

Supplementary Material for “UV Volumes for Real-time Rendering of Editable Free-view Human Performance”

Yue Chen^{1*} Xuan Wang^{2*} Xingyu Chen¹ Qi Zhang²
 Xiaoyu Li² Yu Guo^{1†} Jue Wang² Fei Wang¹

¹Xi'an Jiaotong University ²Tencent AI Lab

Here we provide more implementation details and experimental results. We encourage the reader to view the video results included in the supplementary materials for an intuitive experience of editable free-view human performance.

A. Network Architectures

A.1. Volume Generator

We utilize the volume generator to construct UV volumes, which is presented in Figure 1. We take the human pose as input to the SMPL model and animate human point clouds in different poses. Then we follow the previous work [6] to anchor a set of time-invariant latent codes to the posed human point cloud and voxelize the point cloud. We follow the network architecture of [6] to model the 3D sparse CNN, and reduce the channels from 352 to 64, since the UV volumes only capture low-frequency semantic information.

A.2. Density, IUV and Color Network

We present architectures of density network M_σ , IUV network M_{uv} and color network M_c in Figure 2, Figure 3 and Figure 5, respectively.

A.3. Texture Generator

Figure 4 shows the architecture of convolutional texture generator network G . For each human body part of $\{i\}_{i=1}^{24}$, the texture generator generates a corresponding neural texture stack of $\{\mathcal{E}_i\}_{i=1}^{24}$. To predict the specific pose-dependent \mathcal{E}_i , we concatenate human pose vector θ with a one-hot body part label vector \mathbf{k}_i as input to the texture generator. We forward propagate the generator network G once to predict all the 24 neural textures with a batch size of 24. The CNN-based module is developed to extract the local relation of neural texture stack with respect to the human pose. The output spatial neural texture stacks (NTS) will be used for UV unwrapping subsequently.

*Authors contributed equally to this work.

†Corresponding Author.

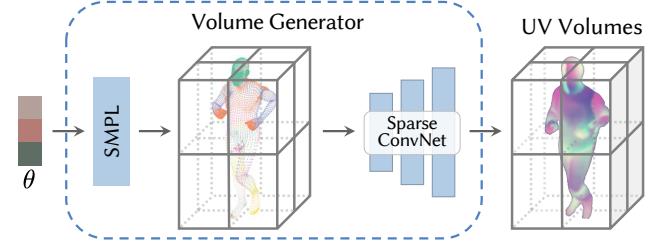


Figure 1. Architecture of the volume generator. It takes human pose θ as input, drives the point clouds parameterized by the SMPL model under the control of θ , and generates the UV volumes using a 3D sparse CNN to encode a set of latent codes \mathbf{z} anchored on the posed point clouds.

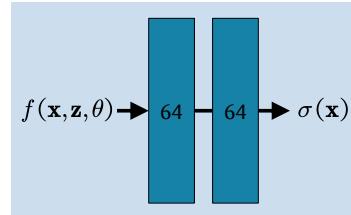


Figure 2. Architecture of the density network. The network takes the feature vector $f(\mathbf{x}, \mathbf{z}, \theta)$ at point \mathbf{x} interpolated by the generated UV volumes and outputs density $\sigma(\mathbf{x})$ using ReLU activation. The shallow density MLP M_σ consists of 2 fully-connected layers with 64 channels.

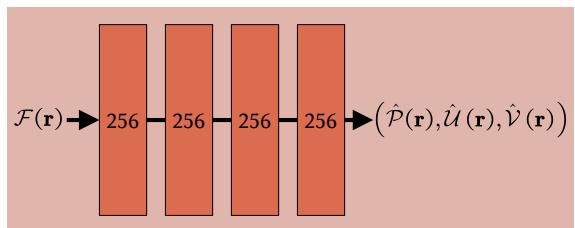


Figure 3. Architecture of the IUV network. The network takes the rendered UV feature vector $\mathcal{F}(\mathbf{r})$ at camera ray \mathbf{r} and outputs view-invariant texture coordinates $(\hat{\mathcal{P}}(\mathbf{r}), \hat{\mathcal{U}}(\mathbf{r}), \hat{\mathcal{V}}(\mathbf{r}))$ using sigmoid activation. The IUV MLP M_{uv} is modeled by 4 fully-connected layers of 256 channels.

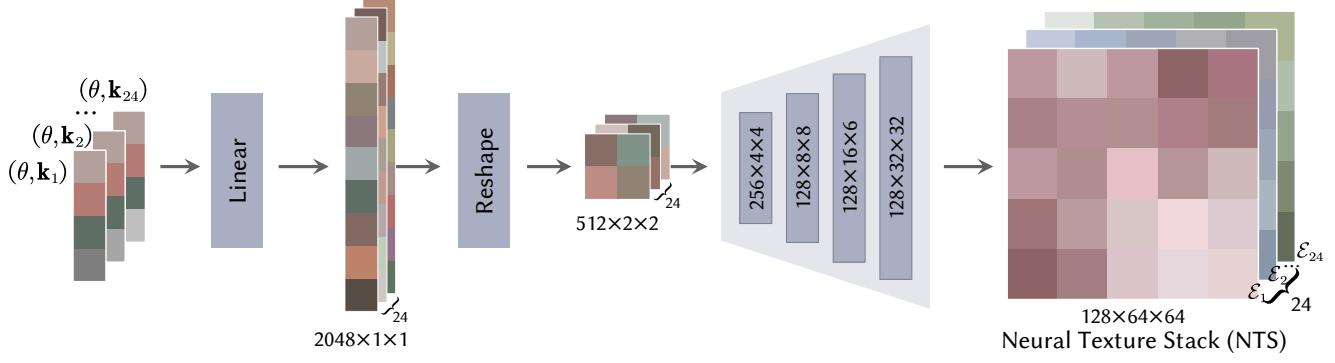


Figure 4. Convolutional texture generator network G consists of 5 convolution layers to get the neural texture stack \mathcal{E}_i with 128 dimensions.

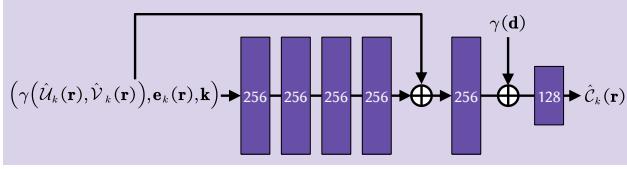


Figure 5. Architecture of the color network. The network takes positional encoding of texture coordinates $\gamma(\hat{\mathcal{U}}_k(\mathbf{r}), \hat{\mathcal{V}}_k(\mathbf{r}))$ along with the sampled texture embeddings at locations $(\hat{\mathcal{U}}_k(\mathbf{r}), \hat{\mathcal{V}}_k(\mathbf{r}))$ and a one-hot part label vector \mathbf{k} . The color MLP M_c is modeled by 5 fully-connected layers of 256 channels, including a skip connection that concatenates inputs to the fourth layer’s activation. The feature vector of the fifth layer is processed by an additional layer with 128 channels, along with positional encoding of input viewing direction $\gamma(\mathbf{d})$. A final layer with a sigmoid activation outputs view-dependent RGB color $\hat{C}_k(\mathbf{r})$ of body part k .

B. Additional Implementation Details

We set λ_{vgg} to 5×10^{-2} and λ_s to 1×10^{-1} . The λ_p and λ_{uv} are exponential annealing from 1×10^{-1} and 1×10^{-0} to 1×10^{-3} and 5×10^{-2} with $k = 4 \times 10^{-2}$:

$$\begin{aligned} \lambda_p &= \max(1 \times 10^{-1} e^{-k \cdot \text{epoch}}, 1 \times 10^{-3}) \\ \lambda_{\text{uv}} &= \max(1 \times 10^{-0} e^{-k \cdot \text{epoch}}, 5 \times 10^{-2}) \end{aligned} \quad (1)$$

As shown in Figure 6a, the weight of UV-metric is large at the beginning because UV volumes require a warm-start to satisfy the UV unwrap defined by DensePose [2], and then drop rapidly within 100 epochs because DensePose outputs are not accurate. After 100 epochs, UV-metric becomes a regular term used to constrain the solution space of UV volumes.

UV volumes should be learned primarily in the early stages of training, because The NTS makes sense only after the UV volumes warm-start and a coarse geometry is constructed. Conversely, later in training, we optimize NTS to fit high-frequency signals rather than UV coordinates. Therefore, we use two optimization strategies to train the UV volumes and NTS branch. Their learning rates

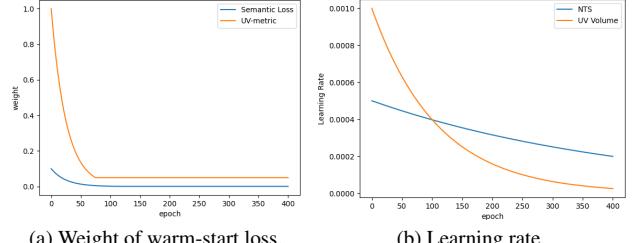


Figure 6. (a) The blue represents the weight of semantic loss λ_p and the orange represents the weight of UV-metric loss λ_{uv} . (b) The orange represents the learning rate of the UV volumes branch and the blue represents the learning rate of the NTS branch.

start from 1×10^{-3} and 5×10^{-4} with a decay rate of $\gamma = 1 \times 10^{-1}$, respectively, and decay exponentially along the optimization, as shown in Figure 6b.

$$\begin{aligned} l_{\text{nts}} &= 5 \times 10^{-4} \gamma^{\frac{\text{epoch}}{1000}} \\ l_{\text{uv}} &= 1 \times 10^{-3} \gamma^{\frac{\text{epoch}}{250}} \end{aligned} \quad (2)$$

In our experiments, we sample camera rays all over an entire image and 64 points along each ray between near and far bounds. The scene bounds are estimated based on the SMPL model. We adopt the Adam optimizer [3] for training our model. We conduct the training on a single NVIDIA A100 GPU. The training on 26-view videos of 100 frames at 960×540 resolution typically takes around 200k iterations to converge (about 20 hours).

C. Additional Baseline Method Details

DyNeRF (DN) [4]. We reimplement DN by following the procedure in their paper to train the model on video sequences of a moving human.

Neural Body (NB) [6]. We use the NB code open-sourced by the authors at <https://github.com/zju3dv/neuralbody> and follow their procedure for training on video sequences of a moving human.

Animatable NeRF (AN) [5]. We use the AN code open-sourced by the authors at <https://github.com/zju3dv/animatable> and follow their procedure for training on video sequences of a moving human.

D. Additional Ablation Studies

We conduct ablation studies on performer p1 of CMU dataset. As shown in Table 1, Table 2, Table 3, Figure 10, Figure 11, and Figure 12, we analyze the effects of different losses for our proposed approach, different types of NTS, different resolutions of NTS, different methods to model the view-dependent color, and different experimental settings of video frames and input views.

D.1. Effects of Different Losses

Impact of warm-start loss. No Warm-start Loss (w/o \mathcal{L}_{uv}) is an ablation built upon our full model by eliminating the warm-start loss. As shown in Figure 10 and Figure 11, w/o \mathcal{L}_{uv} suffers from ambiguity, like the belt on the pants and meaningless texture. This comparison indicates that the warm-start loss yields better information reuse of different frames by transforming the observation XYZ coordinates to canonical UV coordinates defined by the consistent semantic and UV-metric loss.

Impact of perceptual loss. No Perceptual Loss (w/o \mathcal{L}_{vgg}) is an ablation that uses the same model but training without the perceptual loss. As shown in Figure 10 and Figure 11, w/o \mathcal{L}_{vgg} suffers from blur, like the number 9 on the shirt and distorted number 9 on the texture. This comparison illustrates that perceptual loss can improve the visual quality of synthesized images by supervising the structure of the renderings from local to global during training.

Impact of silhouette loss. No Silhouette Loss (w/o \mathcal{L}_s) is an ablation built upon our full model by eliminating the silhouette loss. As shown in Figure 10, w/o \mathcal{L}_s suffers from artifacts around the performance because there is no the warm-start supervision of semantic and UV-metric labels around the boundary. This comparison demonstrates that silhouette loss is essential for us to model fine-grained geometry.

D.2. Neural Texture Stacks

We performed two ablations: 1) different types of NTS; 2) NTS at different resolutions to illuminate the design decisions for the proposed *Neural Texture Stacks*.

Different Types of NTS. We evaluate our proposed CNN-based Spatial NTS against three ablations: Global NTS, Local NTS, and Hyper NTS. Global NTS (in Figure 7) is built upon our full model by replacing local texture embedding $e_k(r)$ with global pose θ . Local NTS (in Figure 8) transforms observation UV coordinates $(\hat{U}_t^k(r), \hat{V}_t^k(r))$

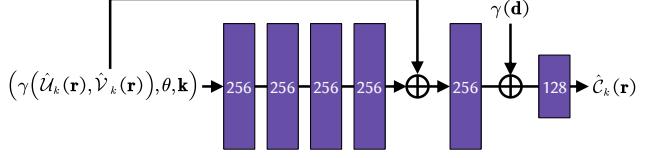


Figure 7. The Global-NTS model directly takes a pose θ as a condition to generate the color. Its architecture is similar to our color MLP M_c (in Figure 5) except for replaces texture embedding $e_k(r)$ with pose θ .

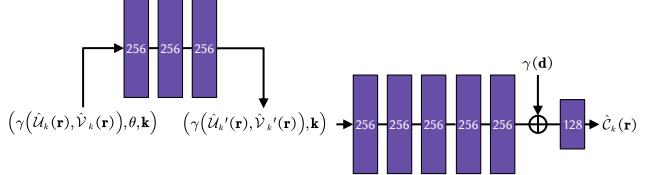


Figure 8. The Local-NTS model transforms UV coordinates $(\hat{U}_t^k(r), \hat{V}_t^k(r))$ to $(\hat{U}_t^{k'}(r), \hat{V}_t^{k'}(r))$ using a deformation field conditioned on pose θ and modeled by three fully-connected layers of 256 channels. Then we use the transformed UV coordinates, part label vector k and viewing direction d as inputs to the subsequent MLP modeled by five fully-connected layers of 256 channels and one fully-connected layer of 128.

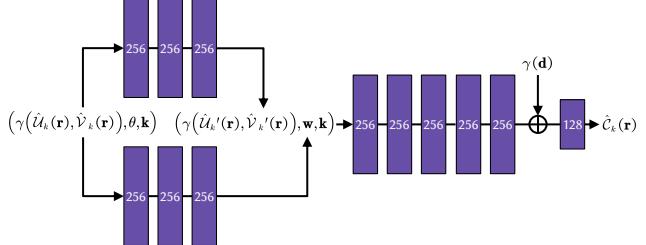


Figure 9. The Hyper-NTS model adds an ambient MLP to the local-NTS model to parameterize the deformation field, which yields warped UV coordinates $(\hat{U}_t^{k'}(r), \hat{V}_t^{k'}(r))$ and a coordinate w in the ambient space. Both outputs, part label vector k and viewing direction d , are concatenated to the subsequent MLP to produce view-dependent colors. The ambient MLP has the same architecture as the deformation MLP.

to canonical UV coordinates $(\hat{U}_t^{k'}(r), \hat{V}_t^{k'}(r))$ using a deformation field. Hyper NTS (in Figure 9) adds an ambient MLP to the local-NTS model to model a slicing surface in hyperspace, which yields a coordinate w in an ambient space.

Table 1 shows the quantitative results on different types of NTS (i.e., global, local, hyper and spatial). It can be seen that the local-NTS model has the worst performance, which is the most limited among these methods. Local NTS only allows coordinate transformation but cannot generate new topological space, which is totally incapable of modeling the topologically varying texture given different poses.

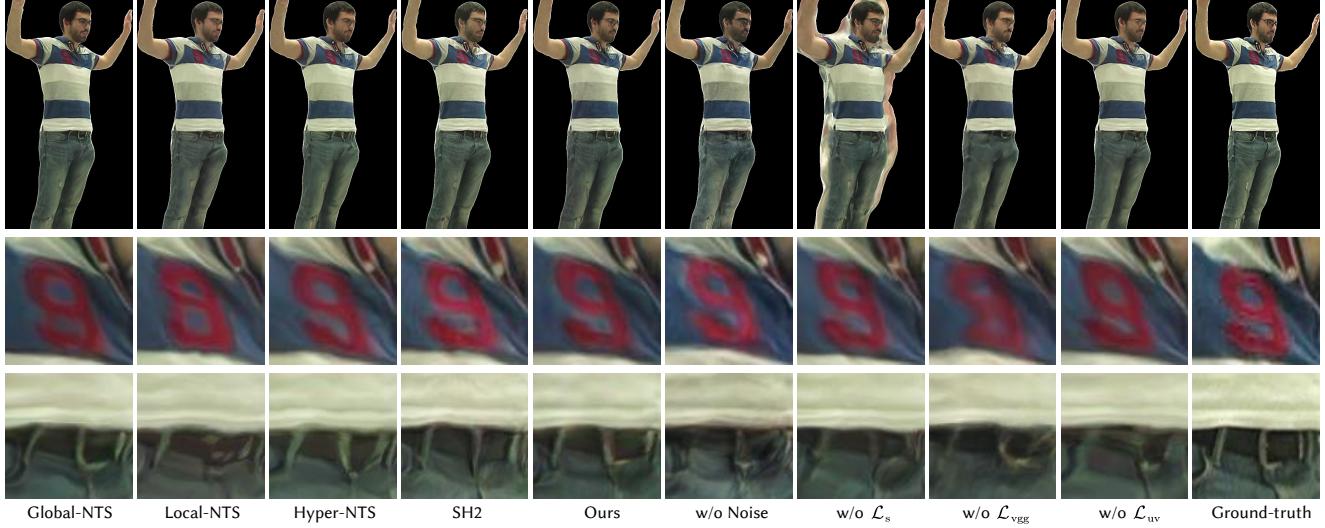


Figure 10. Renderings of our model against ablations.

Type	Novel View Synthesis			Novel Pose Generation		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Global NTS	29.86	0.966	0.062	26.18	0.927	0.074
Local NTS	29.82	0.965	0.061	26.15	0.926	0.074
Hyper NTS	30.08	0.966	0.059	26.19	0.927	0.073
Spatial NTS	30.38	0.966	0.036	26.20	0.927	0.073

Table 1. Effects of different types of NTS on our model.

As shown in Figure 10, it fails to reconstruct the belt on the pants due to the topological variation ((the belt appears when the performer raises his hand and disappears when he puts his hand down because the shirt covers it).

Since the texture of Global-NTS directly condition on global pose without restriction, it is easy for Global-NTS to generate a new topological space, as shown in Figure 10, Global-NTS successfully reconstructs the belt on the pants. However, the method that globally models the texture variation is hard to reuse the information of different observation spaces, which leads to ambiguous textures. As shown in Figure 10, the outline of the number 9 on the shirt is not clear and even connected, making it look like an 8. The ambiguous textures are shown in Figure 11, where the number looks more like a red stain. Lack of local mapping makes a relatively poor performance of Global-NTS as demonstrated in Table 1.

Hyper-NTS can model local texture changes by coordinate transformation and generate new topological spaces simultaneously, so it performs better than Global-NTS and Local-NTS. However, it is a thorny issue to tune the dimension of coordinate w . If the dimension of the coordinates w is too high, Hyper-NTS works as Global-NTS, while functioning as Local-NTS if too low.

In contrast, as shown in Table 1, our CNN-based Spatial NTS outperforms all other NTS, which benefits from the nature of convolution operation capturing local 2D texture changes. At the same time, the MLP only needs to model the local mapping between the neural texture stack and RGB color.

As demonstrated in Figure 10 and Figure 11, our CNN-based Spatial NTS can accurately capture high-frequency details like numbers or wrinkles on shirts, glasses, and a belt on pants.

NTS at Different Resolutions. As shown in Figure 4, we generate an NTS at a resolution of 64×64 . Choosing the resolution of NTS provides a trade-off between quality and memory. We analyze the impacts of resolution in Figure 12, where we report test quality vs. resolution for the dataset of CMU-p2 on PSNR, SSIM and LPIPS metrics. Restricted to memory limitations, NTS has a maximum resolution of 128. It can be observed that the larger resolution of NTS, the better the model performed on novel view synthesis and the novel pose generalization tasks. In this analysis, we found 64×64 to be a favorable optimum in our applications, where NTS at 128×128 resolution is not much better than at 64×64 resolution but costs more memory and time, so we choose 64×64 resolution in all other experiments and recommend it as the default to practitioners.

D.3. View-dependent Color

Ray Direction Noise. To model view-dependent RGB color of human performances, we apply positional encoding $\gamma(\cdot)$ [7] to the viewing direction, and pass the encoded viewing direction, UV map and the sampled texture embedding into the color network M_c to decode the view-dependent color $\hat{C}_k(r)$ of camera ray r .

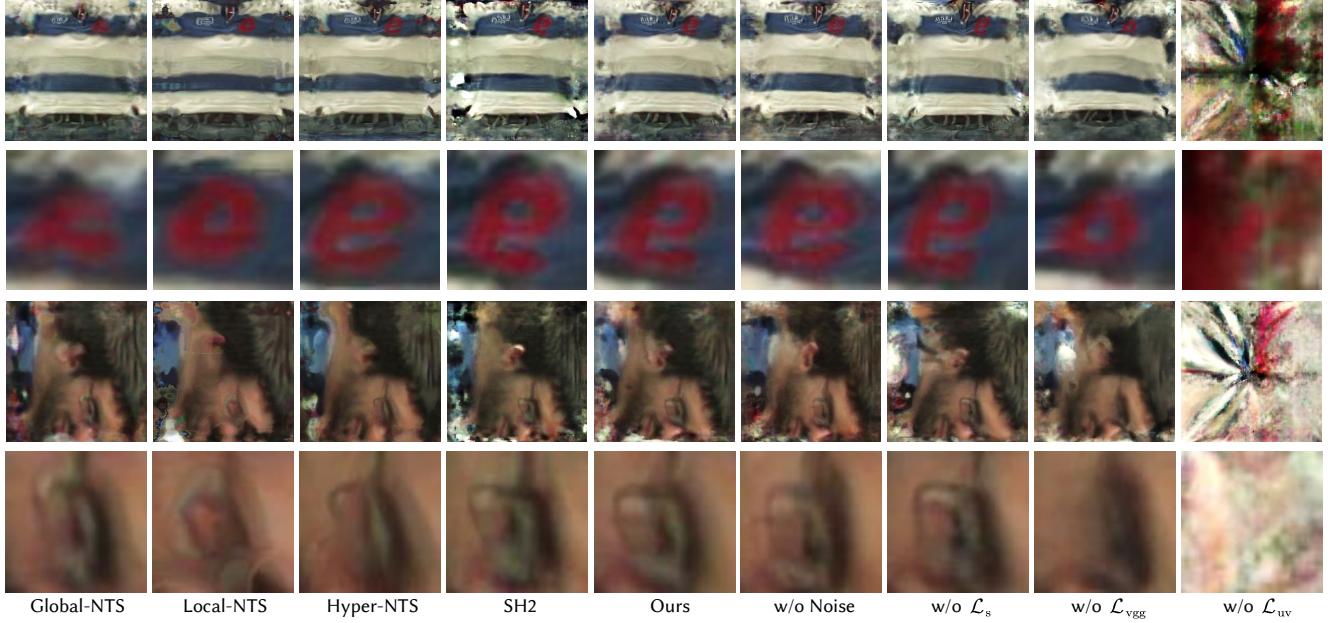


Figure 11. Textures of our model against ablations.

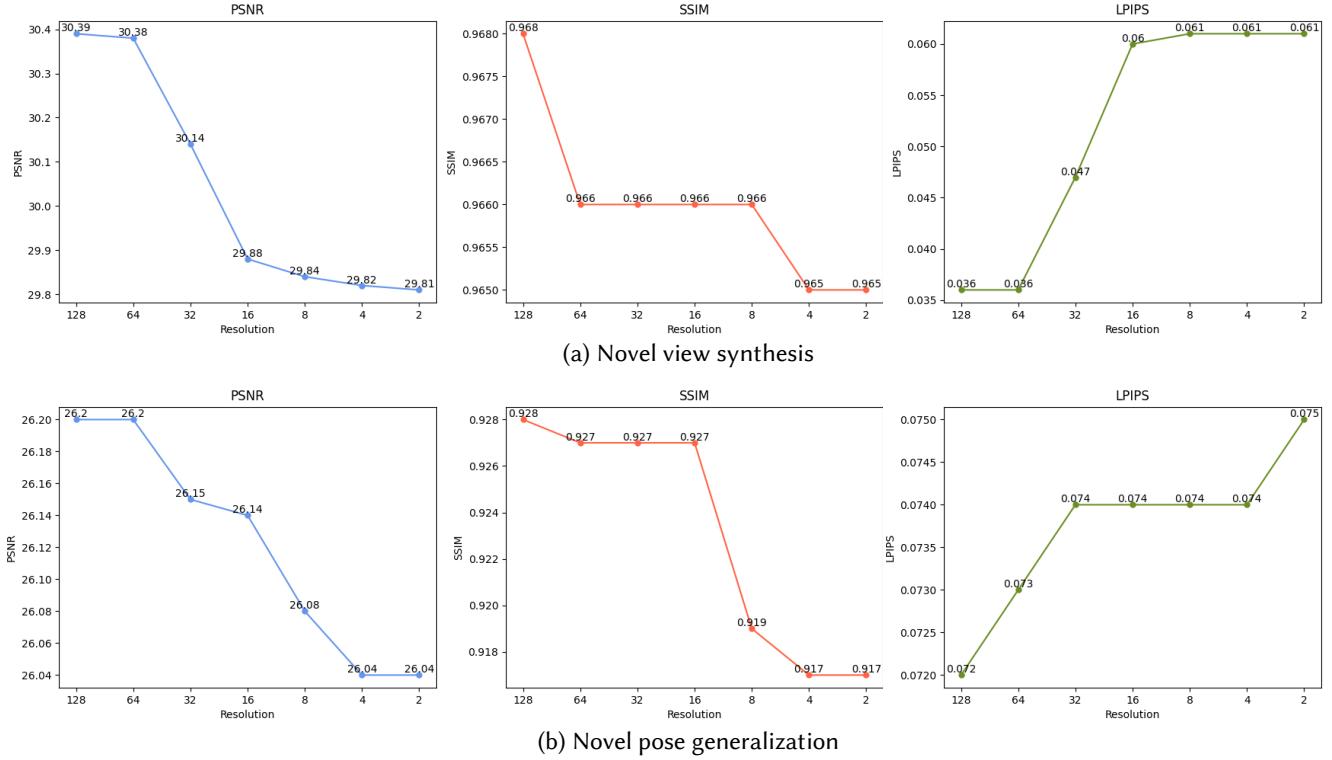


Figure 12. Impacts of NTS at different resolutions.

Since the UV map is generated in a learning-based fashion rather than using direct sampling locations, the color network tends to overfit training viewing directions directly sampled during training. To improve the generalisability of the color network, we apply a sub-pixel noise to the ray

direction. Here, instead of shooting in the pixel centers, a noise ψ is used as follows:

$$x_i = \mathbf{o} + t_i(\mathbf{d} + \psi), \quad (3)$$

where noise ψ can be interpreted as a locality condition,

Type	Novel view synthesis			Novel pose generation		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SH1	29.41	0.966	0.052	26.00	0.926	0.078
SH2	30.02	0.966	0.051	26.15	0.927	0.076
SH3	29.43	0.965	0.051	26.14	0.926	0.075
SH4	27.83	0.962	0.054	26.10	0.926	0.076
w/o noise	28.56	0.962	0.058	25.90	0.924	0.078
Ours	30.38	0.966	0.036	26.20	0.927	0.073

Table 2. Comparison of different methods to model view-dependent RGB color.

i.e., in similar view conditions, RGB color should not be too different. It allows the model to learn smoother transitions between different views.

The ablation of ours w/o noise is presented in Table 2 and demonstrates the effectiveness of the proposed ray direction noise. Figure 10 and Figure 11 show qualitative results of ours w/o noise tests on the novel views. Obviously, ours w/o noise tends to exhibit artifacts in the rendering and textures. **Spherical Harmonic Functions.** Another way to reconstruct the view-dependent color of human performance is using the spherical harmonic (SH) functions. We pass the encoded UV map with the sampled texture embedding into the color network M_c to decode spherical harmonic coefficients η for each color channel:

$$\begin{aligned} (\hat{\eta}_k^0(\mathbf{r}), \hat{\eta}_k^1(\mathbf{r}), \dots, \hat{\eta}_k^n(\mathbf{r})) = \\ M_c\left(\gamma(\hat{\mathcal{U}}_k(\mathbf{r}), \hat{\mathcal{V}}_k(\mathbf{r})), \mathbf{e}_k(\mathbf{r}), \mathbf{k}\right), \end{aligned} \quad (4)$$

where spherical harmonics $(\eta^0, \eta^1, \dots, \eta^n)$ form an orthogonal basis for functions defined over the sphere, with zero degree harmonics η^0 encoding diffuse color and higher degree harmonics encoding specular effects. The view-dependent color $\hat{\mathcal{C}}_k(\mathbf{r})$ of camera ray \mathbf{r} can be determined by querying the specular spherical functions SH at desired viewing direction \mathbf{d} :

$$\hat{\mathcal{C}}_k(\mathbf{r}) = S\left(\frac{\hat{\eta}_k^0(\mathbf{r})}{2}\sqrt{\frac{1}{\pi}} + \sum_{m=1}^n SH^m(\hat{\eta}_k^m(\mathbf{r}), \mathbf{d})\right), \quad (5)$$

where S is the sigmoid function for normalizing the colors.

A higher degree of harmonics results in a higher capability to model high-frequency color but is more prone to overfit the training viewing direction.

The ablation of different harmonics degrees is presented in Table 2, which demonstrates that the harmonics degree of 2 model achieves the best performance among all the SH models but still cannot reach the performance of ours. Figure 10 illustrates the qualitative results of SH2 tests on the novel views, which shows that SH2 suffers from global color shifts (head of the performer) and artifacts (the belt on pants), while ours does not.

Task	4 Views			20 Views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
60 Frames	29.56	0.967	0.045	31.12	0.975	0.038
300 Frames	29.23	0.963	0.048	29.90	0.968	0.046
600 Frames	29.37	0.964	0.049	29.50	0.966	0.048
1200 Frames	28.96	0.961	0.052	29.26	0.963	0.053

Table 3. Ablation study on the number of training frames and views.

D.4. Video Frames and Input Views

To analyze the impacts of the number of camera views and video length, we show the results of our models trained with different numbers of camera views and video frames in Table 3. We conduct the experiments on performer 313 of ZJU dataset. All the results are evaluated on the rest two views of the first 60-frame video. The results show that although the number of training views improves the performance on novel view synthesis, sparse four views are good enough for our model to reconstruct dynamic human performances. In addition, the ablation study of frame numbers indicates that training on too many frames may decrease the performance as the network cannot fit such a long video, which is also mentioned in NeuralBody [6].

E. Additional Results

E.1. Novel View Synthesis

For comparison, we synthesize images of training poses in hold-out test-set views. More qualitative results of novel view synthesis are shown in Figure 13, Figure 14 and Figure 15. Our method produces photo-realistic images with sharp details, particularly letters on clothes (in Figure 13), stripes on T-shirts, and wrinkles in clothes (in Figure 14), which benefits from our proposed Spatial NTS that encodes high-frequency appearance information.

Figure 15 shows the results of comparisons on ZJU Mocap and H36M dataset, which are trained on sparse-views video sequences. Here, we use four training views on ZJU Mocap dataset and three for the most challenging H36M dataset. Our model obviously performs much better in details and sharpness than all other baselines. Furthermore, DyNeRF fails to render plausible results with sparse training views because taking time-varying latent codes as the conditions are hard to reuse information among frames.

E.2. Novel View Synthesis of Dynamic Humans

We present more results on novel view synthesis of dynamic humans in Figure 16. As presented, our model can handle a dynamic human with rich textures and challenging motions, and preserve sharp image details like letters and wrinkles, while keeping inter-view and inter-frame consistency. Note that the last row is the result of our model on

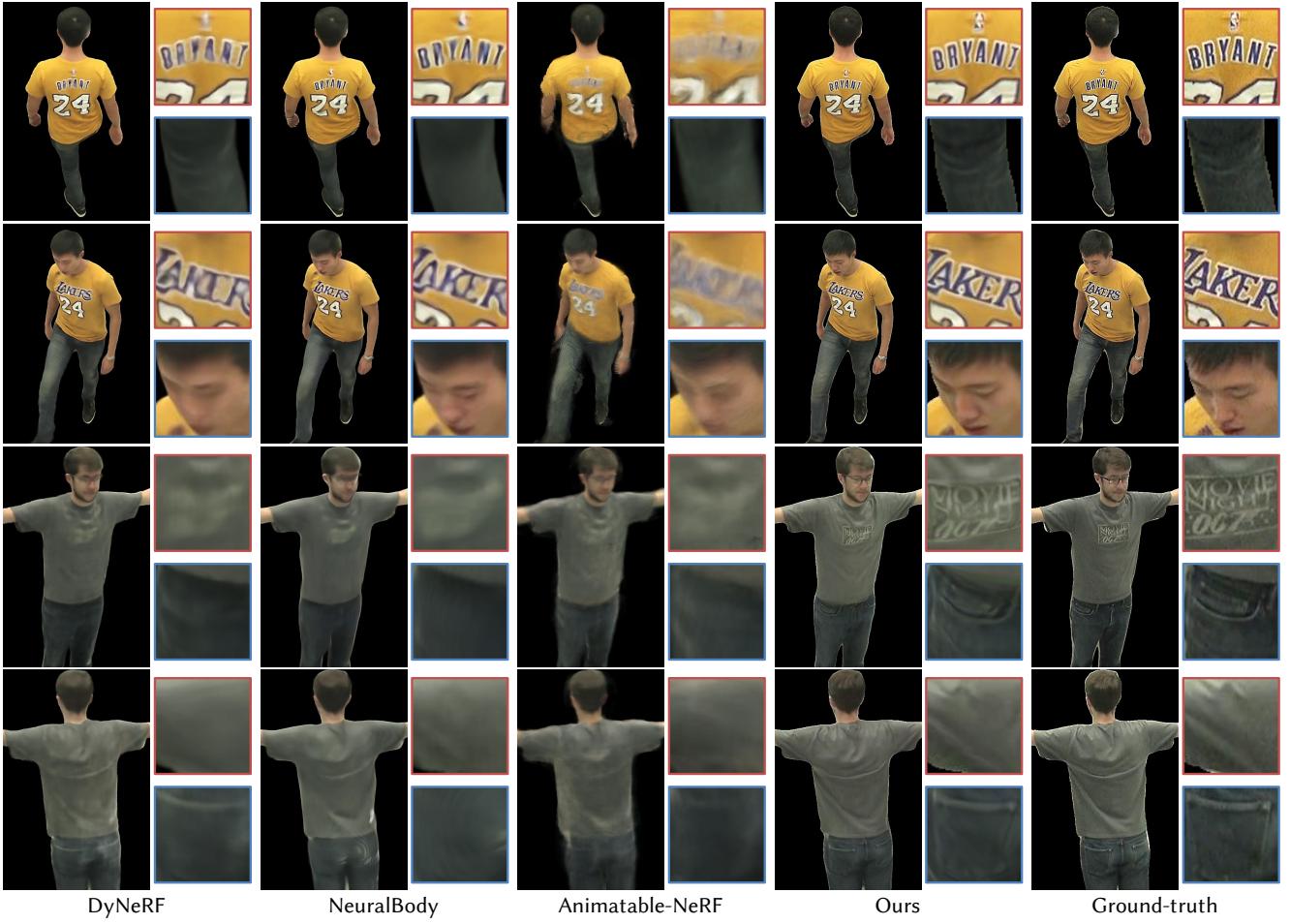


Figure 13. Comparisons on test-set views for performers from CMU Panoptic dataset with 960×540 images. Our model generates photo-realistic appearance images even with rich textures, particularly letters on the performers’ clothes. By contrast, baselines give blurry results while missing a lot of high-frequency details. Here, we present results on two different test views at the same time for each performer.

the H36M dataset, demonstrating that our model can still recover high-fidelity free-view videos under sparse training views.

In addition, we show the intermediate UV images and final RGB images rendered by our model varying with views and human poses in Figure 17, which demonstrates that our model can synthesize photo-realistic view-consistent RGB images that condition on view-consistent UV images rendered by UV volumes.

E.3. Novel Pose Generalization

More qualitative results of novel pose generalization are shown in Figure 18 and Figure 19, where the latter are the results of comparisons on H36M dataset where only three cameras are available for training.

E.4. Reshaping

By changing the SMPL parameters, we can conveniently deform the human performer. We present the performer

whose size is getting smaller and the shoulder-to-waist ratio is getting smaller from left to right in Figure 20. With the help of view-consistent UV coordinates generated by UV volumes, our model still renders view-consistent images with challenging shape parameters. These rendered images maintain highly appearance consistency across changing shapes thanks to the neural texture stacks.

E.5. Visualization of NTS

In contrast to [6] learning a radiance field in the 3D volumes, we decompose a 3D dynamic human into 3D UV volumes and 2D neural texture stacks, as illustrated in Figure 21. The disentanglement of appearance from geometry enables us to achieve real-time rendering of free-view human performance. We learn a view-consistent UV field to transfer neural texture embeddings to colors, which guarantees view-consistent human performance. Details like the folds of clothing vary from motion to motion, as does the topology, so we require a dynamic texture representation. Refer-

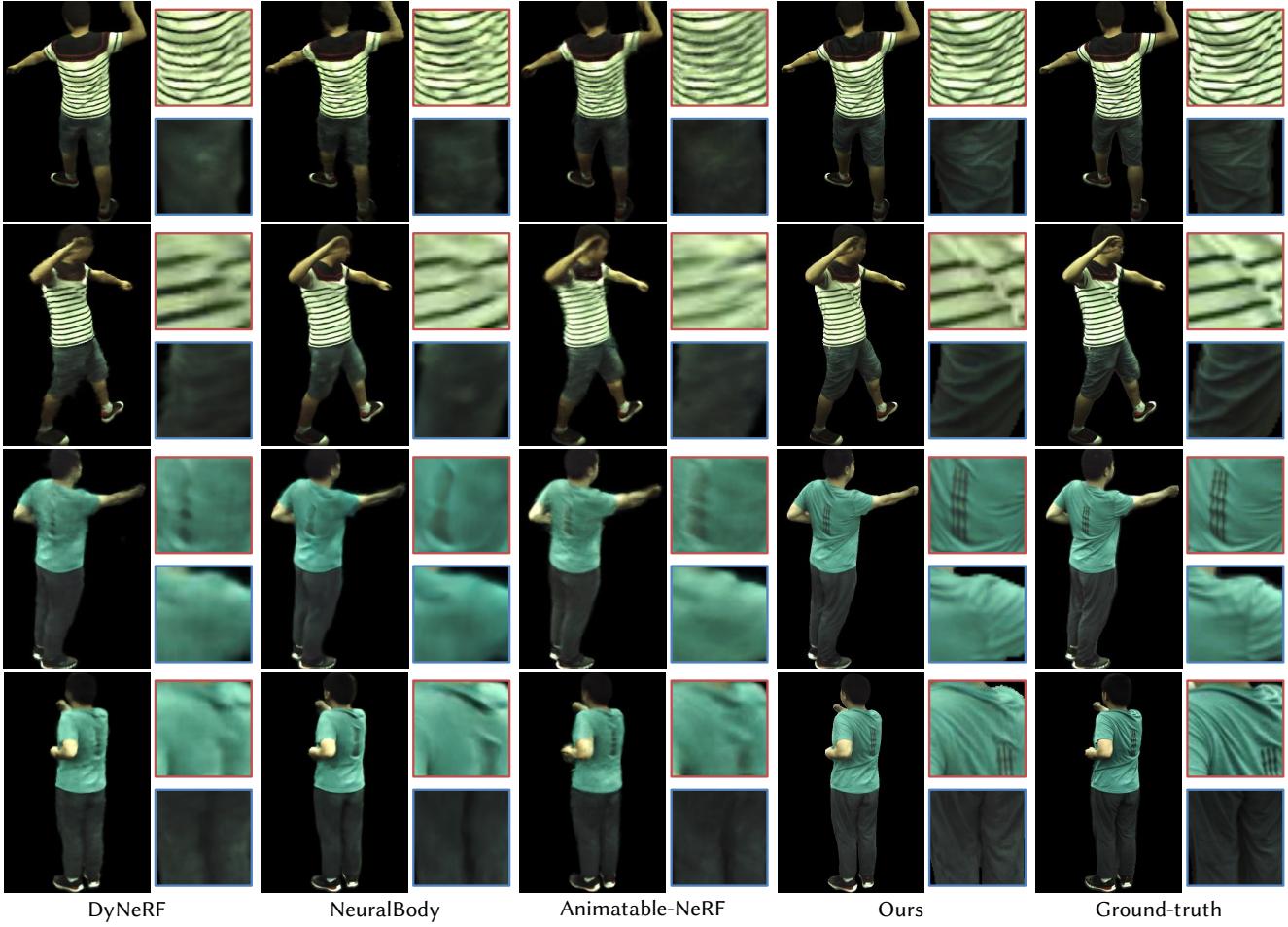


Figure 14. Comparisons on test-set views for performers from ZJU Mocap dataset. Our model obviously perform well in details (e.g., stripes on T-shirts and wrinkles in clothes) and sharpness than all other baselines, which benefits from our proposed Spatial NTS that encodes high-frequency appearance information. Other methods give plausible but blurry and rough synthesized images. Here, we present results on two different test views at the same time for each performer.

ring to Figure 22, we visualize the pose-driven neural texture stacks to describe appearance at different times, which enables us to handle dynamic 3D reconstruction tasks and to generalize our model to unseen poses. It is obvious that our learned NTS preserve rich textures and high-frequency details varying from different poses.

E.6. Retexturing

With the learned dense correspondence of 3D UV volumes and 2D neural texture stacks, we can edit performers’ 3D cloth by user-provided 2D textures. As shown in Figure 23, given arbitrary artistic paintings, we can produce cool stylized dynamic humans leveraging stylizations transferred from the original texture stacks by the network [1]. Visually inspected, the new texture is well painted onto the performer’s T-shirt under different poses at different viewing directions. Besides, we perform some interesting ap-

plications of our model in Figure 24 and Figure 25, which include a 3D virtual try-on implemented by replacing original texture stacks with a user-provided appearance. The visualization results demonstrate that our model can conveniently edit textures preserving the rich appearance and various styles, which benefits from our proposed Neural Texture Stacks, and can render retextured human performance with view consistency well using 3D UV volumes.

References

- [1] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017. 8
- [2] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In

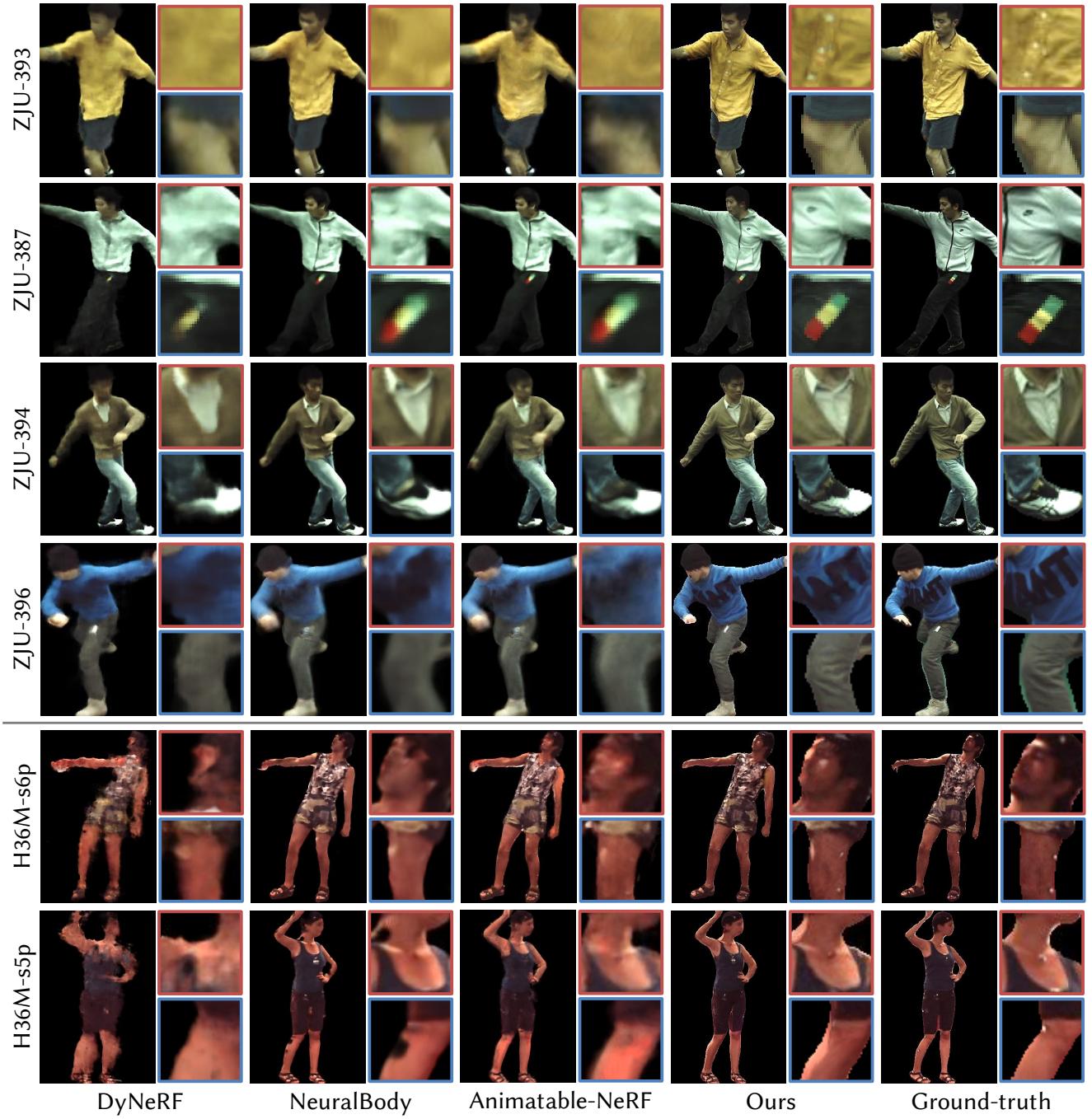


Figure 15. Comparisons on test-set views from the ZJU Mocap dataset with four training views and the most challenging H36M dataset with only three available views for training. Our model generates high-definition results even with rich textures and challenging motions. DyNeRF fails to render plausible results with sparse training views because taking time-varying latent codes as the conditions are hard to reuse information among frames.

- Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2
- [3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

- [4] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*



Figure 16. The rendering of our method on different sequences. Our model can handle a dynamic human with rich textures and challenging motions preserving sharp image details like letters and wrinkles, which benefits from our proposed Spatial NTS that encodes high-frequency appearance information, while keeping inter-view and inter-frame consistency, which benefits from our proposed *UV Volumes*. Note that the last row is the result of our model on the H36M dataset, which demonstrates that our model can still recover high-fidelity free-view videos under sparse training views.

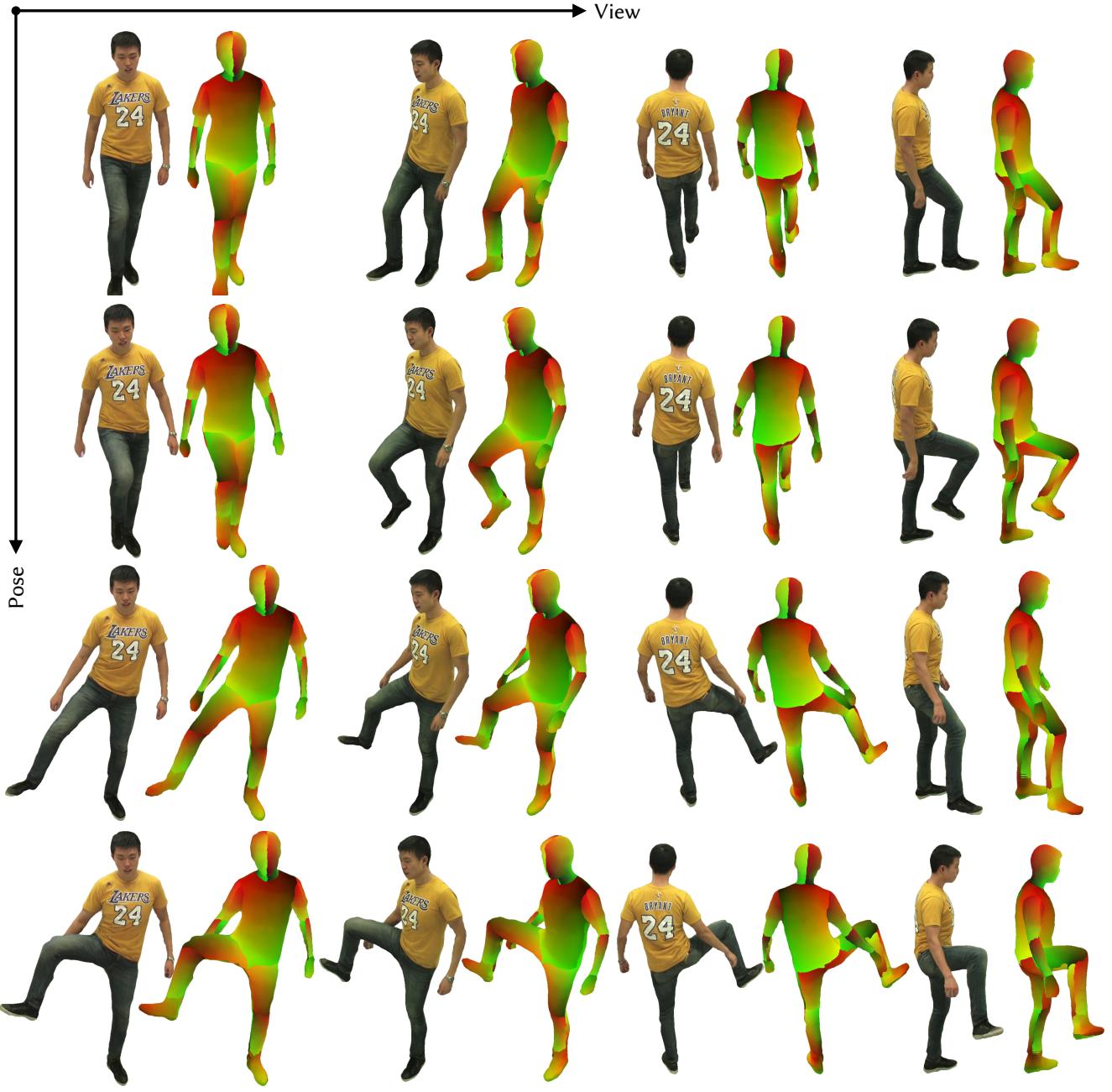


Figure 17. Novel view synthesis of the dynamic human. Our model can synthesize photo-realistic view-consistent RGB images condition on view-consistent UV images rendered by UV volumes. Here, The horizontal axis shows the change in novel views and, the vertical axis shows the change in human poses. All results are rendered from novel views in the training pose sequence.

Recognition, pages 5521–5531, 2022. 2

- [5] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 3

- [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. 1, 2, 6, 7

- [7] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 4

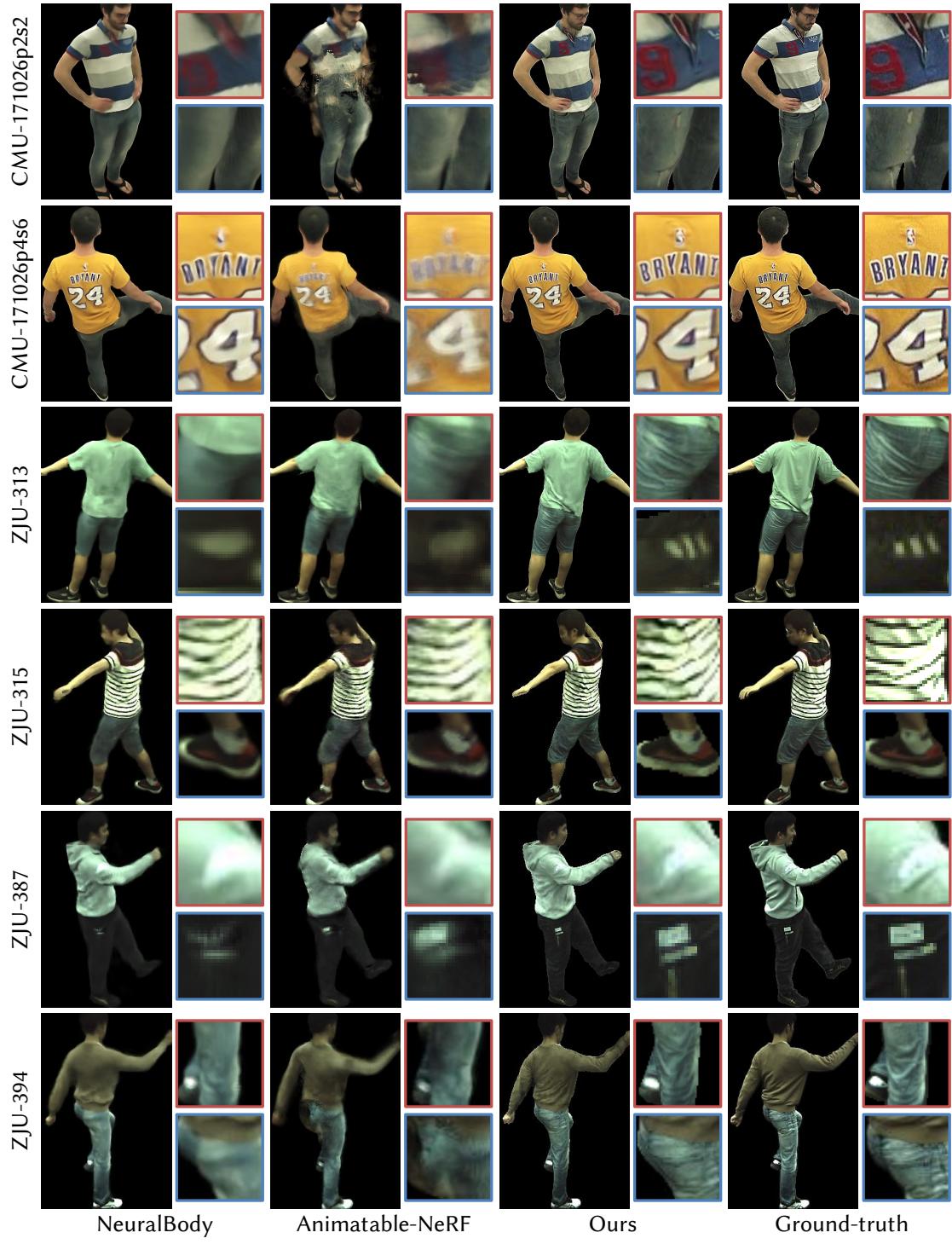


Figure 18. Comparisons on test-set poses for performers from CMU Panoptic and ZJU Mocap dataset.

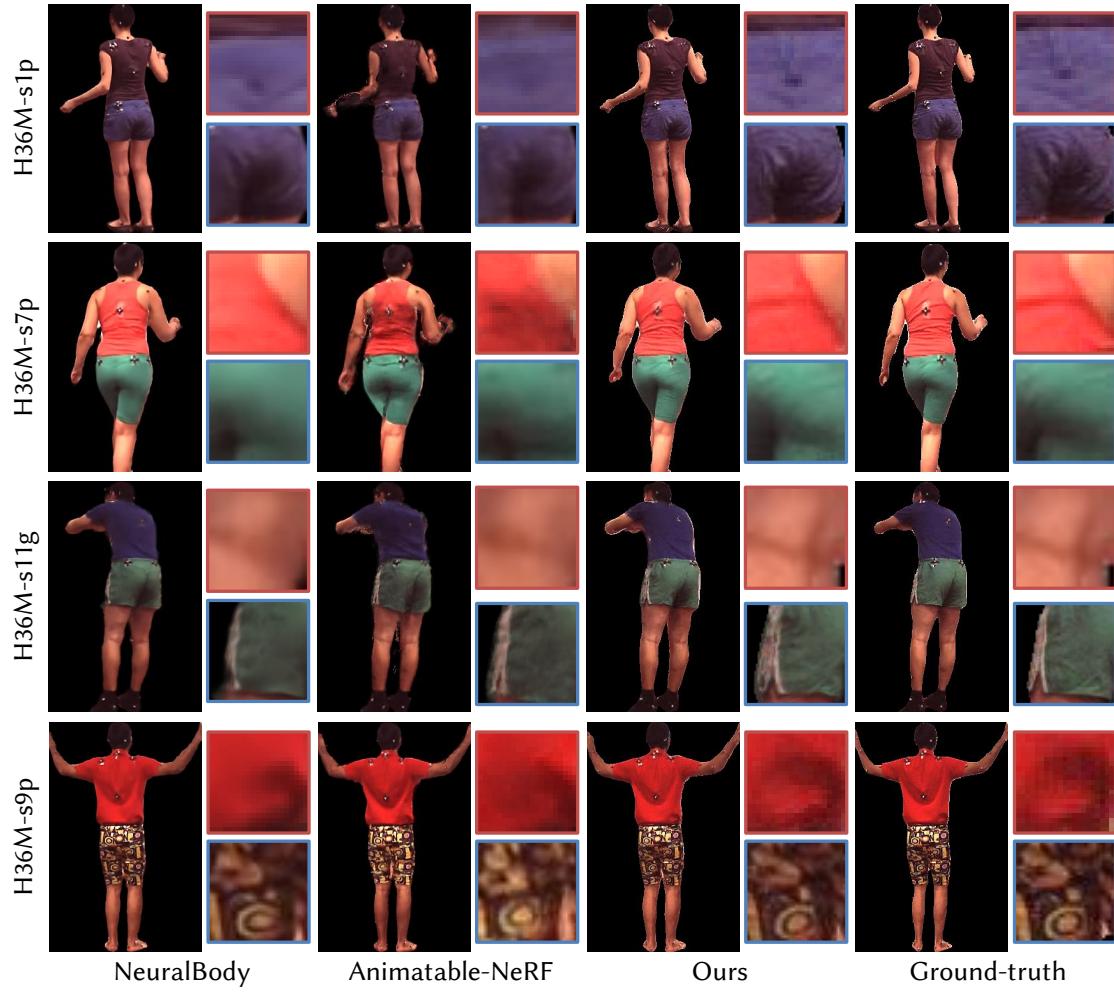


Figure 19. Comparisons on test-set poses for performers from the most challenging H36M dataset.

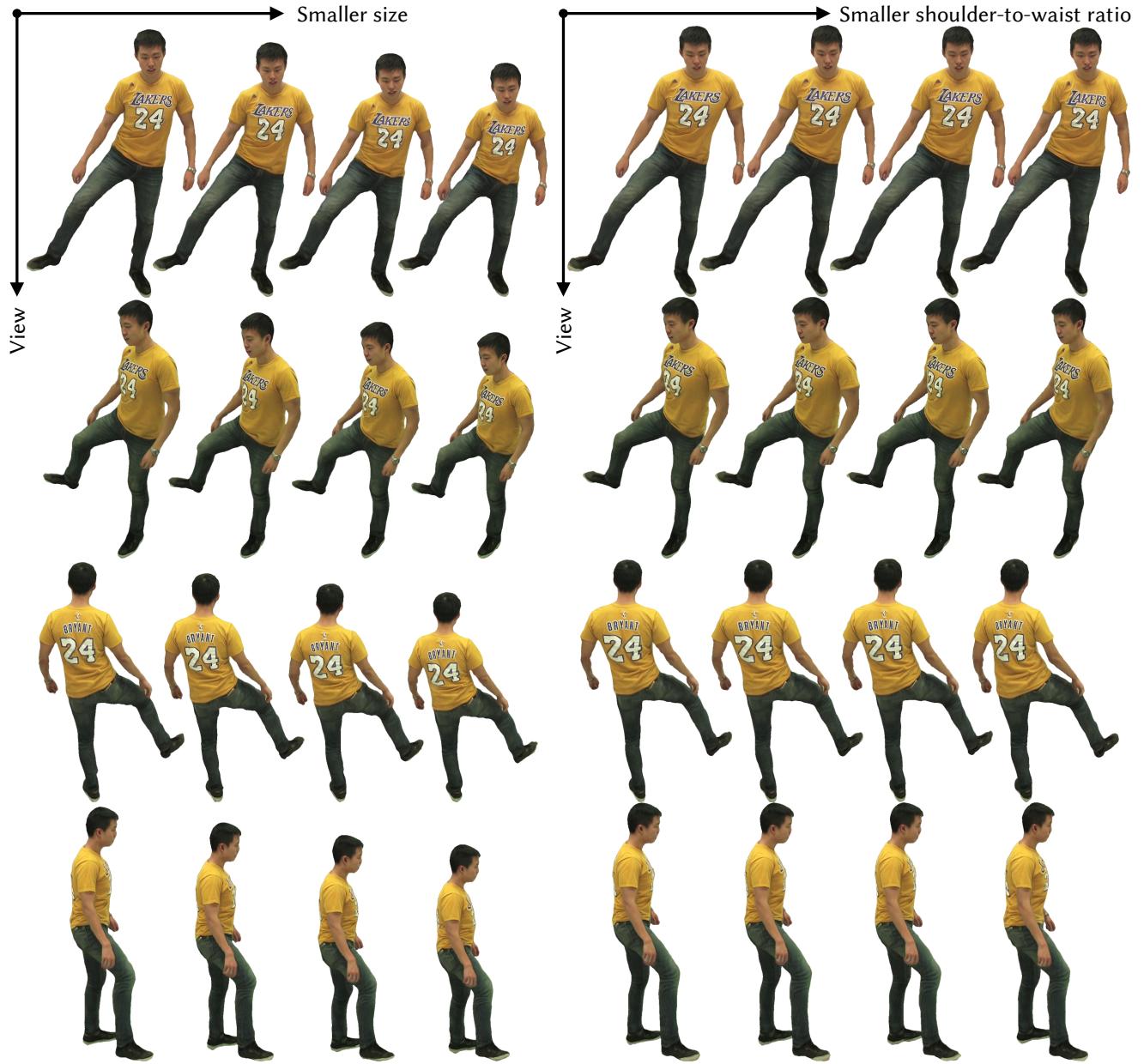


Figure 20. Novel view synthesis results of reshaping. By changing the SMPL parameters, we can conveniently deform the human performer. We present the performer whose size is getting smaller and the shoulder-to-waist ratio is getting smaller from left to right. With the help of view-consistent UV coordinates encoded by UV volumes, our model still renders view-consistent images with challenging shape parameters. Here, The horizontal axis shows shape changes and the vertical axis shows view changes. All results are rendered from novel views.

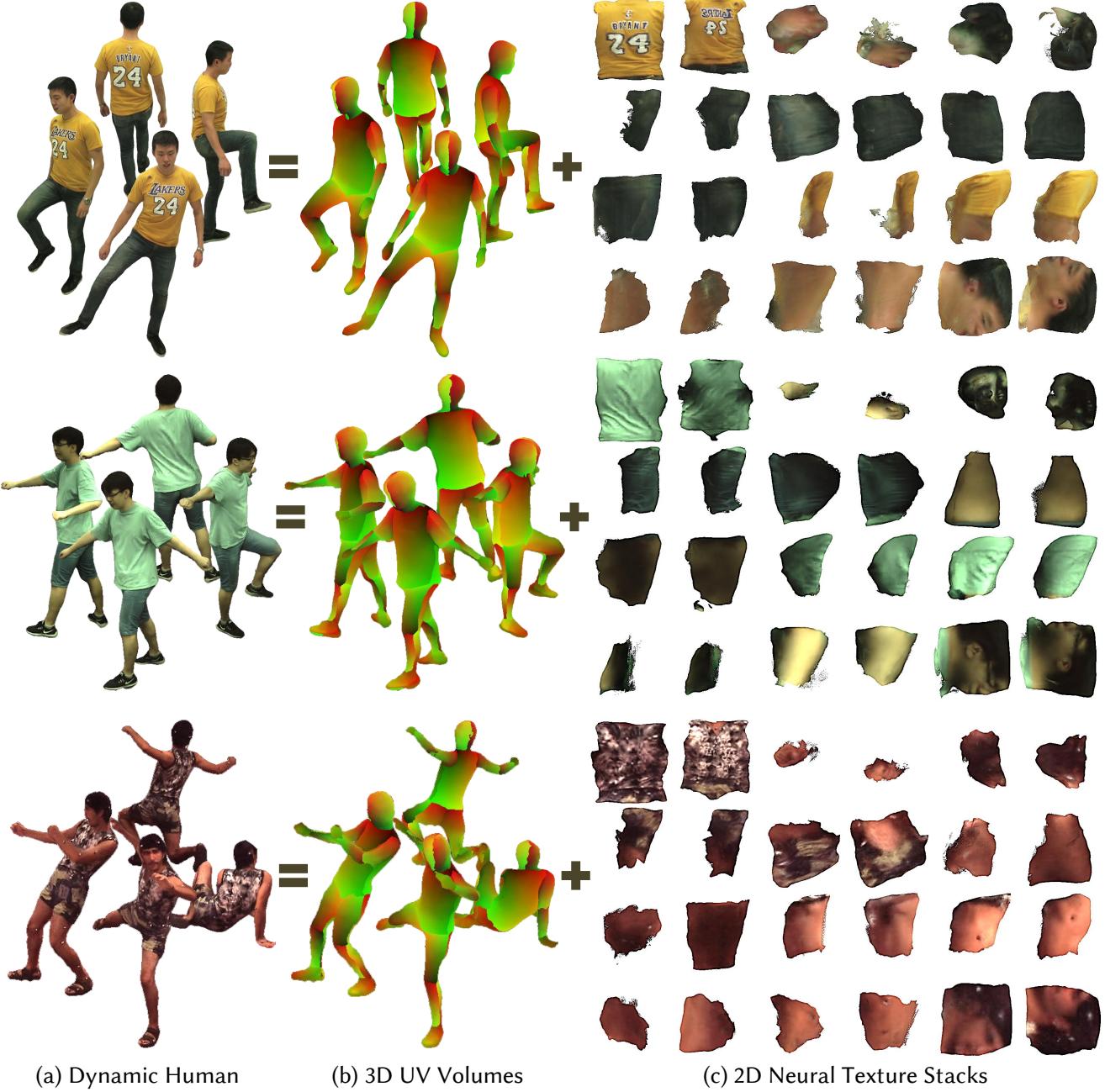


Figure 21. We decompose (a) the dynamic human into (b) 3D UV volumes and (c) 2D neural texture stacks. The disentanglement of appearance from geometry enables us to achieve real-time rendering of free-view human performance. We show performers and their UV avatars with four different poses at four different viewing directions from CMU Panoptic, ZJU-Mocap and H36M datasets. Their neural texture stacks that preserve human appearance with high-frequency details under one of these poses are visualized in the last column. Our method takes smooth UV coordinates to sample neural texture stacks for corresponding RGB value.

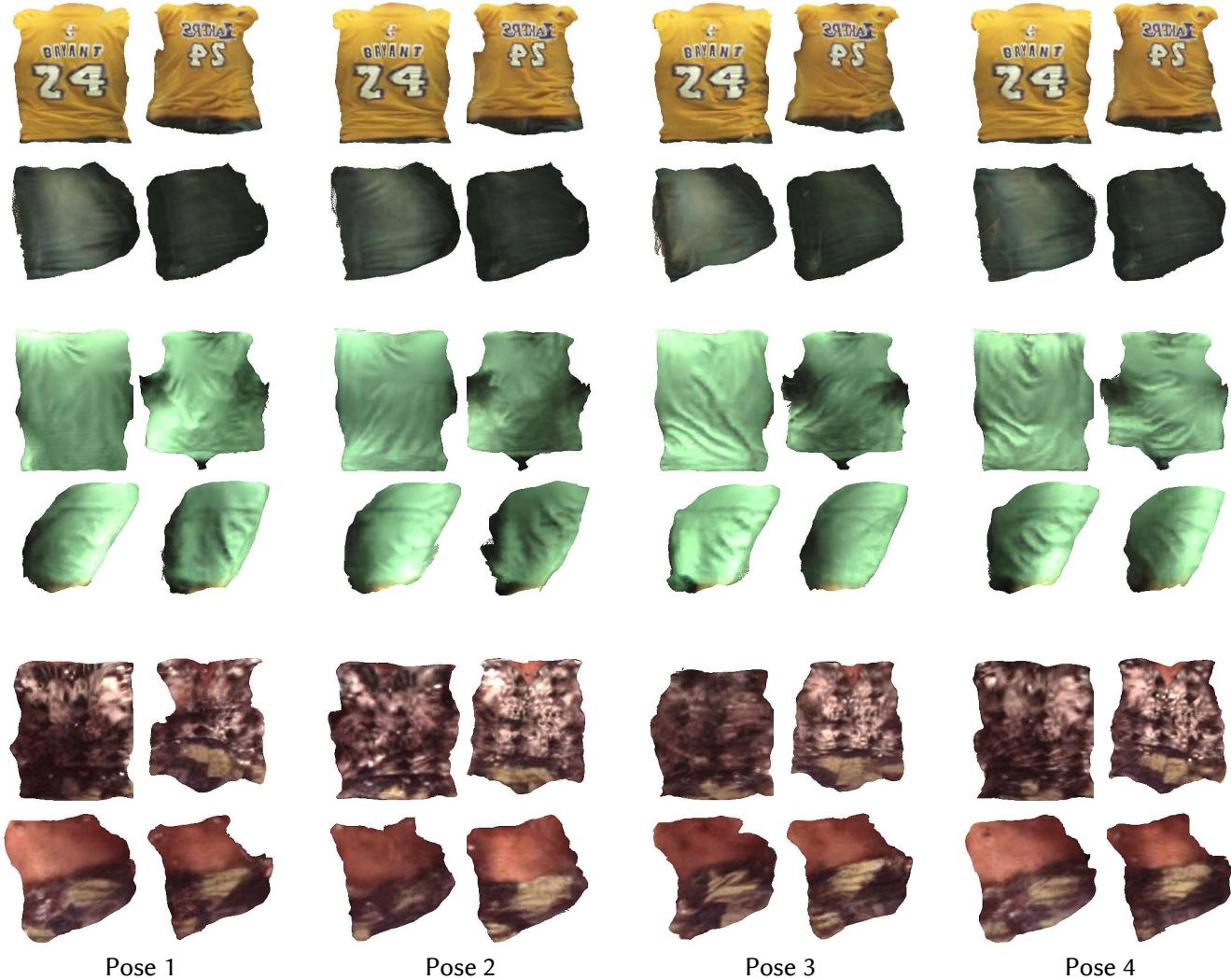


Figure 22. Visualization of neural texture stacks under different poses. Details like the folds of clothing vary from motion to motion, as does the topology. Therefore, we propose pose-driven neural texture stacks to describe textures at different times, which enables us to handle dynamic 3D reconstruction tasks and to generalize our model to unseen poses.



Figure 23. Our model supports rendering a stylized dynamic human with arbitrary artistic painting, which can be applied in controllable 3D style transfer with multi-view consistency.

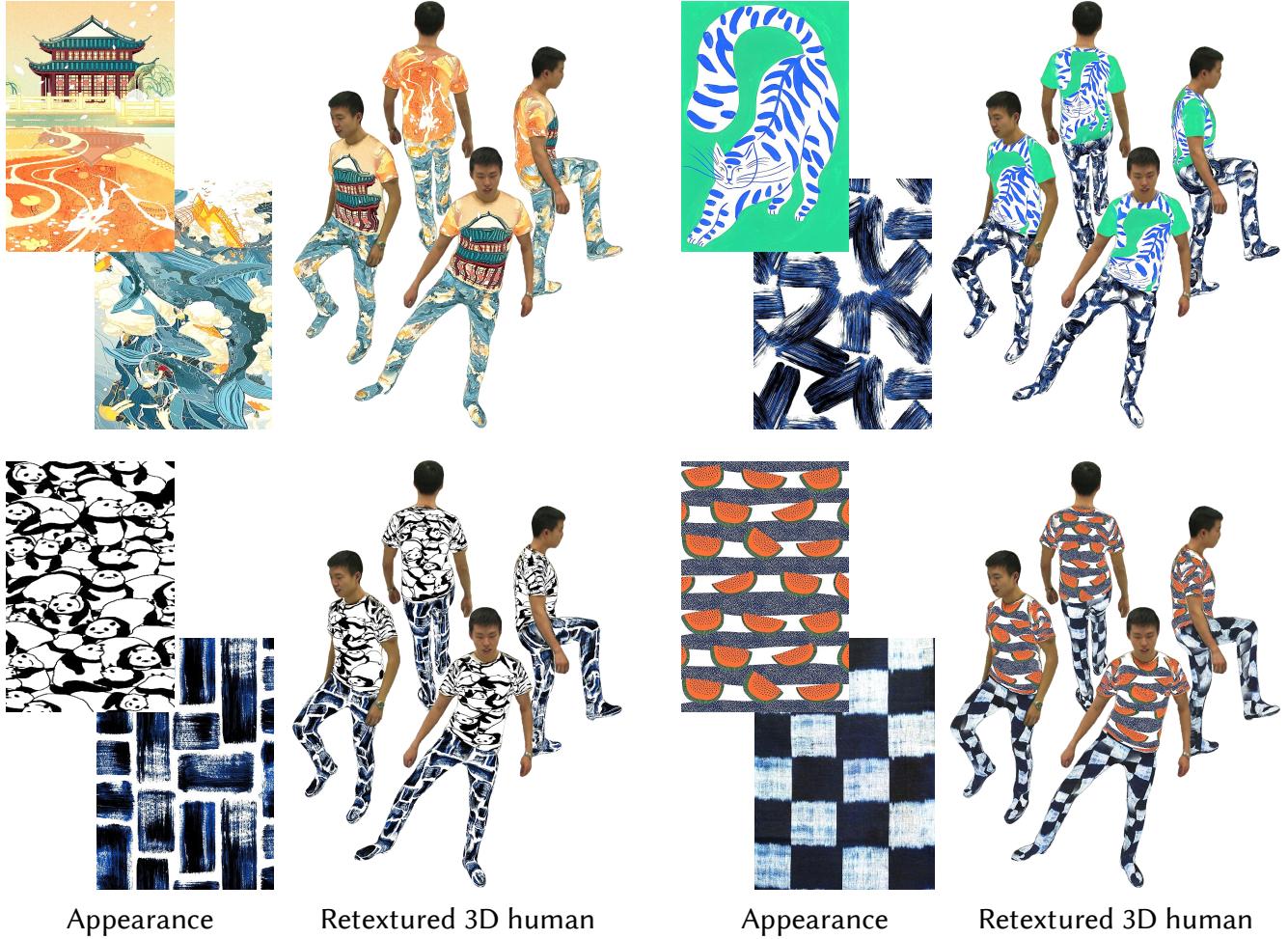


Figure 24. Our model allows us to generate free-view human performance with a user-provided cloth texture image, which enables some interesting applications such as real-time 3D virtual try-on. We collect these appearance images from the Internet.



Figure 25. Our model allows us to generate free-view human performance with a user-provided cloth texture image, which enables some interesting applications such as real-time 3D virtual try-on. We collect these appearance images from the Internet.