Real-time human action recognition from aerial videos using autozoom and synthetic data

Ruiqi Xian[a], Bryan I. Vogel[b], Celso M. De Melo[c], Andre V. Harrison[c], and Dinesh Manocha[a]

[a]University of Maryland, College Park
[b]Booz Allen Hamilton Inc.
[c]DEVCOM Army Research Lab

ABSTRACT

In this paper, we propose a novel approach for real-time human action recognition (HAR) on resource-constrained UAVs. Our approach tackles the limited availability of labeled UAV video data (compared to ground-based datasets) by incorporating synthetic data augmentation to improve the performance of a lightweight action recognition model. This combined strategy offers a robust and efficient solution for UAV-based HAR. We evaluate our method on the RoCoG v2[1] and UAV-Human[2] datasets, showing a notable increase in top-1 accuracy across all scenarios on RoCoG: 9.1% improvement when training with synthetic data only, 6.9% with real data only, and the highest improvement of 11.8% with a combined approach. Additionally, using an X3D backbone further improves accuracy on the UAV-Human dataset by 5.5%. Our models deployed on a Qualcomm Robotics RB5 platform achieve real-time predictions at approximately 10 frames per second (fps) and demonstrate a superior trade-off between performance and inference rate on both low-power edge devices and high-end desktops.

Keywords: Unmanned Aerial Vehicles(UAVs), Real-time human action recognition, Synthetic data

1. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) equipped with high-resolution cameras offer a powerful tool for data collection in diverse environments.[3] Their ability to capture aerial video has revolutionized applications like human detectionn,[4] tracking,[5] and action recognition[6,7] in search and rescue, security, and traffic monitoring. By analyzing video sequences, UAVs can provide valuable insights into human behavior, informing decision-making processes.[8,9]

However, significant challenges exist between the potential of UAVs for human action recognition (HAR) and real-world implementation. These challenges stem from the inherent differences between UAV-captured video and traditional ground-based video:

1. **Reduced Scale of Human Subjects:** Due to high flying altitudes, humans appear smaller in UAV video frames, often occupying less than 5% of the image (e.g., UAV-Human dataset[2]). This makes it difficult for existing models to perceive human movement patterns.
2. **Varied Viewing Angles:** Unlike static ground cameras, UAVs offer oblique and overhead viewpoints. Existing models trained on ground-level activity struggle to generalize to these unseen viewpoints.
3. **Moving Camera Viewpoint:** Continuous UAV movement introduces motion blur and viewpoint changes, further complicating human behavior analysis. These factors, along with the need for image stabilization,[2,10] hinder current models' ability to interpret human behavior.

Further author information: (Send correspondence to Ruiqi Xian)
Ruiqi Xian: E-mail: rxian@umd.edu

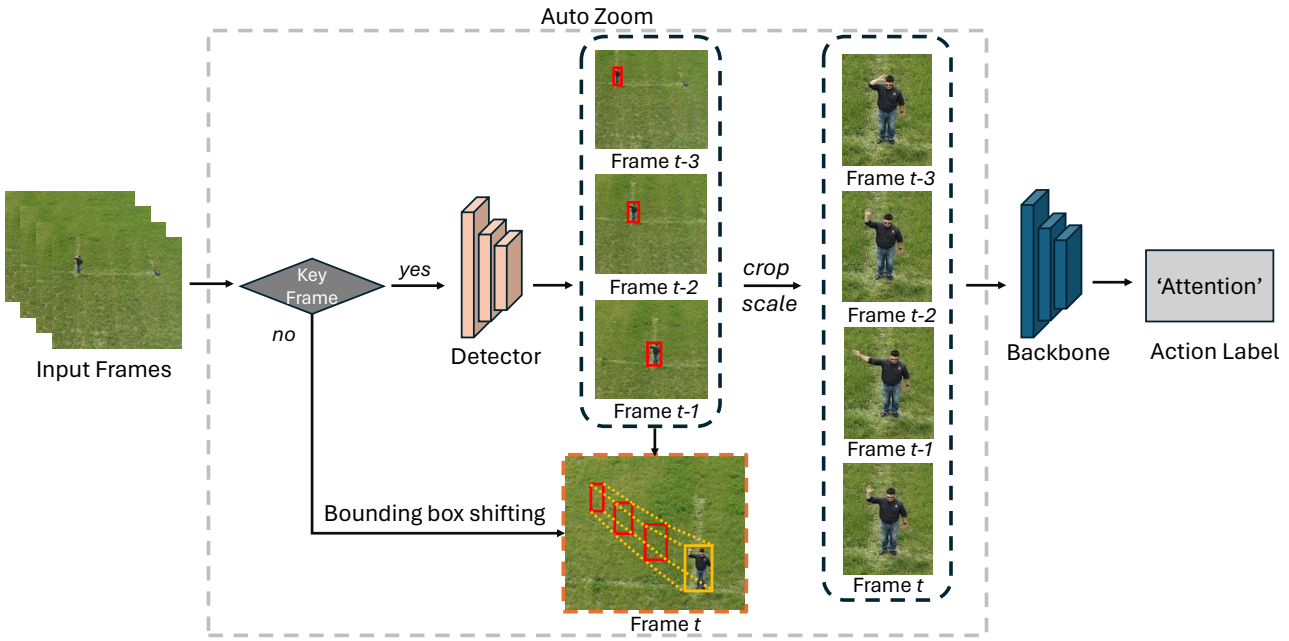

Figure 1. **Overview of our method.** Our method begins with an auto-zoom algorithm employing a lightweight detector that efficiently extracts key information from video frames. To save computational cost, the detector focuses on infrequent target presence and uses bounding box shifting for subsequent frames. Finally, all frames are cropped based on the bounding boxes and rescaled to a uniform size before being fed into the core recognition model for action prediction.

While deep learning has transformed video action recognition, directly applying these methods to UAV videos often leads to a significant drop in performance due to these challenges.[11,12] Furthermore, deploying computationally expensive models on resource-constrained UAV platforms is impractical. Existing mobile recognition methods, while designed for lower power consumption, are primarily optimized for ground camera data and underperform on aerial videos.[13–15] Therefore, developing specialized techniques for the efficient execution on UAV hardware is crucial for unlocking their full potential.

Another challenge is the difficulty and expense of acquiring and labeling high-quality UAV video data. Ground-camera datasets like Kinetics-400[16] boast hundreds of thousands of labeled videos, while the recently introduced UAV-Human dataset,[2] specifically designed for aerial action recognition, contains only a fraction of that size (approximately 22,000 videos). This limited data availability further hinders the development and performance of deep learning models for aerial action recognition.

These unique challenges associated with UAV videos significantly impact the performance of existing human action recognition models. Addressing this research gap is crucial to unlocking the full potential of UAVs for HAR and realizing their broader applicability in real-world scenarios.

Main contribution In this paper, we present a real-time human action recognition algorithm specifically designed for UAV videos, with a strong emphasis on enabling efficient execution on low-power or edge hardware platforms. Our main contributions include:

1. We propose a novel autozoom algorithm that effectively extracts human-centric regions from UAV videos. This algorithm combines autofocus, cropping, and scaling techniques to isolate key action information from the human subject. This approach significantly reduces background noise, facilitating the extraction of more discriminative features for robust human behavior analysis.

2. We introduce a strategy for improving the performance of a lightweight action recognition model through synthetic data augmentation. This approach addresses the challenge of limited training data often encountered in UAV action recognition tasks.

2. RELATED WORK

2.1 Aerial Action Recognition

Despite deep learning's success in recognizing actions from ground-based videos, replicating this feat with Unmanned Aerial Vehicle (UAV) footage presents significant challenges. UAV video analysis contends with camera movement, varying viewpoints, small and scattered objects, and illumination fluctuations.[17]

Researchers have devised various methods to tackle these hurdles. One approach utilizes established 2D Convolutional Neural Networks (CNNs) like ResNet[18] or MobileNet[19] to analyze individual frames within the video. The results are then combined for a final classification.[20–22] Another approach leverages dual-stream CNNs, which process both motion and appearance data simultaneously to enhance recognition capabilities.[6,23]

Moving beyond spatial analysis, 3D Convolutional Networks have emerged to capture both spatial and temporal information. This allows for a more comprehensive understanding of human actions and their environment within aerial videos, addressing the complexities of motion analysis.[2,21,24–26] The latest advancements integrate attention mechanisms with CNNs, specifically designed for resource-constrained devices used in drone applications (e.g., AZTR[27]). Information theory has also been explored to separate crucial motion data from background noise, alongside the use of keyframe selection techniques.[28,29] Unlike most of the previous methods, our method focuses on real-time human action recognition on low-power edge devices.

2.2 Synthetic Data

The scarcity and high cost of real-world data for robotics training have fueled a surge in interest in synthetic data. Unlike real-world data collection, which can be cumbersome, expensive, and limited in its adaptability to new environments, synthetic data offers a controlled and scalable solution.[30]

Recent years have witnessed the development of numerous synthetic datasets specifically tailored to various robotics applications.[31,32] These datasets encompass diverse areas like image classification,[33] segmentation,[34] and action recognition for both ground-based and aerial robots.[35,36] Pioneering datasets like VisDA[37] have played a crucial role in advancing domain adaptation techniques, while specialized collections for tasks like action recognition in aerial videos (e.g., NEC-DRONE dataset[25]) have further propelled research efforts.

Beyond pre-built datasets, the emergence of advanced robotics simulators like CARLA,[38] GTA,[39] and NVIDIA ISAAC/Omniverse[40] empowers researchers to create custom virtual environments for synthetic data generation. These simulators offer a significant advantage by providing error-free environments, ensuring consistency and reliability in the generated data. Additionally, they streamline the process of ground truth annotation, a critical aspect of effective training. This ability to create tailored virtual environments, coupled with the benefits of error-free data, makes these simulators invaluable tools for accelerating advancements in robotics training. By embracing synthetic data and the capabilities of these advanced tools, researchers can overcome the limitations associated with real-world data collection, paving the way for a new generation of more robust, adaptable, and intelligent robots capable of operating effectively in diverse and dynamic environments.

3. METHODOLOGY

3.1 Overall Method

As shown in Figure 2, our method analyzes individual video frames. An initial stage employs an auto zoom algorithm with a lightweight detector to efficiently locate potential targets (likely humans) in the frames. The auto zoom algorithm utilizes sparse detection and bounding box shifting techniques to minimize processing demands. The identified targets are then aligned within the frames for subsequent processing. The recognition model is a Mobile Video Net (MoViNet)[14] stream architecture employing 2D+1 convolution net analyzes the aligned frames, incorporating temporal information from previous frames to achieve real-time inference without sacrificing accuracy, see Figure 3. This approach ensures efficient target identification and robust action recognition within resource-constrained environments.

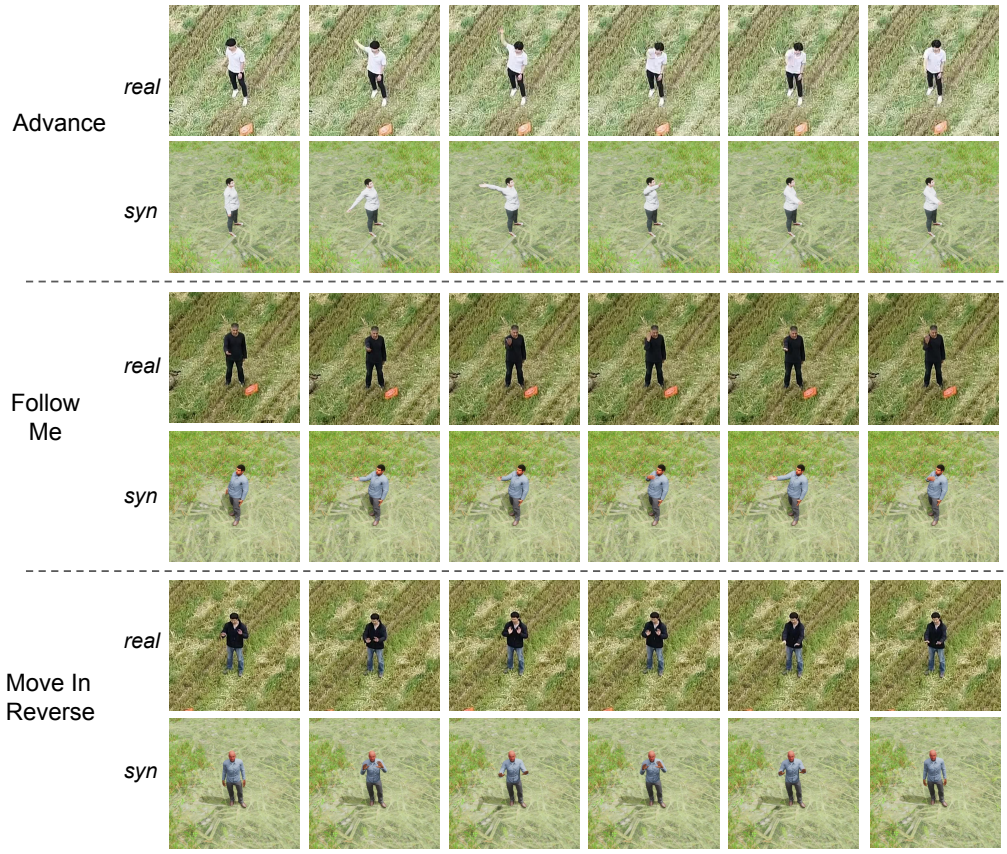

Figure 2. **Real and Synthetic UAV data from RoCoG.**[1]

3.2 Auto Zoom

We present a novel auto-zoom algorithm that significantly improves the efficacy of aerial video action recognition models. Traditional approaches often struggle with the inherent challenge of small target objects in high-altitude aerial footage. Our auto zoom algorithm tackles this issue by first automatically identifying and zooming in on the target object, dynamically adjusting the zoom level to achieve a target pixel representation of 16-22% of the frame compared to the typical 2-5% in raw footage. This magnified view provides richer details for feature extraction, enabling the model to concentrate on the action itself rather than redundant background information. By centering the target object within each frame, the zoom effectively mitigates the influence of UAV motion on feature extraction. This leads to a more robust model with improved performance.

Specifically, the algorithm employs a lightweight human detector to generate the bounding boxes and operates through a dynamic three-step process to optimize efficiency and accuracy without demanding extensive computational resources. Initially, it dynamically selects the size of the cropping region based on the target's bounding box, ensuring the target occupies an ideal 15% to 20% of the frame to balance detail and contextual information. Subsequently, to further conserve computational resources, the algorithm performs detection only on key frames, significantly reducing the processing load. The algorithm predicts the target's position in upcoming frames using a high-confidence threshold from previous detections to maintain focus and accuracy, especially in sequences where UAV movement could potentially introduce noise or outliers. To address potential inaccuracies in initial key frames, a confidence score filter (>0.8) is applied to eliminate unreliable bounding boxes. Subsequent bbox locations are predicted using a pre-defined equation below based on previous key frames, ensuring smoother and more accurate target tracking throughout the video.

$$
\begin{aligned}
&\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} x_t + (D_t + \delta_{D_t}) \cdot \cos(\theta_{t-1} + \delta_\theta) \\ y_t + (D_t + \delta_{D_t}) \cdot \sin(\theta_{t-1} + \delta_\theta) \end{pmatrix}, \\
&\delta_{D_t} = D_t - D_{t-1}, D_t = distance\left(\begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}, \begin{pmatrix} x_t \\ y_t \end{pmatrix}\right), \\
&\theta_{t-1} = \arctan(\frac{y_t - y_{t-1}}{x_t - x_{t-1}}), \delta_\theta = \theta_{t-1} - \theta_{t-2}.
\end{aligned} \tag{1}
$$

x_t and y_t are the coordinates of the bbox at key frame t. D_t, θ_t represent the shifting distance and angle between key frames $t-1$ and t, respectively. δ_{D_t} and δ_{theta} stand for deviations of the shifting distance and angle at frame t.

After the auto zoom process, the algorithm utilizes the Euclidean distance metric to evaluate the difference between the predicted bounding box (bbox) and the one produced by the detector. This step acts as a verification layer for the detection outcomes, given that the efficacy of a lightweight detector may not be as reliable or on par with that of larger detectors. If this distance is within a set threshold, it indicates a successful detection. On the other hand, if the distance goes beyond this threshold, or if there's no detection at all, the predicted bbox is assigned to the current key frame.

By integrating these techniques, the auto zoom algorithm offers a robust solution for aerial video action recognition. It overcomes the challenges posed by small target objects and UAV motion, while simultaneously optimizing computational efficiency.

3.3 Synthetic Data Augmentation

As mentioned in Section.1, the challenge of acquiring and annotating high-quality UAV video data significantly hampers the development of aerial action recognition models. Ground-camera datasets, such as Kinetics-400,[16] contain hundreds of thousands of labeled videos, whereas aerial-specific datasets like UAV-Human[2] have substantially fewer, with around 22,000 videos. This scarcity of labeled UAV video data restricts the training and performance of deep learning models for aerial action recognition tasks.

To mitigate these limitations, synthetic data generated from game engines like Unity are utilized to augment real-world datasets. This synthetic data,[1] crafted with high-quality 3D assets and animated using both skeleton-based and motion capture techniques, adds nearly 107K videos to the dataset, significantly expanding the volume of training data. However, direct inclusion of this synthetic data in training risks biasing the model towards synthetic features, potentially degrading its real-world performance. Furthermore, the necessity to deploy these models on low-power edge devices, where computational resources are limited, raises additional concerns over model size and the risk of overfitting to the large-scale synthetic dataset.

To mitigate the challenges posed by the difference in the size of real-world and synthetic data, we implement a balanced training strategy. Our model undergoes training with a mix of both real-world and synthetic data, despite the synthetic dataset being significantly larger. To counteract potential biases towards synthetic features, we adjust training batches to maintain a roughly 50/50 split between real and synthetic data. This is achieved by oversampling the real data, ensuring that it is not overshadowed by the synthetic dataset within any given training batch. This balanced approach enhances the model's learning and generalization capabilities by drawing on the comprehensive variety of the combined dataset, while also reducing the risk of overfitting and synthetic data bias. Ultimately, this strategy is designed to achieve an optimal equilibrium, leveraging the extensive synthetic data resources to enhance model performance while maintaining its relevance and effectiveness for real-world UAV-based action recognition tasks.

4. RESULTS

4.1 Datasets

We evaluate our method on RoCoG-v2 dataset,[1] which is an indispensable asset for action recognition research, offering a rich collection of both real and synthetic videos. This dataset comprises 482 real videos alongside

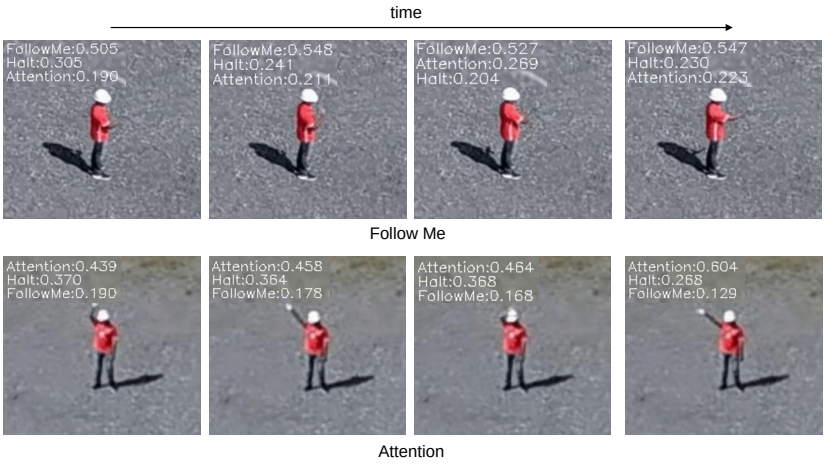

Figure 3. **Inference on real world data.**

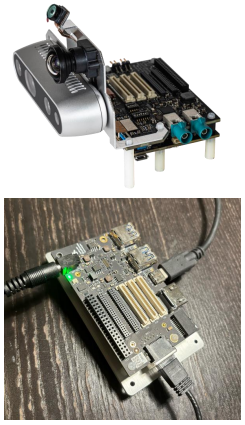

Figure 4. **Qualcomm RB5 Platform**

106,996 synthetic ones, spanning seven distinct action classes captured from both ground-level and aerial viewpoints. These classes are based on seven control signals from the U.S. Army Field Manual, encompassing commands such as "follow me," "advance," "halt," "rally," "attention," "move forward," and "move in reverse." A particular challenge posed by RoCoG-v2 is the visual similarity between certain actions—for instance, "move forward" versus "move in reverse." This similarity presents a significant hurdle for action recognition algorithms, testing their precision in distinguishing and categorizing these closely related actions accurately.

UAVHuman[2] is one of the largest UAV-based human behavior understanding datasets. This benchmark sets a new standard for human behavior analysis with UAVs, offering 67,428 multi-modal video sequences covering action recognition, 22,476 frames for pose estimation, 41,290 frames for person re-identification across 1,144 identities, and 22,263 frames dedicated to attribute recognition. UAVHuman was meticulously compiled over three months, capturing a wide array of urban and rural environments during both day and night, ensuring a rich diversity in subjects, backgrounds, lighting, weather conditions, occlusions, camera movements, and UAV flying attitudes.

4.2 Implementation Details

Evaluation metrics: We evaluate our method and other state-of-the-art methods using Top-1 accuracy scores, where the predictions are considered to be correct if the top 1 highest probability answers match the actual label.

Edge devices: We use a robotic platform (Qualcomm Robotics RB5) with Qualcomm Kryo 585 CPU and Qualcomm Adreno 650, see Figure 4. The efficient models are trained using TensorFlow and deployed using Robot Operating System 2 (ROS2) Galactic. We also test our method on Desktop with an RTX A5000 GPU.

Environmental setup: All models in this paper are trained using NVIDIA GeForce 2080Ti GPUs and NVIDIA RTX A5000 GPUs. The initial learning rate is set at 0.1 for training from scratch and 0.05 for initializing with Kinetics pre-trained weights. Adam is used as the optimizer with 0.0005 weight decay and 0.9 momentum. We use cosine/poly annealing for learning rate decay and multi-class cross-entropy loss to constrain the final predictions.

4.3 Results on RoCoG v2

Our evaluation on the RoCoG v2 dataset employed a fine-tuned MobileNet v2[19] as the pivotal detector within our autozoom algorithm, with Mobile Video Networks (MoViNets),[14] specifically the MoViNet A0 model, serving as the backbone. All the models are initialized with pre-trained Kinetics 400 weights. We tested the efficacy of our method and compared it with others under three distinct training scenarios: exclusively on synthetic data, solely on real data, and a combination of both, with all methods subsequently tested on real data. The results underscore the effectiveness of our approach, revealing a significant enhancement in top-1 accuracy across all

Method	Data	Frames	Input Size	GFLOPS	Params	Top-1
I3D[12]	syn	16	256 × 256	216	12M	48.0
X3D[11]	syn	16	256 × 256	186	3.8M	34.5
Ours	syn	16	172 × 172	5.51	3.1M	**57.1**
I3D[12]	real	16	256 × 256	216	12M	68.1
X3D[11]	real	16	256 × 256	186	3.8M	70.3
MoViNet A0[14]	real	16	172 × 172	2.71	3.1M	78.5
MoViNet A3[14]	real	16	256 × 256	56.9	5.3M	79.8
Ours	real	16	172 × 172	5.51	3.1M	**86.7**
I3D[12]	real+syn	16	256 × 256	216	12M	60.4
X3D[11]	real+syn	16	256 × 256	186	3.8M	63.0
MoViNet A0[14]	real+syn	16	172 × 172	2.71	3.1M	74.6
MoViNet A3[14]	real+syn	16	256 × 256	56.9	5.3M	78.2
Ours	real+syn	16	172 × 172	5.51	3.1M	**90.0**

Table 1. **Results on RoCoG-v2.** Our method shows a notable increase in top-1 accuracy in all three scenarios—9.1% improvement with synthetic data, 6.9% with real data, and 11.8% with a mix of both.

Method	frames	Input Size	Inference Time/frame(ms)	
			RB5	Desktop
MoViNet A0[14]	16	172 × 172	33.2	0.54
MoViNet A2[14]	16	224 × 224	106.4	1.40
MoViNet A3[14]	16	256 × 256	124.0	1.61
I3D[12]	16	256 × 256	-	2.19
X3D[11]	16	256 × 256	-	1.54
Ours	16	172 × 172	**56.5**	**0.76**

Table 2. **Inference Efficiency on RB5 and Desktop.** Our method achieves a better trade-off between model performance and inference rate on low-power edge device and high-end desktop.

scenarios. Specifically, as shown in Table 1, our method shows a notable increase in top-1 accuracy in all three scenarios—9.1% improvement with synthetic data, 6.9% with real data, and 11.8% with a mix of both. This improvement is attributed to our innovative autozoom algorithm, which substantially minimizes background noise and enhances the extraction of discriminative features for a more robust analysis of human behavior.

Moreover, our strategy of augmenting synthetic data further improves the accuracy. While training on a mix of real and synthetic data, our method demonstrated a slight increase in top-1 accuracy(3.3%), contrary to the observed accuracy declines in other methods, likely due to overfitting issues. This indicates not only the effectiveness of the autozoom algorithm in isolating relevant action information but also the benefits of synthetic data augmentation in enhancing the model's comprehension of semantic and motion cues within human activity areas, without succumbing to overfitting.

4.4 Inference Efficiency

In our comparative analysis, we assessed the inference time per video of our proposed methods against other methodologies, utilizing two distinct hardware platforms: the Qualcomm Robotics RB5 platform as the edge device and a high-end desktop equipped with an A5000 GPU in Table 2. For our approach, we incorporated the MoViNet A0 model as the backbone. The findings, detailed in our results, indicate a marked superiority in the speed of inference of our method on both devices compared to previous approaches. It's important to note that due to compatibility issues, the inference times for I3D and X3D models were exclusively reported on the desktop platform, as these models are not supported by the RB5 platform.

While our method demonstrates a slight increase in inference time on both devices when compared to the baseline MoViNet A0 model, this marginal delay is counterbalanced by a substantial improvement in accuracy, as evidenced by the comparative accuracy results. This balance between inference speed and accuracy highlights the

Method	Backbone	Extra data	Input Size	Frames	Views	GFLOPs	Params.	Top-1 Acc ↑
Slowfast[41]	ResNet50	K400	224×224	8	5×3	99	50M	36.3
I3D[12]	ResNet101	K400	540×960	8	10×3	108	28M	21.1
FNet[42]	I3D	K400	540×960	8	10×3	108	28M	24.3
X3D[11]	-	K400	540×960	8	10×3	65	4M	36.6
FAR[43]	X3D	K400	540×960	8	10×3	65	4M	38.6
DiffFAR[44]	X3D	K400	540×960	8	10×3	130	4M	41.9
MViT v1[45]	MViT-B	K400	224×224	16	5×1	71	37M	24.3
ViViT FE[46]	ViT-B	IN-21K	224×224	16	1×1	284	116M	34.1
TimesFormer[47]	ViT-B	K400	224×224	8	1×3	196	131M	38.4
Ours	X3D	K400	224×224	16	10×3	7	4M	**47.4**

Table 3. **Results on UAV-Human.** Our method, when combined with the X3D backbone, demonstrates a significant accuracy improvement of 5.5% on the UAV-Human dataset.

efficacy of our approach, illustrating its potential for real-time application on both edge and high-end computing devices without compromising on performance.

4.5 Results on UAVHuman

Extending the versatility of our approach, we integrated our autozoom algorithm with another backbone model, X3D,[11] and conducted evaluations on the UAV-Human dataset. The outcomes, as presented in our findings, showing 5.5% accuracy improvement, further affirm the superiority of our proposed method. This successful integration demonstrates that our autozoom algorithm can enhance action recognition accuracy across various backbone models, proving its effectiveness on large-scale real datasets. This adaptability underscores the algorithm's potential to significantly improve performance in diverse action recognition tasks, showcasing its broad applicability and the promising direction for future research in enhancing UAV-based human action recognition.

5. CONCLUSION

In conclusion, our research addresses the pivotal challenges in human action recognition (HAR) using Unmanned Aerial Vehicles (UAVs), overcoming obstacles such as the reduced scale of human figures in videos, varying viewing angles, and the dynamic nature of UAV-captured footage. These issues, alongside the scarcity of labeled UAV video data and computational constraints on UAV platforms, necessitate specialized solutions for effective aerial action recognition.

Our contributions, including the novel autozoom algorithm and synthetic data augmentation, significantly enhance the ability to isolate and analyze human actions in UAV videos. These advancements not only improve recognition accuracy but also ensure computational efficiency, enabling real-time processing on edge devices. By tackling the limitations of current models and data availability, our work broadens the potential applications of UAVs in sectors like security, search and rescue, and traffic monitoring, marking a step forward in realizing the full capabilities of UAVs for complex action recognition tasks in varied environments.

ACKNOWLEDGMENTS

This work was supported in part by ARO Grants W911NF2110026, W911NF2310046, W911NF2310352 and Army Cooperative Agreement W911NF2120076

REFERENCES

[1] 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, IEEE (2023).

[2] 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 [*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 16266–16275 (2021).

[3] Zhao, H., Wang, H., Wu, W., and Wei, J., "Deployment algorithms for uav airborne networks toward on-demand coverage," *IEEE Journal on Selected Areas in Communications* **36**(9), 2015–2031 (2018).

[4] Gong, Y., Yu, X., Ding, Y., Peng, X., Zhao, J., and Han, Z., "Effective fusion factor in fpn for tiny object detection," (2020).

[5] Wang, Y., Ding, L., and Laganiere, R., "Real-time uav tracking based on psr stability," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*], 0–0 (2019).

[6] Perera, A. G., Law, Y. W., and Chahl, J., "Drone-action: An outdoor recorded drone video dataset for action recognition," *Drones* **3**(4), 82 (2019).

[7] Perera, A. G., Law, Y. W., and Chahl, J., "Uav-gesture: A dataset for uav control and gesture recognition," (2019).

[8] Hoey, J. and Little, J. J., "Value-directed human behavior analysis from video using partially observable markov decision processes," *IEEE transactions on pattern analysis and machine intelligence* **29**(7), 1118–1132 (2007).

[9] Supancic III, J. and Ramanan, D., "Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning," in [*Proceedings of the IEEE international conference on computer vision*], 322–331 (2017).

[10] Burdziakowski, P., "A novel method for the deblurring of photogrammetric images using conditional generative adversarial networks," *Remote Sensing* **12**(16), 2586 (2020).

[11] Feichtenhofer, C., "X3d: Expanding architectures for efficient video recognition," in [*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 203–213 (2020).

[12] Carreira, J. and Zisserman, A., "Quo vadis, action recognition? a new model and the kinetics dataset," in [*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 6299–6308 (2017).

[13] Pan, J., Bulat, A., Tan, F., Zhu, X., Dudziak, L., Li, H., Tzimiropoulos, G., and Martinez, B., "Edgevits: Competing light-weight cnns on mobile devices with vision transformers," *arXiv preprint arXiv:2205.03436* (2022).

[14] Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., and Gong, B., "Movinets: Mobile video networks for efficient video recognition," in [*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 16020–16030 (2021).

[15] Demir, U., Rawat, Y. S., and Shah, M., "Tinyvirat: Low-resolution video action recognition," in [*2020 25th International Conference on Pattern Recognition (ICPR)*], 7387–7394, IEEE (2021).

[16] 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).

[17] Nguyen, K., Fookes, C., Sridharan, S., Tian, Y., Liu, X., Liu, F., and Ross, A., "The state of aerial surveillance: A survey," *arXiv preprint arXiv:2201.03080* (2022).

[18] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2016).

[19] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," (2017).

[20] Mou, L., Hua, Y., Jin, P., and Zhu, X. X., "Event and activity recognition in aerial videos using deep neural networks and a new dataset," in [*IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*], 952–955, IEEE (2020).

[21] Mliki, H., Bouhlel, F., and Hammami, M., "Human activity recognition from uav-captured video sequences," *Pattern Recognition (PR)* **100**, 107140 (2020).

[22] Mishra, B., Garg, D., Narang, P., and Mishra, V., "Drone-surveillance for search and rescue in natural disaster," *Computer Communications* **156**, 1–10 (2020).

[23] Perera, A. G., Law, Y. W., Ogunwa, T. T., and Chahl, J., "A multiviewpoint outdoor dataset for human action recognition," *IEEE Transactions on Human-Machine Systems* **50**(5), 405–413 (2020).

[24] 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).

[25] Choi, J., Sharma, G., Chandraker, M., and Huang, J.-B., "Unsupervised and semi-supervised domain adaptation for action recognition from drones," in [*IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 1717–1726 (2020).

[26] Wang, X., Xian, R., Guan, T., and Manocha, D., "Prompt learning for action recognition," *arXiv preprint arXiv:2305.12437* (2023).

[27] Wang, X., Xian, R., Guan, T., de Melo, C. M., Nogar, S. M., Bera, A., and Manocha, D., "Aztr: Aerial video action recognition with auto zoom and temporal reasoning," *2023 IEEE International Conference on Robotics and Automation (ICRA)* , 1312–1318 (2023).

[28] Xian, R., Wang, X., and Manocha, D., "Mitfas: Mutual information based temporal feature alignment and sampling for aerial video action recognition," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 6625–6634 (2024).

[29] Xian, R., Wang, X., Kothandaraman, D., and Manocha, D., "Pmi sampler: Patch similarity guided frame selection for aerial action recognition," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 6982–6991 (January 2024).

[30] de Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., and Hodgins, J., "Next-generation deep learning based on simulators and synthetic data," *Trends in cognitive sciences* **26**(2), 174–187 (2022).

[31] Zherdeva, L., Minaev, E., Zherdev, D., and Fursov, V., "Synthetic dataset for navigation tasks of autonomous systems and ground robots," in [*2021 International Conference on Information Technology and Nanotechnology (ITNT)*], 1–4, IEEE (2021).

[32] Kiefer, B., Ott, D., and Zell, A., "Leveraging synthetic data in object detection on unmanned aerial vehicles," in [*2022 26th International Conference on Pattern Recognition (ICPR)*], 3564–3571, IEEE (2022).

[33] Bayraktar, E., Yigit, C. B., and Boyraz, P., "A hybrid image dataset toward bridging the gap between real and simulation environments for robotics: Annotated desktop objects real and synthetic images dataset: Adoreset," *Machine Vision and Applications* **30**(1), 23–40 (2019).

[34] Chu, F.-J., Xu, R., and Vela, P. A., "Learning affordance segmentation for real-world robotic manipulation via synthetic images," *IEEE Robotics and Automation Letters* **4**(2), 1140–1147 (2019).

[35] de Melo, C. M., Rothrock, B., Gurram, P., Ulutan, O., and Manjunath, B. S., "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, IEEE (2020).

[36] Alharbi, F., Ouarbya, L., and Ward, J. A., "Synthetic sensor data for human activity recognition," in [*2020 International Joint Conference on Neural Networks (IJCNN)*], 1–9, IEEE (2020).

[37] Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K., "Visda: The visual domain adaptation challenge," *arXiv preprint arXiv:1710.06924* (2017).

[38] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V., "Carla: An open urban driving simulator," in [*Conference on robot learning*], 1–16, PMLR (2017).

[39] Richter, S. R., Vineet, V., Roth, S., and Koltun, V., "Playing for data: Ground truth from computer games," in [*Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*], 102–118, Springer (2016).

[40] Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al., "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470* (2021).

[41] Feichtenhofer, C., Fan, H., Malik, J., and He, K., "Slowfast networks for video recognition," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], (October 2019).

[42] Lee-Thorp, J., Ainslie, J., Eckstein, I., and Ontanon, S., "Fnet: Mixing tokens with fourier transforms," *arXiv preprint arXiv:2105.03824* (2021).

[43] Kothandaraman, D., Guan, T., Wang, X., Hu, S., Lin, M., and Manocha, D., "Far: Fourier aerial video recognition," in [*Computer Vision – ECCV 2022*], Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., eds., 657–676, Springer Nature Switzerland, Cham (2022).

[44] Kothandaraman, D., Lin, M., and Manocha, D., "Diffar: Differentiable frequency-based disentanglement for aerial video action recognition," in [*2023 IEEE International Conference on Robotics and Automation (ICRA)*], 8254–8261 (2023).

[45] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., and Feichtenhofer, C., "Multiscale vision transformers," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 6824–6835 (October 2021).

[46] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., and Schmid, C., "Vivit: A video vision transformer," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* , 6816–6826 (2021).

[47] Bertasius, G., Wang, H., and Torresani, L., "Is space-time attention all you need for video understanding?," in [*Proceedings of the International Conference on Machine Learning (ICML)*], (July 2021).