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

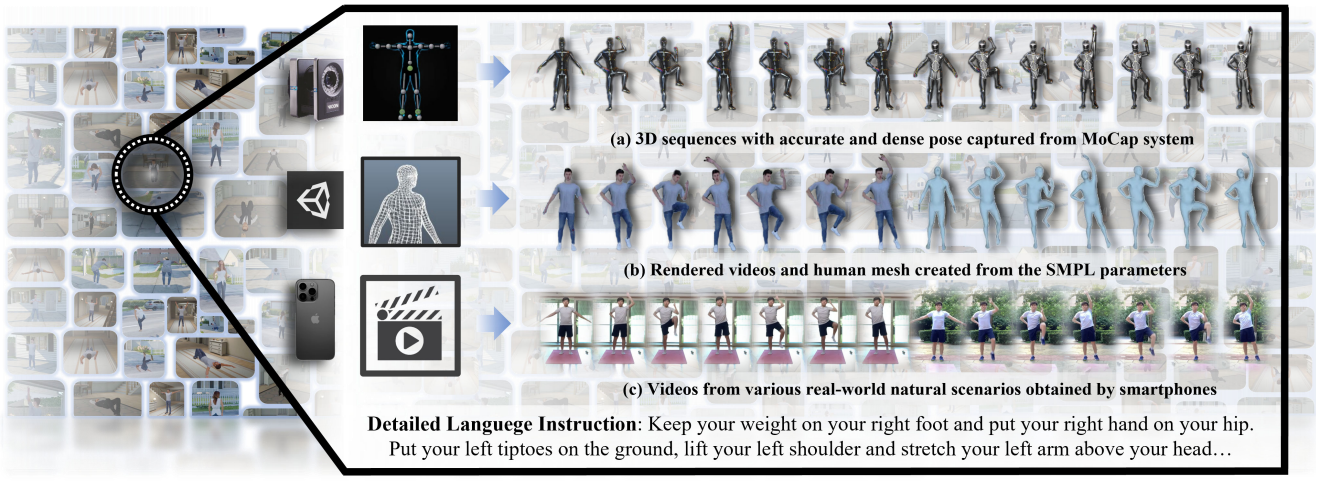


Figure 1. An overview of the proposed FLAG3D dataset, which contains 180K videos of 60 daily fitness activities. Our dataset is comprised of (a) 3D activity sequences captured from advanced MoCap system, (b) rendered videos of different people with their SMPL parameters, and (c) real-world videos obtained by cost-effective phones from both indoor and outdoor natural environments. FLAG3D also provides a series of detailed and professional sentence-level language instructions for each fitness activity. All figures are best viewed in color.

Abstract

With the continuously thriving popularity around the world, fitness activity analytic has become an emerging research topic in computer vision. While a variety of new tasks and algorithms have been proposed recently, there are growing hunger for data resources involved in high-quality data, fine-grained labels, and diverse environments. In this paper, we present FLAG3D, a large-scale 3D fitness activity dataset with language instruction containing 180K sequences of 60 categories. FLAG3D features the following three aspects: 1) accurate and dense 3D human pose captured from advanced MoCap system to handle the complex activity and large movement, 2) detailed and professional language instruction to describe how to perform a specific activity, 3) versatile video resources from a high-tech MoCap system, rendering software, and cost-effective smartphones in natural environments. Extensive experiments and in-depth analysis show that FLAG3D contributes great research value for various challenges, such as cross-domain human action recognition, dynamic human mesh recovery, and language-guided human action generation. Our dataset and source code are publicly available at https://andytang15.github.io/FLAG3D.

1. Introduction

With the great demand of keeping healthy, reducing high pressure from working and staying in shape, fitness activity has become more and more important and popular during the past decades [22]. According to the statistics[1], there are over 200,000 fitness clubs and 170 million club members all over the world. More recently, because of the high expense of coaches and out-breaking of COVID-19, increasing people choose to exclude gym membership and do the workout by watching the fitness instructional videos from fitness apps or YouTube channels (*e.g.,* FITAPP, ATHLEAN-X, The Fitness Marshall, *etc.*). Therefore, it is desirable to ad-

[1]https://policyadvice.net/insurance/insights/fitness-industry-statistics

Table 1. Comparisons of FLAG3D with the relevant datasets. FLAG3D consists of 180K sequences (Seqs) of 60 fitness activity categories (Cats). It contains both low-level features, including 3D key points (K3D) and SMPL parameters, as well as high-level language annotation (LA) to instruct trainers, sharing merits of multiple resources from MoCap system in laboratory (Lab), synthetic (Syn.) data by rendering software and natural (Nat.) scenarios. We evaluate various tasks in this paper, including human action recognition (HAR), human mesh recovery (HMR), and human action generation (HAG), while more potential applications like human pose estimation (HPE), repetitive action counting (RAC), action quality assessment and visual grounding could be explored in the future (see Section 5 for more details.).

Dataset	Subjs	Cats	Seqs	Frames	LA	K3D	SMPL	Resource	Task
PoseTrack [7]	-	-	550	66K	×	×	×	Nat.	HPE
Human3.6M [33]	11	17	839	3.6M	×	✓	-	Lab	HAR,HPE,HMR
CMU Panoptic [37]	8	5	65	594K	×	✓	-	Lab	HPE
MPI-INF-3DHP [57]	8	8	-	>1.3M	×	✓	-	Lab+Nat.	HPE,HMR
3DPW [96]	7	-	60	51k	×	×	✓	Nat.	HMR
ZJU-MoCap [68]	6	6	9	>1k	×	✓	✓	Lab	HAR,HMR
NTU RGB+D 120 [51]	106	120	114k	-	×	✓	-	Lab	HAR,HAG
HuMMan [11]	1000	500	400K	60M	×	✓	✓	Lab	HAR,HMR
HumanML3D [26]	-	-	14K	-	✓	✓	✓	Lab	HAG
KIT Motion Language [71]	111	-	3911	-	✓	✓	-	Lab	HAG
HumanAct12 [28]	12	12	1191	90K	×	×	✓	Lab	HAG
UESTC [35]	118	40	25K	> 5M	×	✓	-	Lab	HAR,HAG
Fit3D [22]	13	37	-	> 3M	×	✓	✓	Lab	HPE,RAC
EC3D [115]	4	3	362	-	×	✓	-	Lab	HAR
Yoga-82 [95]	-	82	-	29K	×	×	×	Nat.	HAR,HPE
FLAG3D (Ours)	10+10+4	60	180K	20M	✓	✓	✓	Lab+Syn.+Nat.	HAR,HMR,HAG

vance current intelligent vision systems to assist people to perceive, understand and analyze various fitness activities.

In recent years, a variety of datasets have been proposed in the field [22, 95, 115], which have provided good benchmarks for preliminary research. However, these datasets might have limited capability to model complex poses, describe a fine-grained activity, and generalize to different scenarios. We present FLAG3D in this paper, a 3D **F**itness activity dataset with **LAnG**uage instruction. Figure 1 presents an illustration of our dataset, which contains 180K sequences of 60 complex fitness activities obtained from versatile sources, including a high-tech MoCap system, professional rendering software, and cost-effective smartphones. In particular, FLAG3D advances current related datasets from the following three aspects:

Highly Accurate and Dense 3D Pose. For fitness activity, there are various poses within lying, crouching, rolling up, jumping *etc.*, which involve heavy self-occlusion and large movements. These complex cases bring inevitable obstacles for conventional appearance-based or depth-based sensors to capture the accurate 3D pose. To address this, we set up an advanced MoCap system with 24 VICON cameras [5] and professional MoCap clothes with 77 motion markers to capture the trainers' detailed and dense 3D pose.

Detailed Language Instruction. Most existing fitness activity datasets merely provide a single action label or phase for each action [95, 115]. However, understanding fitness activities usually requires more detailed descriptions. We collect a series of sentence-level language instructions for describing each fine-grained movement. Introducing language would also facilitate various research regarding emerging multi-modal applications.

Diverse Video Resources. To advance the research directly into a more general field, we collect versatile videos for FLAG3D. Besides the data captured from the expensive MoCap system, we further provide the synthetic sequences with high realism produced by rendering software and the corresponding SMPL parameters. In addition, FLAG3D also contains videos from natural real-world environments obtained by cost-effective and user-friendly smartphones.

To understand the new challenges in FLAG3D, we evaluate a series of recent advanced approaches and set a benchmark for various tasks, including skeleton-based action recognition, human mesh recovery, and dynamic action generation. Through the experiments, we find that 1) while the state-of-the-art skeleton-based action recognition methods have attained promising performance with highly accurate MoCap data under the in-domain scenario, the results drop significantly under the out-domain scenario regarding the rendered and natural videos. 2) Current 3D pose and shape estimation approaches easily fail on some poses, such as kneeling and lying, owing to the self-occlusion. FLAG3D provides accurate ground truth for these situations, which could improve current methods' performance in addressing challenging postures. 3) Motions generated by state-of-the-art methods appear to be visually plausible

and context-aware at the beginning. However, they cannot follow the text description faithfully as time goes on.

To summarize, our contributions are twofold: 1) We present a new dataset named FLAG3D with highly accurate and dense 3D poses, detailed language instruction, and diverse video resources, which could be used for multiple tasks for fitness activity applications. 2) We conduct various empirical studies and in-depth analyses of the proposed FLAG3D, which sheds light on the future research of activity understanding to devote more attention to the generality and the interaction with natural language.

2. Related Work

Table 1 presents a comparison of our FLAG3D with the related datasets. FLAG3D provides detailed language instructions for text-driven action generation compared with other datasets. Here we briefly review numbers of relevant datasets and methods regarding the three tasks we focus on.

Human Action Recognition. As the foundation of video understanding, pursuing diverse datasets has never stopped in action recognition. Existing works have explored various modalities in this area, such as RGB videos [14, 43, 76, 83], optical flows [80], audio waves [100], and skeletons [105]. Among these modalities, skeleton data draws increasing attention because of its robustness to environmental noises and action-focusing nature. During the past few years, various network architectures have been exploited to model the spatio-temporal evolution of the skeleton sequences, such as different variants of RNNs [19, 82, 111], CNNs [18, 20, 50, 104] and GCNs [16, 47, 52, 77, 90, 105]. In terms of skeletal keypoint, current action recognition datasets can be divided into two classes: one is 2D keypoint datasets [20, 105] extracted by pose estimation methods [13, 15, 56, 84], and the other is 3D keypoint datasets [33, 51, 61, 75] collected by sensors or other sophisticated equipment. However, most existing datasets are limited to a single domain of natural scenes. FLAG3D dataset takes a different step towards cross-domain action recognition between rendered videos and real-world scenario videos.

Human Mesh Recovery. Human mesh recovery obtains well-aligned and physically plausible mesh results that human models can parametrize, such as SMPL [53], SMPL-X [66], STAR [62] and GHUM [101]. Current methods take keypoints [9, 66, 114], images [24, 25, 41, 42, 46, 61, 67], videos [17, 39, 54, 58, 60, 86] and point clouds [8, 30, 36, 49, 97] as inputs to recover the parametric human model under optimization [9, 44, 109] or regression [38, 42, 61, 67, 93] paradigm. Besides the above input modalities, there are ground-truth SMPL parameters provided by human datasets. They are registered by marker-less multi-view MoCap [57, 59, 66, 68, 108, 113], or marker/sensor based Mocap [33, 79, 96]. SMPL can also be fitted with the rendered human scan in synthesis datasets [12, 65, 94, 107]. Easily recovered human poses of existing datasets cause the performance of human mesh recovery algorithms not to be fairly evaluated [63], whereas FLAG3D provides human poses with heavy self-occlusion and large movements. The work most related to ours is HuMMan [11], which contains large-scale and comprehensive multi-modal resources captured in a single MoCap room. In comparison, FLAG3D is complementary with rendered and natural videos, as well as more detailed language instructions to describe the activity.

Human Action Generation. In the past several years, various works have utilized multiple forms of information to guide the generation of human actions. Among these, one direction [10, 27, 28, 69, 98, 106] is to explore the underlying data structure of action sequences based on action categories. As real-life movements are often accompanied by audio messages, another direction [32, 45, 78, 87, 88] is to use the audio and motion timing alignment feature. Translating text descriptions to human motion is an emerging topic as well. Several works [6, 23, 26, 29, 70, 72, 74, 81, 91, 92, 103, 110] try to match semantic information and high-dimensional features of action sequences so that it could pursue natural motion sequences guided by language. For traditional motion generation tasks, HumanAct12 [28], UESTC [34] and NTU RGB+D [51] are three commonly used benchmarks. However, the above datasets do not provide paired sophisticated semantic labels to the motion sequences. Moreover, there are not enough motions for the exact text in the language-motion dataset KIT [71]. Recently, BABEL [73] and HumanML3D [26] re-annotates AMASS [55] with English language labels. Nonetheless, they focus on simple actions with uncomplicated descriptions. FLAG3D provides long sequences of actions with detailed and professional language instructions.

3. The FLAG3D Dataset

3.1. Taxonomy

The first challenge to construct FLAG3D is establishing a systematic taxonomy to organize various fitness activities. In previous literature, most existing fitness datasets [22, 95, 115] mix up all the activities. We present a deeper hierarchical lexicon as shown in Figure 2, which contains three levels from roots to leaves, including body parts, fitness activity, and language instructions.

(1) *Body Part.* For the first level, we share our thoughts with HuMMan [11] which uses the driving muscles as basic categories. However, numerous fine-grained muscles exist in the human body, and one activity might be driven by different muscles. We follow the suggestions of our fitness training coaches and choose ten parts of the human body with rich muscles as *chest, back, shoulder, arm, neck, abdomen, waist, hip, leg* and *multiple parts*[2].

[2]Some activities are driven by muscles of various body parts.

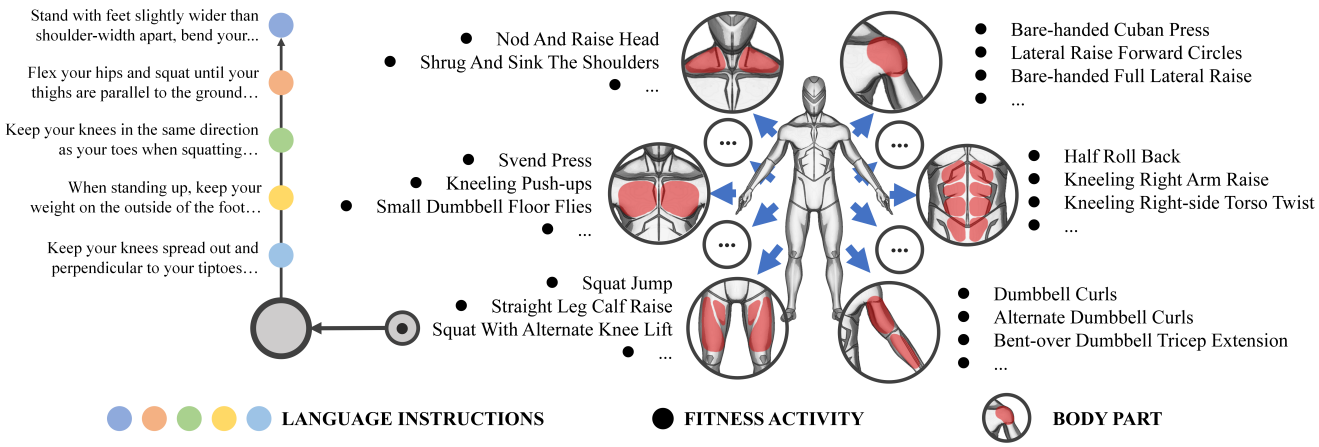


Figure 2. An illustration of the taxonomy of our FLAG3D dataset, which is systematically organized in three levels as *body part*, *fitness activity* and *language instruction*. This figure details a concrete example of the *"Squat With Alternate Knee Lift"* activity that is mainly driven by the quadriceps femoris muscle of the *"Leg"*, while the corresponding language instructions are shown in the left.

(2) *Fitness Activity.* Sixty everyday fitness activities are selected for the second level, linked to the corresponding body parts of the first level. For example, the activity *"Squat With Alternate Knee Lift"* is associated with the quadriceps femoris muscle of the *"Leg"*. We include the complete list of the 60 fitness activities in the Appendix.

(3) *Language Instruction.* We compose the third level of the lexicon with a set of language descriptions from the guidance of the training coaches to instruct users to accomplish the fitness activity. As an example shown in Figure 2, the fitness activity *"Squat With Alternate Knee Lift"* is detailed as *"Stand with feet slightly wider than shoulder-width apart, bend your elbows and put your hands in front of your chest. Flex your hips and squat until your thighs are parallel to the ground, and keep your knees in the same direction as your toes when squatting..."*. There are about 3 sentences and 57 words for each fitness activity on average.

3.2. Data Collection

We deploy high-precision MoCap equipment in an open lab to capture accurate human motion information. To obtain the rendered videos, we purchase kinds of virtual scenes and character models to make full use of the collected 3D skeleton sequences. In different environments, we record real-world natural videos. We agree with the volunteers and ensure that researchers can use these data. More details can be found in the supplementary materials. We detail the data collection process below.

Data from MoCap System. Our MoCap system is equipped in a lab of 20 meters long, 8 meters wide, and 7 meters high. The lab uses the high-tech VICON [5] MoCap system to capture the actors' body part movements through optical motion capture. Cameras used in this system have a maximum resolution of 4096×4096. It is capable of 120fps while maintaining maximum resolution sampling. Ten volunteers perform the actions in the motion capture field. Infrared cameras transmit the high frame rate IR gray-scale images captured over the fiber optic cable to the data switch. They are clock aligned and finally sent to a device with specialized processing software. Through the professional machine, we can monitor the movements of volunteers in various forms, including masks, bones, and marker points. Moreover, we hire professional technicians to perform data restoration and motion retargeting based on high-precision original data so that we can ensure the accuracy and diversity of provided 3D motion data. Meanwhile, we ask each performer to wear MoCap clothes with 77 motion markers listed on a Table in supplementary materials. Dense marker points are also a safeguard for our 3D motion data. Before performing the activity, we ask them to watch the instructional video and read the language instructions. For each action, eight males and two females will perform three times, each containing over eight repetitions. In total, we have 7200 motion sequences, where $7200 = 10(\textit{people}) \times 3(\textit{times}) \times 60(\textit{actions}) \times 4(\textit{motion retargeting})$.

Data from Rendering Software. To fully utilize the 3D MoCap data, we use the rendering software Unity3D [4] to produce synthetic 2D videos with RGB color. For 2D videos, we purchase realistic scene models in Unity Asset Store [3], including indoor and outdoor scenes. As well, we select 6 camera positions in each scene. Our camera positions are dispersed around the avatar. However, we change parameters such as the focal length of each camera to ensure that the viewfinder fits and that the camera parameters are diverse. Specifically, we are greatly appreciated that Renderpeople [2] provides several free character models. We

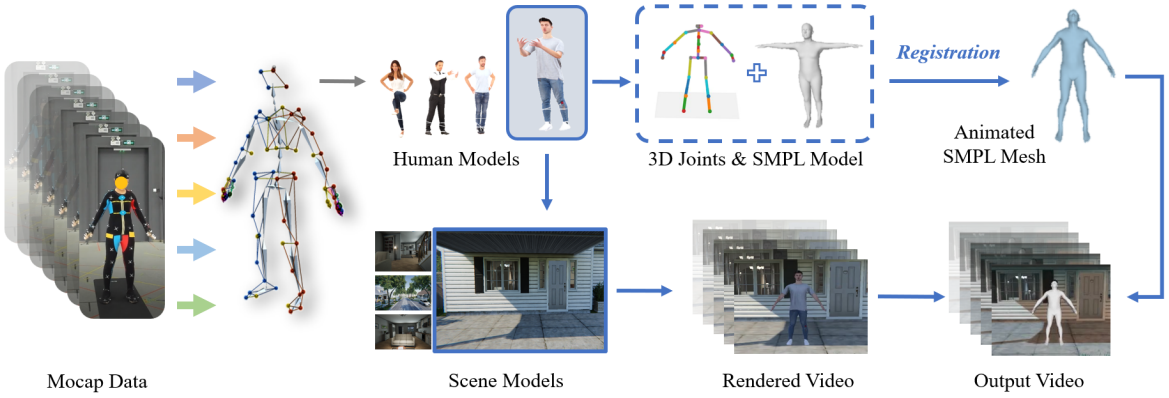


Figure 3. A process of producing the rendered videos and the SMPL parameters. First, we apply the MoCap data to get the dynamic human poses in virtual scenes and render RGB videos with camera parameters. Then we use this joint information from dynamic human models to recover the human mesh in SMPL format. Finally, we combined the SMPL mesh and RGB videos to display the output.

select 4 avatars, import the skeleton information into these avatars and record the motion of the avatars in all directions. The resolution of these videos is 854 × 480, and the frame rate is 30fps. Totally, we have 172,800 videos, where $172,800 = 1800(\textit{mocap sequence}) \times 6(\textit{camera position}) \times 4(\textit{avatar}) \times 4(\textit{virtual scences})$.

Data from Real-world Environment. To obtain versatile data resources and add diverse scenes, we ask 10 extra people to record videos in different real-world scenarios. The recording process is executed using smartphones that can capture 1080p videos from the front view and side view simultaneously, to ensure the diversity of shooting angles. As well, before performing the activity, volunteers are asked to watch the instructional video and read the language instructions carefully. We have 7200 videos, where $7200 = 10(\textit{people}) \times 3(\textit{times}) \times 60(\textit{actions}) \times 2(\textit{views}) \times 2(\textit{scenes})$.

Therefore, we have 24 subjects in all of the videos, where $24 = 10(\textit{mocap}) + 10(\textit{real-world}) + 4(\textit{render-people})$.

3.3. The Body Model

To facilitate different applications (*e.g.,* human mesh recovery and human action generation), FLAG3D adopts the SMPL [53] parametric model because of its ubiquity and generality in various downstream tasks.

Specifically, the SMPL parameters comprise pose parameters $\theta \in \mathbb{R}^{N\times 72}$, shape parameters $\beta \in \mathbb{R}^{N\times 10}$ and translation parameters $t \in \mathbb{R}^{N\times 3}$, where N is the number of frames for each video. We obtain the SMPL parameters based on the captured keypoints and an optimization algorithm. In particular, the optimization process is composed of two stages, where the first stage is to get the shape parameter $\beta \in \mathbb{R}^{N\times 10}$, and the second stage is to gain the pose $\theta \in \mathbb{R}^{N\times 72}$ as well as translation parameter $t \in \mathbb{R}^{N\times 3}$. We denote the E_s and E_p as two objective functions for shape and pose optimization. In the first stage, the objective function is formulated as follows:

$$E_s(\beta) = \frac{\lambda_1}{N} \sum_{(i,j)\in\mathcal{L}} ||\boldsymbol{J}_i(\mathbb{M}(\beta)) - \boldsymbol{J}_j(\mathbb{M}(\beta)) - \mathcal{P}(\boldsymbol{g}_i - \boldsymbol{g}_j)||_2^2 + \lambda_2||\beta||_2^2.$$

Here J_i is the joint regressor for joint i, $\boldsymbol{g}$ is the ground truth skeleton, and $\mathbb{M}$ is the parametric model [53]. $\mathcal{L}$ and $\mathcal{J}$ represent the body limbs and joint sets, respectively. $\mathcal{P}$ is the projection function that projects the $g_i - g_j$ in the direction of $J_i - J_j$. Similarly, the objective function in the second stage is as follows:

$$E_p(\theta, t) = \lambda_3 \frac{1}{N} \sum_{j\in\mathcal{J}} \lambda_{p1}||\boldsymbol{J}_j(\mathbb{M}(\theta, t)) - \boldsymbol{g}_j||_2^2 + \lambda_4||\theta||_2^2.$$

In the equations above, different weights of $\lambda_k (k = 1, 2, 3, ...)$ are denoted for each loss term (see supplementary materials for details). We adopt the L-BFGS [48] method where the search step-length satisfies the strong Wolfe conditions [99] for solving this optimization problem because of its memory and time efficiency.

4. Experiments

4.1. Human Action Recognition

FLAG3D contains both RGB videos and 3D skeleton sequences, maintaining abundant resources for 2D and 3D skeleton-based action recognition. Moreover, FLAG3D also provides videos from different domains, allowing us to evaluate different models' generalized abilities. We first report human action recognition accuracy with the 3D skeleton data captured from the MoCap system. Then we use

Table 2. Action recognition accuracy on the FLAG3D dataset.

Method	In-domain	Out-domain
ST-GCN [105]	97.8	69.9
2s-AGCN [77]	98.6	81.6
MS-G3D [52]	97.7	73.6
CTR-GCN [16]	97.5	77.2
PoseC3D [20]	-	79.9

Table 3. Results of transfer learning on FineGym and NTU60.

Method	FineGym	NTU60-XSub
ST-GCN	91.4 / **92.0** (**+0.6**)	89.0 / **89.0** (**+0.0**)
2s-AGCN	91.8 / **92.1** (**+0.3**)	89.7 / **91.0** (**+1.3**)
MS-G3D	92.7 / **93.4** (**+0.7**)	92.2 / **92.3** (**+0.1**)
CTR-GCN	92.9 / **93.5** (**+0.6**)	90.6 / **90.8** (**+0.2**)
PoseC3D	95.4 / **95.8** (**+0.4**)	93.7 / **93.9** (**+0.2**)

the 2D skeleton data extracted from both rendered and real-world videos to test the transferable ability of the models.

Experiment Setup. For the in-domain evaluation, we use the 5040 skeleton sequences of 7 subjects for training, while the other 2160 skeleton sequences of 3 subjects for testing. For the cross-domain evaluation, we follow [20] to use the Top-Down approach for 2D pose extraction in both rendered and real-world videos. For data selection, we select the rendered videos from the front and side view in one scene (7200×3) for training samples and take all 7200 real-world videos for testing. We evaluate five state-of-the-art methods as ST-GCN [105], 2s-AGCN [77], MS-G3D [52], CTR-GCN [16] and PoseC3D [20]. We exclude PoseC3D from in-domain experiments since it only supports 2D keypoint input. Table 2 presents the compared results. We also test our model using the Mindspore [1].

Result and Analysis. For the in-domain experiments, the Top-1 accuracy of all models is high, which shows that our 3D skeleton data is effective with recently advanced algorithms. Regarding the out-domain experiments, the accuracy drops drastically when transferring the models from rendered to real-world scenarios. On the widely used NTU RGB+D 60 [75] and 120 benchmark [51], Top-1 accuracy achieves and 96.6% and 89.6% respectively with PoseC3D [20], but 79.9% only on FLAG3D. Unlike NTU RGB+D, which has a large proportion of daily actions in the indoor environment, FLAG3D focuses more on the classification of fitness actions, which requires more attention to fine-grained action distinctions. Figure 4 shows these actions share sufficient similarities in motion patterns, such as bending over and swinging arms. As for *"Lying Shoulder Joint Downward Round"*, the counterpart - *"Lying Shoulder Joint Upward Round"* challenges the model in the aspect of temporal modeling. These categories require the models

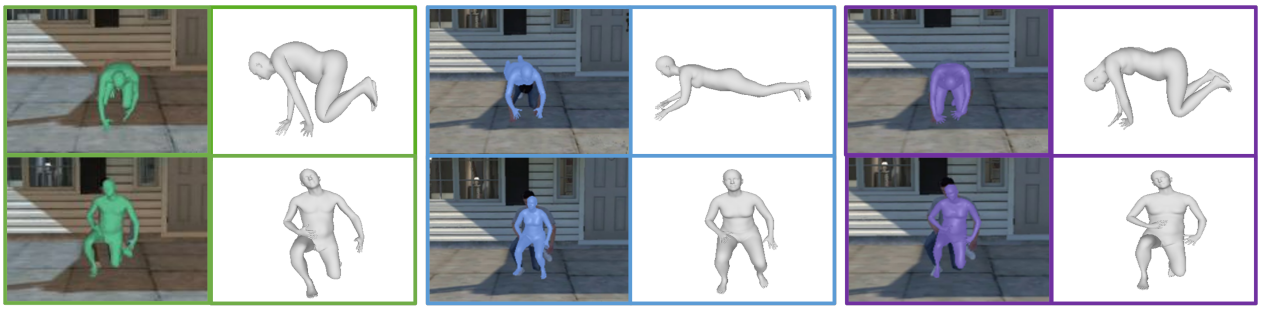
Figure 4. Case study of 2s-AGCN prediction results. The blue boxes are the selected frames of the target category, and the yellow boxes are the confusing categories. From top to bottom are *"Bent-over Dumbbell Tricep Extension"*, *"Right-side Bent-over Tricep Extension With Resistance Band"*, *"Bent-over W-shape Stretch"* and *"Bent-over Y-shape Stretch"*.

to focus on fine-grained action differences. The FLAG3D dataset can be served as a new benchmark for out-domain and fine-grained action understanding. Moreover, we fine-tune the models (pre-trained on FLAG3D) on FineGym and NTU60. In Table 3, pre-trained models achieved better performance (**in bold**), especially on FineGym, which shares some common grounds with FLAG3D such as the fine-grained nature of sports. Promising results show that our FLAG3D dataset can transfer beneficial signals for pre-trained models to boost the performance of other datasets.

4.2. Human Mesh Recovery

FLAG3D provides the SMPL [53] annotations, which are the prevalent ground truth in human mesh recovery. It is available to perform and evaluate popular methods for estimating 3D human poses and shapes. In this section, we first evaluate deep learning-based regression algorithms to verify that our dataset is qualified as a benchmark. Then we use the SMPL [53] annotation data to train ROMP [21] to improve its performance on our test set.

Experiment Setup. To ensure diversity in the subset, we opt for 300K frames for each scene and view during data selection. In order to avoid potential continuity issues and information leakage (*e.g.*, two videos with the same action and human model but different repetitions are in different datasets), we select the first 20% videos for each scene and view them as the test set. We benchmark three typical methods: VIBE [40], BEV [85], and ROMP [21] on FLAG3D using MPJPE (mean per joint position error) and PA-MPJPE (Procrustes-aligned mean per joint position error) metrics. Results are presented in Table 4.



Ground Truth ROMP ROMP-ft

Figure 5. Examples of SMPL prediction results. The top activity is prostrating and the bottom one is taking the knee. ROMP-ft (*i.e.,* fine-tuning ROMP on the training set of FLAG3D) improves its ability to correctly estimate complex postures like kneeling.

Table 4. Human mesh recovery accuracy on the FLAG3D dataset. "↓" indicates that the lower value is better. "ft" represents that we have fine-tuned this method on our trainset.

Method	MPJPE ↓	PA-MPJPE ↓
VIBE [40]	376.67	106.27
BEV [85]	382.77	117.62
ROMP [21]	379.44	100.48
ROMP-ft [21]	**114.73**	**62.29**

Table 5. Performance on challenging cases using two protocols.

	P1		P2	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
w/o	260.918	132.574	490.248	111.654
w. ft-FLAG3D	**119.109**	**81.179**	**131.001**	**75.428**

Result and Analysis. These methods without training achieved unsatisfactory MPJPE and PAMPJPE on the dataset. One of the most important reasons is that when the person in the rendered video is kneeling or lying, the task could be challenging because of the occlusion in the visual view. As displayed in Figure 5, ROMP [21] interprets kneeling as lying and interprets taking on the knee as squatting on the ground. For the evaluation part, VIBE [40] and ROMP [21] achieved the best MPJPE and PA-MPJPE, respectively. But these metrics are still high, indicating that there is still much room for improvement of 3D shape estimation methods on the FLAG3D dataset. Therefore, FLAG3D can be served as a new benchmark for 3D pose and shape estimation tasks. Since ROMP [21] achieved the top-1 PA-MPJPE, we fine-tuned it on FLAG3D with HRNet [84] backbone. ROMP [21] could handle challenging cases and reach better MPJPE and PA-MPJPE after being fine-tuned on our dataset. It indicates that our dataset could benefit 3D pose estimation approach to improve their performance. To verify our ideas, we also test videos involving challenging actions in protocol 1 and challenging views in protocol 2 as shown in table 5. Both situations with self-occlusion can be mitigated after fine-tuning on FLAG3D.

Table 6. Results of MDM in KIT dataset.

	R-Precision ↑	FID ↓
w/o FLAG3D	0.396	0.497
w. FLAG3D	**0.407**	**0.491**

4.3. Human Action Generation

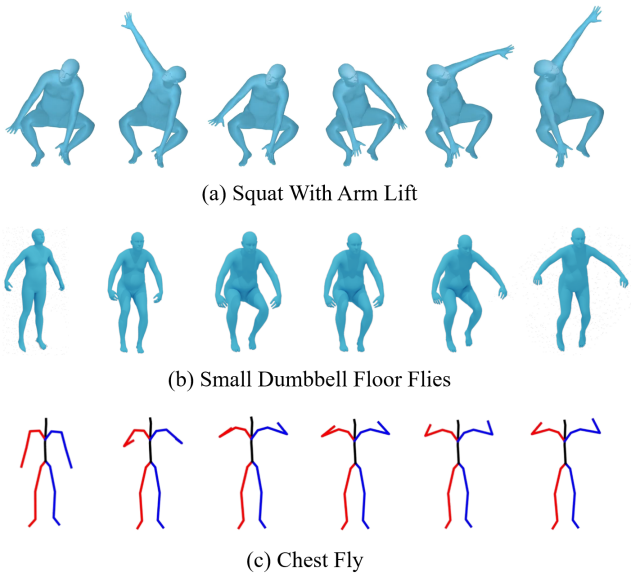
Detailed language instructions and 3D skeleton-based motion sequences with SMPL [53] annotations are included in FLAG3D, which facilitate the application of human action generation. This section reports the results of action-conditioned 3D human motion synthesis under both category-based settings and language-based settings.

Experiment Setup. For category-based action generation, we test the results in ACTOR [69]. We first selected skeleton sequences of 5 subjects for training, while the other 5 subjects were for testing. Same as UESTC [34] in ACTOR [69], we use ST-GCN [105] as the feature extractor. For language-based action generation, we evaluate the methods of Guo *et al.* [26] and TEMOS [70]. We use 90% language-motion pairs for training and 10% for testing.

Result and Analysis. Owing to the different architectural designs of algorithms, we carry over the metrics set of the original paper in each method. Results are shown in Table 7. *For category-based settings*, the FID of FLAG3D is 14.77, and the Multimod index is 6.53. In some cases, ACTOR [69] provides satisfactory results as shown in Figure 6 (a). *For language-based settings*, due to the different complexity of actions in FLAG3D, the action durations are relatively long and they do not satisfy a uniform distribution. As shown in Figure 6 (b), TEMOS [70] appears to be visually plausible and context-aware at the beginning. However,

Table 7. Results of human action generation. Multi.: MultiModality. APE: Average Position Error. AVE: Average Variance Error.

Method	FID ↓	Acc. ↑	Multi.↑	Method	APE_{root} ↓	AVE_{root} ↓	Method	FID ↓	R-precision↑	Multi.↑
ACTOR [69]	14.77	94.50	6.53	TEMOS [70]	0.61	0.66	Guo *et al.* [26]	15.12	0.10	1.20



(a) Squat With Arm Lift

(b) Small Dumbbell Floor Flies

(c) Chest Fly

Figure 6. Qualitative results in FLAG3D. **(a)** "Squat With Arm Lift" visualization results in the category-based method. **(b)** Small Dumbbell Floor Flies: *"Lie flat on your yoga mat, bend your knees, spread your legs shoulder-width apart, and keep your feet firmly planted on the ground. Sink your shoulder blades so that your upper back is flat against the mat..."* At first, the avatar bent the knee correctly, however, it fails to faithfully follow the text description as time goes on. **(c)** Chest Fly: *"Raise your head, keep your chest out, and tighten your abdominal muscles. Keep your palms forward and open your arms alternately at the same time..."* When comes to the word "alternately", Guo *et al.* [26] fails to capture the semantic information and comes to a standstill. (Due to the different forms of data organization in different methods. In Guo *et al.* [26], we use skeletons to demonstrate effects.)

it fails to follow the text description faithfully as time goes on. More structure about characterization in temporal dependencies should be designed. Guo *et al.* [26] designed a list to record keywords such as body parts and movements. So that the model could focus on specific words. FLAG3D is semantically informative and has many professional descriptions. As shown in Figure 6 (c), Guo *et al.* [26] cannot capture the information of the word "alternately" after the execution of movement "keep your palms forward". These cases require models of more generalization so that they could suffer from out-of-distribution descriptions. Moreover, Flag3D increased the performance of existing methods. As shown in Table 6, MDM [92] achieved a better effect in KIT [71] after pretraining on FLAG3D. Results show that the FLAG3D dataset contains beneficial information that can transfer to other datasets.

5. Future Works and Discussion

Based on the high-quality and versatile data resources of FLAG3D, some other potential directions could be further explored. We discuss some of them below:

Visual Grounding. The language instructions of FLAG3D involve the critical steps of specific body parts to accomplish an activity. Grounding these key phases with the corresponding spatial-temporal regions could better bridge the domain gap between linguistic and visual inputs.

Repetitive Action Counting. Counting the occurring times of the repetitive actions benefits users for fitness training [22,31]. While it requires more fine-grained annotations of the temporal boundary, it is desirable to explore unsupervised or semi-supervised learning methods in this direction.

Action Quality Assessment. This task aims to assess how well a fitness activity is performed and give feedback to users to avoid injury and improve the training effect. Unlike previous works [64,89,102,112], the future effort could be devoted to FLAG3D by evaluating whether a 3D activity meets the rule described by the language instruction.

6. Conclusion

In this paper, we have proposed FLAG3D, a large-scale comprehensive 3D fitness activity dataset which shares the merits over previous datasets from various aspects, including highly accurate skeleton, fine-grained language description, and diverse resources. Both qualitative and quantitative experimental results have shown that FLAG3D poses new challenges for multiple tasks like cross-domain human action recognition, dynamic human mesh recovery, and language-guided human action generation. We hope the FLAG3D will promote in-depth research and more applications on fitness activity analytics for the community.

Acknowledgments. This work was sponsored in part by the National Natural Science Foundation of China (Grant No. 62206153, 62125603), CAAI-Huawei MindSpore Open Fund, Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No: ZDSYS20210623092001004), Young Elite Scientists Sponsorship Program by CAST (No. 2022QNRC001), and Shenzhen Stable Supporting Program (WDZC20220818112518001).

References

[1] Mindspore. https://www.mindspore.cn/. 6

[2] Renderpeople. https://renderpeople.com/. 4

[3] Unity asset store. https://assetstore.unity.com/. 4

[4] Unity3d. https://unity.com/. 4

[5] Vicon. https://www.vicon.com/hardware/cameras. 2, 4

[6] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728, 2019. 3

[7] 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

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *ECCV*, pages 311–329, 2020. 3

[9] 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 *ECCV*, pages 561–578, 2016. 3

[10] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *ECCV*, pages 374–390, 2018. 3

[11] 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, 3

[12] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Jiatong Li, Zhengyu Lin, Haiyu Zhao, Shuai Yi, Lei Yang, et al. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 3

[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 3

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[15] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018. 3

[16] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021. 3, 6

[17] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, pages 1964–1973, 2021. 3

[18] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, pages 7024–7033, 2018. 3

[19] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, pages 1110–1118, 2015. 3

[20] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, pages 2969–2978, 2022. 3, 6

[21] 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. 6, 7

[22] 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. 1, 2, 3, 8

[23] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, pages 1396–1406, 2021. 3

[24] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 3

[25] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, pages 10884–10894, 2019. 3

[26] 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, 3, 7, 8

[27] Chuan Guo, Xinxin Zuo, Sen Wang, Xinshuang Liu, Shihao Zou, Minglun Gong, and Li Cheng. Action2video: Generating videos of human 3d actions. *IJCV*, 130(2):285–315, 2022. 3

[28] 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, 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 3

[30] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Garment4d: Garment reconstruction from point cloud sequences. In *NeurIPS*, pages 27940–27951, 2021. 3

[31] Huazhang Hu, Sixun Dong, Yiqun Zhao, Dongze Lian, Zhengxin Li, and Shenghua Gao. Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In *CVPR*, pages 18991–19000, 2022. 8

[32] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 3

[33] 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, 3

[34] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *ACM MM*, pages 1510–1518, 2018. 3, 7

[35] 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

[36] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. Skeleton-aware 3d human shape reconstruction from point clouds. In *ICCV*, pages 5431–5441, 2019. 3

[37] 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

[38] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 3

[39] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, pages 5614–5623, 2019. 3

[40] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 6, 7

[41] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 3

[42] 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 *ICCV*, pages 2252–2261, 2019. 3

[43] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 3

[44] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, pages 6050–6059, 2017. 3

[45] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *NeurIPS*, pages 3581–3591, 2019. 3

[46] 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. In *CVPR*, pages 3383–3393, 2021. 3

[47] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 3595–3603, 2019. 3

[48] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. pages 503–528, 1989. 5

[49] Guanze Liu, Yu Rong, and Lu Sheng. Votehmr: Occlusion-aware voting network for robust 3d human mesh recovery from partial point clouds. In *ACM MM*, pages 955–964, 2021. 3

[50] Hong Liu, Juanhui Tu, and Mengyuan Liu. Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv preprint arXiv:1705.08106*, 2017. 3

[51] 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, 3, 6

[52] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. 3, 6

[53] 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, 5, 6, 7

[54] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, pages 324–340, 2020. 3

[55] 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. 3

[56] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2640–2649, 2017. 3

[57] 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, 3

[58] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *TOG*, 39(4):82–1, 2020. 3

[59] 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 *3DV*, pages 120–130, 2018. 3

[60] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single RGB image. In *ECCV*, pages 752–768, 2020. 3

[61] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, pages 484–494, 2018. 3

[62] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: sparse trained articulated human body regressor. In *ECCV*, pages 598–613, 2020. 3

[63] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *NeurIPS*, 2022. 3

[64] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPR*, pages 20–28, 2017. 8

[65] 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. 3

[66] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 3

[67] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 3

[68] 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, 3

[69] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 3, 7, 8

[70] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497, 2022. 3, 7, 8

[71] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 3, 8

[72] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *RAS*, 109:13–26, 2018. 3

[73] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021. 3

[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[75] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 3, 6

[76] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 3

[77] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 3, 6

[78] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *CVPR*, pages 7574–7583, 2018. 3

[79] 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. *IJCV*, 87(1):4–27, 2010. 3

[80] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014. 3

[81] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[82] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, pages 4263–4270, 2017. 3

[83] 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. 3

[84] 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, 7

[85] 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. 6, 7

[86] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, pages 5349–5358, 2019. 3

[87] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional lstm. In *HAI*, pages 365–369, 2017. 3

[88] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *ACM MM*, pages 1598–1606, 2018. 3

[89] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, pages 9839–9848, 2020. 8

[90] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *CVPR*, pages 5323–5332, 2018. 3

[91] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022. 3

[92] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3, 8

[93] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NeurIPS*, pages 5236–5246, 2017. 3

[94] 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. 3

[95] 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, 3

[96] 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, 3

[97] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *CVPR*, pages 7639–7648, 2021. 3

[98] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *AAAI*, pages 12281–12288, 2020. 3

[99] Philip Wolfe. Convergence conditions for ascent methods. pages 226–235, 1969. 5

[100] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 3

[101] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, pages 6184–6193, 2020. 3

[102] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *CVPR*, pages 2949–2958, 2022. 8

[103] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *RAL*, 3(4):3441–3448, 2018. 3

[104] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *CVPR*, pages 7922–7931, 2019. 3

[105] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018. 3, 6, 7

[106] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. Structure-aware human-action generation. In *ECCV*, pages 18–34, 2020. 3

[107] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, pages 5746–5756, 2021. 3

[108] 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. 3

[109] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 3

[110] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[111] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *ICCV*, pages 2117–2126, 2017. 3

[112] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *CVPR*, 2023. 8

[113] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, pages 7376–7385, 2020. 3

[114] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *ICCV*, pages 5560–5569, 2021. 3

[115] 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, 3