PersonNeRF: Personalized Reconstruction from Photo Collections Supplementary Material

A. Network Architecture

Fig. 1 and Fig. 2 show the network design of our canonical MLP and pose correction MLP. Specifically, we provide the details of how we incorporate appearance embedding ℓ^{app} as well as pose embedding ℓ^{pose} vectors into the corresponding networks.

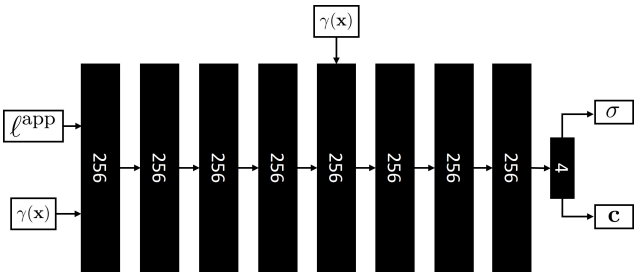


Figure 1. Canonical MLP netwrok. We use an 8-layer MLP with width=256 that takes as input positional encoding γ of position $\mathbf{x}$ and appearance embedding vector ℓ^{app} with dimension=256. The network outputs color $\mathbf{c}$ and density σ. There is a skip connection that concatenates $\gamma(\mathbf{x})$ to the fifth layer. We use ReLU activations after each fully connected layer. For the output layer, we use a ReLU activation for the density value σ to ensure non-negativity and a *sigmoid* activation for the color $\mathbf{c}$ to constrain values between 0 and 1.

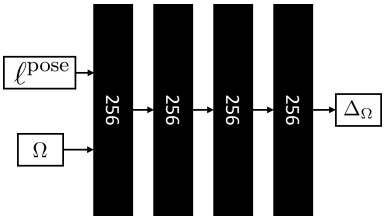


Figure 2. Pose correction MLP network. We use a 4-layer MLP with width=256 that takes as input joint angles Ω and a pose embedding vector ℓ^{pose} with dimension=16. The network produces the residuals of joint angles that are added back to the input pose to refine the body pose prediction.

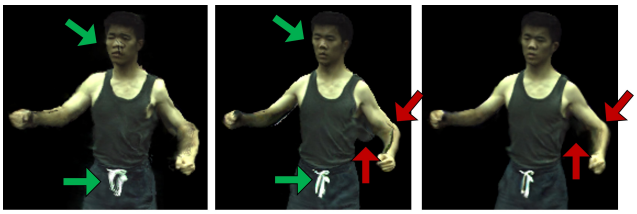
B. Experiments on ZJU-MoCap dataset

B.1. Experimental Setup

We additionally performed experiments on the ZJU-MoCap dataset [1], which provides ground-truth unseen views that enable computation of metrics and analysis of performance on sparse/dense data inputs. We selected subjects 377, 392, and 393–the same individual in different clothing. We evenly selected 10 frames from camera-1 videos to represent "sparse data" (ZJU-Sparse). For "dense data", we used the entire video (ZJU-Dense). The remaining 22 camera views were used for evaluation. We report PSNR, SSIM, and LPIPS* (LPIPS $\times 10^3$) metrics and highlight the best and second-best values.

B.2. Results on ZJU-Sparse dataset

Table 1 shows comparisons on the ZJU-Sparse dataset. Our method outperforms HumanNeRF and our separate-network of our approach in SSIM and LPIPS, with the largest margins in LPIPS, a better measure of visual quality as seen in Fig. 3).



(a) HumanNeRF (b) Ours (Separate Nets) (c) Ours (Single Net)

Figure 3. Our approach enhances details in the face and clothing (green). Single network training further improves shape and appearance consistency (red).

B.3. Results on ZJU-Dense dataset

We conducted an analysis on the ZJU-Dense dataset. As shown in Table 2, our method, which was designed for sparse inputs, still demonstrates improvement. The improvement is particularly noticeable when we remove the regularization designed for handling sparse observations,

	Subject **377**			Subject **392**			Subject **393**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
HumanNeRF	29.59	0.9721	33.53	30.38	0.9626	51.03	27.56	0.9535	55.69
Ours (Separate Nets)	29.61	0.9734	27.66	29.48	0.9640	42.65	27.28	0.9537	47.53
Ours (Single Net)	29.55	0.9737	26.62	30.03	0.9665	38.79	27.59	0.9558	46.16

Table 1. Comparison on ZJU-Sparse dataset (10 images per subject).

indicating that the shared latent space is a promising area for exploration even for dense video.

	PSNR ↑	SSIM ↑	LPIPS* ↓
HumanNeRF	29.92	0.9684	30.97
Ours	29.81	0.9692	30.40
Ours w/o reg.	29.98	0.9700	28.47

Table 2. Our method outperforms HumanNeRF on the ZJU-Dense dataset (an entire video per subject). The best quality is achieved when the regularization designed for sparse input is removed

B.4. Ablation Study of Photo Numbers

In addition, we analyzed how the performance is affected by the number of training images. We do see improvement with more photos on ZJU-MoCap dataset, though with diminishing returns. Table 3 shows numerical results.

# of images per subject	PSNR ↑	SSIM ↑	LPIPS* ↓
a video (∼ 600 frames)	29.81	0.9692	30.40
20 images	29.45	0.9679	32.38
10 images	29.06	0.9653	37.19

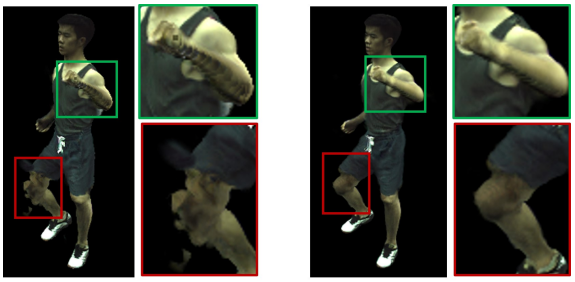
Table 3. The ablation study of photo numbers run on ZJU-MoCap.

B.5. Novel Pose Evaluation

	PSNR ↑	SSIM ↑	LPIPS* ↓
Separate Nets	29.08	0.9691	31.05
Single Net	30.06	0.9727	27.75

Table 4. We achieve better performance for unseen poses when training all photos with different appearances in a single network.

Our focus was on maintaining original poses, not reposable avatar creation, avoiding, e.g., making a famous tennis player perform actions they never did. That said, experiments suggest that our method is capable of handling poses that have not been previously encountered, especially when all photos are trained within a single network. We performed an analysis on ZJU-Sparse dataset (10 frames per subject) where we applied the learned model to body poses from the unseen frames (∼600 frames per subject). As presented in Table 4, single-network training achieves better performance in all metrics. This is because the optimized single, universal motion weight volume can be constrained by a much larger number of poses compared to the separate ones, resulting in a better solution. Fig. 4 shows the visual comparison.



(a) Separate Nets (b) Single Net

Figure 4. Single-network training improves appearance consistency (green) and maintains body shapes (red) for unseen poses.

C. Additional Results

In addition to Roger Federer, we demonstrate our method on a wide variety of subjects that cover different genders and skin tones. In particular, we show results on three tennis athletes, Novak Djokovic, Serena Williams, and Rafael Nadal where each has three appearance sets in the datasets we collected. Moreover, we applied our method to self-captured data (*rugby, hoodie*) provided by HumanNeRF [2] where we evenly select 15 frames from the videos. We present quantitative results in FID in Table 5 and visually compare them with HumanNeRF [2] in Fig. 5. The quality improvement over the related work is similar to the case of Roger Federer.

D. More Visualizations of Personalized Space

In the paper, we show a visualization of (appearance, camera view) plane of the reconstructed space of Roger Federer. Here we show the other two planes, (appearance, body pose) plane in Fig. 6 and (body pose, camera view) plane in Fig. 7 where we keep the camera view and appearance fixed, respectively.

Additionally, we show visualizations of the rebuilt personalized spaces of the other 3 persons, Novak Djokovic in Fig. 8, 9 and 10, Serena Willams in Fig. 11, 12 and 13, and Rafael Nadal in Fig. 14, 15 and 16.

References

[1] 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. 1

[2] 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 (CVPR)*, pages 16210–16220, June 2022. 2, 3

	Novak Djokovic			Serena Willams			Rafael Nadal			Self-captured [2]	
	2013	**2016**	**2019**	**2009**	**2010**	**2011**	**2014**	**2019**	**2022**	***rugby***	***hoodie***
HumanNeRF [2]	87.07	62.01	64.17	104.23	100.52	113.41	90.04	64.95	76.68	102.13	109.31
Our method	81.38	57.14	58.74	87.81	90.70	85.17	80.95	62.75	61.71	97.00	96.99

Table 5. Comparison to related work: FID is computed per subject per year. In addition, We evenly select 15 frames from self-captured data (*rugby, hoodie*) provided by HumanNeRF [2]. Lower FID score is better.

Figure 5. Visual comparisons to HumanNeRF [2] on the tennis athletes (*Novak Djokovic, Serena Willams, Rafael Nadal*) and self-captured subjects (*rugby, hoodie*) from HumanNeRF. *Photo credits to Getty Images.*

Figure 6. Visualization of the (appearance, body pose) plane of the reconstructed space of Roger Federer.

Figure 7. Visualization of the (body pose, camera view) plane of the reconstructed space of Roger Federer.

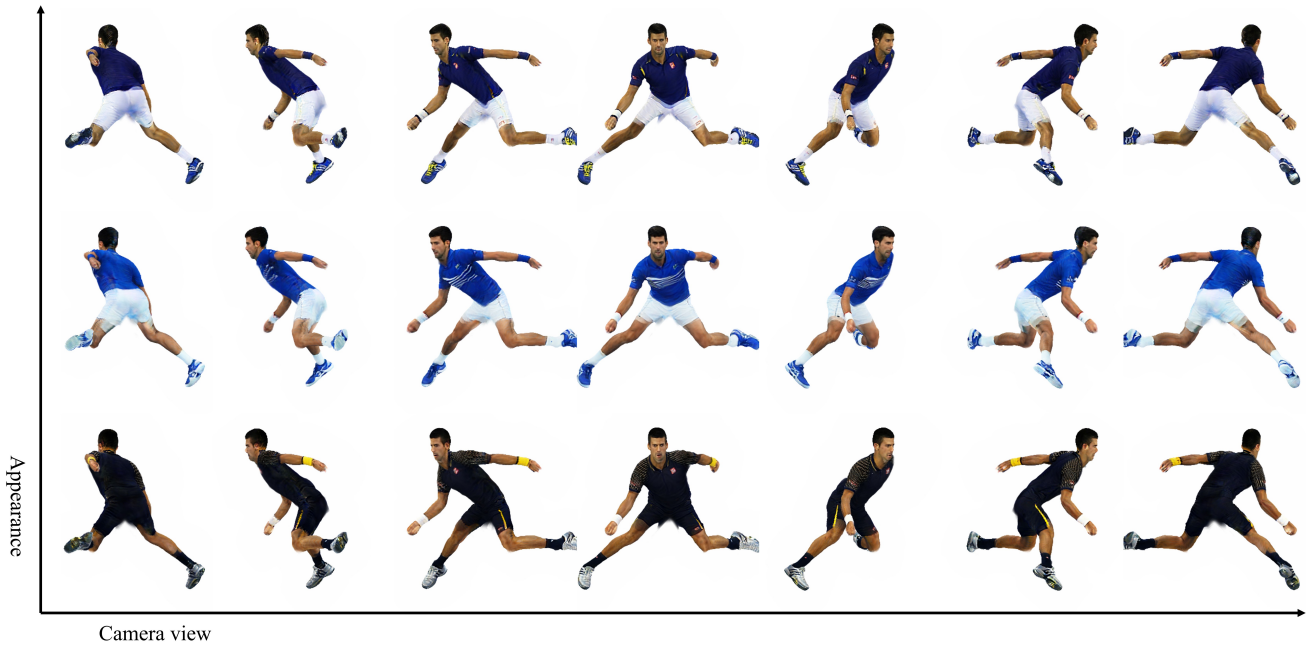


Figure 8. Visualization of the (appearance, camera view) plane of the reconstructed space of Novak Djokovic.

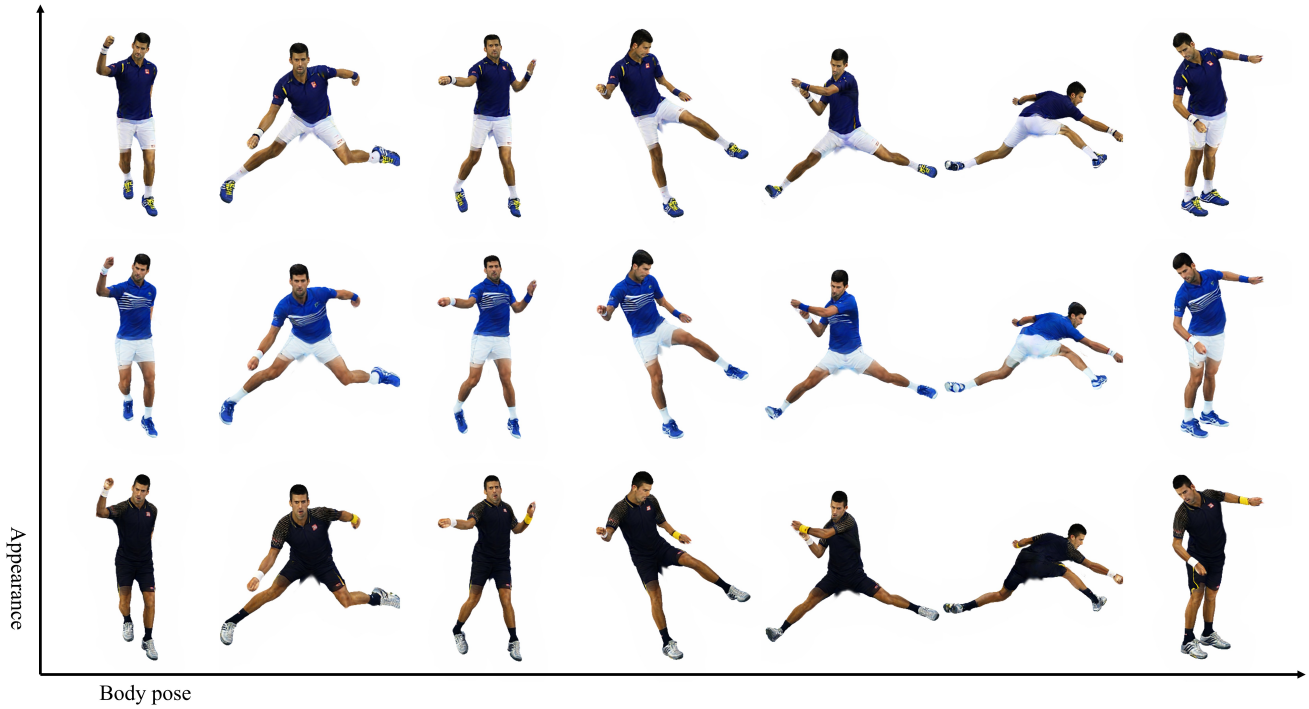


Figure 9. Visualization of the (appearance, body pose) plane of the reconstructed space of Novak Djokovic.

Figure 10. Visualization of the (body pose, camera view) plane of the reconstructed space of Novak Djokovic.

Figure 11. Visualization of the (appearance, camera view) plane of the reconstructed space of Serena Willaims.

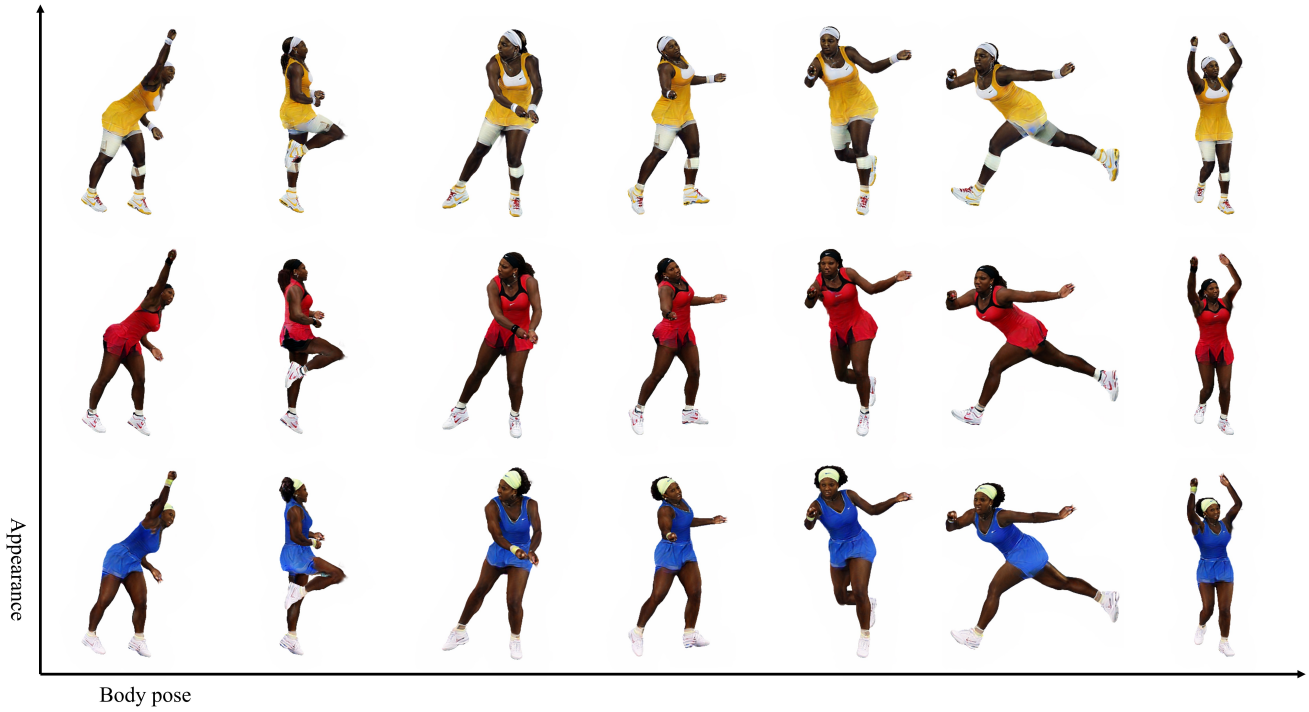


Figure 12. Visualization of the (appearance, body pose) plane of the reconstructed space of Serena Williams.

Figure 13. Visualization of the (body pose, camera view) plane of the reconstructed space of Serena Willaims.

Figure 14. Visualization of the (appearance, camera view) plane of the reconstructed space of Rafael Nadal.

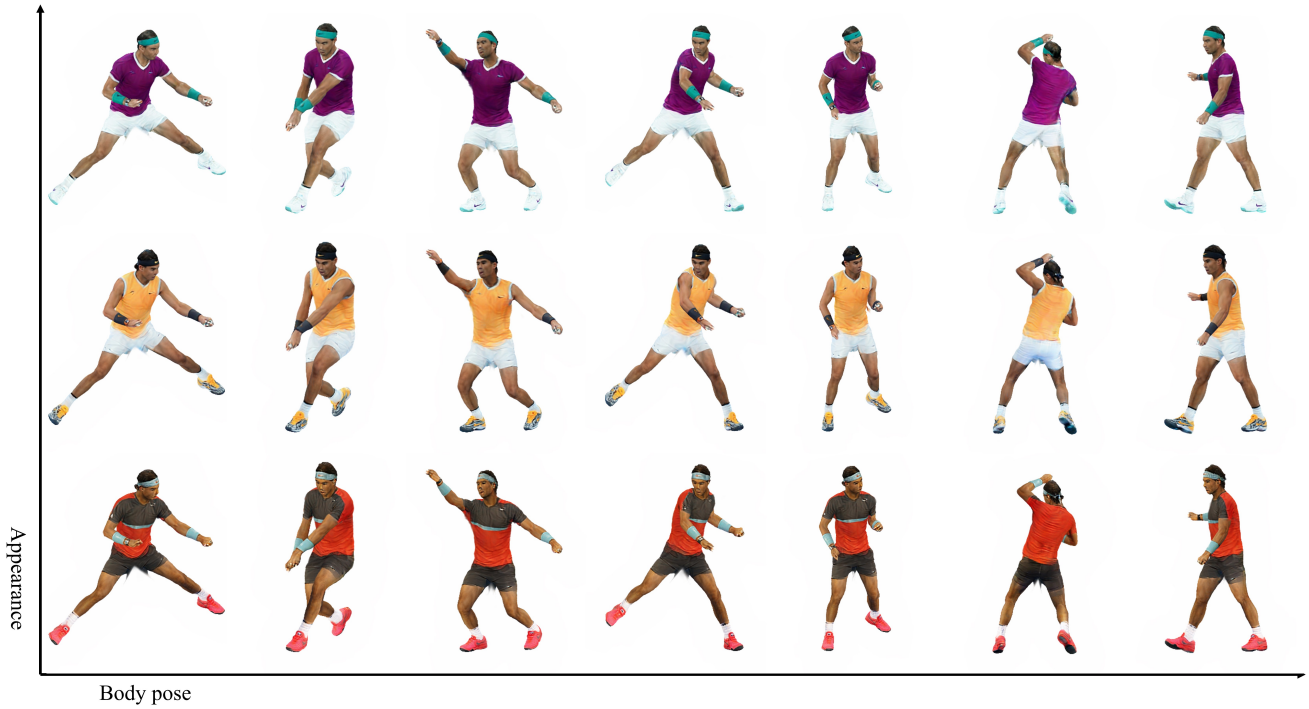


Figure 15. Visualization of the (appearance, body pose) plane of the reconstructed space of Rafael Nadal.

Figure 16. Visualization of the (body pose, camera view) plane of the reconstructed space of Rafael Nadal.