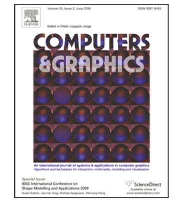
Preprint (2024)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag



NOVAction23: Addressing the Data Diversity Gap by Uniquely Generated Synthetic Sequences for Real-World Human Action Recognition

ARTICLE INFO

Article history:
Received May 8, 2024

Keywords: Human Action Recognition, Data Diversity Gap, Synthetic Data, Procedural Generation

ABSTRACT

Recognition of human actions using machine learning requires extensive datasets to develop robust models. Nevertheless, obtaining real-world data presents challenges due to the costly and time-consuming process involved. Additionally, existing datasets mostly contain indoor videos due to the challenges of capturing pose data outdoors. Synthetic data have been used to overcome these difficulties, yet the currently available synthetic datasets for human action recognition lack photorealism and diversity in their features. Addressing these shortcomings, we develop the NOVAction engine to generate highly diversified and photorealistic synthetic human action sequences. We use NOVAction to create the NOVAction23 dataset comprising 25,415 human action sequences with corresponding poses and labels. In NOVAction23, the performed motions and viewpoints are varied on the fly through procedural generation, to ensure that, for a given action class, each generated sequence features a distinct motion performed by one of the 1,105 synthetic humans captured from a unique viewpoint. Moreover, each synthetic human is unique in terms of body shape (height and weight), skin tone, gender, hair, facial hair, clothing, shoes and accessories. To further increase data diversity, the motion sequences are rendered under various weather conditions and at different times of day, across three outdoor and two indoor settings. We evaluate NOVAction23 by training three state-of-the-art recognizers on it, in addition to the NTU 120 dataset, and corroborating using real-world videos from YouTube. Our results confirm that the NOVAction23 dataset can improve the performance of state-of-the-art human action recognition.

1. Introduction

The analysis of spatio-temporal features is a crucial aspect of understanding videos. To leverage these features, deep architectures including convolutional neural networks (CNNs) have been widely used [1]. Such approaches require a comprehensive training process that can only be achieved with the availability of large datasets. For this, the lack of task-specific data poses a difficult challenge, even more so in the domain of human action recognition [2], which is a complex computer vision problem that requires careful consideration of both the data and the classifier.

Despite extensive research, the performance of human action recognition systems is still problematic. The main reason is the complexity of processing sequences containing diverse human actions, *s.t.*, each person performs actions uniquely, and each sequence is captured with distinct camera views. Training a bias-free model with high generalization capability requires large amounts of data with diversity in actions, viewpoints and subjects. This cannot be easily achieved with real-world datasets, as providing such diverse data in large volumes with accurately annotated labels is quite a challenge.

Large action datasets Kinetics-400 [3] or Kinetics-700 [4], which are curated from real-world videos, provide a wide variety of data made up of image sequences without explicit pose information. Using image-only data in training can lead to problems such as representation bias. To illustrate, if there is a soccer net in the video background, the action could be directly inferred as playing soccer [5]. There have been attempts at capturing large human action datasets in real-world scenes with explicit pose information, such as NTU RGB+D (NTU 60) [6] and its extended version NTU RGB+D 120 (NTU 120) [7], but their data variety have been limited due to having low actor count (40 in NTU 60 and 106 in NTU 120). In addition, they feature only a handful of backgrounds (classrooms, campus gardens, and places in between).

To address the problem of data diversity, we present a versatile synthetic data generation engine named NOVAction, which can create massive human action datasets by generating arbitrarily large number of human action sequences, each unique in terms of acting human, acted motion and camera viewpoint, with pixel-accurate pose information and attribute labels. To this end, NOVAction extends the photorealism of the previous work [8] by including more stable illumination and improved post-processing, offers an additional indoor scene for more diverse backgrounds and lighting conditions, and features a procedural animation system to achieve motion diversity.

We use the NOVAction engine to generate the NOVAction23 dataset consisting of 25,415 unique human action sequences with corresponding poses and labels (available at https://github.com/celikcan-cglab/NOVAction23). While there have been previous synthetic datasets [9–13] that addressed the data annotation problem with automatically generated labels and pose information, these have had limited diversity in terms of camera viewpoints, subjects or motion characteristics. NOVAction23 is a comprehensive photorealistic dataset that specifically addresses these shortcomings by providing sequences of human actions in 20 action classes captured from 125 different base views and performed by 1,105 synthetic humans in five different scenes, three of which comprise expansive outdoor environments, providing a diverse array of backgrounds. Furthermore, the acted motions and the base views are varied on the fly through procedural generation, so that, for a given animation class, each generated action sequence features a unique motion acted by one of the 1,105 synthetic humans captured from a unique viewpoint. Thus, NOVAction23 also addresses the arbitrary-view action recognition problem, the challenge of accurately recognizing human actions from any viewpoint [11, 14], more extensively than the previous synthetic human action datasets.

We demonstrate the efficacy of the NOVAction23 data in improving action recognition performance through experiments using three state-of-the-art action recognizers, namely TimeSformer (TS) [15], Temporal Pyramid Network (TPN) [16] and SlowOnly [17]. We also conduct an ablation study using different data partitions of NOVAction23 to evaluate the effects of lighting conditions, backgrounds and data modality, and to compare the performance of NOVAction23 with another synthetic dataset.

The remainder of this paper is organized as follows. Section 2 provides an overview of prior research on human action datasets, action recognition, and synthetic datasets. Details of the NOVAction engine and the NOVAction23 dataset are given in Sections 3 and 4, respectively. Section 5 presents the experiments where we test NOVAction23 in various settings. Finally, Section 6 outlines the limitations of the present work and concludes the paper.

2. Previous Work

Human Action Datasets. A number of RGB human action recognition datasets, such as UCF101 [18], HMDB51 [19], ActivityNet [20], Kinetics 400, 600 and 700 [3, 4, 21] have been made publicly available. AVA [22] and AVA-Kinetics [23] offer action labeling with bounding boxes. While some of these datasets are relatively high-scale, they suffer from representation bias [5]. In addition to the RGB datasets, several multimodal datasets have also been made available for understanding human activity, such as UTD-MHAD [24] and Diving48 [25], as well as several that are also multi-view, such as MMI [26], SYSU 3D HOI [27], UWA3D [28], FineGYM [29], NTU 60 [6], and NTU 120 [7]. These multimodal datasets provide depth maps and 3D skeletons estimated from the captures by the Kinect sensor [30]. As such, they are widely used for skeleton-based human action recognition, which reduces representational bias since skeletal data is devoid of any background information. However, these datasets have two major shortcomings. First, the 3D skeletons they provide are only estimated with Kinect 3D's own means, therefore are prone to errors [31]. Second, since Kinect, using infrared projection, can not capture depth images accurately in outdoor lighting [32], their data mostly consists of indoor backgrounds and lighting.

Synthetic Datasets. In recent years, synthetic datasets have been created for a variety of purposes, including autonomous driving and object recognition [33–37], person reidentification [38–40] and head pose estimation [41]. VirtualPTB1 [8] and PTAW217Synth [42] were procedurally generated by the NOVA framework for tracking people in normal and adverse weather conditions, respectively.

Synthetic data is also available to support human action recognition research, as real datasets are difficult to collect or assemble. SURREACT [11] provides non-photorealistic video sequences, utilizing 3D pose data provided by the NTU 120 dataset. ActionSim [9] data includes sequences in five action classes created with Unity. Sims4Action [10] offers recorded action videos from The Sims 4 video game featuring 10 action classes with eight different subjects. It features multiple examples per class, but the actions of the classes are nearly identical, as Sims 4 only features a handful of different animations per action. The ElderSim [12] platform used Unreal Engine 4 to generate KIST SynADL, which includes videos of elderly people performing daily activities in 55 classes. Even though they produced a large number of videos, the action variety of the dataset is limited by the motion capture animations of 100 individuals from different angles and times of the day. Mixamo Kinetics [13] is a hybrid dataset containing both synthetic and real data. The synthetic data was generated using six different pre-built avatars performing 14 classes of actions obtained from the Mixamo website.

Action Recognition. After the introduction of inflated 3D convolutional networks (I3D) [3, 43], 3D convolutional networks became the standard for action recognition tasks. Later, many models [17, 44–46] have been built with the same principles and outperformed the original I3D architecture. Recent classifiers TPN [16] and TS [15] have achieved better top-1 classification accuracy in Kinetics 400 [3] compared to priors.

In addition to the 3D convolutional and convolution-free networks, skeleton-based action recognition models using pose estimation of individuals as input have been proposed. PoseC3D [31] introduces a top-down pose extraction method to re-estimate the 2D skeletal information of the datasets since 3D skeletal information obtained from the Kinect sensor may be faulty in some cases. For human recognition, they utilize Faster R-CNN [47], while they use HRNet for pose estimation [48]. This approach aims not only to remove the erroneous information obtained from the Kinect but also to alleviate the domain adaptation problems that may arise from using different types of sensors.

3. NOVAction Engine

NOVAction is an expansion of the NOVA synthetic data generation framework [8]. Both were developed using Unity.

The original NOVA engine is a multifaceted framework for automatically generating arbitrarily large amounts of synthetic data for a wide range of low and high-level computer vision tasks. It can render realistic-looking virtual worlds containing procedurally generated humans together with pixel-level ground truth annotations, including body pose, bounding box, instance segmentation, semantic segmentation, depth map, and optical flow. In addition, NOVA can simulate various environmental factors such as different weather conditions and times of day and bring to life an exceptionally diverse set of unique humans at runtime using procedural generation.

In the following, we detail the extensions made to NOVA in order to realize the NOVAction engine featured in this paper.

3.1. Additional Scene and Lighting

NOVA engine is able to produce sequences in four different scenes (a town square, a suburban street a metropolitan urban district, and a subway station). To increase the variety of the generated data and the compatibility with datasets such as NTU, an office environment, including a lobby and a meeting room, was added. Similar to the existing environments, the new environment has multiple points where synthetic individuals are randomly spawned during data generation. Further, all environments have been configured to use real-time lightning, instead of the previously used baked lightning, for improved photorealism, as illustrated in the third row of **Fig. 1**.

3.2. Improved Image Post-Processing

The NOVA engine uses fast approximate anti-aliasing (FXAA) to advance image sharpness by sampling every pixel in a frame [49]. While FXAA efficiently improves image quality, it does not consider the following or previous frames when rendering the image. On the other hand, temporal anti-aliasing (TAA) [50] improves the sharpness for scenes with more flow compared to FXAA. Therefore, in NOVAction, we replaced

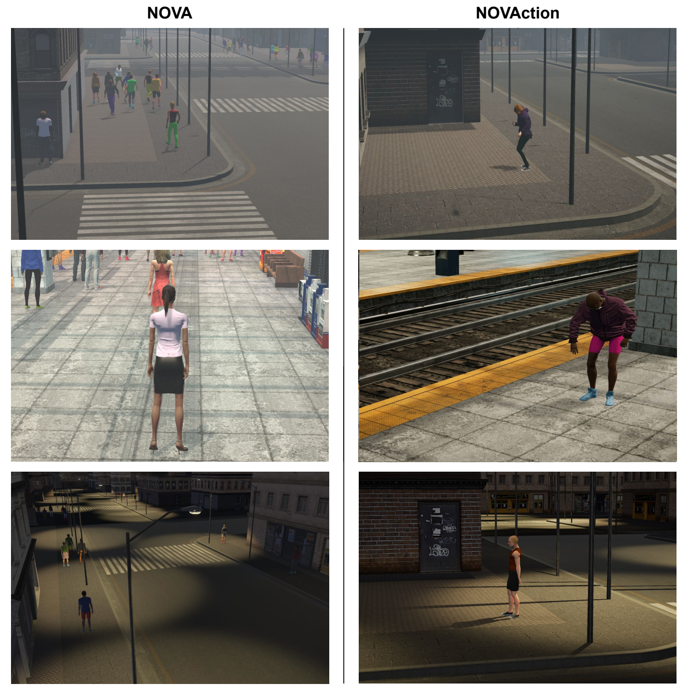


Fig. 1: Sample frames generated by NOVA (left) and NOVAction (right).

FXAA with TAA to acquire image sequences in enhanced quality. In addition, we have implemented bloom, color grading, eye adaptation, and vignetting, as illustrated in the second row of Fig. 1.

3.3. Procedural Animation System

The foremost improvement of NOVAction is the addition of a procedural animation system. We used 23 actions from the Mixamo library [51] and grouped these into 20 different action classes corresponding to the ones in the NTU 120 dataset, by

Table 1: Action class correspondences between the NTU 120 and NOVAction sequences.

NTU 120	NOVAction23	Description
A022	0, 2	Cheer up
A010	1	Clapping
A035	3, 14	Nod head (yes)
A006	4, 6	Pick up
A036	5	Shake head (no)
A038	7	Salute
A104	8	Stretch
A069	9	Thumb up
A009	10	Stand up
A103	11	Yawn
A023	12	Hand wave
A029	13	Tablet/phone interaction
A046	15	Back pain
A007	16	Throw an object
A037	17	Wipe face
A080	18	Squat
A043	19	Falling down
A049	20	Fan self
A102	21	Side kick
A027	22	Jump

using three actions out of 23 as alternatives for similarity in context to the implemented classes. The class correspondences are itemized in **Table 1**. These actions were reformatted to make them compatible with the synthetic human generation system of the NOVA engine, so that any synthetic human generated by NOVAction can perform the added 20 action classes. Sample frames for the action classes are given in **Fig. 3**.

For every individual Mixamo action, we procured animations with distinct subject arm space and speed settings that are commonly available in the Mixamo library. Each animation was acquired in four different versions: one with the fastest motion and widest arm space; one with the slowest motion and widest arm space; one with the fastest motion and narrowest arm space; and one with the slowest motion and narrowest arm space. Then, to generate each action sequence, these four animations were mixed using two-dimensional animation blend trees, where the two parameters were represented by the two axes of the tree and were randomly determined. The outcome of this process significantly augments the diversity of the generated data. As a result, NOVAction can generate distinctively unique actions in each action class, which sets it apart from synthetic action generation systems [9–13] that rely solely on premade motion-captured sequences, severely limiting the variety of performed actions. In addition, NOVAction can automatically produce the corresponding pose information in both 2D and 3D.

Providing a variety of actions was aimed at enhancing NOVAction23's realism by aligning it with real-world action data, thereby improving accuracy when employed as a training dataset for action recognition models, especially in uncommon scenarios. For example, although most side kick actions in reality are executed rapidly, some side kick sequences also exhibit individuals executing the action slowly. The presence of correspondingly timed training data can improve classification accuracy, particularly when used in conjunction with methods that involve pointed temporal inference.

4. NOVAction23 Dataset

It is essential to vary the attributes of classification datasets to improve their potential in model training with higher generalization capability. Our dataset encompasses a diverse range of motions, subjects, camera views and locations (i.e., backgrounds), providing a greater degree of variety in comparison to state-of-the-art real and synthetic datasets.

While most human action recognition datasets consist of sequences taken indoors, outdoor sequences are very limited. This can severely restrict action recognition performance in related cases, such as video footage captured with outdoor cameras. Therefore, we made it a point to generate more data using the outdoor scenes for NOVAction23.

In each scene, there exists five spawn points, at one of which a uniquely generated subject is spawned randomly. And, there are five base camera viewpoints for each spawn point. This brings about 125 base views in total. Once a camera is generated, it focuses directly on the generated subject. Finally, small random perturbations are made to the camera's view angle and position. Hence, the camera viewpoint is unique to each generated action sequence due to the random variations added on top of the base views.

Ensuring subject diversity in real-world datasets is typically challenging, especially in terms of recruiting and/or compensating subjects. When videos are collected via web scraping or similar means, there are usually ethical or legal issues regarding privacy and data protection [53]. We see that many public datasets are either taken down or significantly reduced over time due to these issues. NOVAction combines a large set of attributes (skin tone, gender, height, weight, hair, facial hair, clothing, shoes, accessories, etc.) by making use of several layers including a predefined set of categorizable, annotatable features as well as low-level randomizations on these features, to generate unique human models at runtime. This eliminates privacy concerns and significantly reduces the experimental budget.

Thanks to the diverse generation capabilities of the NOVAction engine, each synthetic human in the NOVAction23 dataset is truly unique. In total, 1,105 synthetic humans were generated. Every one of these synthetic humans performed each of the 23 actions in a specific scene at a specific time of the day. The actions performed were also uniquely varied on the fly, as described in Section 3.3. In this process, over three million raw images were generated in 1920x1080 resolution. The raw images were combined to create the 25,415 action sequences. Action class, environment attributes (weather, time, scene), and

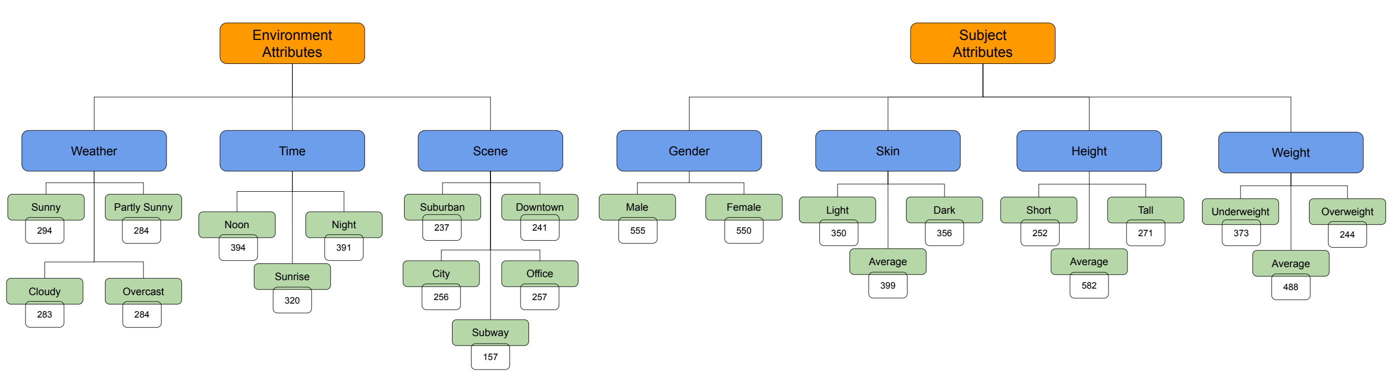


Fig. 2: The main attributes used in the NOVAction23 dataset and their distributions. Height, weight, and skin tone are not discrete values but are grouped into three sets for data labeling purposes.

Fig. 3: Sample frames from NOVAction23, NTU 120, and YouTube sequences for each action class. The first 15 actions show three frames from NOVAction23 and one frame from NTU 120. The last five actions show two frames from NOVAction23, one frame from YouTube, and one frame from NTU 120.

Table 2: Dataset comparison. § indicates that there is no clear statement about the characteristic. † indicates that the given values are base values and there are additional variations per sequence on top of the base values.

Dataset	Type	Scene Type (Scene Count)	Views	Subjects	Actions	Classes	Outdoor Scenes	Resolution	Pose	Videos
ActionSim [9]	Synthetic	2D (§)	2	§	§	5	No	1280x720	✓	100
Sims4Action [10]	Synthetic	3D (2)	24	8	10	10	No	640×368	✗	942
Mixamo [13]	Synthetic	2D (200)	8	6	14	14	Yes	512x512	§	24,533
SURREACT [11]	Synthetic	2D (§)	8	118	§	60	Yes	320x240	✓	105,503
KIST SynADL [12]	Synthetic	3D (4)	28	15	5,500	55	No	640×360	✓	462000
NTU 120 [7]	Real	Real (§)	155	106	§	120	No	1920x1080	✓	114,480
Smarthome [52]	Real	Real (§)	7	18	§	31	No	640×480	✓	16,129
NOVAction23	Synthetic	3D (5)	125†	1,105	23†	20	Yes	1920x1080	✓	25,415

certain synthetic human attributes (gender, skin, height, weight) were automatically labeled along with each generated sequence and became instantly usable. Fig. 3 demonstrates samples from NOVAction23, next to the samples from NTU 120 and YouTube in each action class, and **Fig. 2** gives the distribution of the generated data. Also, a video demonstrating sample action sequences from the NOVAction23 dataset in comparison to the ones from the NTU 120 and Youtube Action sets is provided as supplemental material.

In **Table 2**, we provide a comparison of NOVAction23 to the previously released human action recognition datasets. The table shows that the NOVAction23 dataset stands out especially with a large number of camera views and subjects (synthetic humans), and a high video resolution. It also includes 3D backgrounds of indoor and outdoor environments, photorealistic humans and illumination, bringing it closer to the real data compared to previous synthetic data. Accordingly, the whole dataset includes 25,415 unique action sequences. The complete attribute variety of the dataset is given in Fig. 2 along with their distributions within the 1,105 different settings, i.e., per synthetic human. As the distributions indicate, the variations of the attributes were kept balanced.

5. Experiments

To assess the capabilities of NOVAction23 in improving action recognition models, a series of experiments was conducted using the state-of-the-art action recognizers, as detailed below.

5.1. Datasets

In addition to the full NOVAction23 dataset described in the previous section, we made use of the following as training, test and validation data in our experiments.

NTU 120. We utilize the dataset as NTU 120 train and NTU 120 test with the original cross-subject training and testing split (53 subjects for training and 53 subjects for testing) as proposed in [7].

NTU 20. To have real data compatible with the NOVAction23 dataset, so that they can be deployed together in training and testing, we make use of a modified version of NTU 120 by retaining the original cross-subject partitioning (53 subjects for training and 53 subjects for testing) but removing the sequences for the 100 classes that are not present in NOVAction23. We denote the modified dataset NTU 20, which includes the 20 action classes that coexist in NOVAction23 (Table 1).

YouTube Action. We use this set mainly for validating the performance of the trained models with real-world videos. To this end, we have compiled a collection of 100 videos from YouTube for a set of five action classes (*stand up*, *jump*, *falling down*, *squat*, and *side kick*) selected from Table 1 (20 videos per class). The set was restricted to these five classes due to the limited availability of public videos with full body shots of individuals performing the actions. The collected videos were edited to ensure that each video covered a single action from start to finish, similar to the videos in NTU and NOVAction23. Videos that were not in 1920x1080 resolution were also scaled to 1920x1080 for compatibility. Since we used the YouTube Action videos for validation purposes only, they do not have designated training or test partitions.

SURREACT. To test the performance of NOVAction23 in comparison to other synthetically generated action data, we make use of the SURREACT (HMMR) dataset, as it contains 15 action classes that are also found in NOVAction23 and NTU 120. SURREACT consists of non-photorealistically animated videos of the NTU 120 pose sequences. In the evaluation, we keep the original training split [11], which contains 105,503 sequences from the first 60 action classes of NTU 120.

¼ NOVAction23. To assess the impact of the amount of data when training with NOVAction23, we also used a quarter of NOVAction23, i.e., the action sequences performed by a randomly selected set of 275 subjects (approximately a quarter) out of the total 1,105. This set was used for training only.

All experiments were conducted using a cross-subject setup, i.e., the training, testing, and validation partitions for each experiment included data from distinct groups of subjects.

5.2. Evaluation Setup

The overview of the setup that we used for the skeleton-based action recognition tests is given in **Fig. 4**. The experiments were performed on a cloud server with Intel Gold 5315Y

CPU, Nvidia RTX A6000 GPU, and 45 GB of RAM, utilizing MMAction2 [54], an open source video understanding toolbox based on PyTorch [55]. MMAction2 provides a variety of algorithms for different action recognition approaches, including skeleton-based, spatiotemporal, and RGB-based recognition. Additionally, it offers a wide range of data manipulation tools to facilitate loading and pre-processing.

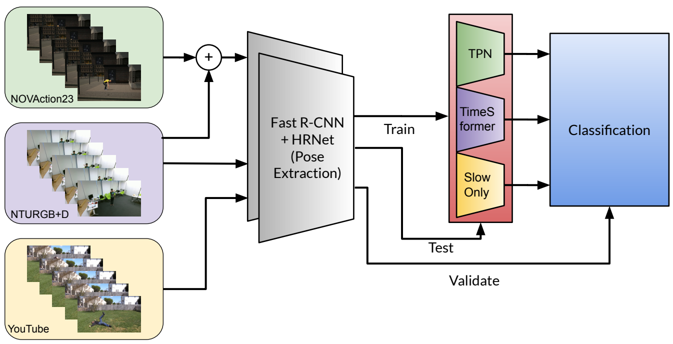


Fig. 4: Diagram of our experimental setup for skeleton-based action recognition.

In order to use the RGB-only video sequences of the YouTube Action set in 2D pose-based (skeleton-only) recognition, it is necessary to apply pose estimation to extract the explicit per-frame pose information. Although the NOVAction engine provided precise pose information as ground truth for the generated action sequences, we also utilized pose estimation for the NOVAction23 sequences to be used in all skeleton-based recognition tasks. This was done to avoid domain adaptation issues and to introduce noise to the otherwise sterile NOVAction23 data, which has been reported to improve overall classification performance [56]. It has also been shown that the use of pose estimation keypoints, rather than pre-processed ground truth pose information, can boost accuracy for skeleton-based action recognition by up to 1.5% when synthetic videos with uniformly sampled frames are used as training data [9]. To circumvent potential errors in the NTU 120 poses, which are commonly attributed to limitations of the Kinect sensor [31], we also applied pose estimation for the NTU 120 sequences. Accordingly, we estimated pose information from the RGB frames using the top-down pose estimation technique, as it provides more accurate pose estimation compared to bottom-up alternatives [57]. This approach, which we used for all skeleton-based models, consists of Faster R-CNN [47] with ResNet50 backbone as the person detector and HRNet [48] with ResNet50 backbone as the pose estimator. Both were pre-trained on the MS COCO dataset [58].

For skeleton-based action recognition, Duan et al. [31] proposed using SlowOnly [17] with the backbone of ResNet50 [59] as the action classifier. In addition to SlowOnly, we also included TPN [16] and TS [15] as alternative recognizer architectures since these networks perform better compared to SlowOnly in certain benchmark tasks, such as the RGB-only action recognition on Kinetics 400 [15, 16]. We used the same hyperparameters utilized in [31] for our TPN and SlowOnly experiments: a dropout rate of 0.5, stochastic gradient descent with a learning rate of 0.05, weight decay of 0.0003, the momentum of 0.9, batch size of 16 and cosine annealing [60] as the learning rate schedule. In TS, however, we did not use dropout as it lowers the accuracy. We used 32 as patch size, AdamW [61] as optimizer with a learning rate of 0.001 and weight decay of 0.1. All models were trained with 48 frames of uniformly sampled 64x64 heatmap inputs from 17 different joint points for 240 epochs.

For the RGB-only modality action recognition tests, we employed SlowOnly with a ResNet50 backbone, which was pre-trained on the Kinetics 400 dataset for 256 epochs. Input data for this modality consists of videos with 8 uniformly sampled frames and a resolution of 224x224, as pre-training for Kinetics 400 was conducted using these parameters. We opted for a constant learning rate of 0.001, the dropout rate of 0.5, and the batch size of 16 for the RGB-only modality models. We trained our RGB-only networks for either 15 or 30 epochs in different types of ablation experiments.

Other training, testing, and validation settings were used the same as the default settings provided in MMAction 2 version 0.24.1.

5.3. Benchmark with Different Recognizers

For the benchmark evaluation with the three action recognizers SlowOnly, TPN and TS, the experiments were conducted in the skeleton-only modality using a cross-subject data split and the results are reported in top-1 and top-5 classification accuracies. Since no large-scale human action data with the pose keypoint structure is currently available for pre-training, we report the results of training our classifiers from scratch in **Table 3**.

Our first evaluation was conducted to determine the benchmark performance of the recognizers on the NTU 120 dataset. The results are given in the NTU 120 Test column of Table 3. In this test, we trained each network using only the NTU 120 training data. Here, it is seen that TPN and SlowOnly have similar results, as TPN outperforms SlowOnly by a slight margin. However, the performance of TS is inferior, suggesting that it

Table 3: Test and validation results for the benchmark evaluation. The best top-1 and top-5 accuracies for each dataset are highlighted in bold, same as the fastest inference speed. Also, + NOVAction23 indicates that all NOVAction23 data is included in the training, while + ¼ NOVAction23 indicates that only a quarter of the NOVAction23 data is included.

		NTU 120 Test		NTU 20 Test		YouTube Action Validation		
Recognizer	**Trained On**	**top-1**	**top-5**	**top-1**	**top-5**	**top-1**	**top-5**	**video/s**
SlowOnly	NTU 120	0.84	**0.97**			0.25	0.66	3.3
	NTU 20			0.95	0.99	0.40	0.89	
	NTU 20 + ¼ NOVAction23			**0.96**	0.99	0.67	0.94	
	NTU 20 + NOVAction23			**0.96**	0.99	**0.75**	**0.97**	
TS	NTU 120	0.75	0.93			0.15	0.45	**3.7**
	NTU 20			0.92	0.99	0.50	0.92	
	NTU 20 + ¼ NOVAction23			0.92	0.99	0.49	0.85	
	NTU 20 + NOVAction23			0.93	0.99	0.63	0.88	
TPN	NTU 120	**0.85**	**0.97**			0.22	0.76	3.4
	NTU 20			0.95	0.99	0.46	0.81	
	NTU 20 + ¼ NOVAction23			0.95	0.99	0.61	0.92	
	NTU 20 + NOVAction23			0.95	0.99	0.68	0.93	

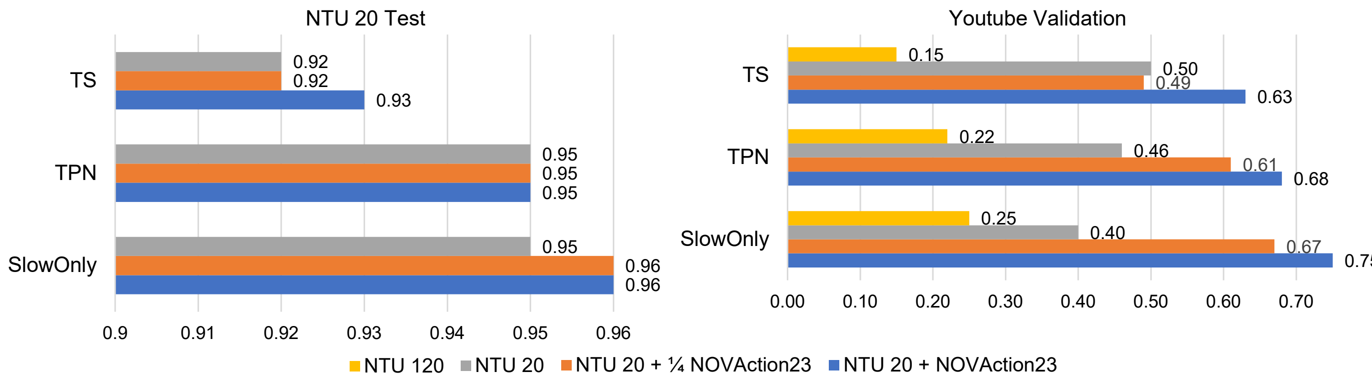


Fig. 5: Top-1 accuracies of the NTU 20 test (left) and the YouTube Action validation (right). Color labels indicate training sets. Also, + NOVAction23 indicates that all NOVAction23 data is included in the training, while + ¼ NOVAction23 indicates that only a quarter of the NOVAction23 data is included.

requires either more pre-training or additional training data in comparison.

The second evaluation involves training the action recognition models using only the 20 action classes that coexist in both the NTU 120 and NOVAction23 to determine whether training with the addition of the synthetic data generated by NOVAction can improve the action recognition performance on the real test data from a well-known dataset. On the NTU 20 test data, the first set of results for each model was obtained by training with only the NTU 20 training data, while the second one was obtained by training with ¼ NOVAction23, *i.e.*, a quarter of the NOVAction23 data, in addition to the NTU 20 training data, and the third one was obtained by training with all of the NOVAction23 data in addition to the NTU 20 training data. The results are shown in **Fig. 5** and in the NTU 20 Test column of Table 3. It can be seen that the addition of the synthetic training data slightly improves the top-1 and top-5 scores, which are already quite high without the addition. The best results are obtained with SlowOnly, closely followed by TPN.

We conducted the third evaluation to assess the ability of NOVAction23 to improve action recognition in-the-wild. For this, we validated our models with the YouTube Action data. The results are reported in Fig. 5 and the YouTube Action column of Table 3, which also includes inference speed. The most notable finding is that using our synthetic data in addition to the real data in training increased the scores by a substantial margin. This also implies that the pose extraction strategy proposed in [31] can also be used as a domain transformation method, since it allows synthetic video to directly improve inference accuracy on arbitrary videos without explicit pose information. In addition, TS performed best in terms of inference speed, but only provided decent accuracy on the YouTube Action set when all NOVAction23 data was used together with the NTU 20 training data. TS was outperformed by TPN and SlowOnly, so further experiments are needed to decide whether TS can be effectively used for real-world skeleton-based human action recognition tasks.

In previous studies [15, 16], both TPN and TS were reported to outperform SlowOnly in the tests performed on Kinetics 400 with the RGB-only modality. However, in our experiments involving skeleton-based modality, SlowOnly achieved the highest top-1 score in all data partitions except NTU 120. The results suggest that SlowOnly has superior generalization capabilities even when working with small datasets. It is hypothesized that this may be due to the complexity of the models employed. When performing recognition on RGB videos, the models use a large number of features compared to those based on the skeleton modality. As such, they can benefit from more complex architectures. Conversely, the skeleton-only modality does not require such complex architectures. In fact, using complex deep learning architectures for the skeleton-only modality can be counterproductive, resulting in reduced accuracy. With large training datasets, TPN produces results comparable to SlowOnly. In addition, TPN offers a modest improvement in inference speed over SlowOnly.

Our findings demonstrate that augmenting the training data with sequences exhibiting diverse motion characteristics captured from varied viewpoints improves action recognition performance on real-world videos. It is evident that although the models performed optimally on the NTU data, this does not necessarily translate into recognition accuracy in the real world, as observed in the YouTube Action validation results. Incorporating diverse synthetic data, in addition to real datasets such as NTU 120, can yield improved classification accuracy in-the-wild. Furthermore, it is also seen that using only a quarter of NOVAction23 in addition to the NTU 20 test data did not provide tangible benefits compared to the scenario where we added all of the NOVAction23 data. This suggests that for the best action recognition performance on real-world videos, the entire NOVAction23 dataset should be used in combination with a real-world dataset.

5.4. Ablation Study

In this section, we present the results of our ablation study using only the SlowOnly recognizer, which performed best in the benchmark evaluation detailed above. For this evaluation, SlowOnly was used with the pose estimator networks Faster R-CNN [47] and HRNet [48] for the skeleton-only modality, and without the pose estimator networks for the RGB-only modality. The same hardware setup was used as described in Section 5.2.

Lighting conditions can vary substantially between our indoor and outdoor 3D scenes. Global illumination was used in both settings to simulate sunlight, but its effect is less pronounced in the indoor scenes. Furthermore, environmental factors such as cloud cover and time of day have minimal impact on local lighting conditions in the indoor scenes, which are primarily lit by multiple light sources in close proximity to the subject. The hue of the lighting in the indoor scenes is also similar to that of the NTU 120 videos. On the other hand, the lighting in the outdoor scenes is mainly influenced by global lighting at sunrise and midday, as well as street lighting in the night sequences. These light sources are farther away from the subject than those in the indoor scenes.

Our first ablation study sought to examine the effects of these different lighting conditions on different data partitions of NOVAction23. To this end, we used seven different data partitions and trained a model in the RGB-only modality using each partition for 15 epochs, after which changes in accuracy became mostly negligible. The NOVAction23 Indoor and NOVAction23 Outdoor partitions each consisted of 8000 indoor and 8000 outdoor videos from NOVAction23, respectively. NOVAction23 Both consisted of 4000 indoor and 4000 outdoor videos from NOVAction23, while NTU 20 (8k) consisted of 8000 videos from the NTU 20 training split. NOVAction23 Indoor + NTU 20 consisted of 4000 videos from NOVAction23 Indoor and 4000 videos from the NTU 20 training split. Similarly, NOVAction23 Outdoor + NTU 20 consisted of 4000 videos from NOVAction23 Outdoor and 4000 videos from the NTU 20 training split. Finally, NOVAction23 Both + NTU 20 consisted of 2000 videos from NOVAction23 Indoor, 2000 videos from NOVAction Outdoor, and 4000 videos from the NTU 20 training split. With these splits, we ensured that all models were trained with a total of 8000 videos to avoid data imbalance. After training the models, we tested them on the entire NTU 20 test split and validated them on the entire YouTube Action set. The results are given in **Table 4** and **Fig. 6**.

Table 4: Results of the first ablation study, where we examine the effects of different illumination conditions of indoor and outdoor scenes of NOVAction23 on action recognition. The best accuracies achieved are given in bold. The Mean column shows the average of the top-1 scores of the NTU 20 test and the YouTube Action validation. In the Trained On column, next to the partition names, the amount of video taken from the designated source is shown in brackets.

	NTU 20 Test		YouTube Action Validation		Mean
Trained On (Data Size)	**top-1**	**top-5**	**top-1**	**top-5**	**top-1**
NOVAction23 Indoor (8k)	0.14	0.36	0.21	0.40	0.18
NOVAction23 Outdoor (8k)	0.14	0.42	**0.36**	**0.67**	0.25
NOVAction23 Both (4k Indoor + 4k Outdoor)	0.13	0.42	0.21	0.42	0.17
NTU 20 (8k)	**0.41**	**0.66**	0.26	0.61	0.34
NOVAction23 Indoor (4k) + NTU 20 (4k)	0.32	0.59	0.16	0.24	0.24
NOVAction23 Outdoor (4k) + NTU 20 (4k)	0.36	0.63	0.33	0.59	0.35
NOVAction23 Both (2k Indoor + 2k Outdoor) + NTU 20 (4k)	0.36	0.58	0.35	0.58	**0.36**

The results revealed that the outdoor videos from NOVAction23 more closely resemble the lighting and overall realism of the real-world videos than its indoor videos. Accordingly, using only the outdoor videos from NOVAction23 substantially improved action recognition in the real-world videos. In addition, NOVAction23 Outdoor outperformed NTU 20 (8k) in recognizing actions in real-world videos. This can be attributed in part to the fact that some actions in NTU 20 were poorly performed by the actors. For instance, the side kick actions performed in the NTU 20 videos were not executed with sufficient accuracy, with the actors lifting their legs only slightly. For best performance, the results suggest that both the indoor and outdoor videos from NOVAction23 should be used in conjunction with other datasets.

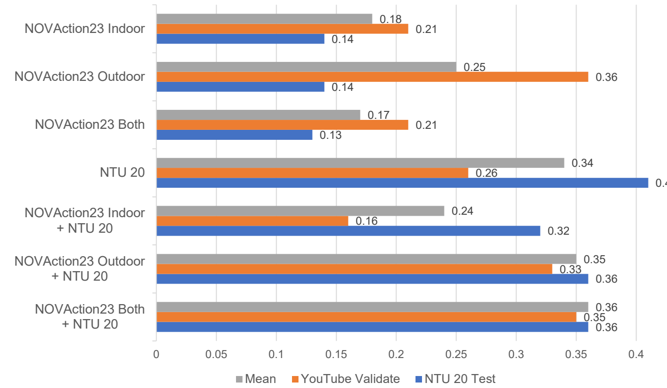


Fig. 6: Top-1 accuracies of the first ablation study.

In the second part, our goal was to compare the performance of the RGB-only and skeleton-only modalities when trained with our synthetic data. For both modalities, we used the entire NOVAction23 data of the corresponding modality for training for 30 epochs, as there was no significant improvement in accuracy thereafter. Both models were tested on the entire NTU 20 test data and validated on the entire YouTube Action set. The results are given in **Table 5**.

Table 5: Results of the second ablation study, in which we compare the action recognition performance of the RGB-only and Skeleton-only modalities when trained with NOVAction23. The best accuracies achieved are given in bold. The mean column shows the average of the top-1 scores of the NTU 20 test and the YouTube Action validation.

	NTU 20 Test		YouTube Action Validation		Mean
Modality	**top-1**	**top-5**	**top-1**	**top-5**	**top-1**
RGB-only	0.17	0.44	0.35	0.67	0.26
Skeleton-only	**0.21**	**0.60**	**0.74**	**0.92**	**0.48**

The second part of the ablation study showed that the model trained in the skeleton-only modality outperformed the model trained in the RGB-only modality. These results suggest that skeleton-based classifiers should be considered for real-world action recognition tasks. Additionally, it was observed that a model trained exclusively with NOVAction23 performed exceptionally well on YouTube Action.

In the third part of our ablation study, the goal was to assess the RGB-only generalization performance of the model, which was pre-trained on Kinetics 400, by fine-tuning it with NOVAction23. We compare this performance to fine-tuning the same pre-trained model using another synthetically generated human action recognition dataset, SURREACT. For this purpose, we kept the same RGB-only training settings used in the previous test, but removed the five action classes (*Side Kick*, *Squat*, *Yawn*, *Thumb Up*, and *Stretch*) that are not shared by the two sets. Hence, the tests were carried out using the 15 action classes that coexist in SURREACT, NTU 120 Test, and NOVAction23. For the same reason, we also removed two classes from YouTube Action, and validated with the remaining three classes (*Stand Up*, *Jump*, and *Fall Down*) that are also available on the aforementioned data partitions. Accordingly, we ended up with 9,346 sequences from SURREACT and 20,317 sequences from NOVAction23. To ensure balanced training with respect to the amount of data used, we trained NOVAction23 for 30 epochs and SURREACT for 65 epochs. The reason for using these epoch numbers is that NOVAction23 contains approximately 2.17 times more video sequences than SURREACT in the specified classes. The results are given in **Table 6**.

Table 6: Results of the third ablation study, where the Kinetics 400 pre-trained model is fine-tuned with the NOVAction23 and SURREACT synthetic datasets separately. The best accuracies achieved are given in bold. The mean column shows the average of the top-1 scores of the NTU 20 test and the YouTube Action validation.

	NTU 20 Test		YouTube Action Validation		Mean
Trained On	**top-1**	**top-5**	**top-1**	**top-5**	**top-1**
NOVAction23	0.17	0.49	**0.34**	**0.62**	**0.26**
SURREACT	**0.18**	**0.51**	0.09	0.30	0.14

The last part revealed that the model trained with NOVAction23 has a higher average accuracy. Even though SURREACT employs identical pose sequences as NTU 20, it only slightly outperforms NOVAction23 on the NTU 20 test data. In contrast, NOVAction23 significantly outperforms SURREACT on the YouTube Action validation. Overall, these results suggest that fine-tuning with NOVAction23 is more effective at generalization than fine-tuning with SURREACT, making NOVAction23 a better candidate for augmenting real-world training data in human action recognition tasks.

6. Conclusion

In this paper, we first introduced the NOVAction engine, a novel tool to automatically generate massively diverse and photorealistically synthetic human action datasets. NOVAction is capable of creating arbitrarily large amounts of unique action sequences, each performed by a distinct synthetic human generated at runtime and captured from diverse camera views.

Next, we presented the NOVAction23 dataset generated using the NOVAction engine. NOVAction23 includes 25,415 video sequences featuring 1,105 synthetic humans performing 20 distinct action classes across five different 3D scenes from 125 base viewpoints. Along with the video sequences, automatically generated precise pose and label information is also included. The NOVAction23 dataset offers a level of diversity that exceeds current state-of-the-art synthetic human action recognition datasets. We make this dataset publicly available at the paper website. In addition, we provide a video demonstrating sample action sequences from the NOVAction23 dataset in comparison to those from the NTU 120 and Youtube Action datasets as supplemental material.

To evaluate the efficacy of the NOVAction23 data in improving recognition performance, we conducted a series of benchmark tests using three state-of-the-art action recognizers (TS [15], TPN [16] and SlowOnly [17]), by training them on both the NTU 120 and NOVAction23 datasets and subsequently validating their performance on videos collected from YouTube. Our results indicated that training on the synthetic NOVAction23 data in addition to the real data leads to improved action recognition performance on real-world data, for which SlowOnly outperforms the other recognizers.

We also conducted a three-part ablation study. In the first part, where we evaluated the effects of lighting conditions using RGB-only training, the results indicated that the outdoor videos from NOVAction23 may be more similar to real-world videos in terms of lighting and overall realism, while it is recommended that both indoor and outdoor videos from NOVAction23 be used in conjunction with other real-world datasets for best action recognition performance. The second part, where we trained with synthetic data only, showed that the skeleton-only modality outperformed the RGB-only modality. For the last part, we compared NOVAction23 with SURREACT in RGB-only training performance, as both are synthetic datasets that aim to address the problem of arbitrary-view human action recognition. The model trained with NOVAction23 had better generalization compared to SURREACT, illustrating the benefits of using more photorealistic data to train human action recognition models and showing that NOVAction23 data is better suited to address this problem.

Our experiments were limited to evaluating the image (RGB-only) and pose (skeleton-only) modalities of the NOVAction23 dataset separately. In future work, it would provide valuable insights to study the effects of using both modalities with training architectures that employ them together. Another limitation was the focus of the present evaluation on a set of 20 action classes that are relatively more common than the other classes found in real action datasets. Since the procedural animation system of the NOVAction engine allows the use of arbitrary motion sequences, future work should benefit from the evaluation of an even more extensive set of data created by using a larger number of action classes.

Although NOVAction23 provides varied action sequences using 1,105 synthetic human actors with unique combinations of attributes including gender, height, weight, skin tone, and clothing, yielding an unprecedented level of diversity in an action recognition dataset, there is potential to further expand this diversity. The KIST SynADL dataset [12] provided synthetically generated data for the recognition of actions by elderly subjects, yet, to our knowledge, no dataset has explicitly addressed the recognition of actions by infants or toddlers. Likewise, neither androgynous body types nor non-binary appearances have been specifically addressed. Future efforts to incorporate additional synthetic data addressing such inadequate representations would be beneficial to enhance recognition performance while still accounting for privacy concerns. In addition, to increase the level of photorealism, an interesting research direction would be incorporating ray tracing -based post-processing approaches or utilizing generative models.

Declarations

Data Availability. The datasets used in this work are provided on the paper's GitHub pagehttps://github.com/celikcan-cglab/NOVAction23.

Code availability. The code used to process the data is available at the paper website.

Conflicts of interest/Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Ji, S, Xu, W, Yang, M, Yu, K. 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence 2012;35(1):221–231.

[2] Sun, C, Shrivastava, A, Singh, S, Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 843–852.

[3] Carreira, J, Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, p. 6299–6308.

[4] Smaira, L, Carreira, J, Noland, E, Clancy, E, Wu, A, Zisserman, A. A short note on the kinetics-700-2020 human action dataset. arXiv preprint arXiv:201010864 2020;.

[5] Choi, J, Gao, C, Messou, JC, Huang, JB. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. Advances in Neural Information Processing Systems 2019;32.

[6] Shahroudy, A, Liu, J, Ng, TT, Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 1010–1019.

[7] Liu, J, Shahroudy, A, Perez, M, Wang, G, Duan, LY, Kot, AC. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence 2019;42(10):2684–2701.

[8] Kerim, A, Aslan, C, Celikcan, U, Erdem, E, Erdem, A. Nova: Rendering virtual worlds with humans for computer vision tasks. In: Computer Graphics Forum; vol. 40. Wiley Online Library; 2021, p. 258–272.

[9] Ludl, D, Gulde, T, Curio, C. Enhancing data-driven algorithms for human pose estimation and action recognition through simulation. IEEE Transactions on Intelligent Transportation Systems 2020;21(9):3990–3999. doi:10.1109/TITS.2020.2988504.

[10] Roitberg, A, Schneider, D, Djamal, A, Seibold, C, Reiß, S, Stiefelhagen, R. 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, p. 8563–8569.

[11] Varol, G, Laptev, I, Schmid, C, Zisserman, A. Synthetic humans for action recognition from unseen viewpoints. International Journal of Computer Vision 2021;129(7):2264–2287.

[12] Hwang, H, Jang, C, Park, G, Cho, J, Kim, IJ. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. IEEE Access 2021;:1–1doi:10.1109/ACCESS.2021.3051842.

[13] da Costa, VGT, Zara, G, Rota, P, Oliveira-Santos, T, Sebe, N, Murino, V, et al. Dual-head contrastive domain adaptation for video action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022, p. 1181–1190.

[14] Gedamu, K, Ji, Y, Yang, Y, Gao, L, Shen, HT. Arbitrary-view human action recognition via novel-view action generation. Pattern Recognition 2021;118:108043. doi:https://doi.org/10.1016/j.patcog.2021.108043.

[15] Bertasius, G, Wang, H, Torresani, L. Is space-time attention all you need for video understanding? In: Meila, M, Zhang, T, editors. Proceedings of the 38th International Conference on Machine Learning; vol. 139 of *Proceedings of Machine Learning Research*. PMLR; 2021, p. 813–824. URL: https://proceedings.mlr.press/v139/bertasius21a.html.

[16] Yang, C, Xu, Y, Shi, J, Dai, B, Zhou, B. Temporal pyramid network for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 591–600.

[17] Feichtenhofer, C, Fan, H, Malik, J, He, K. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 6202–6211.

[18] Soomro, K, Zamir, A, Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. ArXiv 2012;abs/1212.0402.

[19] Kuehne, H, Jhuang, H, Garrote, E, Poggio, T, Serre, T. Hmdb: A large video database for human motion recognition. 2011 International Conference on Computer Vision 2011;:2556–2563.

[20] Fabian Caba Heilbron Victor Escorcia, BG, Niebles, JC. Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, p. 961–970.

[21] Carreira, J, Noland, E, Banki-Horvath, A, Hillier, C, Zisserman, A. A short note about kinetics-600. 2018. URL: https://arxiv.org/abs/1808.01340. doi:10.48550/ARXIV.1808.01340.

[22] Gu, C, Sun, C, Ross, DA, Vondrick, C, Pantofaru, C, Li, Y, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, p. 6047–6056.

[23] Li, A, Thotakuri, M, Ross, DA, Carreira, J, Vostrikov, A, Zisserman, A. The ava-kinetics localized human actions video dataset. 2020. URL: https://arxiv.org/abs/2005.00214. doi:10.48550/ARXIV.2005.00214.

[24] Chen, C, Jafari, R, Kehtarnavaz, N. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP). 2015, p. 168–172. doi:10.1109/ICIP.2015.7350781.

[25] Li, Y, Li, Y, Vasconcelos, N. Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018, p. 513–528.

[26] Xu, N, Liu, A, Nie, W, Wong, Y, Li, F, Su, Y. Multi-modal & multi-view & interactive benchmark dataset for human action recognition. In: Proceedings of the 23rd ACM International Conference on Multimedia. MM '15; New York, NY, USA: Association for Computing Machinery. ISBN 9781450334594; 2015, p. 1195–1198. URL: https://doi.org/10.1145/2733373.2806315. doi:10.1145/2733373.2806315.

[27] Hu, JF, Zheng, WS, Lai, J, Zhang, J. Jointly learning heterogeneous features for rgb-d activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 5344–5352.

[28] Rahmani, H, Mahmood, A, Huynh, D, Mian, A. Histogram of oriented principal components for cross-view action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2016;38(12):2430–2443. doi:10.1109/TPAMI.2016.2533389.

[29] Shao, D, Zhao, Y, Dai, B, Lin, D. Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, p. 2616–2625.

[30] Shotton, J, Girshick, R, Fitzgibbon, A, Sharp, T, Cook, M, Finocchio, M, et al. Efficient human pose estimation from single depth images. Trans PAMI 2012;.

[31] Duan, H, Zhao, Y, Chen, K, Lin, D, Dai, B. Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, p. 2969–2978.

[32] Suarez, J, Murphy, RR. Using the kinect for search and rescue robotics. In: 2012 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). 2012, p. 1–2. doi:10.1109/SSRR.2012.6523918.

[33] Gaidon, A, Wang, Q, Cabon, Y, Vig, E. Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 4340–4349.

[34] Cabon, Y, Murray, N, Humenberger, M. Virtual kitti 2. arXiv preprint arXiv:200110773 2020;.

[35] Dosovitskiy, A, Ros, G, Codevilla, F, Lopez, A, Koltun, V. Carla: An open urban driving simulator. In: Conference on robot learning. PMLR; 2017, p. 1–16.

[36] Ruiz, N, Bargal, S, Xie, C, Saenko, K, Sclaroff, S. Finding differences between transformers and convnets using counterfactual simulation testing. In: Koyejo, S, Mohamed, S, Agarwal, A, Belgrave, D, Cho, K, Oh, A, editors. Advances in Neural Information Processing Systems; vol. 35. Curran Associates, Inc.; 2022, p. 14403–14418. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/5ce3a49415f78db65a714b4f05c62f4e-Paper-Conference.pdf.

[37] Shah, S, Dey, D, Lovett, C, Kapoor, A. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In: Field and service robotics. Springer; 2018, p. 621–635.

[38] Uner, OC, Aslan, C, Ercan, B, Ates, T, Celikcan, U, Erdem, A, et al. Synthetic18k: Learning better representations for person re-id and attribute recognition from 1.4 million synthetic images. Signal Processing: Image Communication 2021;97:116335.

[39] Xiang, S, Fu, Y, You, G, Liu, T. Taking a closer look at synthesis: Fine-grained attribute analysis for person re-identification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2021, p. 3765–3769.

[40] Li, S, Chen, H, Yu, S, He, Z, Zhu, F, Zhao, R, et al. Cocas+: Large-scale clothes-changing person re-identification with clothes templates. IEEE Transactions on Circuits and Systems for Video Technology 2023;33(4):1839–1853. doi:10.1109/TCSVT.2022.3216769.

[41] Ariz, M, Bengoechea, JJ, Villanueva, A, Cabeza, R. A novel 2d/3d database with automatic face annotation for head tracking and pose estimation. Comput Vis Image Underst 2016;148(C):201–210.

[42] Kerim, A, Celikcan, U, Erdem, E, Erdem, A. Using synthetic data for person tracking under adverse weather conditions. Image and Vision Computing 2021;111:104187.

[43] Wang, X, Girshick, R, Gupta, A, He, K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 7794–7803.

[44] Tran, D, Wang, H, Torresani, L, Feiszli, M. Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, p. 5552–5561.

[45] Ghadiyaram, D, Tran, D, Mahajan, D. Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019, p. 12046–12055.

[46] Feichtenhofer, C. X3d: Expanding architectures for efficient video recognition. 2020. arXiv:2004.04730.

[47] Ren, S, He, K, Girshick, R, Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 2015;28.

[48] Sun, K, Xiao, B, Liu, D, Wang, J. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 5693–5703.

[49] Gao, Y, Lin, P, Liu, R. Comparison and analysis between different versions of fxaa. In: 2022 14th International Conference on Computer Research and Development (ICCRD). 2022, p. 299–310. doi:10.1109/ICCRD54409.2022.9730249.

[50] Korein, J, Badler, N. Temporal anti-aliasing in computer generated animation. SIGGRAPH Comput Graph 1983;17(3):377–388. URL: https://doi.org/10.1145/964967.801168. doi:10.1145/964967.801168.

[51] Adobe Mixamo. 2008. URL: https://www.mixamo.com; accessed: 2022-11-07.

[52] Das, S, Dai, R, Koperski, M, Minciullo, L, Garattoni, L, Bremond, F, et al. Toyota smarthome: Real-world activities of daily living. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019, p. 833–842. doi:10.1109/ICCV.2019.00092.

[53] Derry, SJ, Pea, RD, Barron, B, Engle, RA, Erickson, F, Goldman, R, et al. Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. Journal of the Learning Sciences 2010;19(1):3–53. URL: https://doi.org/10.1080/10508400903452884. doi:10.1080/10508400903452884. arXiv:https://doi.org/10.1080/10508400903452884.

[54] Contributors, M. Openmmlab's next generation video understanding toolbox and benchmark. 2020.

[55] Paszke, A, Gross, S, Massa, F, Lerer, A, Bradbury, J, Chanan, G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 2019;32.

[56] Shorten, C, Khoshgoftaar, TM. A survey on image data augmentation for deep learning. Journal of big data 2019;6(1):1–48.

[57] Ning, G, Pei, J, Huang, H. Lighttrack: A generic framework for online top-down human pose tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, p. 1034–1035.

[58] Lin, TY, Maire, M, Belongie, S, Bourdev, L, Girshick, R, Hays, J, et al. Microsoft coco: Common objects in context. 2015. arXiv:1405.0312.

[59] He, K, Zhang, X, Ren, S, Sun, J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, p. 770–778. doi:10.1109/CVPR.2016.90.

[60] Loshchilov, I, Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:160803983 2016;.

[61] Ilya, L, Frank, H, et al. Decoupled weight decay regularization. Proceedings of ICLR 2019;.