Supplementary Material for
“FLAG3D: A 3D Fitness Activity Dataset with Language Instruction”

Yansong Tang[*,†,1], Jinpeng Liu[*,1], Aoyang Liu[*,1],
Bin Yang[1], Wenxun Dai[1], Yongming Rao[2], Jiwen Lu[⋄,2], Jie Zhou[2], Xiu Li[⋄,1]
[*]equal contribution, [†]project lead, [⋄]corresponding authors
{[1]Shenzhen International Graduate School, [2]Department of Automation}, Tsinghua University

1. Comparision With More Datasets

Owing to the page limitation, we did not display numerous relevant datasets, especially for some image-based datasets. Thus, in this section, we provide a more comprehensive comparison for FLAG3D in Table 1.

2. FLAG3D Dataset

Figure 4 displays more examples from our dataset. From left to right, the MoCap data, rendered RGB image, and RGB image with SMPL fitting are shown, respectively. From top to bottom, we demonstrate several frames from FLAG3D Dataset. They represent action *“Svend Press”*, *“Standing Straight-arm Chest Press With Resistance Band ”*,*“Jumping Jacks”*,*“Kneeling Left Knee Lift”*,*“Left-side Knee Raise And Abdominal Muscles Contract ”* and *“Prone Press Up With Torso Rotation”* respectively.

3. Labeling System

There are 60 kinds of fitness actions in FLAG3D. Selected activities exercise most parts of our body, including the chest, back, shoulder, arm, neck, abdomen, waist, hip, and leg. For example, there are six kinds of actions: *“Svend Press”*, *“Keeling Push-ups”*,*“Standing Straight-arm Chest Press With Resistance Band”*, *“Small Dumbell Floor Flies”*, *“Chest Fly”*, and *“ Push-ups that exercise chest muscles”*. As shown in Table 4, we label each action from A001 to A060.

4. Additional Details of Data Collection

From right to left in Figure 1, each part of the bottom branch is point cloud data, rigid body constraints data based on the point cloud data, skeleton data, and skin data. This showes the whole data process at the software level in the MoCap system. The point cloud data and the skeleton data are involved in our dataset to implement various experiments. The rigid body constraints data are the intermediate products to produce the skeleton data. And the skin data displays a real-time capture effect to help us track possible problems in the capture process.

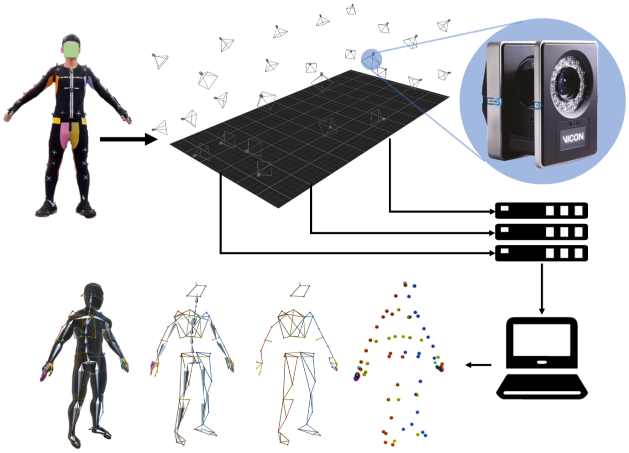


Figure 1. An overview of the MoCap system. From right to left, each part of the bottom branch is point cloud data, rigid body constraints data based on point cloud data, skeleton data, and skin data. When volunteers perform actions in the motion capture field, infrared cameras transmit the high frame rate IR grayscale images to the data switch. At the same time, monitors check whether the data is captured properly by the dynamic data displayed on the device. The point cloud data and the skeleton data are involved in our dataset to implement various experiments. The rigid body constraints data were the intermediate products to produce the skeleton data. And the skin data displayed a real-time capture effect to help us track possible problems in the capture process.

4.1. MoCap Marker

Volunteers wear special MoCap apparel. MoCap costumes have dense markers to ensure that all parts of the body are accurately recorded. Marker numbers for body parts are listed in Table 2. Totally, there are 77 markers on the volunteer body to ensure that our dataset could provide accurate 3D skeleton information.

Table 1. Comparisons of FLAG3D with the various datasets. *Type* represents the dataset consists of images or videos. Other abbreviations are the same as body part.

Dataset	Subjs	Cats	Seqs	Frames	Type	LA	K3D	SMPL	Resource	Task
UCF101 [22]	-	101	13K	>2M	Video	×	×	×	Nat.	HAR
Penn Action [29]	-	15	2326	-	Video	×	×	×	Nat.	HAR,HPR
MPII [3]	-	410	-	24K	Image	×	✓	-	Nat.	HAR,HPE
COCO [14]	-	-	-	104k	Image	×	×	×	Nat.	HPE
AMASS [17]	>300	-	11K	16M	Video	×	✓	✓	Lab.	HMR
AGORA [19]	>350	-	-	17K	Image	×	×	✓	Syn.	HMR
SURREAL [25]	145	-	2K	6M	Video	×	✓	✓	Nat.+Syn.	HMR
HUMBI [28]	>700	-	-	26M	Video	×	✓	✓	Lab.	HMR
THuman [31]	200	-	-	28K	Image	×	✓	✓	Syn.	HMR
PoseTrack [2]	-	-	550	66K	Video	×	×	×	Nat.	HPE
Human3.6M [9]	11	17	839	3.6M	Video	×	✓	-	Lab	HAR,HPE,HMR
CMU Panoptic [11]	8	5	65	594K	Video	×	✓	-	Lab	HPE
MPI-INF-3DHP [18]	8	8	-	>1.3M	Video	×	✓	-	Lab+Nat.	HPE,HMR
3DPW [27]	7	-	60	51k	Video	×	×	✓	Nat.	HMR
ZJU-MoCap [20]	6	6	9	>1k	Video	×	✓	✓	Lab	HAR,HMR
NTU RGB+D 120 [15]	106	120	114k	-	Video	×	✓	-	Lab	HAR,HAG
HuMMan [4]	1000	500	400K	60M	Video	×	✓	✓	Lab	HAR,HMR
HumanML3D [7]	-	-	14K	-	-	✓	✓	✓	Lab	HAG
KIT Motion Language [21]	111	-	3911	-	-	✓	✓	-	Lab	HAG
HumanAct12 [8]	12	12	1191	90K	Video	×	×	✓	Lab	HAG
UESTC [10]	118	40	25K	> 5M	Video	×	✓	-	Lab	HAR,HAG
Fit3D [6]	13	37	-	> 3M	Video	×	✓	✓	Lab	HPE,RAC
EC3D [30]	4	3	362	-	Video	×	✓	-	Lab	HAR
Yoga-82 [26]	-	82	-	29K	Image	×	×	×	Nat.	HAR,HPE
FLAG3D (Ours)	10+10+4	60	180K	20M	Video	✓	✓	✓	Lab+Syn.+Nat.	HAR,HMR,HAG

4.2. Hardware

We use the VICON [1] optical motion capture system to capture the character's motion, which consists of 24 VICON V16 cameras. Our system works with Vicon's powerful software platforms to enable a light touch while we are running a capture session. This MoCap system is highly robust, it can perform bump detection while capturing motions. Bump detection means that the system knows when a camera needs recalibrating and auto-heals without operator input. MoCap system's technology IP advances allow Vantage+ to hit resolution and speed sweet spots. Specific parameters of the camera are shown in Table 3.

Table 2. Marker numbers for body parts. In order to capture fine-grained human movements, we add sufficient markers on all parts of the human body.

Body Parts	Markers	Body Parts	Markers
Head	5	Finger	10 × 2
Waist	6	Thorax and back	12
Arm	4 × 2	Wrist and Palm	4 × 2
Leg	4 × 2	Ankle and Foot	5 × 2
Total	**77**		

Table 3. Hardware Parameters. Cameras used in this system have a maximum resolution of 4096×4096. It is capable of 120fps while maintaining maximum resolution sampling.

	Vantage+ Normal Mode
Model	V16
Resolution (MP)	16
Max Frame Rate (Hz)	120 @ 16MP
Max Frame Rate (Hz)	2000
On-Board Marker Processing	Yes
Standard Lens	18 mm
Wide Lens	12.5 mm
Minimum Standard FOV (H x V)°	54.7 × 54.7
Minimum Wide FOV (H x V)°	76.4 × 76.4
Camera Latency	8.3 ms
Strobe	IR
Shutter Type	Global
Connection Type	Cat5e / RJ45
Power	PoE+
Max Power Consumption	24W
Dimensions (mm) (H x W x D)	166.2 × 125 × 134.1
Weight (kg)	1.6
Updateable Firmware	Yes

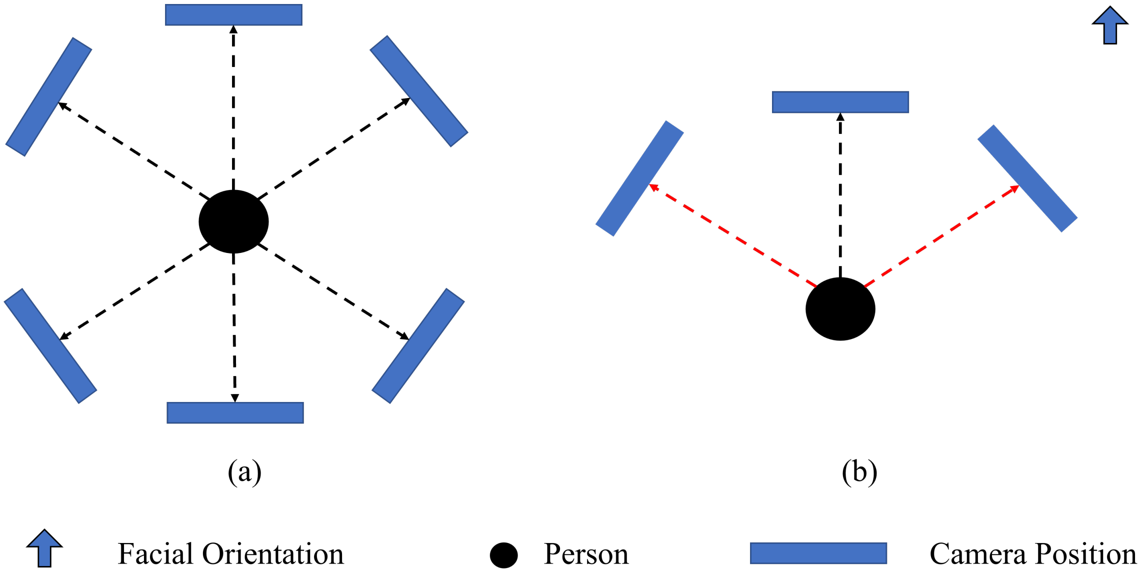


Figure 2. Camera Position Setting. In rendering software, the camera positions are settled as shown in figure (a). The avatar is surrounded by six cameras in a circle to allow for multi-angle video. In the nature scene, we selected two camera positions as shown in figure (b). The red dotted line represents a random selection of one of these perspectives.

4.3. Camera Position

In Figure 2, we show the camera settings for rendered videos and natural videos. We put 6 cameras around the avatars to obtain the multi-view videos and Figure 2(a) displays the arrangement. And for natural videos, we require volunteers to shoot their videos from two views. One is the front view and another is any front side view as shown in Figure 2(b). The red dotted line represents we only need one camera in the side view.

5. Details of Experiments

5.1. Action Recoginition

For in-domain experiments, we employ an Adam [12] optimizer for 30 epochs using a cosine decay learning rate. A batch size of 64 is used. Specifically, the initial learning rate and weight decay are set to 0.1 and 5e-4, respectively. For out-domain experiments, we use a larger batch size of 128. The learning rate, weight decay, optimizer and scheduler are the same as that for in-domain experiments. Unless otherwise mentioned, for all models, we sample 500 frames uniformly from the entire clip and only adopt the joint stream. We take the 10-clip testing results to report our Top-1 accuracy.

5.2. Human Mesh Recovery

For the optimization process to obtain the SMPL mesh, the loss weights λ_1, λ_2, λ_3, λ_4 in the objective function are $1, 5 \times 10^{-3}, 1, 1 \times 10^{-3}$ respectively. We verify the SMPL parameter registration process using the MPJPE and PA-MPJPE and their values are 56.09 and 27.02 respectively.

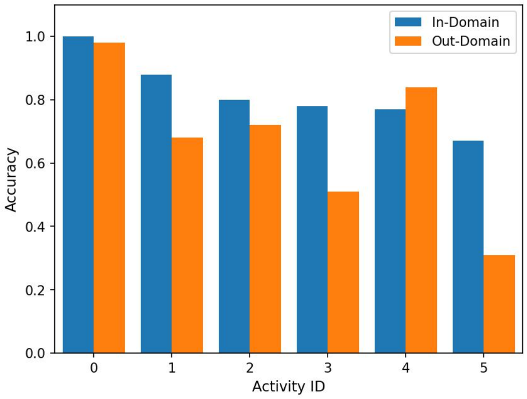


Figure 3. Per-class performances of 2s-AGCN under the in-domain and out-domain scenarios. 0: Wrist Joint Warm-up; 1: Shoulder Bridge; 2: Prone Press Up With Torso Rotation; 3: Lying Shoulder Joint Downward Round; 4: Kneeling Right Arm Raise; 5: Bent-over Dumbbell Tricep Extension.

We used the pre-trained HR-Net [23] backbone to fine-tune the ROMP method on FLAG3D Dataset. We employ an Adam [12] optimizer. An initial learning rate of 5e-5 and a batch size of 64 are used. We use the 3D SMPL joints [16], 2D COCO-17 keypoints [14] obtained from HR-Net [23] method and shape parameters to supervise the training.

6. Analysis of Evaluation

As shown in Figure 3, the Top-1 accuracy of 2s-AGCN on *"Bent-over Dumbbell Tricep Extension"* and *"Lying Shoulder Joint Downward Round"* are inferior compared with other categories, the most confusing classes for *"Bent-over Dumbbell Tricep Extension"* are *"Bent-over W-shape Stretch"*, *"Bent-over Y-shape Stretch"* and *"Right-side Bent-over Tricep Extension With Resistance Band"*.

A qualitative comparison of different methods for human mesh recovery on FLAG3D is shown in Figure 5. As we can see in this picture when the human is lying or kneeling the result will be dissatisfactory. We have mentioned this phenomenon in the body part. In this section, we will explore more about it. The first question is, are these results caused by the reason that the person is too tiny in the picture to recognize or the reason that the action is too complex to recover an accurate pose? The answer is both but the latter more matters. We crop the original picture into 1/2 and 1/4 to make a larger character in the image and then evaluate methods on it. In Figure 5, we showed the inference results from different methods on three cropped pictures. The results indicate that for the original image with the tiny person the challenge lies in recognizing the right person in the picture but for the cropped image because the person is no longer tiny, some algorithms could recognize the person in the right location which is an improvement as we find algorithm couldn't recognize the person in many cases with default threshold 0.2 and we often have to reduce it to let the algorithm work but the side effect is the method will often misrecognize the background as a person or recognize the person which is not in the proper location(see examples in Figure 5 row 4). And for the cropped pictures, the challenge is to understand the right posture in this case. But there is still a large distance to the acceptable result. From Figure 5, BEV [24] and ROMP [5] indeed estimate the lying posture of one body but can't infer the poses of the limbs accurately. Even if the algorithm could recognize the right person in the right location, it is also very challenging to recover some complex activities.

The second question is, What kind of action is difficult to recover? So we explore the effects of different movements on human mesh recovery based on FLAG3D. We compare the results trying to estimate three activities in Figure 5. To control the unrelevant variable, we choose three actions from the same person, same scene and same view. From left to right, the Figure 5 displays the results from standing to lying. We find as people's center of gravity dropping, the task is more challenging because the visible part of the whole body is less and less. From the visual results of the experiment, standing is easy to recover pose than kneeling and kneeling is easy than lying. The character's size in column 2 and column 3 is approximately the same but the methods' performance on these are distinct, which confirms that complex activities are the main challenge for this task.

Another reason that these methods can't deal with these cases is that they never learn about them because past datasets for training hardly include these actions about fitting. So our dataset could serve as a difficult and long-term benchmark for human mesh recovery task.

References

[1] Vicon. https://www.vicon.com/hardware/cameras. 2

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, pages 5167–5176, 2018. 2

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2

[4] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. *arXiv preprint arXiv:2204.13686*, 2022. 2

[5] Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, and Mustafa Mukadam. Revitalizing optimization for 3d human pose and shape estimation: a sparse constrained formulation. In *ICCV*, pages 11457–11466, 2021. 4, 7, 8

[6] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *CVPR*, pages 9919–9928, 2021. 2

[7] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 2

[8] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, pages 2021–2029, 2020. 2

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 2

[10] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. *arXiv preprint arXiv:1904.10681*, 2019. 2

[11] 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*, pages 3334–3342, 2015. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[13] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 7, 8

Table 4. Action comparison table of FLAG3D. In total, there are 60 kinds of actions. Selected activities exercise most parts of the body, including the chest, back, shoulder, arm, neck, abdomen, waist, hip, and leg.

Chest		**Abdomen**	
A001	Svend Press	A030	Lying Alternate Upper-half Leg Raise
A002	Kneeling Push-ups	A031	Half Roll Back
A003	Standing Straight-arm Chest Press With Resistance Band	A032	Kneeling Right-side Torso Twist
A004	Small Dumbbell Floor Flies	A033	Left-side Knee Raise And Abdominal Muscles Contract
A005	Chest Fly	A034	Kneeling Right Leg Backward Stretch
A006	Push-ups	A035	Kneeling Right Arm Raise
Back		**Waist**	
A007	Bent-over W-shape Stretch	A036	Shoulder Bridge
A008	Bent-over Y-shape Stretch	A037	Prone Press-ups
A009	Bent-over A-shape Stretch	A038	Bent-over Torso Rotation
A010	Squat With Arm Lift	A039	Lying Arm Pull
A011	Breaststroke Arm Pull	A040	Prone Press Up With Torso Rotation
A012	Prone Y-shape Stretch	A041	Breaststroke Push-ups
		A042	Sit-ups
Shoulder		**Hip**	
A013	Lateral Raise Forward Circles	A043	Side-lying Left Leg Forward Raise
A014	Lateral Raise Backward Circles	A044	Right Leg Reverse Lunge
A015	Lying Shoulder Joint Upward Round	A045	Kneeling Left Knee Lift
A016	Lying Shoulder Joint Downward Round	A046	Alternate Reverse Lunge
A017	Bare-handed Full Lateral Raise	A047	Side-lying Right Leg Backward Kick
A018	Bare-handed Cuban Press	A048	Left Leg Lunge With Knee Lift
A019	Fortune Cat	A049	Sumo Squat
Arm		**Leg**	
A020	Dumbbell Curls	A050	Straight Leg Calf Raise
A021	Alternate Dumbbell Curls	A051	Squat Jump
A022	Right-side Kettlebell Bent-over Row	A052	Squat With Alternate Knee Lift
A023	Wrist Joint Warm-up	A053	Standing Alternate Butt Kick
A024	Right-side Bent-over Tricep Extension With Resistance Band	A054	Knee Warm-up
A025	Bent-over Dumbbell Tricep Extension		
Neck		**Whole body**	
A026	Nod And Raise Head	A055	Butt Kicks
A027	Two-way Head Turn	A056	Jump Left and Right
A028	Shrug And Sink The Shoulders	A057	Jumping Jacks
A029	Four-way Nod Head	A058	High Knee
		A059	Clap Jacks
		A060	Run In Place With Arm Swing

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2, 3

[15] 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. *TPAMI*, 42(10):2684–2701, 2019. 2

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3

[17] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 2

[18] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516, 2017. 2

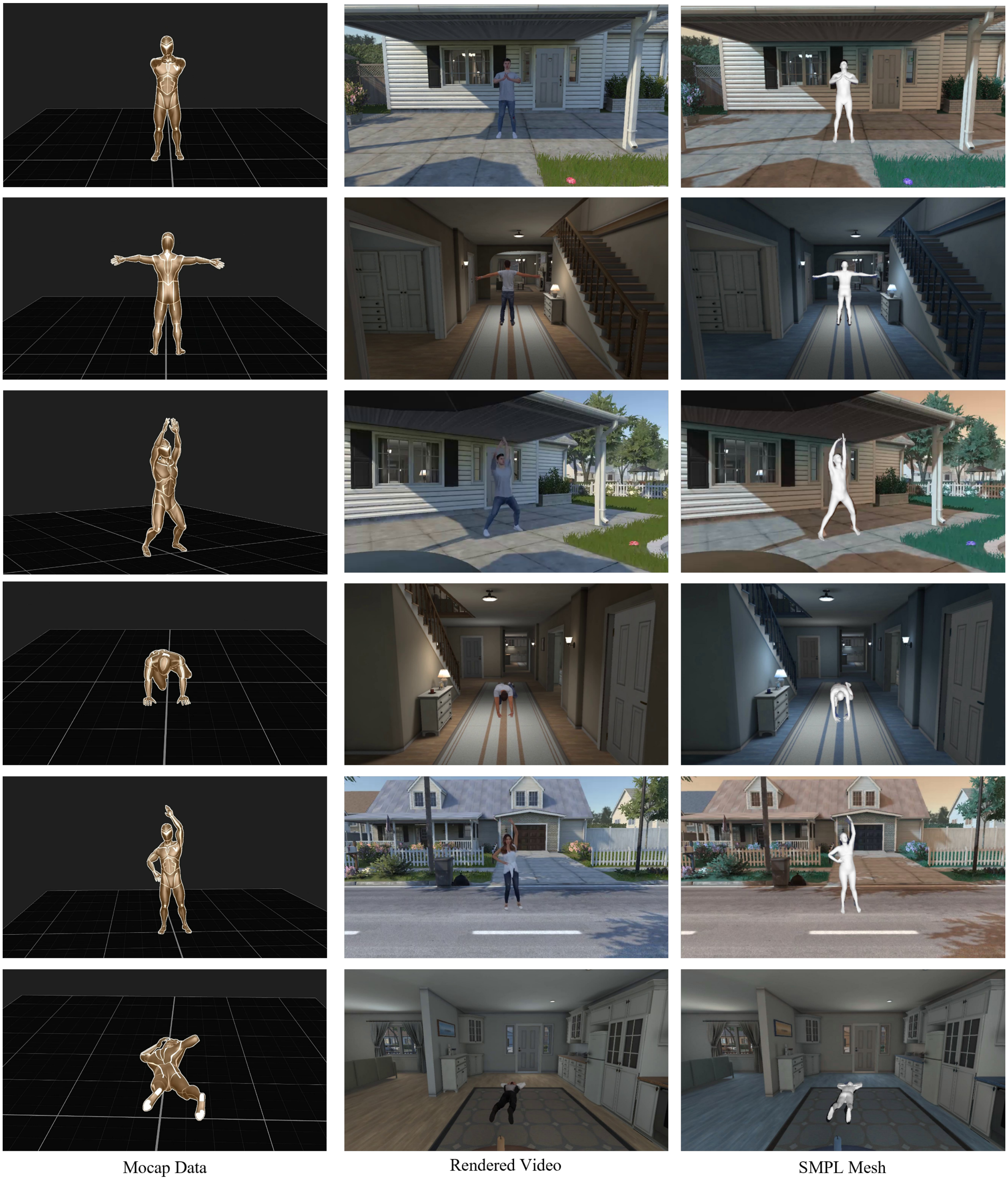


Figure 4. Several frames from FLAG3D Dataset. From left to right we display the MoCap data, Original Rendered RGB Videos, and Rendered RGB Videos with SMPL Mesh fitting results.

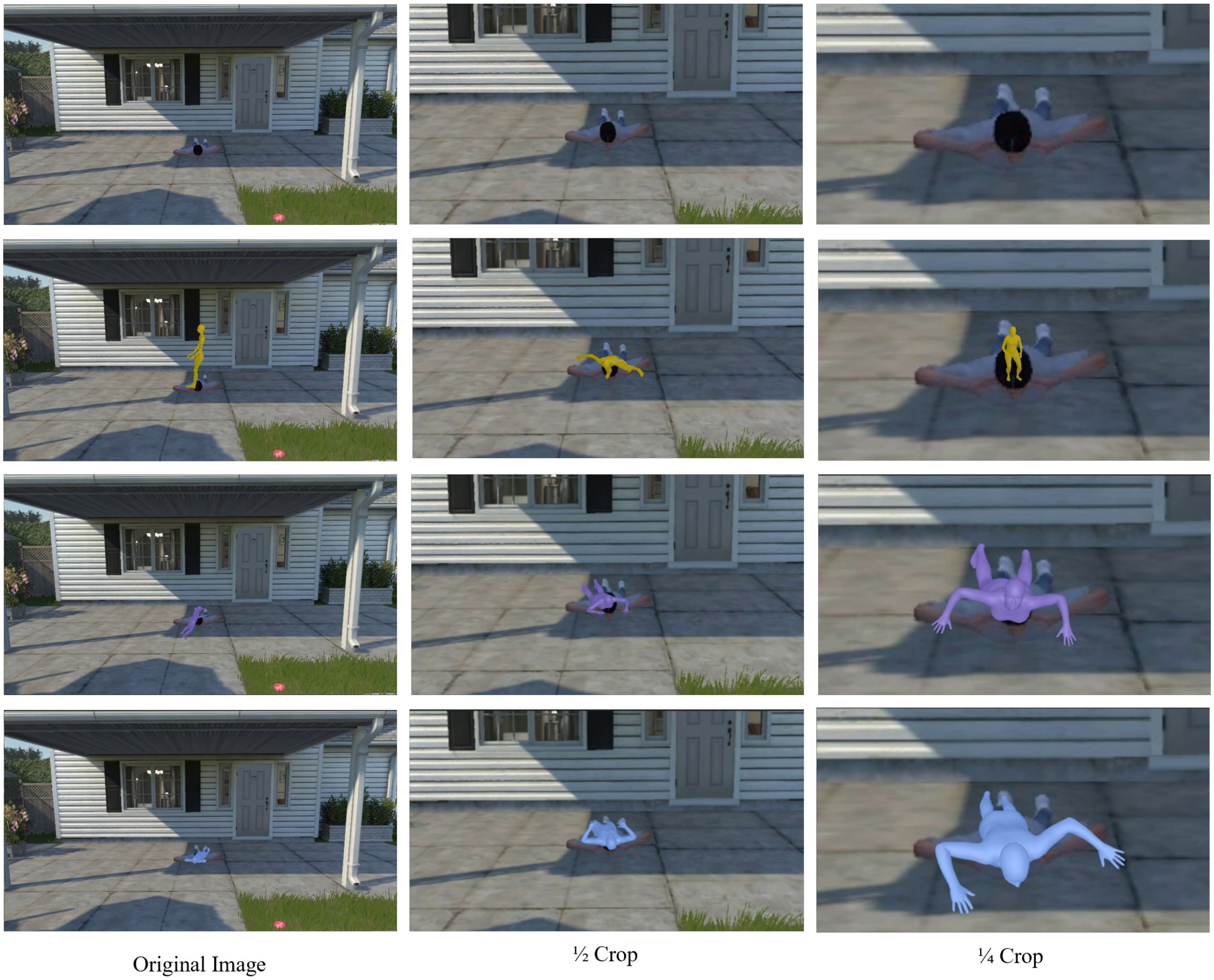


Figure 5. Results from different methods on three cropped pictures. From top to bottom: Original Picture(row 1), VIBE [13](row 2), ROMP [5](row 3), BEV [24] (row 4). From left to right: Original Image, Cropped Image 1(down to a half of the original), Cropped Image 2(down to 1/4 of the original).

[19] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, 2021. 2

[20] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pages 9054–9063, 2021. 2

[21] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2

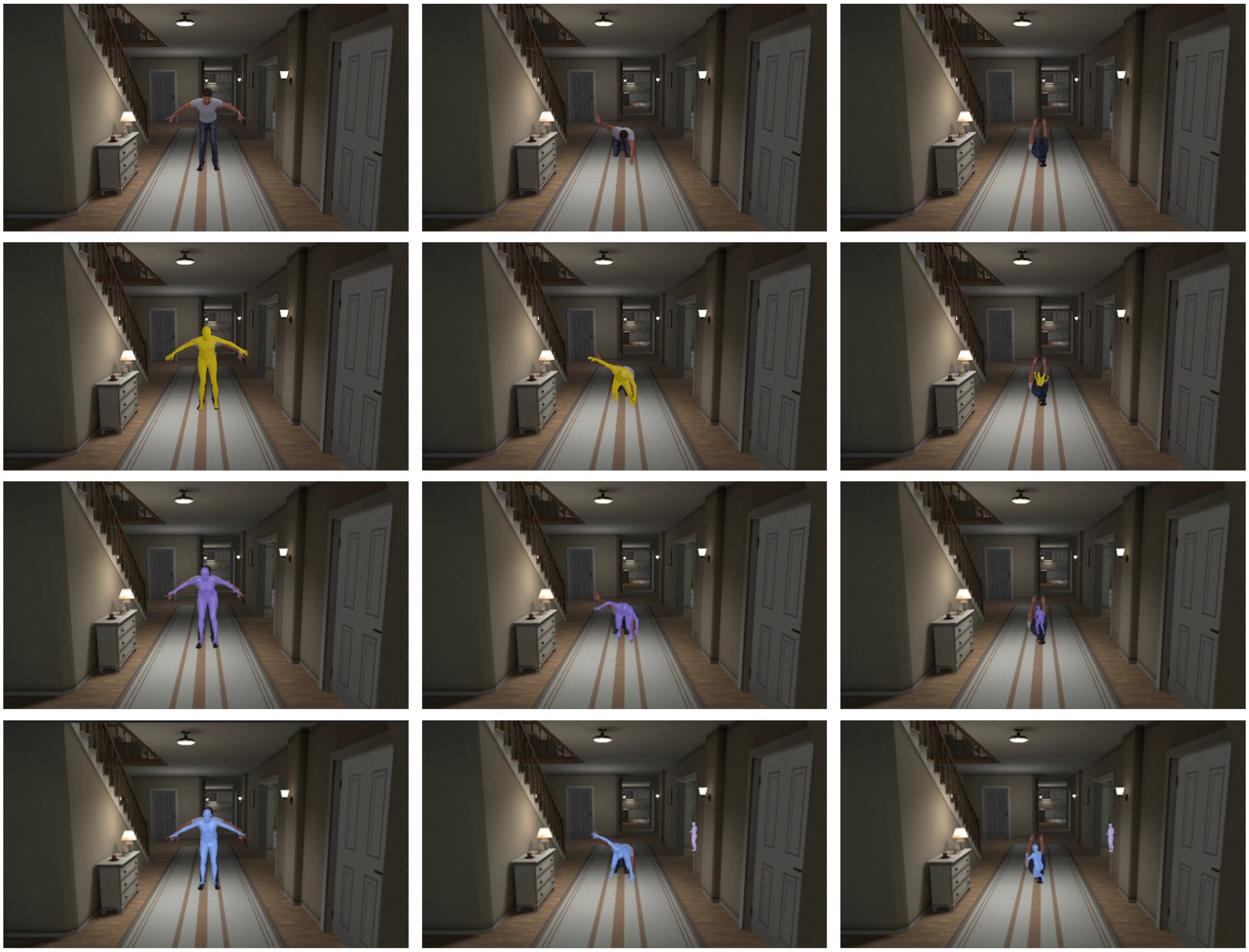
[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 3

[24] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252, 2022. 4, 7, 8

[25] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, 2017. 2

[26] Manisha Verma, Sudhakar Kumawat, Yuta Nakashima, and Shanmuganathan Raman. Yoga-82: a new dataset for fine-grained classification of human poses. In *CVPRW*, pages 1038–1039, 2020. 2



Standing Kneeling Lying

Figure 6. From left to right the activities are “*Bent-over W-shape Stretch*”,“*Kneeling Right Arm Raise*” and “*Lying Shoulder Joint Upward Round*” respectively. For actions in column 1, each method could estimate the shape and pose properly. For actions in column 2, there are some biases for recovering the posture. And for actions in column 3, they all can’t handle situations. From top to bottom:Original Picture(row 1), VIBE [13](row 2), ROMP [5](row 3), BEV [24] (row 4).

[27] 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 *ECCV*, pages 601–617, 2018. 2

[28] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *CVPR*, pages 2990–3000, 2020. 2

[29] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE ICCV*, pages 2248–2255, 2013. 2

[30] Ziyi Zhao, Sena Kiciroglu, Hugues Vinzant, Yuan Cheng, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. 3d pose based feedback for physical exercises. *arXiv preprint arXiv:2208.03257*, 2022. 2

[31] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *The IEEE ICCV*, October 2019. 2