FreeMan: Towards Benchmarking 3D Human Pose Estimation in the Wild Appendix

In this Appendix, we present more examples of comparisons between Human3.6M, HuMMan and FreeMan in Section. 1, including test results in the wild and source data. Section. 2 gives more statistics about subjects, scenarios and action sets. Then details of experiments in benchmark are shown in 4. Finally, usage and license of FreeMan are illustrated in Section. 5.

1 Dataset Comparisons

1.1 Results In the Wild

We present more comparison result of monocular models trained on different datasets in Figure 1. Test samples are from 3DPW [1], including indoor and outdoor scenes, and all actions are not included in training data. It can be seen that models trained on FreeMan has the best generalization ability for data in the wild.

1.2 Source Data

In Fig. 2, we provide more example images from different existing multi-view human datasets for extensive comparison. Most datasets are collected in dedicated laboratories, hence background are simple and fixed for a specific view. Only MuCo-3DHP and our dataset, FreeMan, consist of background in the wild. However, MuCo-3DHP only includes 4 scenes and 3 objects and is too small to support large-scale training, while our data is designed to cover 11M frames and 10 scenarios, which refer to 29 locations, and 40 subjects.

2 FreeMan Statistics

2.1 Subject Frame Numbers

As we presented in Sec. 3 of the main paper, our dataset consists of data of 40 subjects and we split dataset by subjects. In Fig. 3, we show frame number of each subject in the whole dataset and every subset. It can be seen that distributions of frame number with respect of subjects are in shape of long-tail.

2.2 Scenarios

Overall, we divide the 29 locations in our dataset into 10 groups based on semantics. To further demonstrate diversity of our data, we show three examples for each type of scenario in Fig. 4. For 6 scenarios at left, we present images from different views to show diverse background. Meanwhile, the other four scenarios refer to multiple locations and examples from a single view are shown. Furthermore, as shown in Fig. 5, FreeMan includes data collected in different time periods, making the environment more diverse and complex.

Preprint. Under review.

3 Toolchain

In this section, we present the detail of human pose generation toolchain in Algorithm. 1, which is presented in Sec. 4.2 of main paper, and the camera adjustment mechanism to deal with failure case of camera calibration in Algorithm. 2.

After Algorithm. 1, the initial annotation of re-projected 2D poses, 3D poses and SMPL annotations. Then the mean squared error between $\tilde{K}_{2D}$ and K_{2D} are calculated as re-projection error E_{rep} for each view and views with re-projection larger than threshold σ are considered as failed to calibrate. If more than 2 views out of all 8 views failed to calibrate, the session will be discarded. Otherwise, the 2D poses of uncalibrated views are ignored and triangulation process are repeated with poses from calibrated views only. Then camera parameters of uncalibrated ones are re-computed by 2D-3D keypoint pairs. And 2D poses are updated by re-projecting 3D poses. In practical implementation, keypoints threshold ϕ is set to 0.5 and re-projection error threshold σ is set to 50 pixels while the whole image is in the resolution of 1920×1080.

4 Experiments

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

4.0.1 Monocular 3D Human Pose Estimation

For training of HMR[7], 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[8], 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[9].

We also make cross-dataset test on FreeMan, Human3.6M and HuMMan, the results of HMR on all test sets are shown in the Tab. 1. Based on the above advantages in FreeMan, it makes existing models more generalizable across datasets. And FreeMan is still challenging in in-domain experiments.

4.0.2 2D-to-3D Pose Lifting

For training data in 2D-to-3D pose lifting, we use Human3.6M data provided by VideoPose3D[10], 70% of released data from HuMMan and training split of FreeMan, respectively. Following the

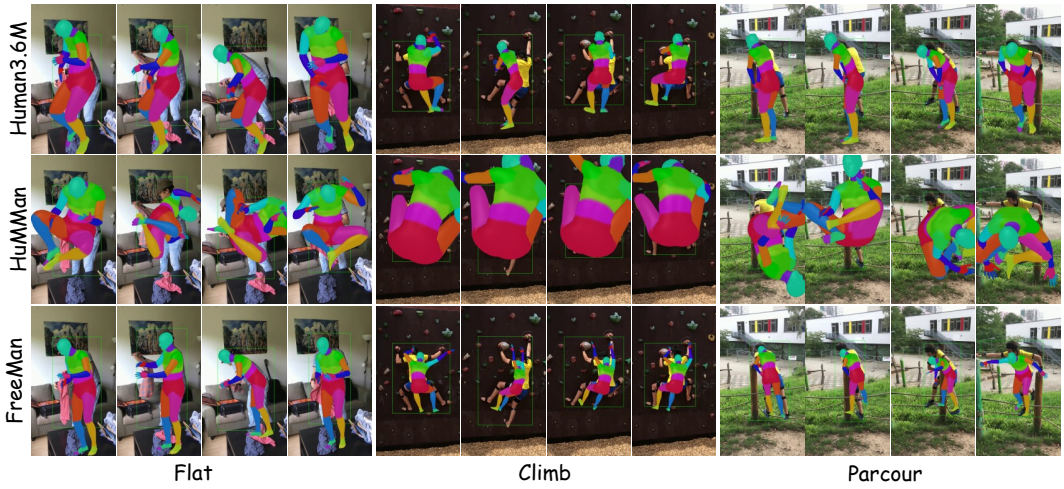

Figure 1: Zero-shot test result of monocular HMR in 3DPW [1] test set. Each row refer to results of models trained on corresponding datasets. $Flat$ refers to indoor scenes while $Climb$ and $Parcour$ are from outdoor scenes.

Figure 2: Comparison between our dataset and existing multi-view 3D human dataset. **MuCo refer to MuCo-3DHP[2], Panoptic refer to CMU Panoptic[3], H-Eva refer to HumanEva[4], H3.6M stands for Human3.6M[5], M-I-3D stands for MPI-INF-3DHP[6],**. All the images are resized to keep length of short edges the same without cropping. Only FreeMan includes more than 5 scenes.

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.

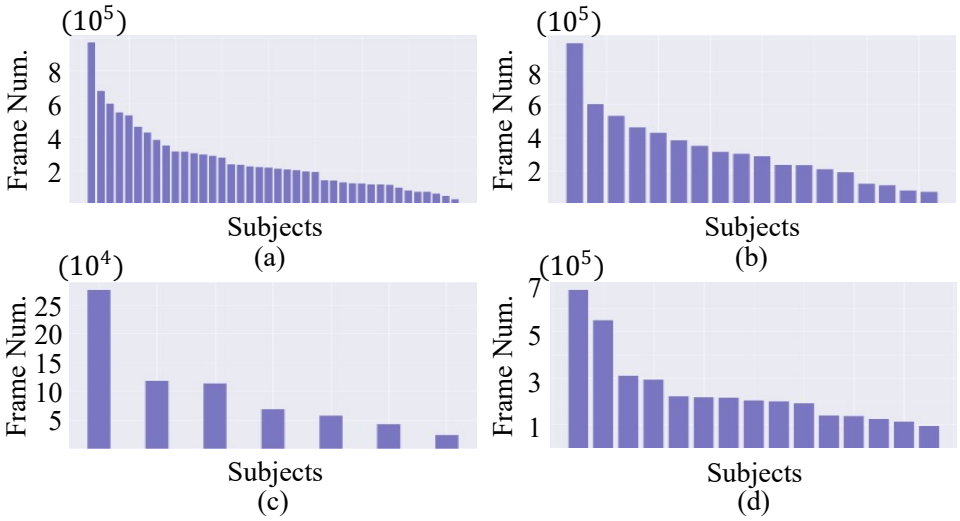

Figure 3: Frame number with respect of subjects. (a) Statistics of the whole dataset; (b) Statistics of train set; (c) Statistics of validation set; (d) Statistics of test set.

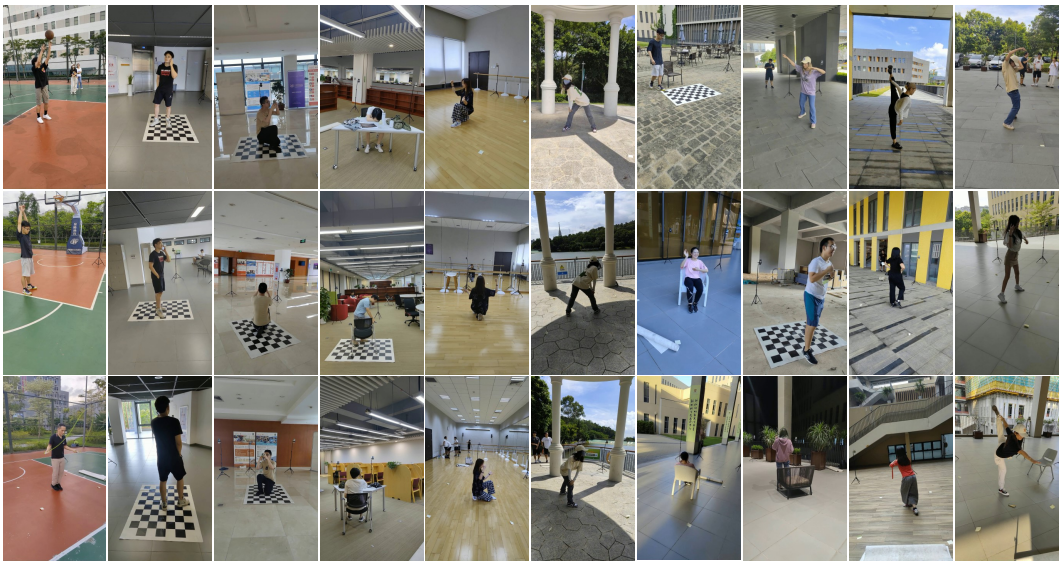

Figure 4: Example images of 10 kinds of scenarios that corresponds to 29 locations. For scenes shown in leftmost six columns, various of background from different views are presented. For outdoor scenes such as cafe, platform, courtyard and square, various locations are included.

original setting in HuMMan[11], we split released data of 100 subjects into training and test set by subjects. And test set of AIST++ and FreeMan are used for evaluation. During training, coordinates of 2D keypoints are normalized by height and width of corresponding images. Ground 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[9].

All models are optimized using Adam optimizer with learning rate of 10^{-4} on one NVIDIA RTX-3090-24GB. SimpleBaseline[12], VideoPose3D[10] are trained for 80 epochs with batch size of 1024 and PoseFormer[13], MHFormer[14] are trained for 25 epochs with batch size of 256 following their original settings. We show the results of VideoPose3D and PoseFormer in Tab. 3.

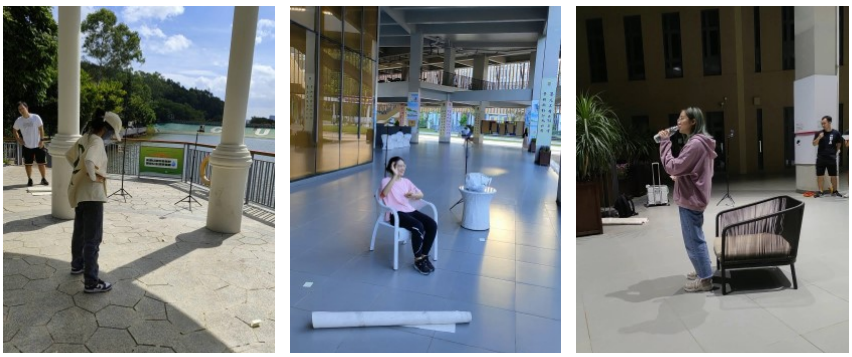

Figure 5: Example images of data collected at different time with various lighting condition.

Algorithm 1 Pose Annotation Pipeline

Require:
Multi-view human frames
$I_{Human} \in \mathbb{R}^{8\times N_{human}\times H\times W\times 3}$
Multi-view chessboard frames
$I_{board} \in \mathbb{R}^{8\times N_{board}\times H\times W\times 3}$
2D Human Pose Estimator $\mathcal{M}$
Keypoints threshold ϕ
Ensure:
Human frame number $N_{human} > 30$
Chessboard frame number $N_{board} > 30$
1: Estimate 2D human pose K_{2D} = $\mathcal{M}(I_{Human})$, $K_{2D} \in \mathbb{R}^{8\times N_{human}\times 17\times 2}$;
2: Estimate camera parameters for 8 cameras $\{C_i\}, i \in \{1,2,3,4,5,6,7,8\}$;
3: Filter 2D pose $\hat{K}_{2D}$ = $K_{2D}[score > \phi]$;
4: K_{3D} = Triangulation($\hat{K}_{2D}$, $\{C_i\}$);
5: $\tilde{K}_{3D}$ = Smooth(K_{3D});
6: M_{SMPL} = MeshFit($\tilde{K}_{3D}$);
7: Re-projected 2D pose $\tilde{K}_{2D}$ = Reprojection($\tilde{K}_{3D}$, $\{C_i\}$);
8: **return** 3D Pose $\tilde{K}_{3D} \in \mathbb{R}^{N_{human}\times 17\times 3}$, 2D Pose $\tilde{K}_{2D} \in \mathbb{R}^{N_{human}\times 17\times 2}$, SMPL Mesh M_{SMPL}, Camera Parameters $\{C_i\}$

4.0.3 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[16] 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. 2, we report recall and MPJPE of each experiment. Recall@500mm means that 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.

Algorithm 2 Camera Adjustment

Require:
Keypoints threshold ϕ
Reprojection error threshold σ

$E_{rep,i}\{\tilde{K}_{2D,i}, K_{2D,i}\} = MSE(\tilde{K}_{2D,i}, K_{2D,i}), i \in \{1, 2, ..., 8\}$
if $E_{rep,i} < \sigma$ for $i \in \{1, 2, 3, ..., 8\}$ **then**
    return $\tilde{K}_{3D}$, $\tilde{K}_{2D}$, C_i.
else
    Visualize & find calibrated C_T & Failed views C_F;
    if $len(C_F) > 2$ **then**
        return None
    else
        $\hat{K}_{2D}$ = $K_{2D,C_T}[score > \phi]$;
        K_{3D} = Triangulation($\hat{K}_{2D}$, C_T);
        $\tilde{K}_{3D}$ = Smooth(K_{3D});
        $\hat{C}_F$ = $argmin_{C_F}(E_{rep}\{K_{3D}, C_F, K_{2D}\})$;
        $\tilde{K}_{2D}$ = Reprojection($K_{3D}, \{C_T, \hat{C}_F\}$);
        return $\tilde{K}_{3D}$, $\tilde{K}_{2D}$, $\{C_T, \hat{C}_F\}$
    end if
end if

Train	Test	Recall@500mm (%)	MPJPE (mm) ↓
Human3.6M	Human3.6M	100	25.29
Human3.6M	FreeMan	0.00	-
Human3.6M	FreeMan (w/ GT Root)	96.20	103.02
FreeMan	FreeMan	99.97	26.07
FreeMan	Human3.6M	96.68	61.29
FreeMan	Human3.6M (w/ GT Root)	100.00	58.29
FreeMan	FreeMan†	99.98	35.04

Table 2: Results of VoxelPose[17] for Multi-View 3D Pose Estimation. Recall@500mm refer to ratio of predictions with MPJPE smaller than 500mm. 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.

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.

4.1 Neural Rendering of Human Subjects

4.1.1 Implementation Details

We use 128 samples per ray and train for $400K$ iterations with the Adam optimizer as the setting in [18]. 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 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).

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$\uparrow_{21.15\%}$
	FreeMan†	AIST++	141.84	94.59$\uparrow_{24.66\%}$
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$\uparrow_{7.09\%}$
	FreeMan†	AIST++	133.89	84.68$\uparrow_{14.52\%}$

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.

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

Table 4: Neural rendering in 60FPS results by using HumanNeRF[18]. Note that $\texttt{LPIPS}^* = \texttt{LPIPS} \times 10^3$.

4.1.2 Visualization Results

We show the visualization results of the reconstruction of the selected videos in FreeMan dataset at Fig. 6. 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 our FreeMan has sufficient diversity and challenges for NeRF reconstruction on human.

4.1.3 Experiments on 60FPS

Due to the occurrence of blur as Fig. 7 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. 4. The results indicate that FreeMan remains highly challenging for human neural rendering in natural lighting conditions.

5 Dataset Usage

FreeMan is available for academic communities to boost related researches. Dataloader to load FreeMan in pytorch is open sourced at `https://github.com/wangjiongw/FreeMan` and data storage structure are illustrated in this section.

5.1 Data Format

Overall, FreeMan provided multi-view human motion data and corresponding 2D / 3D human pose annotations. All data are separated by data type. For each session, human motion videos from 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 poses, are given 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

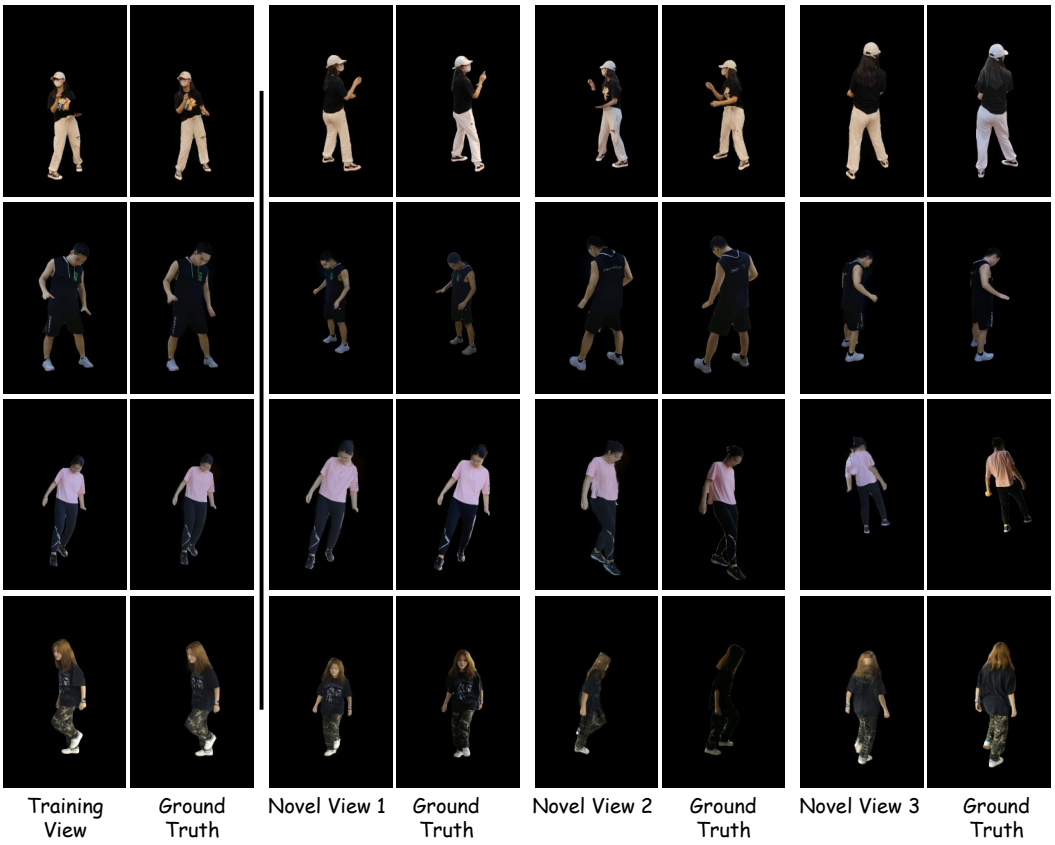

Figure 6: Rendering results in 30FPS dataset.

Figure 7: Example images of motion blur on human body in 30FPS dataset.

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).

5.2 Statement of responsibility

All actors involved in FreeMan are recruited on basis of voluntary and well informed of data collection purpose. As FreeMan is contructed for research purpose only, personal information are disclosed. FreeMan adopts lisence of CC BY-NC (allowing only non-commercial use).

To use the released data, users will be required to sign a data use agreement. After submitting signed agreement and purpose via an online form, then they can download FreeMan by provided sharing link. All users should abide the relevant data use agreement and use rights will be terminated for any violations of data use agreement.

References

[1] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[2] 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, sep 2018.

[3] 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.

[4] 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.

[5] 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, jul 2014.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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. October 2022.

[12] 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.

[13] 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.

[14] 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.

[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.

[16] 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.

[17] 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.

[18] 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.