# Enhancing human action recognition with GAN-based data augmentation

Prasanna Reddy Pulakurthi[a], Celso M. de Melo[b], Raghuveer Rao[b], and Majid Rabbani[a]

[a]Rochester Institute of Technology, Department of Electrical and Computer Engineering
[b]DEVCOM U.S. Army Research Laboratory

.

## ABSTRACT

Deep Neural Networks (DNNs) have emerged as powerful tools for human action recognition, yet their reliance on vast amounts of high-quality labeled data poses significant challenges. The traditional approach of collecting and labeling large volumes of real-world data is not only costly but also raises ethical concerns. A promising alternative is to generate synthetic data. However, the existing synthetic data generation pipelines require complex simulation environments. Our work presents a novel solution by employing Generative Adversarial Networks (GANs) to generate synthetic yet realistic training data from a small existing real-world dataset, thereby bypassing the need for elaborate simulation environments. Central to our approach is a training pipeline that extracts motion from each training video and augments it across varied subject appearances within the training set. This method increases the diversity in both motion and subject representation, thus significantly enhancing the model's performance in accurately recognizing human gestures. The model's performance is rigorously evaluated in diverse scenarios, including ground and aerial views, to demonstrate the method's versatility and effectiveness. The findings of our study highlight the efficiency of GAN-based data augmentation, utilizing a minimal real dataset to create synthetic data without relying on complex simulators. Moreover, useful insights are provided by analyzing the critical factors influencing gesture recognition performance, such as the diversity in gesture motion and the diversity in subject appearance. The code is available at *https://github.com/PrasannaPulakurthi/EHAR-GAN*.

**Keywords:** Human Action Recognition, Generative Adversarial Networks, Deep Neural Networks, Synthetic Data, Data Augmentation

## 1. INTRODUCTION

Human Action Recognition (HAR) is an important task in the field of computer vision, with numerous applications,[1, 2] including surveillance for safety and security,[3] patient monitoring in healthcare,[4] sports analysis for performance enhancement,[5] autonomous vehicles for pedestrian detection,[6] and human-computer interaction.[7] At its core, HAR aims to automatically recognize human actions from a series of observations (image frames), typically a video. The advent of Deep Neural Networks (DNNs) has significantly advanced the state-of-the-art in this domain. However, the efficiency of these models is heavily dependent on the availability of large, high-quality labeled datasets. The conventional method of manually collecting and analyzing real-world data is not only time-consuming and costly but also raises potential ethical and privacy concerns.

In an effort to address these challenges, researchers have explored the generation of synthetic data as an alternative to real-world data collection. One of the primary benefits of synthetic data is its ability to be rapidly scaled and easily labeled, which is particularly useful for data augmentation purposes. However, traditional synthetic data generation approaches require either advanced computer graphics or complex simulation environments with physics engines, which are not only challenging but also require significant resources. These limitations underscore the need for a solution to circumvent the complexities associated with conventional synthetic data generation systems.

[*]©2024 SPIE. This paper was published in the Proceedings of SPIE, Volume 13035, Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications II. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, [https://doi.org/10.1117/12.3021572]

Our work introduces a novel approach to generate synthetic yet realistic training data for human action recognition using Generative Adversarial Networks (GANs) from a small existing real-world dataset. By leveraging minimal real-world data, we bypass the need for complex simulation environments and significantly reduce the barriers to entry for synthetic data generation in HAR. Our method focuses on extracting motion from training video and augmenting it across varied subject appearances, thus enhancing both the diversity and the volume of the available training data. This paper presents a comprehensive evaluation of our model's performance across diverse scenarios and demonstrates the effectiveness of GAN-based data augmentation in overcoming the challenges associated with traditional data collection and synthetic data generation methods.

## 2. RELATED WORK

The landscape of Human Action Recognition (HAR) has been profoundly shaped by the integration of deep learning techniques, with Deep Neural Networks (DNNs) at the forefront, to achieve state-of-the-art performance. The need for extensive, labeled datasets has been a persistent hurdle, sparking interest in synthetic data generation as a viable solution. Synthetic data augmentation is crucial in enhancing DNN performance for HAR by expanding the training dataset with modified versions of the existing data to include more variation.

### 2.1 Synthetic Human Action Recognition Datasets

Several synthetic video datasets have been created in recent years to train HAR deep learning models. The Game Action Dataset (GAD)[8] dataset explores the gaming domain for data generation. It comprises recordings from gaming sessions in GTA5 and FIFA performed by human players. This dataset derived from the gaming environment includes synchronized ground and aerial views. A player's command controls the character's motions in the game. The PHAV[9] dataset uses a modern and accessible game engine (Unity®Pro) to synthesize a labeled dataset. This dataset is produced using an interpretable parametric generative model of human action videos, which depends on computer graphics techniques (procedural generation) of modern game engines.

SURREACT (Varol et al., 2021[10]) utilizes 3D human motion estimation models, including HMMR (Kanazawa et al., 2019[11]) and VIBE (Kocabas et al., 2019[12]), for reconstructing human body meshes and motions from single-view RGB videos. The dataset employs the SMPL (Loper et al., 2019[13]) statistical model for the body mesh, augmented with randomized cloth textures, lighting conditions, and body shapes to enhance diversity.

The SynADL[14] dataset, focusing on the Activities of Daily Living (ADL) of elders, incorporates 3D human characters created through Kinect sensor scans of 15 participants, whose Motion Capture (MoCap) data animate the characters, aiming at detecting ADL with precision. A significant amalgamation by Kim et al. (2022) resulted in the SynAPT[15] dataset, integrating PHAV, SURREACT, and SynADL for pre-training models adaptable to novel downstream tasks encompassing entirely new categories. BABEL[16] introduces a comprehensive dataset featuring detailed language labels for MoCap sequences. It uniquely categorizes actions into sequence and frame labels, aligning each action with its exact duration in the MoCap data across over 250 distinct action categories.

The RoCoG[17] and its successor RoCoG-v2[18] datasets, designed for human-robot interaction, are based on seven gestures from the US Army Field Manual.[19] While RoCoG offers a static ground view, RoCoG-v2 extends this with static ground and aerial views. Moreover, RoCoG contains only manually designed motions, whereas RoCoG-v2 introduces MoCap data for some motions. These methods use virtual environments to generate synthetic data and examine how variations in the environment and character models affect recognition accuracy. Their findings underscored the benefits of diversity in synthetic datasets and the challenges in achieving sufficient realism and variability.

Panev et al. (2024)[20] explored the impact of different rendering methods and motion quality on the effectiveness of synthetic data for action recognition tasks. Their study highlighted the potential of high-quality renderings to improve model performance and noted the complexity and computational demands of creating such environments. It introduces four unique synthetic datasets generated through a synthesis of MoCap and video-based motions with rendering techniques like Computer Graphics (CG) and neural rendering.

Prior work in this area used complex simulation environments to create synthetic data. Unlike previous methods that require elaborate setup and customization, our approach generates diverse and realistic synthetic data from a small set of real videos, effectively circumventing the need for complex simulation environments.

## 2.2 Real Human Action Recognition Datasets

HMDB51[21] marked a significant leap forward with its collection of 51 diverse action categories, while UCF-101[22] further enriched the landscape with 101 activity classes across five categories. NTU-RGB+D120,[23] which introduced a large multi-view dataset with 120 action classes and included depth map sequences, 3D skeletal data, and infrared videos for each sample. The Charades[24] dataset, consisting of 157 action classes, is composed of hundreds of people recording videos in their own homes, acting out casual everyday activities. The ActivityNet[25] dataset contains 200 different types of activities of videos collected from YouTube.

The introduction of the Kinetics series (Kinetics-400,[26] Kinetics-600,[27] Kinetics-700,[28] and Kinetics-700-2020[29]) was pivotal, offering a vast array of categories and nearly a 1,000 videos per category, significantly boosting the field's development. Further diversification in dataset perspectives was achieved with Charades-Ego[30] and HOMAGE,[31] presenting daily activities from both first and third-person views.

The Okutama-Action,[32] UAV-Gesture,[33] PRAI-1581,[34] and AVI[35] datasets offer focused insights into human action recognition, Unmanned Aerial Vehicle (UAV) control gestures, person ReID, and violent action recognition, respectively. The UAV-Human dataset,[36] notable for its aerial capture of human activities through UAVs in various settings, became the largest real HAR dataset featuring aerial views. Additionally, the YouTube-Aerial Dataset (YAD)[37] contributed aerial videos from YouTube, showcasing dynamic camera movements and varying altitudes. RoCoG[17] and its successor, RoCoG-v2,[18] enriched the dataset landscape by including real data for the same categories as their synthetic counterparts, with RoCoG-v2 adding aerial perspectives to the mix.

## 2.3 Generative Adversarial Networks (GANs) in Data Generation

A rapidly growing area of research is the application of Generative Adversarial Networks (Goodfellow et al., 2014[38]) for data argumentation in HAR. Unlike traditional data augmentation techniques, GANs can generate new data instances that retain the underlying structure of the action while varying in appearance and context. This capability is particularly advantageous for HAR, where diversity in motion and diversity in subject appearance are crucial for improving model performance, robustness, and generalization.

GANs have emerged as a powerful class of machine learning frameworks that are capable of generating high-quality, realistic synthetic data. GANs consist of two neural networks, a generator and a discriminator, that are trained simultaneously through a competitive training process. The generator aims to produce data samples that are indistinguishable from real data, while the discriminator aims to distinguish between real and generated data samples accurately. This paradigm has been applied in various domains, including image generation,[39] style transfer,[40] domain adaptation,[41] and more recently, in augmenting datasets for machine learning tasks.[42]

Models such as Liquid Warping GAN (Liu et al., 2019[43]) and Everybody Dance Now (Chan et al., 2019[44]) have revolutionized the generation of action videos through the transfer of body poses from one individual's performance video to another person, either from a different video or a single image. These models facilitate the creation of diverse action sequences by adapting the appearance of the subject based on the selected visual source image, thereby enabling a wide range of applications in video synthesis and modification.

Unlike the traditional computer graphics workflow, which is both time-consuming and manually intensive for scene design, GANs utilize a model that has been pre-trained on human motion datasets. This significantly reduces the time required to produce synthetic video sequences for HAR. This method not only addresses the limitations associated with the collection of large real-world datasets and the use of complex simulation environments but also paves the way for more generalized and robust HAR models.

# 3. DATASET

Our experiments are performed on real video data from the RoCoG-V2[18] dataset, which contains both ground and aerial views. The dataset consists of 11 adult subjects (10 males and one female) and includes diverse age ranges and clothing types. The subjects are divided into train, validation, and test split according to Table 1. The authors carefully selected a set of four subjects to form the test set, which included the only female participant and reflected the diversity in the background location, body build, and skin color.

Table 1. Subjects divided into training, validation, and test sets.

| | | **Train** | **Validation** | **Test** |
|---|---|---|---|---|
| Ground | Subject No. | 1, 3, 5, 7, 8 | 2, 10 | 0, 4, 6, 9 |
| | No. Training Samples | 25, 20, 62, 34, 21 | 21, 21 | 21, 34, 24, 21 |
| Air | Subject No. | 5, 7, 8 | 10 | 0, 4, 6, 9 |
| | No. Training Samples | 22, 23, 21 | 21 | 28, 22, 20, 21 |

The ground view dataset, shown in Figure 1, comprises the training set consisting of 162 videos of five subjects, the validation set consists of 42 videos of two subjects, and the test set consists of 100 videos of four subjects.

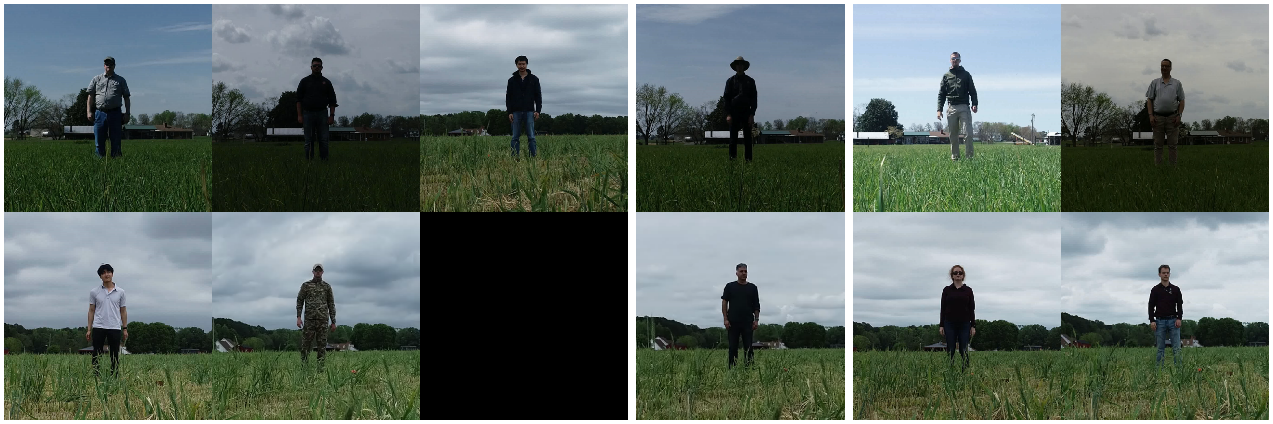

Figure 1. Ground view dataset: a) training subjects (left), b) validation subjects (middle), and c) test subjects (right).

The aerial view dataset, shown in Figure 2, comprises the training set consisting of 66 videos of three subjects, the validation set consists of 21 videos of one subject, and the test set consists of 91 videos of four subjects. All the subjects perform multiple instances of the seven gesture classes in both ground and air perspectives.

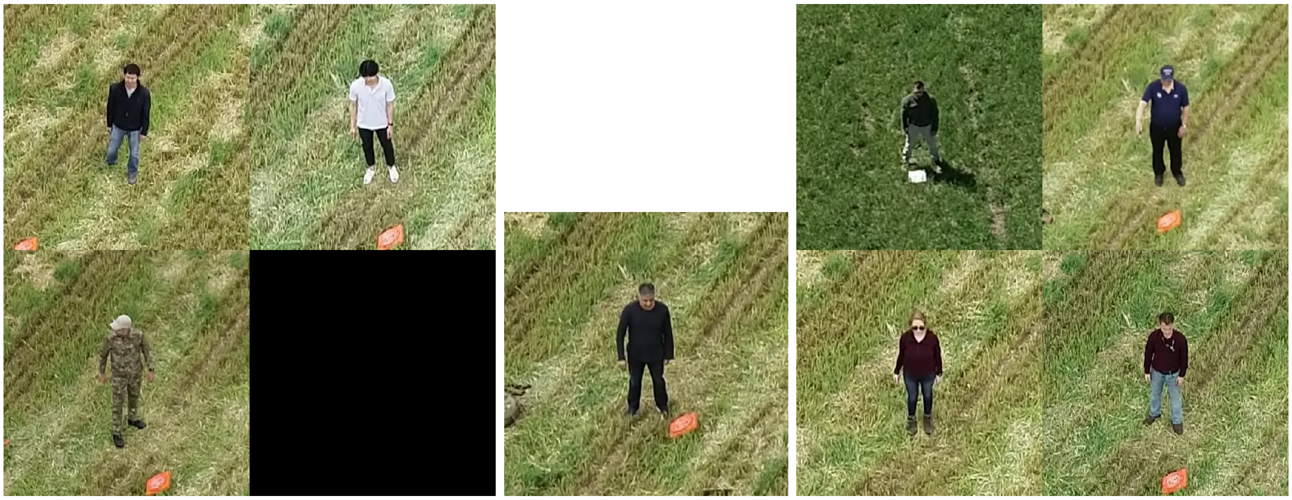

Figure 2. Aerial view dataset: a) training subjects (left), b) validation subject (middle), and c) test subjects (right).

# 4. EXPERIMENTS

The OpenMMLab's MMAction2[45] framework is used for training and evaluating human action recognition models. More specifically, the Inflalted 3D ConvNet (I3D)[46] (ResNet50-based) model, which is a popular deep video activity recognition model, is employed. The input to the model is a 32-frame sequence of 256x256 pixels, where the video is resized as needed. Data augmentation techniques include random horizontal flip and scale augmentations for all the approaches. All the models are trained for 20 epochs, starting with the Kinetics-400[26] pre-trained weights and use top-1 classification accuracy to compare their performance. Training employs a Stochastic Gradient Descent (SGD) optimizer with specific parameters: learning rate = 0.002, momentum = 0.9, weight_decay = 0.0001.

For generating the dataset, the Liquid Warping GAN (LWG)[43] with Attention is used. The LWG serves as a unified framework for synthesizing human images, mimicking human motion, transferring appearances, and creating new viewpoints. It also features a 3D body mesh recovery module that employs SMPL[13] to separate and analyze the pose and body shape aspects effectively. Unlike the complex and time-consuming CG pipelines, the model in LWG has been pre-trained on priors to generate human motion videos, thus reducing the processing time for generating synthetic video sequences for HAR. Here LWG is used to extract motion from a subject in one video and then apply it to the visual appearances of different subjects within the training dataset.

Four experiments are devised to explore the impact of synthetic data augmentation. The first experiment establishes a baseline action recognition performance by training the model on the original real data. The next two experiments aid in evaluating the influence of variations in subject appearance and variations in motion on recognition performance. Finally, the last experiment trains the model with a complete synthetic dataset generated by translating the motion of all the subjects to the appearance of all the subjects.

## 4.1 Experiment 1: Original Training Data

The first experiment aims to establish the baseline action recognition performance on the real dataset presented in Section 3 without any synthetic data.

## 4.2 Experiment 2: Motion of one subject →Appearance of all subjects

The second experiment, "One Motion All Appearances (1MAA)," aims to analyze the influence of diverse variations in subject appearance on action recognition performance. To achieve this, the motion of one subject is extracted and transferred to the appearance of all the subjects in the training dataset. An example is shown in Figure 3, in which the motion of subject number 1 is transferred to the appearance of all the subjects. Only a single image of the target subject is required to perform the translation.

## 4.3 Experiment 3: Motion of all subjects →Appearance of one subject

The third experiment, "All Motions One Appearance (AM1A)," is designed to analyze the influence of diverse variations in motion on action recognition performance. The underlying motion of all the subjects is transferred to the appearance of a single subject. For example, the actions performed by all the subjects are translated to the appearance of subject number 5 as shown in Figure 4.

## 4.4 Experiment 4: Motion of all subjects →Appearance of all subjects

The final experiment, "All Motions All Appearances (AMAA)," shows the effectiveness of data augmentation on the HAR performance. This experiment consists of extracting the motion of all the subjects and transferring it to the appearance of all the subjects. Thus increasing the diversity in both the motion and appearance of subjects.

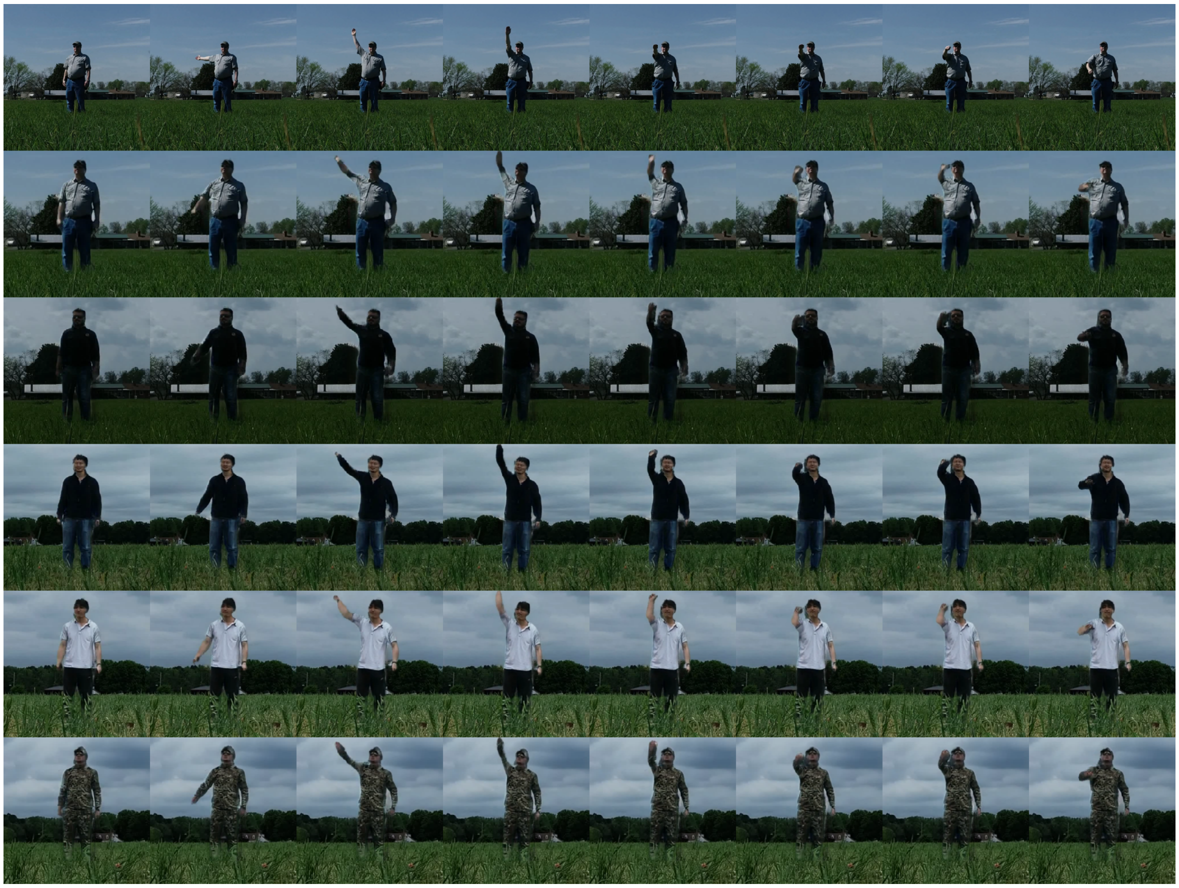

Figure 3. The gestures performed by subject number 1 are translated to the appearance of all the subjects.

Table 2. The results (top-1 accuracy %) of all four experiments.

| Experiment No. | Experiment Name | Ground % | Aerial % |
|---|---|---|---|
| 1 | Original Data | 84.00$_{\pm 2.74}$ | 65.93$_{\pm 3.56}$ |
| 2 | One Motion All Appearances (1MAA) | 87.04$_{\pm 4.09}$ | 68.79$_{\pm 5.86}$ |
| 3 | All Motions One Appearance (AM1A) | 85.64$_{\pm 3.96}$ | 67.99$_{\pm 6.45}$ |
| 4 | All Motions All Appearances (AMAA) | **88.40**$_{\pm 0.55}$ | **73.63**$_{\pm 5.04}$ |

## 5. RESULTS

The results of all four experiments are presented in Table 2.

Below are specific observations from each experiment.

**Experiment 1:** The baseline action recognition performance achieved when training on the real dataset is (84.00 ± 2.74 %) and (65.93 ± 3.56 %) for the ground and air datasets, respectively. The results show the mean and standard deviation over five runs with different random seed values.

**Experiment 2:** For each subject, the synthetic data is generated as described in Section 4.2 and is augmented to the real dataset. Each experiment uses the motion from only one subject, therefore creating diversity in subject appearance but not the motion. Similarly, this is performed for all the subjects. The results of these experiments are shown in Table 3.

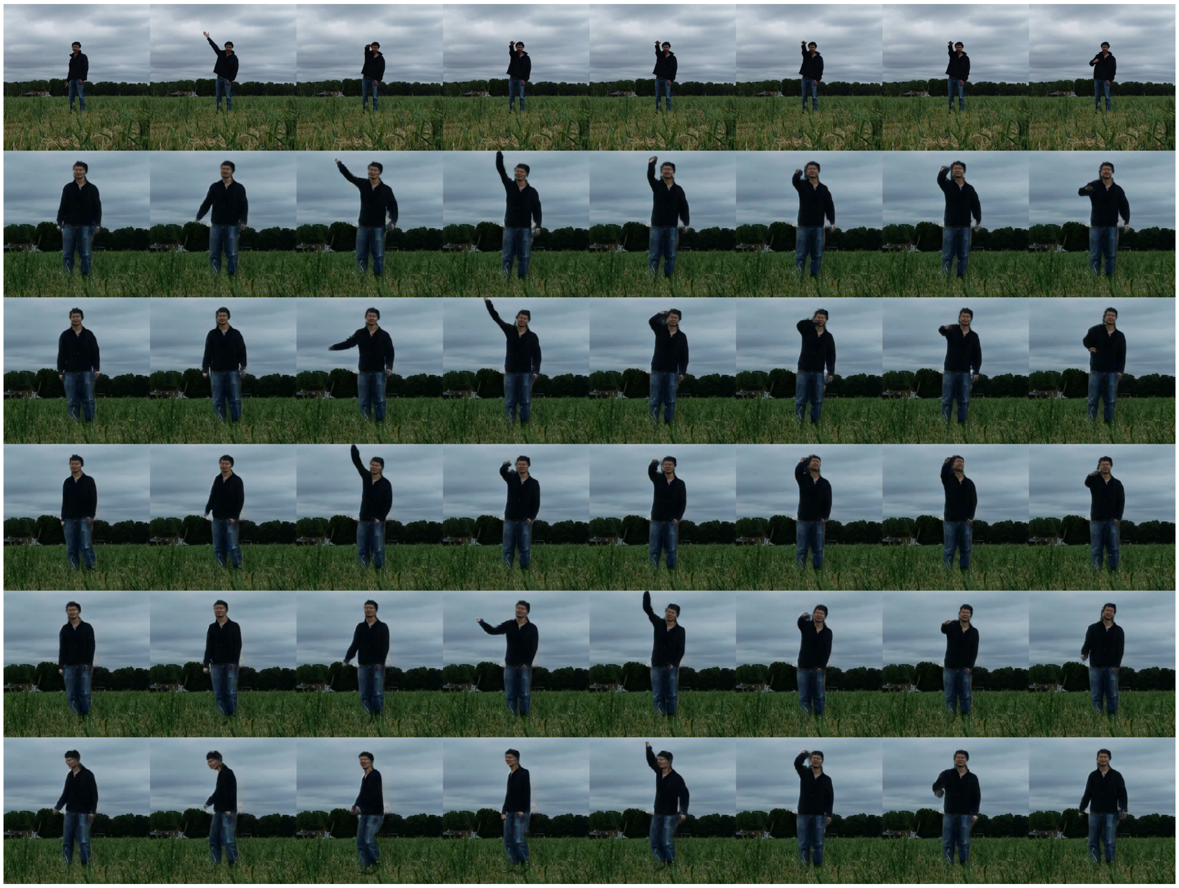

Figure 4. The gestures performed by all the subjects are translated to the appearance of subject number 5.

Table 3. The results (top-1 accuracy %) of Experiment 2.

| Subject No. | Ground % | Aerial % |
|---|---|---|
| 1 | $89.4_{\pm 2.41}$ | - |
| 3 | $85.2_{\pm 5.26}$ | - |
| 5 | $87.6_{\pm 3.97}$ | $70.11_{\pm 5.29}$ |
| 7 | $88.2_{\pm 2.05}$ | $71.87_{\pm 6.80}$ |
| 8 | $84.8_{\pm 5.22}$ | $64.40_{\pm 2.76}$ |
| Average | $87.04_{\pm 4.09}$ | $68.79_{\pm 5.86}$ |

The average performance over all the subjects is (87.04 ± 4.09 %) and (68.79 ± 5.86 %) for the ground and aerial datasets, respectively. The results in Experiment 2 are significantly better than in Experiment 1 (baseline), implying that increasing the variations in subject appearance alone can improve the model's action recognition performance.

**Experiment 3:** For each subject, the synthetic data generated as described in Section 4.3 is augmented to the real dataset. Each experiment uses the appearance of only one subject, therefore creating diversity in subject motion but not the appearance. This is performed in a similar fashion for all the subjects. The results of these experiments are summarized in Table 4.

Table 4. The results (top-1 accuracy %) of Experiment 3.

| Subject No. | Ground % | Aerial % |
|---|---|---|
| 1 | $84.0_{\pm 4.64}$ | - |
| 3 | $84.4_{\pm 4.51}$ | - |
| 5 | $86.2_{\pm 2.95}$ | $68.13_{\pm 2.69}$ |
| 7 | $88.0_{\pm 4.64}$ | $66.37_{\pm 7.36}$ |
| 8 | $85.6_{\pm 3.05}$ | $69.45_{\pm 8.84}$ |
| Average | $85.64_{\pm 3.96}$ | $67.99_{\pm 6.45}$ |

The average performance over all the subjects for Experiment 3 is (85.64 ± 3.96 %) and (67.99 ± 6.45 %) for the ground and aerial datasets, respectively. The average performance of high diversity in motion (Experiment 3) is slightly better than the baseline (Experiment 1). However, high diversity in appearance (Experiment 2) is better than that of high diversity in motion (Experiment 3). Therefore, it can be inferred that augmenting the real data with higher diversity in subject appearance performs better than higher diversity in subject motion.

Table 5 shows the number of training samples used in each experiment. In Experiment 2, named "One Motion All Appearances (1MAA)," each subject had a different amount of synthetic data generated. This varied amount is because each subject has a different number of training samples to begin with. On the other hand, in Experiment 3, "All Motions One Appearance (AM1A)," since we use motion from all the videos, the same amount of synthetic data is generated for all the subjects.

Crucially, the dataset augmentation with a synthetic but smaller subset (Experiments 2 and 3) yields an enhancement in HAR performance. This enhancement is particularly pronounced when the augmented dataset is characterized by a greater variance in appearances rather than motions. The rationale underlying this observation is that the augmented motions, despite their variance, are not novel constructs but are instead derivations extrapolated from the original dataset. The final experiment contains the full spectrum of motions and appearance and yields the highest accuracy for both datasets, suggesting that a more varied training set that includes all motions and all appearances tends to result in better model performance.

**Experiment 4:** For this experiment, the motion extracted from each subject is transferred to the appearance of all the subjects, thus significantly increasing the diversity in motion and appearance. The generated dataset is augmented with the original dataset and used to train the model. From the results in Table 2, it is observed that the model achieves a classification accuracy of (**88.40** ± **0.55** %) and (**73.63** ± **5.04** %) for ground and aerial datasets, respectively. Therefore, GAN-based synthetic data augmentation achieves a significant performance improvement in human action recognition over the baseline.

## 6. CONCLUSION

Our research presents a novel solution to the generation of synthetic yet realistic training data from a small existing real-world dataset by employing Generative Adversarial Networks (GANs), which avoids the need for elaborate simulation environments. The proposed method increases the diversity in both motion and subject representations, thus significantly improving the model's gesture recognition accuracy. Our work further explores the critical factors influencing the action recognition performance, such as variations in motion and variations in subject appearance. Finally, our experimental results indicate that augmenting the real data with synthetic data with higher diversity in subject appearance performs better than higher diversity in subject motion.

## ACKNOWLEDGMENTS

This work was supported by DEVCOM U.S. Army Research Laboratory through Booz Allen Hamilton under contract W911QX-21-D-0001.

Table 5. Number of Training Samples by Experiment and Subject

| Experiment Name | Subject No. | No. Training Samples | |
|---|---|---|---|
| | | Ground | Aerial |
| 1MAA | 1 | 287 | - |
| | 3 | 262 | - |
| | 5 | 472 | 132 |
| | 7 | 332 | 135 |
| | 8 | 267 | 129 |
| AM1A | 1 | 324 | - |
| | 3 | 324 | - |
| | 5 | 324 | 132 |
| | 7 | 324 | 132 |
| | 8 | 324 | 132 |
| Original Data | | 162 | 66 |
| Average 1MAA | | 324 | 132 |
| Average AM1A | | 324 | 132 |
| AMAA | | 972 | 264 |

## REFERENCES

[1] Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O., "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video technology* **18**(11), 1473–1488 (2008).

[2] Shah, K., Shah, A., Lau, C. P., de Melo, C. M., and Chellappa, R., "Multi-view action recognition using contrastive learning," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 3381–3391 (2023).

[3] Sultani, W., Chen, C., and Shah, M., "Real-world anomaly detection in surveillance videos," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 6479–6488 (2018).

[4] Lotfi, A., Langensiepen, C., Mahmoud, S. M., and Akhlaghinia, M. J., "Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour," *Journal of ambient intelligence and humanized computing* **3**, 205–218 (2012).

[5] Parmar, P. and Tran Morris, B., "Learning to score olympic events," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*], (July 2017).

[6] Enzweiler, M. and Gavrila, D. M., "Monocular pedestrian detection: Survey and experiments," *IEEE transactions on pattern analysis and machine intelligence* **31**(12), 2179–2195 (2008).

[7] Rautaray, S. S. and Agrawal, A., "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review* **43**, 1–54 (2015).

[8] Sultani, W. and Shah, M., "Human action recognition in drone videos using a few aerial training examples," *Computer Vision and Image Understanding* **206**, 103186 (2021).

[9] De Souza, C. R., Gaidon, A., Cabon, Y., and López, A. M., "Procedural generation of videos to train deep action recognition networks," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 2594–2604 (2017).

[10] Varol, G., Laptev, I., Schmid, C., and Zisserman, A., "Synthetic humans for action recognition from unseen viewpoints," in [*International Journal of Computer Vision*], (2021).

[11] Kanazawa, A., Zhang, J. Y., Felsen, P., and Malik, J., "Learning 3d human dynamics from video," in [*Computer Vision and Pattern Recognition (CVPR)*], (2019).

[12] Kocabas, M., Athanasiou, N., and Black, M. J., "Vibe: Video inference for human body pose and shape estimation," in [*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2020).

[13] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J., "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* **34**, 248:1–248:16 (Oct. 2015).

[14] Hwang, H., Jang, C., Park, G., Cho, J., and Kim, I.-J., "Eldersim: A synthetic data generation platform for human action recognition in eldercare applications," *IEEE Access* **11**, 9279–9294 (2023).

[15] Kim, Y.-w., Mishra, S., Jin, S., Panda, R., Kuehne, H., Karlinsky, L., Saligrama, V., Saenko, K., Oliva, A., and Feris, R., "How transferable are video representations based on synthetic data?," in [*Advances in Neural Information Processing Systems*], Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., eds., **35**, 35710–35723, Curran Associates, Inc. (2022).

[16] Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., and Black, M. J., "BABEL: Bodies, action and behavior with english labels," in [*Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*], 722–731 (June 2021).

[17] de Melo, C. M., Rothrock, B., Gurram, P., Ulutan, O., and Manjunath, B., "Vision-based gesture recognition in human-robot teams using synthetic data," in [*2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*], 10278–10284 (2020).

[18] Reddy, A. V., Shah, K., Paul, W., Mocharla, R., Hoffman, J., Katyal, K. D., Manocha, D., de Melo, C. M., and Chellappa, R., "Synthetic-to-real domain adaptation for action recognition: A dataset and baseline performances," in [*2023 IEEE International Conference on Robotics and Automation (ICRA)*], 11374–11381 (2023).

[19] Department of the Army, U. S., [*Visual Signals*], Field manual, Headquarters, Department of the Army (1987).

[20] Panev, S., Kim, E., Namburu, S. A. S., Nikolova, D., de Melo, C., De la Torre, F., and Hodgins, J., "Exploring the impact of rendering method and motion quality on model performance when using multi-view synthetic data for action recognition," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 4592–4602 (January 2024).

[21] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T., "Hmdb: A large video database for human motion recognition," in [*2011 International Conference on Computer Vision*], 2556–2563 (2011).

[22] Soomro, K., Zamir, A. R., and Shah, M., "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402* (2012).

[23] Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G., "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in [*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2016).

[24] Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A., "Hollywood in homes: Crowdsourcing data collection for activity understanding," in [*Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 2016, Proceedings, Part I 14*], 510–526, Springer (2016).

[25] Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C., "Activitynet: A large-scale video benchmark for human activity understanding," in [*2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 961–970 (2015).

[26] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950* (2017).

[27] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., and Zisserman, A., "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340* (2018).

[28] Carreira, J., Noland, E., Hillier, C., and Zisserman, A., "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987* (2019).

[29] Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., and Zisserman, A., "A short note on the kinetics-700-2020 human action dataset," *arXiv preprint arXiv:2010.10864* (2020).

[30] Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., and Alahari, K., "Actor and observer: Joint modeling of first and third-person videos," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 7396–7404 (2018).

[31] Rai, N., Chen, H., Ji, J., Desai, R., Kozuka, K., Ishizaka, S., Adeli, E., and Niebles, J., "Home action genome: Cooperative compositional action understanding," in [*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 11179–11188, IEEE Computer Society (jun 2021).

[32] Barekatain, M., Marti, M., Shih, H.-F., Murray, S., Nakayama, K., Matsuo, Y., and Prendinger, H., "Okutama-action: An aerial view video dataset for concurrent human action detection," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*], (July 2017).

[33] Perera, A. G., Wei Law, Y., and Chahl, J., "Uav-gesture: A dataset for uav control and gesture recognition," in [*Proceedings of the European Conference on Computer Vision (ECCV) Workshops*], (September 2018).

[34] Zhang, S., Zhang, Q., Yang, Y., Wei, X., Wang, P., Jiao, B., and Zhang, Y., "Person re-identification in aerial imagery," *IEEE Transactions on Multimedia* **23**, 281–291 (2021).

[35] Singh, A., Patil, D., and Omkar, S., "Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*], 1710–17108 (2018).

[36] Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., and Li, Z., "UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 16266–16275 (June 2021).

[37] Sultani, W. and Shah, M., "Human action recognition in drone videos using a few aerial training examples," *Computer Vision and Image Understanding* **206**, 103186 (2021).

[38] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," in [*Advances in Neural Information Processing Systems*], Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., eds., **27**, Curran Associates, Inc. (2014).

[39] Pulakurthi, P. R., Mozaffari, M., Dianat, S. A., Rabbani, M., Heard, J., and Rao, R., "Enhancing gan performance through neural architecture search and tensor decomposition," in [*ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 7280–7284 (2024).

[40] Azadi, S., Fisher, M., Kim, V. G., Wang, Z., Shechtman, E., and Darrell, T., "Multi-content gan for few-shot font style transfer," in [*IEEE conference on computer vision and pattern recognition*], 7564–7573 (2018).

[41] Pulakurthi, P. R., Dianat, S. A., Rabbani, M., You, S., and Rao, R. M., "Unsupervised domain adaptation using feature aligned maximum classifier discrepancy," in [*Applications of Machine Learning 2022*], Zelinski, M. E., Taha, T. M., and Howe, J., eds., **12227**, 1222707, International Society for Optics and Photonics, SPIE (2022).

[42] Lu, Y., Chen, D., Olaniyi, E., and Huang, Y., "Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review," *Computers and Electronics in Agriculture* **200**, 107208 (2022).

[43] Liu, W., Zhixin Piao, M. J., Wenhan Luo, L. M., and Gao, S., "Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in [*The IEEE International Conference on Computer Vision (ICCV)*], (2019).

[44] Chan, C., Ginosar, S., Zhou, T., and Efros, A. A., "Everybody dance now," in [*IEEE International Conference on Computer Vision (ICCV)*], (2019).

[45] Contributors, M., "Openmmlab's next generation video understanding toolbox and benchmark." https://github.com/open-mmlab/mmaction2 (2020).

[46] Carreira, J. and Zisserman, A., "Quo vadis, action recognition? a new model and the kinetics dataset," 4724–4733 (07 2017).