Supplementary: Geometry-Guided Progressive NeRF for Generalizable and Efficient Neural Human Rendering

Mingfei Chen[1,2], Jianfeng Zhang[3], Xiangyu Xu[1], Lijuan Liu[1], Yujun Cai[1], Jiashi Feng[1], and Shuicheng Yan[1]

[1] Sea AI Lab
[2] University of Washington
[3] National University of Singapore

1 Pose Estimation

As Table 1, we evaluate the keypoints MPJPE (Mean Per Joint Position Error) metric on human bodies of ZJU-MoCap dataset [6] for pose assessment. Our GP-NeRF achieves 47.6mm on the unseen bodies. Specifically, we use OpenPose [1] to generate groundtruth poses and predicted poses from groundtruth images and the corresponding rendered novel view images respectively. Previous baselines NT [7], NHR [8] and NB [6] perform per-scene training, so we evaluate their results on seen human bodies for comparison. NT, NHR and NB get 57.3mm, 56.05mm and 51.3mm, respectively. The generalizable baseline NHP [2] gets 56.05mm on the unseen human bodies. Our method achieve over 7% improvement on the unseen human body pose assessment task, even comparing to per-scene training methods.

Table 1. Pose Error Comparison. MPJPE (Mean Per Joint Position Error) results comparison on human bodies of ZJU-MoCap dataset [6].

Method	Unseen	MPJPE ↓
NB [8]	✗	51.3 mm
NHR [8]	✗	56.0 mm
NT [7]	✗	57.3 mm
NHP [2]	✓	56.1 mm
GP-NeRF (Ours)	✓	47.6 mm

2 More Discussion on Method Comparisons

In this section, we further compare our contributions with some related human rendering methods: Neural Actor [4], Animatable SDF [5], MonoPort [3] and [9].

Our Geometry-guided Progressive NeRF differs from these previous methods [4,5,3,9] in that it addresses two important issues of them with high efficiency:

1) Inaccurate geometry prior and hard to adapt to the target human body. Neural Actor and Animatable SDF transform the geometry prior (e.g., SMPL) to a canonical space, or even directly use SMPL to guide point sampling for the 'coarse' step. However, the SMPL estimation does not always fit well to the target body, such as missing the clothes regions, and hard to generalize accurately over different body shapes. Differently, we learn query for each SMPL vertex to aggregate the multi-view feature adaptively, and use a learnable network to expand the sparse geometry feature volume to adapt to the target human body. After the 'coarse' step, our aggregated geometry volume can generalize over different body shapes more accurately and even cover the clothes regions.

2) Violating normal human body geometry structures. Methods like MonoPort and [9] directly sample points from the whole 3D feature space (or generated 3D occupancy fields (MonoPort), and then decide the human body opacity for later rendering. However, these methods neglect coherence between the geometry prior and visual input, which might generate results that violate the normal human body structures (e.g., legs / arms in the wrong places). Differently, our multi-view integration is guided by learnable queries of the sparse vertices in SMPL. Therefore, we can adaptively fuse the visual and geometry information, and construct the geometry volume that conforms to the normal body geometry constraints with much fewer sampling points.

3 Result Videos

We provide some result videos on the ZJU-MoCap dataset under various generalization levels. Specifically, we provide some demo result examples on the training frames in *seen.mp4*, result examples on the unseen human body frames in *unseen.mp4*. In each frame of these videos, we show two columns of images, where the left one is the groundtruth and the right one is our synthesis results. For each video, the three input cameras are uniformly set following the implementation details and fixed to their locations, and we render different target views that roughly cover the 360 degrees surrounded the human body to show our synthesis results.

4 Source Code

We provide our implementation code based on PyTorch in the following GitHub link: https://github.com/sail-sg/GP-Nerf. Our code can also support custom data input from the real life calibrated cameras.

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017) 1
2. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. In: NeurIPS (2021) 1
3. Li, R., Xiu, Y., Saito, S., Huang, Z., Olszewski, K., Li, H.: Monocular real-time volumetric performance capture. In: ECCV (2020) 1
4. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM Trans. Graph. **40**(6) (dec 2021) 1
5. Peng, S., Zhang, S., Xu, Z., Geng, C., Jiang, B., Bao, H., Zhou, X.: Animatable neural implicit surfaces for creating avatars from videos (2022) 1
6. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021) 1
7. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. In: ACM Trans. on Graphics (2019) 1
8. Wu, M., Wang, Y., Hu, Q., Yu, J.: Multi-view neural human rendering. In: CVPR (2020) 1
9. Zins, P., Xu, Y., Boyer, E., Wuhrer, S., Tung, T.: Data-driven 3d reconstruction of dressed humans from sparse views. In: 2021 International Conference on 3D Vision (3DV). pp. 494–504 (2021) 1, 2