OPEN: Order-preserving Pointcloud Encoder Decoder Network for Body Shape Refinement *Supplementary Material*

Haidong Zhu[*], Ye Yuan[†], Yiheng Zhu[†], Xiao Yang[†] and Ram Nevatia[*]
[*]Department of Computer Science, University of Southern Califormia, Los Angeles, California 90089
[†]Bytedance Inc., Mountain View, California 94041
haidongz@usc.edu

In this supplementary document, we include some discussions and details for the experiments that couldn't be fit in the main paper. We discuss the main differences between our model and other state-of-the-art models, followed by some ablation studies for different modules in our network.

I. Discussion

In this discussion part, we discuss the major difference of our proposed network OPEN compared with other point cloud processing networks and reconstruction networks.

A. *Difference with other point cloud processing networks*

Existing point cloud backbone networks, such as PointNet [1] and PointNet++ [2], are the order equivalent networks for the original point cloud during reconstruction. That is to say, different input orders of the point do not have any effect on the reconstructed point cloud and cannot preserve the order of the point. Our task needs an order-preserving point cloud since each point has its specific location and meaning for SMPL meshes. Thus existing point cloud processing network is not feasible for our method. When the point cloud has specific order to preserve, OPEN can refine its shape while preserving its original point order.

B. *Comparison with existing 3-D reconstruction models*

Existing 3-D reconstruction networks can be divided into two different branches. For some of them, the networks can preserve the original order of the point cloud while their reconstruction is not fine enough with parametric human prior (like SMPL [3] and SMPL-X [4]). For the others, networks generate finegrained of its out shape while cannot preserve its original order (like PIFu [5] and PIFuHD [6]). Compared with these methods, our method can preserve its original order while refining the point cloud, so that it can more precisely describe the outer shape of the object.

II. Experiments

A. *Implementation Details*

For the experiment, we use the pretrained ResNet-18 [7] and finetune it with the data used in our experiments. We train our whole pipeline end-to-end and set the starting learning rate as 1e-4. This learning rate decay to half after every 250 epochs, and the maximum number of epochs for the experiment is 2,000. The first multi-layer perceptron (MLPs) used by the point encoder is a two-layer MLP with 64 as its hidden space. takes the N_s-by-3 point cloud as input and concatenates it with its local covariance (N_s-by-9) to extend to a N_s-by-12 matrix for each example. It projects the matrix to a N_s-by-64 feature matrix, followed by a graph-convolution layer to project the whole feature into a 1024-d feature space. After this, we first max-pool the N_s-by-1024 local feature to a 1-by-1024 feature vector and use a second MLP to project it to 512-d space with 512 as the feature dimension of the hidden layer. For each 3-D point set used for supervision during training, we randomly rotate the shape between -45 degrees to +45 degrees from the front view angle for augmentation. We evaluate the reconstructed results on the front view results for the test split of the RenderPeople Posed subset during inference.

TABLE I
Ablation results on RenderPeople test set. 'rot.' is the abbreviation of rotation augmentatation of the original 3-D shapes in RenderPeople Posed dataset. The first half of the table are the results compared with OPEN on supervised settings, while the second half are the results compared with using external datasets for 2-D weakly supervision.

Method	CD (↓)	Mesh Acc (↓)
OPEN	0.60	0.33
OPEN w/o RGB input	0.75	0.43
OPEN w/o global feat. f_g	0.88	0.51
OPEN w/o local feat. f_l	0.83	0.49
OPEN w/o rot.	0.73	0.45
OPEN w/o rot. w/o L2	0.77	0.46
OPEN w/ Rigged	0.53	0.29
OPEN w/ Rigged w/o rot.	0.55	0.30

B. *Ablation studies*

Besides comparing with the state-of-the-art human body shape reconstruction models, we show some ablations for the composition of the OPEN model in Table I. We make some comparisons on i) excluding L2 loss ii) excluding the RGB input of the reconstruction, iii) excluding using of rotation augmentation for figures with 3-D human body labels, iv) using only f_g or f_l instead of using both of them. We have the following observations:

1) Loss and augmentation: We note that using the rotation augmentation, since both of them can significantly improve the variety of the dataset compared with using the 3-D Chamfer Distance only for supervision. Also, using L2 in the model helps improve the CD and mesh accuracy. Introducing L2 might limit the power of using only Chamfer Distance, but it will be treated as a regulator to avoid the network focusing too much on the shapes in the training set.

2) Feature used for OPEN: In the experiment, we show that using all three features for decoding the refined shape can have the best performance for the OPEN model, and removing any of them will harm the results. Note that the model still generates better results than the original SMPL model when not using any RGB images as feature input. If no 2-D RGB images are available during training, OPEN will generate an average modification for all the points for the reconstructed SMPL model. The significant improvement here compared with SMPLify shows that the reconstructed human body shape by SMPLify tends to have some general errors, such as the waist is always thinner than the groundtruth shape, which makes the average modification works for the correction.

References

[1] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *CVPR*, 2017, pp. 652–660.

[2] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *arXiv preprint arXiv:1706.02413*, 2017.

[3] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: A skinned multi-person linear model,” *TOG*, vol. 34, no. 6, pp. 1–16, 2015.

[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it smpl: Automatic estimation of 3d human pose and shape from a single image,” in *ECCV*, 2016, pp. 561–578.

[5] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in *ICCV*, 2019, pp. 2304–2314.

[6] S. Saito, T. Simon, J. Saragih, and H. Joo, “Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization,” in *CVPR*, 2020, pp. 84–93.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.