Rendering Humans from Object-Occluded Monocular Videos Supplementary Materials

Tiange Xiang*, Adam Sun, Jiajun Wu, Ehsan Adeli, Li Fei-Fei
Stanford University

A. Network Architecture

OccNeRF renders humans through the regression MLP network $\mathcal{F}(\cdot)$. We designed this network by following the architecture suggested by NeRF [5]. Specifically, there are 8 linear layers in this network, each with 256 neurons and the ReLU non-linearity. The input to this network is the same as the input described in Equation 9. For each input tensor, this network outputs two values indicating the density and radiance at the corresponding position. The density values σ are returned at the end of the fourth linear layer, while the radiance values $\mathbf{c}$ are returned at the end of the eighth linear layer. There is also a skip connection, which concatenates the input to the activation of the fifth layer. The architecture is outlined in Figure 1.

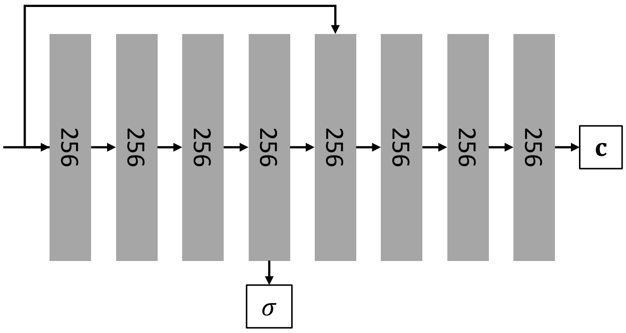

Figure 1. Network architecture for the regression MLP $\mathcal{F}(\cdot)$ used in OccNeRF.

B. Definition of Occlusion

For simulated occlusions (ZJU-MoCap), we define the extent of occlusion as $1 - \frac{\#\text{occluded pixels}}{\#\text{valid pixels}}$. For real-world occlusions (OcMotion), since there is no reference for the occluded body, we rely on 2D projections of the GT SMPL mesh: $1 - \frac{\#\text{visible pixels} \cap \text{SMPL pixels}}{\#\text{SMPL pixels}}$. The occlusion extent for different OcMotion videos are as follows:

Video	Mild	Severe	A	B
Where?	Sec. 4	Sec. 4	Supp. F	Supp. F
Extent	17%	79%	55%	6%

Table 1. Occlusion extents for OcMotion videos.

C. Relevance to SMPL Methods

Implicit modeling (NeRF) and explicit modeling (SMPL) are two distinct approaches. NeRF methods provide high-fidelity and photo-realistic renderings at customized angles, while SMPL methods can generate a complete human body but often suffer from low resolution textures. In fact, most advanced human NeRF methods rely on SMPL predictions as geometry priors to achieve more realistic results. However, recovering *occluded appearances* is an unsolved problem in both approaches. Exactly as the reviewer pointed out, our method considers the completeness of SMPL mesh and designed $\mathcal{L}_{completeness}$ (Eq. 10) to combine advantages of both implicit and explicit approaches, which is novel and effective.

D. Metric on Completeness

In addition to the most commonly used metrics: PSNR/SSIM, here we calculate an extra metric to measure completeness of human renderings. Detailedly, we compute IoU between 2D GT segmentations and the rendered masks on the ZJU-MoCap dataset as an indication of completeness. An occlusion-aware method is supposed to yield high IoU scores.

Subject	377	386	387	392	393	394
[7]	0.8523	0.7908	0.8397	0.8103	0.8135	0.8002
Ours	0.8573	0.8755	0.8677	0.8327	0.8317	0.8458

E. Occlusion Sensitivity

In section 4.4 of the main paper, we demonstrated experimental results under simulated occlusions by masking 50% of the valid pixels. In this section, we explore the sensitivity of our proposed method under different levels of occlusion. We designed experiments by masking 10%,

*Correspondence to xtiange@stanford.edu

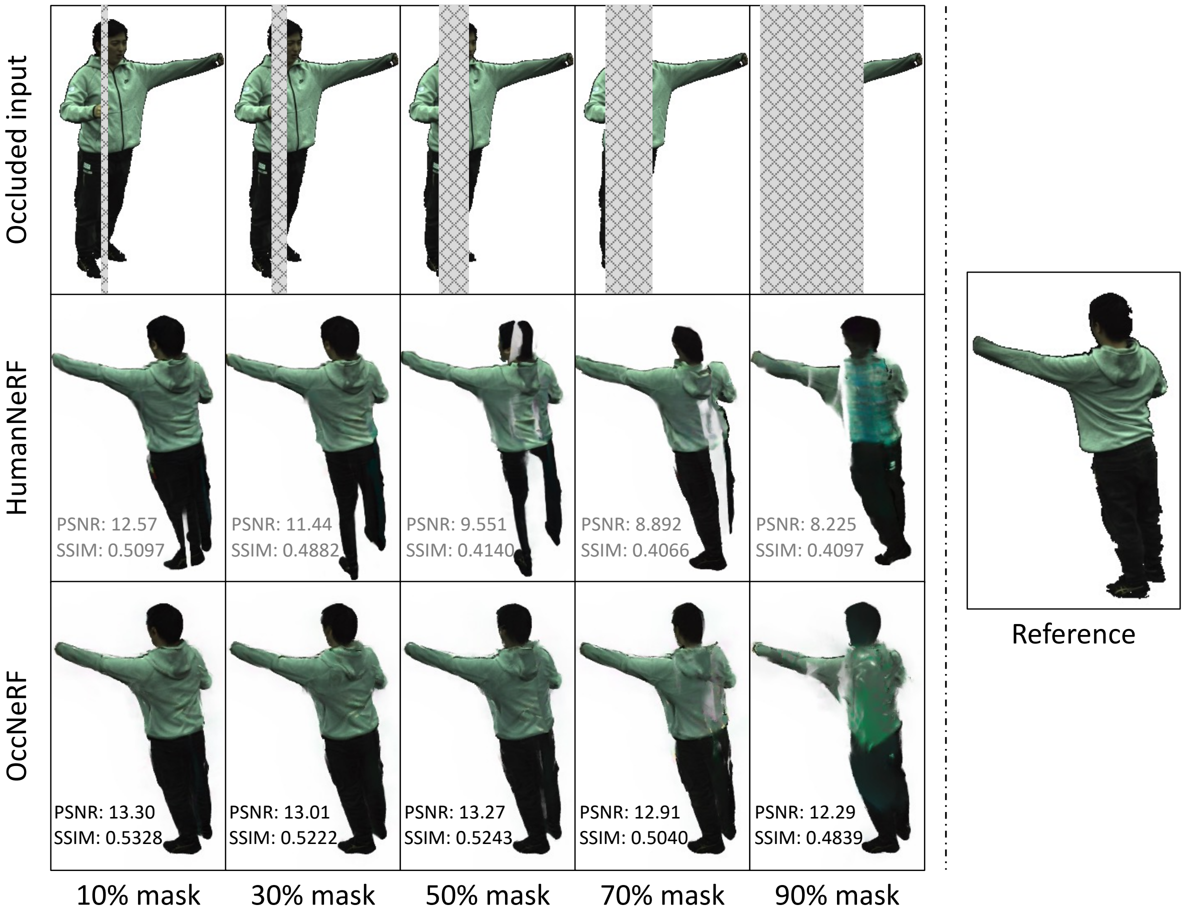

Figure 2. Comparison results for different levels of occlusions.

30%, 50%, 70%, and 90% of the valid pixels in the video frames. The rendering results under various levels of occlusions are compared in Figure 2. For most occlusion levels (i.e. 10% - 70%), OccNeRF not only produces a complete human geometry but also in-paints the occluded appearances very well, surpassing HumanNeRF considerably. Severe occlusions (i.e. 90%) are challenging for both OccNeRF and HumanNeRF. Nevertheless, OccNeRF still outperforms HumanNeRF by a safe margin due to less artifacts and more natural renderings. The quantitative metrics also validate our claims.

F. More Comparisons on Simulated Occlusions

In section 4.4 of the main paper, we compare OccNeRF against the state-of-the-art method HumanNeRF [7]. However, we are also interested in comparing with Neural Body, another standard benchmark [6]. We trained Neural Body by following their suggested implementation, while adding the same simulated occlusions to the first 80% of training frames. In Figure 3, it is clear that Neural Body completely fails on occluded training data. The renderings are highly corrupted by unexpected artifacts and splotches due to the occlusion. OccNeRF, on the other hand, demonstrates its rendering superiority in terms of both geometry modeling and appearance recovering.

G. More Results on Real-World Occlusions

Simulated occlusions cannot fully reflect the challenge of real-world scenes. In section 4.5 of the main paper, we presented results on two videos from the OcMotion [2] dataset, which contain real-world occlusions from various obstacles.

One of the main challenges is that it is sometimes difficult to acquire accurate priors (binary human mask, SMPL parameters) from the occluded videos. Inaccurate priors impair network performance that leads to poor rendering quality. In the main paper, we acquired the binary mask from Mask2Former [1] and the SMPL parameters from [3]. Here, we conduct additional experiments on two more challenging videos from the OcMotion dataset while using inaccurately estimated priors. Specifically, the binary mask is acquired from Mask2Former [1] again, but we calculate the SMPL parameters using PARE [4], an occlusion-

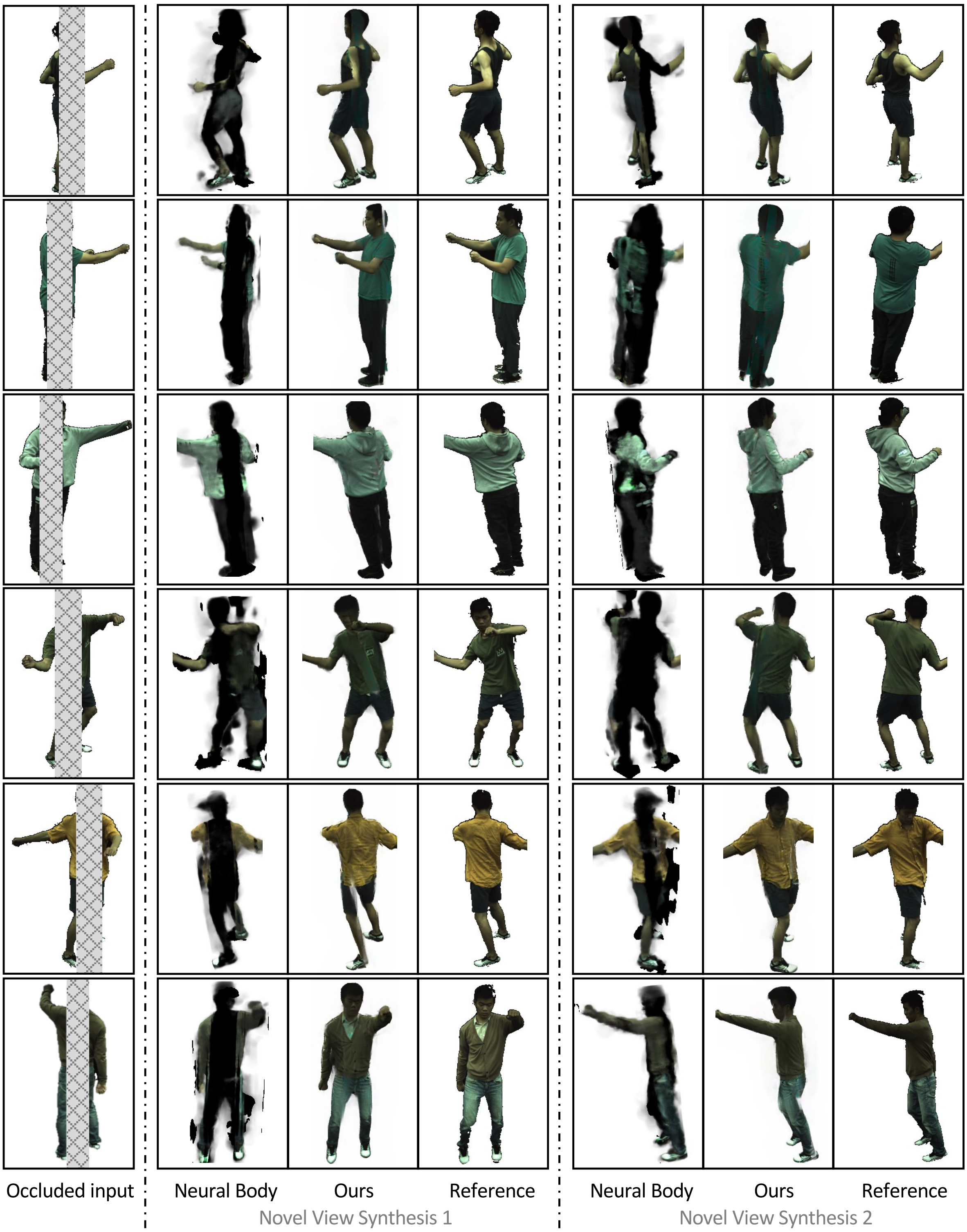

Figure 3. Qualitative results on **simulated** occlusions in the OcMotion dataset [2]

robust monocular video based method. As shown in Figure 4, inaccurate priors pose additional troubles to both of the methods. OccNeRF tends to produce more smoothed results lacking of high-frequency details, whereas HumanNeRF fails to generate reasonable renderings at all.

ZJU-MoCap	Subject **377**				Subject **386**			
	PSNR$_{\text{vis}}$	SSIM$_{\text{vis}}$	PSNR$_{\text{full}}$	SSIM$_{\text{full}}$	PSNR$_{\text{vis}}$	SSIM$_{\text{vis}}$	PSNR$_{\text{full}}$	SSIM$_{\text{full}}$
Neural Body [6]	7.784	0.4214	16.61	0.9173	8.641	0.3120	19.02	0.9318
OccNeRF	13.23	0.6097	23.43	0.9642	13.44	0.5974	23.66	0.9639

ZJU-MoCap	Subject **387**				Subject **392**			
	PSNR$_{\text{vis}}$	SSIM$_{\text{vis}}$	PSNR$_{\text{full}}$	SSIM$_{\text{full}}$	PSNR$_{\text{vis}}$	SSIM$_{\text{vis}}$	PSNR$_{\text{full}}$	SSIM$_{\text{full}}$
Neural Body [6]	9.164	0.3520	18.74	0.9232	7.830	0.3698	16.73	0.9115
OccNeRF	13.27	0.5243	22.26	0.9513	13.00	0.5692	22.13	0.9575

ZJU-MoCap	Subject **393**				Subject **394**			
	PSNR$_{\text{vis}}$	SSIM$_{\text{vis}}$	PSNR$_{\text{full}}$	SSIM$_{\text{full}}$	PSNR$_{\text{vis}}$	SSIM$_{\text{vis}}$	PSNR$_{\text{full}}$	SSIM$_{\text{full}}$
Neural Body [6]	9.146	0.3861	18.13	0.9163	10.15	0.3830	19.73	0.9265
OccNeRF	12.00	0.4655	21.58	0.9489	13.12	0.5317	22.06	0.9532

Table 2. Quantitative comparison against Neural Body [6] on ZJU-MoCap. We color cells that have the best metric values.

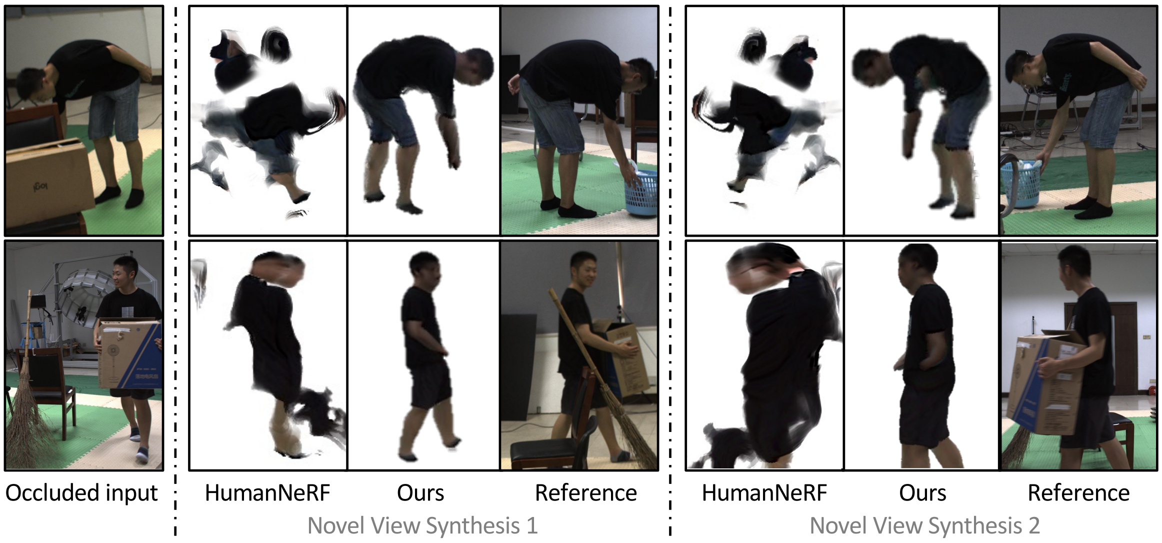

Figure 4. More qualitative results on **real-world** occlusions with inaccurately estimated priors in the OcMotion dataset [2].

References

[1] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 2

[2] Buzhen Huang, Yuan Shu, Jingyi Ju, and Yangang Wang. Occluded human body capture with self-supervised spatial-temporal motion prior. *arXiv preprint arXiv:2207.05375*, 2022. 2, 3, 4

[3] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 710–720. IEEE, 2021. 2

[4] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2

[5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[6] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 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*, pages 9054–9063, 2021. 2, 4

[7] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 1, 2