2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

SynthAct: Towards Generalizable Human Action Recognition based on Synthetic Data

David Schneider[1,*], Marco Keller[1], Zeyun Zhong[1,2], Kunyu Peng[1], Alina Roitberg[3], Jürgen Beyerer[1,2] and Rainer Stiefelhagen[1]

***Abstract*— Synthetic data generation is a proven method for augmenting training sets without the need for extensive setups, yet its application in human activity recognition is underexplored. This is particularly crucial for human-robot collaboration in household settings, where data collection is often privacy-sensitive. In this paper, we introduce SynthAct, a synthetic data generation pipeline designed to significantly minimize the reliance on real-world data. Leveraging modern 3D pose estimation techniques, SynthAct can be applied to arbitrary 2D or 3D video action recordings, making it applicable for uncontrolled in-the-field recordings by robotic agents or smarthome monitoring systems. We present two SynthAct datasets: AMARV, a large synthetic collection with over 800k multi-view action clips, and Synthetic Smarthome, mirroring the Toyota Smarthome dataset. SynthAct generates a rich set of data, including RGB videos and depth maps from four synchronized views, 3D body poses, normal maps, segmentation masks and bounding boxes. We validate the efficacy of our datasets through extensive synthetic-to-real experiments on NTU RGB+D and Toyota Smarthome. SynthAct is available on our project page[4].**

I. Introduction

In the field of computer vision and machine learning, the ethical implications of data collection are increasingly important. Large-scale datasets like Kinetics [5], often scraped from the web, have advanced the field of human action recognition but come with drawbacks. These datasets frequently include videos collected without the knowledge or consent of the individuals involved, a practice raising concerns. Particularly in domestic or human-robot interaction scenarios, as robotic systems become more common in household settings, the need for privacy preserving data collection becomes more urgent. Focusing exclusively on body pose sequences in the data collection process represents a significant advancement in this direction which mitigates privacy concerns. While pose-based action recognition methods make use of this paradigm, they can not maintain the advantages of video based human action recognition.

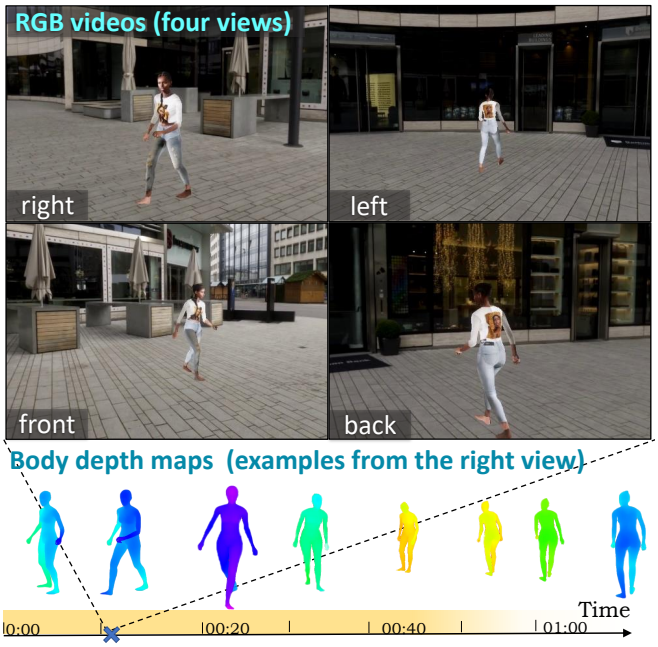

Fig. 1: Recording snapshots of the four camera views of SynthAct (top) and corresponding depth maps (below).

Creating a synthetic dataset for human behavior is a challenging task due to the wide range of variables such as body types, appearances, clothing styles, and backgrounds, both indoor and outdoor, as well as different activities and motion styles. Existing databases are often limited, focusing on specific use-cases [19], [50], [24] or offering a narrow range of categories [9], [14]. To address these limitations, we introduce *SynthAct*, a synthetic data generator for human action recognition. Developed using the Unity engine, SynthAct leverages the body shape and movement models from SMPL-H [39] to create a diverse range of human appearances.

For large-scale pretraining, we leverage SynthAct to introduce the **A**rchive Of **M**otion Capture **A**s **R**endered **V**ideos (AMARV), which is based on the AMASS [26] motion capture dataset. By incorporating state-of-the-art 3D pose estimation techniques, SynthAct can be applied to arbitrary human action datasets without the need for labeled ground truth poses. To demonstrate the ablicability of SynthAct to in-the-wild scenarios, we create the Synthetic Smarthome dataset as a counterpart to Toyota Smarthome [11], avoiding the use of ground truth poses.

This work was supported by the JuBot project which was made possible by funding from the Carl-Zeiss-Foundation and performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. A. Roitberg is supported by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy - EXC 2075 (SimTech).

*Corresponding author. (Email: david.schneider@kit.edu)
[1]Authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany.
[2]Authors are with the Fraunhofer IOSB, Germany.
[3]Author is with the Institute for Artificial Intelligence, University of Stuttgart, Germany.

[4]https://simplexsigil.github.io/synthact/

In summary, our contributions include:

- A data generation pipeline integrating 3D pose estimation with our Unity-based generator, SynthAct, reducing the dependency on real-world action video data.
- Two synthetic action datasets, AMARV and Synthetic Smarthome, each providing multiple perspectives and modalities such as RGB and depth.
- Comprehensive experiments that assess the effectiveness of large-scale pretraining on synthetic data and its applicability for fine-tuning on real-world tasks.

This work aims to reconcile the need for data collection in everyday living scenarios with the ethical obligation to respect individual privacy as well as to alleviate the requirement of large scale datasets for human action recognition.

II. Related work

Leveraging synthetic training data is well-established in object recognition, semantic segmentation and body pose and -shape estimation communities [48], [6], [30], [36], [35], [26], [51], [33], [47], [31], [16], [29] but is a rather new direction in human activity recognition. One of the first works addressing synthetic-to-real activity recognition was the framework of de Souza et. al. [14] which utilizes a combination of motion capture data and computer graphics techniques to generate single-view videos of 35 different actions. In the following years, multiple pipelines for generating synthetic activity examples have been developed for specific domains, including elderly assistance [19], recognition from new viewpoints for indoor daily living activities [50], [22] and unusual human activities in urban areas [24].

Table I provides an overview of publicly available datasets for synthetic-to-real activity recognition. However, these datasets have several limitations if they are intended to be used as a universal source of additional training data. They are often small in size, especially in regard to the number of captured classes, and are built for very domain-specific use cases, making them of limited use in a general scenario. For example, the largest available dataset, Eldersim [19], tackles an important application, yet its focus on indoor observation of elderly subjects is narrow. PHAV [14] and Mixamo [9] cover both indoor and outdoor activities but have only 35 and 14 action categories, respectively. The BEDLAM dataset [3] bears similarities to AMARV, as both utilize the AMASS dataset, RGB backgrounds, SMPL-based 3D bodies, and diverse clothing styles. While our data generator is built on the Unity perception library [49], BEDLAM employs Unreal Engine. Despite these technical distinctions, the two datasets serve different purposes: BEDLAM focuses on supplying training data for 3D human pose and shape estimation, whereas AMARV aims to be a comprehensive resource for synthetic training data in the domain of generalizable activity recognition. Due to its use case, BEDLAM does not focus on representing the full set of actions in AMASS and samples 2311 motions while AMARV covers the more than 10k motion sequences which are labeled by BABEL [33]. Since AMARV also provides simultaneous recordings from four perspectives, this results in a significantly larger dataset.

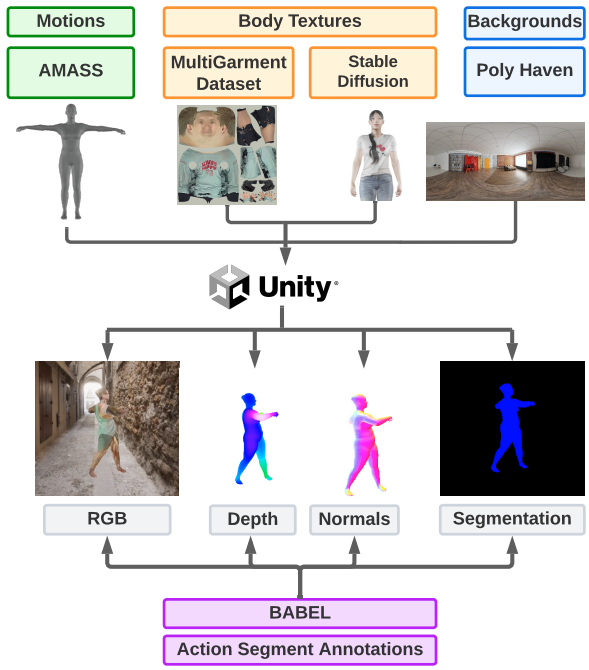

Fig. 2: Overview of our data generation pipeline using multiple publicly available data sources.

From the methodological perspective, past works on activity recognition from simulations either focus on (1) complementing real training data by mixing it with generated data (often posed as a domain adaptation problem) [14], [24], [19], [50], [41], or (2) learning action categories on synthetic data only [37], [27], [28] (synthetic-to-real domain generalization). [14] propose to mix virtual-world and real-world examples within the same minibatch, while [50], [22] follow a similar strategy to improve activity recognition from new viewpoints by augmenting real data with its synthetic variants from new perspectives. While we make use of simulation to address action recognition in the real world, we also need to mention the research of Puig et al. [32] which addresses learning compositions of actions within virtual domains only.

III. Implementation

An overview of our Unity-based data generation pipeline is provided in Figure 2. In our experiments we use the publicly available AMASS database [26] in combination with BABEL labels [33] or pose sequences extracted with 4DHumans [18] from Toyota Smarthome [11] as motion source. Diverse clothing styles and textures are either adapted from MultiGarment Net [2] or generated through stable diffusion [38] and adapted with Blender [7], and virtual humans are placed into heterogeneous 3D environments in form of license free HDRI images obtained from Poly Haven [1].

Motion diversity Using the AMASS dataset [26] as a source of action animation sequences for our data generator

TABLE I: Overview on existing synthetic action recognition datasets with various properties.

Dataset	Year	Actions	Motion Acq.	Motion source	Size / Dur.	Simult. Views (All)	Modalities
PHAV [14]	2017	35	MC	CMU Mocap	55h	1 (1)	RGB
ActionSim [24]	2021	5	MC	Mocap	0.1k / 1h	1 (6)	RGB,2DP
SynADL [19][1]	2021	55	PE	EtriAct.3D	462k	- (28)	RGB, 3DP
Sims4ADL [37]	2021	10	-	Ingame	1k / 10h	1 (25)	RGB
SURR. NTU [50]	2021	60	PE	NTU RGB+D	105k	3 (3)	RGB, OF, 2DP
SURR. UESTC [50]	2021	40	PE	UESTC	3k	8 (8)	RGB, OF, 2DP
Mixamo [9]	2022	14	PE	Kinetics	25k / 70h	8 (8)	RGB, 2DP
BEDLAM [3]	2023	-	MC	AMASS	10k	1 (n/a)	RGB, IS, D, 3DP
AMARV	2023	260	MC	AMASS	800k	$4\pm30°$(4)	RGB, IS, SN, D, 3DP
Synthetic Smarthome	2023	31	PE ([18])	Toyota Smarthome	45k	$4\pm30°$(4)	RGB, IS, SN, D, 3DP

IS Instance segmentation **OF** Optical Flow **D** Depth Maps **2DP** 2D-Poses **3DP** 3D-Poses **SN** Surface Normals **MC** Motion Capture **PE** Pose Estimation

[1] While eldersim is the dataset generator, SynADL is the dataset name.

provides high quality motion capture animations. We limit ourselves to the 10892 motion capture sequences which are provided with dense annotations by the BABEL language label dataset [33]. For Synthetic Smarthome we make use of 4DHumans, a 3D pose and shape estimation network. By estimating 3D pose from 2D video data, 4DHumans induces significant jitter, especially on the z-axis of the root node as well as false positive recognitions. We make use of simple thresholding and SLERP interpolation smoothing to restrict to higher quality motions.

Body model and animation AMASS provides its motions in form of the SMPL-H body model [39]. SMPL-H is a parametric model with Shape Blend Shapes, which enable the depiction of people with varying body proportions. In our dataset we specifically try to address discriminating biases and therefore actively cover a wide range of body shapes by randomly varying blend shape parameters within manually chosen bounds in order improve the generalization of action recognition models.

Body and background textures Our main source of clothing textures is the data published by Multi Garment Net [2] which includes 98 body clothing textures. A downside of this data is that it mainly depicts young to middle aged males, mostly caucasian and always with exactly the same face. In order to address this, we generate front and back views of realistic body textures with stable diffusion [38] for eight younger and eight older female characters with varying ethnic backgrounds. Part of our dataset consists of scrambled textures; we crop individual subregions of full body textures like left and right arm or leg and randomly mix them with each other, resulting in a total of 1240 different textures. We made sure to balance male and female source textures for this process. To set up multiple different background scenes for the motions, we make use of a total of 201 HDRI images from Poly Haven [1] as skybox backgrounds which we selected to represent a roughly equal amount of 103 indoor scenes and 98 outdoor scenes.

Varying lighting Variations in lighting are known to have a major impact on the quality of vision based recognition systems[20]. In our scenes, part of the lighting comes from HDRI images, the intensity can be adjusted to create strongly or weakly lit environments. However, the background light is not sufficient to properly illuminate the person. Therefore, four light sources were integrated at the positions of the cameras to obtain similar lighting conditions from all angles. In order to create diverse and realistic lighting settings, we randomly parameterise the foreground and background lighting intensity within manually selected bounds.

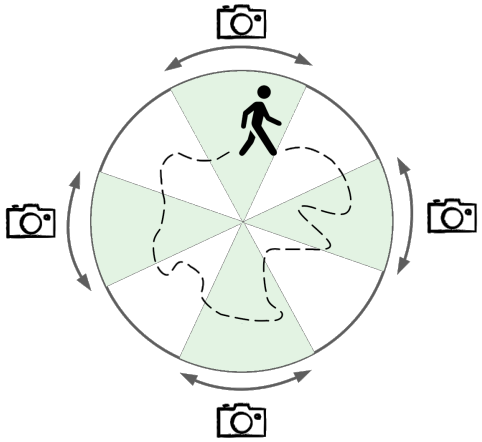

Fig. 3: Illustration of the positioning of the four cameras.

Multi view recordings One of the features of our synthetic data generator is the use of four simultaneously recording cameras (front, left, right and back camera). While keeping the actor in the image center, each camera is allowed to randomly position itself for each individual clip within $\pm30°$ in a 2D plane, resulting in a large range of possible camera angles, see also Figure 3. It is possible to recover the exact angle of a camera from recorded metadata, allowing for a finer differentiation of multiple views. We evaluate the covered space of each animation before rendering it and then arrange the cameras around the center of that area with a minimal distance, ensuring that an actor coming up close to the camera will never walk past it.

Multi-modal data output SynthAct captures depth infor-

mation, 2D and 3D bounding boxes, body keypoints, pixel normals and segmentation maps for each person in the scene in addition to the ground truth labels of the animations. Furthermore, meta-information such as camera and light source pose, light source intensity values and background information are saved frame wise.

Raw depth data requires significant amounts of storage up to a point which is infeasible for a dataset like AMARV. To reduce the required storage space and enable us to use common video compression algorithms, we apply the slightly adapted approach of Sonoda et al. [45] to convert depth values to RGB by mapping them to hues. The mapping induces a resolution of about 1530 different values. To minimize the loss of accuracy during conversion, we analyze the motion range in the z-axis and perform z-cropping, meaning that we map the closest and furthest point in the scene to the available range. By doing so, we achieve a varying precision for our videos ranging from sub-millimeter precision for motions with little z-movement to 10 mm precision for motions with 15 m of z-movement. Both values are within the expected precision of a modern real world depth estimation camera.

IV. Experiments

We aim to provide a pipeline for generating synthetic human action videos from in the wild human action data in order to partially or fully replace it during training. In section IV-A, we list results on pre-training on our large scale dataset AMARV. Fine-tuning on NTU RGB+D for both, RGB and depth based action recognition is evaluated in Section IV-B and Section IV-C. In Section IV-D we make use of Synthetic Smarthome and supplement as well as replace real world data with the intention to maintain performance and in Section IV-E we analyze the domain gap between data generated by SynthAct and real world datasets.

Architectures and Training We make use of multiple neural network architectures for our experiments. For end-to-end pre-training on AMARV in Section IV-A) as well as for downstream experiments in Section IV-B, we select a commonly used convolutional neural network, X3D-S, as well as a commonly used visual transformer architecture, MViTv2-S. For each network we choose the small configuration which allows for training on AMARV in a reasonable amount of time without access to large scale hyperscalers. In Section IV-D we also use MViTv2-S, but since we focus on synthetic data in the downstream dataset we initialize the model with Kinetics pre-trained weights. For feature extraction based experiments in IV-E as well as depth based action recognition on AMARV and NTU RGB+D, we make use of the pre-trained Omnivore[17] network which is based on Swin-B. All our experiments can be reproduced with four NVIDIA H100 or A100 GPUs or comparable. Unless noted otherwise, fine tuning is performed for 100 epochs with batch size 128 for X3D-S and 48 for MViTv2-S, we provide experiment setup scripts with further details.

Datasets For our experiments we make use of Toyota Smarthome [11] and AMASS [26] in combination with BABEL [33] as motion source and for experiments. NTU RGB+D [42] is used as downstream dataset.

A. AMARV Pre-Training

TABLE II: Results of supervised training on AMARV.

Model	Init	Accuracy Top-1	Accuracy Top-5	Bal. Acc.
Val-as-test-260				
X3D-S		41.35	73.10	12.1
MViTv2-S	MF-A	41.3	73.4	12.7
Val-as-test-120				
X3D-S		42.2	71.7	21.8
MViTv2-S	MF-A	42.3	72.4	23.2
Val-as-test-60				
X3D-S		44.8	75.4	28.8
MViTv2-S	MF-A	45.1	76.3	28.5

The BABEL annotations from Punnakkal et al. [33] provide more than one label for many AMASS segments which does imply a multi-label problem but the authors decided to ignore the multi-label property and instead chose to sample segments multiple times, once for each possible label. We follow them in this decision. This inhibits a perfect Top-1 score on AMARV, since a clip will be evaluated multiple times with different labels. Likewise to [33], we list Top-5 accuracy to mitigate this problem. BABEL does not provide the test set labels, for this reason we propose to publish results by using the official BABEL validation set for testing and refer to this protocol by *val-as-test*.

X3D-S is trained from scratch for 300 epochs, for MViTv2-S we apply the self-supervised masked feature approach from [55] for 230 epochs which we refer to as MF-A, followed by 70 epochs of supervised training.

We list our results in Table II, for all as well as the most frequent 120 or 60 classes. Despite their different architectures, both networks perform very similar on AMARV.

B. Fine-tuning on NTU RGB+D

TABLE III: Fine-tuning comparison of different synthetic data approaches on NTU RGB+D.

Method & Training		NTU 60 CS	NTU 60 CV
HPM_{RGB} + Traj [22]	S	75.8	83.2
SURR[50]	R	89.0	93.1
SURR[50]	S+R	89.6	94.1
X3D-S	A $\rightarrow$ R	88.54	97.2
MViTv2-S	MF-A $\rightarrow$ R	**93.8**	**98.6**

We consider AMARV as a dataset for large scale synthetic human action video pre-training and therefore evaluate models by fine-tuning performance on downstream datasets. For this we either list fine-tuning of X3D-S pre-trained on AMARV in a supervised way (A $\rightarrow$ R) or MViTv2-S with masked feature pre-training (MF-A $\rightarrow$ R) in Table III.

TABLE IV: Cross-view-subject evaluation on NTU RGB+D, comparison with SURREACT.

Training data		0°	45°	90°
SURREACT [50]	R	**86.9**	74.5	53.6
X3D-S	R	86.4	**77.8**	**60.4**
SURREACT [50]	S → R	84.1	77.5	66.2
X3D-S	**A → R**	89.9	81.8	**68.0**
MViTv2-S	**MF-A → R**	**94.4**	**84.5**	65.1
SURREACT [50]	S + R	**89.7**	**82.0**	**69.0**

TABLE V: Action recognition performance on AMARV using depth only.

Model	Top-1	Top-5
Swin-B (OMNIVORE)	22.30	53.44
Swin-B (AMARV)	**37.22**	**72.85**

On NTU RGB+D we can compare to the experiments from Varol et al. [50] who publish full fine-tuning results based on mixed synthetic SURREACT and real data training (S + R). X3D-S pre-trained on AMARV outperforms [50] on the cross-view split and performs almost equal on the cross-subject split. MViTv2-S significantly improves over the results of [50] on both NTU RGB+D splits. Note, that SURREACT [50] make use of mixed synthetic and real data for training while we only use synthetic data of AMARV for pretraining.

Cross-view-subject comparison. Varol et al. [50] specifically designed SURREACT to improve cross-view performance. They extended the cross-subject evaluation protocol on NTU RGB+D by only allowing training on front view videos while testing individually on the 0°, 45° and 90° views. Our results are listed in Table IV, we do not use synthethic data on NTU RGB+D but only make use of our synthetic multi-view pretraining dataset. X3D-S with synthetic pre-training on AMARV (A → R) not only outperforms the results of [50] on the comparable synthetic pre-training based setting (S → R) but even performs on par with mixed synthetic and real fine-tuning performed by [50] (S + R). MViTv2-S with self supervised pre-training on AMARV significantly outperforms all other methods on the front view but is not as good for the 90° view.

C. Depth-based Fine-Tuning

In this section, we describe our depth-based experiments on NTU RGB+D. We use OMNIVORE-B [17], which is based on Swin-B [23], as our baseline model. OMNIVORE is pre-trained jointly on three datasets with different visual modalities: ImageNet-1K (images), Kinetics-400 (videos) and SUN RGB-D (images and depth maps, but no humans). We use the disparity maps as input for the Swin-B model. Specifically, we first reconstruct depth maps from the depth videos coded in RGB format, and take the inversed depths, i.e., disparity maps, as input for the model. We use standard image augmentation methods, such as, RandAugment [8], Mixup [57], and Random Erasing [58], following [17]. First,

TABLE VI: Action recognition performance on NTU RGB+D [42] using depth only. Results demonstrate that the models pre-trained on AMARV are more generalizable and show superior fine-tuning performance.

Method	Cross Sub.		Cross View	
	Top-1	Top-5	Top-1	Top-5
TSN [54]	73.00	–	68.32	–
Dhiman et al. [15]	68.30	–	72.40	–
HPM_{3D} [22]	71.50	–	70.50	–
DDIs [34]	84.00	–	82.06	–
Wang et al. [53]	87.73	–	87.37	–
Swin-B (OMN., frozen)	46.89	78.50	45.24	78.11
Swin-B (AMA., frozen)	62.01	91.20	63.70	92.00
Swin-B (OMNIVORE)	92.80	**99.42**	92.84	99.61
Swin-B (AMARV)	**92.89**	99.41	**93.38**	**99.65**

Fig. 4: Gain of pre-training on AMARV over OMNIVORE baseline on NTU RGB+D [42] using depth only. We plot the gain in per-class Top-1 accuracy for the top ten and bottom ten classes. Pre-training on AMARV improves the Top-1 accuracy on 52 out of the 60 total classes.

we compare the Swin-B model (initialized with OMNIVORE-B weights) on AMARV with full fine tuning, further referred to as Swin-B AMARV or frozen backbone fine-tuning. Results in Table V show that full fine-tuning (AMARV) is necessary due to the large domain gap between AMARV and pre-training datasets of OMNIVORE, fine tuning of the patch embeding as well as the classifier alone (OMNIVORE) does not provide good performance.

In order to assess the generalization capability, we carry out experiments on the NTU RGB+D [42] dataset using the masked depth maps provided by the authors. We first freeze the main portion while keeping the patch embedding layer and classifier of the two models, i.e., Swin-B (AMARV) and Swin-B (OMNIVORE), trainable. Results in the middle compartment of Table VI show that the model fine-tuned on AMARV is more generalizable, demonstrating the effectiveness of the proposed dataset. Additionally, we compute

TABLE VII: Toyota Smarthome comparison on cross-subject and cross-view 1 splits.

Methods	Modality		mPCA (%)	
	Pose	RGB	CS	CV1
P-I3D[10]	✓	✓	54.2	35.1
Separable STA [11]	✓	✓	54.2	35.2
VPN [13]	✓	✓	60.8	43.8
VPN+SSTA-PRS [56]	✓	✓	65.2	-
MMNet [4]	✓	✓	70.1	37.4
VPN++ [12]	✓	✓	71.0	-
LSTM [25]	✓	×	42.5	13.4
MS-AAGCN [44]	✓	×	56.5	-
2s-AGCN [43]	✓	×	57.1	22.1
5C-AGCN+SSTA-PRS [56]	✓	×	62.1	22.8
DT [52]	×	✓	41.9	20.9
I3D [5]	×	✓	53.4	34.9
AssembleNet++ [40]	×	✓	63.6	-
VPN++ [12]	×	✓	69.0	-
Ours (R)	×	✓	**71.2**	42.2
Ours (S + R)	×	✓	70.5	**47.8**

the per-class Top-1 accuracy of these two models, and plot the gain for the top ten and bottom ten classes in Figure 4. Furthermore, we fine-tune these two models fully on the NTU RGB+D dataset, and show the results in the bottom compartment of Table VI. While both models outperform all existing depth-based methods in all settings, the model pre-trained on AMARV shows even superior results over the OMNIVORE baseline, especially in the cross view setting.

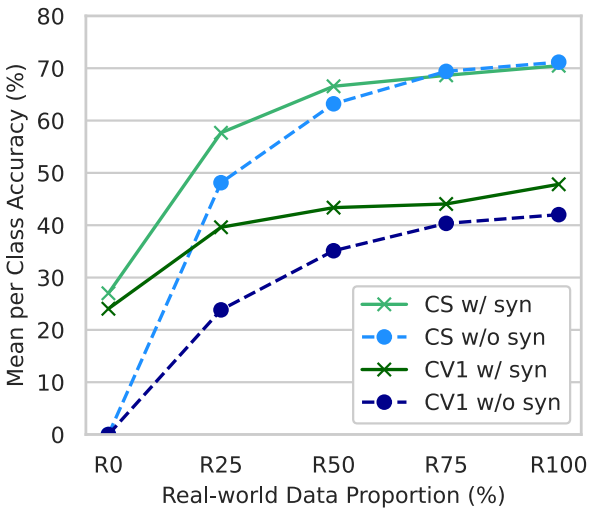

Fig. 5: Comparison between synthetically supplemented training with real-world-only experiments considering varying shares of real world data on Toyota Smarthome.

D. Synthetic data in Downstream Datasets

This experiment assesses if synthetic data can minimize the need for real-world data in downstream fine-tuning, particularly in privacy-sensitive, indoor scenarios like Toyota Smarthome. Figure 5 shows performance retention on real world data reduction across two training splits: *cross-subject*, with a larger multi-view training set, and *cross-view one*, only offering a single camera perspective for training. Our generator's multi-view video output from single view input poses notably boosts performance in the sparse cross-view setting, allowing almost 75% reduction in real-world data (remaining with only 469 real world samples) while mostly maintaining real world performance. In Table VII we compare our results with existing state-of-the-art approaches for Toyota SmartHome. Even on comparison with multi-modal methods in inference, the additional synthetic data leads to a significant improvement on the cross-view 1 split.

E. Synthetic-to-Real Domain Similarities

TABLE VIII: Similarity of synthetic datasets (including our proposed AMARV dataset) and three real datasets with OMNIVORE features.

Real	**Synthetic**	**MMD**	**KL-Div**	**JS-Div**
NTU RGB+D [42]	Surreact	**24.89**	393.89	**67.05**
	Mixamo	51.44	502.72	103.96
	AMARV	33.81	**300.68**	74.97
UCF-101 [46]	Surreact	33.28	501.58	118.98
	Mixamo	38.49	518.28	141.23
	AMARV	**20.40**	**353.08**	**104.98**
HMDB-51 [21]	Surreact	32.79	539.80	88.85
	Mixamo	37.99	414.55	86.91
	AMARV	**18.27**	**195.61**	**50.52**

In this section we assess the domain gap between our data generator and real world datasets as well as existing synthetic datasets. Using the OMNIVORE Swin-L model [17] as a feature extractor, we quantify domain discrepancy with MMD, KL-Div, and JS-Div metrics. Videos were segmented into overlapping windows (32 frames, stride 16) for feature generation and features were generated for HMDB-51, NTU RGB+D, UCF101, AMARV, Mixamo, and Surreact. Table VIII shows AMARV minimizes domain gaps, except against NTU RGB+D where SURREACT [50] excels. Note that SURREACT renders its synthetic data onto the original NTU RGB+D backgrounds.

V. CONCLUSION

We introduce SynthAct, a versatile synthetic action video generator developed using the Unity game engine and publicly accessible data. When integrated with 3D pose estimation techniques, SynthAct is applicable to a wide range of human motion recordings. Models pre-trained on our generated data achieve state-of-the-art performance on two of the three SURREACT cross-view-subject metrics, as well as on depth-based action recognition on the NTU RGB+D dataset. Additionally, we demonstrate the efficacy of synthetic data in reducing the need for real-world recordings, particularly in the context of the Toyota Smarthome dataset, also achieving state-of-the-art results. This is especially crucial in privacy-sensitive scenarios where minimal data collection is desired. With our generator we aim to drive human machine cooperation in environments with sparse action recordings by significantly decreasing the need for real world training data.

REFERENCES

[1] Poly haven hdri. http://web.archive.org/web/20230223210457/https://polyhaven.com/hdris. Accessed: 2023-03-01.

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019.

[3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023.

[4] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[7] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[9] Victor G Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1181–1190, 2022.

[10] Srijan Das, Arpit Chaudhary, Francois Bremond, and Monique Thonnat. Where to focus on for human action recognition? In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–80. IEEE, 2019.

[11] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 833–842, 2019.

[12] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9703–9717, 2021.

[13] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 72–90. Springer, 2020.

[14] César Roberto de Souza12, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López. Procedural generation of videos to train deep action recognition networks. 2017.

[15] Chhavi Dhiman and Dinesh Kumar Vishwakarma. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Transactions on Image Processing*, 29:3835–3844, 2020.

[16] Mona Fathollahi Ghezelghieh, Rangachar Kasturi, and Sudeep Sarkar. Learning camera viewpoint using cnn to improve 3d body pose estimation. In *2016 fourth international conference on 3D vision (3DV)*, pages 685–693. IEEE, 2016.

[17] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022.

[18] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023.

[19] 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*, pages 1–1, 2021.

[20] Amir Kolaman, Dan Malowany, Rami R Hagege, and Hugo Guterman. Light invariant video imaging for improved performance of convolution neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6):1584–1594, 2018.

[21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[22] Jian Liu, Hossein Rahmani, Naveed Akhtar, and Ajmal Mian. Learning human pose models from synthesized data for robust rgb-d action recognition. *International Journal of Computer Vision*, 127:1545–1564, 2019.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[24] Dennis Ludl, Thomas Gulde, and Cristóbal Curio. Enhancing data-driven algorithms for human pose estimation and action recognition through simulation. *IEEE transactions on intelligent transportation systems*, 21(9):3990–3999, 2020.

[25] Behrooz Mahasseni and Sinisa Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3054–3062, 2016.

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019.

[27] Zdravko Marinov, Alina Roitberg, David Schneider, and Rainer Stiefelhagen. Modselect: Automatic modality selection for synthetic-to-real domain generalization. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 326–346. Springer, 2023.

[28] Zdravko Marinov, David Schneider, Alina Roitberg, and Rainer Stiefelhagen. Multimodal generation of novel action appearances for synthetic-to-real recognition of activities of daily living. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11320–11327. IEEE, 2022.

[29] Dennis Park and Deva Ramanan. Articulated pose estimation with tiny synthetic videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 58–66, 2015.

[30] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018.

[31] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012.

[32] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018.

[33] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021.

[34] Ziliang Ren, Qieshi Zhang, Xiangyang Gao, Pengyi Hao, and Jun Cheng. Multi-modality learning for human action recognition. *Multimedia Tools and Applications*, 80:16185–16203, 2021.

[35] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2213–2222, 2017.

[36] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[37] Alina Roitberg, David Schneider, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelhagen. Let's play for action: Recognizing activities of daily living by learning from life simulation video games. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser,

and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
[39] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017.
[40] Michael S Ryoo, AJ Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 654–671. Springer, 2020.
[41] David Schneider, Saquib Sarfraz, Alina Roitberg, and Rainer Stiefelhagen. Pose-based contrastive learning for domain agnostic activity representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3433–3443, 2022.
[42] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.
[43] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.
[44] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020.
[45] Tetsuri Sonoda and Anders Grunnet-Jepsen. Depth image compression by colorization for intel® realsense™ depth cameras. *Intel RealSense*, 2020.
[46] 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.
[47] Baochen Sun and Kate Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, volume 1, page 3, 2014.
[48] Min Sun, Pushmeet Kohli, and Jamie Shotton. Conditional regression forests for human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3394–3401. IEEE, 2012.
[49] Unity Technologies. Unity Perception package. `https://github.com/Unity-Technologies/com.unity.perception`, 2020.
[50] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. In *IJCV*, 2021.
[51] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017.
[52] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.
[53] Huogen Wang, Zhanjie Song, Wanqing Li, and Pichao Wang. A hybrid network for large-scale action recognition from rgb and depth modalities. *Sensors*, 20(11):3305, 2020.
[54] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*, 2016.
[55] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.
[56] Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, Gianpiero Francesca, and Francois Bremond. Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2363–2372, 2021.
[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
[58] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.