MonoHuman: Animatable Human Neural Field from Monocular Video Supplementary Material

Zhengming Yu[1], Wei Cheng[1,2], Xian Liu[3], Wayne Wu[2], Kwan-Yee Lin[2, 3] ✉

[1]SenseTime Research [2]Shanghai AI Laboratory

[3]The Chinese University of Hong Kong

yuhuaijin36@gmail.com chengwei@sensetime.com

alvinliu@ie.cuhk.edu.hk wuwenyan0503@gmail.com junyilin@cuhk.edu.hk

A. Network Architecture

A.1. Blend MLP

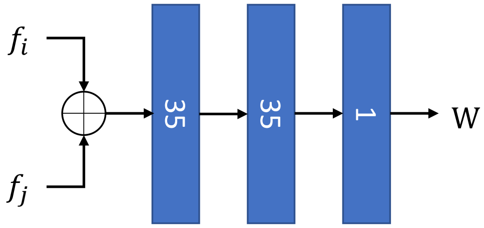

Figure 1. **Visualization of Blend MLP.** We use 3 layers MLP to calculate the belnd weights of correspondence features. It takes the concatenation of two correspondence features f_i and f_j as input, and output the final blend weight $\mathbf{W}$ of these two features.

A.2. Rendering Network

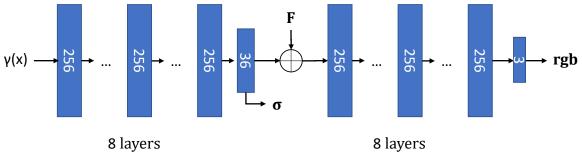

Figure 2. **Visualization of Rendering Network.** We use total 16 layers MLP to calculate the final color rgb and density σ. The first 8 layers of MLP take $\gamma(x)$ which means the positional encoding of point x as input, and output a 36 dimensional vector. One dimension of first 8 layers MLP output density value σ, and the remaining dimension is concatenated with the blended feature F. The second 8 layers MLP take the concatenation as input, and output the final rgb color value.

B. Ablation Study

B.1. Ablation Study on Shared Bidirectional Deformation Module

Fig. 3 illustrates that the Shared Bidirectional Deformation Module with consistent loss we proposed help produce more accurate deformation in regions like arms. Without this loss, the deformation in arms area tends to be bending and produce obvious artifacts.

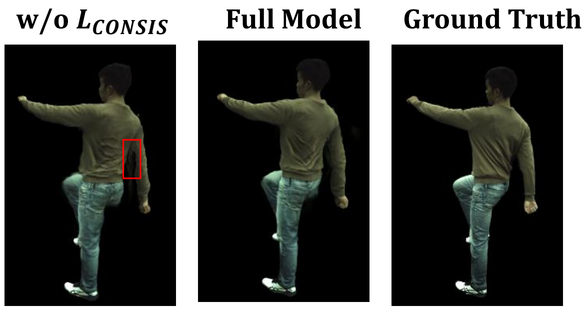

Figure 3. **Ablation of $\mathcal{L}_{\text{CONSIS}}$.** We compare the qualitative result without consistent loss.

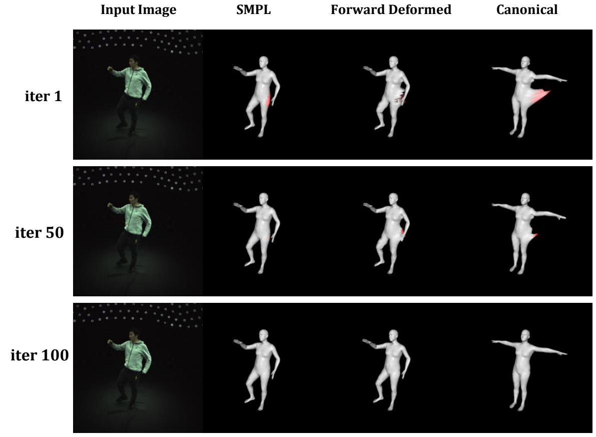

Figure 4. **Visualization of $\mathcal{L}_{\text{CONSIS}}$.** We optimize the $\mathcal{L}_{\text{CONSIS}}$ separately and visualize its effectiveness.

In Fig. 4 we use SMPL vertex as input of Shared Bidirectional Deformation Module and optimize it with learning rate of 5×10^{-6} separately. Points calculated by consistent loss which value $>= 0.05$ is colored red. The second column is the projection of the input image's SMPL vertex.

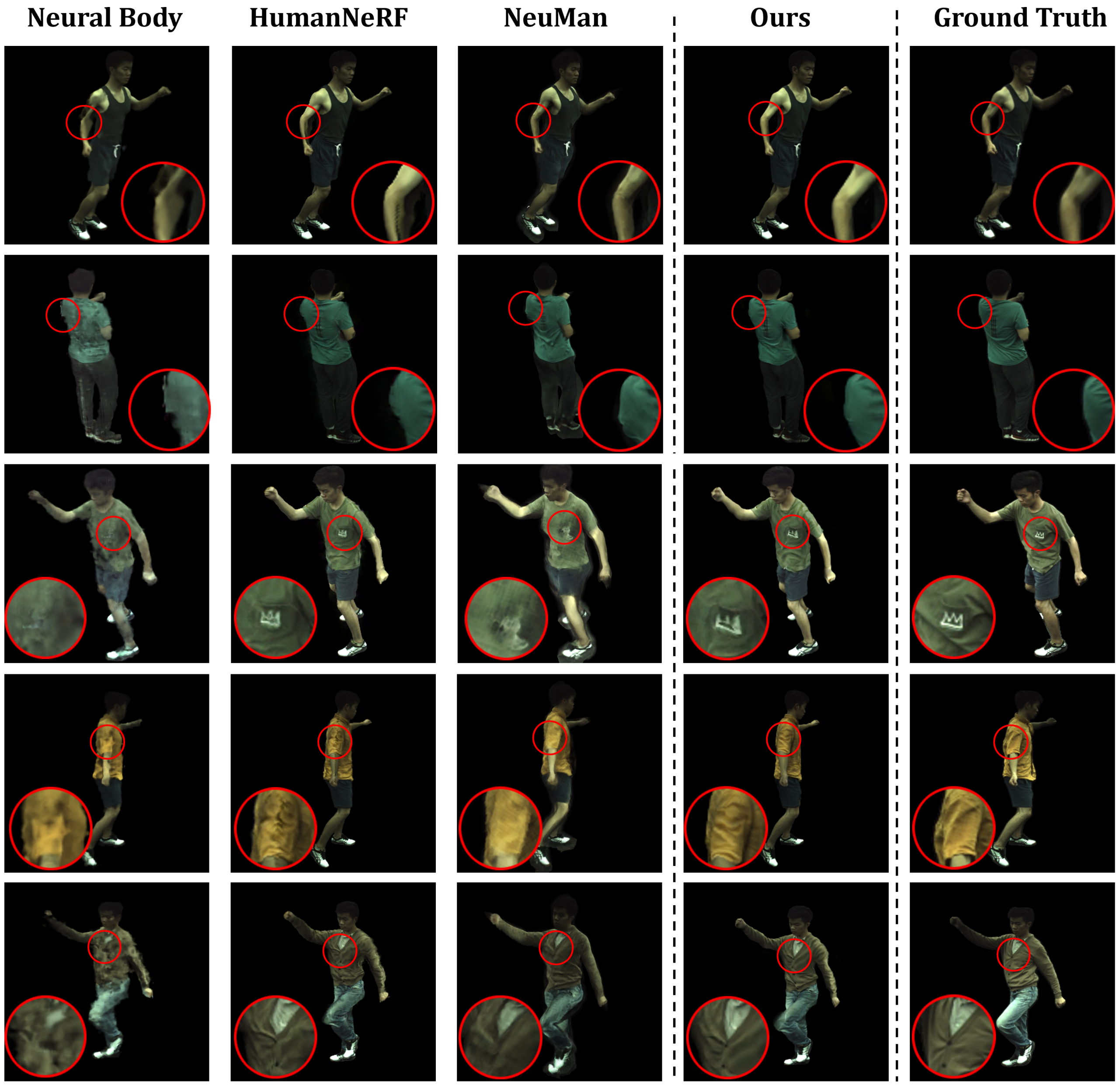

Figure 5. **Qualitative result of novel pose setting in ZJU-MoCap.** We compare the novel pose synthesis quality with baseline methods in ZJU-Mocap. Result shows that our method synthesise more realistic images in novel pose.

Canonical Vertex is the points deformed from SMPL vertex using backward deform of Shared Bidirectional Deformation Module. Forward Deformed Vertex means the points forward deformed from canonical vertex. The red points illustrate that the consistent loss we proposed can detect the points that are deform incorrectly. We further visualize the results of iteration1, 50 and 100. It shows that the red points reduce gradually, which means that the consistent loss correct and regularize the shared deformation weight.

B.2. Ablation Study on Forward Correspondence Search Module

Fig. 6 shows that the correspondence features produced by the Forward Correspondence Search module we proposed help produce more accurate color and texture in cloth regions. Without these features, the synthesis result tends to produce an unnatural texture in the cloth.

B.3. Ablation Study on Sequence Length

In order to explore our method's performance in different training sequence lengths, we evaluate the novel view synthesis results in different sample rates. The result is shown

	PSNR			SSIM			LPIPS		
frame_nums	w/o $\mathcal{L}_{\text{CONSIS}}$	w/o feat	full	w/o $\mathcal{L}_{\text{CONSIS}}$	w/o feat	full	w/o $\mathcal{L}_{\text{CONSIS}}$	w/o feat	full
380	27.87	27.84	**27.89**	0.9387	0.9389	**0.9389**	55.85	**54.91**	55.28
76	27.44	27.71	**27.77**	0.9348	0.9368	**0.9390**	63.16	60.94	**56.40**
38	27.65	27.71	**27.88**	0.9366	0.9358	**0.9390**	61.94	62.32	**56.40**
19	27.27	27.35	**27.52**	0.9341	0.9359	**0.9361**	67.56	62.82	**62.56**

Table 1. **Ablation study on Sequence Length.** We compare the novel view synthesis results in different training frame length. LPIPS* = LPIPS $\times 10^3$.

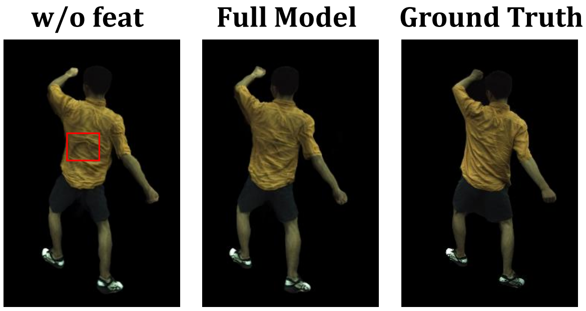

Figure 6. **Qualitative ablation of correspondence features.** We compare the qualitative result without correspondence features.

	Subject **377**			Subject **386**			Subject **387**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [3]	29.29	0.9693	39.40	30.71	0.9661	45.89	26.36	0.9520	62.21
HumanNeRF [5]	29.91	0.9755	23.87	32.62	0.9672	39.36	**28.01**	0.9634	35.27
Ours	**30.77**	**0.9787**	**21.67**	**32.97**	**0.9733**	**32.73**	27.93	**0.9633**	**33.45**

	Subject **392**			Subject **393**			Subject **394**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [3]	28.97	0.9615	57.03	27.82	0.9577	59.24	28.09	0.9557	59.66
HumanNeRF [5]	30.95	0.9687	34.23	28.43	0.9609	36.26	28.52	0.9573	39.75
Ours	**31.24**	**0.9715**	**31.04**	**28.46**	**0.9622**	**34.24**	**28.94**	**0.9612**	**35.90**

Table 2. **Novel view synthesis quantitative comparison on ZJU-MoCap dataset.** We show the results of each subject in the table. LPIPS* = LPIPS $\times 10^3$.

	Subject **377**			Subject **386**			Subject **387**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [3]	29.08	0.9679	41.17	29.76	0.9647	46.96	26.84	0.9535	60.82
HumanNeRF [5]	29.79	0.9714	28.49	32.10	0.9642	41.84	28.11	0.9625	37.46
Ours	**30.46**	**0.9781**	**20.91**	**32.99**	**0.9756**	**30..97**	**28.40**	**0.9639**	**35.06**

	Subject **392**			Subject **393**			Subject **394**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [3]	29.49	0.9640	51.06	28.50	0.9591	57.072	28.65	0.9572	55.78
HumanNeRF [5]	30.20	0.9633	40.06	28.16	0.9577	40.85	29.28	0.9557	41.97
Ours	**30.98**	**0.9711**	**30.80**	**28.54**	**0.9620**	**34.97**	**30.21**	**0.9642**	**32.80**

Table 3. **Novel pose synthesis quantitative comparison on ZJU-MoCap dataset.** We show the results of each subject in the table. LPIPS* = LPIPS $\times 10^3$.

in Table 1. We use subject 394 in ZJU-MoCap as testing and sample in rates of 1, 5, 10, and 20, and the number of frames is 380, 76, 38, 19 respectively. We follow HumanNeRF [5] to evaluate in 22 cameras not seen in training in 30 sample rates. And follow NeuralBody [3] to evaluate the subject in a 3d bounding box to avoid getting an inflated PSNR value. The result shows that the relationship between sequence length and generated quality is not linear. And When in small training frame numbers like 19 frames, the module we proposed helps to retain the more realistic result.

C. Evaluation details

In order to evaluate novel view and novel pose synthesis, we split the frames in camera 1 in a 4:1 ratio as Set A and B. We only use Set A for training. For novel view evaluation, we sample the synchronous video frames of Set A for all unseen 22 cameras at the rate of 30. For novel pose evaluation, we sample at the same rate for the synchronous frames of Set B for all cameras. In general, the evaluation frames for the novel view would be 242 frames, and 184 frames for the novel pose setting.

D. Condition on view direction

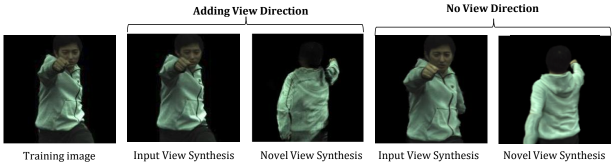

Figure 7. **Comparison of adding view direction.** We compare the results of conditioning blend weights on view direction and no condition.

We have tried to condition our feature blend weights on view direction when designing our model, but we found overfitting problems that were also found by works like StyleSDF [2] and StyleNeRF [1]. As shown in Fig. 7, conditioning weights on view direction can help to overfit in the training frames and synthesize realistic results even in some highly reflective areas, but generate lots of artifacts when synthesizing novel view images. We find conditioning blend weights on view direction weakens the generalization ability to the novel view.

E. More results

To compare the generated results, we visualize the novel pose synthesis results in ZJU-MoCap dataset of NeuralBody, HumanNeRF and our method in Fig. 5. NeuralBody tends to generate vague images with large noise in novel poses. HumanNeRF tends to produce some black line artifacts in some detailed areas. We also show the extract comparison with HumanNeRF under the challenge poses

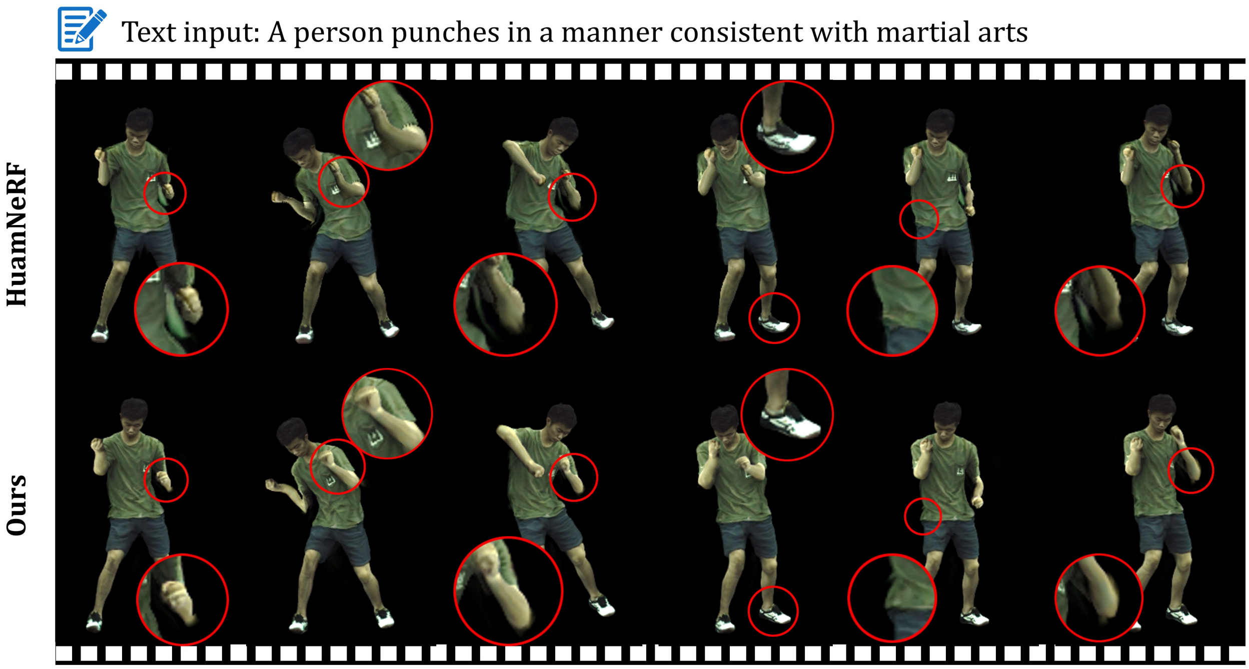
Text input: A person punches in a manner consistent with martial arts
HuamNeRF
Ours

Figure 8. **Qualitative results on challenge poses generated by MDM [4] through text input.** We evaluate our method driven by challenge pose sequence generated by MDM model.

generated by MDM [4] model in Fig. 8. The result shows that our methods can retain these detail due to the correct deform and help with guided features.

We show the detailed quantitative results of each subject we compare in ZJU-MoCap dataset in Table. 3 and Table. 2. Though the PSNR metric of NeuralBody seems good in value, they synthesize poor visual quality images in both novel view and novel pose (as PSNR prefers smooth results). Our improvement in PSNR over HumanNeRF is not significant, and slightly lower in subject 387 when testing in the novel pose. But in LIPIS, our method has larger improvement in novel view and novel pose setting respectively. We can see that our MonoHuman framework outperforms existing methods in most metrics in both settings.

References

[1] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3

[2] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 3

[3] 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. *CVPR*, 2021. 3

[4] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 4

[5] 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. 3