FreeMan: Towards Benchmarking 3D Human Pose Estimation under Real-World Conditions

Jiong Wang[1,2]*, Fengyu Yang[1]*, Bingliang Li[1], Wenbo Gou[1], Danqi Yan[1],
Ailing Zeng[3], Yijun Gao[2], Junle Wang[2], Yanqing Jing[2], Ruimao Zhang[1]†
[1]The Chinese University of Hong Kong, Shenzhen [2]Tencent [3]IDEA
jiongwang@link., fengyuyang1@link.cuhk.edu.cn, ruimao.zhang@ieee.org

Figure 1. The left displays sample frames from Human3.6M [18] and HuMMan [7], which were collected under laboratory conditions, and contrasted with our FreeMan dataset that was collected in real-world scenarios. Frames from FreeMan have been cropped into a square format for visualization purposes, with the original resolution being 1920×1080 pixels. The right-hand side demonstrates the **test results on 3DPW of the HMR model** [23] trained on these three datasets. Notably, the model trained using FreeMan is able to adapt flawlessly to real-world conditions, demonstrating its superior generalization ability. Visualization uses implementation of mmHuman3D [11].

Abstract

Estimating the 3D structure of the human body from natural scenes is a fundamental aspect of visual perception. 3D human pose estimation is a vital step in advancing fields like AIGC and human-robot interaction, serving as a crucial technique for understanding and interacting with human actions in real-world settings. However, the current datasets, often collected under single laboratory conditions using complex motion capture equipment and unvarying backgrounds, are insufficient. The absence of datasets on variable conditions is stalling the progress of this crucial task. To facilitate the development of 3D pose estimation, we present FreeMan, the first large-scale, multi-view dataset collected under the real-world conditions. FreeMan was captured by synchronizing 8 *smartphones across diverse scenarios. It comprises* $11M$ *frames from* 8000 *sequences, viewed from different perspectives. These sequences cover* 40 *subjects across* 10 *different scenarios, each with varying lighting conditions. We have also established an semi-automated pipeline containing error detection to reduce the workload of manual check and ensure precise annotation. We provide comprehensive evaluation baselines for a range of tasks, underlining the significant challenges posed by FreeMan. Further evaluations of standard indoor/outdoor human sensing datasets reveal that FreeMan offers robust representation transferability in real and complex scenes. FreeMan is publicly available at* https://wangjiongw.github.io/freeman.

1. Introduction

Estimating 3D human poses from real scene input is a long-standing yet active research topic since its huge potential in real applications, such as animation creation [60, 63], virtual

*First two authors contributes equally. Work done during Jiong Wang's MPhil study at CUHK(SZ). Wenbo Gou and Danqi Yan were research assistant at CUHK(SZ).

†Corresponding author. Email: ruimao.zhang@ieee.org

reality [15, 56], the metaverse [30, 39, 62] and human-robot interaction [17]. Specifically, it aims to identify and determine the spatial positions and orientations of the human body's parts in 3D space from input data such as the image or the video. Despite numerous models proposed in recent years [27, 34, 58], practical implementation in real scenes remains challenging due to the varying conditions such as viewpoint, occasions, human scale, uneven light conditions and complex background. Some challenges may stem from the disparity between the recent benchmarks and real-world scenarios. As shown in Fig. 1, the widely recognized Human3.6M [18], along with the currently largest dataset HuMMan [7], are usually in laboratory settings utilizing intricate equipment, which maintains constant camera parameters and offers minimal variation in background conditions. The effectiveness of the trained models when trained using these datasets often decline significantly in real-world environments.

From a data-oriented perspective, we have identified several constraints that hinder the performance of the existing models. **(1) Insufficient Scene Diversity.** Existing datasets, as shown in Tab. 1, are mainly collected in controlled laboratory conditions, which may not be optimal for robust model training due to static lighting conditions and uniform backgrounds. This limitation becomes especially crucial when the objective is to estimate 3D pose in real-world scenarios, where scene contexts exhibit substantial variability. In certain datasets, even though the data is collected from outdoor scenes, *e.g.,* MuCo [43] and 3DPW [55] in Tab. 1, the variety of scenarios remains remarkably limited. This constraint significantly hampers the applicability of trained models across a broader range of situations. **(2) Limited Actions and Body Scales.** In existing datasets, the range of human actions tends to be rather limited. Even in the currently largest dataset, HuMMan [7], the variety of actions in the publicly available data is quite restricted. Additionally, these large datasets typically employ fixed cameras to capture data from various perspectives. The distance from the camera to the actor is relatively constant, which results in a relatively fixed human body scale across different videos. **(3) Restricted Scalability.** The annotation of current datasets primarily relies on expensive manual processing, which greatly restricts the scalability of the datasets. Especially when the camera used for collection is movable, how to effectively align data from different cameras and perform efficient annotation remains an open issue.

To address these above issues, this work presents FreeMan, a novel large-scale benchmark for 3D human pose estimation under real-world conditions. FreeMan contains $11M$ frames in 8000 sequences captured by 8 smartphone cameras from different views simultaneously, as illustrated in Fig. 2. It covers 40 subjects in 10 kinds of scenes. To our best knowledge, it is the current largest multi-view 3D pose estimation dataset, with variable camera parameters and complex background environments. It is $215\times$ of the famous outdoor dataset 3DPW [55]. From a practical perspective, it has several appealing strengths: **Firstly**, a large number of scenes introduce diversity in both backgrounds and lighting, enhancing the generalization ability of models trained on FreeMan in real-world scenarios. This makes it particularly suitable for evaluating algorithmic performance in practical applications. **Secondly**, the distances between the 8 cameras and the actors are variant (*i.e.,* 2 to 5.5 meters) across sequences, resulting in significant scale changes in human bodies. **Thirdly**, although we employed mobile RGB cameras to collect data, we propose a semi-automated annotation pipeline and erroneous frame detection, thereby significantly reduce manual workload and enhance the scalability and annotation accuracy of the dataset. **Lastly**, the proposed FreeMan encompasses a wide range of pose estimation tasks, which include monocular 3D estimation, 2D-to-3D lifting, multi-view 3D estimation, and neural rendering of human subjects. We present thorough evaluation baselines for the aforementioned tasks on FreeMan, highlighting the inherent challenges of such a new benchmark.

In summary, this paper has made three contributions:

Dataset	Environment	#Subj	#Action	#Scene	#Seq	#Frame	#Camera	FPS
HumanEva[51]	Laboratory	4	6	1	168	80K	7	30
CMU Panoptic[16]	Laboratory	8	5	1	65	154M	31	30
MPI-INF-3DHP[42]	Real Scene	8	8	1	16	1.3M	14	30
3DPW[55]	Real Scene	7	47	4	60	51K	1	30
Mirrored Human[13]	Laboratory	-	-	-	-	1.5M	1	30
Human3.6M[18]	Laboratory	9	15	1	840	3.6M	4 (Fixed)	30
AIST++[32]	Laboratory	30	10	1	1408	10.1M	9 (Fixed)	30
HuMMan[7]	Laboratory	1000	500	1	400K	60M†	11 (Fixed)	30
HuMMan-released[7]	Laboratory	132	20	1	4466	278K†	11 (Fixed)	30
FreeMan	Real Scene	40	123	10‡	8000	11.3M	8 (Movable)	30 / 60

Table 1. Overview of 3D human pose datasets. [1] Comparison of our proposed FreeMan dataset with existing 3D Human Pose datasets. Only HD Cameras counted for CMU Panoptic[16]. † Only 1% of the HuMMan dataset (600K frames) is made publicly available. ‡ FreeMan includes 10 types of scenes that correspond to 29 locations. Fixed means cameras are fixed within the whole dataset, while our cameras are movable and camera poses vary among video sequences.

Figure 2. Equipment setting of data collection using 8 cameras. Cameras are attached to tripods.

- We have constructed a large-scale dataset for 3D human pose estimation under varied real-world conditions. The impressive transferability of the models trained on FreeMan to real-world scenarios has been demonstrated.
- We have showcased a simple yet effective toolchain that enables the semi-automatic annotation and efficient manual correction.
- We provide comprehensive benchmarks for human pose estimation and modeling on FreeMan, facilitating downstream applications. These baselines highlight potential directions for future algorithmic enhancements.

2. Related Work

Human Pose Datasets. Human modeling is a significant task in computer vision. Existing datasets predominantly rely on 2D and 3D keypoint annotations, with 3D keypoint datasets available in two forms: monocular and multi-view. For 2D keypoint, there are some single-frame datasets such as MPII [3] and COCO [35], which provide diverse images with 2D keypoints annotations, while video datasets such as J-HMDB [21], Penn Action [66] and PoseTrack [4] provide 2D keypoints with temporal information. In contrast, 3D keypoint datasets are often constructed in indoor scenes, such as Human3.6M [18], CMU Panoptic [22], MPI-INF-3DHP [42], AIST++ [33] and HuMMan [7] for multi-view. There also exists some outdoor datasets such as 3DPW [55] for monocular cases. Details of these datasets are shown in Tab. 1. However, the majority of outdoor datasets such as MPI-INF-3DHP, MuCo-3DHP, and 3DPW exhibit a limited variety of acquisition scenes, and the datasets that involve fixed camera poses such as AIST++.

3D Human Pose Estimation. The present study categorizes the task of 3D pose estimation into three distinct types, namely 2D-to-3D pose lifting, monocular 3D pose estimation, and multi-view 3D pose estimation. In the 2D-to-3D pose lifting task, Martinez [41] proposed a simple baseline to regress the 3d keypoints based on a convolutional neural network from 2D keypoints. However, subsequent works, such as Videopose3D [49], PoseFormer [69] and MHFormer [34], have improved upon this baseline by integrating temporal information into their models. In monocular 3D pose estimation task, HMR [23], SPIN [29] takes a single RGB image as input to perform 3D huna pose estimation, which is often used as baselines for comparison with other algorithms, such as PARE [27], SPEC [28] and HybrIK [31]. Additionally, multi-view methods are proposed to accommodate potential body parts overlapping in monocular view. Iqbal's [19] and MCSS [46] adopt weak supervision to reduce the dependence on the 3D annotated pose, while Canonpose [57] and EpipolarPose [26] turned to self-supervise fashion to deal with multi-view data.

Neural Rendering of Human Subjects. With the development of NeRF [44] in dynamic scene rendering, people also focus on the dynamic rendering of humans. Compared to dynamic scenes, the non-rigid property of humans has more challenges. The prior knowledge of body movements can provide a good prior for rendering, and many methods use SMPL [40] as a prior for body rendering. Most methods reconstruct human bodies through multi-view videos [38, 47, 59], while recent works have also employed single-view videos, such as HumanNeRF [61], FlexNeRF [20], YOTO [24].

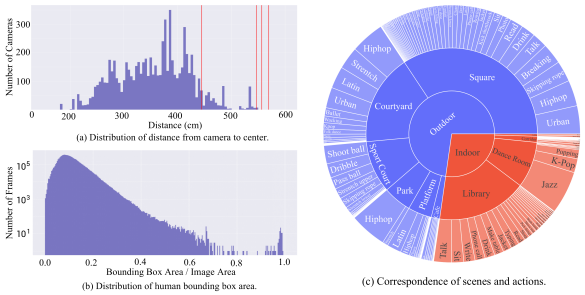

Figure 3. (a) Distribution of distance from the camera to the center of the system, indicated by translation along the z-axis in camera parameters. Four vertical red lines represent the distance of 4 cameras in Human3.6M [18]. (b) Distribution of human bounding box areas. The horizontal axis represents the ratio of the bounding box area over the image area. The vertical axis is in log scale. (c) Correspondence of scenes and actions. Areas of blocks represent the scale of the respective frame number. The outmost circle shows actions and the circle in the middle present 10 type of scenes in our dataset. Zoom in $10\times$ for the best view.

3. FreeMan Dataset

FreeMan is a large-scale multi-view dataset under real-world conditions with precise 3D pose annotations. It comprises $11M$ frames from 1000 sessions, featuring 40 subjects across 10 types of scenes. The dataset includes $10M$ frames recorded at 30FPS and an additional $1M$ frames at 60FPS. Next, we highlight the diversity of FreeMan, from various scenario selections, actions, camera settings and subjects.

Scenarios. We design 10 types of real-world scenes, including 4 indoor and 6 outdoor scenes, for our data collection. Fig. 3 (c) illustrates the scene diversity of our FreeMan. The blue section represents the outdoor part, while the red part refers to frames captured in indoor scenes. Specifically, there are 2.76 million frames captured indoors and 8.45 million frames captured outdoors. In the outdoor data, there are different frame numbers collected under varying lighting conditions, with 1 million frames captured at night or dusk and 7.45 million frames captured during daytime. Moreover, the central block of the circle denotes different scenarios, while the blocks on the outermost circle refer to actions. The areas of the blocks are proportional to frame number. Please refer to supplementary material for more details.

Action Set. Following the popular action recognition dataset

NTU-RGBD120 [37], we compose our action set with several common actions corresponding to scenes in daily life, $e.g.$, drinking and talking in a cafe, reading in the library. Meanwhile, subjects interact with real objects to make data as close to real world as possible. As shown in the topmost row of Fig. 4, interaction with objects brings complicated occlusions, making our data more challenging. For outdoor scenarios, we set the data collection field as large as possible to help subjects perform actions with little restriction.

Camera Positions. Cameras in previous 3D human pose datasets [7, 16, 18] are fixed, resulting that only a few camera poses being included. As shown in Fig. 2, cameras are attached to tripods and are newly placed from time to time, and translation from the center of the system to camera d, which is the physical distance between the camera and the system center, can vary from 2m to 5.5m. Fig. 3 (a) shows the distribution of d and the corresponding number of cameras. Most cameras are located around 4 meters far away from the system center. Besides, we show the distribution of the human bounding box area in Fig. 3 (b), in a unit of ratio to the whole image area, to demonstrate the variation of human size. With variations in camera translation and human actions, the area of human bounding boxes varies from 0.01 to 0.7 of the whole image area.

Subjects. There are 40 subjects participating in the construction of FreeMan and recruitment is completely based on voluntary. Among them, 22 actors are trained dancers for dance actions. All of them are well-informed and signed the agreement to make data public for research purposes only.

4. Data Acquisition & Annotation Pipeline

Overview. To collect a large-scale dataset from real-world environments, we developed a comprehensive toolchain, as shown in Fig. 5. Unlike previous toolchains used in controlled or idealized conditions, we carefully accounted for potential challenges in outdoor settings, including calibration and synchronization errors. To overcome these issues, we proposed an semi-automated pipeline including error detection and manual correction to ensure efficient data collection and annotation.

4.1. Hardware Setup

Cameras. We collect FreeMan via 8 Mi11 phones [1] indexed from 1 to 8 as our data collection devices. ***Note*** *8* ***collection of one action as one session, which corresponds to 8 RGB sequences from 8 views,*** and each phone is attached to a tripod to keep stable during data collection. As shown in Fig. 2, all devices are positioned in a circle around a human at a height of approximately 1.6 meters above the ground, and the distance from cameras to the system center varying from 2 to 5.5 meters, which is similar to real-life usage scenarios. Each smartphone captures RGB sequences using its main camera at 1920×1080 resolution and $30/60$ FPS. During the data collection process, actors perform actions facing the cameras with odd-numbered indices. As shown in Fig. 5 (a), the only requirement beyond devices is a stable network connection to server for data transmission.

Device Synchronization. Previous works [7, 16, 18] have synchronized devices using wired interfaces in a laboratory. However, the complexity of the entire system coupled with the difficulty in deploying it in real-world environments, has prompted us to consider alternative methods. To address issues related to usability and device constraints, we connect all devices wirelessly to a single server and developed an Android app that utilizes the Network Time Protocol (NTP) [45] to calculate the time difference between each device and the server's clock. During the capture process, temporal information is stored locally on each device as a timecode, while the server records the synchronized capture interval for all devices. The starting frame is determined by matching the timecode to the frame closest to the server's clock time. As shown in Fig. 5(b), synchronization errors are smaller than a single frame during our testing, corresponding to 33ms and 16ms for 30FPS and 60FPS, respectively.

Chessboard-based Calibration. At the beginning of each session, we first shoot a chessboard with known size tiled at the center of the system, then calculate the intrinsic and extrinsic camera parameters following the standard implementation in OpenCV [6, 68]. Please refer to supplementary meterial for details of data flow.

Pixel Alignment However, calibration with coarse matching points on chessboard is not accurate enough. After data collection and synchronization, we extract one frame from all synchronized videos, and then use LightGlue [36] to calculate dense matching points across views. Then dense matching points are used to further refine the camera extrinsic parameters resulted from chessboard-based calibration.

4.2. Pose Annotation.

Once videos are collected, we use a state-of-the-art detector YOLOX [14] to detect human bounding box and HRNet-w48 [52] to detect 2D keypoints of 8 views $K_{2D} \in \mathbb{R}^{8\times17\times2}$ in COCO [35] format. To eliminate the effect of potential wrong keypoints output, keypoint predictions with confidence lower than a threshold ϕ are removed. Then remaining 2D keypoints are used for triangulation to get 3D human pose $K_{3D} \in \mathbb{R}^{17\times3}$ with pre-computed camera parameters. Here, we set ϕ to be 0.5. Furthermore, we optimize K_{3D} with smoothness constraints and bone length constraints introduced in HuMMan [7] resulting in optimized 3D pose $\tilde{K}_{3D} \in \mathbb{R}^{17\times3}$. Then we fit a standard SMPL [40] model to the estimated 3D skeleton by SMPLify [5] to produce a rough mesh annotation. After that, we project 3D keypoints to 2D image planes of each view using corresponding camera parameters. With regularization in triangulation and optimization along the temporal axis, the re-projected 2D

Figure 4. The diverse frames in FreeMan. The topmost two rows presents a range of indoor and outdoor scenes, highlighting human-object interactions and the diversity of scene contexts, lighting conditions, and subjects. The third row exhibits frames from different views. The final row illustrates the temporal variation of human poses from a consistent viewpoint, emphasizing the dynamism of motion capture.

poses $\tilde{K}_{2D}$ is more accurate than K_{2D}, especially for occlusion cases. Comparison between original K_{2D} and $\tilde{K}_{2D}$ are shown in the left part of Fig. 7.

Erroneous Pose Detection & Correction. Although 2D pose estimator has been well developed, pose with heavy occlusions can be inaccurate. Thus, we propose a pipeline to filter erroneous 2D keypoints among vast millions of frames and then correct them **by human annotators**. As shown in Fig. 5, estimated 2D poses are feed into a pre-trained image generator to generate human images. Then we use SAM [25] to get human mask of original and generated images and intersection-over-union (IoU) between these mask are calculated. Poses correspond to IoU lower than a threshold α are considered as erroneous ones and then checked by human annotator. Specifically, we choose Stable Diffusion 1.5 and ControlNet [64] as conditional image generator and DeepDataSpace [2] are used as annotation tools. Fig. 6 presents examples of correct and erroneous cases. More detailed processes and results are displayed in the supplementary material.

4.3. Keypoint Quality Assessment

To demonstrate the effectiveness of our toolchain, we test it on Human3.6M [18]. We select 3 different actions of each subject in the training set, which covers 10% sequences of the whole training set and all kinds of actions. Following [7], keypoint quality is assessed by Euclidean distance between estimated 2D poses and ground truth 2D poses in units of pixels. The error results in less than 1% of pixels for images of 1000×1000, indicating that our toolchain can generate annotations with an accuracy that is acceptable considering the cognitive errors inherent in human labeling.

5. Benchmarks

We have constructed four benchmarks utilizing images and annotations derived from our dataset. The data is subdivided based on subjects, allocating 18 subjects for training, 7 for validation, and 15 for testing purposes. This partitioning results in three subsets composed of $5.87M$, $700K$, and $3.69M$ frames, respectively. For each benchmark, subject lists of each subset are shared, and only views selected from the session vary for each task.

Monocular 3D Human Pose Estimation (HPE). This task involves taking a monocular RGB image or sequence as input and predicting 3D coordinates in camera coordinate system. We randomly select one view from each session for this task. The performance of algorithms is measured by widely used Mean Per Joint Position Error (MPJPE) [18] and Procrustes analysis MPJPE (PA-MPJPE) [40].

2D-to-3D Lifting. Given that 2D human poses can be pre-

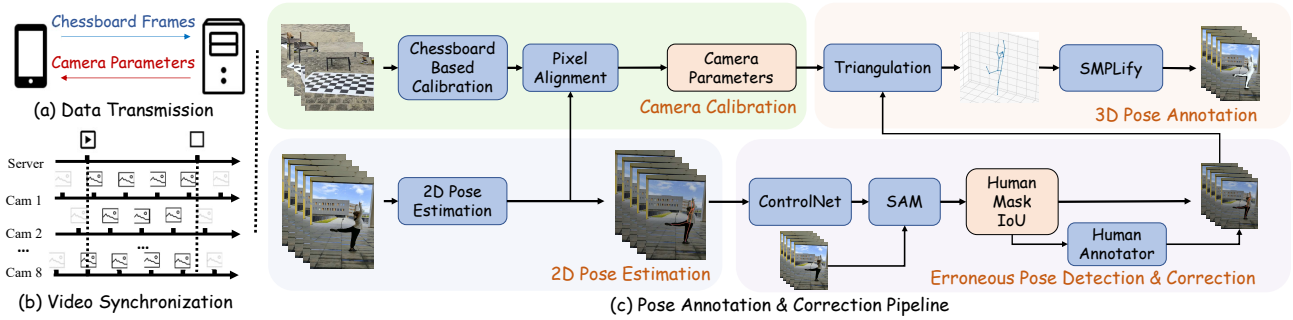

Figure 5. The illustration of data collection and annotation toolchain: (a) depicts the transmission of signals between cameras and servers for camera calibration, where chessboard frames are sent to the server, and camera parameters are returned. (b) demonstrates the synchronization process among devices. (c) showcases the pipeline for pose annotation.

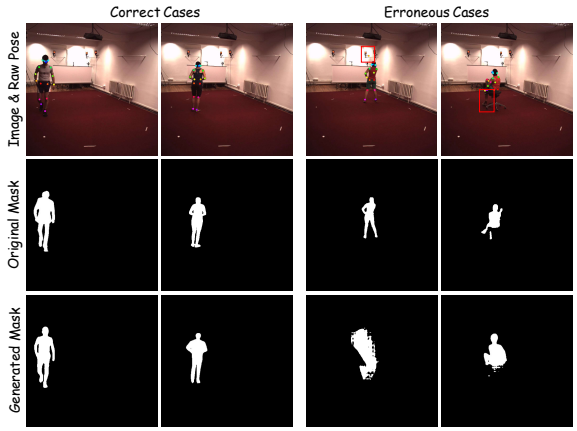

Figure 6. Demonstration of erroneous pose detection in Sec. 4. Human3.6M examples shown for quality assessment. The first row shows input frame and 2D keypoints by Pose estimator, and the last two rows show segment mask of original image and generated image by SAM. The left two columns are examples of correct poses, while the right two columns refer to cases with erroneous keypoints as highlighted by the red boxes. Please zoom in for details.

dicted using existing 2D keypoint detectors [8, 9, 12, 52], the primary goal of this task is to effectively elevate these 2D poses into the 3D space within the camera coordinate system. The evaluation metrics are the same as HPE.

Multi-View 3D Human Pose Estimation. Estimating the 3D human pose from multiple views presents a natural solution to overcome occulusion in motion capture. For this task, models are provided with images or videos from multiple views, along with corresponding camera parameters. The objective is to predict the 3D coordinates of human joints in the same world coordinate system as the cameras. following implementation of [54], metrics of the task is MPJPE and average precision (AP) with specific thresholds.

Human Neural Rendering. The free-viewpoint rendering of humans is a significant issue in human modeling. With the rise in popularity of neural radiance fields (NeRF) [44] for the novel view rendering task, several methods, including HumanNeRF [61], have emerged. These methods utilize monocular human motion videos as input to synthesize novel views of dynamic humans through NeRF. The widely used metrics of prediction are PSNR, SSIM [67] and LPIPS [65].

Method			HMR		PARE	
Train	Supervision	Test	MPJPE	PA	MPJPE	PA
Human3.6M	2D+3D KPTs	3DPW	279.92	133.13	118.54	81.22
HuMMan	2D+3D KPTs	3DPW	407.57	192.75	110.99	63.11
HuMMan	2D KPTs+SMPL	3DPW	475.73	184.15	114.20	66.19
HuMMan	2D+3D KPTs+SMPL	3DPW	437.52	203.17	114.33	72.12
FreeMan	2D+3D KPTs	3DPW	157.46	87.93$\uparrow_{33.95\%}$	118.31	68.72$\uparrow_{15.39\%}$
FreeMan	2D KPTs+SMPL	3DPW	151.85	88.85$\uparrow_{51.75\%}$	94.27	60.39$\uparrow_{8.76\%}$
FreeMan	2D+3D KPTs+SMPL	3DPW	159.31	91.33$\uparrow_{55.04\%}$	98.33	64.51$\uparrow_{10.55\%}$

Table 2. Monocular 3D HPE performance of HMR [23] and PARE [27] trained on different dataset for monocular Human Pose Estimation. PA stands for PA-MPJPE and both metrics are in unit of millimeters. The lower metrics is, the better performance model obtains. All released part of HuMMan is used for training. ↑ refers to the improvement relative to HuMMan and ↑ refers to the improvent relative to Human3.6M.

6. Experiments

In this section, we experiment with the four benchmarks. In human 3D pose estimation tasks, we conduct several transfer tests with other standard datasets to evaluate the effectiveness and transferability of our proposed FreeMan dataset. Existing similar datasets, Human3.6M [18] & HuMMan [7], are used for comparison. Since *HuMMan only releases 1% of data for pose estimation**, we only involve it in monocular 3D human pose estimation and 2D-to-3D lifting. As for the neural human rendering, we train the model from one of the 8 views and test on the other the views in selected sessions.

6.1. Monocular 3D Human Pose Estimation

Implementation details. For the Human3.6M [18] and HuMMan [7] datasets, all views in their training set are utilized. To balance number of frames, we randomly sample a single view from sessions in the training split for FreeMan, resulting that the frame numbers of all three datasets

*https://opendatalab.com/OpenXDLab/HuMMan

Figure 7. The examples of human pose annotations are presented as follows. At left, the first row displays 2D keypoints directly generated by HRNet-w48 [52], while the second row presents re-projected 2D poses. For heavy occlusions, our pipeline corrects the erroneous keypoints effectively. The right part showcases the SMPL annotation examples for each view in our dataset.

being $312K$, $253K$, and $233K$, respectively. Videos of all three datasets are downsampled to 10FPS, following the implementation of MMPose [10]. Following [48], we select HMR [23] and PARE [27] as models to evaluate and implement experiments with configurations open-sourced by [48]. Please refer to supplementary material for more.

Results. We perform testing on the test set of 3DPW [55]. The performance of the models trained on different datasets, with varying types of supervision, are reported in Tab. 2. Notably, the HMR models trained on FreeMan exhibit significantly better performance on the 3DPW test set compared to those trained on Human3.6M and HuMMan with PA-MPJPE 133.13mm and 192.75mm respectively, which indicates that FreeMan demonstrates superior generalizability compared to the others. The same results are obtained with PARE, further confirming that FreeMan outperforms even in more advanced algorithms. This can be attributed to the diversity of scene contexts and human actions present in our dataset, which provides better transferability in real-world scenarios.

6.2. 2D-to-3D Pose Lifting

Implementation Details. For this task, we employ CNN-based methods, SimpleBaseline [41] and VideoPose3D [49], and Transformer-based methods, PoseFormer [69] and MHFormer [34], and all methods follow corresponding official implementations. To verify the effect of the dataset scale, we also train our model on the whole training set. The results of SimpleBaseline and MHFormer are presented in Tab. 3, and more details can be found in supplementary material.

Results. As shown in Tab. 3, results of the in-domain test on FreeMan are provided as a baseline for future work. For in-domain testing, MPJPE of SimpleBaseline on FreeMan (79.22mm) is larger than that on HuMMan [7] (78.5mm[†]) and Human3.6M [18] (53.4mm[‡]), demonstrating that FreeMan is a more challenging benchmark. Besides, all the methods trained on FreeMan tend to generalize better than that on HuMMan and Human3.6M when testing on AIST++ under the same setting as MPJPE and PA-MPJPE are much smaller in cross-domain test. Although the scale of FreeMan training set is of a similar magnitude as HuMMan's, which is much smaller than Human3.6M's, models trained on FreeMan outperform models trained on the other two by a large margin. Furthermore, when the training set is expanded to all frames in training split, FreeMan can further boost models to achieve better performance, proving that our large-scale data helps to improve model performance.

Algorithm	Train	Test	MPJPE (mm)	PA (mm)
SimpleBaseline	FreeMan	FreeMan	90.53	54.17
	FreeMan[†]	FreeMan	79.22	49.11
	Human3.6M	AIST++	212.57	138.98
	HuMMan	AIST++	255.5	116.86
	FreeMan	AIST++	156.96	105.85↑$_{10.30\%}$
	FreeMan[†]	AIST++	126.23	88.07↑$_{24.64\%}$
MHFormer	FreeMan	FreeMan	93.00	63.50
	FreeMan[†]	FreeMan	77.06	53.38
	Human3.6M	AIST++	171.19	133.37
	HuMMan	AIST++	188.73	101.52
	FreeMan	AIST++	132.99	88.79↑$_{12.54\%}$
	FreeMan[†]	AIST++	124.34	79.22↑$_{21.97\%}$

Table 3. Performance of methods with different training and testing datasets in 2D-to-3D Pose Lifting. PA stands for PA-MPJPE. [†] refer to experiments with the whole training set of FreeMan. Smaller MPJPE and PA-MPJPE indicate better performance. Highlighted rows show training on our dataset achieves the best performance in the transfer test. ↑ refers to the improvement relative to HuMMan.

6.3. Multi-View 3D Human Pose Estimation

Implementation Details. We conduct in-domain and cross-domain tests between Human3.6M and FreeMan to evaluate the effectiveness and generalization ability. We conduct the experiments with VoxelPose [53], which locates the human root first and then regresses 3D joint location accordingly. COCO-format poses in FreeMan are interpolated to match

[†]As full data not accessible, we use result from HuMMan [7] directly.
[‡]Results of our implementation.

Train	Test	AP@25mm (%) ↑	AP@50mm (%) ↑	AP@75mm (%) ↑	AP@100mm (%) ↑	Recall@500mm (%) ↑	MPJPE@500mm (mm) ↓
Human3.6M	Human3.6M	32.32	97.47	98.61	98.99	100.00	25.29
Human3.6M	FreeMan	0.00	0.00	0.00	0.00	0.06	89.85
Human3.6M	FreeMan (w/ GT Root)	0.00	1.27	11.44	21.40	96.20	103.02
FreeMan	FreeMan	43.38	88.77	97.73	99.12	99.97	26.07
FreeMan	Human3.6M	0.00	5.77	82.85	92.62	96.68	61.29
FreeMan	Human3.6M (w/ GT Root)	0.00	6.60	87.91	95.38	100.00	58.30

Table 4. Multi-View 3D Pose Estimation results of VoxelPose [54]. Ground truth root position (GT Root) is not used if not specified. Recall@$500mm$ shows the percentage that falls within the threshold, and the MPJPE@$500mm$ indicates the average MPJPE values within the threshold. Rows highlighted shows the best setting in cross-domain test.

Scene	PSNR↑	SSIM↑	LPIPS* ↓
Square	25.98	0.9501	58.38
Corridor	24.57	0.9340	**81.39**
Sports Port	26.33	0.9662	30.09
Park	23.86	0.9439	73.61
Courtyard	28.56	0.9630	53.99
Dance Room	**30.11**	0.9658	43.34
Library	29.41	**0.9665**	31.53
Platform	26.79	0.9439	70.01
Lobby	25.41	0.9387	78.80
Cafe	27.32	0.9644	37.88

Table 5. Neural rendering results by using HumanNeRF [61] on 10 scenes of FreeMan. Note that $\texttt{LPIPS}^* = \texttt{LPIPS} \times 10^3$. The highest values are bolded and underlined ones refer to the lowest.

that in Human3.6M. We trained VoxelPose [53] following official implementation. For Human3.6M, bounding box annotations are from [50] and its validation set is used for the test. For FreeMan, we only use 4 odd-indexed views from the training set.

Results. Results of all experiments are reported in Tab. 4. For in-domain testing, the model trained on FreeMan achieves MPJPE@500mm of 26.61mm on test set consisting of *odd-indexed* views. For cross-domain testing, the model trained on FreeMan achieves Recall@500mm of 96.68% and MPJPE@500mm is 61.29mm on Human3.6M validation set. However, the model trained on the Human3.6M dataset fails to locate human on FreeMan test set, resulting zero AP with threshold smaller than 100mm. To get rid of the effects of root location, we input the ground truth root locations to model directly. With this setting, the model trained on Human3.6M obtains MPJPE@500mm of 103.02mm on FreeMan test set, while the model trained on FreeMan can obtain MPJPE@500mm of 58.30mm on Human3.6M validation set. Results show that the model trained on FreeMan has a much better generalization ability, while that on Human3.6M struggles in transfer testing.

6.4. Neural Rendering of Human Subjects.

Implementation Details. We employ 10 scenes captured by FreeMan to train HumanNeRF [61]. To obtain human body segmentation annotations, we utilize the SAM [25] algorithm with our bounding boxes as prompts. Throughout the training step, we randomly select one view for each session and render the rest 7 view as novel views for testing. We then calculate metrics including PSNR, SSIM, and LPIPS, to evaluate the performance of the model. Please refer to supplementary material for results of data at 60FPS.

Results. The reconstruction results in 10 scenes are shown in Tab. 5. The best reconstruction achieves a high PSNR of 30.11dB which indicates FreeMan contains contents that the models can learn and fit very well. While the performance varies, the lowest PSNR of 23.86 shows FreeMan also contains contents that are outside of model's learning scope and challenging. Additionally, the results in 10 scenes including both easy contents that the model can handle well and challenging contents demonstrating the diversity of FreeMan.

7. Conclusion

We present FreeMan, a novel large-scale multi-view 3D pose estimation dataset 3D human pose annotations. We elaborately develop a simple yet effective semi-annotation pipeline to automatically annotate frame-level 3D landmarks at a much lower cost, and build a comprehensive benchmark for 3D human pose estimation.

Extensive experimental results demonstrate the difficulty of test in varied conditions and the strengths of the proposed FreeMan. As a large-scale human motion dataset, our FreeMan addresses the existing gap between the current datasets and real-scene applications, and we hope that it will catalyze the development of algorithms designed to model and sense human behavior in real-world scenes.

Limitations. Prompts to generation model require careful tuning for high quality and accuracy of error pose detection can be limited by human image generation models.

Acknowledgement

We sincerely thank all volunteers and MaxDancingClub from CUHK(SZ) for participation, and Mr. Ruipeng Cao for software development. The work is partially supported by the Young Scientists Fund of the National Natural Science Foundation of China under grant No. 62106154, by the Natural Science Foundation of Guangdong Province, China (General Program) under grant No.2022A1515011524, and by Shenzhen Science and Technology Program JCYJ20220818103001002 and ZDSYS20211021111415025 and by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong (Shenzhen).

References

[1] Mi11. https://www.mi.com/global/product/mi-11/, 2022. 4

[2] International Digital Economy Academy. Deepdataspace. https://github.com/IDEA-Research/deepdataspace, 2023. 5

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3

[4] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *CoRR*, abs/1710.10000, 2017. 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 4

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[7] 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. 1, 2, 3, 4, 5, 6, 7

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 6

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 6

[10] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 7

[11] MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark. https://github.com/open-mmlab/mmhuman3d, 2021. 1

[12] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 6

[13] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, 2021. 2

[14] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4

[15] Shivam Grover, Kshitij Sidana, and Vanita Jain. Pipeline for 3d reconstruction of the human body from ar/vr headset mounted egocentric cameras. *arXiv preprint arXiv:2111.05409*, 2021. 2

[16] TomasSimon HanbyulJoo, HaoLiu XulongLi, LinGui LeiTan, and TimothyGodisart SeanBanerjee. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1), 2019. 2, 4

[17] Abdelfetah Hentout, Mustapha Aouache, Abderraouf Maoudj, and Isma Akli. Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017. *Advanced Robotics*, 33(15-16):764–799, 2019. 2

[18] 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. 1, 2, 3, 4, 5, 6, 7

[19] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5243–5252, 2020. 3

[20] Vinoj Jayasundara, Amit Agrawal, Nicolas Heron, Abhinav Shrivastava, and Larry S Davis. Flexnerf: Photorealistic free-viewpoint rendering of moving humans from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21118–21127, 2023. 3

[21] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199, 2013. 3

[22] 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. 3

[23] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 3, 6, 7

[24] Jaehyeok Kim, Dongyoon Wee, and Dan Xu. You only train once: Multi-identity free-viewpoint neural human rendering from monocular videos. *arXiv preprint arXiv:2303.05835*, 2023. 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 5, 8

[26] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1077–1086, 2019. 3

[27] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: part attention regressor for 3d human body estimation. *CoRR*, abs/2104.08527, 2021. 2, 3, 6, 7

[28] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC:

seeing people in the wild with an estimated camera. *CoRR*, abs/2110.00620, 2021. 3

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[30] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352*, 2021. 2

[31] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. *CoRR*, abs/2011.14672, 2020. 3

[32] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 2

[33] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 3

[34] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. *CoRR*, abs/2111.12707, 2021. 2, 3, 7

[35] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 3, 4

[36] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 4

[37] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2684–2701, 2019. 4

[38] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *arXiv preprint arXiv:2001.04947*, 2020. 3

[39] Doina Popescu Ljungholm. Metaverse-based 3d visual modeling, virtual reality training experiences, and wearable biological measuring devices in immersive workplaces. *Psychosociological Issues in Human Resource Management*, 10 (1), 2022. 2

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3, 4, 5

[41] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. *CoRR*, abs/1705.03098, 2017. 3, 7

[42] 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 *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2, 3

[43] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, 2018. 2

[44] 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. 3, 6

[45] D.L. Mills. Internet time synchronization: the network time protocol. *IEEE Transactions on Communications*, 39(10): 1482–1493, 1991. 4

[46] Rahul Mitra, Nitesh B Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 6907–6916, 2020. 3

[47] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 3

[48] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 7

[49] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. *CoRR*, abs/1811.11742, 2018. 3, 7

[50] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4342–4351, 2019. 8

[51] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.*, 87(1-2):4–27, 2010. 2

[52] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 4, 6, 7

[53] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, pages 197–212. Springer, 2020. 7, 8

[54] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision (ECCV)*, 2020. 6, 8

[55] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 2, 3, 7

[56] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1643–1652, 2018. 2

[57] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2021. 3

[58] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3d human pose estimation. *Advances in Neural Information Processing Systems*, 2021. 2

[59] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020. 3

[60] 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 (CVPR)*, pages 16210–16220, 2022. 1

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 3, 6, 8

[62] Jae Shin Yoon. *Metaverse in the Wild: Modeling, Adapting, and Rendering of 3D Human Avatars from a Single Camera*. PhD thesis, University of Minnesota, 2022. 2

[63] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021. 1

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 5

[65] 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. 6

[66] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *2013 IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 3

[67] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 6

[68] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 4

[69] Ce Zheng, Sijie Zhu, Matías Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *CoRR*, abs/2103.10455, 2021. 3, 7