Supplementary Material – NeuMan: Neural Human Radiance Field from a Single Video

Wei Jiang†[1,2], Kwang Moo Yi[1,2], Golnoosh Samei[1],
Oncel Tuzel[1], and Anurag Ranjan[1]

[1] Apple
[2] The University of British Columbia
{jw221, kmyi}@cs.ubc.ca, {golnoosh, otuzel, anuragr}@apple.com

A.1 Dataset Details

The dataset details are as follows.

Sequence	Total Frames	Train Frames	Validatation Frames	Test Frames
Seattle	41	33	4	4
Citron	37	30	4	3
Parking	42	34	4	4
Bike	104	83	11	10
Jogging	102	82	10	10
Lab	103	82	11	10

Table 1: Number of frames in each dataset used for training, validation and test.

A.2 SMPL Refinement

Given an image, we regress the 2D joints j_{2d} and segmentation mask m of the human using HigherHRNet [1] and DensePose [2]. We further estimate the SMPL mesh $M = (V, F)$, a collection of vertices and faces using ROMP [7]. The mesh M is parametrized by SMPL parameters θ such that $M = \mathrm{SMPL}(\theta)$ and includes the 3D joints j_{3d}. The regressed SMPL parameters θ are noisy. Therefore, we use soft-rasterizer [4], Π to refine these estimates. Given a mesh, M and camera θ_c, the rasterizer renders a silhouette $\hat{m} = \Pi(\theta_c, M)$. We also project the 3D joints in the image plane using camera matrix $\hat{j}_{2d} = \mathbf{p}(j_{3d})$ where $\mathbf{p}$ is a projection operator. We obtain the refined SMPL parameters and camera estimates by minimizing

$$\theta^* = \min_{\theta} \parallel m - \hat{m} \parallel + \parallel j_{2d} - \hat{j}_{2d} \parallel . \tag{1}$$

†Work done while interning at Apple.

Notice that we use the estimates from DensePose [2] as the target silhouettes in the preprocessing, while using the estimates from Mask-RCNN [3] as the target masks during the training of human NeRF model. It's because in the preprocessing phase, the rendered masks are from naked SMPL mesh and DensePose [2] is trained on such dense SMPL correspondences. However, we wish to learn extra geometry details beyond the SMPL model with the human NeRF model, and Mask-RCNN [3] better estimates those 2D details such as the hair and clothes.

A.3 Network Architecture

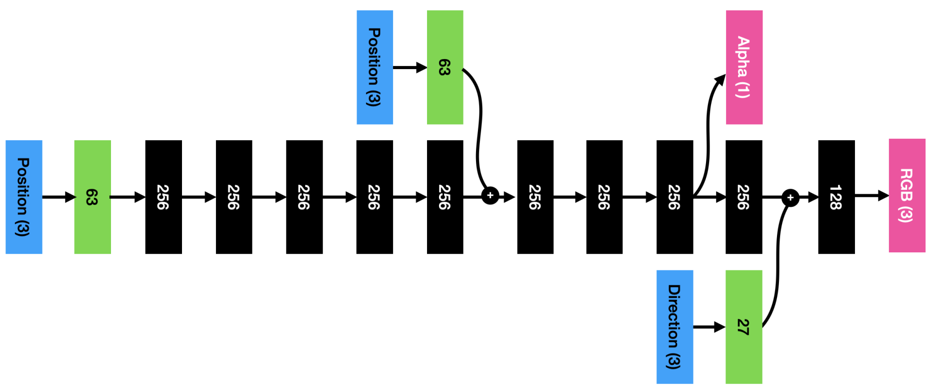

Fig. 1: **Visualization of NeRF architecture** – Blue boxes denote the raw inputs to the network, green boxes denotes the positional encoded values, black boxes denotes the hidden layers, and pink boxes denote the final outputs. The number in each box indicates the feature dimensions. + sign indicates the concatenation operations. ReLU is applied after each hidden layer.

Following [5], our scene NeRF models consists of a coarse sub-model and a fine sub-model. However, with the prior geometry provided by SMPL estimates, we only use one sub-model for the human NeRF model. Each sub-model and the error-correction network has the same architecture as shown in 1.

A.4 Comparison with previous works

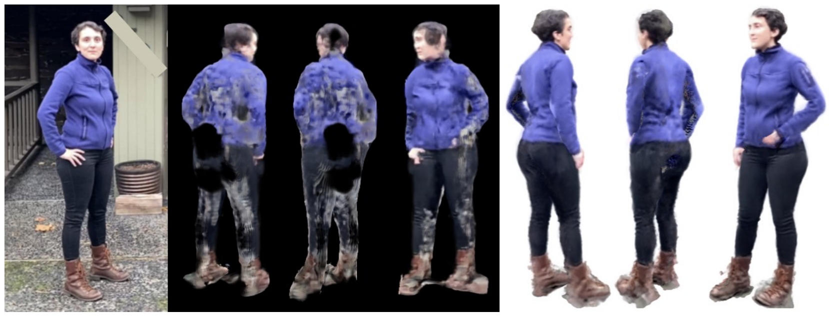

Fig. 2: **Novel view rendering of NeuralBody vs Ours on Seattle sequence.** – The pose being rendered is from the left training image. Three images in the middle are the novel view renderings from NeuralBody, and three images on the right are from ours.

We apply NeuralBody [6] to our dataset in a monocular setting. The results are shown in 2. NeuralBody [6] overfits to the training observations, and produce poor rendering on the back of the subject, while ours generalize better and can faithfully render the back.

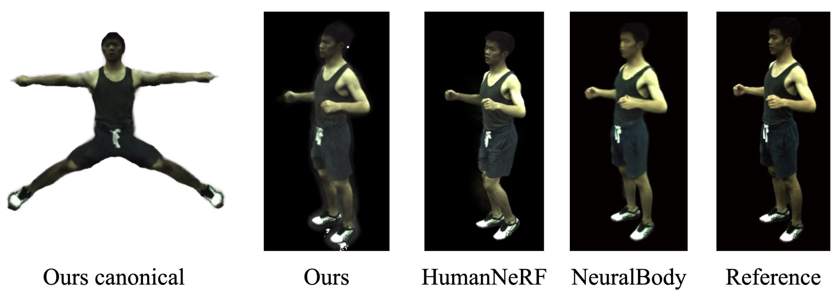

Fig. 3: **Novel View Reconstructions on public ZJU Mocap dataset [6]** – Ours and HumanNeRF [8] use only one camera view, NeuralBody [6] uses multiple camera views.

we also compare our method with HumanNeRF [8] and NeuralBody [6] on a ZJU Mocap dataset, as shown qualitative comparisons in Figure 3. Our method renders high quality novel view renderings with the ability to extrapolate in pose space.

A.5 Error Correction Network and Scene Model Conditioning

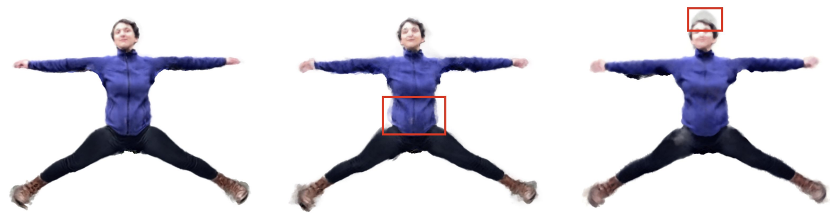

Fig. 4: Canonical renderings of our full model, model without error correction network, and model without conditioning on scene NeRF.

Without the error correction network, the the canonical NeRF lacks details on the cloth and face. Training only the human NeRF in isolation leads to worse performance as the human NeRF model may encode the background pixels into its radiance fields due to segmentation errors. In either case, the canonical NeRF creates fogs around the human to hallucinate the clothing dynamics or the background colors, as shown in 4.

References

1. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In: CVPR (2020) 1
2. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018) 1, 2
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 2
4. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7708–7717 (2019) 1
5. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020) 2
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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021) 3
7. Sun, Y., Bao, Q., Liu, W., Fu, Y., Michael J., B., Mei, T.: Monocular, One-stage, Regression of Multiple 3D People. In: ICCV (October 2021) 1
8. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. arXiv preprint arXiv:2201.04127 (2022) 3