Drone-HAT: Hybrid Attention Transformer for Complex Action Recognition in Drone Surveillance Videos

Mustaqeem Khan
MBZUAI, Abu Dhabi, UAE
mustaqeem.khan@mbzuai.ac.ae

Jamil Ahmad
MBZUAI, Abu Dhabi, UAE
jamil.ahmad@mbzuai.ac.ae

Abdulmotaleb El Saddik
University of Ottawa, Canada
elsaddik@uottawa.ca

Wail Gueaieb
University of Ottawa, Canada
wgueaieb@uottawa.ca

Giulia De Masi
Technology Innovation Institute, UAE
giulia.demasi@tii.ae

Fakhri Karray
University of Waterloo, Canada
karray@uwaterloo.ca

Abstract

Ultra-high-resolution aerial videos are becoming increasingly popular for enhancing surveillance capabilities in sparsely populated areas. However, analyzing human activities automatically, such as "who is doing what?" in these videos, is desirable to realize their surveillance potential. In contrast, atomic visual action detection has successfully recognized such activities in movie data. However, adapting it to ultra-high resolution aerial videos is challenging because the target persons appear relatively tiny from overhead views and are sparsely located. Additionally, existing atomic visual action detection methods are based on single-label actions. However, people can perform multiple actions simultaneously, so a multi-label approach would be more appropriate. To address these problems, we propose a multi-label action detection/recognition framework using a hybrid attention vision transformer (HAT) to recognize recurrent actions more efficiently. Additionally, a multi-scale, multi-granularity module inside the action recognition transformer block extracts relevant features without redundancy. Using the Okutama Dataset, we demonstrated that our method performs better than existing state-of-the-art methodologies for interpreting aerial videos for human activity.

1. Introduction

Surveillance cameras are commonly deployed in cities to ensure public safety, but not in sparsely populated regions with limited safety concerns. Drones may monitor such areas periodically because there are no tall trees or buildings. The mobility of drones allows them to monitor a wide range of sparsely populated areas, and it is advantageous to analyze the surveillance videos automatically to determine

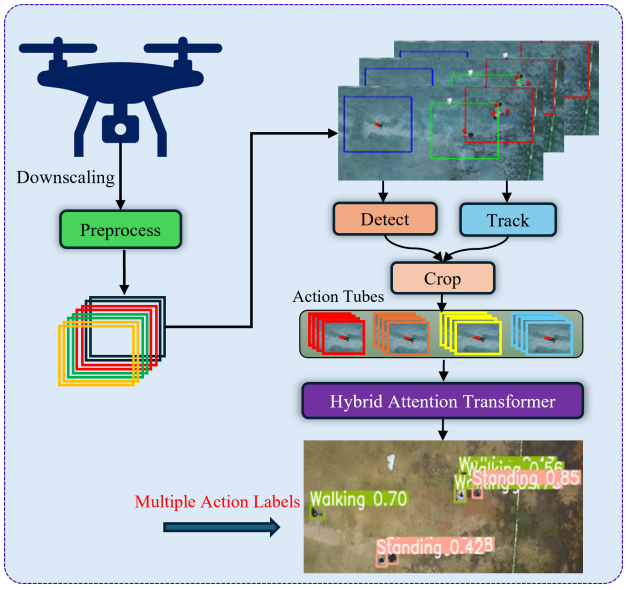

Figure 1. Overview of the proposed system that uses ultra-high-resolution aerial images acquired by drones to detect and recognize complex recurrent actions. As a result of the detected actions, attention maps are generated, highlighting the target individuals to estimate the action label.

"who is doing what?". According to atomic visual action, people can discover their spatiotemporal location and actions in videos at each frame to determine human actions [14]. However, aerial drone videos make accurate detection challenging due to their unique characteristics, such as ultra-high-resolution images, tiny object appearance, sparse location, and fast movements. To streamline action detection in aerial videos, we propose a novel framework that seamlessly integrates object detection, multi-object tracking, and action recognition (illustrated in Figure 1).

Action detection in surveillance videos relies heavily on

object detection, the detected object's quality, and effective motion capturing. Though drone-captured aerial images are usually ultra-high-resolution, the objects typically captured from a distance appear tiny with unusual backgrounds, making them hard to detect using simple object detection methods [8, 21, 22, 28]. Furthermore, processing such high-resolution imagery is computationally expensive. Downsampling is typically used to reduce the computational burden; it further reduces the spatial resolution of the objects, thereby decreasing performance. Alternatively, some existing methods use a sliding window to make patches from an image before object detection. In contrast, approaches crop an ultra-high-resolution aerial image into smaller patches (proposals) before object detection [10, 16, 26]. Object detection performance has improved significantly with the use of these methods. However, these methods are inefficient when the target objects are sparsely located. To address the issue, we utilize regional locations by exclusively selecting areas that contain the target objects. This method may result in fewer selected areas compared to using a sliding window in situations where people are sparsely located. However, it's important to note that atomic visual actions (AVA) detection provides estimates of actions for each frame, which essentially identifies "who is doing what."

To gather information about a person's activity, it is important to consider their spatiotemporal context. Typically, spatiotemporal tubes are utilized in recent models for this purpose [13, 17]. However, in drone-recorded aerial videos, the person being recorded may appear to shift positions due to the drone's movement, even though they are staying in the same location. To avoid this issue, we use a multi-object tracking method to create spatiotemporal tubes that capture the person's movements over time. We then align the spatiotemporal tube to its first frame to eliminate any discrepancies caused by the drone's movement. We focus on the target person within recurrent tubes to ensure accurate action recognition. Unlike potentially inconsistent observations of other objects, we assume consistent observation of the target person throughout their tube.

We propose a novel hybrid attention vision transformer (HAT) that utilizes multi-scale and multi-granularity fusion. This approach efficiently recognizes recurrent actions performed by the target person. Our team's contributions are threefold. Firstly, we propose a new framework for multi-label action detection and recognition on aerial surveillance videos that outperforms existing baseline methods based on experiments. Secondly, we present a vision transformer-based action recognition model that utilizes a fused vision strategy of multi-scale and granularity. Thirdly, we introduce a new granularity layer that combines coarse-grained and fine-grained information to more efficiently and quickly identify human action. Furthermore, we conducted extensive experiments and evaluated our model on the widely used Okutama recurrent action recognition dataset [4] and achieved higher accuracy than existing models, making it more efficient for multi-label action detection and recognition tasks. An overview of the proposed approach is illustrated in Figure 1

2. Related Work

Detecting small objects is a challenging problem, and many studies have attempted to address it. There are two main scenarios for small object detection and recognition. In one scenario, the image has low resolution, resulting in tiny objects containing only a few pixels. Techniques like amplification [15], and resolution enhancement have been applied to improve detection performance [3]. Another scenario contains many pixels on a relatively small area of the image, which makes it appear relatively small, although it contains numerous pixels. High-resolution aerial images exemplify this second scenario, where object detection and recognition directly on the original image are preferred. While prior approaches have utilized patch-based object detection for aerial videos for some time [10, 16, 26], the recent advancement lies in jointly employing region proposals and clustering to reducing the number of necessary patches, particularly when dealing with sparsely distributed objects [19].

Density map regression on downsized aerial images can learn promising regions likely containing objects. After defining an image size, these regions can be further clustered based on their relative distances. An effective clustering strategy should satisfy two key conditions - first, reducing the number of images, and second, completely preserving the object's appearance. However, these two conditions can somewhat conflict. Solely meeting the first condition may result in partially cropped objects, while assigning each object to an individual patch can satisfy the second condition but introduce redundant patches.

Previous work used grid-based clustering, but it is constrained by predefined grid size and location, potentially resulting in incomplete object cropping and affecting bounding box detection [19]. To address this issue, we utilize peak point Non-Maximum Suppression and hierarchical clustering in our pre-processing steps to ensure each object is fully contained in at least one frame.

Architectures based on Transformer models [31] have emerged recently. Beyond its natural language processing roots, the Transformer excels at capturing long-range dependencies and modeling complex relationships. People started exploring Transformer applications in computer vision after introducing Vision Transformer (ViT) in 2020, which initially applied it to image classification [9]. This method recognizes actions in videos by treating image sequences as temporally evolving frames, capturing both spatial content and motion.

Building upon the success of ViT as a backbone for action recognition, the Swin Transformer (released in 2021) introduces a novel computational strategy. This strategy utilizes shifting windows to restrict self-attention calculations within local areas of the image, promoting efficiency [23, 24]. TimeSFormer captures spatiotemporal features from frame-level patches that use the main architecture of ViT for video processing [6]. ViViT is a model that uses Transformers to extract spatiotemporal tokens from videos. It is a purely Transformer-based model with multiple Transformer layers [2]. These models rapidly developed and used for action detection and recognition [5, 30, 35].

However, due to the Transformer's large size, recent research focuses on reducing attention complexity to make it more efficient. For example, MViT combines multi-scale hierarchies with Transformers using lower input/channel dimensions, progressively increasing capacity at lower resolutions [11, 20]. Some studies, like Video Transformer Network [25] and MoViNets [18], explore enhancing Transformer efficiency. These efforts aim to reduce computational/memory costs while maintaining performance for real-time processing.

We were motivated by [12, 29] and multi-granularity methods [37, 38] in diverse domains. Therefore, we present a new ViT with a Multi-Scale and Multi-Granularity Fused vision strategy for action recognition in aerial video. This design integrates a module within Transformer blocks to efficiently reduce secondary attention computations, resulting in lower computational costs and memory requirements while maintaining performance.

3. Proposed Architecture

Our proposed framework coherently generates frames, bounding boxes, and sequence tubes processed by transformers with multi-scale fusion to recognize multiple actions per frame (as illustrated in Fig. 2). Using a frame 2160 x 3840 input size, our system first downsamples frames to 1280×720. The modified YOLOv8p object detector, which is based on YOLOv8 [27], then generates bounding boxes for each detected person within the frames. A multi-person tracker connects the bounding boxes to form recurrent tubes over successive frames. Sample frames from the tracks are then input into our proposed action recognition module to obtain corresponding visual features. A multi-granularity hybrid attention module leverages these features to generate fused attention maps at multiple scales and granularities, focusing on target persons' actions. A second transformer block finally uses the concatenated attention maps and their multiplied transformed features to estimate multi-label action classes, as visually illustrated in Fig. 2. In summary, our end-to-end framework performs multi-label action detection within ultra-high resolution aerial surveillance videos by coherently processing frames, generating bounding boxes and recurrent tubes, and recognizing multiple concurrent activities via a multi-scale feature fusion transformer.

3.1. Proposed Action Recognition Module

A multiple-object tracking method was used in our approach to link bounding boxes into recurrent tubes, called DeepSort [33]. In Deep SORT, two descriptors (IoU and appearance) are combined with a Kalman filter to calculate bipartite bounding boxes. To overcome occlusions and long-time tracking problems, an appearance descriptor is derived from a CNN network trained by a Cosine Softmax Classifier [32]. In the next step, we use a novel module called Action Recognition to obtain the actions of each person at each frame once we have obtained their recurrent tubes. We present a novel Transformer unit that utilizes a combined multi-scale and multi-granularity fused vision-based strategy. As a result of combining multiple scales and granularities, a large amount of salient information is saved, both in terms of sample length and sequence length.

In the proposed actions recognition module, we utilize the transformer blocks to extract the salient information through a novel architectural paradigm to process parallel and extract global context across the entire input that is effective in the computer vision field. The attention part of the transformer quantifies how well the query matches parts of the input, guiding the model to focus on salient features. This significantly enhances the ability to recognize pertinent visual patterns and objects by simultaneously modeling inter-dependencies across the entire scene or image content.

3.1.1 Proposed Multi-Granularity Attention Module

Current transformer models heavily depend on direct attention computations to recognize actions. This approach can be extremely naive, particularly with multi-label action detection and recognition complexity. To overcome limitations in existing approaches, we introduce a novel vision transformer unit called the Multi-Granularity Attention Module (MGAM). MGAM combines two strategies for efficiency: multi-scale processing and multi-granularity sampling.

Multi-scale processing: This approach shortens the sequence length required for self-attention calculations. It starts with the original input resolution and smaller channel sizes. It then hierarchically increases the channel capacity while transforming the spatial resolution into one-dimensional signals through patches. Pooling operations are applied on intermediate tensors to capture information from different spatial scales.

Multi-granularity sampling: This strategy focuses on the most informative parts of the data. MGAM employs a

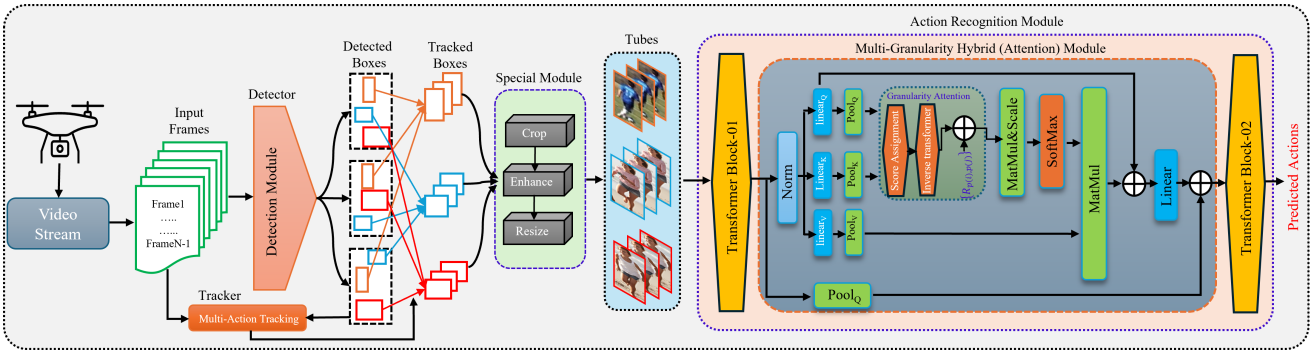

Figure 2. The proposed architecture begins with each ultra-high-resolution aerial image (3840 × 2160), wherein the system converts the video into frames. Subsequently, detectors are applied to these selected frames to produce precise bounding boxes for each person. These bounding boxes are then interconnected as tubes using a multi-person tracking algorithm. The frames are sampled from these tubes and fed to the action recognition module to generate attention maps, highlighting the target individuals to estimate multi-label action classes.

scale-granularity fusion module with score assignment and inverse transformation. This allows the model to select only the most critical cues for subsequent calculations within the multi-scale framework. Relative positional information is further incorporated by adding relative positions to the self-attention computation, considering the distance between tokens within the attention matrix.

Further enhancing efficiency, the multi-granularity computation employs a two-part approach: coarse-grained and fine-grained. **Coarse-grained computation** operates directly on the attention matrix within the multi-scale module, and **Fine-grained computation** involves performing attention computation followed by operations on the output tokens. The multi-granularity module at the core of this computation calculates the significance score of each input token. This score serves as the basis for differentiating between coarse- and fine-grained levels of analysis.

The significance score S_J uses the self-attention matrix. Each row in this matrix sums to 1 due to the attention mechanism, and the output token is obtained through a weighted summation of the attention weights. These weights represent the importance of input tokens relative to output tokens. The computation of output tokens at the self-attention level heavily relies on the attention matrix, denoted by A, and the pooling operator, denoted by $p(V;\Theta_V)$. The formula for calculating the attention significance score is presented as follows:

$$S_J = A_{1,J} \times \|P(V_J;\Theta_{V_J})\| \times \frac{\sum_{I=2}^{L'} A_{1,I} \times \|p(V_I;\Theta_{V_I})\|}{I}, \quad J \in \{2...l'\} \tag{1}$$

The output token from the attention layer is passed through a Hybrid granularity layer to calculate the significance score as:

$$I(x_j^{(t)})I(x_j) = \sum_{i=1,i\neq j}^{n} A_{i,j} = \frac{1}{h}\sum_{t=1}^{h} I(x_j^{(t)}) \tag{2}$$

Following the significance score calculation, a two-stage process integrates coarse and fine granularities through sampling. The fine-grained attention reduction leverages the attention map to eliminate redundant output tokens using the significance scores. Similarly, granularity fusion reduction reduces the length of the token sequence by focusing on tokens with less informative content. Sampling employs the Cumulative Distribution Function (CDF) based on the normalized significance scores to interpret probability distribution. The CDF allows us to sample tokens based on their importance. The inverse transform method is then applied using the following formula to perform the sampling:

$$CDF_I = \sum_{J=2}^{J=I} S_J \tag{3}$$

$$\Psi(K) = CDF^{-1}(K), \quad K \in [0,1] \tag{4}$$

To obtain the desired K samples, we employ a two-step approach that avoids excessive randomization often encountered in the Top-k method, especially with a large K value. We deviate from a purely random approach and opt for a fixed sampling strategy. This involves selecting k values from a pre-defined set: $K = \{2_1^K, 2_3^K, ..., 2_{2K-1}^K\}$. Since the significance score function ($\Psi(.)$) operates on real numbers, we identify the index of the token with the closest significance score to each chosen k value. This approach mitigates potential issues associated with exact matches in a continuous domain.

Following this sampling process, we obtain a refined attention matrix, $A_S \in \mathbb{R}^{(K'+1)\times(N+1)}$, by selecting the corresponding rows of the original attention matrix based on

the chosen indices. Finally, the output token is computed using the following formula:

$$O = A_S P(V; \Theta_V) + P(q; \Theta_q) \tag{5}$$

Building upon the significance score calculation (Equation (2)), we leverage this information to categorize tokens at the individual level. This categorization distinguishes between more informative and less informative tokens within the sequence. Let's denote the sequence of input token vectors as $X_{\text{cls}} \oplus X$. Here, $X = [X_1, X_2, ..., X_S]$ represents the length of the output token sequence after the adaptive sampling step in the previous section.

We employ a Top-k approach to identify the top K tokens with the maximum significance scores (Equation (2)). These tokens denoted as X_{in}, form the set of more informative tokens and have dimensions $X_{\text{in}} \in \mathbb{R}^{K \times D}$. The remaining tokens with lower significance scores are grouped into a separate sequence, denoted as X_{low}, representing less informative tokens. This sequence has dimensions $X_{\text{low}} \in \mathbb{R}^{(K'-K) \times D}$, where K' is the total number of unique indices selected after sampling. Unlike pruning operations that completely remove uninformative tokens, our approach retains them in the sequence X_{low}.

$$X'_{low} = pool(X_{low}) \tag{6}$$

OR average weighted pooling:

$$\alpha \cdot x_k X'_{low} = softmax(I(x_{low})) = pool(\alpha X_{low}) \tag{7}$$

The resulting sequence incorporates informative and less informative tokens following the token classification and aggregation steps. The length of the token sequence becomes $[x_{cls} \oplus X_{in} \oplus X'_{low}]$ after granularity mixing.

where $X_{\text{low}'} \in \mathbb{R}^{(K'-K)'} \times D$ represents the aggregated version of sequence X_{low}

This process effectively reduces the original sequence length while preserving essential information.

To determine the optimal number of informative tokens (K), we introduce a new set of learnable parameters, denoted by $R = [R_1, ..., R_S]$. These parameters are constrained to the range $[0, 1]$ through a uniform distribution. The learnable parameter R_I associated with each token X_I influences its representation in the final sequence. Here's how:

- If R_I is close to 1, the token's influence (X_I) remains relatively unchanged.
- If $R_{=}I$ is closer to 0, the token's influence is scaled down, emphasizing the focus on more informative tokens.

The specific value of K_l for the l^{th} layer is determined by the following formula:

$$K_l = ceil(sum(l; R)) \times S.\text{K}.K_{l+1} \leq k_l \tag{8}$$

The final tensor feeds further to the final transformer block for the final prediction and recognition of recurrent actions in the input sequence, as mentioned in Fig. 2

3.1.2 Proposed Action Module Configuration

Our proposed multi-granularity hybrid module builds upon the MViTv2 [20] architecture with related stages/components. Every stage utilizes transformer blocks with a channel of a consistent size. The input data undergoes a preprocessing step, which is divided into small patches and then cropped using an inception approach. This processed data is then projected into specific cubes before feeding into the network. As the network progresses through the four scale stages, the spatial resolution of the features is reduced while the channel dimension is concurrently increased. Importantly, pooling operations only affect the feature maps, excluding the processed class token information.

The network's attention strategy employs several heads that grow proportionally with the increasing channel dimension. Our innovative multi-granularity module is designed to seamlessly integrate with these scale modules. Additionally, a specific sampling parameter is set within the attention layer for optimal performance. Finally, a learnable parameter sequence is utilized to determine the most informative tokens at each level, enhancing the overall efficiency of the model.

4. Experimental Setup & Results

4.1. Dataset

The Okutama-action [4] is a large and popular dataset, enabling the evaluation and experimentation of a multi-label visual action detection system for aerial videos 3. We utilized the Okutama dataset for our proposed action detection and recognition system. This dataset consists of 43 minutes of footage recorded by two drones in the morning and evening, meticulously annotated with bounding boxes encompassing relevant subjects in each frame. Notably, the annotations encompass 12 distinct categories of human actions, such as Handshaking, Hugging, Reading, Drinking, etc. Significantly, the multi-label nature of the annotations allows for the assignment of multiple action classes to a single subject simultaneously, accurately reflecting real-world scenarios where individuals may engage in multiple actions concurrently. For instance, a person could be labeled with both "Reading" and "Sitting" actions simultaneously. By

Figure 3. Sample frames from Okutama drone action detection dataset [4]

leveraging this comprehensive and representative dataset, the researchers aim to develop and validate a robust multi-label visual action detection system tailored for aerial video analysis.

4.2. Experimental Details

The experimental setup and implementation details for the proposed multi-label visual action detection system using the Okutama-action dataset are meticulously outlined [4]. Adhering to previous work, the dataset is divided into a training set comprising 33 aerial videos and a testing set with 10 aerial videos. Crucial hyper-parameters, such as the peak point confidence threshold and distance threshold, are determined through validation experiments, ensuring optimal performance. To train the action recognition module effectively, we employ a two-pronged approach. First, we sample 32 ground-truth recurrent tubes from each action class to ensure the model learns diverse action sequences. In order to significantly enhance the performance of our machine learning model, we confidently employ the Adam optimizer coupled with a decaying learning rate schedule. This optimizer helps the model converge efficiently while the decaying learning rate prevents overfitting. Additionally, we leverage data augmentation techniques such as flipping, rotation, resizing, and cropping. These techniques artificially expand the training dataset, improving the model's generalization ability to unseen data. Finally, the pre-normalization setup with residual connections and layer normalization further aids in training by facilitating the flow of gradients and enhancing model stability. Remarkably, despite the challenges posed by large-size aerial videos, the framework ingeniously decomposes the problem into multiple simpler tasks, enabling a plug-and-play model implementation on a single NVIDIA TITAN X GPU. This modular approach streamlines the training process and demonstrates the system's scalability and efficiency in handling complex aerial video data.

4.3. Ablation Study

The experimental evaluation process is rigorously designed to analyze the proposed multi-scale and multi-granularity action recognition module's performance on the targeted Okutama dataset. The model undergoes random initialization and training on the dataset in typical scenarios. However, the researchers conduct ablation experiments to gain comprehensive insights, meticulously examining the Top-1 accuracy and complexity metrics. The multi-head pooled self-attention mechanism is a critical design element within the multi-scale module. Here, the selection of the pooling function is paramount for optimal performance. After a comprehensive evaluation, we have determined that channel-wise convolution with layer normalization pooling is the most effective approach among the three methods examined.

Average Pooling: This method reduces the sequence length by averaging elements within a sliding window. However, our experiments revealed that average pooling significantly hinders performance (by 2.3 % and 3.5 % compared to max pooling and convolutional pooling, respectively) as it can overlook crucial information within the sequence. **Channel-wise Convolution with Layer Normalization Pooling:** This method emerged as the most effective approach, demonstrably outperforming both max pooling and average pooling. As shown in Table 1, applying channel-wise convolution with layer normalization pooling resulted in a noteworthy 1.2 % improvement compared to max pooling.

Table 1. Ablation study about the impact of Pooling Function (Max, Average, ConvLN) on Accuracy (Acc)

Kernel Size (s+1)	Pooling Function	Acc
s+1	Max	54.22
s+1	Average	57.44
s+1	Conv	58.85
3x3x3	ConvLN	60.76

We investigate coarse-grained pooling within the multi-granularity module's granularity fusion layer to optimize the system for real-time applications. Pooling aggregates tokens processed coarsely. We explore two methods, average pooling and weighted average pooling, and evaluate their effectiveness through ablation studies. Using one and five units, we test different configurations for coarse-grained units in the mixed-granularity setting. Pooling reduces the computational sequence length by aggregating coarsely sampled tokens while retaining information. Experiments show that weighted average pooling achieves 1.4 % better than average pooling. This is because it considers each token's importance, leading to a more comprehen-

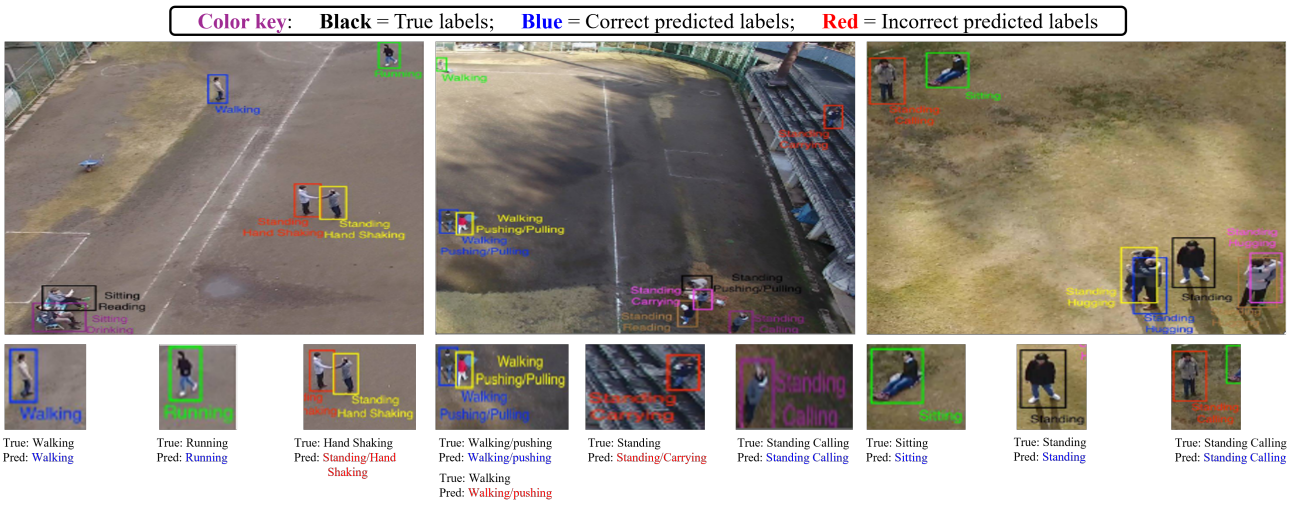

Figure 4. Visual result examples of the proposed multi-label action detection/recognition system.

sive representation of the input sequence and better feature capture. Additionally, the later stages of coarse-grained sampling compensate for reducing the fine-grained attention matrix, providing the model with more recognition features. This configuration improves performance by 2.8 % compared to using one unit (See Table 2).

Table 2. Ablation study about the impact of Aggregation Method (Average, Weighted Average) on Performance

Pool Method	Acc
Avg Pool-1	59.10
Avg Pool-5	59.53
Weight Avg-1	59.55
Weight Avg-5	60.76

We analyze the impact of the attention significance score within our model's multi-scale mechanism. This score determines which tokens are most important. We compared several methods to calculate this score: random selection, summing weights of self-attention for all tokens, and using the attention weights specifically for recognition. We found that using the attention weights yielded the best results (as shown in Table 3). This approach identifies tokens significantly influencing the final classification because classification tokens directly contribute to category selection. By focusing on these crucial tokens, the model becomes more sensitive to key features, ultimately enhancing its performance.

In our ablation study for efficient token selection within the model, we evaluated three approaches: Top-k subsampling, inverse transform sampling, and a novel combined method utilizing the mixed granularity layer. Top-k subsampling, while fast, discards potentially valuable tokens

Table 3. Ablation study on Attention Significance Score Strategies and their Influence on Model Accuracy

Scoring Strategy	Acc
Random	52.63
Self-Attention	56.30
Proposed	60.76

and lacks flexibility in later stages, restricting overall performance. Conversely, inverse transform sampling retains a broader range of information, benefiting lower layers and the final classification, but may not be the most computationally efficient. To address this trade-off, we propose a combined method. It leverages the mixed granularity layer to replace discarded tokens from Top-k with more efficient computational units. This ingenious approach, validated through ablation studies (refer to dedicated section), allows us to reduce computation without sacrificing significant information, ultimately enhancing model efficiency.

4.4. Discussion and Comparison

This section analyzes the performance of our model on the Okutama action recognition dataset (Table 3). We begin by evaluating a baseline model using an MViT-2 module [20] to capture spatial features without the proposed multi-granularity module. This achieves 48.34 % accuracy, exceeding previous baseline results on Okutama. Introducing the multi-granularity module significantly improves performance. By feeding the network "fused" frames instead of originals, the model achieves 60.76 % accuracy, a 25.26 % increase over the previous state-of-the-art and an 11.42 % improvement from the baseline. This suggests the module effectively focuses on action regions while disregarding

Table 4. Significant Performance Gains on Okutama Dataset [4]: Comparison with Prior Work

Method	Year	Backbone	Accuracy (%)
AARN [34]	2019	C-RPN + YOLOv3-tiny	33.75
Lite ECO [39]	2018	BN-Inception + 3D-Resnet-18	36.25
13D(RGB) [7]	2017	3D CNN backbone	38.12
3DCapsNet-DR [36]	2021	3D CNN + Capsule	39.37
3DCapsNet-EM [36]	2021	3D CNN + Capsule	41.87
DroneCaps [1]	2022	3D CNN + BVC + Capsule	47.50
Ours Baseline	2024	**MViTv2**	**48.34**
Ours (HAT)	2024	**MViTv2**	**60.76**

irrelevant background information.

Our revolutionary approach achieves 60.76 % accuracy on Okutama, surpassing existing methods. The granularity attention module with scoring and inverse transformation allows the network to concentrate on action areas and ignore background noise, unlike previous 3D CNN-based methods [1]. Additionally, our plug-and-play design offers superior computational efficiency compared to prior works. This efficient and accurate model, capable of distinguishing similar classes, paves the way for real-time applications involving multi-label action recognition with recurrent tube processing. A visual illustration of the result is shown in Fig 4.

While our model significantly reduces computation and improves recognition compared to existing methods, there's room for further development. Firstly, for challenging datasets like Okutama, there's still potential to improve recognition accuracy. Secondly, the current model focuses solely on action recognition and could be adapted to broader video understanding tasks. Finally, incorporating additional performance metrics would provide a more comprehensive evaluation. Our future work will address these limitations. We aim to refine the model for even higher accuracy on demanding datasets. Additionally, we plan to explore the application of our proposed modules to diverse video understanding domains.

5. Conclusion

This work addresses the challenge of recurrent action detection and recognition in aerial surveillance videos, particularly for sparsely populated areas where public safety is a concern. We propose a novel multi-label framework that offers several advantages. Firstly, it allows for flexible replacement of detection and tracking modules based on specific needs. This enables training and inference of all modules on a single system, making it more adaptable than existing solutions for multi-label action detection in aerial videos. Furthermore, our framework tackles the crucial challenge of improving efficiency in multi-label action recognition by employing a multi-level refinement strategy that integrates multi-scale and multi-granularity mechanisms.

This strategy leverages attention and token-based approaches to optimize performance in real-world scenarios. By introducing multi-granularity selection on top of the multi-scale approach, we shorten the computational length of the sequence effectively, leading to increased efficiency for action recognition tasks. Our proposed method has been extensively tested, and the results of our experiments demonstrate its effectiveness. We envision its future development for a wider range of application scenarios, where a balance between retaining feature information and reducing computational costs remains paramount.

6. Acknowledgments

This work is part of the project "Intelligent Object Detection, Dynamic Scene and Activity Recognition for Real-Time UAV Applications," developed at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) and funded by the Technology Innovation Institute (TII), Abu Dhabi, UAE. Furthermore, we acknowledge the invaluable contribution of artificial intelligence (AI) tools in enhancing the efficiency and advancements in our research endeavors.

References

[1] Abdullah M Algamdi, Victor Sanchez, and Chang-Tsun Li. Dronecaps: recognition of human actions in drone videos using capsule networks with binary volume comparisons. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3174–3178. IEEE, 2022. 8

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, Montreal, BC, Canada, October 2021. 3

[3] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem. Finding tiny faces in the wild with generative adversarial network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 21–30, June 2018. 2

[4] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger. Okutama-action:

An aerial view video dataset for concurrent human action detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, pages 28–35, July 2017. 2, 5, 6, 8

[5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv*, 2020. 3

[6] Gedas Bertasius, Hanwen Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the ICML Conference*, volume 2, page 4, Virtual, July 2021. 3

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2019. 8

[8] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 379–387, 2016. 2

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 2

[10] A. Van Etten. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv*, 2018. 2

[11] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yeqing Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, Montreal, BC, Canada, October 2021. 3

[12] Muhammad Fayyaz, Seyed Amir Hossein Koohpayegani, Faezeh Riahi Jafari, Suraj Sengupta, Hamid Reza Vaez Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers. In *Proceedings of the European Conference on Computer Vision*, pages 396–414, Tel Aviv, Israel, October 2022. Springer. 3

[13] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. A better baseline for AVA. *arXiv*, 2018. 2

[14] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6047–6056, June 2018. 1

[15] P. Hu and D. Ramanan. Finding tiny faces. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 951–959, July 2017. 2

[16] P. C. Hytla, K. S. Jackovitz, E. J. Balster, J. R. Vasquez, and M. L. Talbert. Detection and tracking performance with compressed wide area motion imagery. In *Proc. IEEE Nat. Aerosp. Electron. Conf. (NAECON)*, pages 163–170, July 2012. 2

[17] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *Proc. ICCV*, pages 4405–4413, Oct. 2017. 2

[18] Dmytro Kondratyuk, Li Yuan, Yeqing Li, Lei Zhang, Mingxing Tan, Tara Brown, and Bo Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, Nashville, TN, USA, June 2021. 3

[19] R. LaLonde, D. Zhang, and M. Shah. ClusterNet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Proc. IEEE Comput. Vis. Pattern Recognit.*, pages 4003–4012, June 2018. 2

[20] Yeqing Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MVITv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, New Orleans, LA, USA, June 2022. 3, 5, 7

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. to be published. 2

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot MultiBox detector. In *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, pages 21–37, Amsterdam, The Netherlands, Oct. 2016. Springer. 2

[23] Zitao Liu, Yutong Lin, Yuting Cao, Han Hu, Yixuan Wei, Zhaohui Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, Montreal, BC, Canada, October 2021. 3

[24] Zitao Liu, Junjie Ning, Yuting Cao, Yixuan Wei, Zhaohui Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition*, pages 3202–3211, New Orleans, LA, USA, June 2022. 3

[25] Dan Neimark, Omri Bar, Matan Zohar, and Dirk Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, Montreal, BC, Canada, October 2021. 3

[26] R. Porter, A. M. Fraser, and D. Hush. Wide-area motion imagery. *IEEE Signal Process. Mag.*, 27(5):56–65, Sept. 2010. 2

[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[28] J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. *arXiv*, 2018. 2

[29] Dan Shi, Yizhou Zhong, Qi Cao, Lin Ma, Jianqiang Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, Vancouver, BC, Canada, June 2023. 3

[30] Dan Shi, Yizhou Zhong, Qi Cao, Junge Zhang, Lin Ma, Jianqiang Li, and Dacheng Tao. React: Temporal action detection with relational queries. In *Proceedings of the European Conference on Computer Vision*, pages 105–121, Tel Aviv, Israel, October 2022. Springer. 3

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 30, 2017. 2

[32] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, 2018. 3

[33] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and real-time tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3

[34] Fan Yang, Sakriani Sakti, Yang Wu, and Satoshi Nakamura. A framework for knowing who is doing what in aerial surveillance videos. *IEEE Access*, 7:93315–93325, 2022. 8

[35] Chuan Lin Zhang, Jie Wu, and Yiran Li. Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 492–510, Tel Aviv, Israel, October 2022. Springer. 3

[36] Ping ZHang, Ping Wei, and SHuhuan Han. Capsnets algorithm. In *Journal of physics: conference series*, page 012030. IOP Publishing, 2021. 8

[37] Xinyi Zhang, Peng Li, and Hao Li. Ambert: A pre-trained language model with multi-grained tokenization. *arXiv*, 2020. 3

[38] Jia Zhao, Yuchen Wang, Jianmin Bao, Yuxi Wu, and Xiaodong He. Fine-and coarse-granularity hybrid self-attention for efficient bert. *arXiv*, 2022. 3

[39] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2020. 8