Neural Networks 185 (2025) 107182

Contents lists available at ScienceDirect

Neural Networks

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

Full Length Article

Paying more attention on backgrounds: Background-centric attention for UAV detection

Xiuxiu Lin [a], Yusu Niu [a], Xinran Yu [a], Zhun Fan [b], Jiafan Zhuang [a], An-Min Zou [a,c,*]

[a] College of Engineering, Shantou University, Shantou, 515063, China
[b] University of Electronic Science and Technology of China, Chengdu, 611731, China
[c] Key Laboratory of Intelligent Manufacturing Technology, Ministry of Education, Shantou University, Shantou, 515063, China

ARTICLE INFO

Keywords:
Unmanned Aerial Vehicle
Background-centric Attention Module
Object detection

ABSTRACT

Under the advancement of artificial intelligence, Unmanned Aerial Vehicles (UAVs) exhibit efficient flexibility in military reconnaissance, traffic monitoring, and crop analysis. However, the UAV detection faces unique challenges due to the UAV's small size in images, high flight speeds, and limited computational resources. This paper introduces a novel Background-centric Attention Module (BAM) to address these challenges. Unlike traditional methods relying on UAV visual features, the BAM utilizes complex background information to identify UAV presence. The BAM seamlessly integrates into existing UAV detection frameworks, improving accuracy with no significant increase in the computation time. Extensive experiments on challenging datasets, Naval Postgraduate School Drones (NPS), and Flying drones (FLDrones) using mainstream detectors YOLOv5 and TphPlus demonstrate the effectiveness of the BAM in significantly enhancing detection accuracy. This research emphasizes the importance of background information in the UAV detection and proposes a method aligning with human perceptual processes, paving the way for further advancements in the field.

1. Introduction

Recently, Unmanned Aerial Vehicles (UAVs) have been used in many fields, *e.g.*, military reconnaissance (Chen, Du, Zhang, Han, & Wei, 2022; Shumeye Lakew, Sa'ad, Dao, Na, & Cho, 2020; Xiao et al., 2022), civil surveillance (Asadzadeh, de Oliveira, & de Souza Filho, 2022; Mehta, Gupta, & Tanwar, 2020; Tsao, Girdler, & Vassilakis, 2022), disaster response (Pan, Chen, Yin, & Huang, 2022; Wan, Zhong, Ma and Zhang, 2023; Yang et al., 2022), environmental monitoring (McCabe et al., 2017; Mohamed, Al-Jaroodi, Jawhar, Idries, & Mohammed, 2020; Román et al., 2022), *etc.* The trend in UAV development is towards greater intelligence and swarm collaboration. To achieve an effective collaboration of swarm UAVs and avoid collisions, an accurate and reliable detection of surrounding UAVs plays an important role.

Different from the conventional detection like in the Visual Object Classes (VOC) dataset (Everingham, 2008) and the Microsoft Common Objects in Context (MS-COCO) dataset (Lin et al., 2014), the UAV detection has three main challenges. Firstly, UAVs typically occupy only a tiny portion of the image. For instance, foreground objects in the MS-COCO dataset occupy nearly 20% of the image area, whereas UAVs account for just 0.05% in the Naval Postgraduate School Drones (NPS) dataset (Li et al., 2016). Such a small target size classifies drone detection as a typical small-object detection task (Cheng et al., 2023; Leng et al., 2023). Small objects often lack distinctive visual features, a challenge that is particularly evident in drone detection, especially under complex scenarios (Li et al., 2016; Rozantsev, Lepetit, & Fua, 2017).

Secondly, UAVs commonly have high flight speed. Therefore, the UAV detection needs to be real-time, otherwise, the target would be out of sight in the next frame. Thirdly, the platform on UAVs has limited computing resources, which requires the detector to be lightweight and low-cost. In general, the difficult-to-detect characteristics and real-time processing requirement make the UAV detection unique and challenging.

Based on successful practices on the conventional detection (He, Zhang, Ren, & Sun, 2016; Redmon, Divvala, Girshick, & Farhadi, 2015; Ren, He, Girshick, & Sun, 2017), existing UAV detection methods can be roughly divided into two families. The first group of works (Tang et al., 2023; Wang et al., 2022; Xie, Yu, Wu, Shi, & Chen, 2020) focus on discovering and aggregating visual information of UAVs to enhance the feature discriminatively. The second group of works (Ashraf, Sultani, & Shah, 2021; Rozantsev, Lepetit, & Fua, 2015; Sangam, Dave, Sultani, & Shah, 2023) focus on exploiting motion information of UAVs, which can provide extra clues for detection. In general, existing works follow

* Correspondence to: 243 Daxue Road, Shantou 515063, Guangdong Prov., China.
E-mail address: amzou@stu.edu.cn (A.-M. Zou).

https://doi.org/10.1016/j.neunet.2025.107182
Received 6 July 2024; Received in revised form 4 December 2024; Accepted 14 January 2025
Available online 21 January 2025

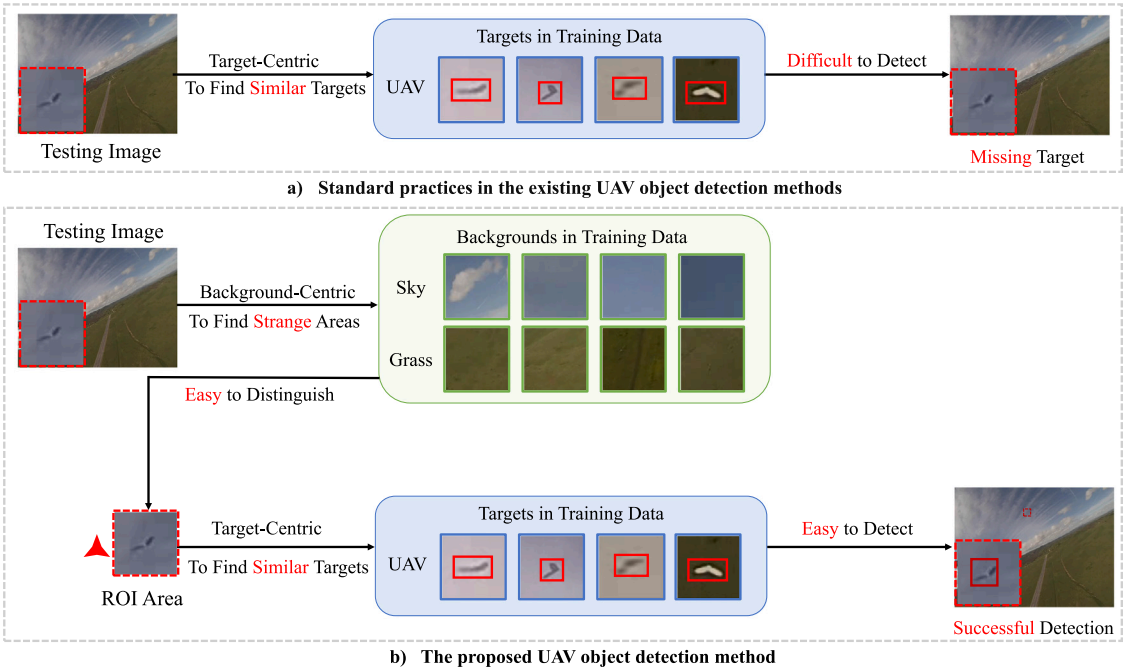

Fig. 1. Comparison of two different UAV detection perspectives. Previous works follow a target-centric procedure to find similar objects according to seen targets in training data, which is difficult since UAVs are usually tiny and lack of distinct visual information. Differently, in this work, we propose a novel background-centric procedure to discover strange areas-Region of Interest (ROI) areas according to typical backgrounds in training data, which is relatively easier to distinguish.

a target-centric procedure, which is inherited from the conventional detection. As illustrated in Fig. 1, the detector typically learns discriminative UAV representations from training data and then searches for similar targets in the tested image. However, due to the tiny size of UAVs and the lack of distinct visual information, the detector cannot obtain sufficient supervision signals for representation learning, which leads to error-prone UAV representations and thus failed detection during testing.

Essentially, an image for the UAV detection typically consists of two key components, i.e., UAVs and backgrounds. Then a question naturally arises: *since UAVs are too tiny to detect, can we seek clues from their surrounding backgrounds?* The answer is yes and we found that it is more consistent with the human-like recognition procedure, which tends to be background-centric (Collins & McDougle, 2021; Heald, Lengyel, & Wolpert, 2021; Tran, Vu, Vo, Nguyen, & Nguyen, 2022; Xie et al., 2024). 'Background-centric' can be concretized according to different circumstances. For example, in the human world, 'Background' can be expressed as 'Context', and the research by Collins and McDougle (2021) and Heald et al. (2021) has clarified the critical role of context in human learning and the execution of motor skills. In industrial anomaly detection, it should be defined as the 'Nominal Example Images', learning feature representations from normal data to detect deviations and anomalies is a widely recognized and successfully applied approach (Tran et al., 2022; Xie et al., 2024). Similarly, in the context of UAV background detection tasks: When faced with an image, humans typically do not carefully search for targets based on the UAV appearance information stored in their memory. In contrast, they first observe whether there are any abnormal areas in the background. If so, they focus on the specific region for further UAV recognition. This background-centric recognition process essentially shares a similar motivation with anomaly detection. As illustrated in Fig. 1, in the tested image, after discovering an unusual black spot in a clear sky, a human would focus on the corresponding areas for further UAV detection. Therefore, backgrounds can provide an important guidance for the UAV detection, which has been ignored in previous studies. Besides, contrary to tiny UAVs, backgrounds occupy most of the areas in the image and thus are sufficient for effective and reliable representation learning. In general, this background-centric motivation provides a new thought for the UAV detection.

Therefore, in this work, we follow the human-like recognition procedure and design a novel **B**ackground-centric **A**ttention **M**odule (BAM) for the UAV detection. Specifically, the BAM can effectively model different background regions in the image and then discover abnormal areas, which provides an important guidance for detecting potential UAVs. The key to the BAM is to effectively represent background regions, especially in complex scenarios. Considering that there are usually multiple background elements present in an image, e.g., sky, grass, etc., we calculate feature prototypes as background representations after feature clustering. It is noteworthy that we discard clusters with fewer pixels and only reserve top-k clusters, which essentially utilizes the characteristics of the UAV detection that background areas usually occupy large portions of an image. Based on the constructed background representations, the BAM can calculate a spatial attention map to highlight the dissimilar regions, which commonly have a high probability of containing the target UAV. To aid applicability, the BAM is designed as a lightweight and plug-and-play module, which can be easily inserted into existing UAV detection frameworks.

We implement the BAM on two mainstream detectors, i.e., YOLOv5 (Jocher et al., 2021) and TphPlus (Zhao, Liu, Lyu, Wang, & Zhang, 2023), and conduct evaluation on two challenging benchmarks, i.e., NPS (Li et al., 2016) and Flying drones (FLDrones) (Rozantsev et al., 2017). The results show that the BAM can bring a significant accuracy improvement and is suitable to be deployed in real-time UAV detection tasks in practical situations, which validates its effectiveness.

The main contributions of this work can be summarized as follows:

1. We find that background information can provide an important guidance for the UAV detection, which is consistent with the human-like recognition procedure but usually ignored in previous studies.
2. We design a novel BAM to effectively model background regions of the image and discover dissimilar areas as potential targets. Besides, the BAM is designed in a plug-and-play style and can be easily implemented in existing detectors.

3. We implement the BAM on two mainstream detectors and evaluate it on two representative and challenging benchmarks. The results demonstrate its effectiveness and generalizability.

2. Related work

2.1. UAV detection

Due to the UAV's small size and fast-moving characteristics, the UAV detection is a challenging task. The UAV detection has been widely studied, and existing works can be roughly divided into two groups, i.e., appearance-based (Tang et al., 2023; Xie et al., 2020) and motion-based methods (Ashraf et al., 2021; Rozantsev et al., 2015; Sangam et al., 2023). Appearance information, *e.g.*, shape, color, and texture, is crucial for the UAV detection. Adaptive switching spatial–temporal feature maps (ASSTFM) (Xie et al., 2020) was proposed to aggregate object appearance information by designing a spatial feature map, which can improve the visibility and discrimination of UAVs. Tang et al. (2023) selected multiple state-of-the-art detectors and optimized each detector for the unique appearance features of UAV objects, which can achieve an ensembled performance while introducing high computational cost. To supplement appearance information, researchers also worked on exploiting motion information, which is crucial, especially for fast-moving UAVs. Rozantsev et al. (2015) proposed to construct motion-stabilized spatiotemporal cubes to highlight moving objects. However, its reliance on a perfect UAV-centered cube limits its performance in practice, especially for tiny targets with structural distortions. DogFight (Ashraf et al., 2021) was proposed to use a two-stage segmentation-based approach employing spatiotemporal attention cues. However, due to Dogfight's reliance on non-parallelized connected component analysis and tracking on the CPU, it experiences lower frame rates and throughput. Additionally, the involvement of non-differentiable components hinders its ability to function as a fully end-to-end model. TransVisdrone (Sangam et al., 2023) enhances feature integration by utilizing a temporal transformer to fuse features from multiple frames, implicitly encoding motion information. Nevertheless, TransVisDrone's adoption of a hybrid spatio-temporal transformer architecture, integrating both the CSPDarknet53 and VideoSwin models, introduces its own set of challenges. This complex structure not only amplifies the difficulty of training the model but also necessitates more expensive computational resources, presenting significant barriers to accessibility and scalability.

Although existing methods attempt to overcome the challenges of the UAV detection, they usually follow the object-centric paradigm, which still suffer from the extremely small size and complex motion of UAVs. In this work, we focus on the commonly ignored background information and attempt to find important clues for discovering potential UAVs.

2.2. Attention mechanism

In the field of the object detection, the attention mechanism is a key technique that draws inspiration from the way human vision allocates attention, focusing on important information within images to improve detection accuracy and efficiency. For example, Squeeze-and-Excitation Network (SENet) (Hu, Shen, Albanie, Sun, & Wu, 2018) first introduces a channel attention mechanism that effectively aggregates global information. The Convolutional Block Attention Module (CBAM) (Woo, Park, Lee, & Kweon, 2018) takes a step further by combining both spatial and channel attention simultaneously. Attention CoupleNet (Zhu et al., 2019) introduces a fully convolutional coupling structure that can integrate global and local information of an object to enhance its representation. The Co-Attention mechanism (Hsieh, Lo, Chen, & Liu, 2019), by establishing a shared attention relationship between examplar images and query images, enhances feature learning and achieves high-precision detection of novel objects. Similarly, PrIme Sample Attention (PISA) (Cao, Chen, Loy, & Lin, 2020) emphasizes different treatments based on the sample importance, which would face challenges in recognizing tiny UAVs that are obviously not the "main" samples in scenes. Neural Attention Learning (NEAL) (Ge, Song, Ma, Qi, & Luo, 2023) generates attention response maps to guide the network towards features that significantly impact prediction outcomes. Overall, despite the attention-based methods (Cao et al., 2020; Ge et al., 2023; Hsieh et al., 2019; Hu et al., 2018; Woo et al., 2018; Zhu et al., 2019) achieve promising performance in the conventional detection, they mostly focus on aggregating information about important and salient objects, which commonly occupy substantial regions in the image. However, the aforementioned attention-based methods are not suitable for the UAV detection since tiny UAVs in a complex scenarios can only occupies negligible regions and easily be concealed by background signals. Different from previous works, we propose a background-centric attention mechanism and focus on modeling background information to discover potential targets, which is more suitable for the UAV detection.

3. Methods

3.1. Overview

In this section, we introduce how to utilize the BAM in object detection frameworks, as illustrated in Fig. 2, aimed at optimizing feature map analysis during the object detection process. The BAM precisely focuses on the background information with stronger supervisory signals in the input feature maps, indirectly indicating the potential positions of UAVs. The module consists of two main parts: background prototype generation and potential target region search. Firstly, background prototypes are extracted from the input feature maps using feature clustering techniques, while ignoring the least representative clusters to reduce noise interference. Subsequently, the similarity between background prototypes and input feature maps is computed to indirectly identify potential target regions. This approach not only enhances object detection accuracy in complex backgrounds, but also offers a novel perspective and technical pathway for the object detection in computer vision.

3.2. Background prototype generation

In drone target detection scenarios, the Background Prototype Generation component serves as the core of the BAM, playing a vital role in distinguishing between background features and potential target features within an image. This process starts with the application of a clustering algorithm to the feature map of the given image, denoted by $F \in \mathbb{R}^{h\times w\times d}$, where h and w represent the height and width of the feature map, respectively, and d indicates the feature dimensionality at each pixel. For each pixel located at position (i, j), where $i \in \{1, 2, \ldots, h\}$ and $j \in \{1, 2, \ldots, w\}$, the feature vector is denoted by F_{ij}, emphasizing the extraction of a d-dimensional feature vector from F. We aim to better represent and differentiate between background and target foreground areas through feature learning on the feature map.

By applying a clustering algorithm on the feature map at each pixel, it is assumed that similarities among features representing the same category will be discovered, manifesting in the form of clustering. The centroids obtained from the clustering algorithm are considered representative prototypes of their respective categories, expressed as $C = \{C_1, C_2, \ldots, C_K\}$, where $C_l \in \mathbb{R}^d$, $l \in \{1, 2, \ldots, K\}$ is the representative prototype of the l category, and K represents the total number of categories.

The category representation for each centroid C_l can be defined as follows:

$$C_l = \begin{cases} \text{Background Prototype}, & \text{if } l \in \{1, 2, \ldots, K-1\} \\ \text{Minimal Clustering}, & \text{if } l = K \end{cases} \tag{1}$$

When performing K-means clustering, which refers to the total number of clusters, it is expected that at least one cluster will represent the

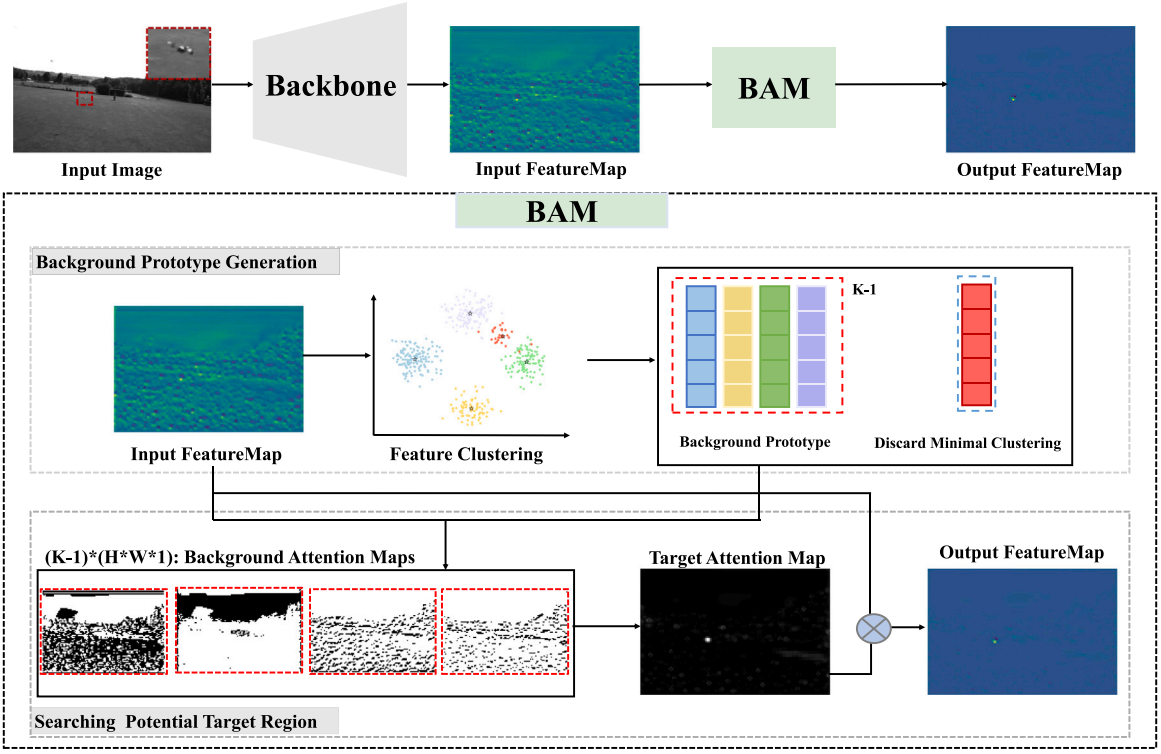

Fig. 2. Illustration of our BAM. Our BAM module is designed to highlight anomalous regions in the original feature map that differ from the typical background in the training data. Our approach first clusters the features of the input feature maps, extracts the typical background prototype, and removes the least salient clusters. Then, the similarity is computed by background prototype with the original feature maps, and after obtaining a number of background attention maps, they are compared with the original feature maps to obtain the anomalous regions, which constitute our target attention maps. Finally, we fuse the target attention maps with the original feature maps.

characteristics of the drone, while the other clusters will represent the typical features of the remaining background. As illustrated in Fig. 2, in the specific scenario of the drone detection, the background area occupies the majority of the image, so we can consider that the larger clusters, i.e., the clusters $\{1, 2, \ldots, K-1\}$, actually represent typical background prototypes. The strategy of excluding the minimal cluster (i.e., the Kth cluster) aims to ignore the least representative prototypes, which are typically associated with noises or minor elements (e.g., the drones) that are not characteristics of the main background. By focusing on the larger, more representative background clusters, we aim to better locate the target drone area, thus improving the detection of drones in complex backgrounds.

3.3. Searching for potential target regions

Building upon the Background Prototype Generation, the subsequent step in the BAM workflow is to identify and extract characteristics background features, aiding in the distinction between background and potential foreground target regions. By leveraging the predefined pixel positions (i, j), where $i \in \{1, 2, \ldots, h\}$ and $j \in \{1, 2, \ldots, w\}$, we compute the similarity score $S[i, j]$ as the maximum cosine similarity with the $K-1$ background clusters:

$$S[i,j] = \max\left(\left\{\frac{C_l \cdot F_{ij}}{\|C_l\| \cdot \|F_{ij}\|}\right\}_{l=1}^{K-1}\right), \quad i \in \{1,2,\ldots,h\}, \quad j \in \{1,2,\ldots,w\}. \tag{2}$$

This calculation enables the model to identify regions corresponding to the background within F. Learning these typical background features allows the model to concentrate on anomalous areas, the non-background regions that are potential target sites. The attention map $AttentionMap \in \mathbb{R}^{h \times w}$ is then constructed by taking the complement of the similarity scores:

$$\text{AttentionMap}[i,j] = 1 - S[i,j], \quad i \in \{1,2,\ldots,h\}, \quad j \in \{1,2,\ldots,w\}. \tag{3}$$

By computing the similarity scores between F and the background prototypes, corresponding to the $K-1$ clusters, we can infer that higher similarity scores indicate a greater likelihood of the entire region being the background (Eq. (2)). By calculating the regions with high similarity between the original feature map and the background representation, we identify the background areas. Our goal is to locate these background areas and then focus on the regions outside of them, as these are the potential foreground regions. As described in Eq. (3), we invert the attention map: the background regions with initially high similarity are transformed into low-weight regions after inversion. Similarly, the regions with high attention map values correspond to areas with low background similarity, which are more likely to contain targets.

3.4. Pseudocode

In this subsection, a detailed explanation of the pseudo-code for the BAM is provided, as shown in Algorithm 1. The BAM aims to enhance the focus on background information in the task of object detection, thereby improving detection performance. Initially, we subject the input feature map F to clustering algorithms to identify clustering patterns of background information automatically. Subsequently, based on the similarity between each pixel and the centroids of background clusters, we generate the background attention map $AttentionMap$. Specifically, we input the feature map F into the clustering algorithm to obtain the cluster centroids C, as well as the sizes of each cluster, which are denoted by $cluster\_sizes$. Then, we sort the cluster centroids C and $cluster\_sizes$ in descending order of cluster sizes and select the top $K-1$ clusters as background clusters C. Next, we traverse the entire feature map, compute the similarity between each pixel and the centroids of the background clusters C_l, and then aggregate these similarities using aggregation functions, followed by taking the negation to obtain the

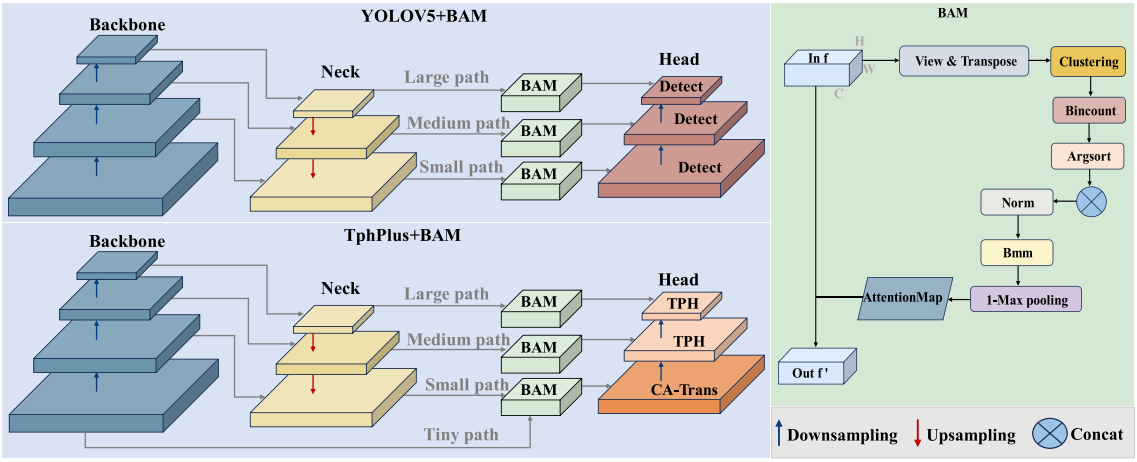

Fig. 3. Details of the incorporation of BAM. BAM acts as a plug and play attention in the detection network that is inserted before the detection head. The difference between YOLOV5 and TphPlus is the different design of the detector head, and the fact that Tphplus has designed a cross-layer asymmetric transformer (CA-Trans) to replace the additional tiny detector head. BAM network architecture first needs to input the original feature map, the attention map obtained after processing is weighted with the original feature map and fed into the detection head.

attention map for potential target regions $AttentionMap$.

We choose to employ the K-means clustering algorithm within our framework. However, although K-means clustering serves as our baseline clustering method, it does not imply it as the ultimate choice among all clustering techniques. Instead, our decision highlights the flexibility of this approach, allowing researchers to adjust their clustering algorithms according to the characteristics of their datasets. Furthermore, based on our experimental results, we determine the optimal value of the number of clusters K to be 5. We encourage researchers to adjust the top-K parameter based on the specifics of their datasets, as this customization does not compromise the efficacy or utility of the proposed method.

Algorithm 1 Background-Centric Attention Module

Input: $F \in \mathbb{R}^{h\times w\times d}$
Output: $TargetAttentionMap \in \mathbb{R}^{h\times w}$
$C, cluster\_sizes \leftarrow$ ClusteringAlgorithm(F)
Sort C and $cluster\_sizes$ in descending order based on $cluster\_sizes$
Let K be the total number of clusters, i.e., the length of C
Initialize $S \in \mathbb{R}^{h\times w}$
for $i = 1$ to h **do**
  for $j = 1$ to w **do**
    Initialize $max\_similarity \leftarrow 0$
    for $l = 1$ to $K - 1$ **do**
      $similarity \leftarrow \frac{C_l \cdot F_{ij}}{\|C_l\| \cdot \|F_{ij}\|}$
      if $similarity > max\_similarity$ **then**
        $max\_similarity \leftarrow similarity$
      end if
    end for
    $S[i, j] \leftarrow max\_similarity$
  end for
end for
for $i = 1$ to h **do**
  for $j = 1$ to w **do**
    $TargetAttentionMap[i, j] \leftarrow 1 - S[i, j]$
  end for
end for
return $TargetAttentionMap$

3.5. Integration details

The BAM is a plug-and-play attention mechanism that is seamlessly integrated into detection networks, specifically positioned before the detection head. As illustrated in Fig. 3, and further elaborated in the experimental subsection, we will use YOLOv5 and TphPlus as exemplary cases to demonstrate the application of BAM. Unlike YOLOv5, TphPlus distinguishes itself with a uniquely designed detection head and incorporates a Cross-Layer Asymmetric Transformer (CA-Trans) to replace the additional tiny detection head typically found in such architectures. The network architecture of BAM begins with the input of the original feature map. After processing, the resulting attention map is weighted and concatenated with the original feature map, which is then fed into the detection head. This integration not only enhances the feature representation by emphasizing salient background information but also improves the model's ability to distinguish between background and foreground objects, leading to more accurate detection capabilities.

4. Experiments

4.1. Datasets and metrics

Datasets. The NPS Drones dataset (Li et al., 2016) comprises a collection of 50 high-definition video sequences designed to enhance the detection capabilities of aerial aircraft, particularly small drones. These videos encompass a total of 70,250 frames and were captured using a GoPro 3 camera mounted on a custom-made triangular-winged aircraft. The NPS dataset consists of color images with resolutions of either 1920 × 1280 or 1280 × 760. The videos are recorded at a frame rate of 30 frames per second. Targets in the images occupy a maximum area of 6.6×10^{-4} and a minimum area of 8.2×10^{-5}, with an average proportion of only 0.05%. Annotations utilized in this study were derived from the clean version released by Dogfight (Ashraf et al., 2021). The training/validation/testing split follows the methodology of Dogfight (Ashraf et al., 2021) and TransvisDrone (Sangam et al., 2023).

The FLDrones dataset (Rozantsev et al., 2017) consists of footage captured by cameras mounted on flying drones, featuring a mix of indoor and outdoor scenes. This dataset contains 14 videos with a total of 38,948 frames, with grayscale resolutions of 640 × 480 or 752 × 480. Targets in the images occupy a maximum area of 1.4×10^{-1} and a minimum area of 2.6×10^{-4}, with an average proportion of only 0.07%. Annotations utilized in this study were derived from the clean version released by Dogfight (Ashraf et al., 2021) and TransvisDrone (Sangam et al., 2023).

Metrics. We utilize metrics such as average precision (AP), precision, and recall to gauge detection quality, while also considering frames per second (FPS) to measure computational efficiency.

Table 1
Compared with several state-of-the-art methods on the NPS and FLDrones datasets. The best results are highlighted in red, while the second-best results are highlighted in blue. Evaluation metrics include MAP (precision) and FPS (speed).

Methods	Venue	AP-NPS	AP-FLDrones	FPS
Mask-RCNNs (He, Gkioxari, Dollár, & Girshick, 2017)	ICCV'17	0.89	0.68	17.55
SCRDet-H (Yang et al., 2019)	ICCV'19	0.65	0.52	–
SCRDet-R (Yang et al., 2019)	ICCV'19	0.61	0.52	–
FCOS (Tian, Shen, Chen, & He, 2019)	ICCV'19	0.83	0.62	–
SLSA (Wu, Chen, Wang, & Zhang, 2019)	ICCV'19	0.46	0.61	–
MEGA (Chen, Cao, Hu, & Wang, 2020)	CVPR'20	0.83	0.65	–
De-DETR (Zhu et al., 2020)	ICLR'21	0.76	–	10.69
VisTR (Wang et al., 2021)	CVPR'21	0.66	–	1.6
TPH-YOLOv5 (Zhu, Lyu, Wang, & Zhao, 2021)	ICCV'21	0.92	0.69	25
Dogfight (Ashraf et al., 2021)	CVPR'21	0.89	0.72	1.0
TransVisDrone (Sangam et al., 2023)	ICRA'23	0.95	0.75	24.6
CFINet (Yuan, Cheng, Yan, Zeng, & Han, 2023)	ICCV'23	0.90	0.63	19.3
YOLOV8l (Jocher, Qiu, & Chaurasia, 2023)	Github'23	0.93	0.68	27.4
YOLOV9c (Wang, Yeh, & Mark Liao, 2025)	ECCV'25	0.91	0.67	30.8
YOLOV10l (Wang et al., 2024)	ArXiv'24	0.92	0.67	27.9
YOLOV11l (Jocher & Qiu, 2024)	Github'24	0.92	0.63	188
RT-DETR (Zhao et al., 2024)	CVPR'24	0.95	0.66	44.9
YOLOv5l (Jocher et al., 2021)	Github'21	0.93	0.66	46
YOLOv5l+Ours	–	0.94	0.76	34.1
TphPlus (Zhao et al., 2023)	RS'23	0.94	0.71	40.2
TphPlus+Ours	–	0.96	0.79	28.7

4.2. Implementation details

We implement our model in PyTorch. All our models are trained and tested using an NVIDIA RTX3090 GPU. Our baseline model choice is the TphPlus (Zhao et al., 2023) model, specifically designed for ground object detection of small targets, employing consistent hyper-parameters with TphPlus (Zhao et al., 2023). Drawing inspiration from prior works (Ashraf et al., 2021; Sangam et al., 2023), we initiate training of our model by employing pre-trained weights available for YOLOv5l (Jocher et al., 2021) on MS-COCO (Lin et al., 2014). Additionally, as the current UAV dataset does not provide annotations for empty frames - frames devoid of target UAVs, we adhere to the previous methodology. Thus, we evaluate using only frames with provided annotations.

Our input frame size is 1920 × 1280. We adopt the Adam optimizer (Kingma & Ba, 2014), setting momentum to 0.843. To better control the adjustment of learning rates during training, we employ a cosine learning rate scheduler (He et al., 2019), initializing the learning rate to 3×10^{-4}, and decaying it to 0.12 times the initial learning rate in the final epoch of training. The total number of training epochs is set to 80. For the configuration of the non-maximum suppression module, we set the IoU threshold to 0.6 and the confidence threshold to 0.001, consistent with prior studies. Considering GPU memory limitations, we set the batch size to 4. For the clustering algorithm used, we opted for the PyTorch implementation of the K-means clustering algorithm.

4.3. Comparison with state-of-the-art methods

As shown in Table 1, we evaluated various state-of-the-art techniques (Ashraf et al., 2021; Chen et al., 2020; He et al., 2017; Jocher & Qiu, 2024; Jocher et al., 2023, 2021; Sangam et al., 2023; Tian et al., 2019; Wang et al., 2024, 2021, 2025; Wu et al., 2019; Yang et al., 2019; Yuan et al., 2023; Zhao et al., 2023, 2024; Zhu et al., 2021, 2020) on the NPS and FLDrones datasets.

We also compared other advanced models, such as CFINet (Yuan et al., 2023), which has shown excellent performance in small object detection tasks by outperforming mainstream methods on the SODA-D and SODA-A datasets. CFINet achieves this by enhancing detection through coarse-to-fine proposal generation and feature imitation learning. However, in our specific task, it did not achieve the expected performance. We attribute this to the scale and diversity of the dataset. While the SODA datasets are larger and more comprehensive, providing greater generalizability, currently available UAV detection datasets, including ours, are relatively limited in these aspects. This likely prevents CFINet from fully leveraging its strengths.

We further analyzed the YOLO series, including YOLOv5, YOLOv8, YOLOv9, YOLOv10, and YOLOv11 (Jocher & Qiu, 2024; Jocher et al., 2023, 2021; Wang et al., 2024, 2025; Yuan et al., 2023). YOLOv5 adopts a lightweight design and incorporates the CSPNet architecture to optimize computational efficiency, making it widely used in industrial inspection and autonomous driving scenarios and particularly well-suited for real-time deployment. While YOLOv8 introduces dynamic anchoring adjustments to enhance adaptability to targets of varying sizes, this enhancement provides limited benefits in our dataset, which primarily features small and uniformly sized drones. YOLOv9 employs programmable gradient information (PGI) to reduce information loss, but the minimal valid information available from small drone targets reduces the impact of this improvement in our context. YOLOv10 eliminates the need for non-maximum suppression (NMS) in end-to-end reasoning, yet this innovation did not yield noticeable performance gains on our dataset or the RTX 3090 platform. YOLOv11 introduces depth-wise separable convolution, significantly reducing computational load and improving FPS, which we recognize as a meaningful advancement.

Despite the rapid evolution of the YOLO series, we selected YOLOv5 as our baseline model to incorporate and compare our attention mechanisms due to its robust performance and stability across diverse scenarios.

As highlighted in Table 1, our proposed BAM module enhances detection performance in UAV environments. Integrating BAM into YOLOv5 increased AP metrics for the NPS and FLDrones datasets by 1% and 10%, respectively. Similarly, applying BAM to TphPlus resulted in AP metric improvements of 2% and 8% on the same datasets. Although BAM introduces a slight reduction in FPS, the models still achieve real-time performance, with YOLOv5 and TphPlus achieving frame rates of 34.1 and 28.7 FPS, respectively. These results demonstrate that BAM significantly improves UAV object detection accuracy in complex scenarios while maintaining real-time performance.

4.4. Ablation analysis

In this subsection, we compare our attention mechanism with classical attention mechanisms, discuss the design choices regarding attention mechanisms, and report their performances. The comparison is conducted on the FLDrones dataset.

Table 2
Compared with other classical attention mechanisms on the FLDrones dataset.

Methods	AP	Precision	Recall
TphPlus (Zhao et al., 2023)	0.714	0.726	0.691
TphPlus (Zhao et al., 2023) + SE (Hu et al., 2018)	0.716	0.731	0.687
TphPlus (Zhao et al., 2023) + CBAM (Woo et al., 2018)	0.712	0.737	0.683
TphPlus (Zhao et al., 2023) + MLCA (Wan, Lu et al., 2023)	0.711	0.728	0.694
TphPlus (Zhao et al., 2023) + CAA (Cai et al., 2024)	0.630	0.678	0.580
TphPlus (Zhao et al., 2023) + CAFM (Hu, Gao, Zhou, Dong, & Du, 2024)	0.709	0.738	0.705
TphPlus (Zhao et al., 2023) + AFGCAttention (Han et al., 2025)	0.700	0.716	0.667
TphPlus (Zhao et al., 2023) + **Ours**	**0.790**	**0.799**	**0.711**

Table 3
Ablation on minimal and residual clustering for background prototyping.

Minimal clustering	Residual clustering	AP	Precision	Recall
		0.714	0.726	0.691
✓		0.754	0.766	0.671
	✓	0.751	0.772	0.709
✓	✓	0.790	0.799	0.711

Table 4
Impact of the cluster number K on Model AP.

Cluster number	3	5	7	10
AP (%)	77.9	79.0	77.8	77.4

Comparison With Other Attention Mechanisms. According to the results presented in Table 2, we conducted a comprehensive performance comparison of various attention mechanisms integrated with the TphPlus baseline model on the FLDrones dataset. We observed that traditional attention models such as SE and CBAM had a minimal impact on model performance, with SE only increasing the Average Precision (AP) by 0.2%, and CBAM causing a 0.2% decrease in AP. Furthermore, MLCA, CAA, CAFM, and AFGCAttention all led to a decrease in AP, with reductions of 0.3%, 8.4%, 0.5%, and 1.4%, respectively. In contrast, the introduction of BAM significantly boosted AP by 7.6%, and also improved Precision and Recall rates.

These results highlight the performance variance among different attention mechanisms and underscore the importance of selecting an appropriate attention mechanism for specific tasks. Classical attention models, such as SE and CBAM, are primarily designed to identify important and salient foreground objects in images, which are usually more discernible across the entire image. However, this premise does not apply to the detection of drones against complex backgrounds, where the foreground regions of drones are relatively small and often difficult to detect effectively. MLCA and CAFM aim to enhance the model's sensitivity to key features to improve recognition accuracy, but in this task scenario, the foreground features of drones provide very limited effective information. CAA attention is more suitable for multi-task learning scenarios. AFGCAttention, despite its adaptive feature enhancement through graph structures, is insufficient for amplifying the already limited supervisory signals of target features.

In summary, we recognize that simply transferring other attention mechanisms is inadequate for the task of detecting small drones against complex backgrounds. It is imperative to develop an attention mechanism that leverages the robustness of background information and guides the model to focus on the differences between targets and backgrounds. This realization led us to propose the BAM attention mechanism, which harnesses the robustness of background information to guide the detection and recognition of drones by highlighting the disparities between targets and backgrounds. This approach not only addresses the specific challenges of drone detection but also enhances the model's capability to detect small targets within complex environments.

Impact of Minimal and Residual Clustering on Background Prototyping. As described in the methods section, we have a strategy for selecting the generated feature clusters. When handling the background prototype for small UAV detection, we proposed a strategy of discarding the minimal cluster and using the remaining clusters as the background prototype. To validate the effectiveness of this strategy, we conducted ablation experiments to evaluate the impact of different clustering strategies on background and small UAV recognition. The experimental results are shown in Table 3.

The rows in the table represent four different background prototype selection strategies: the first row represents the baseline performance with no clustering method, where the original data is used directly for small UAV detection without any background prototype selection. The second row refers to the background prototype based on the minimal cluster, meaning the class with the fewest feature points is selected from all clusters. This strategy may introduce more background noise, affecting the accuracy of small UAV detection. The third row denotes the background prototype based on all clusters without discarding any clusters, which may lead to confusion between small UAV and background features, further hindering small UAV recognition. The last row represents our proposed optimal strategy, where the minimal cluster is discarded during training and only the remaining clusters are used as the background prototype. This strategy effectively avoids the background interference from the minimal cluster and helps the model better distinguish between background and small UAVs.

The experimental results show that our strategy (the last row) outperforms the other strategies in terms of AP, Precision, and Recall, particularly excelling in small UAVs detection. Specifically, discarding the minimal cluster significantly enhances the distinction between background and small UAV, allowing the model to more accurately identify small UAVs. In contrast, other strategies (especially using the minimal cluster or all clusters as the background prototype) introduce more background noises, reducing small UAV detection performance. Therefore, the ablation experiment validates the effectiveness of our proposed background prototype selection strategy.

The Number of Clusters K Chosen. Compared to the innovative BAM proposed in this paper, employing K-means clustering is merely a baseline choice, not the core of our innovation. We advocate for the exploration of clustering algorithms better suited to specific work scenarios. Consequently, we have conducted a series of ablation studies on the selection of the cluster number K, testing various configurations at $K = 3$, $K = 5$, $K = 7$, and $K = 10$. As shown in Table 4, the experimental results indicate that when adjusting the number of clusters and the selected background feature quantity, the MAP of the model exhibits only slight fluctuations. The highest value of 79% and the lowest value of 77.4% represent the current state-of-the-art level compared to the model without BAM, with a difference of only 1.6% between them.

4.5. Edge deployment: NVIDIA Jetson Xavier

To ensure that our drone target detection model performs excellently in real-world applications, we conducted a series of tests on the NVIDIA Jetson Xavier NX hardware. Our Jetson Xavier NX is equipped with 7505MB of video memory and a 6-core CPU, capable of running our model at 640 resolution in 10PW power mode while maintaining a smooth output of **21 FPS**. Notably, throughout the entire testing process, we did not utilize TensorRT optimization, and the device's real-time operating temperature was consistently maintained below 25 °C.

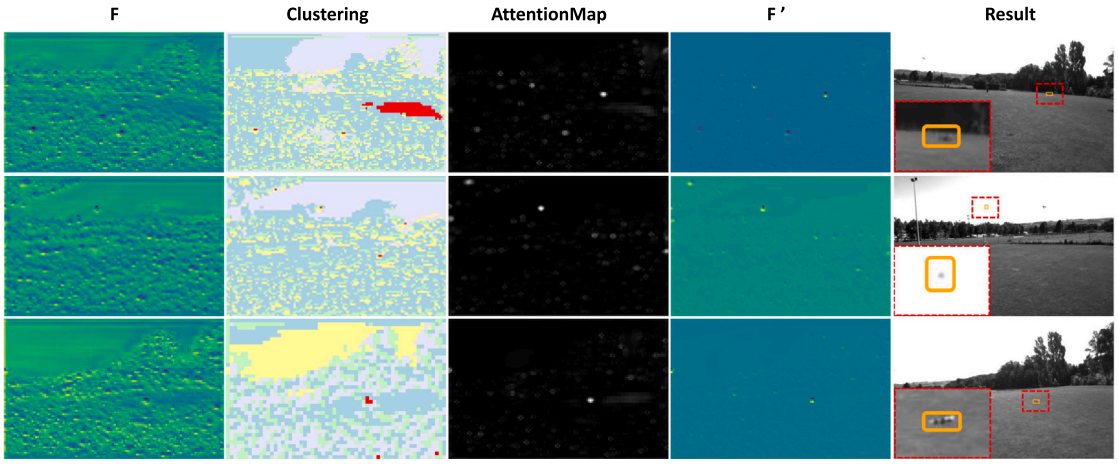

Fig. 4. Visualization of the feature refinement process. From left to right: F - original feature map, Clustering - clustering result, Attention Map - target attention map, F′ - refined feature map obtained by integrating the attention mechanism with the original feature map, and Result - final detection result.

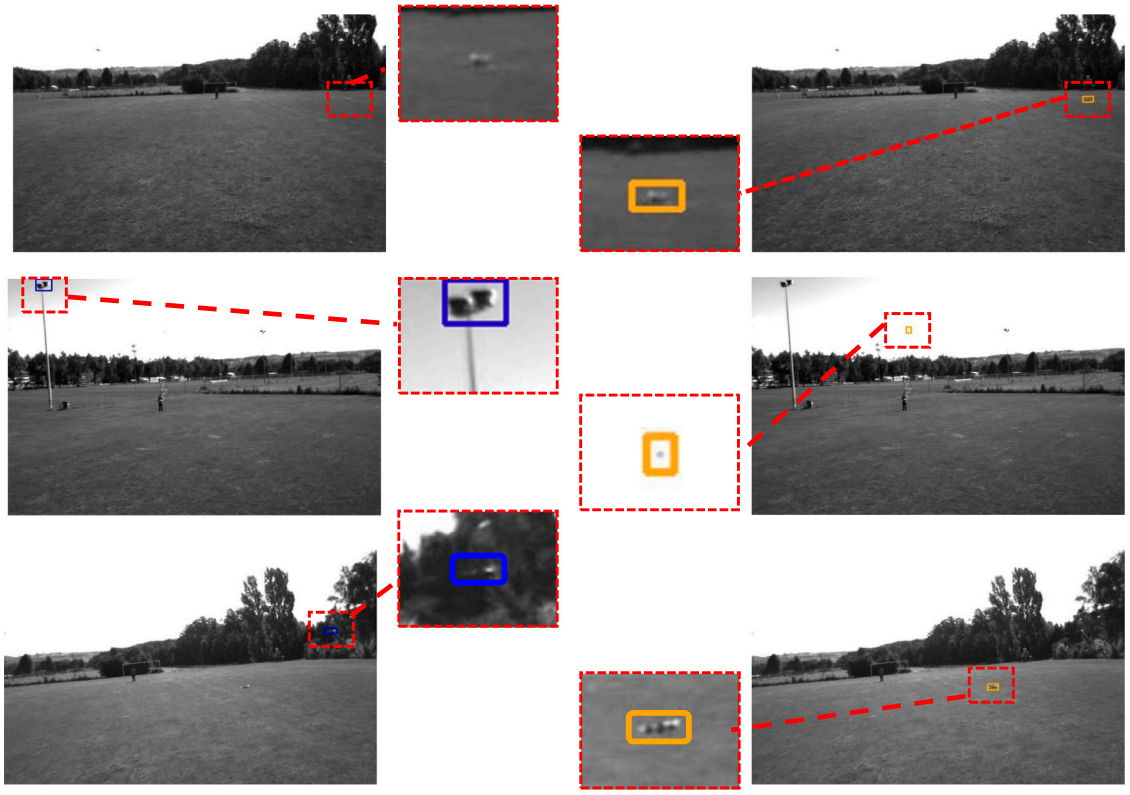

Fig. 5. Comparison of detection results with and without BAM. Visualization of the original baseline (left) and the baseline enhanced with BAM (right).

4.6. Visualization

In this subsection, we present the visualization of our BAM, which plays a pivotal role in enhancing the detection of small targets like UAVs amidst complex backgrounds.

As shown in Fig. 4, the feature refinement process of BAM is visualized step by step. From left to right, it displays the original input feature map (F), the background clustering map generated through clustering (Clustering), the background-based attention map (Attention Map), the optimized feature map fused with attention (F′), and the final detection result (Result). The attention map highlights areas that significantly differ from the background, guiding the model to focus on potential target locations, ultimately generating the optimized feature map (F′) and significantly improving detection performance. Fig. 5 presents a comparative analysis of detection results with and without BAM. With BAM, the model can more effectively distinguish between small targets and the background, significantly reducing the occurrences of missed and false detections. Specifically, the incorporation of BAM allows the model to focus on anomalous areas within the background, leading to more accurate target localization.

The visual analysis above demonstrates that the introduction of BAM not only enhances detection accuracy but also strengthens the model's robustness towards small targets in complex scenarios. This approach emphasizes the crucial role of background information in UAV detection, aligning more closely with human cognitive processes compared to traditional target-centric methods.

5. Conclusion and future work

Inspired by the UAV detection scenario, we prioritized background information for guiding the UAV detection and recognition. We developed a plug-and-play BAM that effectively models and characterizes complex backgrounds. By identifying regions significantly different from background features, the BAM indirectly generates spatial attention maps highlighting key areas containing UAVs with high probability. Furthermore, we seamlessly integrated this attention module into two popular detection frameworks and validated its performance through qualitative and quantitative analyses of challenging datasets. Our results demonstrate a significant improvement in detection accuracy while meeting the real-time application requirements of UAVs. This underscores the ability of background robustness to enhance the model's sensitivity to UAV targets, thereby boosting the UAV detection accuracy.

Although the BAM has demonstrated superior performance in UAV target detection tasks, it does have its limitations. For instance, in UAV target detection tasks, compared to natural image datasets used in general object detection, UAV datasets are often smaller in scale and have fewer target varieties. This somewhat limits the generalization capabilities that the network model can learn. Therefore, how to further improve the quality of learning samples by combining image features from a UAV perspective with existing data augmentation strategies is a highly valuable direction for future research.

CRediT authorship contribution statement

Xiuxiu Lin: Writing – original draft, Validation, Methodology, Conceptualization. **Yusu Niu:** Validation, Software. **Xinran Yu:** Validation, Software. **Zhun Fan:** Validation, Methodology. **Jiafan Zhuang:** Writing – review & editing, Methodology. **An-Min Zou:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgments

The work of J. Zhuang was supported in part by the STU Scientific Research Foundation for Talents (Grant No. NTF22030).

Data availability

Data will be made available on request.

References

Asadzadeh, S., de Oliveira, W. J., & de Souza Filho, C. R. (2022). UAV-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives. *Journal of Petroleum Science and Engineering*, *208*, Article 109633.

Ashraf, M. W., Sultani, W., & Shah, M. (2021). Dogfight: Detecting drones from drones videos. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 7063–7072).

Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., & Yao, Y. (2024). Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 27706–27716).

Cao, Y., Chen, K., Loy, C. C., & Lin, D. (2020). Prime sample attention in object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11580–11588).

Chen, Y., Cao, Y., Hu, H., & Wang, L. (2020). Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10334–10343).

Chen, J., Du, C., Zhang, Y., Han, P., & Wei, W. (2022). A clustering-based coverage path planning method for autonomous heterogeneous UAVs. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 25546–25556.

Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., et al. (2023). Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Collins, A. G., & McDougle, S. D. (2021). Context is key for learning motor skills.

Everingham, M. (2008). The PASCAL visual object classes challenge 2008 (VOC2008) results. http://www.pascal-network.org/challenges/VOC/voc2008/year=workshop/index.html.

Ge, C., Song, Y., Ma, C., Qi, Y., & Luo, P. (2023). Rethinking attentive object detection via neural attention learning. *IEEE Transactions on Image Processing*, 1–1.

Han, D., Ye, T., Han, Y., Xia, Z., Pan, S., Wan, P., et al. (2025). Agent attention: On the integration of softmax and linear attention. In *European conference on computer vision* (pp. 124–140).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 558–567).

Heald, J. B., Lengyel, M., & Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, 489–493.

Hsieh, T.-I., Lo, Y.-C., Chen, H.-T., & Liu, T.-L. (2019). One-shot object detection with co-attention and co-excitation. *Advances in Neural Information Processing Systems*, *32*.

Hu, S., Gao, F., Zhou, X., Dong, J., & Du, Q. (2024). Hybrid convolutional and attention network for hyperspectral image denoising. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5.

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7132–7141).

Jocher, G., & Qiu, J. (2024). Ultralytics YOLO11. URL https://github.com/ultralytics/ultralytics.

Jocher, G., Qiu, J., & Chaurasia, A. (2023). Ultralytics YOLO. URL https://github.com/ultralytics/ultralytics.

Jocher, G., Stoken, A., Borovec, J., NanoCode012, Chaurasia, A., TaoXie, et al. (2021). Ultralytics/YOLOv5: v5.0-YOLOv5-P6 1280 models AWS supervise.ly and YouTube integrations. *Zenodo*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. ArXiv Preprint arXiv:1412.6980.

Leng, J., Zhou, Y., Ye, Y., Gao, C., Gao, X., et al. (2023). Research progress on object detection from the UAV perspective. *Journal of Image and Graphics (China)*, *28*(9), 2563–2586.

Li, J., Ye, D. H., Chung, T., Kolsch, M., Wachs, J., & Bouman, C. (2016). Multi-target detection and tracking from a single camera in unmanned aerial vehicles (UAVs). In *IEEE/RSJ international conference on intelligent robots and systems* (pp. 4992–4997).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In *Computer vision – ECCV 2014* (pp. 740–755).

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., et al. (2017). The future of earth observation in hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3879–3914.

Mehta, P., Gupta, R., & Tanwar, S. (2020). Blockchain envisioned UAV networks: Challenges, solutions, and comparisons. *Computer Communications*, *151*, 518–538.

Mohamed, N., Al-Jaroodi, J., Jawhar, I., Idries, A., & Mohammed, F. (2020). Unmanned aerial vehicles applications in future smart cities. *Technological Forecasting and Social Change*, *153*, Article 119293.

Pan, M., Chen, C., Yin, X., & Huang, Z. (2022). UAV-aided emergency environmental monitoring in infrastructure-less areas: LoRa mesh networking approach. *IEEE Internet of Things Journal*, *9*(4), 2918–2932.

Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2015). You Only Look Once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Román, A., Tovar-Sánchez, A., Roque-Atienza, D., Huertas, I., Caballero, I., Fraile-Nuez, E., et al. (2022). Unmanned aerial vehicles (UAVs) as a tool for hazard assessment: The 2021 eruption of Cumbre Vieja volcano, La Palma island (Spain). *The Science of the Total Environment*, *843*, Article 157092.

Rozantsev, A., Lepetit, V., & Fua, P. (2015). Flying objects detection from a single moving camera. In *IEEE conference on computer vision and pattern recognition* (pp. 4128–4136).

Rozantsev, A., Lepetit, V., & Fua, P. (2017). Detecting flying objects using a single moving camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(5), 879–892.

Sangam, T., Dave, I. R., Sultani, W., & Shah, M. (2023). TransVisDrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. In *IEEE international conference on robotics and automation* (pp. 6006–6013).

Shumeye Lakew, D., Sa'ad, U., Dao, N.-N., Na, W., & Cho, S. (2020). Routing in flying ad hoc networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *22*(2), 1071–1120.

Tang, Z., Gao, Y., Xun, Z., Peng, F., Sun, Y., Liu, S., et al. (2023). Strong detector with simple tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 3047–3053).

Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9626–9635).

Tran, T. M., Vu, T. N., Vo, N. D., Nguyen, T. V., & Nguyen, K. (2022). Anomaly analysis in images and videos: A comprehensive review. *ACM Computing Surveys*, *55*(7).

Tsao, K.-Y., Girdler, T., & Vassilakis, V. G. (2022). A survey of cyber security threats and solutions for UAV communications and flying ad-hoc networks. *Ad Hoc Networks*, *133*, Article 102894.

Wan, D., Lu, R., Shen, S., Xu, T., Lang, X., & Ren, Z. (2023). Mixed local channel attention for object detection. *Engineering Applications of Artificial Intelligence*, *123*, Article 106442.

Wan, Y., Zhong, Y., Ma, A., & Zhang, L. (2023). An accurate UAV 3-D path planning method for disaster emergency response based on an improved multiobjective swarm intelligence algorithm. *IEEE Transactions on Cybernetics*, *53*(4), 2658–2671.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2024). Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458.

Wang, J., Song, T., Zhang, Y., Wang, S., Lin, W., Wang, Z., et al. (2022). Decoupled teacher for semi-supervised drone detection. In *IEEE 8th international conference on computer and communications* (pp. 1869–1873).

Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., et al. (2021). End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8737–8746).

Wang, C.-Y., Yeh, I.-H., & Mark Liao, H.-Y. (2025). Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision* (pp. 1–21).

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision*.

Wu, H., Chen, Y., Wang, N., & Zhang, Z.-X. (2019). Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9216–9224).

Xiao, Z., Zhu, L., Liu, Y., Yi, P., Zhang, R., Xia, X.-G., et al. (2022). A survey on millimeter-wave beamforming enabled UAV communications and networking. *IEEE Communications Surveys & Tutorials*, *24*(1), 557–610.

Xie, G., Wang, J., Liu, J., Lyu, J., Liu, Y., Wang, C., et al. (2024). IM-IAD: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*, 2720–2733.

Xie, J., Yu, J., Wu, J., Shi, Z., & Chen, J. (2020). Adaptive switching spatial-temporal fusion detection for remote flying drones. *IEEE Transactions on Vehicular Technology*, *69*(7), 6964–6976.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., et al. (2019). SCRDet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8231–8240).

Yang, Z., Yu, X., Dedman, S., Rosso, M., Zhu, J., Yang, J., et al. (2022). UAV remote sensing applications in marine monitoring: Knowledge visualization and review. *Science of the Total Environment*, *838*, Article 155939.

Yuan, X., Cheng, G., Yan, K., Zeng, Q., & Han, J. (2023). Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6317–6327).

Zhao, Q., Liu, B., Lyu, S., Wang, C., & Zhang, H. (2023). TPH-YOLOv5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer. *Remote Sensing*, *15*(6), 1687.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., et al. (2024). Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16965–16974).

Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 2778–2788).

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Zhu, Y., Zhao, C., Guo, H., Wang, J., Zhao, X., & Lu, H. (2019). Attention CoupleNet: Fully convolutional attention coupling network for object detection. *IEEE Transactions on Image Processing*, *28*(1), 113–126.