**90.1**	**108.0**

6 Synthesizing the input 2D poses in the training stage

6.1 Detailed description of the synthesis

As described in Section 4.1 of the main manuscript, we synthesize the input 2D poses by adding randomly generated errors on the groundtruth 2D poses in the training stage. For this, we generate errors following Chang et al. [14] and Moon et al. [13] for Human3.6M and COCO body joint sets, respectively. On the other hand, for FreiHAND benchmark, we used detection outputs from [18] on the training set as the input poses in the training stage, since there are no verified synthetic errors for the hand joints.

6.2 Effect of synthesizing the input 2D poses

To demonstrate the validity of the synthesizing process, we compare MPJPE and PA-MPJPE of Pose2Mesh trained with the groundtruth 2D poses, and the synthesized input 2D poses in Table 4. For Human3.6M, only Human3.6M train set is used for the training, and for 3DPW benchmark, Human3.6M and COCO are used for the training. The test 2D input poses used in Human3.6M and 3DPW evaluation are outputs from Integral Regression [19] and HRNet [18] respectively,

using groundtruth bounding boxes. Apparently, when our Pose2Mesh is trained with the synthesized input 2D poses, Pose2Mesh performs far better on both benchmarks. This proves that the synthesizing process makes Pose2Mesh more robust to the errors in the input 2D poses and increases the estimation accuracy.

Table 4. The MPJPE and PA-MPJPE comparison according to input type in the training stage.

input pose when training	Human3.6M		3DPW	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
2D pose GT	70.4	50.6	153.7	94.4
2D pose synthesized (Ours)	**64.9**	**48.7**	**91.4**	**60.1**

7 Train/test with groundtruth input poses

We present the upper bounds of Pose2Mesh, PoseNet, and MeshNet on Human3.6M and 3DPW benchmarks by training and testing with groundtruth input poses in Table 5. Pose2Mesh and PoseNet take the groundtruth 2D pose as an input, while MeshNet takes the groundtruth 3D pose as an input. As the table shows, the upper bound of Pose2Mesh is similar to that of PoseNet, which implies that the 3D pose errors of Pose2Mesh follow those of PoseNet as analyzed in Section 7.2 of the main manuscript. In addition, the upper bound of MeshNet indicates that we can recover highly accurate 3D human meshes if we can estimate nearly perfect 3D poses.

The MPJPE and PA-MPJPE of Pose2Mesh and MeshNet are measured on the 3D pose regressed from the mesh output, while the accuracy of PoseNet is measured on the lifted 3D pose. For the Human3.6M benchmark, only Human3.6M train set is used to train the network. For the 3DPW benchmark, Human3.6M, COCO, AMASS train sets are used to train the network.

Table 5. The upper bounds of Pose2Mesh, PoseNet, and MeshNet on Human3.6m and 3DPW benchmarks.

networks	Human3.6M		3DPW	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
Pose2Mesh with 2D pose GT	51.1	35.3	65.1	34.6
PoseNet with 2D pose GT	50.6	41.3	66.1	43.8
MeshNet with 3D pose GT	13.9	9.9	10.8	8.1

8 Effect of each loss function

We analyze the effect of joint coordinate loss L_{joint}, surface normal loss L_{normal}, and surface edge loss L_{edge} on reconstructing a 3D human mesh in Table 6 and Figure 6. Human3.6M dataset is used for the training and testing. As the table shows, training without L_{joint} has a relatively distinctive effect on MPJPE and PA-MPJPE, while other settings show numerically negligible differences. On the other hand, as the figure shows, training without L_{normal} or L_{edge} clearly decreases the visual quality of the mesh output, while training without L_{joint} has nearly no effect on the visual quality of the meshes. To be specific, training without L_{normal} impairs the overall smoothness of the mesh and local details of mouth, hands, and feet. Similarly, training without L_{edge} ruins the details of body parts that have dense vertices, especially mouth, hands, and feet, by making serious artifacts caused by flying vertices.

Table 6. The MPJPE and PA MPJPE comparison between the networks trained from various combinations of loss functions.

settings	MPJPE	PA-MPJPE
full supervision (Ours)	64.9	48.7
without L_{joint}	66.9	50.1
without L_{normal}	**64.6**	**48.5**
without L_{edge}	64.8	48.7

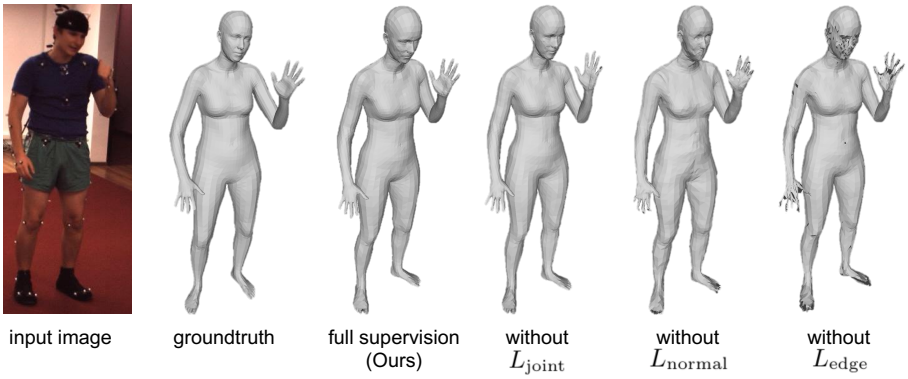

Fig. 6. Qualitative results for the ablation study on the effectiveness of each loss function.

References

1. The Carnegie Mellon University (CMU) Graphics Laboratory Motion Capture Database. http://mocap.cs.cmu.edu/
2. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. The IEEE transactions on pattern analysis and machine intelligence (TPAMI) (2014)
3. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
4. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
5. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
6. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
7. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: The European Conference on Computer Vision (ECCV) (2014)
8. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. In: Special Interest Group on GRAPHics and Interactive Techniques (SIGGRAPH) (2014)
9. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: Special Interest Group on GRAPHics and Interactive Techniques (SIGGRAPH) (2015)
10. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
11. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (2018)
12. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
13. Moon, G., Chang, J.Y., Lee, K.M.: Posefix: Model-agnostic general human pose refinement network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
14. Moon, G., Chang, J.Y., Lee, K.M.: Absposelifter: Absolute 3d human pose lifting network from a single noisy 2d human pose. arXiv preprint arXiv:1910.12029 (2020)
15. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
16. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

17. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. In: Special Interest Group on GRAPHics and Interactive Techniques (SIGGRAPH) (2017)
18. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
19. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: The European Conference on Computer Vision (ECCV) (2018)
20. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
21. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: The IEEE International Conference on Computer Vision (ICCV) (2019)