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

Xiaoyun Zheng [1,2] Liwei Liao [1,2] Xufeng Li [3] Jianbo Jiao [4]
Rongjie Wang [2] Feng Gao [5] Shiqi Wang [3] Ronggang Wang [1†]
[1] Peking University Shenzhen Graduate School [2] Peng Cheng Laboratory
[3] City University of Hong Kong [4] University of Birmingham [5] Peking University
xyun_z@stu.pku.edu.cn {levio, gaof}@pku.edu.cn xufengli2-c@my.cityu.edu.hk
j.jiao@bham.ac.uk wangrj@pcl.ac.cn shiqwang@cityu.edu.hk rgwang@pkusz.edu.cn

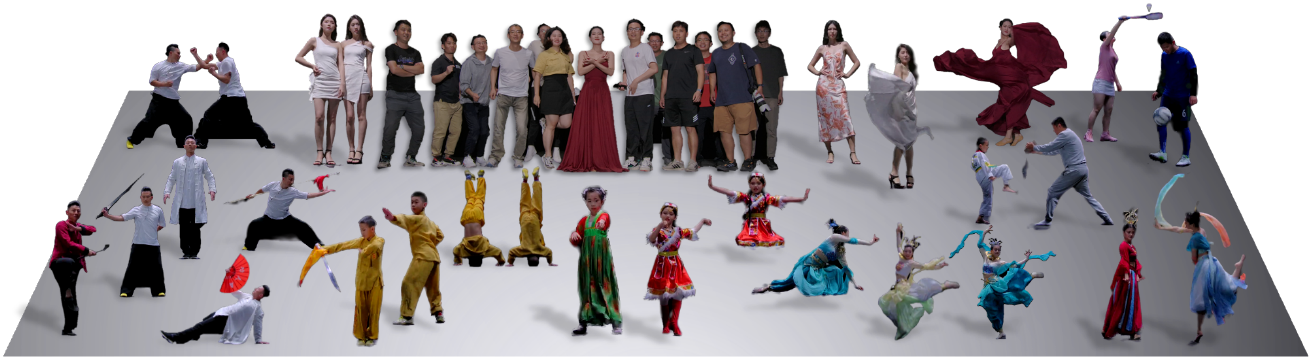

Figure 1. We present **PKU-DyMVHumans**, a versatile human-centric dataset designed for high-fidelity reconstruction and rendering of dynamic human performances from dense multi-view videos. It comprises 32 humans across 45 different dynamic scenarios, each featuring highly detailed appearances and complex human motions.

Abstract

High-quality human reconstruction and photo-realistic rendering of a dynamic scene is a long-standing problem in computer vision and graphics. Despite considerable efforts invested in developing various capture systems and reconstruction algorithms, recent advancements still struggle with loose or oversized clothing and overly complex poses. In part, this is due to the challenges of acquiring high-quality human datasets. To facilitate the development of these fields, in this paper, we present PKU-DyMVHumans, a versatile human-centric dataset for high-fidelity reconstruction and rendering of dynamic human scenarios from dense multi-view videos. It comprises 8.2 million frames captured by more than 56 synchronized cameras across diverse scenarios. These sequences comprise 32 human subjects across 45 different scenarios, each with a high-detailed appearance and realistic human motion. Inspired by recent advancements in neural radiance field (NeRF)-based scene representations, we carefully set up an off-the-shelf framework that is easy to provide those state-of-the-art NeRF-based implementations and benchmark on PKU-DyMVHumans dataset. It is paving the way for various applications like fine-grained foreground/background decomposition, high-quality human reconstruction and photo-realistic novel view synthesis of a dynamic scene. Extensive studies are performed on the benchmark, demonstrating new observations and challenges that emerge from using such high-fidelity dynamic data. The project page and data is available at: https://pku-dymvhumans.github.io.

1. Introduction

We are entering an era in which the distinction between virtual and real worlds is becoming increasingly blurred. High-quality reconstruction and photo-realistic rendering of human activities are crucial for many immersive applications, including AR/VR, 3D content production, and entertainment. However, the process of reconstructing human activities and providing photo-realistic rendering from multiple viewpoints is currently a cutting-edge but challenging technique. The limited availability of real-world datasets is impeding progress in this critical task.

Early solutions adopted multi-view stereo techniques to explicitly reconstruct textured meshes[3, 7, 13], or employed view interpolation for image-based rendering[6, 50]. However, these methods require pre-scanned templates or dense camera rigs (up to 100 cameras [7]), which are expensive and not easily portable or affordable. To simplify the capture systems, some works have used commodity depth sensors to build real-time reconstruction systems [5, 47], but they are limited by inherent self-occlusion constraints. Moreover, these approaches are prone to producing reconstruction errors and rendering artifacts in scenes with thin structures, specular surfaces, and topological changes.

The recent advancements in learning-based techniques have made it possible to achieve robust human attribute reconstruction [14, 31, 32] using only RGB input. Specifically, the methods PIFu [31] and PIFuHD [32] utilize pixel-aligned implicit functions to accurately reconstruct clothed humans with intricate geometric details. However, these approaches struggle with generating realistic appearances, largely due to their reliance on implicit texture representation. On the other hand, recent advancements in neural rendering techniques have showcased impressive capabilities in synthesizing novel views of general static scenes using neural radiance field (NeRF) representations [23]. This approach has also shown promising results in human modeling [18, 30, 52], although it still relies on relatively dense multiview videos as input. To address this limitation, some recent studies [15, 29] propose using human body priors to assist in learning human representations.

Methods that utilize radiance fields aim to optimize both the input and output representation. However, these problems still remain open. Currently, it is still challenging to achieve a balance between realism, robustness to complex poses and clothing, reasonable computation time, and compatibility with complex scenes. One of the challenges stems from the flexibility of humans, as they move in complex ways against natural backgrounds, and their clothing and muscles deform. Additionally, other factors such as occlusion may require comprehensive scene modeling beyond just the humans present. Another pressing issue is the lack of high-quality, high-detail datasets of complex scenarios, as well as the need for accurate human pose and segmentation to training networks, which makes it difficult to accurately evaluate multi-person performance capture systems. Looking ahead, we anticipate the development of a new versatile model that is powerful yet general in terms of both appearance rendering and motion modeling. This model should be capable of generating realistic dynamic details without the need for pre-scanning or pre-processing efforts.

In this work, we propose PKU-DyMVHumans, a versatile human-centric dataset that includes 32 dynamic humans performing various actions and wearing different clothing styles, as shown in Fig. 1. The dataset consists of approximately 8.2 million frames and aims to address the lack of large-scale and high-fidelity human performance datasets. Compared to existing benchmark datasets, PKU-DyMVHumans dataset offers appealing characteristics: **1) High-fidelity human performance**: We construct a professional multi-view system to capture humans in motion, which contains 56/60 synchronous cameras with 1080P or 4K resolution. **2) High-detailed appearance**: It captures complex cloth deformation, and intricate texture details, such as delicate satin ribbon and vivid textures of classical headwear. **3) Complex human motion**: It covers a wide range of special costume performances, artistic movements, and sports activities. **4) Human-object/scene interactions**: These include human-object interactions, as well as challenging multi-person interactions and complex scene effects (*e.g.*, lighting, shadows, and smoking).

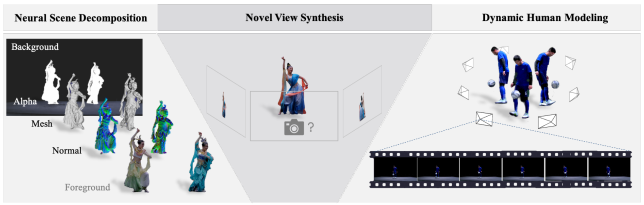

Figure 2. **Research with PKU-DyMVHumans.** It supports various research topics, including neural scene decomposition, novel view synthesis, and dynamic human modeling.

The primary goal of PKU-DyMVHumans is to enable high-fidelity reconstruction and rendering of human performances from dense multi-view videos. To leverage the extensive exploration space offered by our dataset, we develop a unified framework that enables the implementation and evaluation of state-of-the-art NeRF-based approaches on PKU-DyMVHumans. As depicted in Fig. 2, the framework opens up possibilities for various applications, including fine-grained foreground/background decomposition, high-quality human reconstruction, and photo-realistic synthesis of dynamic humans. Extensive studies have been conducted on the benchmark, revealing new insights and challenges associated with the use of such high-fidelity dynamic data.

2. Related Work

Human performance capture and datasets. Human performance capture, which involves capturing the full pose and non-rigid surface deformation of people wearing regular clothing in a space-time coherent 4D manner, has revolutionized the film and gaming industry in recent years [8]. The early existing methods learn from high-quality data obtained from optical marker-based capture systems [11, 48]. Other recent methods utilize depth sensor [34, 38, 46] or volumetric data [8, 9] to achieve reliable human motion and geometry. However, the sophisticated setup restricts their practical deployment. With the emergence of neural rendering techniques, rendering realistic humans

Table 1. **Comparisons of multi-view human datasets.** We compare the proposed dataset with previous human-centric multi-view datasets in terms of scale, attribute, and resolution. The gray color indicates the marker-based capture data with reliable human motion or pre-scanned humans reconstructed from depth sensors or stereo camera arrays. Differing from these efforts, PKU-DyMVHumans features high-fidelity dynamic sequences with high-detailed appearance, complex human motions, as well as challenging multi-person interactions.

Dataset	**#Cam.**	**#Subj.**	**#Scene**	**#Seq.**	**#Frame**	**Dynamic**	**Interaction**	**High-detail**	**Multi-person**	**Resolution**
Human3.6M[11]	4	11	15	839	3.6M	✓	✓	✗	✗	1000×1002
THUman2.0[48]	60	200	-	-	-	✗	✗	✗	✗	512×512
Hi4D[46]	8	40	20	100	11K	✓	✓	✗	✓	940×1280
DNA-Rendering[38]	60	500	1,187	25,320	67.5M	✓	✓	✓	✗	4096×3000, 2448×2048
DynaCap[9]	50-101	4	5	440	11.8M	✓	✗	✗	✗	1285×940
MultiHuman[54]	128	50	-	-	-	✗	✓	✗	✓	-
THuman5.0[34]	32	10	10	320	96K	✓	✗	✓	✗	4096×3000
HuMMan[2]	11	132	20	4,466	278K	✓	✓	✗	✗	1920×1080
UltraStage[56]	32	100	20	-	192K	✗	✗	✓	✗	7860×4320
Actors-HQ[10]	160	8	21	2,560	39.8K	✓	✗	✓	✗	4096×3072
HUMBI[49]	107	772	4	-	26M	✗	✗	✓	✗	1920×1080
NHR[44]	56/72	4	3	320	47K	✓	✓	✗	✗	1024×768
AIST++[37]	9	30	10	1,408	10.1M	✓	✓	✗	✓	1920×1080
ZJU_Mocap[30]	21	9	6	2̃07	185K	✓	✗	✗	✗	1024×1024
THuman4.0[55]	24	3	3	320	250K	✓	✗	✗	✗	1330×1150
ENeRF-Outdoor[18]	18	7	4	72	86K	✓	✓	✗	✓	1920×1080
FreeMan[39]	8	40	123	8,000	11.3M	✓	✓	✗	✗	1920×1080
PKU-DyMVHumans (Ours)	56/60	32	45	2,668	8.2M	✓	✓	✓	✓	3840×2160, 1920×1080

directly from images has become a trend. Such a setting [1] usually requires the dataset equipped with high-quality dense view images [44, 49] or accurate annotations like human body keypoints and foreground segmentation [30, 37, 39, 55]. However, recent methods still often struggle to reconstruct models accurately when the clothes are loose, oversized, or when the poses are too complex. Differing from these efforts, we aim to provide a versatile human-centric dataset designed for high-fidelity reconstruction and rendering of dynamic human performances from dense multi-view videos. The unfold comparisons between PKU-DyMVHumans and the existing datasets are given in Tab. 1. When compared to the published benchmark datasets, PKU-DyMVHumans features high-detailed appearance, complex motion, as well as challenging human-object interactions, multi-person interactions and complex scene effects (*e.g.*, lighting, shadows, and smoking).

Neural implicit representations. In the domain of photo-realistic novel view synthesis and 3D scene modeling, differentiable neural rendering based on various data proxies has achieved impressive results and gained popularity. Traditional methods rely on explicit geometric representations, such as depth maps [27], point cloud [44], meshes [26], and voxel grids [36]. Recently, coordinate-based networks have become a popular choice for implicit 3D scene representations, such as radiance fields [23], signed distance fields (SDF) [25, 28], or occupancy [22]. The emerging neural implicit representation approaches show promising results in novel view synthesis [20, 23, 24] and high-quality 3D reconstruction from multi-view images [12, 40, 41]. In the pioneering work of NeRF [23], a Multi-Layer Perceptron (MLP) is trained to encode a radiance field reconstructed from a set of input RGB images. However, it cannot extract high-quality surfaces due to the lack of surface constraints in the geometry representation. NeuS [40] addresses this limitation by representing the 3D surface as an SDF for high-quality geometry reconstruction. However, the explicit integration in NeuS makes it computationally intensive, resulting in slow training and limited applicability to static scene reconstruction. To overcome the slow training of deep coordinate-based MLPs, Instant-NGP [24] proposes a multi-resolution hash encoding technique, which has been proven effective in accelerating the training process.

Free view synthesis for dynamic scenes. Photo-realistic rendering of dynamic scenes from a set of input images is necessary for many applications. For dynamic scene modeling, some methods [16, 17, 45] consider dynamic scenes as a 4D domain and add the time dimension to the input spatial coordinate. This approach enables the implementation of space-time radiance fields. In the field of human rendering, several approaches [21, 30, 43] utilize human priors to model human motions and achieve free-viewpoint rendering of dynamic human. HumanNeRF [43] demonstrates the ability to render realistic humans from monocular video sequences. While these methods produce impressive rendering results, their training process is time-consuming. To address the need for fast dynamic scene reconstruction, NeuS2 [41] integrates multi-resolution hash encoding into SDF and proposes an incremental training strategy with a global transformation prediction component. This approach leverages shared geometry and appearance information in two consecutive frames. Besides, Tensor4D [35] proposes a hierarchical tri-projection decomposition method to learn high-quality dynamic scenes representation from sparse-view videos. This method models a 4D tensor using nine 2D feature planes, capturing spatio-temporal information in a compact and memory-efficient manner. However, they generally handle videos with short frames.

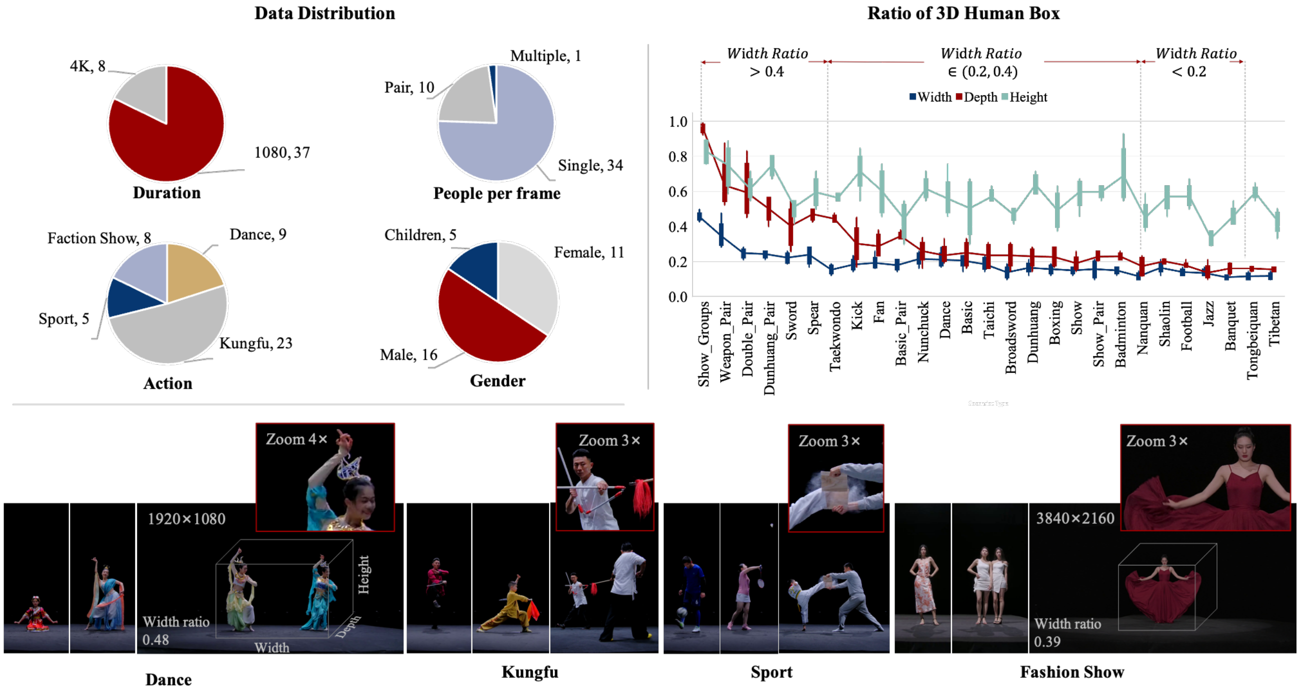

Figure 3. Category definition and distribution of the proposed PKU-DyMVHumans dataset.

3. The PKU-DyMVHumans Dataset

3.1. Data Capturing and Processing

Data capturing setting. Our goal is to design a system that can acquire human-centric datasets for high-quality human reconstruction and photo-realistic novel view synthesis in markerless multi-view capture settings. Our capture system is built on an indoor stage, lit by spotlights from above and equipped with a circular camera array. (1) For the 1080P sequences category, the camera system comprises 60 Z CAM E2 cameras operating at a resolution of 1920×1080 and 25 FPS. The 60 cameras are evenly distributed in a circle around the players, with each camera positioned approximately 6 meters from the system center. (2) For the 4K Studio, players stand on the stage are lit by a follow spotlight. This setup includes 56 calibrated cameras positioned in a large arc around the players. The distance from each camera to the system center is approximately 6 meters. In both situations mentioned above, we capture multiple sequences of challenging human performances, such as dance, kungfu, sport, and fashion shows.

Sparse reconstruction. Following the procedure outlined in [23, 40], we utilize the COLMAP Structure-from-Motion (SfM) algorithm [33] to calibrate camera intrinsic and extrinsic parameters. Our goal is to enable a neural processing pipeline capable of accurately reconstructing both static and dynamic scenes from multiple views using state-of-the-art neural implicit representation methods such as NeuS and NeRF-based approaches. Furthermore, we provide a data conversion tool that transforms the sparse model output from SfM into a format compatible with Instant-NGP [24], NeuS, and Tensor4D [35].

Foreground object segmentation. Each frame extracted from the original video is passed to BGMv2 [19] to generate the binary foreground object mask. This not only improves dense reconstruction but also contributes to the subsequent step of dynamic human reconstruction. In this context, our contemporary work Surface-SOS [53] aims for delicate segmentation by leveraging the cross-view consistency of neural implicit surface representation from a sparse set of posed images (Sec. 4.3).

3.2. Dataset Statistics and Distribution

The PKU-DyMVHumans dataset consists of 45 different dynamic scenarios, totalling approximately 8.2 million frames of recording. The sequences feature performers in different locations, engaging in various actions and clothing styles. There are 32 professional players, including 16 males, 11 females, and 5 children, performing 4 different action types: dance, kungfu, sport, and fashion show[1]. As shown in Fig. 3, the diversity of PKU-DyMVHumans stems from the wide variations in neural human modeling across different motion categories, as well as variances in human size, and character poses. This makes PKU-DyMVHumans a more challenging dataset compared

[1] All the actors have given consent in signed written forms to the use of their recordings.

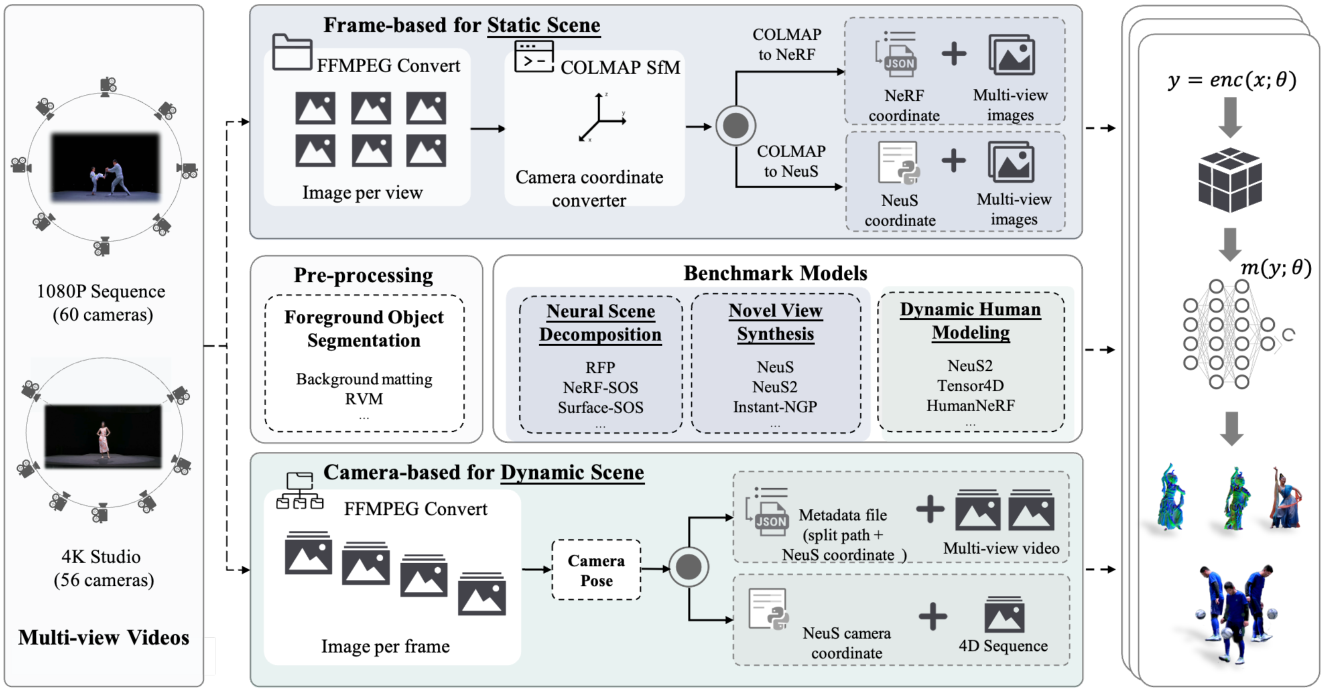

Figure 4. **Benchmarks pipeline of PKU-DyMVHumans.** Given a multi-view video input, the first step is to extract the frames and estimate the foreground object mask and camera parameters. Specifically, BGMv2 [19] is used to generate the binary foreground object mask. Afterwards, COLMAP SfM [33] is used to estimate camera parameters and generate a sparse point cloud. Using these components, we have constructed three benchmarks. (a) The implementation of NeRF by Instant-NGP requires providing initial camera parameters in JSON file format compatible with the original NeRF codebase. (b) In addition to RGB and mask images, the NeuS implementation expects a camera file that contains a projection matrix and a normalization matrix for each image. (c) We also provide data conversion from NeuS to NeuS2 and Tensor4D format for specifying dynamic scenes.

to previous human datasets, as it requires higher generalization abilities from human-object reconstruction and rendering methods. The average ratios for human width, depth and height are 0.31, 0.19, and 0.57, respectively. For the reason that human bodies are typically long and thin, the depth ratios of humans (3D bounding box depth/image width) tend to be more concentrated in smaller proportions. These ratios indicate that our dataset presents challenges beyond scale, including appearance diversity and pose complexity.

In the 1080P sequences category, the dataset consists of 36 scenes with a resolution of 1920×1080. The duration ranges from 10 to 487 seconds. The dance sequences involve complex and highly deformable clothing, sports sequences involve interactions between players and props (such as battledore, ball, and taekwondo plank), and kungfu sequences correspond to different choreography types with corresponding costumes and props (such as Broadsword, Nanquan, Nunchuck, Taichi, *etc*.). In the 4K Studio category, the dataset includes 8 scenes with a resolution of 3840×2160. The duration of each sequence ranges from 10 to 231 seconds. This category covers both solo fashion shows and pair/group performances, as well as basic choreography dances and advanced dances originally choreographed by dancers.

3.3. Benchmark Pipeline

The wide range of shapes and textures available in PKU-DyMVHumans offers a valuable resource for training and evaluating human reconstruction and rendering of both static and dynamic scenes. By leveraging the extensive exploration space offered by PKU-DyMVHumans, we propose an off-the-shelf framework that simplifies the implementation of state-of-the-art NeRF-based methods. This includes neural scene decomposition, 3D human reconstruction, and novel view synthesis of dynamic scenes. Please refer to Fig. 4 for more implementation details.

4. Experiments

The objective of our benchmark is to achieve robust geometry reconstruction and novel view synthesis for dynamic humans under markerless and fixed multi-view camera settings, while minimizing the need for manual annotation and reducing time costs. To validate the effectiveness and potential of our dataset in neural human modeling, we conduct experiments based on different human width ratios. These experiments cover various action motions, clothing styles, and accessories. Our benchmark focuses on three perspectives, each corresponding to a specific type of application

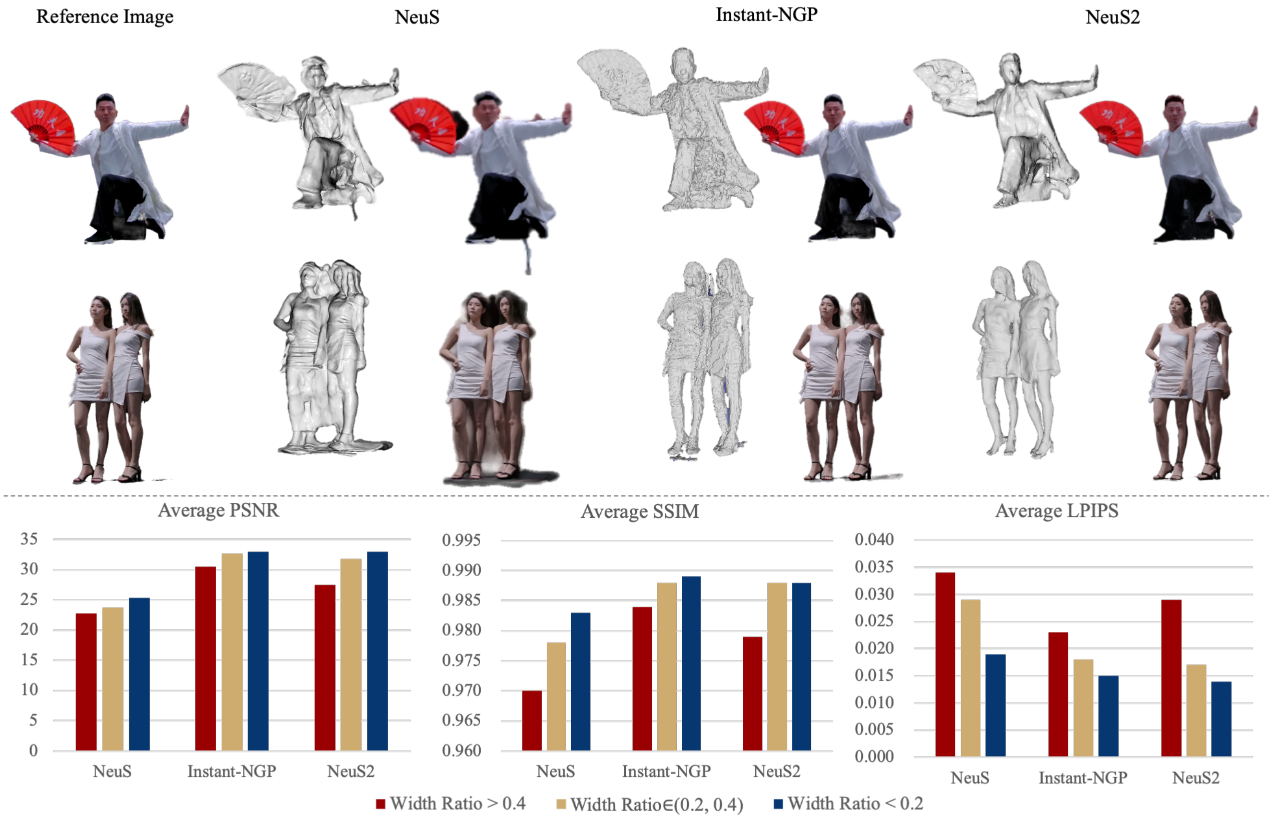

Figure 5. Comparisons on PKU-DyMVHumans dataset for static scene geometry reconstruction and novel view synthesis.

Table 2. Quantitative performance of novel view synthesis on the proposed PKU-DyMVHumans dataset.

Type	Scenes	NeuS [40]			Instant-NGP [24]			NeuS2 [41]		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Width Ratio >0.4	1080_Dance_Dunhuang_Pair_f14f15	21.38	0.959	0.042	26.37	0.974	0.035	25.34	0.967	0.044
	1080_Sport_Taekwondo1_Pair_m11c21	23.72	0.970	0.037	32.50	0.989	0.019	27.12	0.981	0.029
	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
Width Ratio $\in (0.2, 0.4)$	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_Basic_Pair_c24c25	25.48	0.979	0.024	31.96	0.989	0.017	30.99	0.986	0.019
	4K_Studios_Dance_Single_f20	22.67	0.981	0.028	30.50	0.986	0.026	31.67	0.989	0.018
	1080_Dance_Dunhuang_Single_f12	25.19	0.978	0.024	31.46	0.983	0.019	30.61	0.985	0.020
	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_Pair_f18f19	22.80	0.976	0.031	34.24	0.993	0.016	32.23	0.992	0.014
Width Ratio <0.2	1080_Sport_Football_Single_m11	24.91	0.983	0.017	29.83	0.982	0.018	30.50	0.986	0.016
	1080_Dance_Banquet_Single_c23	26.73	0.984	0.016	36.20	0.993	0.009	35.56	0.993	0.009
	1080_Kungfu_Tongbeiquan_Single_m13	23.59	0.980	0.024	32.19	0.991	0.016	31.61	0.988	0.017
	1080_Dance_Tibetan_Single_c22	26.26	0.984	0.020	33.69	0.991	0.016	34.28	0.990	0.014
Average		23.93	0.977	0.027	**32.02**	**0.987**	**0.019**	30.75	0.986	0.020

task. These perspectives include multi-view human reconstruction for novel view synthesis, dynamic human modeling, and self-supervised neural scene decomposition. All experiments are conducted on a single GeForce RTX3090 GPU. For more implementation details and additional results, please refer to the supplementary materials.

4.1. Novel View Synthesis

Implementation details. To evaluate static scene reconstruction, we provide three representative baselines of NeuS [40], Instant-NGP [24] and NeuS2 [41] with PKU-DyMVHumans. All baselines employ publicly available official implementations, and we fine-tune hyperparameters to achieve the best possible results. In the 1080P sequences, we use the same set of 52 training cameras, and 8 validation cameras. For the 4K sequences, we use a set of 48 training cameras and 8 validation cameras. The geometry reconstruction results are generated by training the released code on the training dataset with foreground object mask supervision. We then use the remaining test cameras to obtain

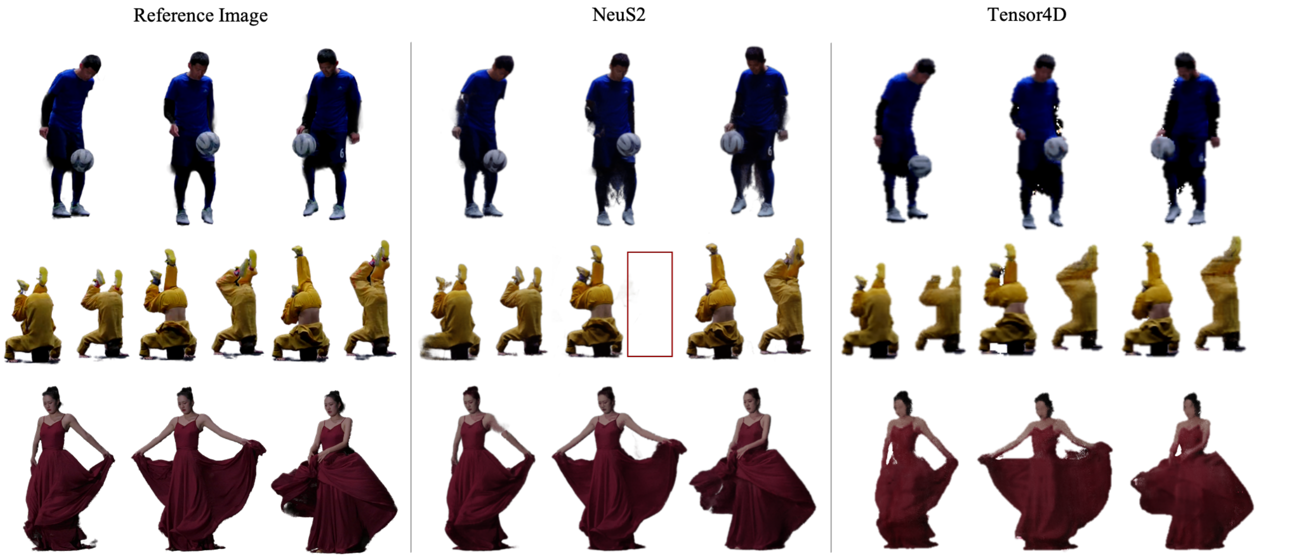

Figure 6. The results of space and time novel view rendering under short frame sequences, with dense (NeuS2 [41]) and sparse (Tensor4D [35]) camera views respectively.

novel view synthesis results and calculate metrics such as PSNR, LPIPS [51], and SSIM [42]. All metrics are computed over the whole image with white background, and the numerical results are averaged over 6 frames for all experiments. Meanwhile, the scene is divided into three sectors based on the human width ratio. Subsequently, multiple numerical analyses are performed for each sector.

Results analysis. A qualitative comparison of geometry reconstruction and novel view synthesis results for all methods is presented in Fig. 5 and supplementary materials. From the observations in Fig. 5, it is evident that NeuS introduces artifacts when encountering concave geometry (e.g., head, knees, and joints) and suffers from inaccurate density field modeling when the foreground object is dark or in shadow. The extracted meshes of Instant-NGP are noisy due to the lack of surface constraints in the geometry representation. NeuS2 exhibits limited performance with excessively smooth surfaces in the 3D geometry reconstruction. In terms of the novel view synthesis results, although NeuS2 performs better on average, it still lacks visual details and tends to produce blurry results compared to Instant-NGP, as evidenced by distinct cloth wrinkles and folding fan slats.

As shown in Tab. 2, Instant-NGP achieves the highest average performance in terms of PSNR, SSIM, and LPIPS. The best reconstruction achieves a high PSNR of 36.20, indicating that PKU-DyMVHumans contains the contents that the model can learn and fit well. However, the performance varies, and the lowest PSNR of 20.38 shows that it also includes contents that are outside of the model's learning scope and are challenging. There is a significant gap in these metrics between NeuS and the other two hash encoding-based methods, namely NeuS2 and Instant-NGP, as multi-resolution hash encoding excels at modeling high-frequency details. Additionally, the sector of $WidthRatio>0.4$ suffers from lower quality due to complex interactions between humans or humans and objects.

4.2. Dynamic Human Modeling

Implementation details. We apply the task of dynamic human reconstruction with two different input settings: 1) Dynamic reconstruction under densely captured videos: In this setting, each sequence contains 10 to 500 frames with 48 or 52 camera views for training and 8 camera views for testing. Given the multi-view videos of a moving object and camera parameters for each view, we use Instant-NGP and NeuS2 to recover per-frame reconstruction and dynamic reconstruction, respectively. 2) Dynamic reconstruction using sparse and fixed cameras: We use 16 cameras focusing on the front face, with the number of training frames ranging from 10 to 20. For this setting, we employ NeuS2 and Tensor4D to represent topologically varying objects. For quantitative evaluation, we calculate and average the PSNR, LPIPS, and SSIM scores over all frames.

Results analysis. We present example results for novel view synthesis in Fig. 6. As shown in Fig. 6, when dealing with long sequences of 10 frames that involve challenging movements or multiple foreground objects, NeuS2 struggles to accurately reconstruct dynamic objects. Similarly, for sparse-view videos, Tensor4D struggles to render high-quality images for dynamic scenes and faithfully recover appearance details such as thin finger motions, human-object interaction, facial expressions, and cloth wrinkles.

The quantitative results are summarized in Tab. 3, which include both familiar contents that the model can handle well and more challenging new contents, demonstrating the diversity of our dataset. Additionally, we observe that the

Table 3. Quantitative comparisons between dense and sparse camera views using long and short frame sequences.

Width ratio of human box	**Instant-NGP** [24] **Long sequence(Dense-view)** PSNR ↑ / SSIM ↑ / LPIPS ↓	**NeuS2** [41] **Long sequence(Dense-view)** PSNR ↑ / SSIM ↑ / LPIPS ↓	**NeuS2** [41] **Short sequence(Dense-view)** PSNR ↑ / SSIM ↑ / LPIPS ↓	**Tensor4D** [35] **Short sequence(Sparse-view)** PSNR ↑ / SSIM ↑ / LPIPS ↓
Width Ratio $>$0.4	29.53 / 0.984 / 0.025	25.87 / 0.974 / 0.032	28.62 / 0.980 / 0.025	27.59 / 0.970 / 0.040
Width Ratio $\in (0.2, 0.4)$	30.55 / 0.983 / 0.022	30.05 / 0.985 / 0.021	32.04 / 0.989 / 0.017	31.56 / 0.982 / 0.028
Width Ratio $<$0.2	29.39 / 0.984 / 0.022	30.52 / 0.987 / 0.018	32.93 / 0.989 / 0.015	28.89 / 0.986 / 0.022
Average	29.82 / 0.984 / 0.023	28.81 / 0.982 / 0.024	**31.20 / 0.986 / 0.019**	29.38 / 0.979 / 0.030

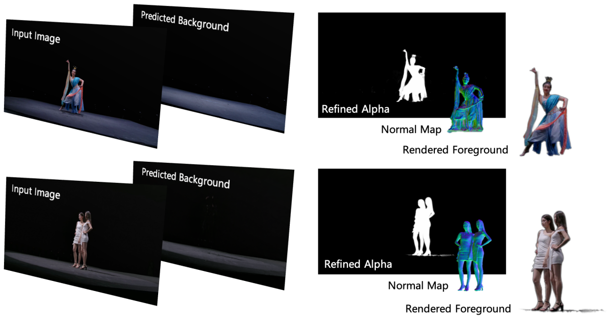

Figure 7. Given multi-view images as input, the goal is to decompose 3D scenes into geometrically consistent foreground objects, texture-completed backgrounds, and generate convincing segmentation and normal maps for them.

diverse motions, appearances, long sequences, and loose garments in PKU-DyMVHumans pose further challenges for 4D neural human rendering.

4.3. Neural Scene Decomposition

Neural scene decomposition aims to effectively separate the foreground and background without any annotations, with downstream applications in multi-view object segmentation [4] and beyond, such as 3D/4D human reconstruction and rendering. Under conditions of multi-camera inputs, the structural, textural and geometrical consistency among each view can be leveraged to achieve fine-grained object segmentation. Surface-SOS [53] addresses this information by combining the neural surface representation from multi-view observations via volume rendering of SDF. It captures the compositional nature of scenes and provides additional inherent information, improving 3D human reconstruction.

We train Surface-SOS [53] for each individual sequence and present some example results in Fig. 7. It shows that Surface-SOS can generates detailed foreground alpha, resulting in high-quality geometry surfaces. This is evidenced by the distinct cloth wrinkles and natural shading. The quantitative results in Tab. 4 also highlights the improvement of the Surface-SOS refinement. The detailed procedures and more comparison result against relevant methods of PKU-DyMVHumans are further illustrated in the supplementary material. It shows that when dealing with scenes involving human-object interactions, multi-person interactions, as well as challenging complex scene effects, it faces even greater challenges.

Table 4. Quantitative evaluation of the foreground segmentation on PKU-DyMVHumans dataset.

Width ratio of human box	**BGMv2** [19] MSE↓ / mIoU↑ / Acc.↑	**Surface-SOS** [53] MSE↓ / mIoU↑ / Acc.↑
Width Ratio $>$0.4	0.345 / 0.853 / 0.925	0.226 / 0.871 / 0.943
Width Ratio $\in (0.2, 0.4)$	0.482 / 0.891 / 0.932	0.219 / 0.936 / 0.970
Width Ratio $<$0.2	0.439 / 0.900 / 0.958	0.228 / 0.917 / 0.973
Average	0.422 / 0.881 / 0.938	**0.224 / 0.908 / 0.962**

5. Conclusion

In this work, we introduced PKU-DyMVHumans, a dynamic human dataset designed for high-fidelity human reconstruction and rendering from dense multi-view videos. It features high-fidelity human performance, including high-detailed appearance, complex human motion, as well as challenging human-object interactions, multi-person interactions and complex scene effects (*e.g.*, lighting, shadows, and smoking). We further presented benchmark tasks, with detailed experiments on several advanced methods. PKU-DyMVHumans further fill in the gap between existing datasets and real-scene applications.

Challenges and future works. While we have validated the complexity and fidelity of our dataset on numerous human-centric reconstruction and rendering scenarios. It is significant to highlight the more challenging and realistic multiple-person/subject modeling that could reflect the rendering differences with respect to multi-person interactivity, complex scene effects, and multi-view consistent performance. Additionally, free-viewpoint rendering of a moving subject from a monocular self-rotating video is a complex yet desirable setup. Our supplementary material provides additional experiments for free-viewpoint rendering of moving subjects, the results are affected by local occlusion and view absence, leading to artifacts in view rendering. With these opportunities and challenges, we believe PKU-DyMVHumans will benefit the development of new approaches in the community.

Acknowledgments. This work is supported by Outstanding Talents Training Fund in Shenzhen, Shenzhen Science and Technology Program (RCJC20200714114435057, SGDX20211123144400001), National Natural Science Foundation of China (U21B2012), and Migu-PKU Meta Vision Technology Innovation Laboratory (R24115SG). Jianbo Jiao is supported by the Royal Society grants IES\R3\223050 and SIF\R1\231009.

References

[1] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. 3

[2] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. 3

[3] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 2

[4] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, and Patrick Pérez. Multi-view object segmentation in space and time. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2647, 2013. 8

[5] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 2

[6] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 453–464. 2023. 2

[7] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 2

[8] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 2

[9] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4):1–16, 2021. 2, 3

[10] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 3

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2, 3

[12] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1261, 2020. 3

[13] Hao Li, Linjie Luo, Daniel Vlasic, Pieter Peers, Jovan Popović, Mark Pauly, and Szymon Rusinkiewicz. Temporally coherent completion of dynamic shapes. *ACM Transactions on Graphics (TOG)*, 31(1):1–11, 2012. 2

[14] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 49–67. Springer, 2020. 2

[15] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 2

[16] 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 Recognition*, pages 5521–5531, 2022. 3

[17] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 3

[18] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 3

[19] 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. 4, 5, 8

[20] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 3

[21] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 3

[22] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 3

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 4

[24] 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. 3, 4, 6, 8

[25] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[26] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 3

[27] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022. 3

[28] Rui Peng, Xiaodong Gu, Luyang Tang, Shihe Shen, Fanqi Yu, and Ronggang Wang. Gens: Generalizable neural surface reconstruction from multi-view images. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[29] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[30] 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. 2, 3

[31] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

[32] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[33] 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. 4, 5

[34] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *European Conference on Computer Vision*, pages 702–720. Springer, 2022. 2, 3

[35] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 3, 4, 7, 8

[36] Sahil Sharma and Vijay Kumar. Voxel-based 3d face reconstruction and its application to face recognition using sequential deep learning. *Multimedia tools and applications*, 79: 17303–17330, 2020. 3

[37] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, 2019. 3

[38] Cheng W. et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *ICCV*, 2023. 2, 3

[39] Jiong Wang, Fengyu Yang, Wenbo Gou, Bingliang Li, Danqi Yan, Ailing Zeng, Yijun Gao, Junle Wang, and Ruimao Zhang. Freeman: Towards benchmarking 3d human pose estimation in the wild. *arXiv preprint arXiv:2309.05073*, 2023. 3

[40] 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. 3, 4, 6

[41] 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. 3, 6, 7, 8

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[43] 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

[44] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 3

[45] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 3

[46] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023. 2, 3

[47] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5746–5756, 2021. 2

[48] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021. 2, 3

[49] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2990–3000, 2020. 3

[50] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021. 2

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[52] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2

[53] 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. 4, 8

[54] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021. 3

[55] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 3

[56] Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. Relightable neural human assets from multi-view gradient illuminations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4327, 2023. 3