PKU-DyMVHumans: A Multi-View Video Benchmark for High-Fidelity Dynamic Human Modeling

Supplementary Material

This supplementary material presents more details and additional results not included in the main paper due to page limitation. The list of items included are:

- More dataset information in Sec. A, including visualizations of data samples, and per-category data distribution;
- Additional experimental details are provided in Sec. B. These include implementation details of neural scene decomposition, additional visualizations and quantitative results of novel view synthesis, as well as more evaluation results on dynamic human modeling.

A. Additional Dataset Information

A.1. More Visualizations of Data Samples

Fig. 1 presents an overview of the sample for each scene in PKU-DyMVHumans. It can be seen that each sample has a distinct texture, motion, and interactions, covering a wide range of fundamental and complex dynamic performances. As shown in Fig. 2 and Fig. 3, a set of multi-view images is illustrated for the 1080P and 4K Studio sequences category, respectively. It clearly shows the differences between each view and provides comprehensive categories in our dataset.

A.2. Per-category Data Distribution

PKU-DyMVHumans contains a massive scale of subjects (32), scene (45), sequences (2,668) and frames (8.2M). We provide a full scene distribution with the number of frames for each action category in Fig. 4.

B. Additional Experimental Details

B.1. Neural Scene Decomposition

Method overview. For the scene captured by multi-view images, we use COLMAP [3] and BGMv2 [1] to get sparse 3D points and coarse object masks as co-inputs, and predict a dense, geometrical consistent object map, as well as a textural, completed background for each image. Fig. 5 shows an overview of our contemporary work Surface-SOS [7], in which multi-view geometric constraints are embedded in the form of dense one-to-one mapping in 3D surface representation. By connecting SDF-based surface representation to geometric consistency, and applying volume rendering to train the network with robustness, it can reconstruct the foreground object geometry and appearance over time.

Implementation details. To train Surface-SOS [7], we introduce photometric and geometric losses to supervise the training process, with the multi-view images serving as the primary supervision signal. Our objective is to achieve fine-grained object segmentation and analyze the correlation between the neural surface representation and object segmentation. In this study, we evaluate two approaches: NeRF-based segmentation, which does not introduce the normal to regularize the output SDF implicitly, and SDF-based segmentation, which provides the SDF-based surface representation for cross-view geometric constraints.

3D Segmentation of scenes with a single/multiple foreground object. As shown in Fig. 6, Surface-SOS successfully refines the segmentation remarkably using two neural representation models. When the normal is not introduced to implicitly regularize the output SDF (i.e., NeRF-based segmentation), it often produces noisy segmentation. However, when providing the SDF-based surface representation, the network is able to learn 3D geometry implicitly and generate an accurate foreground decomposition. These examples demonstrate that accurate prediction of object geometry with SDF-based surface representation is beneficial for object segmentation.

B.2. Novel View Synthesis

We conducted an additional experiment on the remaining scenes in PKU-DyMVHumans dataset. The complete quantitative comparisons are presented in Tab. 1. Additionally, we present additional qualitative comparisons in Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11. Consistent with the results in the main paper, PKU-DyMVHumans dataset offers a wide range of shapes and appearances, providing a comprehensive foundation for evaluating various methods for novel view synthesis in terms of human performance.

B.3. Dynamic Human Modeling

More Analyses of Dynamic Human Modeling. Free-viewpoint rendering of a moving subject from a monocular self-rotating video is a complex yet desirable setup. In the 4K Studio sequences category, we provide monocular self-rotating videos of human performers. These videos demonstrate the versatility of our dataset in synthesizing novel views of dynamic humans from fixed monocular cameras. To further illustrate this, we conducted additional experiments using HumanNeRF [6] baseline, a free-viewpoint rendering method for a moving subject. We selected 4 scenes from the 4K Studio sequences category with diverse motions and appearances and used images captured by camera 27, resulting in sequences ranging from 250 to 300.

We provide four visual examples of our challenging sce-

Table 1. Results of per-scene novel view synthesis on 4 action categories.

Action Type	Scenes	NeuS [4]			Instant-NGP [2]			NeuS2 [5]		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Dance	1080_Dance_Dunhuang_Pair_f14f15	21.38	0.959	0.042	26.37	0.974	0.035	25.34	0.967	0.044
	1080_Dance_Dunhuang_Single_f12	25.19	0.978	0.024	31.46	0.983	0.019	30.61	0.985	0.020
	1080_Dance_Dunhuang_Single_f13	25.37	0.979	0.021	33.91	0.989	0.013	31.31	0.984	0.016
	1080_Dance_Dunhuang_Single_f14	25.03	0.977	0.025	32.02	0.988	0.016	30.34	0.985	0.016
	1080_Dance_Dunhuang_Single_f15	25.61	0.978	0.029	34.91	0.992	0.014	32.27	0.990	0.016
	1080_Dance_Jazz_Single_c22	26.59	0.984	0.018	31.41	0.987	0.011	35.38	0.991	0.010
	1080_Dance_Tibetan_Single_c22	26.26	0.984	0.020	33.69	0.991	0.016	34.28	0.990	0.014
	1080_Dance_Banquet_Single_c23	26.73	0.984	0.016	36.20	0.993	0.009	35.56	0.993	0.009
	Average	25.27	0.978	0.024	**32.50**	**0.987**	**0.017**	31.90	0.986	0.018
Kungfu	1080_Kungfu_Weapon_Pair_m12m13	19.51	0.941	0.061	22.31	0.955	0.014	22.07	0.962	0.048
	1080_Kungfu_Double_Pair_m12m13	18.48	0.939	0.062	20.42	0.948	0.144	22.45	0.965	0.047
	1080_Kungfu_Basic_Pair_c24c25	25.48	0.979	0.024	31.96	0.989	0.017	30.99	0.986	0.019
	1080_Kungfu_Fan_Single_m12	25.72	0.981	0.024	33.16	0.988	0.021	30.56	0.985	0.021
	1080_Kungfu_Taichi_Single_m12	22.16	0.976	0.027	30.94	0.988	0.020	29.60	0.986	0.020
	1080_Kungfu_Shaolin_Single_m12	23.01	0.978	0.029	31.85	0.989	0.017	32.08	0.989	0.016
	1080_Kungfu_Sword_Single_m13	23.24	0.978	0.023	31.10	0.986	0.019	29.39	0.985	0.019
	1080_Kungfu_Spear_Single_m13	22.53	0.973	0.033	31.85	0.989	0.018	28.16	0.985	0.021
	1080_Kungfu_Kick_Single_m13	20.16	0.970	0.032	29.43	0.986	0.032	28.68	0.987	0.024
	1080_Kungfu_Basic_Single_m13	23.17	0.978	0.021	28.18	0.982	0.024	28.53	0.985	0.019
	1080_Kungfu_Tongbeiquan_Single_m13	23.59	0.980	0.024	32.19	0.991	0.016	31.61	0.988	0.017
	1080_Kungfu_Nunchuck_Single_m14	21.21	0.976	0.024	29.62	0.985	0.046	28.32	0.984	0.022
	1080_Kungfu_Nanquan_Single_c24	25.81	0.983	0.026	37.62	0.995	0.013	34.67	0.993	0.014
	1080_Kungfu_Broadsword_Single_c24	25.65	0.985	0.018	35.28	0.993	0.014	37.24	0.994	0.009
	1080_Kungfu_Boxing_Single_c25	24.31	0.981	0.024	38.37	0.996	0.009	36.91	0.995	0.009
	Average	22.94	0.973	0.030	**30.95**	0.984	0.028	30.08	**0.985**	**0.022**
Sport	1080_Sport_Football_Single_m11	24.91	0.983	0.017	29.83	0.982	0.018	30.50	0.986	0.016
	1080_Sport_Taekwondo1_Pair_m11c21	23.72	0.970	0.037	32.50	0.989	0.019	27.12	0.981	0.029
	1080_Sport_Badminton_Single_f11	25.22	0.980	0.028	34.29	0.993	0.011	33.79	0.993	0.014
	Average	24.62	0.977	0.028	**32.20**	**0.988**	**0.016**	30.47	0.987	0.020
Fashion Show	4K_Studios_Show_Pair_f16f17	23.26	0.977	0.036	35.02	0.991	0.020	32.12	0.987	0.025
	4K_Studios_Show_Pair_f18f19	22.80	0.976	0.031	34.24	0.993	0.016	32.23	0.992	0.014
	4K_Studios_Show_Single_f16	20.38	0.975	0.042	34.49	0.990	0.012	34.33	0.993	0.012
	4K_Studios_Show_Single_f17	23.07	0.982	0.036	35.08	0.992	0.021	34.11	0.992	0.018
	4K_Studios_Show_Single_f18	22.95	0.983	0.027	36.73	0.995	0.012	34.32	0.994	0.013
	4K_Studios_Show_Single_f19	24.36	0.986	0.024	38.94	0.996	0.010	37.03	0.995	0.011
	4K_Studios_Dance_Single_f20	22.67	0.981	0.028	30.50	0.986	0.026	31.67	0.989	0.018
	Average	22.79	0.980	0.032	**35.00**	**0.992**	0.017	33.69	**0.992**	**0.016**
Average		23.90	0.977	0.029	**32.66**	**0.988**	**0.019**	31.53	0.987	**0.019**

nario dataset in Fig. 12. While body pose and non-rigid motion were not completely recovered, as the movement of the skirts relied on the temporal dynamics of subject motion. We hope the result points in a promising direction towards modeling humans in complex poses and clothing, and eventually achieving fully photorealistic, freeviewpoint rendering of moving people.

References

[1] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L. Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8762–8771, 2021. 1

[2] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2

[3] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[4] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 27171–27183, 2021. 2

[5] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[6] 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, 11

[7] Xiaoyun Zheng, Liwei Liao, Jianbo Jiao, Feng Gao, and Ronggang Wang. Surface-sos: Self-supervised object segmentation via neural surface representation. *IEEE Transactions on Image Processing*, 33:2018–2031, 2024. 1, 5

Figure 1. **Data overview**. PKU-DyMVHumans is a dynamic, human-centric dataset with diverse subjects, each featuring highly detailed appearances and complex human motions.

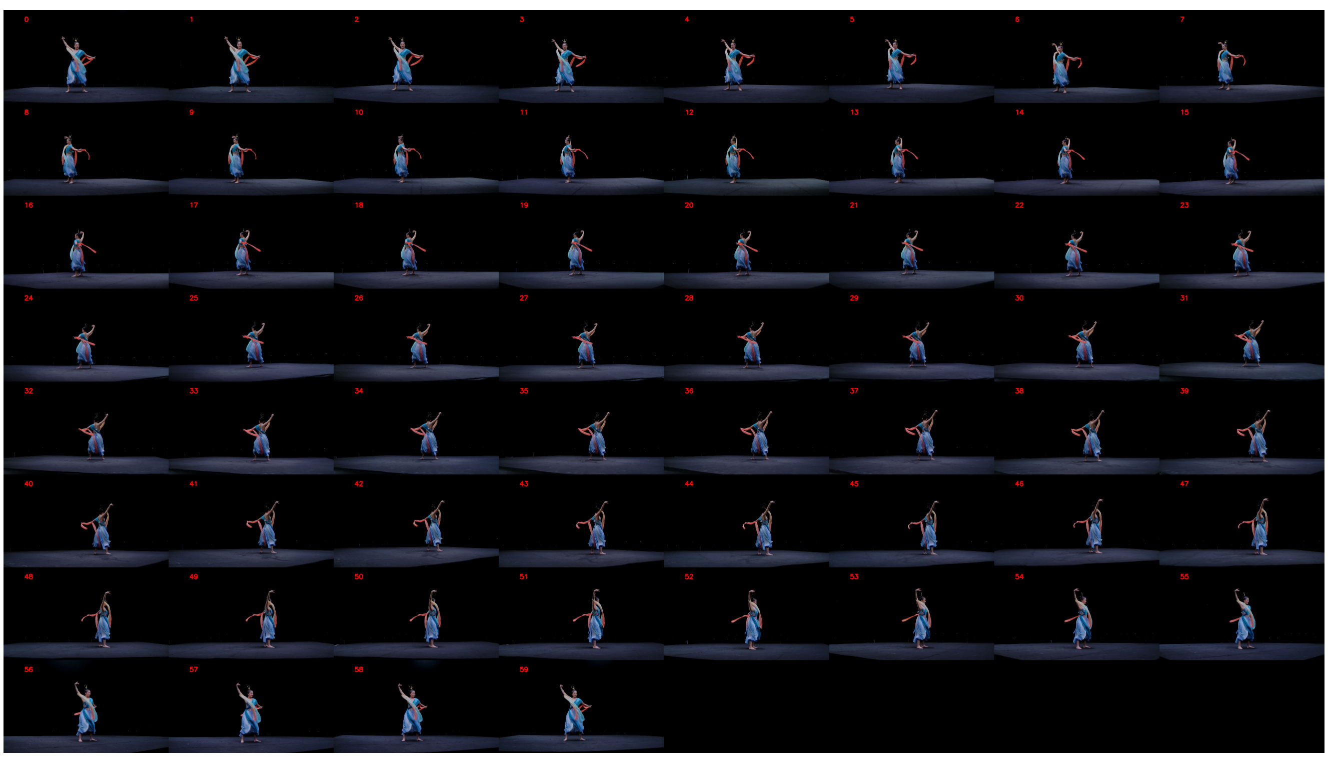

Figure 2. A set of example multi-view images in the 1080P sequences (1080_Dance_Dunhuang_Single_f12).

Figure 3. A set of example multi-view images in the 4K Studio sequences (4K_Studios_Show_Pair_f18f19).

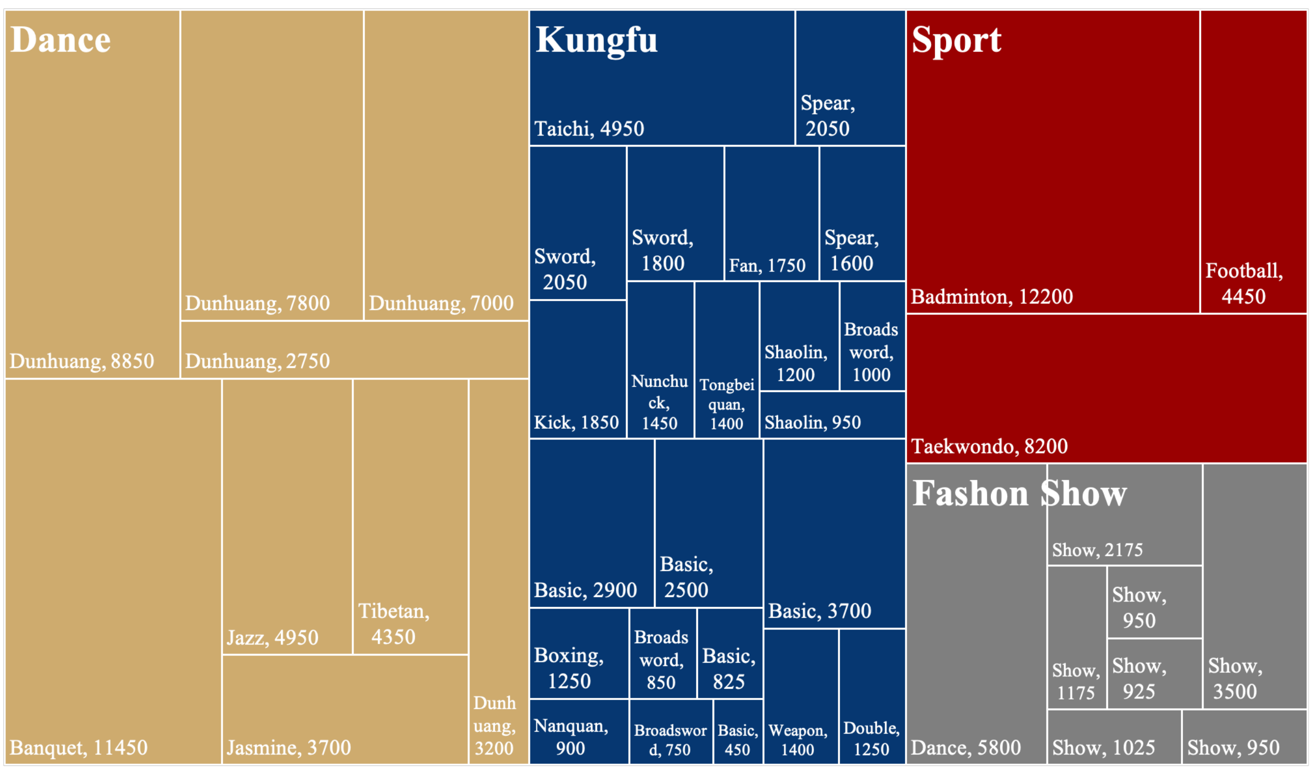

Figure 4. A scene list that provides a detailed breakdown of the number of frames in each category.

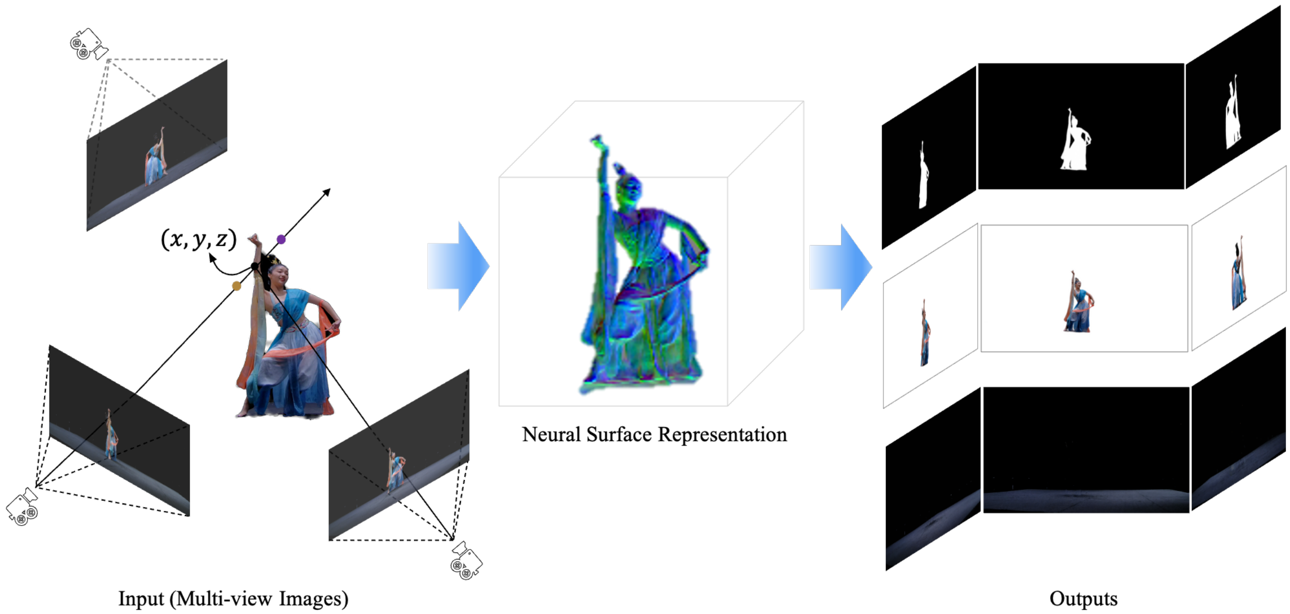

Figure 5. Surface-SOS [7], a self-supervised learning framework towards delicate segmentation by combining 3D neural surface representation power from multi-view images of a scene.

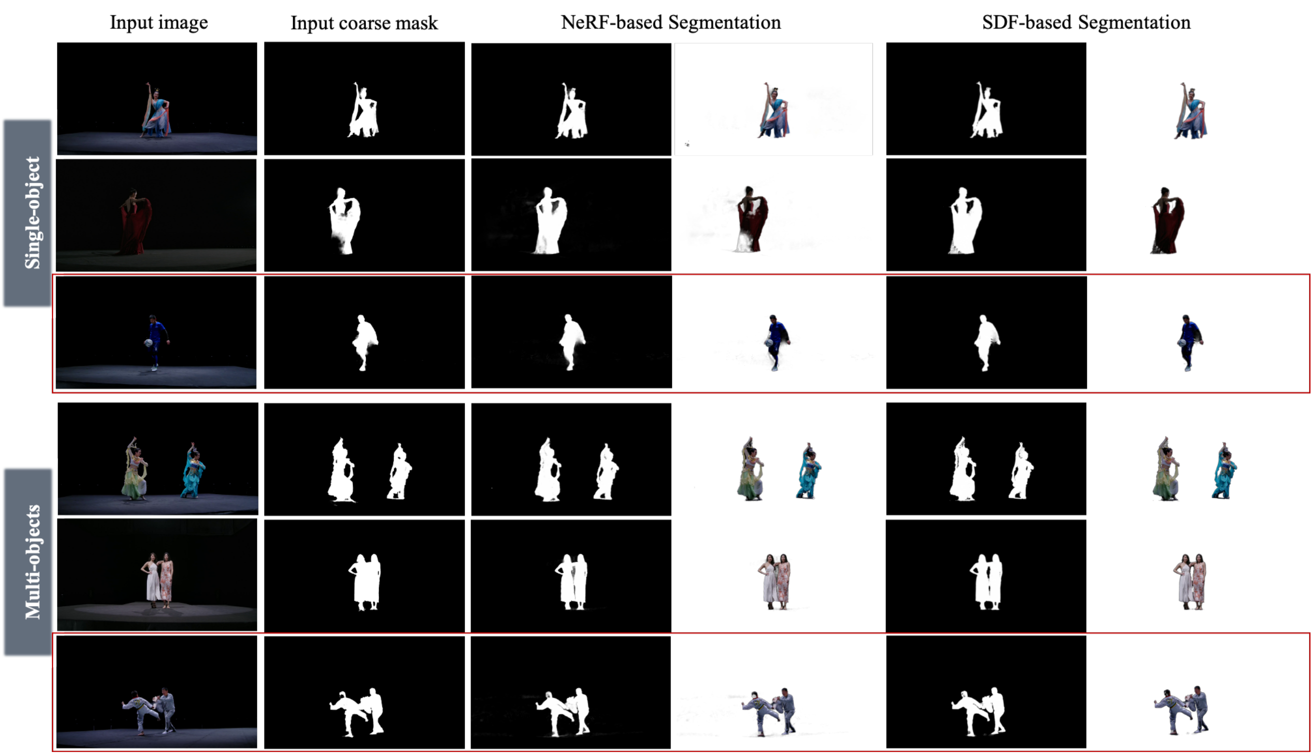

Figure 6. Qualitative comparison on 3D segmentation of scenes with a single/multiple foreground object.

Figure 7. More visualizations results of scene sample (4K Studios).

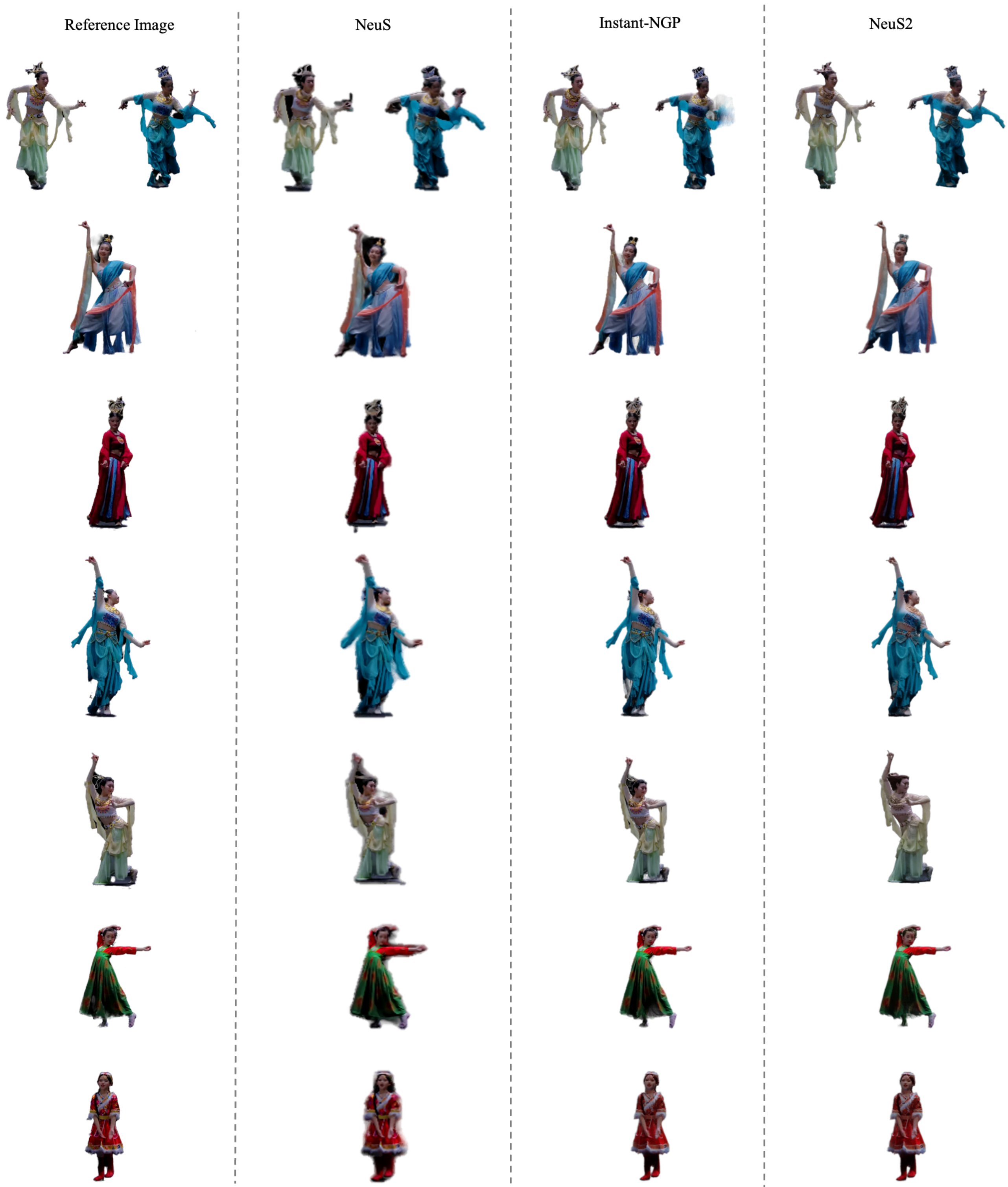

Figure 8. More visualizations results of scene sample (dance).

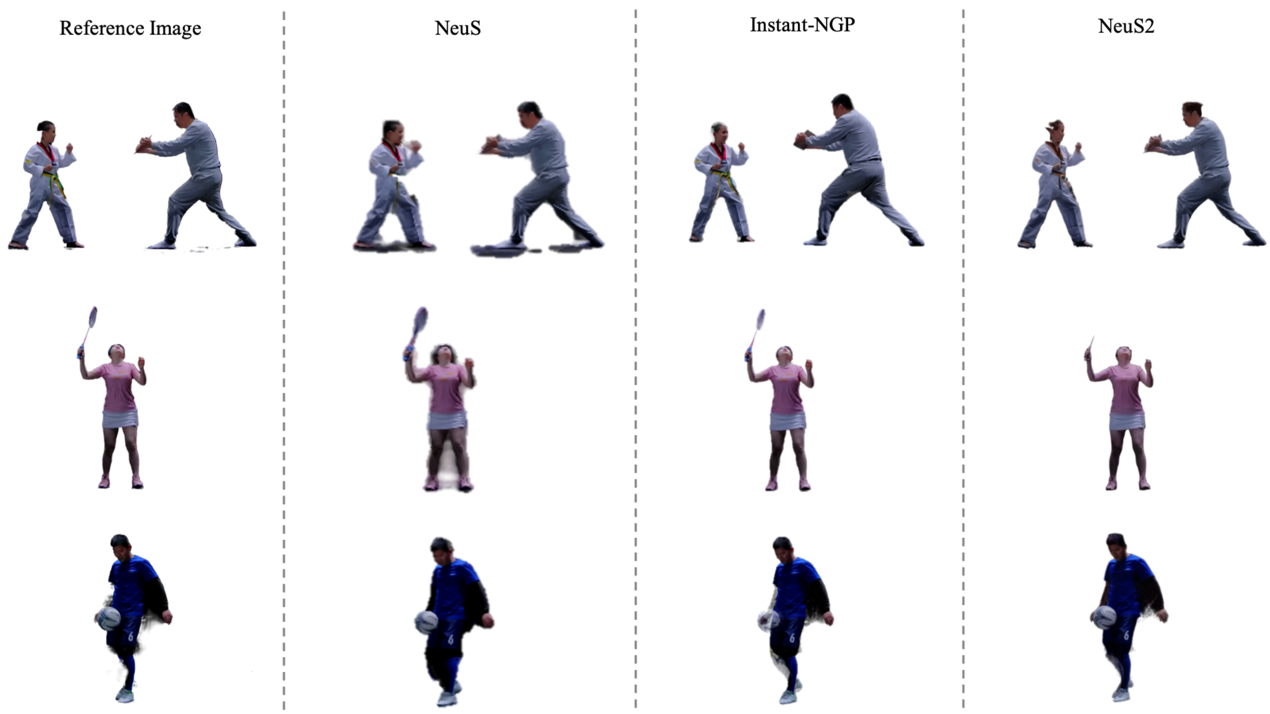

Figure 9. More visualizations results of scene sample (sport).

Figure 10. More visualizations results of scene sample (kungfu).

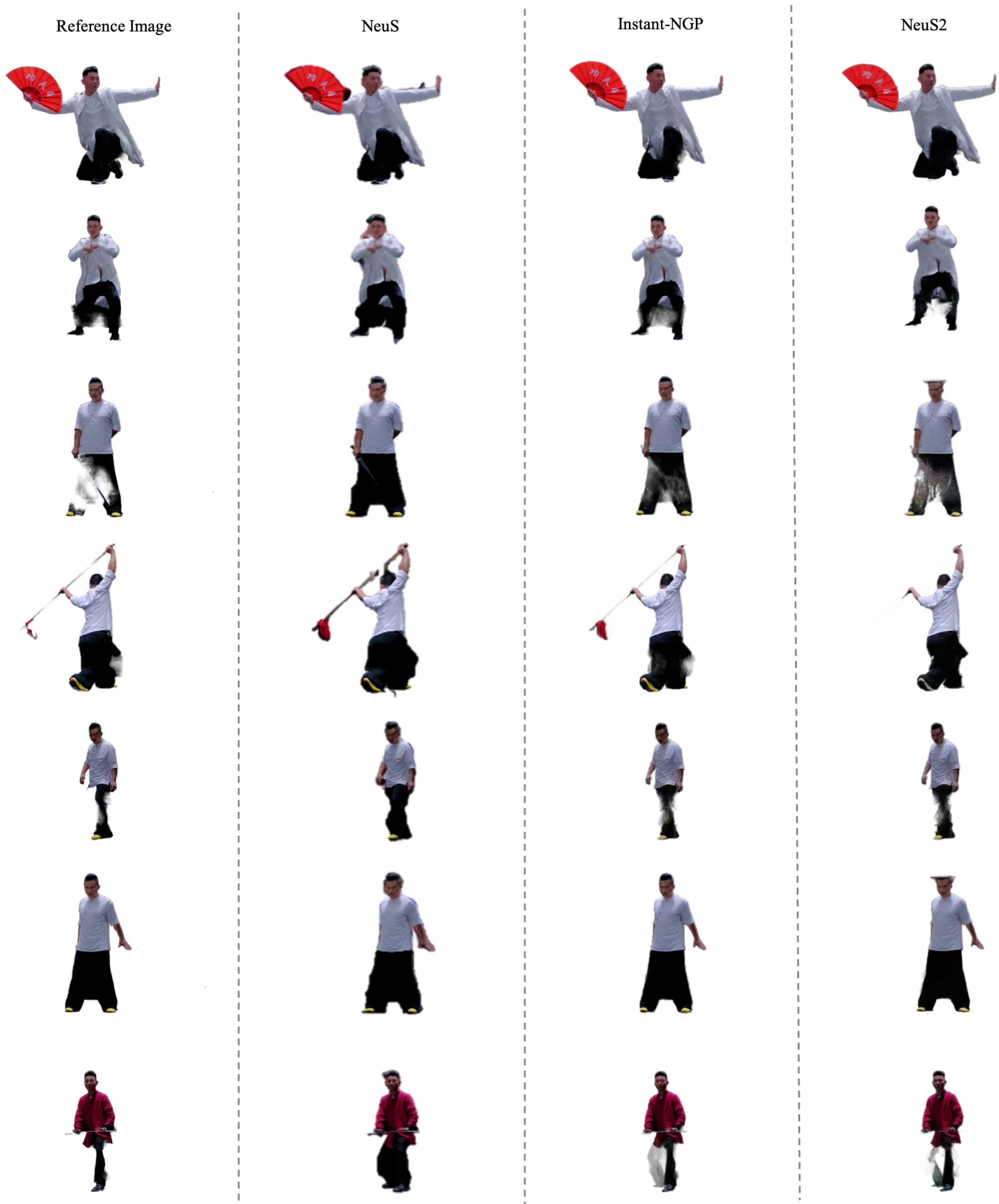

Figure 11. More visualizations results of scene sample (kungfu).

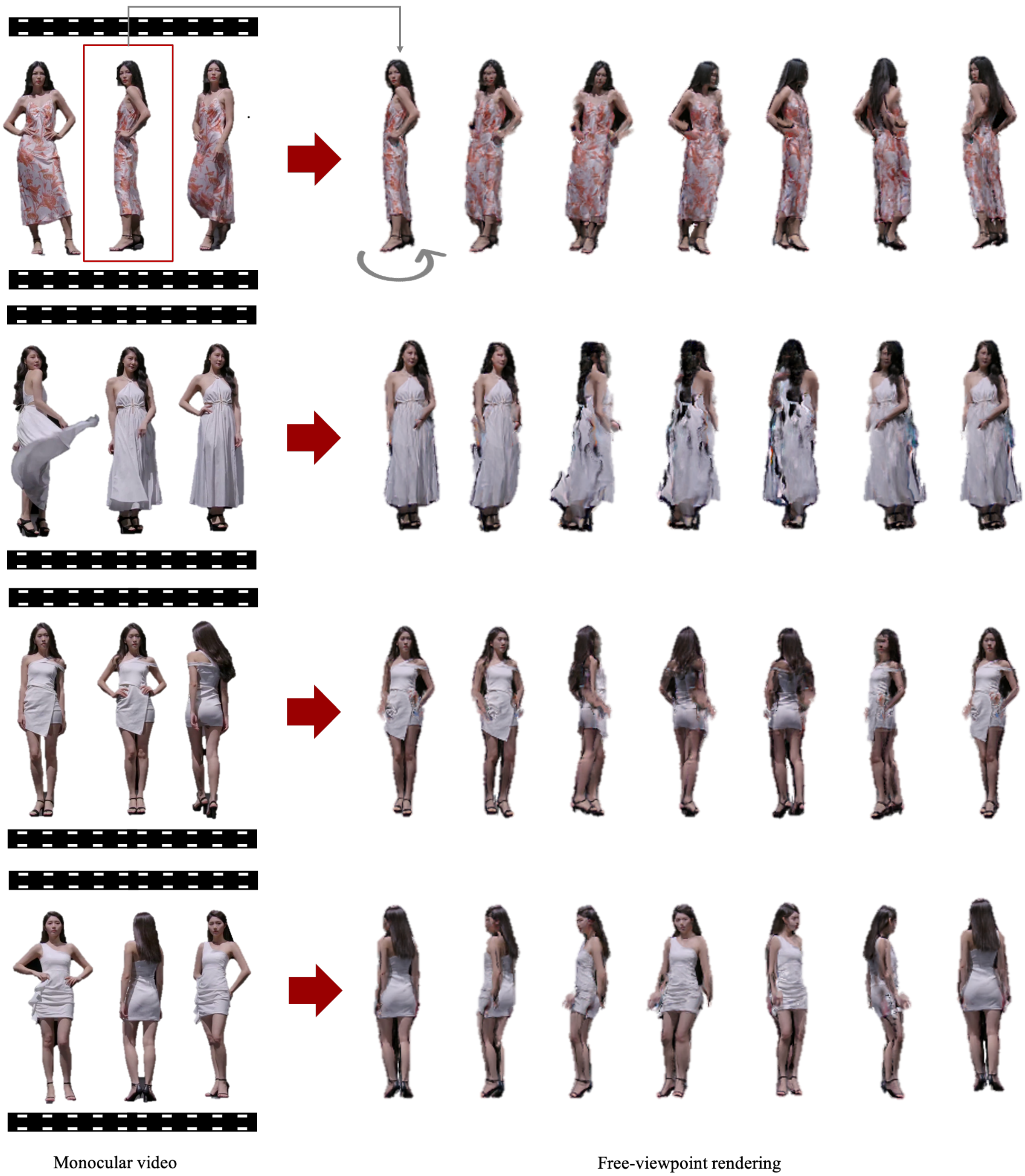

Figure 12. Visual examples of free view synthesis on the four scenes data are provided. The input is a monocular video capturing a human performing complex movements (left). The HumanNeRF [6] generates a free-viewpoint rendering for any frame in the sequence (right).