Exploring the Impact of Rendering Method and Motion Quality on Model Performance when Using Multi-view Synthetic Data for Action Recognition

Stanislav Panev*[1], Emily Kim*[1], Sai Abhishek Si Namburu[1], Desislava Nikolova[2], Celso de Melo[3], Fernando De la Torre[1], Jessica Hodgins[1]

[1]Carnegie Mellon University, [2]Technical University of Sofia, [3]Army Research Laboratory

{spanev, ekim2}@andrew.cmu.edu, snamburu@alumni.cmu.edu, dnikolova@tu-sofia.bg, celso.m.demelo.civ@army.mil, {ftorre, jkh}@andrew.cmu.edu

Abstract

This paper explores the use of synthetic data in a human action recognition (HAR) task to avoid the challenges of obtaining and labeling real-world datasets. We introduce a new dataset suite comprising five datasets, eleven common human activities, three synchronized camera views (aerial and ground) in three outdoor environments, and three visual domains (real and two synthetic). For the synthetic data, two rendering methods (standard computer graphics and neural rendering) and two sources of human motions (motion capture and video-based motion reconstruction) were employed. We evaluated each dataset type by training popular activity recognition models and comparing the performance on the real test data. Our results show that synthetic data achieve slightly lower accuracy (4–8 %) than real data. On the other hand, a model pre-trained on synthetic data and fine-tuned on limited real data surpasses the performance of either domain alone. Standard computer graphics (CG)-rendered data delivers better performance than the data generated from the neural-based rendering method. The results suggest that the quality of the human motions in the training data also affects the test results: motion capture delivers higher test accuracy. Additionally, a model trained on CG aerial view synthetic data exhibits greater robustness against camera viewpoint changes than one trained on real data. See the project page: http://humansensinglab.github.io/REMAG/

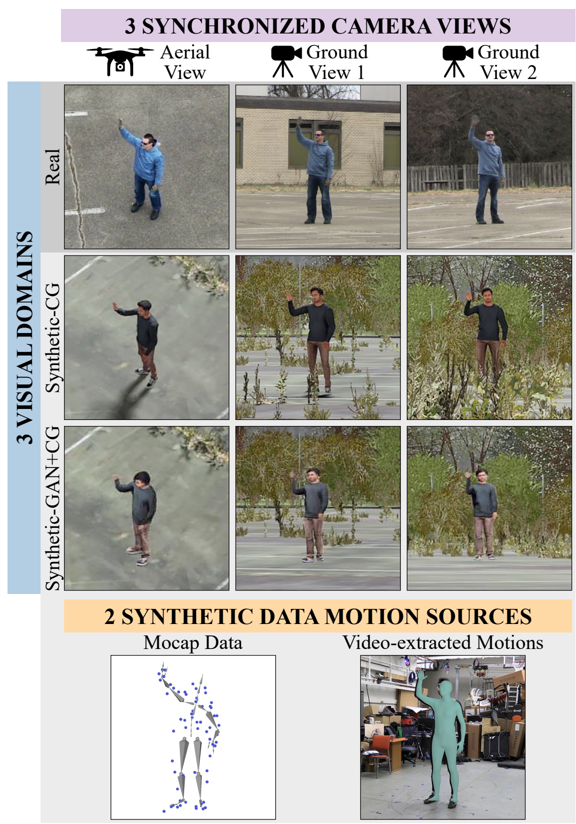

Figure 1. REMAG - an HAR dataset suite comprises five datasets: one real and four synthetic by combining two renderers (CG and neural) with two motion sources (motion capture and video-based). Each of them includes three camera views.

1. Introduction

Human action recognition (HAR) from videos is crucial for numerous applications [37]. For instance, video-based human action tracking has proven useful for surveillance and detecting abnormal behaviors [36]. It has also been used for sports, training, and physical therapy [17, 28, 52].

Robust learning for HAR models relies heavily on diverse, large-scale training data. Collecting data is laborious, time-consuming, and error-prone. To solve these issues, re-

*Equal contribution to this work.

searchers have used synthetic data for training [6, 7, 51]. A significant advantage of synthetic data is that it can be scaled quickly, labels can be added effortlessly, and the data can be diversified while keeping their photometric and geometric qualities consistent for data augmentation.

We collected a dataset with video recordings of eleven activity categories captured in the wild with three cameras—one orbiting small UAV and two fixed ground cameras. We also recorded indoor motion capture data and RGB videos of the same activities and reconstructed the human motions in those videos using *VIBE* [22]. We utilized these two sources of motion data (*mocap* and *VIBE*) to create synthetic datasets by employing two different rendering techniques: a 3D computer graphics (CG) engine *Blender* and a neural human motion imitation generative model *Liquid Warping GAN* (LWG) [29]. Figure 1 presents samples of the waving gesture across the three camera views from the real and synthetic modalities.

These datasets constitute *REMAG—REndering*, *Motion*, *Aerial*, and *Ground* view analysis data suite for the HAR task. With it, we sought to answer the following questions:

1. Does the choice of the ML model make a significant difference in performance?
2. Does the synthetic data rendering method affect the model performance?
3. Does improving the quality of the motion in the synthetic training data improve the performance?
4. Can we combine synthetic training data with a limited amount of real training data to improve performance?
5. Can the models trained on one camera view transfer to a novel camera view?

We ran an extensive set of experiments and compared the performance of models trained on different modalities of the data. We used three *off-the-shelf* activity recognition models—*X3D* [11] (a single video stream), *SlowFast* [12] (dual video streams), and *MViT* [10] (transformer-based).

The contributions of our paper are as follows:

1) A new real-world video dataset for human action recognition has been collected and annotated, including three synchronized camera views, one aerial dynamic and two ground static views, and eleven everyday activities;

2) Four synthetic counterparts of the real dataset were created by combining two different rendering approaches (CG and neural-based) with two human motion sources (motion capture data and video-extracted motions);

3) An extensive set of experiments has been conducted to answer the five research questions listed above using off-the-shelf activity recognition models. We analyzed how the different rendering techniques, the motion sources, and transferring from one camera view to another affect the models' performance. To the best of our knowledge, we are the first to perform such an analysis systematically. Our datasets will be made public.

2. Related Work

Research in video action recognition has progressed quickly in the past decade. Several review papers covered in great detail different types of models and datasets for deep video action recognition [25] [47]. Here, we briefly discuss the most popular public training synthetic HAR datasets, the existing real HAR benchmarks, neural renderers for HAR data, and video-based activity recognition deep learning models.

2.1. Synthetic HAR Datasets

Many synthetic video datasets have been created for training HAR deep-learning models in recent years. Synthetic data has already proven to be a helpful solution when large amounts of annotated data should be produced quickly. In Tables 1 and 2, we compare some of the most recent synthetic and real HAR datasets.

The *PHAV* dataset [8] was created through a procedural workflow based on human action videos parametric generative model. A unique feature of it is the so-called *kite camera dynamics model*, which resembles the view from a person following the actor. The characters were animated by composite actions assembled from a library of atomic motions. It contains motion capture data or manually designed motions. The *Game Action Dataset (GAD)* dataset [46] is a relatively small dataset comprised of recordings of gaming sessions (GTA5 and FIFA) performed by human players. This dataset is the first derived from gaming environments and includes synchronized ground and aerial views. The game generates character motions that are conditioned on players' commands. Verol *et al*. used 3D human motion estimation models, such as *HMMR* [20] and *VIBE* [22], to reconstruct the human body mesh and its motions from a single view RGB videos to create the *SURREACT* [50] dataset. The body mesh is based on the *SMPL* [31] statistical model and is further augmented with randomized cloth textures, lighting conditions, and body shapes for better diversity. *SynADL* [18] is a synthetic dataset focused on detecting elders' *activities of daily living* (ADL). The 3D human characters were created by scanning 15 participants with a *Kinect* sensor. Motion capture data were recorded from the same 15 participants and used to animate the characters. Kim *et al*. combined *PHAV*, *SURREACT*, and *SynADL* to create a new dataset called *SynAPT* [51]. The goal of this dataset was to pre-train a model which would be transferred to a different downstream task involving completely new categories. *RoCoG* [6] and *RoCoG-v2* [40] are two datasets designed for human-robot interaction based on seven gestures from the *US Army Field Manual* [16]. *RoCoG-v2* presents static ground and aerial views, unlike *RoCoG*, which contains only a static ground view. Moreover, *RoCoG* contains only manually-designed motions, whereas *RoCoG-v2* introduces motion capture data

Name	Year	Num. Act.	Motion Source[a,b]				Total Frms	Num. Seq.	FPSq[o]	Rendering		Human Models		Camera Views[c]					Virt. Env.[d]		
			Ma.	Ga.	Vid.	MC				Engine	Type	#	Type	#	St.	Dy.	Ae.	Gr.	#	In.	Out.
PHAV [8]	2017	35	✓			✓	6.0M	39.9k	150	Unity	CG	20	RD15[e]	1		✓		✓	7	✓	✓
<u>RoCoG</u> [6]	2020	8	✓				17.6M	117.0k	150	Unity	CG	4	HQ[n]	8	✓			✓	4		✓
GAD [46]	2021	7		✓			0.3M	1.4k	186	GTA5 FIFA	CG	n/a	game char.[f]	2		✓	✓	✓			✓
SURREACT [50]	2021	100			✓		9.3M	109.0k	85	Cycles[g]	CG		SMPL	8	✓			✓		✓	
SynADL [18]	2021	55				✓[15]	135.0M	462.0k	292	UE4	CG	15	Kinect[h]	28	✓			✓	4	✓	
SynAPT [51]	2022	150	✓		✓	✓		150.0k		*comb.*[i]	CG	*comb.*[i]		*comb.*[i]					*comb.*[i]		
<u>RoCoG-v2</u> [40]	2023	7	✓			✓	19.6M	107.0k	183	Unity	CG	2	HQ[n]	12	✓		✓	✓	2		✓
<u>SynCG-MC</u> (ours)	2023	11				✓**[26]**	31.4M	25.4k	**1,236**	Eevee[j]	CG	**32**	**Pro.**[p]	3	✓	✓	✓	✓	3		✓
<u>SynCG-RGB</u> (ours)	2023	5			✓[15]		5.0M	6.1k	**820**	Eevee[j]	CG	**32**	**Pro.**[p]	3	✓	✓	✓	✓	3		✓
<u>SynLWG-MC</u> (ours)	2023	11				✓**[26]**	31.2M	25.4k	**1,228**	**LWG[k]+** Eevee[j]	**NR[l]+ CG**	**32**[m]	SMPL	3	✓	✓	✓	✓	3		✓
<u>SynLWG-RGB</u> (ours)	2023	5			✓[15]		5.0M	6.1k	**816**	**LWG[k]+** Eevee[j]	**NR[l]+ CG**	**32**[m]	SMPL	3	✓	✓	✓	✓	3		✓

[a] Four motion sources: manually designed motions (Ma.), game engine motions (Ga.), video-extracted (Vid.), and motion capture data (MC)
[b] The numbers depict the number of subjects included in the data [c] View types: Static (St.), Dynamic (Dy.), Aerial (Ae.), Ground view (Gr.)
[d] Virtual Environments: Indoor (In.), Outdoor (Out.) [e] 15-muscle ragdoll model [f] Game character
[g] Blender Cycles render engine [h] 25-joint models made by scanning human participants with a Kinect sensor
[i] A combination of PHAV, SURREACT, and SynADL datasets [j] Blender Eevee render engine [k] Liquid Warping GAN [29]
[l] Neural Renderer [m] SMPL meshes were textured using 32 real human avatars [n] High-quality human assets from public repositories
[o] Mean number of frames per sequence [p] Models have 25 motion capture-animated joints, high-resolution meshes and textures

Table 1. Synthetic HAR datasets comparison sorted by the year of publication. The underlined datasets include real counterparts (Table 2).

Name	Year	Num. Act.	Subj.	Seqs.	Views		Sites	
					#	Ae.	In.	Out.
UCF-ARG [13]	2010	10	12	1.4k	3	✓		✓
HMDB51 [24]	2011	51		6.8k	1		✓	✓
UCF-101 [44]	2012	101		13.3k	1		✓	✓
Kinetics-400 [21]	2017	400		306.2k	1			
Charades-Ego [42]	2018	157	112	8.0k	2		✓	
Kinetics-600 [2]	2018	600		495.4k	1			
Kinetics-700 [3]	2019	700		650.3k	1			
Kinetics-700 2020 [43]	2020	700		647.9k	1			
NTU-RGB+D 120 [27]	2020	120	106	114.5k	3		✓	
<u>RoCoG</u> [6]	2020	8	14	1.5k	1			✓
HOMAGE [38]	2021	75	27	5.7k	2		✓	
UAV-Human [26]	2021	155	119	22.5k	1	✓	✓	✓
YAD [46]	2021	8		0.4k	1	✓		✓
<u>RoCoG-v2</u> [40]	2023	7	10	0.5k	2	✓		✓
<u>Real</u> (ours)	2023	11	24	1.5k	3	✓		✓

Table 2. Real HAR datasets comparison sorted by the year of publication.

for some motions. Both datasets deliver real twin datasets.

We introduce four unique synthetic datasets: *SynCG-MC*, *SynCG-RGB*, *SynLWG-MC*, and *SynLWG-RGB*, generated using a combination of a motion source (motion capture or video-based motions) and a rendering technique (*CG* and *neural rendering*). As evident from Table 1 *SynLWG-MC* and *SynLWG-RGB* are the only HAR datasets based on neurally-generated videos, all other datasets are created using different kinds of CG renderers. To the best of our knowledge, we are the first to generate such HAR datasets and compare them with the standard CG engines. Our synthetic datasets were created using 32 realistic rigged 3D human characters produced by scanning real people [14]. They were animated using a 26-participant motion capture library we created. Furthermore, very few aerial synthetic HAR datasets exist, which is also apparent from Table 1. The two other datasets are *GAD* and *RoCoG-v2*. Still, neither contains nor provides analysis on a more complete set of camera views as we do, namely, synchronized dynamic aerial and static ground views. Moreover, the test results in [51] highlight the need for more synthetic aerial HAR datasets. The authors use a combination of three synthetic datasets (*PHAV*, *SURREACT*, and *SynADL*) without aerial view data to pre-train their models. *UAV Human* [26], the only aerial dataset out of the six real test datasets, delivers the lowest accuracy.

2.2. Real HAR Datasets

We also provide a real dataset containing the same three camera views and action categories as the four synthetic variants mentioned in the previous section. This dataset provides a performance baseline (train and test on real data only), a real test set, and a fine-tuning dataset. More details for creating it are provided in Section 3.2. Table 2 briefly compares our real and other currently existing HAR datasets.

The *UCF-ARG* [13] dataset varied the scope of the camera angles by providing videos from aerial and rooftop cameras. The aerial view was captured by a camera attached to a balloon. *HMDB51* [24] was one of the first datasets with an increased number of action categories, providing 51 action categories from 6766 videos. *UCF-101* [44] dataset introduced a larger dataset with 101 activity classes subdivided into five categories. Introducing massive action datasets from the *Kinetics* family (*Kinetics-400* [21], *Kinetics-600* [2], *Kinetics-700* [3], and *Kinetics-700-2020* [43]) accelerated the progress in the field by providing several hundred activity categories and close to 1000 videos per category. *Charades-Ego* [42] and *HOMAGE* [38] are datasets of videos of daily human activities recorded from first- and third-person perspectives. The *NTU-RGB+D120* [27] dataset also introduced labeled multi-view videos for action recognition with 120 classes from 110 000 clips of video, depth map sequences, 3D skeletal data, and infrared videos for each sample. *UAV-Human* [26] is a dataset of human activities captured by a flying UAV in multiple urban and rural districts in daytime and nighttime. It is the largest real HAR dataset which contains an aerial view. *YAD* [46] (*YouTube-Aerial Dataset*) is a small HAR dataset containing aerial videos from YouTube. It includes large and fast camera motions as well as variable shooting altitudes. *RoCoG* [6] *RoCoG-v2* [40] also contain real data of the same activity categories as their synthetic variants. *RoCoG-v2* contains aerial and ground views, whereas *RoCoG* provides only a static ground view.

2.3. HAR Data Neural Renderers

Models like *Everybody Dance Now* [5] and *Liquid Warping GAN* (LWG) [29] generate videos by transferring the body pose from a video of a person performing an activity to a new person from another video or an image. They can generate action videos with different appearances depending on the input visual source image. However, these models rely on body pose estimation from monocular videos or images, which is inherently less accurate than motion capture. We integrated *MoSh* [30], a model for generating 3D human *SMPL* meshes from a motion capture marker set, into the *LWG* architecture. Thus, we extended its capabilities to generate videos from motion capture data.

In recent years, NLP models have demonstrated the capability to generate different data types from text prompts. Models like *TEMOS* [34], *MotionCLIP* [48], *MDM* [49], *T2M-GPT* [55], and *MotionGPT* [19] can generate realistic human motions by taking activity descriptions (*text-to-motion*). Other models like *Stable Diffusion* [41] and *Dall-E* [39] can generate images from text (*text-to-image*). Ma *et al*. [32] also made one of the first attempts for a pose-guided *text-to-video* generation. However, it is hard to incorporate such models into our pipeline because they do not offer fine-grained control over the generated output and establish precise motion matching across different synthetic sets—an important aspect of our analysis.

2.4. Deep Video Action Recognition

An expansive set of activity recognition models has been developed in recent years [35]. This section reviews some of the most popular and accessible ones used in our analysis.

Architectures like *I3D* [4] and *X3D* [11] were built on image classification networks applied to videos. *I3D* does that by inflating the filters and pooling kernels to 3D, while *X3D* progressively expands on a tiny 2D image classification model with multiple axes (space, time, width, height). The performances of *I3D* and *X3D* are 73 % and 77 %, respectively when tested on *Kinetics* [9]. The *SlowFast* [12] architecture was built upon a different approach, combining the *slow* and *fast* pathways, which take small and large frame rates of the video, respectively, to identify gestures happening at shorter and longer time frames. Its performance on *Kinetics* was 77 %. *MViT* [10] is founded upon multi-scale transformers to create a multi-scale pyramid of features to learn coarse to complex features. So far, *MViT* has shown the best performance on *Kinetics*—83 % [9].

3. Our Datasets

This section describes the creation of our dataset suite consisting of a *real-world* and four *synthetic* video datasets.

3.1. Activity Classes

Our datasets contain eleven daily human activity categories (Figure 2), where five are *Gestures* and six involve *Object-handling*. Every frame that does not fit into any of these categories was labeled as *Idle*. The set of activities was intentionally constructed to contain confounding pairs to increase the difficulty of the recognition task. For instance, *Carrying a shovel on the shoulder* resembles *Carrying a bat* on the shoulder; some instances of *Holding out a flashlight* may be confused with *Talking on the phone*; *Waving "Hello or Goodbye"* is similar to *Shaking fist*.

3.2. Real-World Capture

This dataset contains video recordings of human participants using multiple cameras during daylight hours in three outdoor environments: a *grass field*, a *parking lot*, and a *tennis court*. The data collection spanned eight months (November 2021 to July 2022). Thus, different seasons were captured. All subjects consented with an approved *Institutional Review Board* (IRB) for video or motion capture.

In total, 24 subjects (18 male and 6 female) participated in the video collection process. Most of them were captured with three cameras[1] recording simultaneously in *4K* (3840 px × 2160 px) resolution at 30 fps. One camera was a remotely-controlled small UAV (DJI Mavic 2 Zoom), which circled over the participant at a constant speed as shown in Figure 3. This camera constitutes the *aerial view*. The target radius of the circular UAV trajectory R_{traj} is 15 m. The flying altitude H_{traj} was maintained within the 11 m – 13 m range. These two trajectory parameters were selected such

[1] One and a half participant sessions lack the second ground camera.

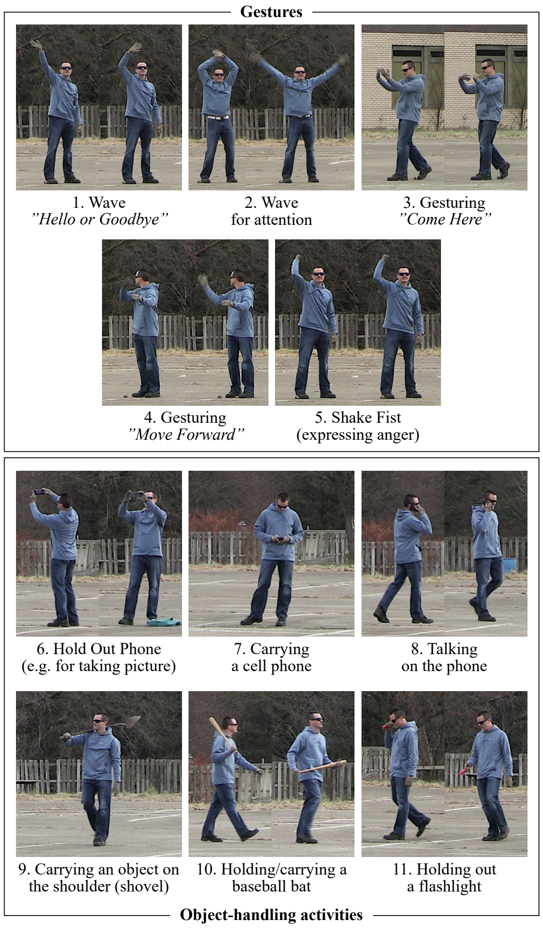

Figure 2. Human activity classes.

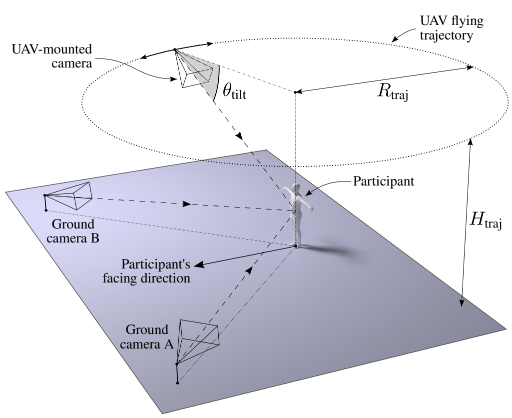

Figure 3. Real data collection setup.

that the camera's tilt angle θ_{tilt} would vary within the range $30° – 45°$. The mean height of the participant in the image frame was kept to about $200\,\text{px}$ by adjusting the lens focal length. The subject performed every activity twice, once for each flying direction of the UAV (clockwise and

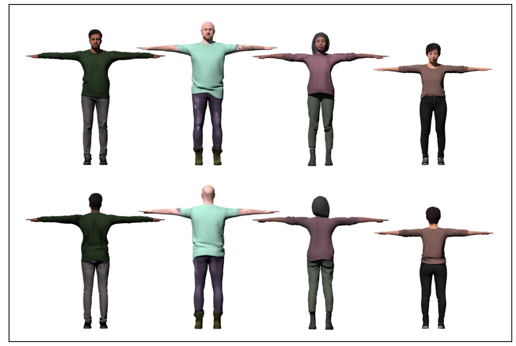

Figure 4. Front and rear view of a subset of two male and two female 3D human characters used to generate our synthetic dataset.

counter-clockwise). The other two cameras were stationary (mounted on tripods) and elevated about $1.3\,\text{m}$ above the ground. They constitute the *ground view*. The three views can be seen in Figure 1.

To synchronize the three video streams, we asked each participant to clap their hands above their head at the beginning of each session. The temporal offsets of the streams were determined manually in post-processing using the clapping actions. The annotation software *ELAN* [53] was used to synchronize and label the activity being performed.

The real-world dataset contains a total of 1538 video sequences. For all videos, we detected the subjects in each frame using *Faster R-CNN* [54] and cropped around the center of the bounding box. The frames were resized to $224\,\text{px} \times 224\,\text{px}$. The videos have a frame rate of $30\,\text{fps}$. The total frame count of the dataset is $1\,844\,253$ ($\sim$17.08 h). The mean video sequence length is $39.97\,\text{s}$.

3.3. Synthetic Data Generation

Two generation methods were employed: a standard computer graphics (CG) pipeline and a deep generative neural model called *Liquid Warping GAN* (LWG) [29]. To animate the human characters, we used two sources of motion: a motion capture (MC) dataset we collected and motions extracted from real video sequences (RGB). We utilize a simple naming convention to differentiate between the four synthetic datasets—*Syn⟨Renderer⟩-⟨MotionSource⟩*: *SynCG-MC*, *SynCG-RGB*, *SynLWG-MC*, *SynLWG-RGB*. Below, we provide more details on how each one was created.

Fully CG-based synthetic data We used *Blender* as the main 3D scene development and rendering environment (Eevee engine). *MotionBuilder* [1] was used as motion data editing and re-targeting software. We used 32 commercially-available rigged and skinned human characters [14] (Figure 4) divided into two gender groups—16 male and 16 female. We collected motion capture data from 26 human actors[2] (19 male, 7 female) performing the

[2]For more information, refer to the supplementary material.

same eleven activities from Figure 2 and re-targeted it to the characters from their gender group. Therefore, each subject's data were used to animate 16 characters. A 3D graphics artist designed three virtual environments (visualizations can be found in the supplementary material) resembling the ones from the real data capture. The ground plane meshes were reconstructed from the shooting locations through photogrammetry [45]. Various vegetation (trees and grass) and structure (buildings and fences) assets were added manually. We randomized the camera location, the lighting conditions[3], and the characters' clothing colors to introduce more diversity in the final synthetic images. The colors were sampled from the real videos.

Each of the three camera views of the *SynCG-MC* dataset is represented by 8648 video sequences that contain 10 453 420 frames—equivalent to almost 100 h of data. In total, this synthetic variant contains 31 million frames, equivalent to about 290 h of data, roughly 17 times more than the real dataset.

While collecting the motion capture data, we also captured RGB videos from 15 subjects. With this data, we generated another variation of the data rendered from the CG pipeline, *SynCG-RGB*. We used *VIBE* [22] to fit the *SMPL* [31] parametric model for each person in the videos and produce 3D meshes and skeletons. They were used to animate the synthetic avatars, similar to *SynCG-MC*. The rendering pipeline follows that of *SynCG-MC*.

LWG-generated Data A major drawback of using the traditional CG pipeline is that it requires time and manual effort to design virtual scenes and render them. To solve this problem, we used a neural-based rendering method to generate videos of the same gestures. Because the model has been pre-trained on priors to generate video of human motions, it reduces the processing for generating a single video sequence.

In particular, we used *LWG* with Attention [29] to generate the second set of synthetic videos. *LWG* is a unified framework for human image synthesis, including human motion imitation, appearance transfer, and novel view synthesis, with a 3D body mesh recovery module that utilizes *SMPL* [31] to disentangle the pose and shape.

We chose *LWG* because it allowed us to easily incorporate motion capture data a second source of human motion data. The original pipeline relies on a pose reconstruction model from 2D videos called *SPIN* [23]. We used the indoor videos we captured while collecting the motion capture data to generate a synthetic HAR dataset with this non-traditional, neural-based method (*SynLWG-RGB*). This approach resulted in lower motion quality compared to that of the motion capture. We also generated a counterpart of the *SynLWG-RGB* dataset by modifying the original *LWG* pipeline by replacing the motions extracted from *SPIN* with *SMPL* fit on motion capture and CG source images (*SynLWG-MC*). We used the *SMPL* parameters fitted on raw marker data from motion capture using *MoSh* [30]. *SynLWG-MC* had a better quality of motion (less jittery) compared to *SynLWG-RGB*.

For both *SynLWG* datasets, we used front and back rendered images of the 3D human character to generate the input source images (Figure 4). For each sequence, we used the same background, avatar, and motion capture sequence that was used in the corresponding CG sequence. All frames were resized to $224\,\mathrm{px} \times 224\,\mathrm{px}$ and were generated at 30 fps.

[3]The lighting conditions of each scene were determined by randomly selecting one out of 13 HDRI environmental spherical panoramas downloaded from *Poly Haven* [15]

4. Experiments

We used three deep video activity recognition models in our experiments: *SlowFast* [12], *MViT* [10], and *X3D* [11]. All models are part of the *PyTorch* [33] implementation of *SlowFast* on GitHub [9].

We trained all models for 300 epochs and used top-1 classification accuracy to compare their performance. We used *Kinetics-400* [21] pre-trained network weights. We used an *SGD* optimizer for *Slowfast* and *X3D*, and an *AdamW* optimizer for *MViT* with a cosine learning schedule of learning rate decay starting with 0.001 following [10–12]. We trained on 4 GPUs (batch size 7/GPU).

During training, we employed three 50 % probability augmentation methods: color-jittering (brightness, hue, contrast, saturation), random Gaussian blur, and sharpness adjustment. We randomly select 64-frame clips from videos, count frame labels, and assign the label with the highest count to the clip. All training samples contain at least 80 % of frames with the same label.

During testing, we evaluated all 64-frame clips from the videos and predicted their labels. Training procedures vary depending on the dataset. For real data, we exclude test subjects. For lab-captured datasets, we train on all data and test on the real data for three subject groups.

4.1. Analysis

This section presents the experiments we conducted using our dataset suite. They were designed to answer the following questions:

Does the choice of the ML model make a significant difference in performance? We set a performance baseline by training three activity recognition models (*SlowFast*, *MViT*, *X3D*) on real-world data. Each model is trained on eleven activity classes from 21 subjects, excluding the test subjects. Three models were trained for each camera view, each with a distinct test group.

The results from this experiment are shown in Figure 5. On average, *SlowFast* outperforms *MViT* and *X3D* for both

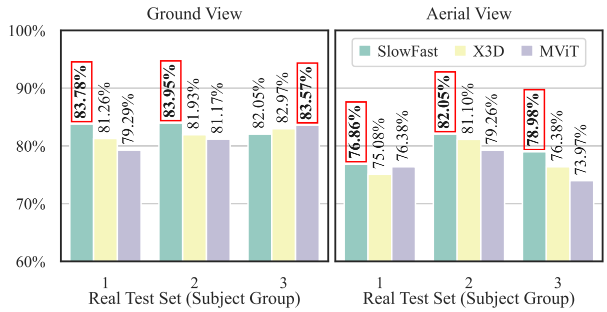

Figure 5. Recognition accuracy achieved by the three activity recognition models trained and tested on the real-world data using the full set of *eleven* activities.

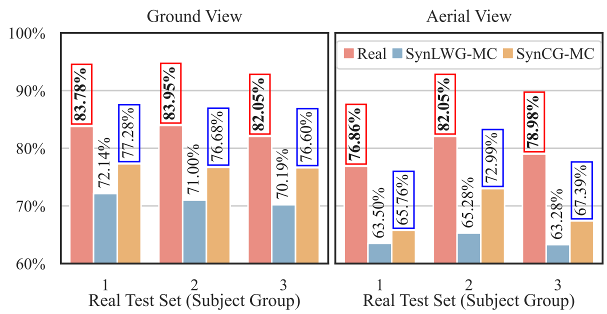

Figure 6. Evaluation of *SlowFast* trained on all datasets for *eleven classes*.

camera views most of the time. We believe that this result is because *MViT* and *X3D* are designed for high spatio-temporal resolution samples. Our input sample size of 64 frames with 224 px × 224 px was too small to take advantage of this functionality. Based on these initial results, we used only *SlowFast* for the remainder of the experiments.

We also observe that the ground view always performs better than the aerial one (Figure 5). We believe this is because the aerial view is recorded with a moving camera, making it more challenging to learn the task-specific spatio-temporal features. The ground cameras provide two views and, therefore, twice as much data. The overhead camera angle may also degrade performance.

Does the synthetic data rendering technique affect the model performance? A unique feature of our dataset suite is that it is generated using two different rendering techniques. We compared the performance of *SlowFast* trained on three datasets, *Real*, *SynCG-MC* and *SynLWG-MC* for eleven activities. The models were tested on three real test groups, excluding overlaps with the training data.

The results in Figure 6 show that the models trained on real data consistently outperform those trained on synthetic data by 4-8 %. *SynCG-MC* data performs strictly better than *SynLWG-MC*, suggesting the distribution gap between the data generated by the standard CG-based method for ren-

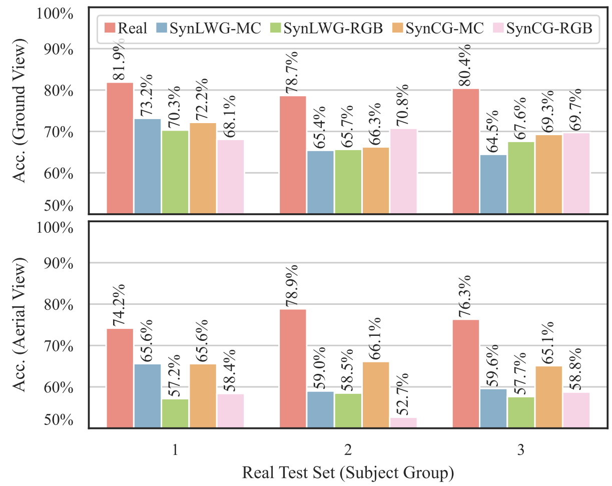

Figure 7. Evaluation of *SlowFast* trained on real and synthetic datasets for *five gesture classes – 15 subjects*. Comparison between different motion data sources.

dering and the real data is smaller than the one between the *LWG* approach and the real data.

Does improving the quality of the motion in the synthetic training data improve the performance? Another property of our dataset suite is that we used two different motion sources to generate the synthetic dataset. Motion capture systems collect the skeletal data in 3D, whereas the motion extracted from the RGB video extrapolates the 3D information from the 2D data. Therefore, the synthetic data generated using motion capture data has superior quality. We compared the performance of *SlowFast* trained on *Real*, *SynCG-MC*, *SynCG-RGB*, *SynLWG-MC*, and *SynLWG-RGB*. Because *SynCG-RGB* has only 15 subjects, we also limited the other datasets to 15 subjects. We evaluated the models for the five gesture-only classes.

Figure 7 show that the model trained on real data performs the best for both views. For ground view data, the motion source does not significantly influence model performance in both *CG* and *LWG*-rendered videos. In aerial data, motion capture-generated datasets outperform video-extracted motion datasets, likely because ground cameras focus on lateral axes, aligning with pose reconstruction methods. In contrast, aerial views encompass diverse angles, making the quality of the motion more important.

Can we combine synthetic training data with a limited amount of real training data to improve performance? To assess this hypothesis, we created two subsets from the real dataset that include 5- and 10-subject data. Then, we used them to train/fine-tune three models—a model untrained on any of our datasets and two pre-trained on *SynCG-MC* and *SynLWG-MC*. Our results (Figure 8) indicate that pre-training a model on synthetic data and subsequently fine-tuning it on a small amount of real data outper-

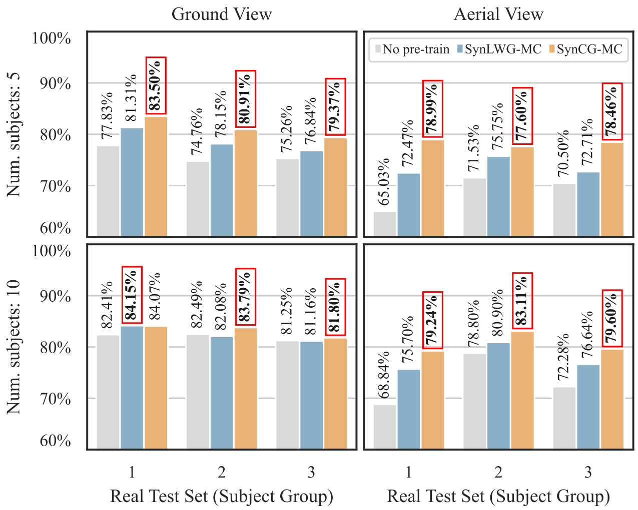

Figure 8. Accuracy of *SlowFast* pre-trained on synthetic data and fine-tuned on limited amounts of real data. Experiments are based on *eleven* activity classes.

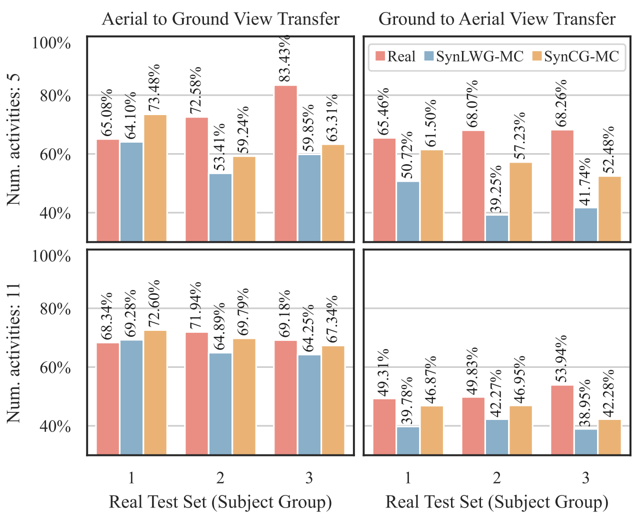

Figure 9. Camera view change analysis. The recognition accuracy delivered by *SlowFast* trained on *five* (gestures only) and *eleven* activity classes and tested on the opposing camera view data.

forms a model trained only on the exact small amounts of real data. That is valid for both camera views. Furthermore, the models pre-trained on *SynCG-MC* consistently outperform the ones pre-trained on *SynLWG-MC*. The improvement is considerably more subtle for the ground view when the larger (10 subjects) real subset is used. The overall limited diversity of the static ground view data most probably causes that effect, which underlines the importance of the synthetic HAR datasets for improving the performance on more dynamic scenes, such as the aerial view.

Can the models trained on one camera view transfer to a novel one? In this experiment, we evaluated the robustness of the model trained on our datasets to the change of camera viewpoints. We did that by cross-view evaluation. We ran two sets of experiments, one with the five gesture-only datasets and the other with all eleven behaviors for *Real*, *SynCG-MC* and *SynLWG-MC* datasets excluding any overlaps with the test set.

The results in Figure 9 show that the model trained on the aerial view is more robust to view change than the model trained on the ground view for all types of training data. We hypothesize that two factors are responsible for this difference in performance. First, while the ground cameras only see the subject from the front, drone cameras see the subject from all sides. This more varied viewpoint allows the model trained on aerial data to transfer to the ground data more easily than in the opposite direction. Second, the difference in the sizes of the ground and the aerial datasets play a role: because the ground view data is twice as large as the aerial view data, it is harder for the model trained on ground view data to transfer to the aerial view.

5. Conclusion

In this paper, we introduced a new *HAR* dataset suite containing real and synthetic data captured from ground and dynamic aerial camera perspectives. Synthetic data was generated using two rendering methods: standard computer graphics (*CG*) and neural-based rendering (*LWG*). We evaluated synthetic data performance against baseline models trained on real-world data, yielding the following findings: (1) Training data rendering method matters (*CG* outperforms *LWG*). (2) Motion quality is more crucial than rendering quality for model performance. (3) Fine-tuning models with a smaller batch of real data after pre-training on synthetic data improves performance, sometimes surpassing models trained solely on the full real dataset. (4) Synthetic data training with diversity enhances model robustness to changes in camera view compared to real data training. In summary, the quality of synthetic data is vital for bridging the domain gap with real-world data, and even small amounts of real-world data can boost performance through fine-tuning. Future experiments should explore broader activities, alternative synthetic data forms, and the impact of fine-tuning and data mixing ratios.

6. Acknowledgement

We thank Melanie Danver and Justin Macey for their valuable help in creating the datasets. This work has been funded in part by the DEVCOM Army Research Lab. Desislava Nikolova did her work while visiting Carnegie Mellon University, sponsored by the Bulgarian-American Commission for Educational Exchange Fulbright.

References

[1] AutoDesk. MotionBuilder. https://www.autodesk.com/products/motionbuilder/. Accessed: 2023-03-27. 5

[2] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 3, 4

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 3, 4

[4] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 4

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody Dance Now. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[6] Celso M. de Melo, Brandon Rothrock, Prudhvi Gurram, Oytun Ulutan, and B.S. Manjunath. Vision-based gesture recognition in human-robot teams using synthetic data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10278–10284, 2020. 2, 3, 4

[7] Celso M. de Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. Next generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, 26(2):174–187, 2022. 2

[8] César Roberto De Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López. Procedural Generation of Videos to Train Deep Action Recognition Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2594–2604, 2017. 2, 3

[9] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 4, 6

[10] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6824–6835, October 2021. 2, 4, 6

[11] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4, 6

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 4, 6

[13] Center for Research in Computer Vision at the University of Central Florida. UCF-ARG Data Set. https://www.crcv.ucf.edu/data/UCF-ARG.php. Accessed: 2023-03-27. 3

[14] Renderpeople GmbH. Renderpeople. https://renderpeople.com/. Accessed: 2023-03-27. 3, 5

[15] Poly Haven. HDRI Library. https://polyhaven.com/hdris. Accessed: 2023-03-27. 6

[16] Headquarters — Department of The Army. *Visual Signals — Field Manual No. 21-60*. US Army, 1987. 2

[17] Kristina Host and Marina Ivašić-Kos. An overview of human action recognition in sports based on computer vision. *Heliyon*, 8(6):e09633, 2022. 1

[18] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. *IEEE Access*, 11:9279–9294, 2023. 2, 3

[19] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 4

[20] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5607–5616, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. 2

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 4, 6

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 5252–5262, Piscataway, NJ, June 2020. IEEE. 2, 6

[23] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6

[24] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563, Nov. 2011. ISSN: 2380-7504. 3

[25] Viet-Tuan Le, Kiet Tran-Trung, Vinh Truong Hoang, and Huihua Chen. A comprehensive review of recent deep learning techniques for human activity recognition. *Intell. Neuroscience*, 2022, Jan 2022. 2

[26] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles. *2021 IEEE/CVF CVPR*, Jun 2021. 3, 4

[27] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 3, 4

[28] Rex Liu, Albara Ah Ramli, Huanle Zhang, Esha Datta, and Xin Liu. An overview of human activity recognition using wearable sensors: Healthcare and artificial intelligence. *CoRR*, abs/2103.15990, 2021. 1

[29] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 4, 5, 6

[30] Matthew Loper, Naureen Mahmood, and Michael J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6), Nov 2014. 4, 6

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), Oct 2015. 2, 6

[32] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 4

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[34] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 4

[35] Hieu H Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A Velastin. Video-based human action recognition using deep learning: a review. *arXiv preprint arXiv:2208.03775*, 2022. 4

[36] Oluwatoyin P. Popoola and Kejun Wang. Video-based abnormal human behavior recognition—a review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):865–878, 2012. 1

[37] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010. 1

[38] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home Action Genome: Cooperative Compositional Action Understanding. *2021 IEEE/CVF CVPR*, Jun 2021. 3, 4

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. 4

[40] Arun V Reddy, Ketul Shah, William Paul, Rohita Mocharla, Judy Hoffman, Kapil D Katyal, Dinesh Manocha, Celso M de Melo, and Rama Chellappa. Synthetic-to-real domain adaptation for action recognition: A dataset and baseline performances. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023. 2, 3, 4

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 4

[42] Gunnar A. Sigurdsson, Abhinav Kumar Gupta, Cordelia Schmid, Ali Farhadi, and Alahari Karteek. Actor and Observer: Joint Modeling of First and Third-Person Videos. *2018 IEEE/CVF CVPR*, pages 7396–7404, 2018. 3, 4

[43] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020. 3, 4

[44] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. Technical report, CRCV-TR-12-01, November 2012. 3

[45] 3Dflow s.r.l. 3D Flow Zephyr. https://www.3dflow.net/3df-zephyr-photogrammetry-software/. Accessed: 2023-11-03. 6

[46] Waqas Sultani and Mubarak Shah. Human action recognition in drone videos using a few aerial training examples. *Computer Vision and Image Understanding*, 206:103186, May 2021. 2, 3, 4

[47] Zehua Sun, Jun Liu, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, and Gang Wang. Human action recognition from various data modalities: A review. *CoRR*, abs/2012.11866, 2020. 2

[48] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 4

[49] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 4

[50] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021. 2, 3

[51] Yo whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 3

[52] Andrew S. Whitford, Emily Kim, Eni Halilaj, Keelan Enseki, Adam Popchak, and Jessica Hodgins. Sensor-based evaluation of physical therapy exercises. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 7556–7561, 2021. 1

[53] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). 5

[54] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[55] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4