This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering

Wei Cheng[1] Ruixiang Chen[2*] Siming Fan[1,2*] Wanqi Yin[2*] Keyu Chen[1*]
Zhongang Cai[3] Jingbo Wang[4] Yang Gao[2]
Zhengming Yu[1] Zhengyu Lin[2] Daxuan Ren[3]
Lei Yang[1,2] Ziwei Liu[3] Chen Change Loy[3] Chen Qian[1]
Wayne Wu[1] Dahua Lin[1,4] Bo Dai[1†] Kwan-Yee Lin[1,4†]
[1] Shanghai AI Laboratory [2] SenseTime Research [3] S-Lab, NTU [4] CUHK

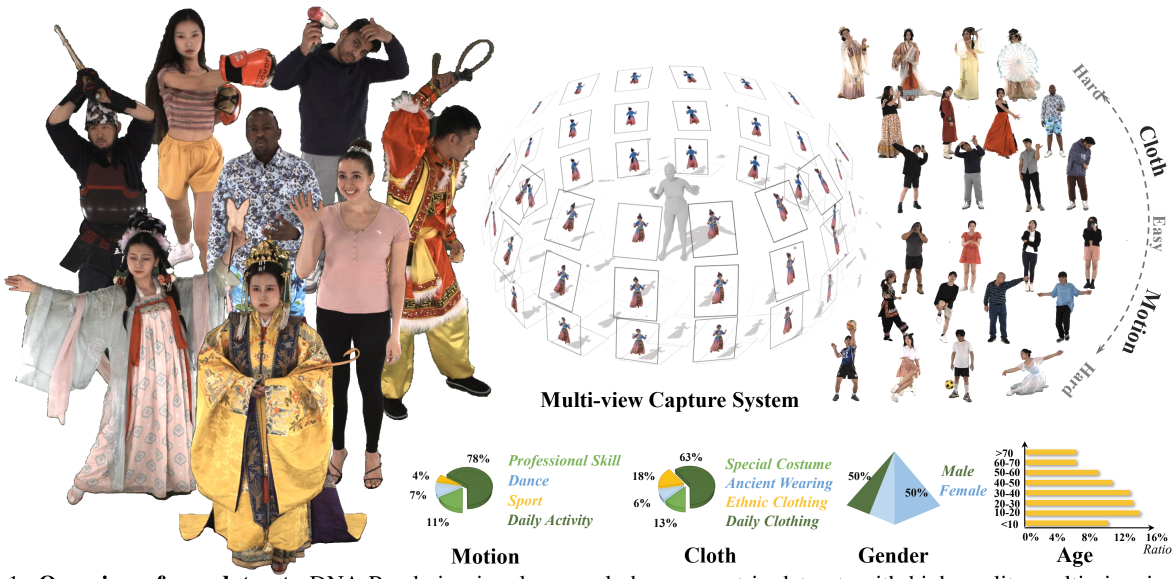


Figure 1: **Overview of our dataset**. DNA-Rendering is a large-scale human-centric dataset, with high-quality multi-view images and videos for various human actors. The dataset comes with grand categories of motion, cloth, accessory, body shape, and human-object interaction. We hope it could boost the development of human-centric rendering and related tasks.

Abstract

Realistic human-centric rendering plays a key role in both computer vision and computer graphics. Rapid progress has been made in the algorithm aspect over the years, yet existing human-centric rendering datasets and benchmarks are rather impoverished in terms of diversity (e.g., outfit's fabric/material, body's interaction with objects, and motion sequences), which are crucial for rendering effect. Researchers are usually constrained to explore and evaluate a small set of rendering problems on current datasets, while real-world applications require methods to be robust across different scenarios. In this work, we present ***DNA-Rendering****, a large-scale, high-fidelity repository of human performance data for neural actor rendering. DNA-Rendering presents several appealing attributes. First, our dataset contains over* 1500 *human subjects,* 5000 *motion sequences, and* 67.5*M frames' data volume. Upon the massive collections, we provide human subjects with grand categories of pose actions, body shapes, clothing, accessories, hairdos, and object intersection, which ranges the geometry and appearance variances from everyday life to professional occasions. Second, we provide rich assets for each subject – 2D/3D human body keypoints, foreground masks, SMPLX models, cloth/accessory materials, multi-view images, and videos. These assets boost the current method's accuracy on downstream rendering tasks. Third, we construct a professional multi-view system to capture data, which contains* 60 *synchronous cameras with max* 4096×3000 *resolution,* 15 *fps speed, and stern camera calibration steps, ensuring high-quality resources for task training and evaluation.*

*Joint-first authors with W. Cheng.
†Equal advising.

Dataset	Attribute					Scale				Realism
	Ethnicity	Age	Cloth	Motion	Interactivity	#ID × #Outfit	#Motions	#View	#Frames	HRes
Human3.6M [19]	✗	✗	✗	✓	✓	11×1	17	4	3.6M	1000P
CMU Panoptic [24]	✓	✓	✗	✓	✓	97×1	65	$31 + 480^*$	15.3M	1080P
ZJU-MoCap [44]	✗	✗	✗	✓	✗	10×1	10	24	180K	1024P
HUMBI [66]	✓	✓	✓	✗	✗	772×1	–	107	26M	1080P
AIST++ [56, 28]	✗	✗	✗	✗	✗	30×1	–	9	10.1M	1080P
THuman 4.0 [51]	✗	✗	✓	✓	✗	3×1	–	24	10K	1150P
HuMMan [5]	✓	✓	✓	✓	✗	1000×1	500	10	60M	1080P
GeneBody [10]	✓	✓	✓	✓	✓	50×2	61	48	2.95M	2048P
DNA-Rendering (Ours)	✓	✓	✓	✓	✓	500×3	1187	60	67.5M	4096P

Table 1: **Dataset comparison on attributes and scales.** We compare the proposed dataset with previous human-centric multiview datasets in terms of attribute coverage, scale, and realism. 'Ethnicity' denotes whether the dataset contains actors from multiple ethnic groups. 'Age' means if there is a wide age range containing elders or infants. 'Cloth' separates datasets with only daily costumes or with extra diverse clothing. 'Attribute-Motion' denotes whether it has human motion in different scenarios. 'Interactivity' tells whether there contains human-object interaction. We mark these attributes with ✓ and ✗. In scale, we list the number of key factors with compared dataset, Note that 'Scale-#Motions' means the number of motion categories, and superscript * means low-resolution VGA cameras, we exclude them during '#View' ranking and '#Frames' calculation. We abbreviate resolution at height as 'HRes'.

Along with the dataset, we provide a large-scale and quantitative benchmark in full-scale, with multiple tasks to evaluate the existing progress of novel view synthesis, novel pose animation synthesis, and novel identity rendering methods. In this manuscript, we describe our DNA-Rendering effort as a revealing of new observations, challenges, and future directions to human-centric rendering. The dataset, code, and benchmarks will be publicly available at https://dna-rendering.github.io/.

1. Introduction

Understanding humans is an everlasting problem in our research community, and extensive literature on perceiving and synthesizing humans shows great efforts toward this goal. Over the decades, many pioneers have constructed large-scale and diverse datasets, such as COCO [31] for human pose estimation, and ActivityNet [4] for analyzing human action. These datasets have been pivotal in advancing the development of human-centric perception algorithms.

Yet, when it comes to human-centric rendering, there is still a noticeable gap in comprehensive datasets. Capturing high-quality and massive 3D/4D human avatars is difficult due to the requirements of high-end equipment as well as an efficient data processing pipeline. Existing datasets [19, 23, 44, 51, 18] partially narrow the gaps but have significant limitations on sample diversity (*e.g.,* clothing, motion, body shape, and human-object interaction), or have insufficient realism (*e.g.,* camera resolution, and capture speed). These factors are crucial to rendering effects.

To drive advance in human-centric rendering, we contribute a large-scale multi-view human performance capture dataset, named DNA-Rendering, which includes the factors that are important to rendering in great diversity and granularity. *On the hardware side*, we build a 360-degree indoor system equipped with 60 calibrated RGB cameras and 8 synchronized depth sensors. The captured videos are under the fidelity of up to 12MP (4096×3000) resolution and recorded at 15 fps. *From the dataset's footage design aspect*, we intend to cover most attributes that could reflect the rendering differences with respect to texture, materials, primary/secondary motion deformation, and category priors. In particular, we design over 1500 outfits and 1187 motion types to ensure the comprehensive coverage of real-world scenarios. We invite 500 actors to participate in the data capture process. We record each person with three different outfits and at least nine unique motions. The full dataset contains 5000 video sequences with over $67.5M$ frames. Compared with the existing human-centric dataset like CMU Panoptic [24], ZJU-MoCap [44], THUman [51], and Human3.6M [19], DNA-Rendering comprises the most multi-view body performance samples and reaches the highest image quality. The unfold comparisons between DNA-Rendering and the others are given in Tab. 1.

Meanwhile, we provide essential annotations attached to each frame to facilitate the application of downstream tasks. We develop an automatic annotation pipeline encompassing camera calibration, color correction, image matting, 2D /3D landmark estimation, and SMPLX model fitting. To ensure the labeling quality, we developed a series of technical refinements to the annotation toolchain. With these efforts, the automatic pipeline can generate faithful data annotations both effectively and efficiently.

The unprecedented richness of DNA-Rendering dataset provides fertile data soil for researchers to develop, and dissect their rendering methods in depth. To set up a kickoff example, we further construct benchmarks upon the dataset with extensive experiments. We evaluate the performances of several state-of-the-art full-body rendering and animation approaches under three major tasks, *i.e., novel view synthesis*, *novel pose animation*, and *novel identity rendering*. To better analyze current methods in terms of the model capacity, module necessity, and methodology generality, we set up multiple test set splits under different levels of challenging aspects. For instance, we divide the *easy, medium,* and *hard* subsets *w.r.t.* the cloth looseness, the texture complexity, the motion difficulty, and the human-object interactivity, respectively. We conclude a series of key observations based on the benchmarks, such as how human prior influences the robustness of rendering, how sensitive the multi-view/frame relationship module design is to data volume/distribution, and how loss design affects the performance in terms of different rendering metrics.

In summary, the DNA-Rendering project fulfills the requirement of a high-fidelity human performance capture dataset for the research community. We establish by far the largest multi-view human body performance dataset for high-fidelity human-centric rendering research, with an emphasis on image quality and data attributes. The attached benchmarks provide baseline standards for three major tasks, with rigorous evaluations and dissections on multiple state-of-the-art methods. We believe the dataset, the attached benchmarks, and the tools will boost a wide range of digital human applications and inspire future research.

2. Related Work

2.1. Human-centric Datasets

Perception Datasets. Perceiving human is a long-standing problem. Over the decades, researchers have kept dedicating their efforts to building relevant datasets. Earlier efforts in the computer vision community present large-scale datasets like COCO [31] for human segmentation or keypoint detection from in-the-wild images. Later research works follow the inspiration to establish open-world datasets [8, 16], while with emphasis on parsing more precise human body parts. Some researchers focus on constructing datasets [7, 40, 50] that capture daily activities and help the perception of human action recognition by using RGB-D cameras. Despite the wild variety of data samples, these datasets are not capable of human rendering tasks, due to a lack of multiview images as groundtruth references for evaluating methods' performance. In computer graphics society, there is another parallel branch that contributes datasets [27, 25] for avatar animation, with recording human motion via maker-based motion capture systems. AMASS [35] further integrates these motion capture databases with fully rigged surface mesh representation. In the last decade, computer vision and graphics society fit in with each other in the field of perceiving 3D humans. Datasets like Human3.6M [19] and MPI-INF-3DHP [36] capture humans under in-door multi-view environment with 3D marker label or multiview segmentation, which further encourage the applications in recovering human in 3D. These databases facilitate the development of numerous algorithms. However, due to the limits of the data sample, camera views, and resolution, they cannot well reflect the pros and cons of rendering methods.

Rendering Datasets. Representing 3D/4D human appearances and performances are important in both research communities and commercial applications. THuman [71, 64, 54] and commercial scan datasets [15, 14] capture static human scan reconstructed by either depth sensors or camera array. [2, 3, 67, 51, 69] provide dynamic human scans with minimal clothing and daily costumes. These datasets are usually biased centering on standing poses due to the sophisticated capture process. With the emergence of neural rendering techniques, rendering realistic humans directly from images has become a trend. CMU Panoptic [23] uses a 30-HD-camera system and annotates the humans with 3D keypoints. HUMBI [66] focuses on local motions like gestures, facial expressions, and gaze movements. ZJU-MoCap [44] is a widely used dataset for human rendering algorithms but with limited motion and clothing diversity, which might lead the evaluation to great bias. AIST++ [56, 28] is a dance database with various dance motions while sticking in the one-fold scenario and lacking view density. Recently proposed HuMMan [5] and GeneBody [10] datasets, expand the motion and clothing diversity, while the effective human resolution is still below 1K. Concurrent works [72, 18] also contribute datasets for human avatar tasks, while centering on detailed human geometry with long-lens cameras to film human body parts.

2.2. Implicit Neural Body Representation

Different from previous works that represent humans with explicit representations, recent work models human appearance as neural implicit function, *e.g.*, neural radiance fields [38] or neural signed distant functions. PIFu [47, 48] presents the orthogonal camera space as an occupancy function conditioned by pixel-aligned features and depth. NeuralBody [44] learns a neural radiance field of dynamic humans conditioned by body structure and temporal latent code from sparse multi-view videos. Recently, many category-agnostic implicit representations, PixelNeRF [63], IBRNet [59], VisionNeRF [29], *etc.*, can generalize NeRF to arbitrary unseen scenes given a set of reference views. The intrinsic differences among these methods are the design of feature aggregation, which varies from average [63], max pooling [45] to more adaptive weighted pooling [59] and vision transformer [29]. Given human rendering is more challenging due to the large variation in pose and appearance, recent generalizable human rendering methods [68, 37, 10, 26] condition such image feature-aligned NeRF with human priors. For example, NeuralHumanPerformer [26] uses structured latent code and KeypointNeRF [37] deploys human keypoints.

2.3. Animatable Digital Human

The challenge of creating realistic animatable human avatars from images is two folds – (1) how to reconstruct the human body from motion sequences and (2) how to disentangle non-rigid deformation. Early seminal work A-NeRF [53] learns dynamic body from sequences, it conditions the radiance field with relative pose coordinate of the query point, which fails to model the non-rigidity of clothed humans. To reconstruct the human body from sequences, AnimatableNeRF [43] learns a static canonical radiance field together with a ray blending network from the current frame to canonical space. To further better disentangle motion deformation from pose recent works [49, 9] use a complementary forward blending network or root-finding algorithm to regularize the learned blending with cycle consistency loss. Other works [61, 20, 65] learn animatable models from more challenging monocular video, with a tighter assumption of Gaussian distributed occupancy along bone or fixed SMPL motion weights.

3. DNA-Rendering

3.1. Dataset Capture

System Setup. Our capture system contains a high-fidelity camera array, with 60 high-resolution RGB cameras and 16 lighting boards uniformly distributed in a sphere with a radius of three meters. The cameras are adjusted to point at the sphere's center, where the participants perform. Concretely, the array consists of 48 high-end 2448×2048 industrial cameras, and 12 ultra-high resolution cameras with up to 4096×3000 resolution. We additionally place eight Kinect cameras to capture additional depth streams as auxiliary geometric data. The high-fidelity video streams and depth streams are synchronized at 15 frames per second. The above designs ensure the system could record the sharp texture edges, fine-grained color changes of clothing patterns, and the reflection effects caused by different clothing materials. Please refer to Sec. A.3 [1]for more details.

Data Collection Protocol. To enable subsequent research probing into the factors that have influences on rendering, we design a data collection protocol with both interlaced and hierarchical data attributes. Specifically, we ask each actor to wear three sets of outfits and perform at least three actions in different hallucinated scenarios for each outfit, which maximize the identity scale and diversity. Each motion sequence is recorded under specific action category instruction with a free-style performance lasting for 15 seconds, which ensures the diversity of action performance. As an auxiliary feature, we also capture a static frame of A-pose for actors in each outfit for canonical pose recording, and a frame with only empty background for image matting. For accurate camera pose annotation, extrinsic calibration data are collected at a daily frequency. The color data and

[1]If not specified, the indexes with the combination of a capital letter and an Arabic numeral refer to the corresponding sections/figures/tables listed in the supplementary material.

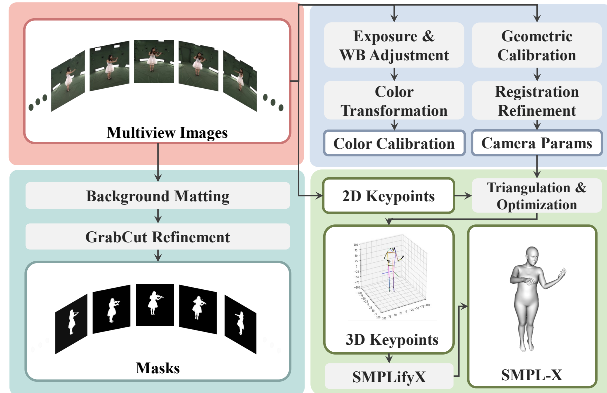


Figure 2: **Annotation pipeline.** The illustration of annotation pipeline for camera calibration, camera color calibration, masks, keypoints, and parametric model.

intrinsic calibration data are collected whenever system adjustments are made. Please refer to Sec. A.4 for details.

3.2. Dataset Statistics

In order to cover diverse attributes that relate to rendering quality, we have carried out a detailed design from the selection of the actors' gender, age, and skin color, to their actions, clothing, and makeup. The key statistics of our dataset are shown at the bottom of Fig. 1. Specifically, to preserve authenticity in action behavior, we invite 153 professional actors to perform special scenes with corresponding costumes/makeup, and 347 normal performers to act under footage of daily-life scenes. The special scenes constitute 153 sub-categories, including sports, dances, and unique event performances such as typical costumes in ancient Chinese dynasties, traditional costumes around the world, cosplay, *etc*. Common scenes can be divided into 269 sub-categories, covering scenes such as daily indoor activities, communication, entertainment, and new trends. We describe the comprehensive distribution of data in Sec. A.1 and the limitation of data in Sec. A.5.

3.3. Data Annotation

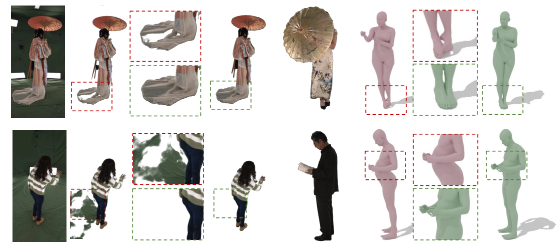
To enable applications in human rendering and animation, DNA-Rendering provides rich annotations attached with the raw data, *i.e.,* camera calibration, camera color calibration, image matting, and parametric model fitting. The overall annotation pipeline is shown in Fig. 2.

Camera Calibration. First, we calibrate the intrinsic parameters of each camera individually. Specifically, we divide the camera's field of view into a 3×3 Sudoku, and capture images with ± 30 degree rotation in pitch, row, and yaw angle of checkerboard in all grids, referring to Fig. S4. Second, for extrinsic calibration, we deploy multiple ChArUco boards and spin the main board in the capture volume. We use open toolboxes [12, 1] to optimize intrinsic parameters, distortion coefficients, and extrinsic parameters with the captured data. To eliminate the depth camera pose error caused by the large resolution gap between industrial cameras and Kinect depth cameras, we further adopt a point cloud registration stage to refine the depth camera extrinsic parameters in the second stage. More concretely, for each depth camera, we project the partial point cloud and estimate a full point cloud from the MVS algorithm such as [57] as a reference. We jointly optimize the pose graph of the depth camera for neighboring pointclouds with overlaps through a multi-way registration [11] with MVS pointcloud as reference. For detailed camera calibration, please refer to Sec. B.1 in the supplementary.

Color Calibration. The identical color response across different cameras could be vital for a multi-view, mixed-type camera system to provide qualified data for rendering applications, as it is an essential data basis for algorithms to render realistic view-dependent effects. Different from other multi-camera datasets, *e.g.*, Multiface [62, 33] which uses a network to optimize the color transformation during model training, we pay attention to ensure the color consistency of data collection across different cameras. First, we conduct careful adjustments on hardware parameters such as exposure and white balance to make the captured color of the color checkerboard under the standard light as close as possible. Then, the 2-order polynomial correction coefficients could be optimized by least square regression of transforming the detected color to the true value on the color checkerboard. Please refer to Sec. B.1 and Fig. S5 for details. We also analyze the impact of color consistency of multi-camera datasets on generalizable rendering in Sec. D.4.

Matting. Considering the large quantities of the captured images, we develop an automatic matting pipeline to extract the foreground objects. We first adopt an off-the-shelf background matting model [30] to eliminate most background pixels. However, due to the complicated nature of the capture settings, the learning-based model inevitably generates unsatisfying results in some challenging cases, leaving some pieces of labeled data with artifacts such as broken holes or noisy patches (Fig. 3). Thus, we further propose a refinement strategy by applying *HSV* filtering and the GrabCut [46] algorithm to improve the matting quality. We compare matting with and without refinement, and visualize detailed manual assessment in Fig. S8.

Keypoints and Parametric Model. Inspired by existing works [5, 6], we develop an automatic pipeline to annotate keypoints and parametric model parameters. 1) First, 2D keypoints in COCO-Wholebody [21] format (including body, hand, and face keypoints) are detected for each camera view, with pretrained model HRNet-w48 [55]. 2) Then, we triangulate 3D keypoints with known camera intrinsic and extrinsic parameters from the multi-view 2D keypoints with optimization and post-processing strategies [13] including keypoint selection, bone length constraint, as well as outlier removal. 3) Finally, we register the SMPLX, a commonly used parametric model, via 3D keypoints. Body shape $\beta \in \mathbb{R}^{n\times 10}$ (or $\beta \in \mathbb{R}^{n\times 11}$ for children [17, 41]), pose parameters (body pose, hand pose, and global orientation) $\theta \in \mathbb{R}^{n\times 156}$, and translation parameters $t \in \mathbb{R}^{n\times 3}$ (n is the number of frames) are estimated via a modified SMPLify-X [42] for dynamic poses.



(a) (b)

Figure 3: **Annotation quality improvements.** The zoom-in boxes with red show the annotation quality before pipeline optimization. The green ones show quality improvements over (a) mask annotation and (b) SMPLX annotation with the optimized pipeline.

Our annotation pipeline is proved effective and robust in getting natural SMPLX model, as shown in Fig. 3. We evaluate the fitting error between 3D keypoints and corresponding regressed SMPLX joints. The mean and median 'Mean Per Joint Position Error' (MPJPE) of our system is 30.20 mm and 29.80 mm. The error is on par with the oracle fitting accuracy of 29.34 mm in Human3.6M [19, 34], which includes data from an optical motion capture system. Detailed analysis is conducted in Sec. B.3 and Sec. B.4. A thorough comparison of our fitting pipeline with other fitting methods [52, 70, 10] is described in Sec. B.5.

4. Benchmarking Human-centric Rendering

DNA-Rendering dataset could be used to boot the developments of research on high-fidelity human body rendering tasks, due to its *large-scale volume, diverse scenarios, multi-level challenges, and high-resolution* properties. To kick off an example of how to utilize this dataset, we set up benchmarks with exclusive experiments centered around three fundamental tasks of human body rendering.

4.1. Data Splits

To unfold each method in depth, and thoroughly evaluate the effectiveness of our dataset, we construct multiple training and testing data splits to conduct level tests for each method. We consider the four most influential factors of rendering quality for the benchmark test, *i.e.,* looseness of clothes, texture complexity, pose difficulty, and interactivity between the human body and manipulated object.

The Cloth Looseness. We define the cloth's challenging levels by the deformation distance between the minimal-cloth human body and the clothing outline, and the softness of cloth materials. The *Easy* level covers cases wearing tight-fitting clothes like yoga wear and sports t-shirts. The *Medium* level includes the daily clothes such as coats, skirts, jeans, loose t-shirts, *etc.* The *Hard* level contains ethical costumes, national clothing, and fancy decorations.

The Texture Complexity. The texture distribution also plays an important role in the dynamic human body rendering tasks. To examine the correlations between texture complexity and rendering performance, we build three data splits for texture evaluation. The *Texture-Easy* split is composed of single-color clothes. The *Texture-Medium* split includes most daily clothes in a few colors and plain patterns. The *Texture-Hard* split contains the most complicated texture clothes with intricate patterns like dots, stripes, *etc.*

The Pose Difficulty. In the novel pose animation task, it is vital to probe if the trained models could handle different levels of motion sequences in terms of difficulties and degree of out-of-distribution. Therefore, we split three levels to pose difficulties. The *Easy* data are simple motions with limited body parts involved, like shaking and waving hands. The *Medium* level refers to casual motions including full-body actions such as walking, eating, sitting, kneeling, stretching, *etc*. Moreover, the *Hard* split is designed to cover the extremely challenging motion cases that are performed by professional sports players or actors, *e.g.,* instrument playing, sports action, yoga, and dancing.

The Human-Object Interactivity. We propose to evaluate the impact of human-object interactivity by object size and non-rigidity. The *Interaction-No* split contains pure human motions with no interactive objects; the *Interaction-Easy* split includes rigid small-size hand-held objects like cellphones, pencils, cigarettes, and cups. The *Interaction-Medium* split has middle-size hand-held objects, *e.g.*, handbag, volleyball, newspaper, *etc*. This split includes both rigid motions and non-rigid object motions; and the *Interaction-Hard* split consists of large-size assets such as yoga mats, desks, chairs, and sofas.

To sum up, we construct an overall train split consisting of 400 sequences with even distribution on all human factors and difficulties, and 13 test factor-difficulty splits in total with three sequences in each test split.

4.2. Task Definition

Depending on the generalizability of the state-of-the-art methods, we categorize the recently published works into two classes: case-specific methods and generalizable ones. We evaluate the methods under multiple problem settings according to their categories. Concretely, we set up *novel view synthesis* and *novel pose animation* tasks for the case-specific methods, and *the novel identity rendering task* for the generalization approaches. In this section, we present the key observations of the benchmarks.

Novel View Synthesis. Recent *dynamic* human rendering works like NeuralBody [44], A-NeRF [53], AnimatableNeRF [43], and NeuralVolumes [32] obtained impressive results by training on a single case with multi-view

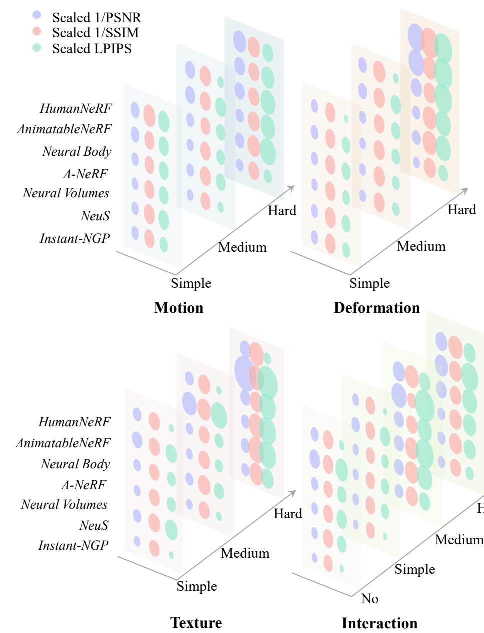


Figure 4: **Quantitative results visualization of novel view synthesis test across benchmarks splits and difficulties.** The colored circles denote different metrics, the smaller the circles indicate the better the novel view quality the method achieved. The numbers are reported in the appendix (Tab. S2).

video data. HumanNeRF [61] demonstrated the ability to render realistic novel view images of humans from monocular video sequences. In this task, we adopt the official implementation of the case-specific methods and train each individual model for every single case in the DNA-Rendering test set. For a fair comparison, we unify the training setting of NeuralVolumes [32], A-NeRF [53], NeuralBody [44], AnimatableNeRF [43], and HumanNeRF [61] with 42 dense training views. We evaluate the image rendering quality of these methods on the other 18 unseen testing camera poses. Meanwhile, we also train two general scene *static* methods – Instant-NGP [39] and NeuS [58], in each testing frame with the same training views. These two methods' performances could serve as the per-frame static reconstruction baseline reference. The rendering results are analyzed based on the difficulty level of data splits.

Novel Pose Animation. Similar to the novel view synthesis task, we conduct novel pose animation benchmark on the four dynamic methods [44, 43, 53, 61]. For each test case, we split the sequence into two parts, where images from the first 80% frames are used for training and the ones from the last 20% are used for testing. Besides, for the SMPL-guided pose animation methods [44, 43, 61], we provide the *SMPL parameters* of test images for the models to infer rendering. As for the SMPL-free method [53], the trained models take the target *pose images* as the input (*i.e.,* the underlying skeletons), and render humans in novel poses.

Novel Identity Rendering. The other category of our

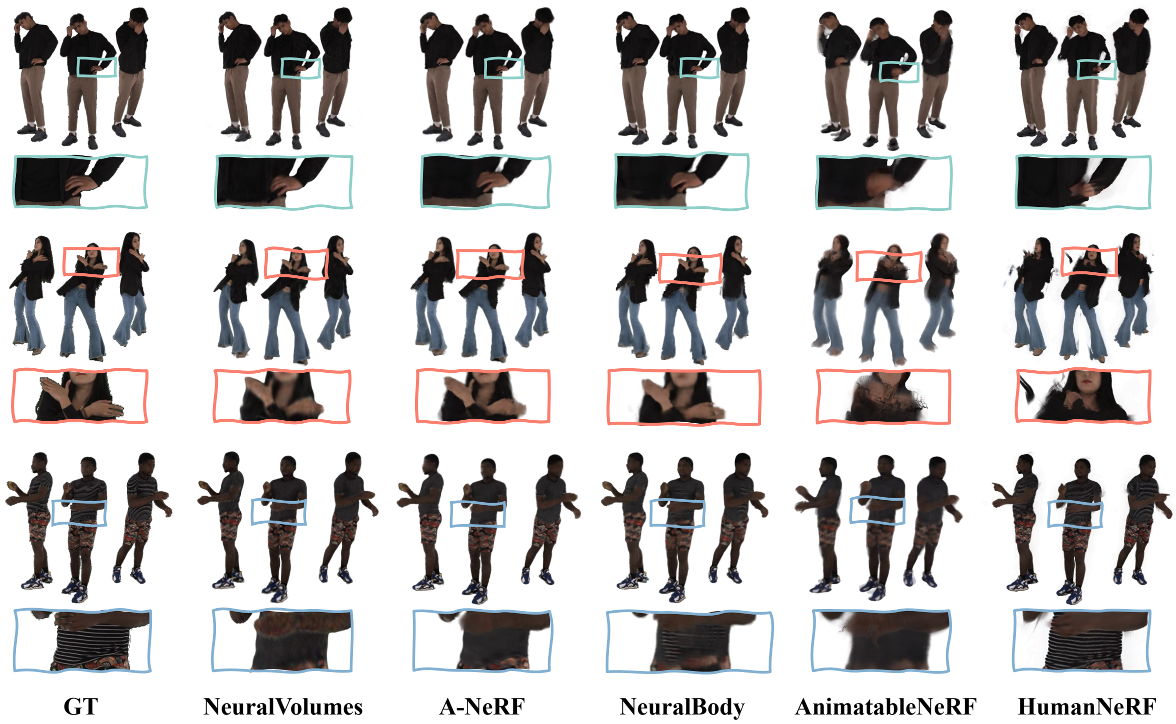


Figure 5: **Visualization of novel view synthesis result samples.** We qualitatively compare the novel views of the test frames in our test splits. More qualitative novel results in each split are shown in Fig. S9.

benchmark methods is the generalizable algorithms that can be trained on multiple cases and infer across different unseen identities. Specifically, we probe three general scene methods – PixelNeRF [63], VisionNeRF [29], IBRNet [59], and two human-centric methods – NeuralHumanPerformer [26], and KeypointNeRF [37]. To fairly compare their performances on unseen identities, we use the same training set (all training samples of the three splits, which results in 400 sequences in total) to train the generalizable models. In the inference stage, we evaluate the rendering quality on novel cases from each test split respectively.

4.3. Benchmark Results

As introduced in Sec. 4.1, we construct a test set with 13 sub-splits according to the four most concerned attributes in different difficulty levels, and an extra *No* level for *Interaction*. This results in a data volume of 39 motion sequences for testing. For all rendered images, three metrics are computed – PSNR, SSIM [60], and LPIPS-Alex [22] (LPIPS* denotes LPIPS$\times$1000). We evaluate more than 10 state-of-the-art methods on these splits and analyze their performances under the same metrics. The experiment analysis is given below. Noted that due to limited space in the main paper, we provide the detailed setting, thorough discussions, and additional results in Sec. C in the supplementary.

Novel View Synthesis. We visualize the bubble diagram of quantitative results across all benchmark splits in Fig. 4. The precise numbers of the quantitative results are reported in Tab. S2 in the supplementary. We conclude three key observations in the main paper: (1) Generally speaking,

Splits	PSNR↑					SSIM↑					LPIPS*↓				
	NV	AN	NB	AnN	HN	NV	AN	NB	AnN	HN	NV	AN	NB	AnN	HN
Motion-Simple	22.05	26.65	25.84	22.78	24.65	0.947	0.965	0.974	0.958	0.953	78.30	58.04	58.33	74.33	62.76
Motion-Medium	19.30	21.73	21.84	21.41	21.14	0.941	0.951	0.969	0.957	0.952	92.80	71.81	65.46	80.97	54.46
Motion-Hard	19.17	21.49	20.43	19.64	22.48	0.938	0.952	0.965	0.949	0.964	105.46	83.58	82.85	97.98	51.18
Deformation Simple	20.42	25.44	24.57	23.62	26.15	0.939	0.957	0.968	0.958	0.967	84.65	53.30	59.04	61.12	30.18
Deformation-Medium	23.09	27.26	27.05	23.52	24.97	0.945	0.963	0.974	0.961	0.958	61.09	48.41	49.91	65.43	33.94
Deformation-Hard	20.11	20.88	20.27	19.41	19.70	0.925	0.926	0.956	0.943	0.924	117.31	108.89	102.84	103.22	102.67
Texture-Simple	20.99	26.21	25.54	23.12	25.65	0.954	0.974	0.982	0.970	0.974	77.68	49.88	50.48	67.40	28.81
Texture-Medium	25.44	27.94	25.77	23.15	27.19	0.959	0.966	0.977	0.962	0.969	56.68	43.94	48.44	67.00	24.04
Texture-Hard	20.95	23.22	22.05	18.45	23.78	0.916	0.927	0.951	0.943	0.945	117.93	98.43	96.01	101.09	41.84
Interaction-No	22.64	26.32	25.41	22.44	25.93	0.957	0.968	0.980	0.967	0.968	71.98	62.55	54.71	69.29	31.61
Interaction-Simple	24.28	27.57	26.42	23.18	27.18	0.965	0.976	0.983	0.968	0.975	55.54	48.35	45.86	65.36	23.31
Interaction-Medium	20.37	23.67	21.96	20.81	23.20	0.934	0.950	0.965	0.953	0.951	95.79	84.74	87.41	97.32	52.10
Interaction-Hard	21.14	25.00	22.10	21.29	22.29	0.931	0.949	0.961	0.953	0.940	94.04	79.40	89.63	91.82	70.54
Overall	21.53	24.88	23.79	21.76	24.18	0.942	0.956	0.970	0.957	0.957	85.33	68.56	68.54	80.18	46.73

Table 2: **Benchmark results on novel pose task.** We abbreviate NeuralVolumes [32] as 'NV', A-NeRF [53] as 'AN', NeuralBody [44] as 'NB', AnimatableNeRF [43] as 'AnN' and HumanNeRF [61] as 'HN'. indicate best, second best, and third best performance in the same split respectively. Although NV is not directly applicable to this task, we list its results as a dynamic method baseline for reference.

Figure 6: **Visualization of novel pose animation result samples.** From top to bottom, we illustrate the reposing results generated by **(a-e)**: NeuralVolumes [32], A-NeRF [53], NeuralBody [44] AnimatableNeRF [43], and HumanNeRF [61].

the rendering quality is inversely proportional to split difficulties, as reported in Fig. 4, where the circles get bigger when the difficulty grows. (2) Among case-specific dynamic methods, A-NeRF [53] achieves the best PSNR, and NeuralBody [44] and HumanNeRF [61] gets the best SSIM and LPIPS respectively. Qualitative results are shown in Fig. 5, NeuralBody [44] and A-NeRF [53] could render novel view image with fewer background artifacts than other methods, while HumanNeRF [61] can better preserve high fidelity textures, especially in high-frequency texture regions. (3) When rendering novel views for trained human action frames, hard *Texture* cases have a large performance gap among dynamic methods (refer to T-shirt case with stripe pattern in Fig. 5). Meanwhile, dynamic methods' performances on *Texture* degrade the most when difficulty rises compared to the static baselines (refer to bubbles in Fig. 4). More qualitative results in each benchmark split are shown in Fig. S9, and we analyze the conceptual difference of these methods in Sec. C.2.1.

Novel Pose Animation. Similar to novel view synthesis, when split difficulty increases the rendering quality decreases as shown in Tab. 2. Among all data factors, we found that *Deformation* and *Interaction* are insurmountable factors for current methods to model in novel poses. Qualitative results are displayed in Fig. 6, none of the methods can generate reasonable deformation in the case of the Peking opera costume. NeuralBody [44] and AnimatableNeRF [43] can not model the interactive objects, and the objects are stretched when given large poses in A-NeRF [53]. Conclusively, current methods can learn reasonable human avatars with even hard *Motion* and *Textures*, while stuck in the imperfectness of modeling hard *Deformation* and *Interaction*. These animation challenges should stimulate the communities for further investigation. More detailed analysis is provided in Sec. C.2.2 in the appendix.

Novel Identity Rendering. We report the quantitative metrics of all 39 novel identities in Tab. 3. Generally, generalizable methods with human piror [37, 26] perform better with higher robustness than category-agnostic methods [63, 59, 29]. Among category-agnostic methods, IBRNet [59] directly blends pixel color from source views, and it outperforms PixelNeRF [63] and VisionNeRF [29] that predict radiance color only from image features. We draw the conclusion that, in generalizable human rendering, human prior and appearance references from observation could help boost the generalization ability on data with large variations of poses and appearances. We illustrate the qualitative results in Fig. S11. We provide additional results and analysis in Sec. C.2.3 in supplementary.

4.4. Cross-dataset Comparison

Apart from the benchmark experiments, we also evaluate the *data generalizability* provided by our dataset and the other competitive ones, *i.e.,* GeneBody [10], ZJU-MoCap [44] and HuMMan [5].

Setting and Implementations. To eliminate the scale and annotation differences across all datasets, we train three general scene generalizable rendering methods [63, 59, 29] on these datasets with the same pixel batch per-iteration and stop training with the same $200K$ global iterations. For each method, we train each individual model on each dataset mentioned above, with a fixed image resolution 512×512 and four balanced views. To thoroughly evaluate the datasets' generalizability, we cross-verify the rendering images of *novel identities* on each dataset.

Results. The experimental results are presented in Fig. 7 in terms of the average PSNR of all three methods. From this colored error map, we conclude that training on DNA-Rendering dataset is beneficial for generalizing to the other datasets. In general, due to the existence of domain gaps, a model would perform better in the situation of an in-domain setting, where the training set and test set follow the same distribution, see diagonal elements in Fig. 7. The off-diagonal numbers report the cross-domain performances of models trained on one dataset and tested directly on other datasets' test sets. We observe an interesting phenomenon that, compared to datasets with limited data diversity and high data bias (like ZJU-MoCap [44] and HuMMan [5]), the proposed dataset enables generalization methods to achieve more plausible results even with large domain gaps. Moreover, opposite to DNA-Rendering, HuMMan [5] generalize poorly on other datasets even on cases with simple motions and appearances in ZJU-MoCap [44], despite the fact that both HuMMan [5] and our DNA-Rendering have large data volume. From a data engineering perspective, this demonstrates the construction of the proposed dataset benefits the community not merely with the amount of data, more importantly, the significant improvement in data completeness and richness. Due to space limit, we provide the detailed setup and additional results in Sec. D of the supplementary. It is worth motioning that, we also unfold the generalization performance across testing cameras and reveal the impact of color consistency for multi-camera datasets in Sec. D.4.

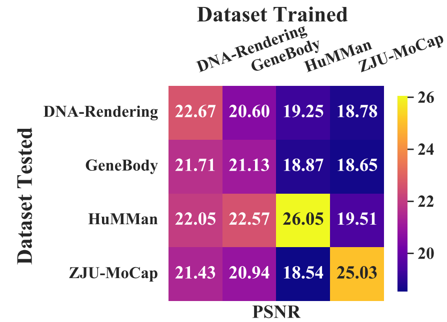


Figure 7: **Results of cross-dataset experiments**. We visualize the 'affinity' matrix of cross-dataset evaluation results.

5. Conclusion

We have presented DNA-Rendering, a large-scale and high-fidelity repository for human-centric rendering. It is a multiview human body capture dataset that covers many diverse factors like ethnicity, age, body shape, clothing, motion, and interactive objects with faithful annotations. We have also presented benchmarks to evaluate state-of-the-art approaches on the DNA-Rendering dataset with in-depth discussions, and compared our dataset with the others via cross-dataset experiments on generalization capability. We hope our DNA-Rendering project could boost the development of human-centric rendering and related domains with new reflections, challenges, and opportunities.

Acknowledgements. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also partially supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20221-0011, MOE-T2EP20221-0012), NTU NAP.

Splits	PSNR↑					SSIM↑					LPIPS*↓				
	IBR	PN	VN	NHP	KN	IBR	PN	VN	NHP	KN	IBR	PN	VN	NHP	KN
Motion-Simple	26.13	26.04	25.90	25.61	24.67	0.964	0.957	0.959	0.961	0.964	65.48	72.68	72.29	65.53	44.77
Motion-Medium	25.56	25.84	25.22	24.63	24.34	0.966	0.960	0.961	0.963	0.971	59.71	63.80	64.83	61.08	38.22
Motion-Hard	23.78	24.49	23.72	23.43	24.36	0.959	0.950	0.949	0.956	0.973	79.18	89.93	93.04	80.20	43.79
Deformation Simple	26.72	26.85	26.31	26.41	26.01	0.965	0.960	0.960	0.963	0.965	53.73	61.92	63.48	48.45	34.68
Deformation-Medium	27.46	27.55	27.90	27.28	25.83	0.965	0.961	0.965	0.963	0.972	57.30	66.82	62.82	56.02	38.00
Deformation-Hard	23.98	20.77	20.05	22.64	23.00	0.942	0.841	0.838	0.936	0.943	87.04	282.36	264.57	92.22	67.76
Texture-Simple	25.70	26.27	25.62	25.43	22.72	0.973	0.967	0.968	0.973	0.971	63.66	69.16	70.93	58.70	50.07
Texture-Medium	26.15	26.76	26.40	26.25	25.14	0.965	0.960	0.963	0.963	0.968	54.45	56.66	56.58	53.95	35.10
Texture-Hard	23.34	24.08	23.61	23.45	22.91	0.932	0.921	0.922	0.933	0.935	98.77	106.05	104.58	90.84	72.87
Interaction-No	26.08	25.91	26.39	25.46	23.43	0.968	0.958	0.963	0.966	0.968	61.99	72.67	66.91	64.35	44.95
Interaction-Simple	27.60	25.76	27.54	26.67	26.12	0.976	0.944	0.973	0.974	0.977	50.50	82.30	53.38	50.77	28.60
Interaction-Medium	24.04	24.44	24.09	23.61	22.70	0.950	0.937	0.941	0.947	0.950	84.64	96.53	93.92	86.29	63.72
Interaction-Hard	24.78	25.76	24.93	24.02	24.12	0.951	0.944	0.942	0.946	0.953	79.06	82.30	82.54	78.82	58.43
Overall	25.49	25.42	25.21	24.99	24.26	0.960	0.943	0.946	0.957	0.962	68.89	92.55	88.45	68.25	47.77

Table 3: **Benchmark results on novel identity task.** We abbreviate IBRNet [59] as 'IBR', PixelNeRF [63] as 'PN', VisionNeRF [29] as 'VN', NeuralHumanPerformer [26] as 'NHP' and KeypointNeRF [61] as 'KN'.

References

[1] Oliver Batchelor. Multi-camera calibration using one or more calibration patterns. https://github.com/oliver-batchelor/multical, 2021.

[2] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *CVPR*, 2014.

[3] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *CVPR*, 2017.

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[5] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *ECCV*, 2022.

[6] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery, 2022.

[7] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *ICIP*, 2015.

[8] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021.

[10] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint*, 2022.

[11] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *CVPR*, pages 5556–5565, 2015.

[12] XRPrimer Contributors. Openxrlab foundational library for xr-related algorithms. https://github.com/openxrlab/xrprimer.

[13] XRMoCap Contributors. Openxrlab multi-view motion capture toolbox and benchmark. https://github.com/openxrlab/xrmocap, 2022.

[14] 3D People Cooperation. 3d people dataset. https://3dpeople.com/.

[15] RenderPeople Cooperation. Renderpeople dataset. http://https://renderpeople.com/, 2017.

[16] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018.

[17] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *MICCAI*, 2018.

[18] 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. *TOG*, 2023.

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013.

[20] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022.

[21] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[23] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015.

[24] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 2017.

[25] Sai Charan Mahadevan Karunanidhi Durai Kumar, Huang Geng. Sfu motion capture database. https://mocap.cs.sfu.ca/.

[26] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*, 2021.

[27] CMU Graphics Lab. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/.

[28] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.

[29] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023.

[30] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *CVPR*, 2021.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[32] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *TOG*, 2019.

[33] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *TOG*, 2021.

[34] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *TOG*, 2014.

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

[36] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017.

[37] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *ECCV*, pages 179–197, 2022.

[38] 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. In *ECCV*, 2020.

[39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 2022.

[40] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *WACV*, 2013.

[41] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *CVPR*, 2021.

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

[43] 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.

[44] 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 *CVPR*, 2021.

[45] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

[46] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *TOG*, 2004.

[47] 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 *ICCV*, 2019.

[48] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020.

[49] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021.

[50] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[51] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*, 2022.

[52] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH*, 2022.

[53] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *NeurIPS*, 2021.

[54] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *TPAMI*, 2023.

[55] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[56] 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 *ISMIR*, 2019.

[57] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, 2021.

[58] 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. *NeurIPS*, 2021.

[59] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021.

[60] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

[61] 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 *CVPR*, 2022.

[62] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, et al. Multiface: A dataset for neural face rendering. *arXiv preprint*, 2022.

[63] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021.

[64] 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 *CVPR*, 2021.

[65] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *CVPR*, 2023.

[66] 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 *CVPR*, 2020.

[67] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, 2017.

[68] 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 *CVPR*, 2022.

[69] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *CVPR*, 2022.

[70] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *TPAMI*, 2021.

[71] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *CVPR*, 2019.

[72] 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 *CVPR*, 2023.