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

Supplementary Material

1. Overview

FreeMan consists of data from 10 types of scenarios and 27 locations, and lighting conditions are various among locations. Meanwhile, we adopt different action sets for scenarios to make FreeMan more diverse. In this section, we first compare FreeMan with existing relevant works to visually demonstrate its diversity. Then, we showcase more examples from various perspectives, including scenes, lighting conditions, and occlusions, to provide a more comprehensive representation of FreeMan.

1.1. Comparison with Existing Datasets

By showing example frames in Fig. 1, we demonstrate the breadth and realism of our 3D human body dataset, which surpasses previous works by encompassing diverse and dynamic real-world conditions. We believe that our dataset will serve as a valuable resource for advancing research in computer vision, human pose estimation, and action recognition.

1.2. Scenes

In this section, we provide an overview of the 10 scene categories included in FreeMan. Fig. 2 presents three images for each scenario categories. It can be seen that background are much different across scenarios. For each category, we present three representative images that capture the essence of the scene and highlight the diversity of actions performed.

1.3. Lighting

This section focuses on the different lighting conditions captured in our dataset. We specifically address challenging lighting scenarios such as backlighting and overexposure, which are common in real-world environments. We include a collection of images in Fig. 3 that demonstrate how FreeMan represents these variations and challenges in lighting. Besides well-lit cases, there are also challenging cases of backlighting, which is also common in real world.

1.4. Occlusion

The third section highlights human occlusion phenomena, encompassing both object interactions and self-occlusion during complex actions as shown in Fig. 4. We provide visual examples showcasing instances where the human body is partially or fully occluded by objects, as well as instances where self-occlusion occurs during intricate movements.

2. Limitations & Future Work

FreeMan is the first attempt to address the challenges of 3D human pose estimation in real-world environments with diverse variations. However, it is important to acknowledge that the real world encompasses a multitude of variables, and there are additional conditions that can be further explored. Despite the limited size of FreeMan, it serves as a catalyst for advancing algorithmic research in this domain and provides a means to evaluate the performance of existing methods under varying conditions.

Currently, our pose annotations are in the form of 17 keypoints following the COCO format. However, given the increasing demand for fine-grained human body modeling, the estimation of whole-body key points has become more crucial. In future work, we can consider expanding our pose annotations to cover the entire body, enabling more comprehensive analysis and modeling.

In addition to pose estimation, we have also extended FreeMan dataset to improve the performance of human body rendering algorithms in real-world scenarios. Leveraging the data collected under diverse and dynamic environmental conditions, FreeMan offers a more realistic input for studying the robustness of human body rendering algorithms. Furthermore, in addition to NeRF-based algorithms [24], future research can explore the application of recent dynamic 3D Gaussian splatting algorithms [14] with FreeMan dataset.

3. Toolchain

In this section, we introduce more details of our toolchain. this section focus on pixel alignment for camera calibration and erroneous pose detection.

3.1. Camera Calibration

To calibrate camera accurately, FreeMan adopts two-stage camera calibration. Before collecting each session, we collect frames of chessboard and conduct standard calibration process in OpenCV [2, 26]. At this stage, we first calibrated the cameras by capturing at least thirty frames of a calibration board with each camera to determine their intrinsic parameters. Next, the cameras were fixed in a stationary position on tripods, and all cameras simultaneously captured images of a calibration board at the central position to calculate the extrinsic parameters. However, due to variations in lighting conditions and distances from certain perspectives, errors in estimating the extrinsic parameters are inevitable in this step.

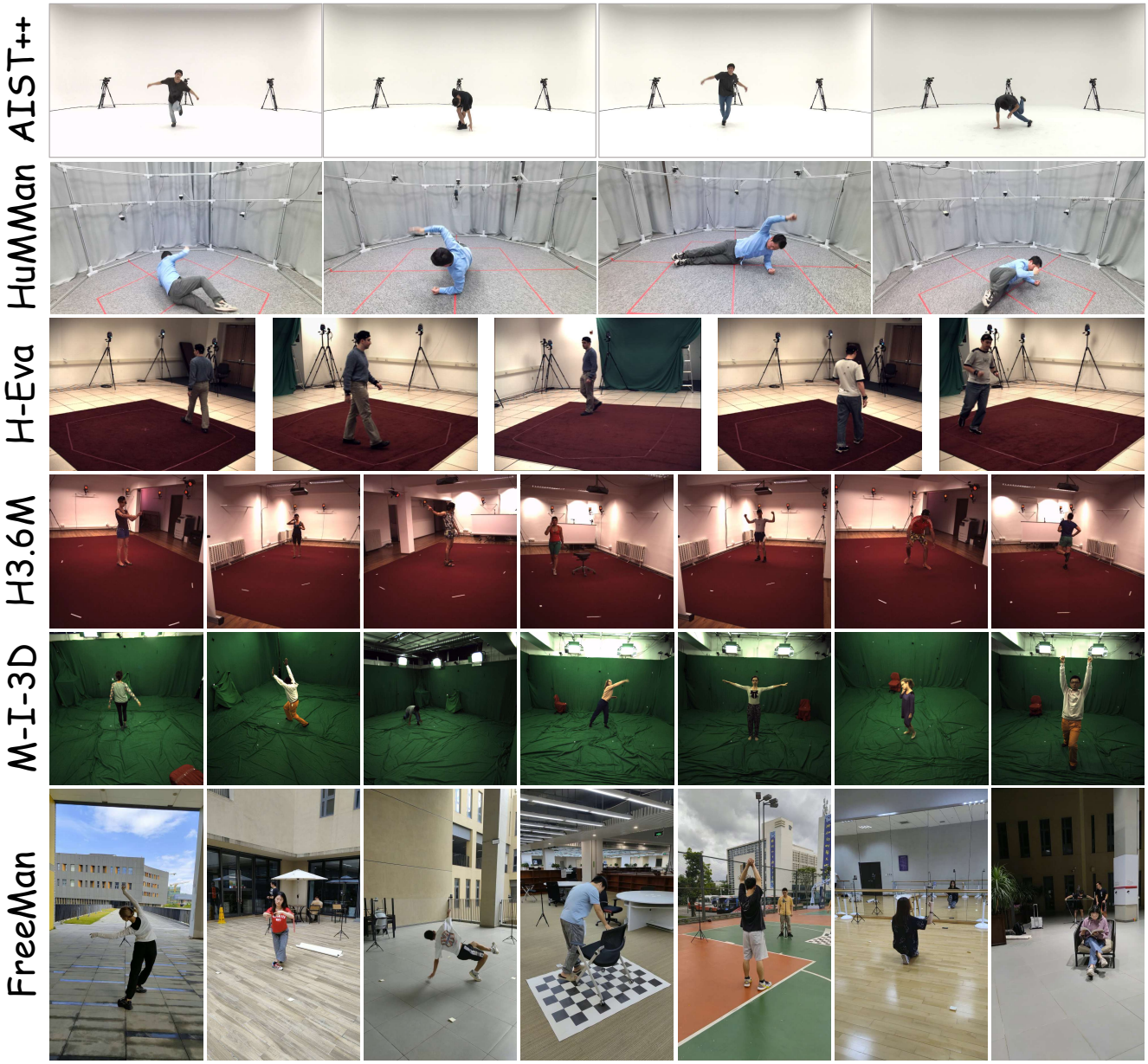


Figure 1. Comparison between FreeMan and existing multi-view 3D human datasets, and all images maintain their original aspect ratio. H-Eva, H3.6M and M-I-3D represents Human-Eva [21], Human3.6M [7] and MPI-INF-3DHP [17], respectively. Datasets except for FreeMan are all collected in fixed laboratory environment. Although MPI-INF-3DHP composites different textures to the background, its human-object interaction and action sets are much simpler.

To address the errors in extrinsic parameter calibration, we introduce pixel alignment and calculate dense image matching points using synchronized video content. By leveraging well-synchronized video data, we can establish correspondence between pixels in different camera views, allowing for more accurate estimation of the extrinsic parameters.

We use LightGlue [13] to calculate image correspondances between adjacent views. For all 8 viewpoints, we take each viewpoint as the center and select three consecutive viewpoints as a group. For each group, the pixel matcher calculates matching pairs of pixels between two adjacent views and then filters out the pairs of pixels that are common to the three views. For each set of videos, we select frames containing at least 50 sets of pixel pairs, and then use these pixel pairs to calculate camera extrinsic parameters. Once this process is completed for all viewpoints, we perform coordinate system matching to align all the viewpoints within the same world coordinate system.

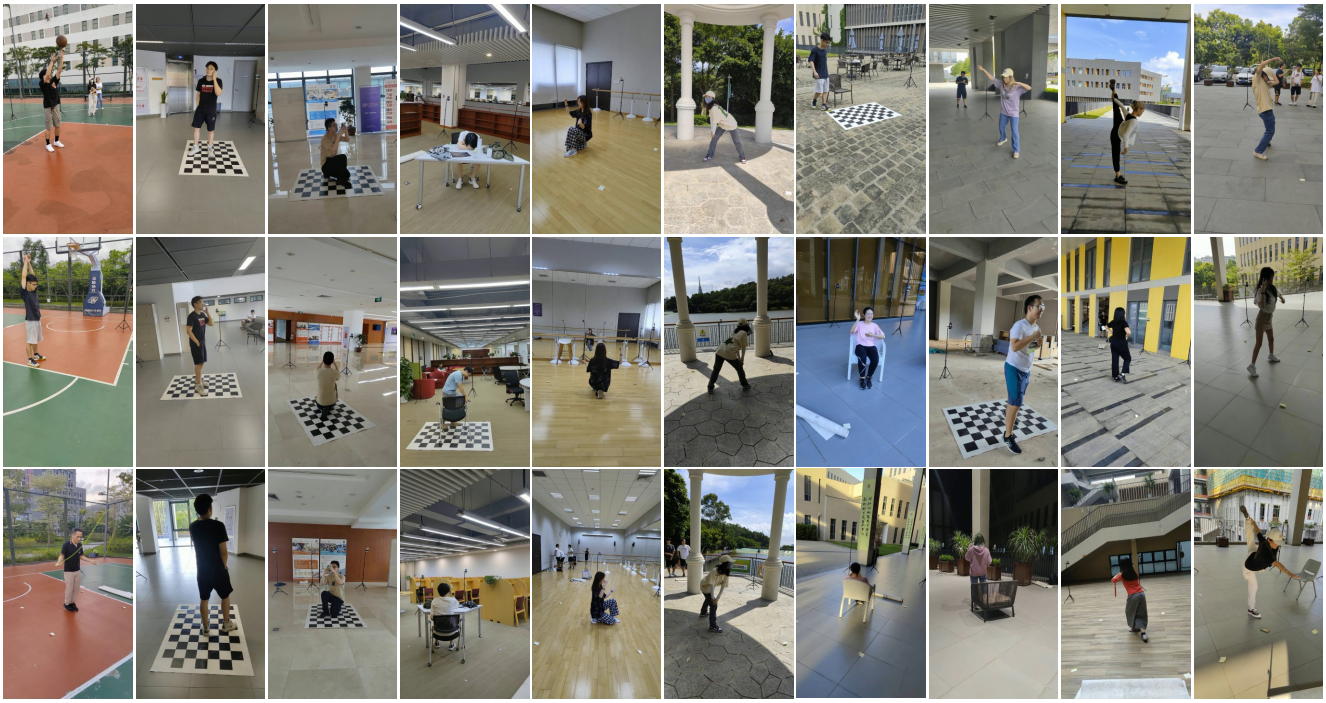


Figure 2. Example images of 10 kinds of scenarios. For scenes shown in leftmost six columns, various of background from different views are presented. For outdoor scenes such as cafe, platform, courtyard and square, **various locations are included**.

Figure 3. Example images for lighting conditions. The first row presents cases of backlit, resulting in reduced visibility and a relatively darker appearance of the person. The second row shows cases of well-lit, which is normal cases in real world.

3.2. Erroneous Detection & Correction

As presented in Sec. 4.2 in paper, we propose an error detection pipeline and then correct detected frames manually to deal with potential error from 2D pose estimator efficiently.

Given an estimated human pose, we feed it to a pre-trained conditional image generator, ControlNet [25] with Stable Diffusion v1.5 [20]. Original frame is also applied as a condition for synthesis by DDIM inversion [18]. Besides, scenario category, actions and brief description of actor in each session are input as text prompts. Then we use SAM [9] to process synthesized and original images and obtain binary human masks, using keypoints as prompts. If intersection-over-union (IoU) between the two masks is lower than pre-defined threshold, corresponding pose is classified as erroneous one and then will be correct by human annotator. Fig. 6 present results of correct and erroneous poses. We use implementation of Diffuser [23] for image generator and official implementation of SAM. Notably, as ControlNet expects poses in OpenPose format [4], we transfer COCO-format annotations to OpenPose and plot the skeleton image in corresponding color pattern. Neck in OpenPose skeleton is defined to be the mid-point between shoulder keypoints. Moreover, we use DeepDataSpace [1] as manual annotation tool which supports annotation by dragging keypoints.

4. Experiments

In this section, we present more details about experiments of benchmarks we set and provide further results of extensive experiments.

Figure 4. Examples of occlusion in FreeMan. Top row shows self-occlusion and the bottom row presents occlusions in human-object interactions.

4.1. Monocular 3D Human Pose Estimation

For training of HMR [8], we use Adam optimizer with fixed learning rate of 2.5×10^{-4}. Training processes are conducted on a single NVIDIA RTX-3090-24GB GPU with batch size of 128. Additionally, for training of PARE [10], we use Adam optimizer with fixed learning rate of 5.0×10^{-5} at the backbone and head of the network. Training processes are conducted on a single NVIDIA RTX-3090-24GB GPU with batchsize of 128. Only 2D and 3D keypoints in FreeMan are converted to the format of *HumanData* provided in mmhuman3D [5]. And for finetuning of PARE, we use Adam optimizer with fixed learning rate of 1.0×10^{-5}. The training process are conducted on a single NVIDIA RTX-3090-24GB GPU with batchsize of 128.

We conduct cross-dataset test on FreeMan, Human3.6M and HuMMan, the results of HMR on all test sets are shown in the Tab. 1. When testing on Human3.6M, in-domain test obtains the lowest error while models trained on FreeMan achieves the second place (192.19mm) and surpass model trained on Human3.6M (465.1mm) greatly. For HuMMan test set, model trained on FreeMan still achieves better performance than than on Human3.6M. Since some recent work [10] improves models' performance through dataset mixture, we further finetune the pre-trained PARE model, and results are shown in Tab. 2. It can be seen that the 3D HPE performance improvement by fintuning on FreeMan is still higher than HuMMan.

We believe that the improvements mentioned above are due to the diversity of the FreeMan dataset mentioned in dataset overview, which makes the model more robust and generalizable. At the same time, the noticeably higher MPJPE on the FreeMan test set compared to other test sets indicates that the FreeMan is also a challenging benchmark.

MPJPE/PA-MPJPE(mm) Train \ Test	H36M	HuMMan	FreeMan
H36M	98.62/59.17	392.89/175.94	350.97/178.85
HuMMan	465.1/224.53	-	413.26/218.28
FreeMan	192.19/112.7	302.09/147.67	148.22/100.56

Table 1. Cross-domain test results of HMR with the same supervision 2D&3D KPTs. MPJPE & PA-MPJPE are presented in unit of mm. Due to limited amount of data, all released part of HuMMan are used as training data.

Datasets	Supervision	Train	MPJPE	PA
HuMMan	2D + 3D KPTs + SMPL	FT	85.4	49.2
HuMMan	2D KPTs + SMPL	FT	78.9	49.4
FreeMan	2D + 3D KPTs+ SMPL	FT	76.5	**48.3**
FreeMan	2D KPTs + SMPL	FT	**76.1**	48.9

Table 2. Finetuning pre-trained PARE model on FreeMan and HuMMan. FreeMan can bring larger improvement compared with HuMMan.

4.2. 2D-to-3D Pose Lifting

For training data in 2D-to-3D pose lifting, we use Human3.6M data provided by VideoPose3D [19], 70% of released data from HuMMan and training split of FreeMan, respectively. Following the original setting in HuMMan [3], we split released data of 100 subjects into training and test set by subjects. For FreeMan, we select one view from every session and down-sample the videos to 15FPS, resulting in the frame number to be $350K$, which is similar to the amount of released part of HuMMan ($253K$) and much smaller than Human3.6M ($1500K$). Following [3], we unify the test set to be AIST++ [11] in order to verify the generalization across datasets. And test set of FreeMan are used for reference.

During training, coordinates of 2D keypoints are normalized by height and width of corresponding images. Ground

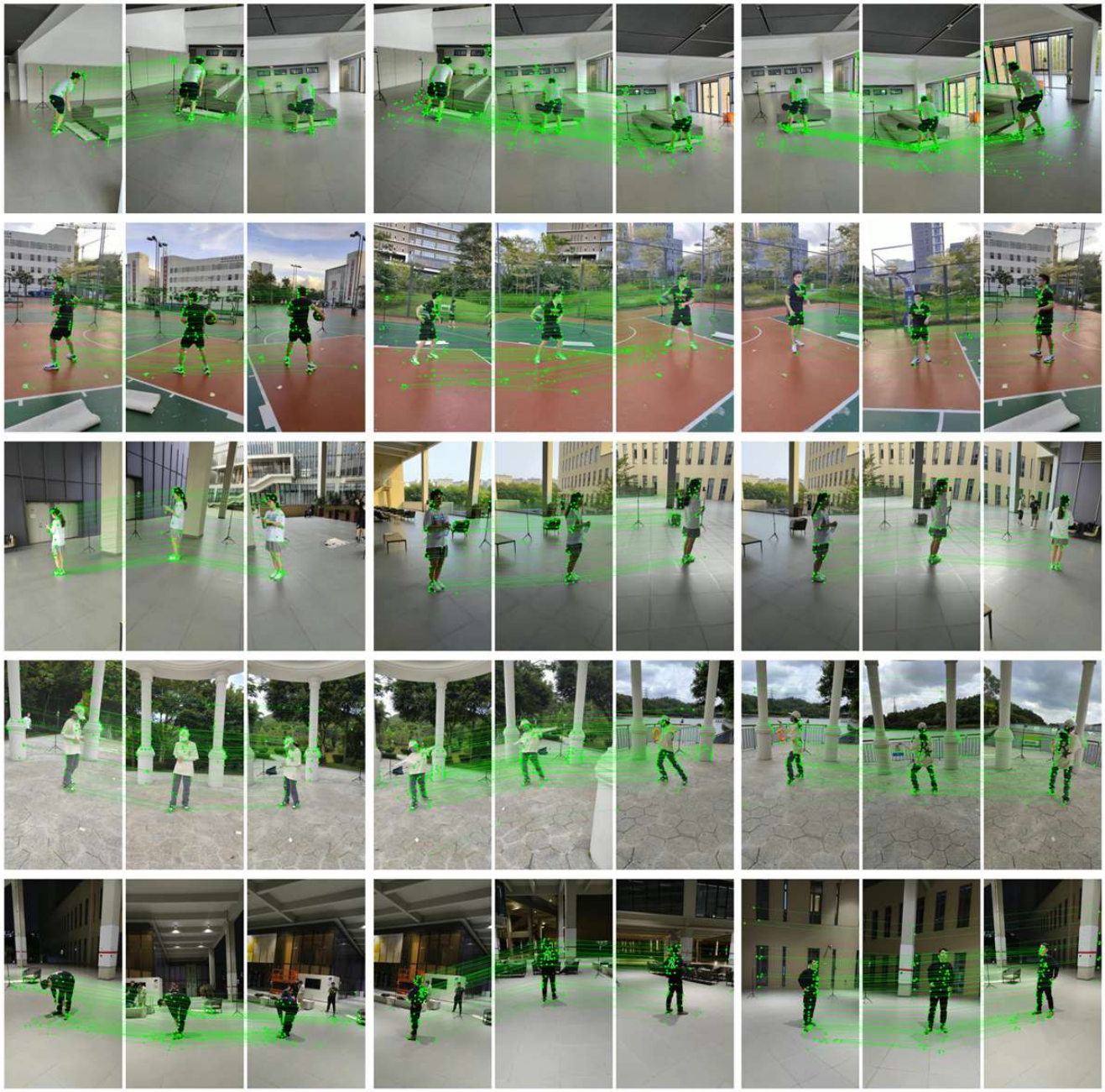


Figure 5. Example of calculated image correspondances of 5 scenes. Three adjacent viewpoints form one group, and three groups for each scenario .

truth 3D poses are transferred into camera coordinate system and root of skeleton is placed to origin. During test, since resolution of images are different among datasets, input 2D keypoints are normalized by resolution of test images. Keypoints in COCO format are mapped to that in Human3.6M format following mmHuman3D [5].

All models are optimized using Adam optimizer with learning rate of 10^{-4} on one NVIDIA RTX-3090-24GB. SimpleBaseline [16], VideoPose3D [19] are trained for 80 epochs with batch size of 1024 and PoseFormer [28], MHFormer [12], PoseFormerV2 [27] are trained for 25 epochs with batch size of 256 following their original settings. We show the results of VideoPose3D, PoseFormer and PoseFormerV2 in Tab. 3.

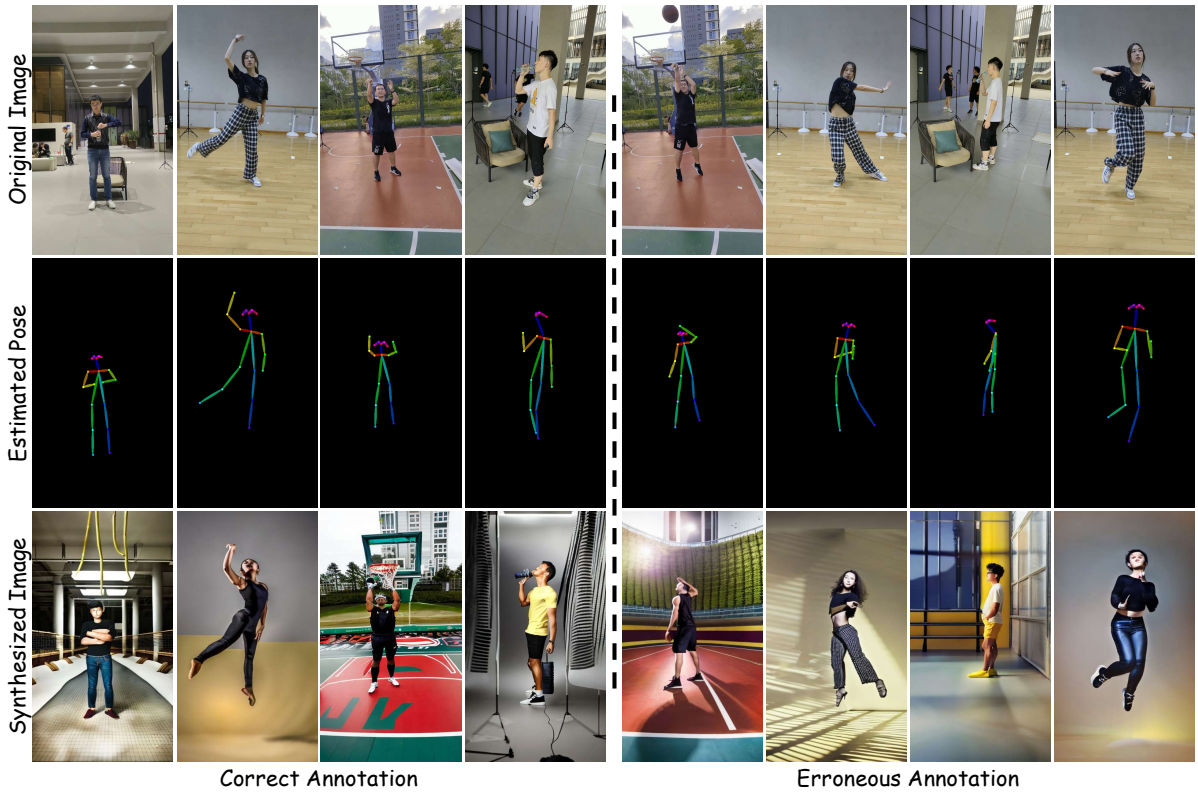


Figure 6. Results of erroneous pose detection. Original images, estimated pose by 2D pose estimator and synthesized images are presented in three rows, respectively. The left part shows cases of correct pose annotation, while the right part presents cases of erroneous poses. In the synthesized image, erroneous pose annotation usually results a completely different body part from the corresponding part in the original image, which means lower IoU (Intersection over Union) for the human body mask compared to the correct annotation.

4.3. Multi-view 3D Human Pose Estimation

In multi-view 3D human pose estimation, we use 4 views from both Human3.6M and FreeMan as input to VoxelPose. For Human3.6M, we follow the same processing steps as Transfusion [15]. For FreeMan, videos from odd-indexed views in training split are downsampled by 5 times to make data scale comparable. *Note single frame from all input views as one group*, Human3.6M and FreeMan include $223K$ and $132K$ groups of training data, respectively.

We first finetune ResNet-50 [6] backbone pre-trained on COCO with each dataset for 10 epochs, and then optimized the latter modules in decoder for additional 15 epochs. Both the two stages use Adam optimizer with a learning rate of $1e-4$ and batch size of 32. Models are trained on 4 NVIDIA A100-80GB GPUs. To solve the difference between joint definitions, we select 13 common joints between Human3.6M and COCO format, and then use the mid-points of the *left & right hips* and *left & right shoulders* to generate *mid-hip* and *neck*. In experiments, *mid-hip* is used as the root joint. The images are all cropped by human bounding boxes and then resized to make short edges the same.

In Tab. 4, we report recall and MPJPE@500mm of each experiment. In calculation of Recall@500mm, only predictions with MPJPE smaller than 500mm are treated as positive predictions and a higher recall value refers to higher successful rate to locate humans in space. And only positive predicted poses contribute to MPJPE in the final column.

Without ground truth root location, the model trained on Human3.6M is unable to locate human in cross-domain test and thus corresponding MPJPE is not available. Even though the ground truth root positions are given, recall value and MPJPE of model trained on Human3.6M are still 96.20% and 103.02mm, which is lower than that of model trained on FreeMan in cross-domain test without GT root (96.68% & 61.29mm), demonstrating that our training set has better transferability and test set is more challenging.

4.4. Neural Rendering of Human Subjects

4.4.1 Implementation Details

We use 128 samples per ray and train for $400K$ iterations with the Adam optimizer as the setting in [24]. Samely, to improve the quality of our results, we have increased the number of rays sampled for the foreground subject, as identified by the segmentation masks. We achieve this by

Algorithm	Train	Test	MPJPE (mm)↓	PA (mm)↓
VideoPose3D	FreeMan	FreeMan	88.68	49.17
	FreeMan†	FreeMan	73.98	45.22
	Human3.6M	AIST++	190.46	146.98
	HuMMan	AIST++	265.10	125.56
	FreeMan	AIST++	146.66	99.01↑21.15%
	FreeMan†	AIST++	141.84	94.59↑24.66%
PoseFormer	FreeMan	FreeMan	92.94	64.91
	FreeMan†	FreeMan	77.68	54.39
	Human3.6M	AIST++	179.54	151.38
	HuMMan	AIST++	158.13	96.98
	FreeMan	AIST++	133.39	90.10↑7.09%
	FreeMan†	AIST++	133.89	84.68↑14.52%
PoseFormerV2	FreeMan	FreeMan	92.11	64.91
	FreeMan†	FreeMan	90.81	55.98
	Human3.6M	AIST++	236.23	154.93
	HuMMan	AIST++	205.73	103.80
	FreeMan	AIST++	131.13	87.24↑15.95%
	FreeMan†	AIST++	113.89	80.61↑22.34%

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.

Train	Test	Recall@500mm (%) ↑	MPJPE (mm) ↓
Human3.6M	Human3.6M	100	25.95
Human3.6M	FreeMan	0.06	-
Human3.6M	FreeMan (w/ GT Root)	96.20	154.41
FreeMan	FreeMan	99.97	26.61
FreeMan	Human3.6M	96.68	62.37
FreeMan	Human3.6M (w/ GT Root)	100.00	58.30
FreeMan	FreeMan†	99.98	35.04

Table 4. Results of VoxelPose [22] for Multi-View 3D Pose Estimation. Recall@500mm refer to ratio of predictions with MPJPE smaller than 500mm, MPJPE here has not threshold for all keypoints. FreeMan† represents test set of even indexed cameras. Ground truth root position (GT Root) is not used if not specified. Rows highlighted shows the best setting in cross-domain test.

implementing a random ray sampling method that assigns a higher probability of 0.8 to foreground subject pixels and a lower probability of 0.2 to the background region. The resize scale of the image is set to 0.5. It takes about 48 hours to train on one NVIDIA RTX-3090-24GB for each one.

In order to ensure the quality of training, the number of frames of video clips in different scenes is in the interval of 300 to 1200 frames. The selected ten clips contain a variety of actions, ranging from slow and deliberate movements (such as warm-up exercises) to fast and energetic ones (such as dancing).

4.4.2 Visualization Results

We show the visualization results of the reconstruction of the selected videos in FreeMan dataset at Fig. 7. The above two lines reflect the results of relatively good reconstruction, while the following two lines reflect the results of relatively poor reconstruction. This indicates that FreeMan has sufficient diversity and challenges for human reconstruction.

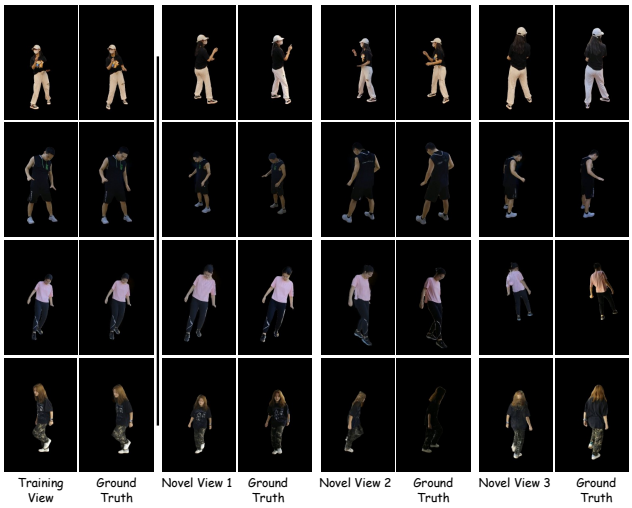


Figure 7. Rendering results for selected sessions of 30FPS. The lower two rows present cases of bad rendering results.

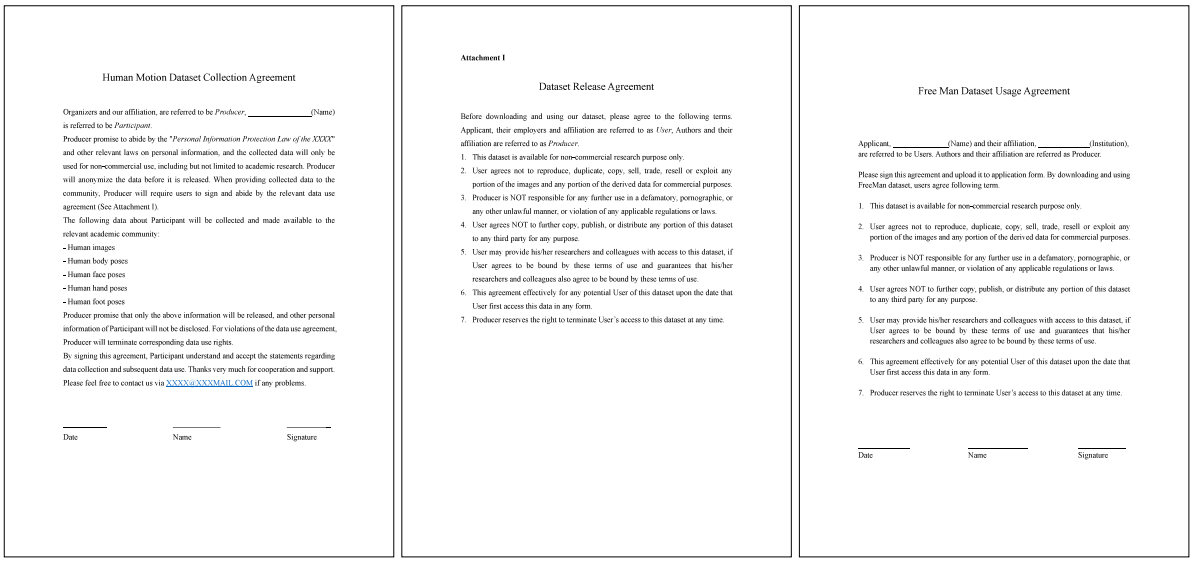
Figure 8. Example images of motion blur on human body in sessions of 30FPS. Image patches within bounding boxes at upper row are shown at lower row.

4.4.3 Experiments on data of 60FPS

Due to the occurrence of blur as Fig. 8 in body parts such as hands and feet when moving at high speeds, we collect videos at 60FPS to provide higher quality ground truth. We conduct experiments on two video clips from Park and Square scenes, and the experimental results are as Tab. 5. The results indicate that FreeMan remains highly challenging for human neural rendering in natural lighting conditions.

5. Dataset Documentations

FreeMan is available for academic communities to boost related researches. Example code to load FreeMan in py-



Human Motion Dataset Collection Agreement

Organizers and our affiliation, are referred to be *Producer*, _____(Name) is referred to be *Participant*.
Producer promise to abide by the "*Personal Information Protection Law of the XXXX*" and other relevant laws on personal information, and the collected data will only be used for non-commercial use, including but not limited to academic research. Producer will anonymize the data before it is released. When providing collected data to the community, Producer will require users to sign and abide by the relevant data use agreement (See Attachment I).
The following data about Participant will be collected and made available to the relevant academic community:
- Human images
- Human body poses
- Human face poses
- Human hand poses
- Human foot poses

Producer promise that only the above information will be released, and other personal information of Participant will not be disclosed. For violations of the data use agreement, Producer will terminate corresponding data use rights.
By signing this agreement, Participant understand and accept the statements regarding data collection and subsequent data use. Thanks very much for cooperation and support.
Please feel free to contact us via XXXX@XXXMAIL.COM if any problems.

_____ _____ _____
Date Name Signature

Attachment I

Dataset Release Agreement

Before downloading and using our dataset, please agree to the following terms. Applicant, their employers and affiliation are referred to as *User*, Authors and their affiliation are referred to as *Producer*.
1. This dataset is available for non-commercial research purpose only.
2. User agrees not to reproduce, duplicate, copy, sell, trade, resell or exploit any portion of the images and any portion of the derived data for commercial purposes.
3. Producer is NOT responsible for any further use in a defamatory, pornographic, or any other unlawful manner, or violation of any applicable regulations or laws.
4. User agrees NOT to further copy, publish, or distribute any portion of this dataset to any third party for any purpose.
5. User may provide his/her researchers and colleagues with access to this dataset, if User agrees to be bound by these terms of use and guarantees that his/her researchers and colleagues also agree to be bound by these terms of use.
6. This agreement effectively for any potential User of this dataset upon the date that User first access this data in any form.
7. Producer reserves the right to terminate User's access to this dataset at any time.

Free Man Dataset Usage Agreement

Applicant, _____(Name) and their affiliation, _____(Institution), are referred to be Users. Authors and their affiliation are referred as Producer.

Please sign this agreement and upload it to application form. By downloading and using FreeMan dataset, users agree following term.

1. This dataset is available for non-commercial research purpose only.
2. User agrees not to reproduce, duplicate, copy, sell, trade, resell or exploit any portion of the images and any portion of the derived data for commercial purposes.
3. Producer is NOT responsible for any further use in a defamatory, pornographic, or any other unlawful manner, or violation of any applicable regulations or laws.
4. User agrees NOT to further copy, publish, or distribute any portion of this dataset to any third party for any purpose.
5. User may provide his/her researchers and colleagues with access to this dataset, if User agrees to be bound by these terms of use and guarantees that his/her researchers and colleagues also agree to be bound by these terms of use.
6. This agreement effectively for any potential User of this dataset upon the date that User first access this data in any form.
7. Producer reserves the right to terminate User's access to this dataset at any time.

_____ _____ _____
Date Name Signature

Figure 9. **Left**: Data collection agreements for actors involved in FreeMan. **Middle**: Illustration of users' responsibility and usage agreement for actors. **Right**: Usage agreement for dataset users. Information collected will not be published are only for backup. Identity information are omitted.

Scene	PSNR↑	SSIM↑	LPIPS* ↓
Square	23.99	0.9389	88.10
Park	24.43	0.9527	61.84

Table 5. Neural rendering in 60FPS results by using HumanNeRF [24]. $\texttt{LPIPS}^* = \texttt{LPIPS} \times 10^3$.

torch is available on Github and data storage structure are illustrated in this section.

5.1. Data Format

FreeMan consists of multi-view human motion data and corresponding 2D / 3D human pose annotations. All data are separated into videos, camera parameters, bounding boxes and keypoints annotations basd on data type. For each session, human motion videos of all views are stored in format of $mp4$ and there are 8 synchronized videos for 8 views. Camera parameters, including image resolution, camera intrinsic parameters and camera extrinsic parameters, are saved in $JSON$ format.

Human keypoint annotations are encoded into format of npy, which is also known as numpy array. 2D poses of one session are stored with an array whose shape is $[V, F, J, 2]$, where V for view indexes, F for frame number, J for total number of joints and keypoint locations are given by (x, y) coordinates in unit of pixels. 3D poses are stored in an array with shape of $[F, J, 3]$, and 3D keypoint locations are provided by (x, y, z) in world coordinate system.

5.2. License & Ethical Impact

As FreeMan is constructed for research purpose only, FreeMan adopts license of CC BY-NC 4.0 (Non-Commercial use only). Furthermore, subsequent users who are granted access to the dataset are required to sign relevant usage agreements and provide backup information. This is done to safeguard the privacy and security of individuals associated with the FreeMan dataset and prevent data misuse.

All actors involved in FreeMan are recruited on basis of voluntary and well informed of data collection purpose. All volunteers signed a data collection agreement which declares project proposal and data to be released, which is shown in Fig. 9. To protect the privacy of all participants, we anonymized all data in the dataset by removing any personally identifiable information.

Major potential social impacts for human related research are about privacy leakage. We have taken several steps to protect the privacy and anonymity of the individuals involved in the dataset. All data subjects provided informed consent before participating, and we have anonymized all data to the best of our abilities to ensure the non-disclosure of any personal information of the actors. Furthermore, we understand the importance of data governance and the need to address potential misuse or unintended consequences. We encourage researchers and users of our dataset to handle the data responsibly, following ethical guidelines and respecting privacy considerations. Before getting access to our data, researchers will be required to sign an agreement to obey our license. We are committed to ongoing discussions and collaborations with experts in the field to address any concerns

and ensure that our work contributes positively to the research community while minimizing any potential negative social impact.

5.3. Maintenance Plan

To access FreeMan data, users are required to sign a dataset usage agreement that illustrates responsibilities and requirements. After submitting signed agreement and basic information via online forms, they can download FreeMan from dataset host. All users should abide the relevant data use agreement and use rights will be terminated for any violations of data use agreement. Application procedure requires applicants to submit their basic information for backup and our data are available on Huggingface and OpenDataLab for research community.

5.4. Agreements

Specifically, we present agreements for both actors and users in Fig. 9. The leftmost two pages are for actors in the project. The first page is to explain this project and show data type while the second page is to show how our data will be publiced and used.

The last page is required for users to sign before access FreeMan, explaining users' responsibility. Information is collected for backup purposes only.

References

[1] International Digital Economy Academy. Deepdataspace. https://github.com/IDEA-Research/deepdataspace, 2023. 3

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 1

[3] 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. 4

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[5] MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark. https://github.com/open-mmlab/mmhuman3d, 2021. 4, 5

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

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

[8] 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. 4

[9] 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. 3

[10] 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. 4

[11] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 4

[12] 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. 5

[13] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 2

[14] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 1

[15] Haoyu Ma, Liangjian Chen, Deying Kong, Zhe Wang, Xingwei Liu, Hao Tang, Xiangyi Yan, Yusheng Xie, Shih-Yao Lin, and Xiaohui Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. *arXiv preprint arXiv:2110.09554*, 2021. 6

[16] 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. 5

[17] 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

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[19] 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. 4, 5

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[21] 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

[22] 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. 7

[23] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 3

[24] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 1, 6, 8

[25] 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. 3

[26] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 1

[27] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886, 2023. 5

[28] 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. 5