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

ACM Reference Format:
. 2022. Supplementary Material for "UV Volumes for Real-time Rendering of Editable Free-view Human Performance". In . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/8888888.7777777

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

1 ADDITIONAL RESULTS

1.1 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 1, Figure 2 and Figure 3. Our method produces photo-realistic images with sharp details, particularly letters on clothes (in Figure 1), stripes on T-shirts and wrinkles in clothes (in Figure 2), which benefits from our proposed spatial neural texture stacks (NTS) that encode high-frequency appearance information.

Figure 3 are the results of comparisons on ZJU Mocap and H36M dataset which are training 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 perform well 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 is hard to reuse information among frames.

1.2 Novel View Synthesis of Dynamic Humans

We present more results on novel view synthesis of dynamic humans in Figure 4. As presented, our model can handle dynamic human with rich textures and challenging motions, and preserve sharp image details like letters and wrinkles, while keeps inter-view consistency and inter-frame consistency. 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.

In addition, we also show the intermediate UV images and final RGB images rendered by our model varying with views and human poses in Figure 5, which demonstrates that our model can synthesise photo-realistic view-consistent RGB images that condition on view-consistent UV images rendered by UV volumes.

SIGGRAPH Conference Proceedings, Dec 2022, Daegu

ACM ISBN 978-1-4503-1234-5/22/07…$15.00
https://doi.org/10.1145/8888888.7777777

1.3 Novel Pose Generalization

More qualitative results of novel pose generalization are shown in Figure 6 and Figure 7, where the latter are the results of comparisons on H36M dataset where only three cameras are available for training.

1.4 Reshaping

By changing the SMPL parameters, we can conveniently deform the human performer. We present the performer whose size is getting smaller and shoulder-to-waist ratio is getting smaller from left to right in Figure 8. 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.

1.5 Visualization of NTS

In contrast to [Peng et al. 2021] learning a radiance filed in the 3D volumes, we decompose a 3D dynamic human into 3D UV volumes and 2D neural texture stacks, as illustrated in Figure 9. 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. Referring to Figure 10, 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 which varying from different poses.

1.6 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 11, given any arbitrary artistic paintings, we can produce cool stylized dynamic humans leveraging stylizations transferred from the original texture stacks by the network [Ghiasi et al. 2017]. Visually inspected, the new texture are well painted onto the performer's T-shirt under different poses at different viewing directions. Besides, we perform some interesting applications of our model in Figure 12 and Figure 13, which includes a 3D virtual try-on implemented by replace original texture stacks with user-provided appearance. The visualization results demonstrates that our model can conveniently edit textures preserving rich appearance and various style, which benefits from our proposed Neural Texture Stacks, and can render retextured human performance with view consistency well using 3D UV volumes.

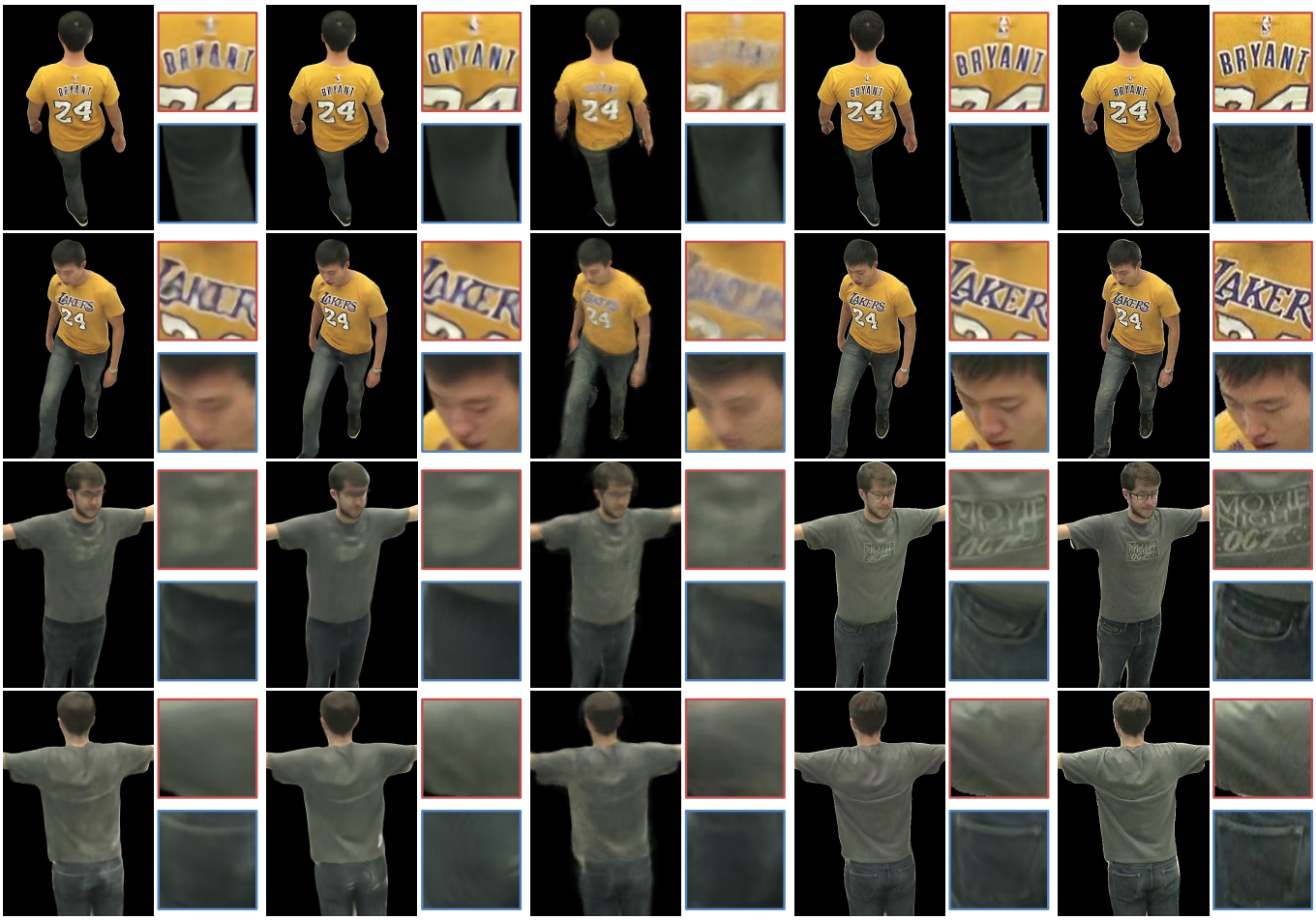

Fig. 1. Comparisons on test-set views for performers from CMU Panoptic dataset with 960×540 images. Our model generates photo-realistic appearance images even with rich textures, particularly letters on the performers' clothes. By contrast, baselines gives blurry results while misses a lot of high-frequency details. Here, we present results on two different test views at the same time for each performer.

REFERENCES

Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830* (2017).

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. 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*. 9054–9063.

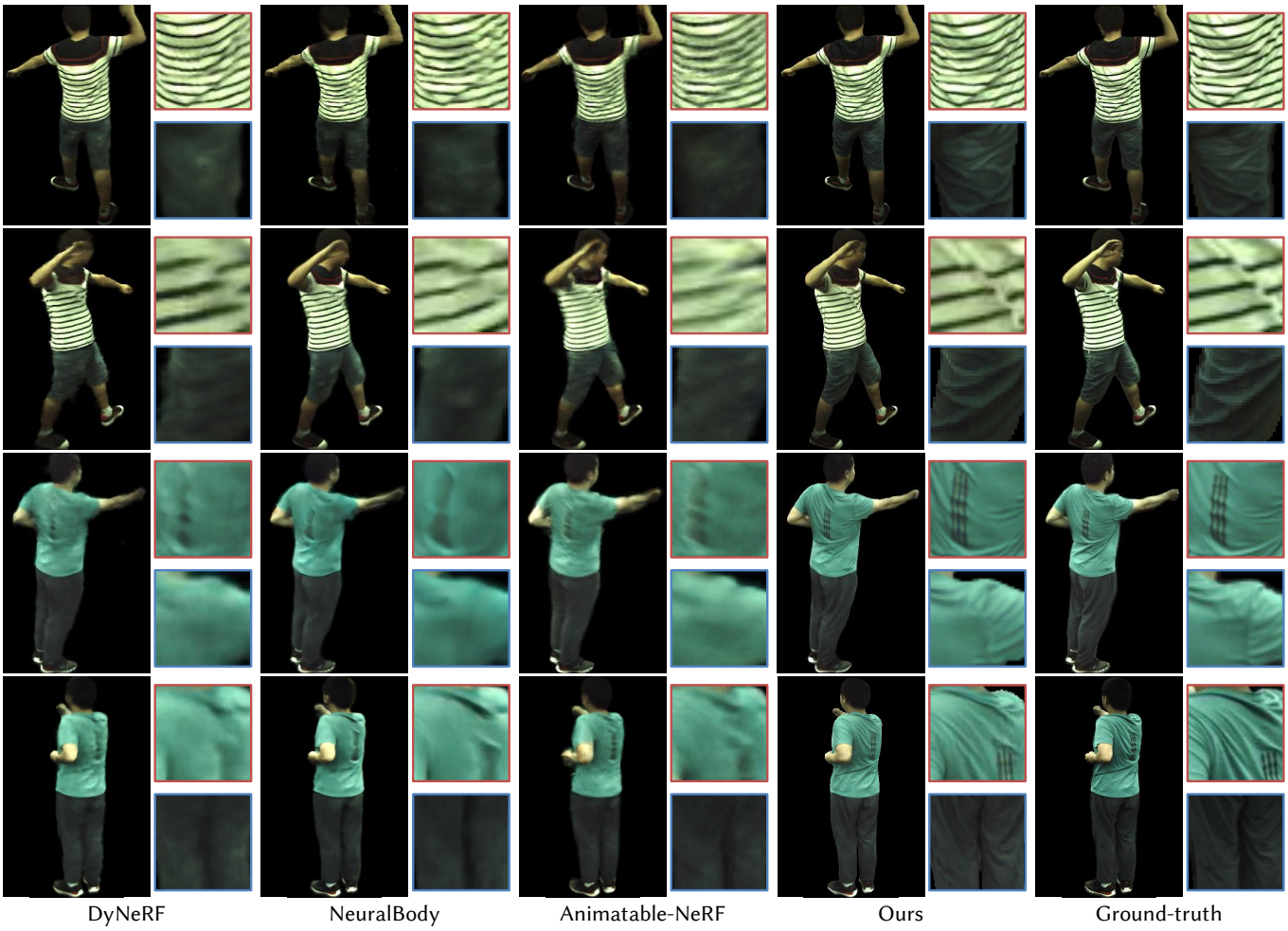

Fig. 2. Comparisons on test-set views for performers from ZJU Mocap dataset. Our model obviously perform well in details (e.g., stripes on T-shirts and wrinkles in clothes) and sharpness than all other baselines, which benefits from our proposed spatial *neural texture stacks (NTS)* that encode high-frequency appearance information. Other methods gives plausible but blurry and rough synthesized images. Here, we present results on two different test views at the same time for each performer.

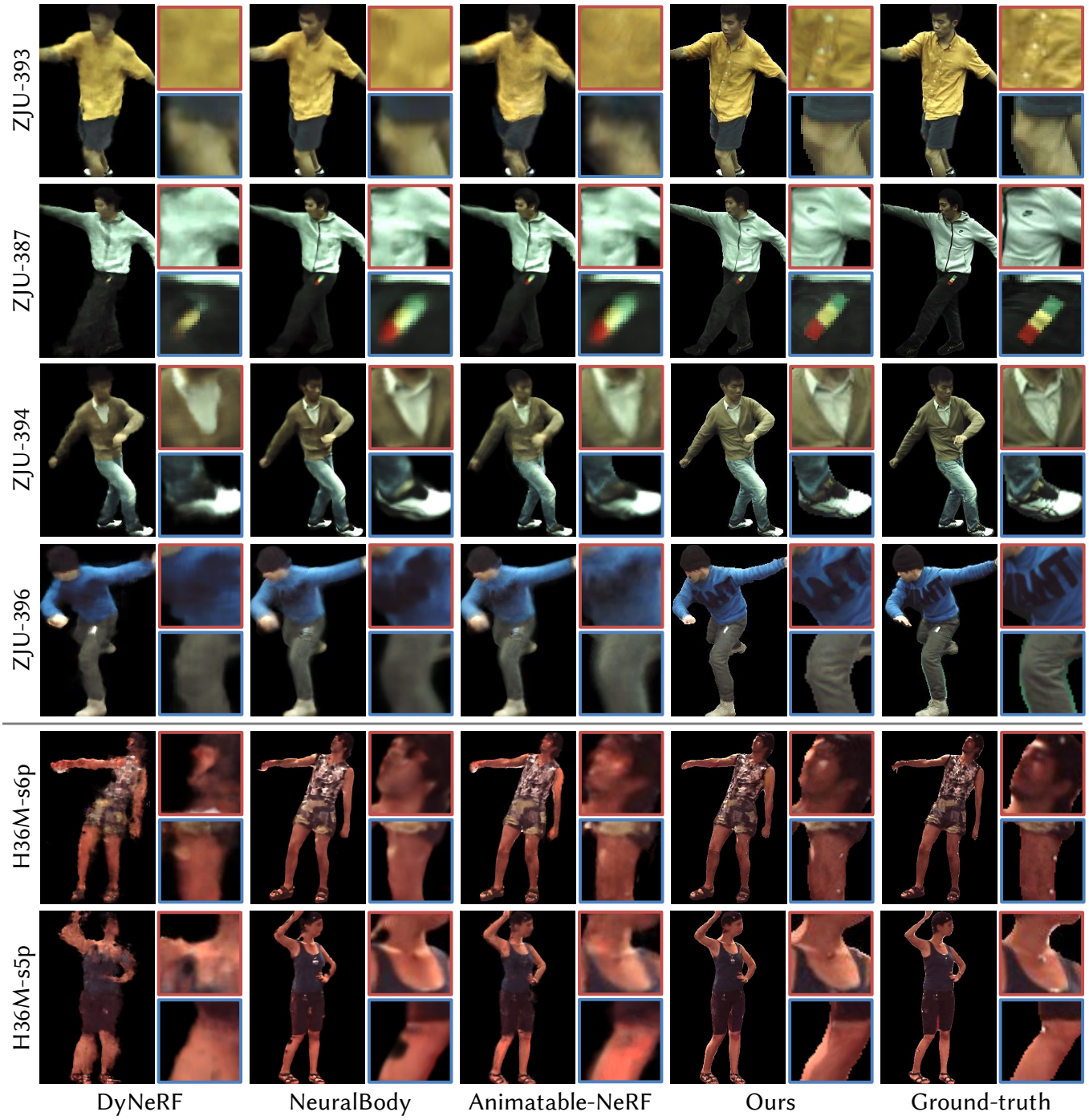

Fig. 3. Comparisons on test-set views from ZJU Mocap dataset that has four training views and the most challenging H36M dataset that only has three available views for training. Our model generates high-definition results even with rich textures and challenging motions. DyNeRF fails to render plausible results with sparse training views because taking time-varying latent codes as the conditions is hard to reuse information among frames.

Fig. 4. The rendering of our method on different sequences. Our model can handle dynamic human with rich textures and challenging motions preserving sharp image details like letters and wrinkles, which benefits from our proposed spatial *neural texture stacks (NTS)* that encode high-frequency appearance information, while keeping inter-view consistency 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.

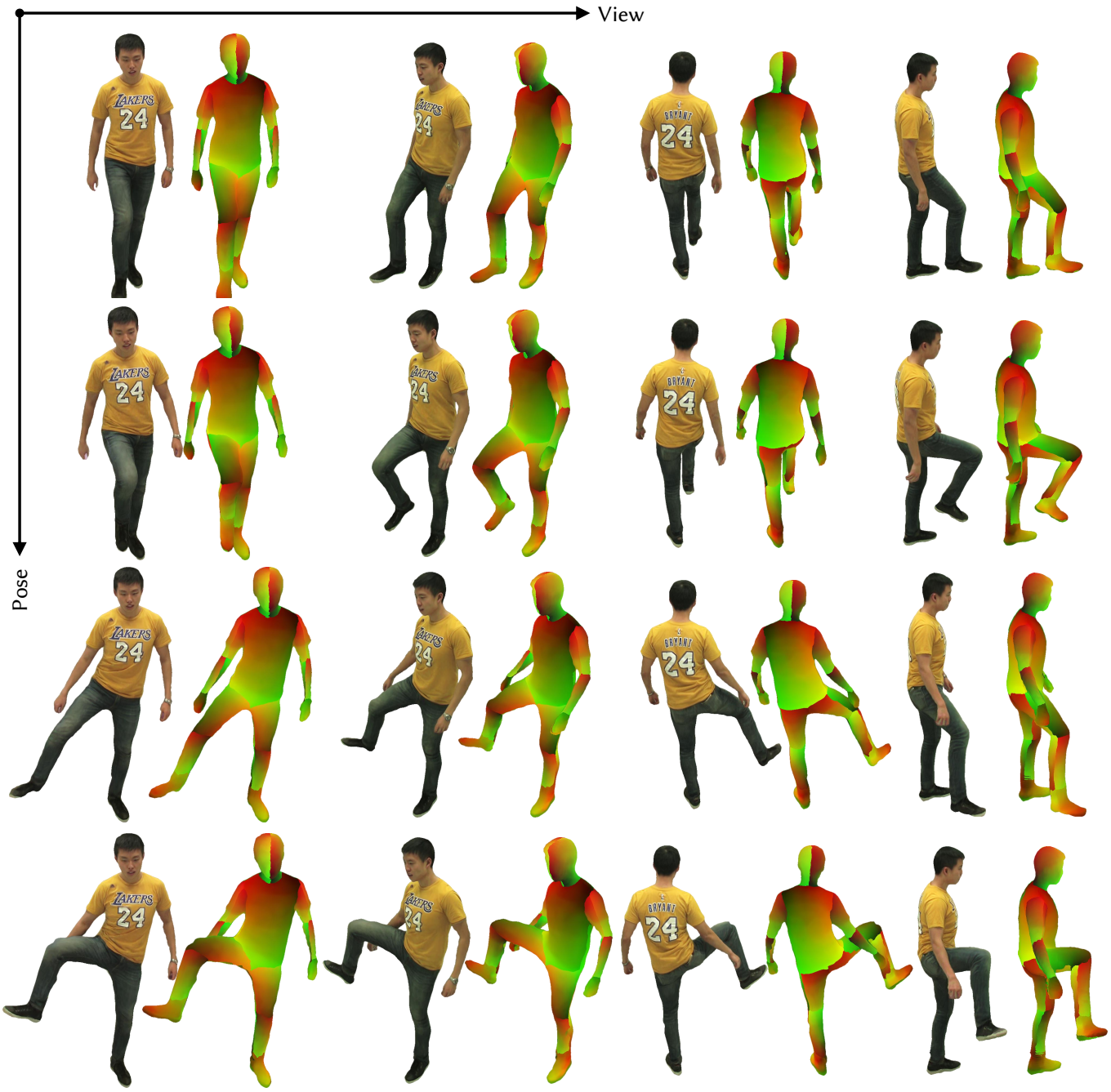

Fig. 5. Novel view synthesis of the dynamic human. Our model can synthesise photo-realistic view-consistent RGB images that condition on view-consistent UV images rendered by UV volumes. Here, The horizontal axis shows the change in novel views and the vertical axis shows the change in human poses. All results are rendered from novel views in the training pose sequence.

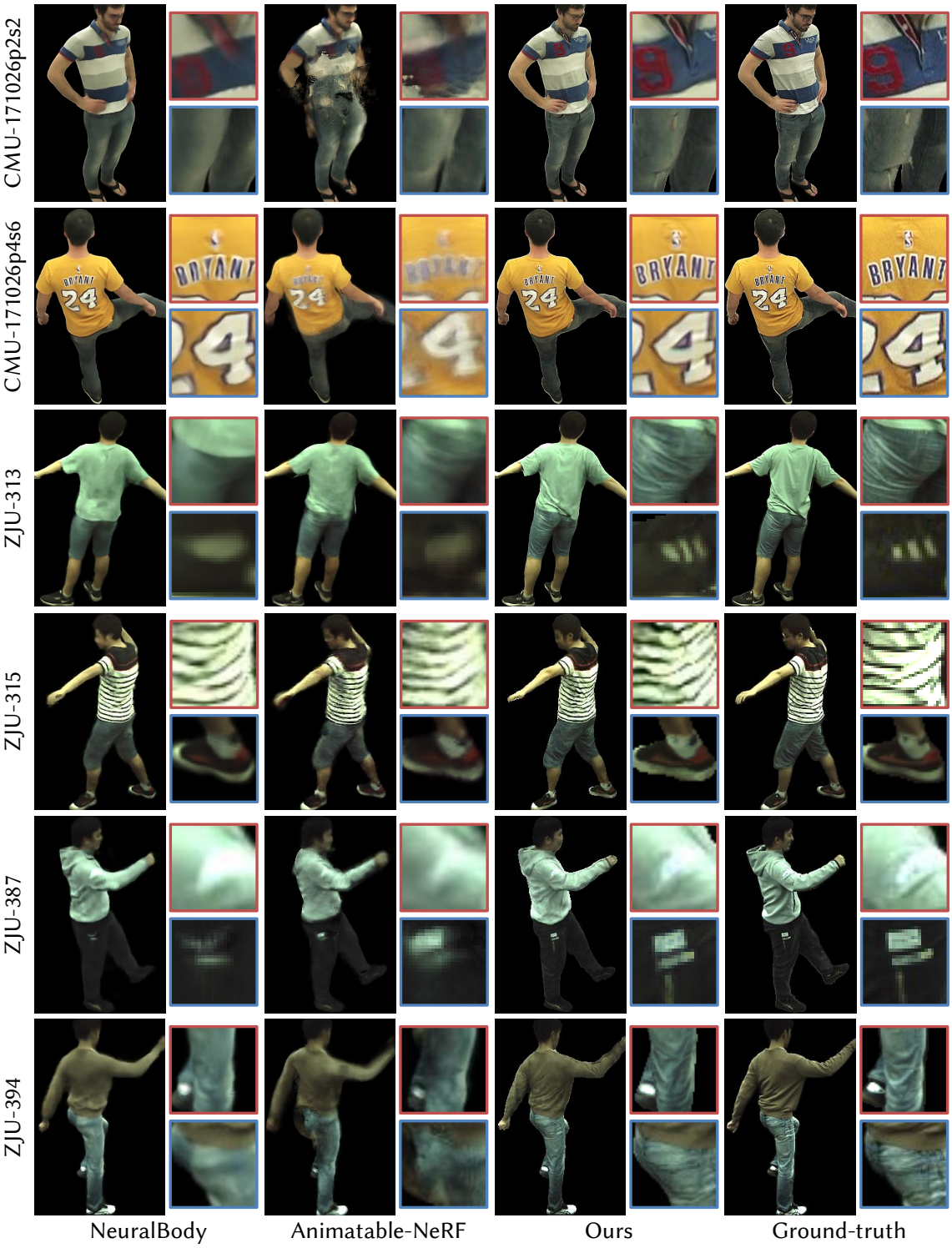

Fig. 6. Comparisons on test-set poses for performers from CMU Panoptic and ZJU Mocap dataset.

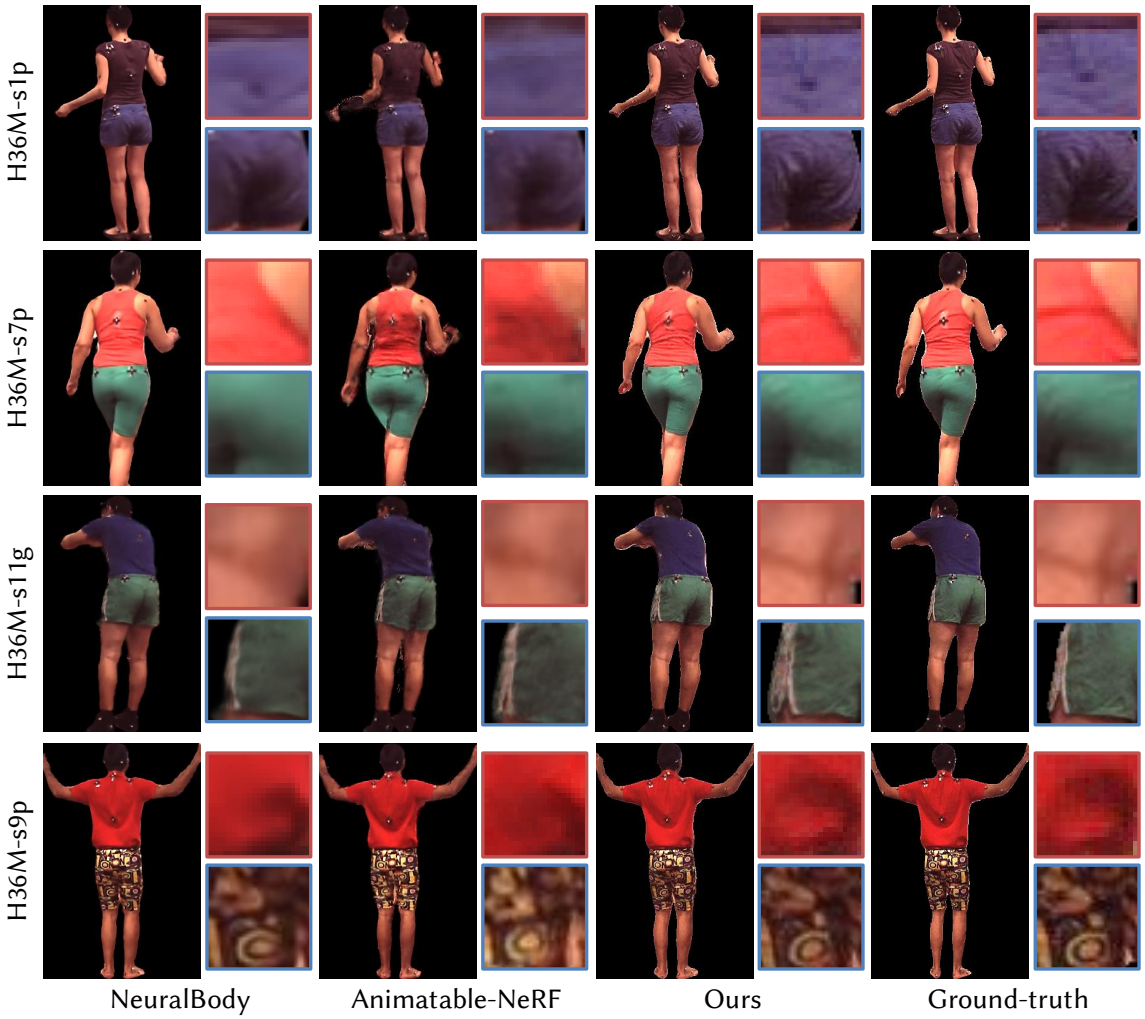

Fig. 7. Comparisons on test-set poses for performers from the most challenging H36M dataset where only three cameras are available for training.

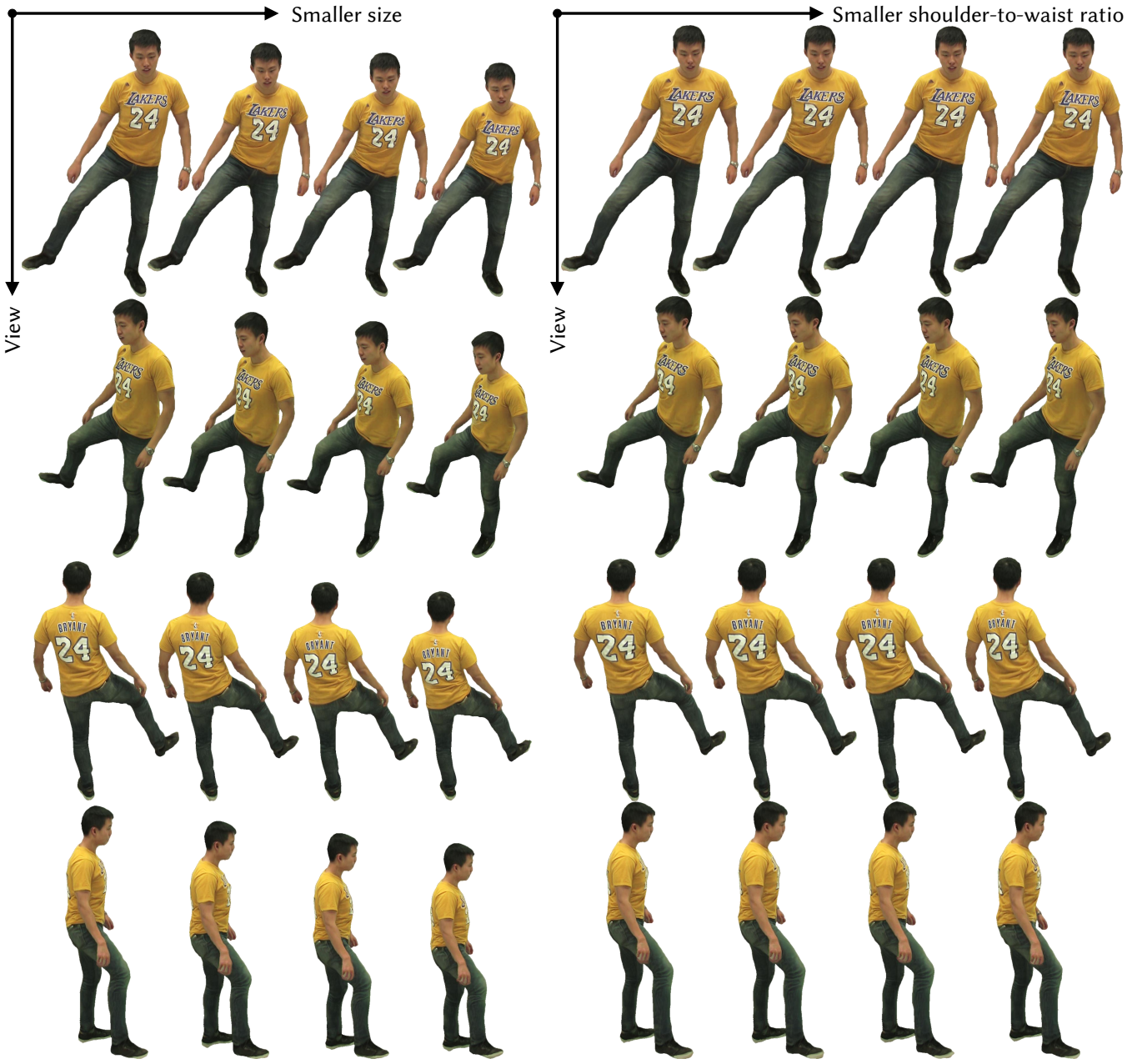

Fig. 8. Novel view synthesis results of reshaping. By changing the SMPL parameters, we can conveniently deform the human performer. We present the performer whose size is getting smaller and shoulder-to-waist ratio is getting smaller from left to right. With the help of view-consistent UV coordinates encoded by UV volumes, our model still renders view-consistent images with challenging shape parameters. Here, The horizontal axis shows shape changing and the vertical axis shows views changing. All results are rendered from novel views.

Fig. 9. 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.

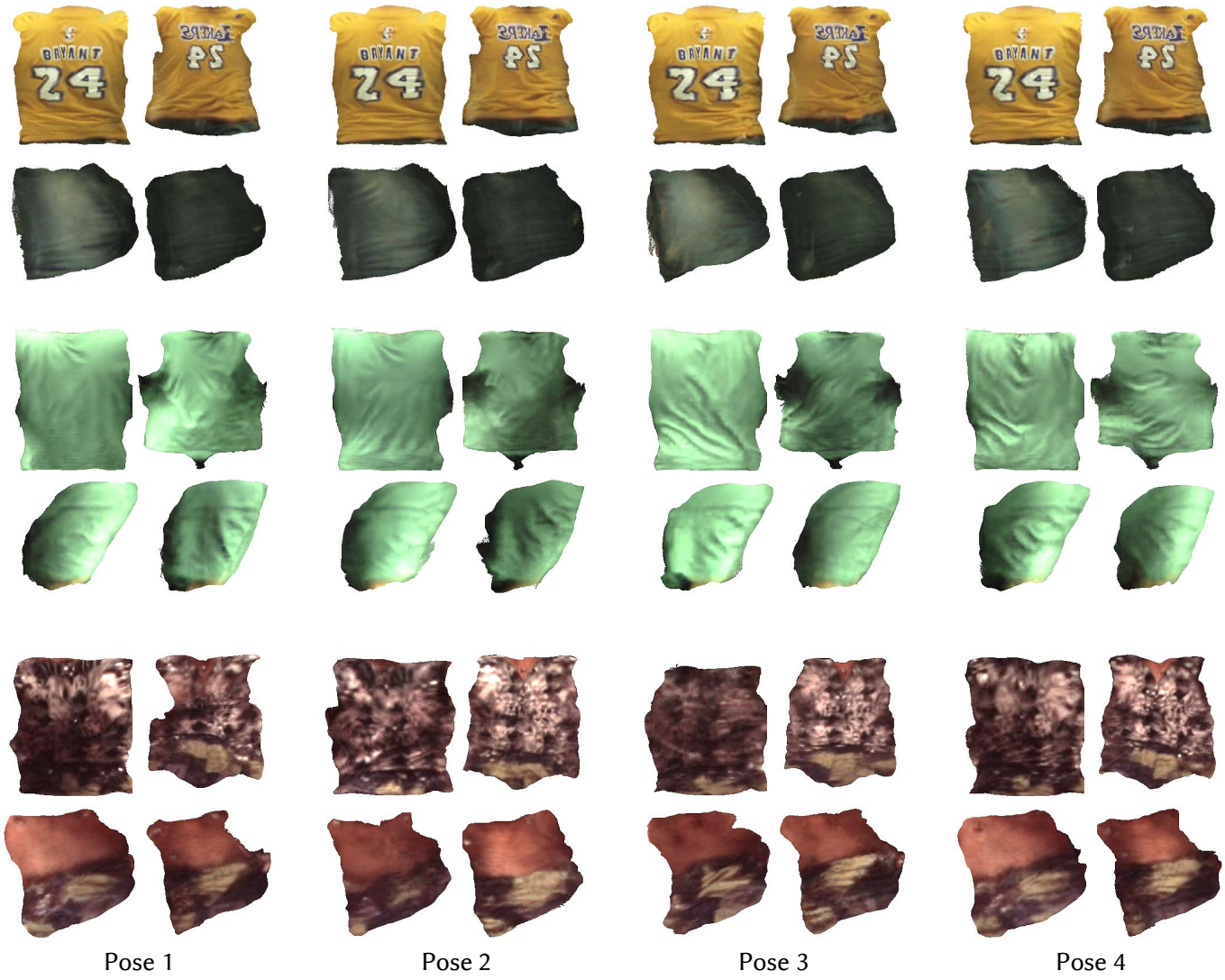

Fig. 10. Visualization of neural texture stacks under different poses. Details like the folds of clothing vary from motion to motion, as does the topology. Therefore, we propose pose-driven neural texture stacks to describe textures at different times, which enables us to handle dynamic 3D reconstruction tasks and to generalize our model to unseen poses.

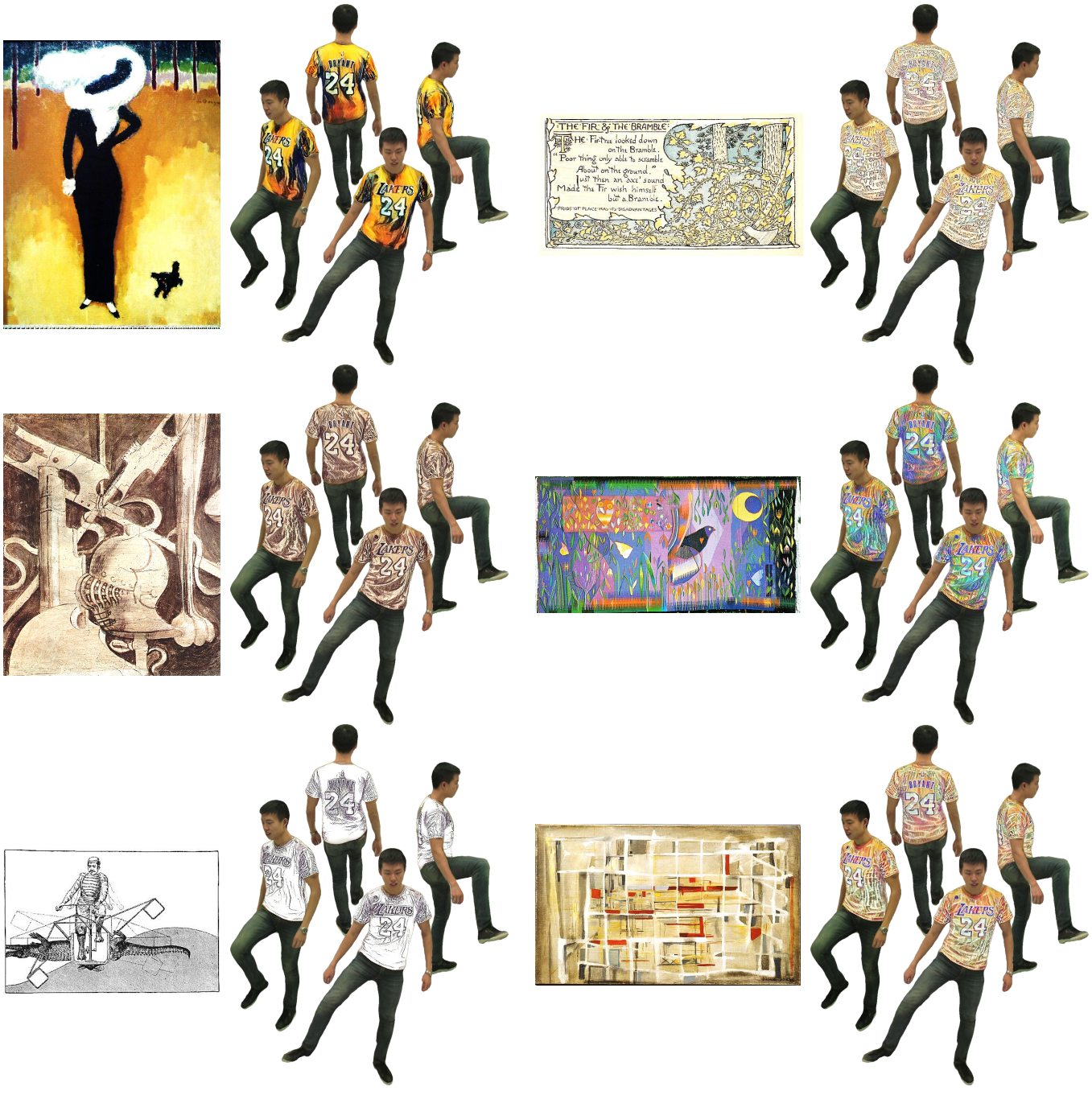

Fig. 11. Our model supports rendering a stylized dynamic human with any arbitrary artistic painting, which can be applied in controllable 3D style transfer with multi-view consistency.

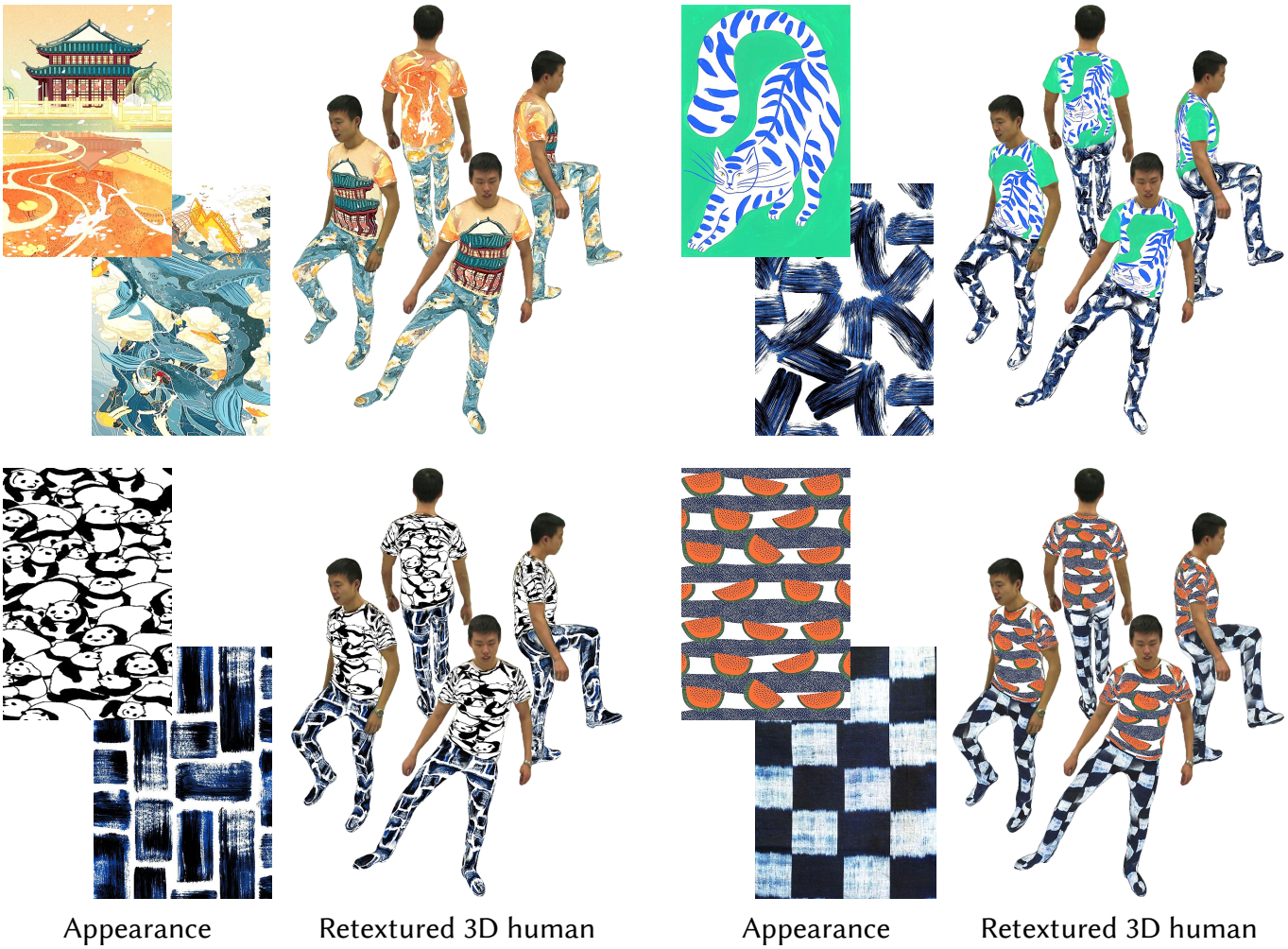

Fig. 12. Our model allows us to generate free-view human performance with a user-provided cloth texture image, which enables some interesting applications such as real-time 3D virtual try-on. We collect these appearance images from the Internet.

Fig. 13. 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.