Mathematical Biosciences and Engineering

MBE, 20(4): 6630–6651.
DOI: 10.3934/mbe.2023285
Received: 17 December 2022
Revised: 06 January 2023
Accepted: 15 January 2023
Published: 02 February 2023

http://www.aimspress.com/journal/MBE

Research article

A-RetinaNet: A novel RetinaNet with an asymmetric attention fusion mechanism for dim and small drone detection in infrared images

Zhijing Xu, Jingjing Su* and Kan Huang

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

*** Correspondence:** Email: 202130310028@stu.shmtu.edu.cn.

Abstract: To solve the problems of texture lacking and resolution coarseness in the detection of dim and small drone targets in infrared images, we propose a novel RetinaNet with an asymmetric attention fusion mechanism for dim and small drone detection. First, we propose a super-resolution texture-enhancement network as an effective solution for the lack of texture-related information on small infrared targets. The network generates super-resolution images and enhances the texture features of the targets. Second, considering the inadequacy of feature pyramids in the feature fusion stage, we use an asymmetric attention fusion mechanism to constitute an asymmetric attention fusion pyramid network for cross-layer feature fusion in a bidirectional manner; it achieves high-quality semantic and location detail information interaction between scale features. Third, a global average pooling layer is employed to capture global spatial-sensitive information, thus effectively identifying features and achieving classification. Experiments were conducted by using a publicly available infrared image dim-small drone target detection dataset; the results show that the proposed method achieves an AP of 95.43% and a recall of 80.6%, which is a significant improvement over the current mainstream target detection algorithms.

Keywords: asymmetric attention fusion; infrared image; drone detection; RetinaNet; super-resolution

1. Introduction

With the development of aerial photography technology, drones are now frequently deployed in military and civilian applications such as intelligent inspection and agricultural drug spraying due to their low cost, simple operation, concealment and mobility. While delivering convenience, drones also bring several issues, including illegal intrusion, interference with civil aviation and the leakage of military secrets, which remain dangers to the security of military regions and public privacy. With the recent development in infrared imaging technology, the detection of small objects via infrared imaging has boosted a variety of applications, including low-altitude aircraft flying, infrared search and tracking [1] and military reconnaissance systems. Therefore, infrared imaging systems are a

critical technology for the identification of drone targets. The following difficulties exist when detecting drone targets in infrared images:

1) Imbalance between the target and background areas. According to the definition of the International Society for Optical Engineering, in a 256 × 256 image, a small target in an infrared image takes up less than 9 × 9 pixels in the image. There may even be a one-pixel target in public databases.

2) Lack of texture information. In complex scenes, dim and small targets are often submerged in the background in infrared images and appear as point targets lacking shape and texture information, which leads to low detection accuracy and high false alarm rates.

3) Contradiction between resolution and semantic information. In the deep network, the semantic information is strong and the feature map resolution is low; but, in complex scenes, the dim and small drones are often submerged in the background in infrared images, which leads us to need high-resolution feature maps to achieve accurate detection of the target information. Therefore, dim and small drone detection in infrared images is a challenging task.

To solve these challenging problems, researchers have proposed various small target detection algorithms for infrared systems in recent years. Traditional small target detection methods for infrared images can be divided into two categories: single-frame detection and multi-frame detection methods. Infrared-purposed small target detection methods based on single-frame detection include local contrast-based methods [2–8], filter-based methods [9] and structure-based methods [10]. All of these approaches make use of the distinction between background and target representations, e.g., grayscale values and continuity. Most of the methods assume that the gray value of the target is higher than the gray value of the background, but this hypothesis is not universal. In fact, in complex scenes, the background is much brighter than the target, making it more difficult to identify the target since it is frequently masked by the background. Infrared-purposed small target detection methods based on multi-frame detection mainly include motion trajectory-based methods [11,12] and energy-based methods [13]. However, these methods cannot satisfy real-time detection requirements.

With the development of deep learning techniques, there has been significant research progress in areas of natural language processing [14] and computer vision [15]. Many researchers have started to use convolutional neural networks (CNNs) to improve the detection performance of infrared-based small targets. Infrared-purposed small target detection methods based on CNNs are divided into two categories: single-stage detection and two-stage detection. The two-stage detectors mainly contain R-CNNs [16], Fast R-CNNs [17] and Faster R-CNNs [18]. The single-stage detectors mainly contain SSD [19], YOLO [20] and RetinaNet models [21]. As a representative single-stage target detection algorithm, the backbone network of RetinaNet is laterally connected to form a feature pyramid with rich semantic information, and its detection network is divided into two parallel sub-networks, i.e., a classification sub-network and regression sub-network for effective classification recognition and detection regression, respectively. In addition, RetinaNet solves the problem of positive and negative sample imbalance in the single-stage detection algorithm by using the focal loss function to achieve the same detection accuracy as the two-stage detection method; additionally, the convergence speed is fast. In this work, we applied the RetinaNet network for dim and small drone detection in infrared images and adopted some improvements to it. We propose a super-resolution texture-enhancement network to improve the resolution of infrared drone images and enhance target texture information. Second, to make multi-scale feature fusion more effective in detecting the point targets of drones in complex infrared backgrounds, an asymmetric attention fusion mechanism (AAFM) is proposed; it effectively combines deep and shallow features. Finally, to accurately identify targets, a classification sub-network has been constructed by using a global average pooling layer. In addition,

we use k-means clustering to generate adaptive anchor boxes, which improves the recall rate of dim and small drones in infrared images.

In summary, the main contributions of this paper are as follows.

1) Due to the lack of texture information for dim and small targets in infrared images, we perform super-resolution texture enhancement to enhance the texture content information and generate super-resolution images before the infrared images enter the backbone network.

2) We inject high-level semantic and low-level location detail information into multi-scale features into the AAFM, which is beneficial for the recognition and localization of drone targets. The AAFM effectively combines different levels of feature maps via a pixel-by-pixel spatial attention module (PAM) and global-channel attention module (GAM), and the size of the feature map generated by the AAFM is the same as the input feature map, which solves the problem of difficult detection for dim and small targets in complex infrared backgrounds.

3) We use a combination of standard convolution and global averaging pooling to capture finer-grained global spatial information and effectively identify the feature maps for classification purposes. Finally, we propose a new RetinaNet-based network, called A-RetinaNet. Compared with RetinaNet, the detection accuracy is higher by 16.32% and the recall rate is higher by 49.34%. Unlike other detection networks, the proposed network achieves a good balance between detection accuracy and real-time performance.

The rest of this paper is organized as follows: Section 2 briefly reviews the related work on infrared-based dim and small target detection. Section 3 details the improvement strategies of A-RetinaNet. Section 4 outlines the experimental results and a comparison with other state-of-the-art methods. Section 5 gives the conclusions of this paper.

2. Related work

In this section, we briefly review the existing methods of infrared-based small target detection.

The detection methods for small targets in infrared images are mainly divided into two categories: traditional infrared-based small target detection methods and CNN-based detection methods.

Traditional infrared-based small target detection can be divided into two categories. 1) Single-frame-based detection. It mainly uses the difference between the target and background information for filtering, or different characteristics such as the gray value and gradient value for local comparison. Fan et al. [2] designed an infrared-purposed small target detection method based on candidate region suggestion and a CNN classifier by exploiting the difference between the target and background in terms of grayscale values for image pre-processing. Finally, it was able to separate the real target from the background. Zhang et al. [3] proposed a new infrared-purposed small target detection algorithm by exploiting the fact that the neighborhood pixel intensities of infrared-based targets show a light-dark-light pattern. Cao et al. [4] found that small target derivatives have positive and negative properties, while background information does not have such properties; so, they designed an infrared-purposed small target detection algorithm based on derivative similarity. Lu et al. [9] found that salt noise and the background are similar, so they used the idea of median filtering to design a marine background filter for removing the target; they then used the difference between the background and the target to extract the target region. Dong et al. [11] developed a model to automatically select the intensity or direction modality, calculate the saliency according to the smoothness of the background and achieve the detection of infrared targets by using the antivibration pipeline filtering method. Chen et al. [5] used a contrast map to detect the target based on the characteristic that the brightness value of the target area is higher than the brightness

value of the background area. The method can be effective for the detection of small targets, but, while enhancing the target brightness, it also enhances some high-brightness noise, which eventually leads to a high false alarm rate; it is calculated pixel by pixel, resulting in a slow detection speed. Therefore, Han et al. [6] used a new local contrast method to achieve the detection of small targets in infrared images. The method uses chunking calculation, which not only takes the maximum value in grayscale values, but it also considers the mean value, improving the shortcomings of the local contrast method. Since the sliding window and small targets have similar sizes, the detection effect is not obvious in complex environments. Qin and Li [7] used sliding windows to segment infrared images; the sliding window size is larger than the target size, solving the drawback of low detection accuracy of infrared images in complex environments. The above methods can only better detect targets with high grayscale values and cannot detect dark targets.

2) Multi-frame-based detection. It not only uses the basic image information, but it also combines the motion track of the target for target detection, so it is computationally extensive. Ding et al. [12] designed an adaptive pipeline filter to correct the detection results by exploiting the property that the motion trajectory of the target is continuous. Guo et al. [13] proposed a parallel computation method based on dynamic programming by exploiting the property of directionality of the target energy, which divides the search area into various small parts in parallel to reduce the computational effort and improve the detection efficiency.

CNN-based methods. Compared with traditional methods, CNNs directly take infrared images as input and learn the features of the small targets by using a neural network model to solve the detection problem of infrared-based small targets. Ju et al. [22] designed a simple and effective small-target detection network. The network modifies the downsampling operation to extended convolution in order to prevent small target information loss, which preserves the image resolution. And, they used the pass-through module to combine the position information and semantic information. Liang et al. [23] modified the backbone for the RetinaNet network target information loss problem and designed a small target detection method based on parallel multi-scale feature enhancement. Hou et al. [24] used multi-scale fusion by stitching convolutional kernels of different sizes to form 16 feature channels, which improved the detection performance of small targets. And, they split the high-resolution feature map by using the pass-through module to form a multi-scale feature map with the low-resolution feature map, which facilitates the detection of small targets of different scale sizes. Li et al. [25] designed a detection network focused on detecting small targets by modifying the residual module, soft-NMS, and the convolution in the YOLOv3 network, which improved the detection speed of small targets, as well as the detection accuracy of obscured targets. Song et al. [26] developed a small target detection algorithm based on multi-scale feature fusion by effectively combining global and local information through the CornerNet network. Compared with YOLOv3, the detection accuracy and speed are improved by more than 11%. Wang et al. [27] designed an infrared-purposed small target detection model based on a feature fusion convolutional network by using dense connectivity to combine low-level position information and high-level semantic information. The network was able to reach a detection speed of 105 FPS, and it could even detect small targets of 2×2 pixel size. Ju et al. [28] designed a CNN-based infrared-based small target detector by using an image filtering module, performing multi-scale feature layer fusion to obtain a confidence map and entering the detection network to obtain the classification and location of the target. Yao et al. [29] designed a modified version of the Fully Convolutional One-Stage Object Detection (FCOS) network by incorporating the maximum filtering method, which balances real-time operability and accuracy. To avoid the loss of targets in the deep network due to the use of pooling layers, Tong et al. [30] proposed an enhanced asymmetric attention (EAA) mechanism based on U-Net. The EAA attention

module is used for cross-layer feature fusion to detect small targets in infrared images. Xu et al. [31] added dilated convolution for dual-channel feature extraction in YOLOv3 to achieve the effect of expanding the perceptual field and generating feature layers of different sizes to detect targets of different sizes.

In conclusion, although the existing infrared-purposed dim and small target detection methods have made significant breakthroughs, they have certain limitations. First, as the number of network layers increases, the deep feature map contains more semantic information, but the resolution of the feature map is low. In addition, in practical scene applications, the background information of infrared images is more complex and the target pixels are extremely small, making them difficult to detect. In this work, we use a super-resolution texture-enhancement network, which not only extracts the image texture information and the semantic information of the deep network, but it also generates super-resolution images, solving the contradictory problem of semantic information and resolution. In addition, we use an AAFM, which considers the fusion of semantic and position information across the whole network and makes full use of contextual information to improve the detection performance of infrared-based dim and small targets. We also improved the detection network and generated adaptive anchor boxes. Finally, the improved model was tested on a publicly available infrared-based dim and small drone dataset. The results show that the detection accuracy and recall of the method in this paper can achieve good results when compared with other infrared-purposed target detection methods.

3. Proposed method

3.1. Our network structure

In this section, our network structure is described in detail, and the A-RetinaNet network is shown in Figure 1. The network is composed of four components: an image pre-processing module, backbone network, AAFM module and asymmetric detection (AD) module. The image pre-processing section performs super-resolution texture enhancement of the image, the backbone network is used to extract target features, the AAFM module is used to enhance target features submerged in the background and the AD module is used to efficiently identify target categories and locations. In addition, k-means clustering is used to generate adaptive anchor box size and scale. Infrared-based drone images are transmitted to our network and then a series of operations are performed; eventually, the detection network obtains the specific location and confidence level of the drones.

3.2. Super-resolution texture-enhancement module

The traditional RetinaNet network directly feeds image data into the backbone network. However, in complex infrared image scenes, small targets are easily submerged in the background, resulting in a shortage of texture and shape information, poor image resolution and low detection accuracy. To solve these issues, we first perform a pre-processing operation on the image by using a super-resolution texture enhancement (SRTE) module. Unlike the feature textures transfer [32], we use a densely connected network structure [33] to extract all levels of feature layers of the high-resolution image to obtain global features in the image, as shown in Figure 2. Fixing the input image size to 256 × 256, the input of the SRTE module is divided into two parts: source features and reference features. The SRTE module uses a content extractor and texture extractor to effectively combine depth semantic information and shallow spatial feature information. SRTE enhances feature texture content information and generates super-resolution images while removing image background noise.

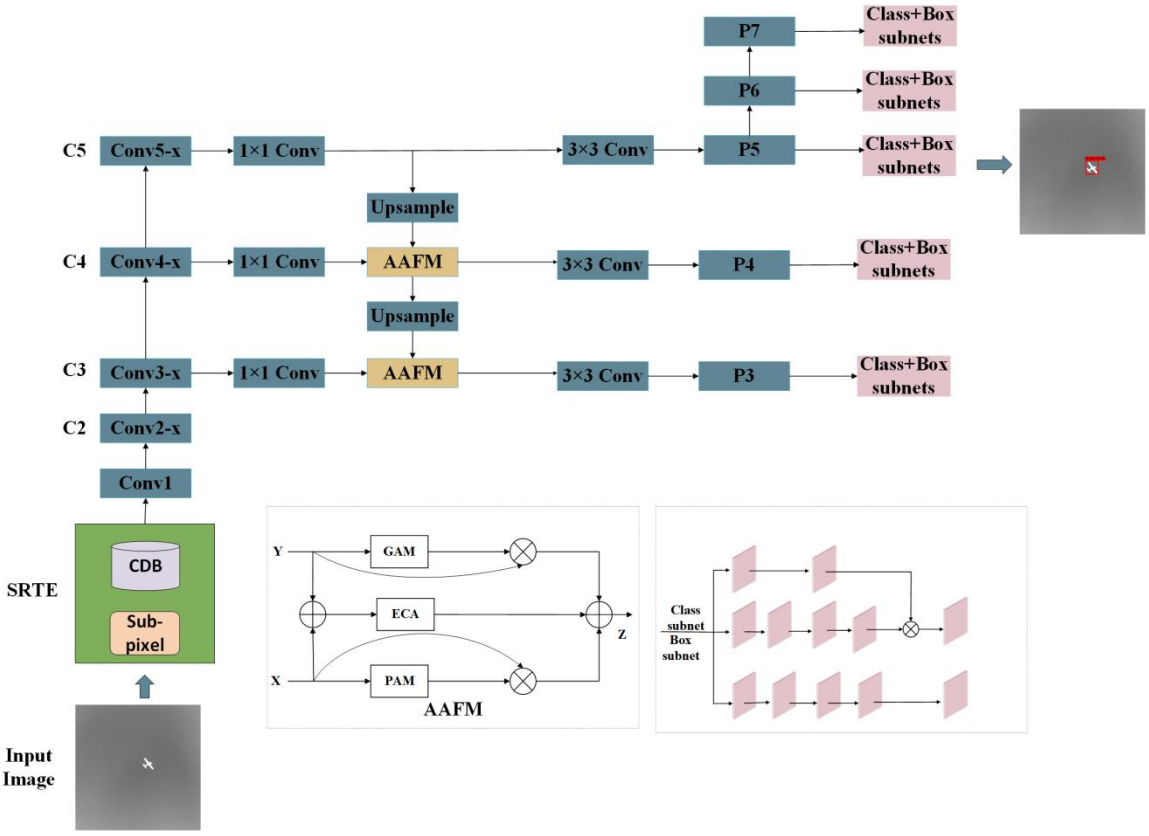


Figure 1. Structure of the A-RetinaNet network.

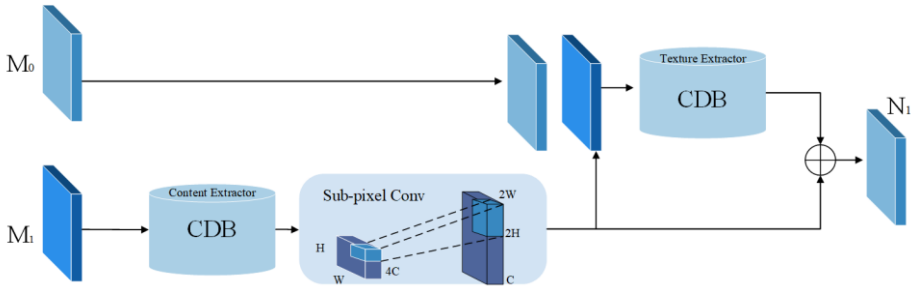


Figure 2. Overall flow of the SRTE module.

N_1 is the output of the SRTE module, denoted as

$$N_1 = T_t(M_0 \,||\, T_c(M_1)\uparrow_{2\times}) + T_c(M_1)\uparrow_{2\times} \tag{1}$$

where T_c and T_t denote the content extractor and texture extractor, respectively; || denotes feature fusion, $\uparrow_{2\times}$ denotes bifurcation by subpixel convolution, M_0 denotes the reference features and M_1 represents the source features.

For the input infrared drone image, the input is first divided into source features and reference features. The source features first go through a content extractor structure to extract the main content information of the input features, and then go through sub-pixel convolution to double the resolution of the content features and generate super-resolution images. Second, the reference features and the source features are fused pixel by pixel and then sent to the texture extractor, which extracts the semantic information from the input and reliable texture information from the reference features while suppressing noise. Finally, we fuse the texture features and high-resolution content features via

residual concatenation to obtain a rich texture information feature map. The final generated super-resolution feature map has both reliable texture information selected in the shallow features and semantic information in the deep features.

The structures of the content extractor and texture extractor are similar. As shown in Figure 3, the convolution-dense block (CDB) is used as the basic module of the content and texture extractors. The CDB enhances the feature delivery and reduces the number of parameters to improve the detection efficiency while mapping its deep feature information. The CDB constructs dense connections from four sets of standard convolutions and the activation function Leaky ReLU, which improves the expressiveness of the deep network and avoids overfitting.

The difference with ResNet is that our CDB possesses 3 × 3 convolution and Leaky ReLU. Since the batch normalization (BN) layer would ignore the absolute difference between image pixels and destroy the original contrast information of the image, we do not use the BN layer. And, we use the Leaky ReLU function to solve the dying ReLU problem. In addition, unlike the residual connections of ResNet, our content extractor uses cascaded dense connections. These densely connected modules increase the capacity of the network, enhance feature conduction and ensure training stability.

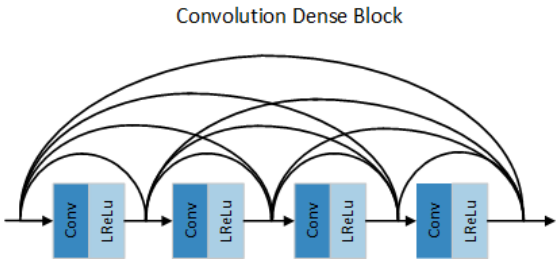


Figure 3. Structure of the content extractor.

3.3. AAFM module

Since the feature pyramid network lacks an analysis of the overall structure of the feature map, it cannot take into account both shallow position and high-level semantic information. We constructed an AAFM and applied it to the feature fusion module of the original feature pyramid. Specifically, we were inspired by asymmetric contextual modulation [34], which allows it to focus on both high-level semantic understanding information and low-level position analysis information through the use of a bidirectional approach. First, the super-resolution images are sent to the backbone network to extract features and the obtained feature maps {C3, C4, C5} are used to construct the AAFM. The summation module of each layer feature map is modified as an AAFM. Through top-down global channel attention modulation, the high-level feature maps assign different weight values according to the importance of each channel, allowing the network to use global information to propagate effective semantic information from the deep network to the shallow layer and improve feature representation. Through bottom-up pixel-by-pixel attention modulation, the low-level feature maps embed position information into the high-level semantic features to highlight feature details of the infrared-based target and avoid the problem of position information being overwhelmed in the complex background. The shallow and deep feature maps are fused pixel by pixel to obtain the optimized global channel information; then, the effective channel attention (ECA) module [35] is used to improve the input performance of the network and enhance the detection accuracy. It facilitates the network to detect the location of dim and small drones in infrared images more easily. The structure of the AAFM is shown in Figure 4.

As can be seen in Figure 4, the low-level feature map *X* extracts shallow position information through the PAM. The high-level feature map *Y* is assigned weight values according to the channel importance in the channel dimension by the GAM, which enhances the weight share of effective features and reduces the weight of interfering features. The low-level feature map and the high-level feature map are fused pixel by pixel and then passed through the ECA module to solve the feature mismatch problem and enhance the detection accuracy. Finally, feature aggregation is performed to add the contextual features obtained from the high-level feature map to the low-level feature so that the low-level feature map can receive richer contextual information from the high-level feature map. The computational properties of the AAFM are as follows.

$$\begin{cases} M=X+Y \\ Z=G(Y)\otimes Y+P(X)\otimes X+E(M) \end{cases} \tag{2}$$

where *X* denotes the low-level feature map, which contains a lot of location information, and *Y* denotes the high-level feature map, which contains semantic information in the image. *G*(*Y*) denotes channel attention modulation, *P*(*X*) denotes pixel-by-pixel convolution and *E*(*M*) denotes the ECA module.

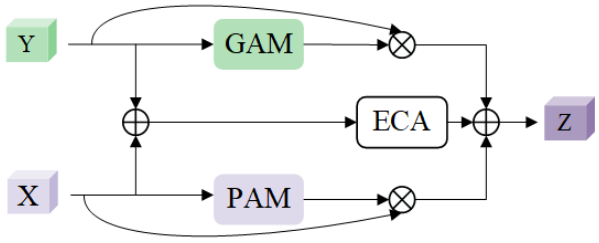


Figure 4. Overall flow of the AAFM module.

The GAM is shown in Figure 5. The input features pass through two parallel branches, and the branches include the global average pooling layer and the convolutional layer. After passing through the global average pooling layer, the global feature information of the feature map is obtained by using a 1 × 1, C/4 convolutional layer down-scaling operation and a 1 × 1, C convolutional layer up-scaling operation. After the sigmoid activation layer, it is multiplied with the input layer to generate the attention weight map and calculate the weight value of each channel. Finally, it is summed and fused with the input layer that has undergone 1 × 1 convolution. The global channel attention modulation is calculated with the following characteristics.

$$\begin{cases} Y_1=\sigma(B(C_2(B(C_1(G(Y)))))) \\ Y_2=YY_1 \\ G(Y)=C_2Y+Y_2 \end{cases} \tag{3}$$

where *G* is the global average pooling, C_1 is the 1 × 1 convolution with channel number C/4 and *B* is the batch normalization followed by the activation function ReLU. C_2 is the 1 × 1 convolution with channel number C, and σ is the sigmoid function.

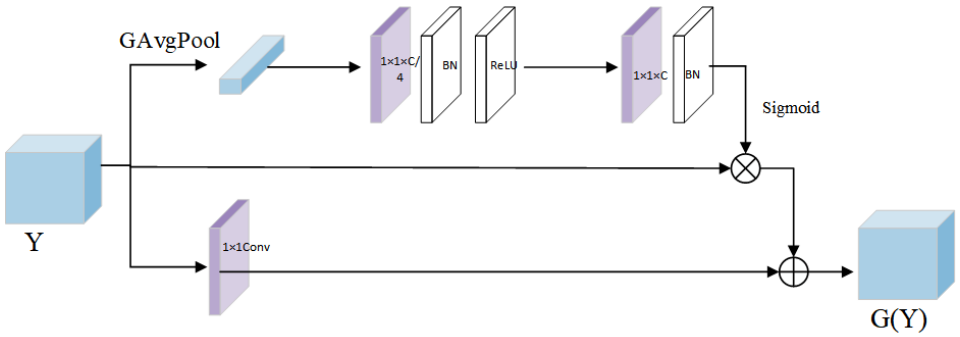


Figure 5. Overflow of the GAM.

To highlight the detailed information on small targets in infrared images, we use the PAM as shown in Figure 6. The input features are convolved pixel by pixel, while the BN-ReLU function is added to the $PWConv_1$ convolution operation to make the network more effective for feature selection. Finally, the sigmoid activation layer (σ) is passed so that the weight values of the spatial regions are calculated. The computational properties of the PAM are as follows:

$$P(X)=\sigma\left(B(PWConv_2(\delta(B(PWConv_1(X))))\right) \tag{4}$$

where PWConv denotes pixel-by-pixel convolution, where the convolutional kernel sizes of $PWConv_1$ and $PWConv_2$ are $C/4 \times 1 \times 1$ and $C \times 1 \times 1$, respectively. σ denotes the sigmoid function, *B* denotes BN and δ denotes the rectified linear unit (ReLU).

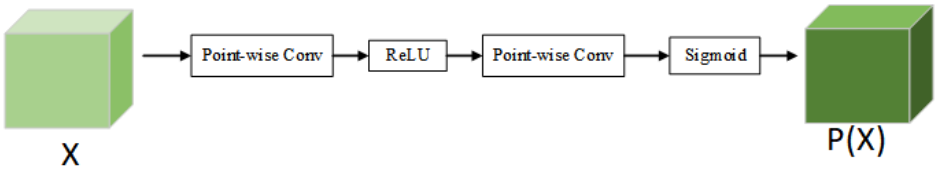


Figure 6. Overall flow of the PAM.

To capture the information content locally across channels and enhance the detection performance for small targets submerged in the background, an ECA mechanism is used in this algorithm, as shown in Figure 7. First, the input features are pooled through a global average pooling layer that reduces the number of model parameters while retaining a large amount of spatial information. Second, the information interaction and weight sharing between adjacent layer channels are achieved by one-dimensional convolution. Then, the input feature is processed by the sigmoid function. Finally, the input feature maps and the processed feature map weights are multiplied to obtain the output.

$$E(M)=\sigma(CD1_k(G(M))) \otimes M \tag{5}$$

where *M* denotes the feature map generated by fusing the low-level feature map and the high-level feature map, *G* denotes the global average pooling, $CD1_k$ denotes the one-dimensional convolution, σ denotes the sigmoid function and ⊗ indicates pixel-by-pixel multiplication.

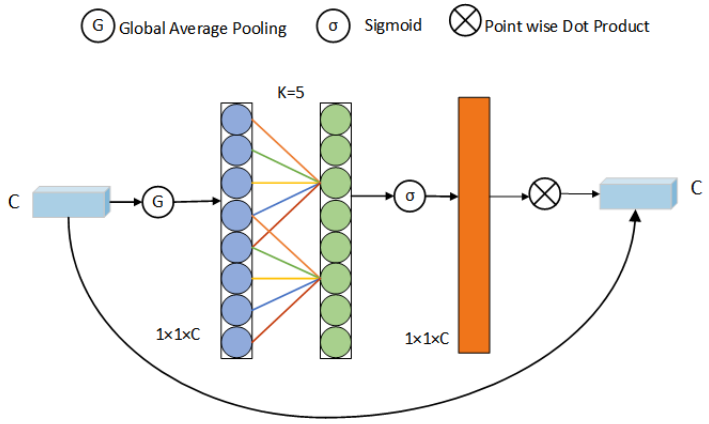


Figure 7. Overall flow of ECA mechanism.

3.4. AD module

To increase the detection accuracy of our network, we used an AD module, as shown in Figure 8. For classification detection, the fully connected layer [36] is ideally suited since it is more sensitive to spatial information and can more clearly discriminate between whole and partial targets. The convolutional layer can effectively extract object information, and it is more suitable for regression detection. However, the large number of fully connected parameters and large computational effort can lead to poor real-time target detection, so we chose to use a global average pooling layer [37] instead of the fully connected layer, which reduces the number of parameters, improves the detection efficiency and, on the other hand, mitigates the occurrence of overfitting. Therefore, we modified the classification sub-network to fuse the global average pooling layer and the convolutional layer for classification detection and achieved good results. After each layer of the feature map enters the classification sub-network, it first passes through a 3 × 3 convolutional layer with 256 channels and then divides into two branches, one of which is the same structure as the RetinaNet classification sub-network, through three 3 × 3 convolutional layers; the other passes through the global average pooling layer. Finally, the target classification result is obtained by the sigmoid activation function and a 3 × 3 convolutional layer of K × A channels.

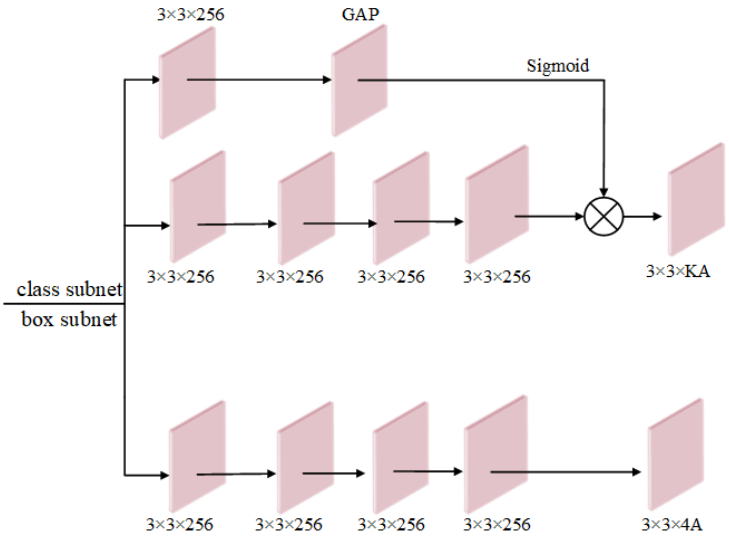


Figure 8. Overall flow of AD. GAP: global average pooling layer.

3.5. Generation of adaptive anchor boxes

The original RetinaNet network selected five effective feature layers as prediction feature layers, with sizes [32 × 32, 64 × 64, 128 × 128, 256 × 256, 512 × 512], and it set nine anchor boxes with three area ratios $\{2^0, 2^{1/3}, 2^{2/3}\}$ and three aspect ratios {1:1, 1:2, 2:1} on each feature map. The size range for anchor boxes is 32 to 813 pixels. The image size of the infrared-based drone dataset used in this work was 256 × 256, and the targets in the image were divided into extended targets and point targets, so the anchor box size setting of the original RetinaNet network generally has poor performance in terms of detecting targets in this dataset. Our method uses adaptive clustering to select anchor boxes of a suitable size and aspect ratio based on the characteristics that the drone targets are smaller in the images and there are more point targets, which reduces the problem of large computation caused by the original RetinaNet model due to unsuitable anchor boxes.

We use the k-means algorithm to cluster the real label boxes and aspect ratios of the infrared-based drone dataset to achieve appropriate anchor box proportions and sizes for better detection of the model. The k-means clustering algorithm is a data clustering method based on the Euclidean distance measure, and it assigns the sample data to its closest cluster center. However, because it often generates more error for large anchor boxes rather than small ones, we use the intersection of union (IOU) approach to reduce the error between distances. The IOU is the intersection proportion of label boxes and cluster centers (anchor boxes), as shown in Eq (6), where the larger the IOU, the smaller the distance.

$$d_{\text{box, cluster}} = 1 - IOU(\text{box, cluster}) = 1 - \frac{x_{min}(box,cluster) * y_{min}(box,cluster)}{\text{box\_area} * \text{cluster\_area} - x_{min} * y_{min}} \tag{6}$$

We select all of the XML file information and cluster the label boxes of the dataset into k clusters, requiring the cluster center coordinates to be the same as the center coordinates of the label boxes. First, we calculate the distance between each labeled box and the center of the k clusters separately, as shown in Eq (6), where *box_area* and *cluster_area* are the area of the label box and anchor box, respectively. x_{min} (box, cluster) and y_{min} (box, cluster) are the minimum coordinate values x and y of the label box and anchor box, respectively. Then, the label boxes are assigned to the cluster centers that are closest to them. Finally, after categorizing all label boxes, the cluster centers are updated and the cluster centers are calculated as shown in Eq (7).

$$\begin{aligned} w_i' &= \text{medium}(w_i) \\ h_i' &= \text{medium}(h_i) \end{aligned} \tag{7}$$

where (w_i, h_i) respectively denote the width and height of the label box, (w_i', h_i') respectively denote the width and height of the anchor box, $i \in \{1, 2, ..., k\}$ and medium is the median of the width and height of all label boxes in the same cluster. The new cluster center is the median value of the label boxes in each cluster. The algorithm repeats the above steps until the elements in each cluster no longer change.

In this work, we divided the dataset label boxes into nine clusters, i.e., k = 9. From Figure 9, the aspect ratios of the anchor boxes were {0.7, 1.0, 1.0, 0.7, 1.0, 0.9, 2.1, 0.6, 1.1}; so, in this study, [0.6, 0.8, 1.0] was used as the aspect ratio of anchor boxes, and the anchor box sizes were set to [16 × 16, 32 × 32, 64 × 64, 128 × 128, 256 × 256]. The generation of adaptive anchor boxes not only improves the recall of the model, enabling the detection of more positive samples, but it also improves the detection accuracy and efficiency of the algorithm.

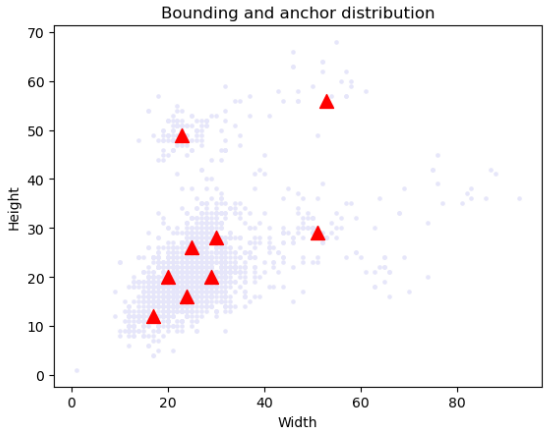


Figure 9. Distribution of adaptive anchor boxes.

3.6. Loss function

We need to calculate the error between the predicted and true values while training the model; the lower the error value, the more accurate the model. The sample types are separated into positive and negative samples throughout the training stage. Positive samples are those where there is an overlap between the predicted box and the true box of more than 0.5, negative samples are those where there is an overlap of less than 0.4 and ignored samples are those where there is an overlap between 0.4 and 0.5. Due to the unbalanced distribution of positive and negative samples in infrared images, we chose to employ a novel loss function developed for the RetinaNet network, called focal loss. This modification of the cross-entropy loss function ended up causing the model to concentrate more on sparse and hard-to-classify samples during training by lowering the weight of many simple negative samples.

First, we recall the cross-entropy loss function, as shown in Eq (8).

$$L=-y\log y^{'}-(1-y)\log\left(1-y^{'}\right)=\begin{cases}-\log y^{'}, & y=1\\ -\log\left(1-y^{'}\right), & y=0\end{cases} \tag{8}$$

where y is the sample label and y' is the probability that the sample is positive as predicted by the model. If y is a positive sample, then $y = 1$; otherwise, $y = 0$. Equation (8) demonstrates that the model loss is less when the probability that the sample is positive is greater; conversely, the model loss is bigger when the probability that the sample is negative is greater.

In the case of the cross-entropy loss function, a large number of negative samples will result in a slower iteration of the loss function that is not optimal. The loss function's weight is adjusted by the addition of the factor α to make it focus more on positive samples. Introducing the α factor adjusts the influence of positive and negative samples on the weight of the loss function to make it focus more on positive samples. However, if there is a large number of simple and easy-to-classify negative samples, the training process will be carried out in favor of negative samples and a large number of simple samples will be superimposed, which may exceed the loss of difficult samples, resulting in poor training; so, the γ factor has been introduced to focus the network training on sparse and difficult-to-classify samples.

$$L_{fl}=\begin{cases}-\alpha\left(1-y^{'}\right)^{\gamma}\log y^{'}, & y=1\\ -(1-\alpha)y^{'\gamma}\log\left(1-y^{'}\right), & y=0\end{cases} \tag{9}$$

Our total loss function contains both classification loss and regression loss, as shown in Eq (10).

$$L=L_{cls}+L_{reg}=L_{fl}+L_{smoothL_1} \tag{10}$$

Focal loss is used for classification loss, and smooth L_1 loss is used for regression loss, as shown in Eq (11).

$$L_{smoothL_1}(x)=\begin{cases} 0.5x^2 & |x|<1 \\ |x|-0.5 & \text{other} \end{cases} \tag{11}$$

where x represents the difference in values between the predicted and true boxes.

4. Experimental results and analysis

In this study, we designed a series of experiments to demonstrate the effectiveness of our proposed method; here, we evaluate the proposed method quantitatively and qualitatively on a publicly available infrared-based dim and small drone dataset. The details of our experimental setup and experimental results are presented in this section, along with an analysis of the results.

4.1. Introduction of dataset and evaluation indicators

In this study, we used a dataset for infrared detection and tracking of dim-small aircraft targets under a ground/air backdrop [38]. We arbitrarily selected 5206 images from the publicly available dataset, and all images were labeled with drone labels and bounding boxes. The dataset scenes are divided into backgrounds such as sky and ground, and also includes multiple scenes; example images for each scene are shown in Figure 10, and Table 1 describes the background information of these small infrared targets. The images in the simple sky background account for about 15% of the images, and the images in the complex background account for 85%.

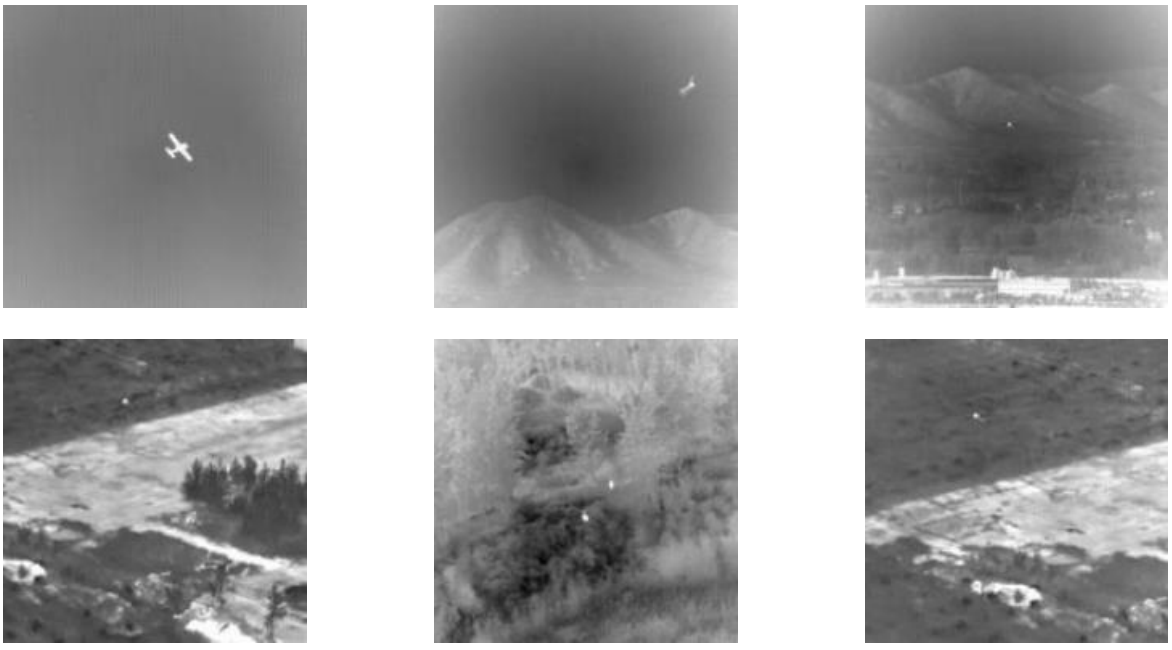


Figure 10. Sample images of the dataset with targets shown as extended targets or point targets in the image.

Table 1. Background information of infrared-based small targets.

	Details of targets and background	Image size	Number
Simple sky background	Clear extended target	256 × 256	613
	Shape-blurred extended target	256 × 256	133
Complex background	Point target, ground background	256 × 256	400
	From near to far, dim-small target, ground background	256 × 256	1779
	Point target, low SNR, ground background	256 × 256	1120
	Ground background, the background is brighter than the target	256 × 256	394
	Ground-space junction background and others	256 × 256	767

We evaluate the algorithm in terms of precision, recall, speed and precision. The precision, recall, F1 score and average precision (AP) are defined as follows.

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}}\times 100\% \tag{12}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}}\times 100\% \tag{13}$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{14}$$

$$\text{AP}=\int_0^1 \text{P(r)dr} \tag{15}$$

4.2. Network training

All models for the experiments in this study were conducted by using an Ubuntu 18.04 operating system with an I7-7700x CPU and NVIDIA GTX1080Ti GPU, and all experiments were conducted under the PyTorch framework. Our network structure is modified by RetinaNet, and our models were trained with the same hyperparameters as RetinaNet, using the Adam optimizer, with a maximum learning rate of 10^{-4}, a minimum learning rate of 10^{-6}, a batch size of 32, a momentum of 0.9 and a total of 200 epochs trained.

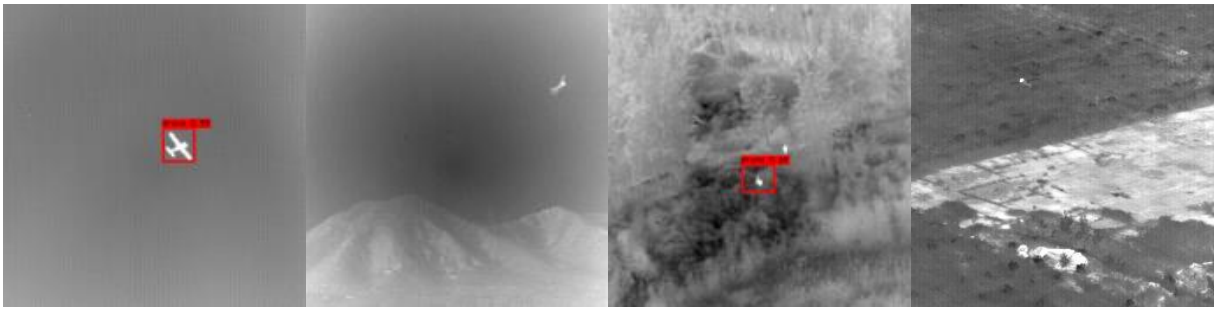
4.3. Results and analysis

4.3.1. Ablation experiments

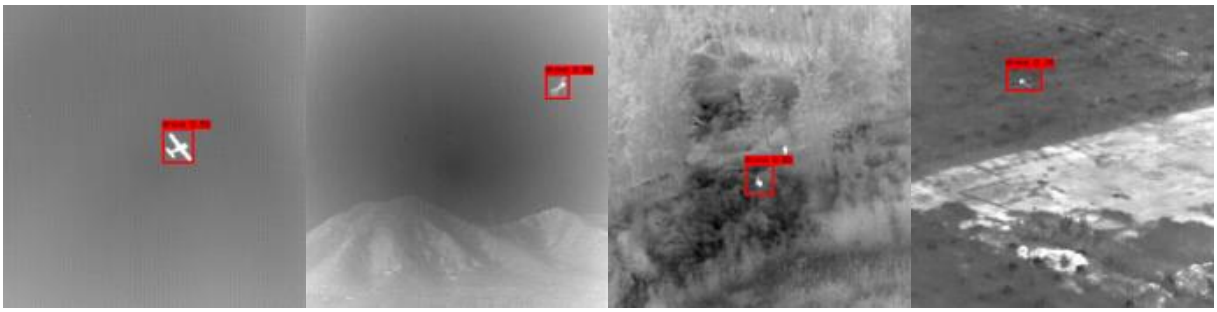
We have added detailed technical improvements to RetinaNet to allow our network to better detect infrared-based dim and small drone targets and improve overall accuracy. As shown in Table 2, the baseline network is the RetinaNet network. On the infrared-based dim-small drone dataset, the SRTE module improved the baseline by 1.74%, which confirms that the SRTE module can improve the network's ability to detect small targets by enhancing the texture information to emphasize the detailed information of small targets. The accuracy increased by 1.33% after the addition of the AAFM module to the baseline and SRTE module. This is because it integrates deep features and shallow

features, fully utilizes contextual data to ensure that the semantic information of the deep features is integrated into the low-level features and enhances the deep network's capacity to characterize small targets. Meanwhile, the introduction of the classification sub-network improved the detection accuracy by 0.36%, which confirms that the global average pooling can capture spatial information more sensitively and identify small targets more efficiently by combining the global average pooling with the convolutional layer. Finally, via adaptive anchor box generation, we reset the anchor box size and scale, which increased the AP value by 12.89%. In terms of detection accuracy, the resetting of the adaptive anchor box size had a great impact on the improvement of detection accuracy because we reset the anchor boxes that fit the dataset by using k-means clustering; this allowed the model recall to greatly improve and finally bring about the improvement of detection accuracy. The ablation experimental results show that our model improved the detection accuracy by 16.32%, which exceeded the AP value of the YOLOv5 model, thus proving that our model is more robust for the detection of dim-small drones in complex infrared backgrounds.

As shown in the example image in Figure 11, our proposed method greatly improved the detection accuracy of drone targets. Both the baseline network and the A-RetinaNet network could successfully detect the target when the target is extended in a simple background. The original RetinaNet could not detect the target when the drone was offset, resulting in a blurred shape of the target in the sky background. And, when part of the background was brighter than the drone target in a flat ground backdrop, the original RetinaNet also failed to successfully identify the target. In contrast, the A-RetinaNet network could successfully detect the target with high confidence in both point target and complex scenes.



(a)



(b)

Figure 11. Some drone detection results. (a) RetinaNet and (b) our proposed model A-RetinaNet.

Table 2. Effects of the settings of each module on the detection accuracy. AAB: adaptive anchor box.

RetinaNet	SRTE	AAFM	AD	AAB	AP
√					79.11%
√	√				80.85%
√	√	√			82.18%
√	√	√	√		82.54%
√	√	√	√	√	95.43%

4.3.2. Comparison with other advanced models

Our A-RetinaNet model is based on the RetinaNet network with ResNet50 as the backbone. To test the performance of our detection models, we conducted experiments using the infrared-based dim-small drone dataset and provide here a comparison of the detection performance of our models with YOLOv5, SSD, RetinaNet, Faster R-CNN , YOLOX, YOLOv7 and CenterNet [39]. To achieve a fair comparison, the hyperparameters and training parameters of our eight models were set to be the same as those of the benchmark experimental RetinaNet.

We recorded the detection performance of the eight models based on the evaluation metrics, as shown in Tables 3 and 4. In terms of detection accuracy, the feature map resolution extracted by Faster R-CNN was too low to detect infrared-based small targets, so Faster R-CNN had the worst accuracy value and is not suitable for infrared-based small target detection with complex backgrounds. Since dim and small drone infrared images lack color and texture shape information and YOLOv5 uses many image data enhancement methods, its detection performance was significantly better than that of Faster R-CNN, as well as other baseline models. As shown in Table 3, the AP value of YOLOv5 was 43.32% higher than that of Faster R-CNN, and 13.85% higher than that of RetinaNet. Our proposed method is based on RetinaNet, and the original RetinaNet is ineffective in detecting drone targets in complex backgrounds, with a detection accuracy of only 79.11%. Our method achieved good improvement in the AP value relative to RetinaNet due to the acquisition of texture information of the target, effective feature fusion and classification detection. The AP value of A-RetinaNet was 16.32% higher than that of the original RetinaNet, and even 2.47% higher than the best YOLOv5 value.

In addition, we also present the parameters of recall, precision, F1 score and detection speed for different models, as seen in Table 4. Regarding precision, RetinaNet, CenterNet and A-RetinaNet showed high precision, and Faster R-CNN had the worst detection. For the recall rate, YOLOv5 and A-RetinaNet had an over 80% recall rate; and, RetinaNet had the worst recall rate. Relative to RetinaNet, our recall was improved by 49.34% because we specifically set up adaptive anchor boxes that match the dataset. The F1 score is an evaluation metric to balance the precision and recall. Regarding the F1 score, YOLOv5 and A-RetinaNet detected the infrared-based dim-small targets optimally while ensuring detection accuracy.

In terms of detection speed, it can be seen in Table 4 that the detection speed of CenterNet was significantly faster than that of other models at 54.4 FPS; this is due to the direct use of a heat map to predict the centroid of the object without the need of Non-Maximum Suppression (NMS) for screening; but, its detection was slightly inferior to that of A-RetinaNet. The speed of SSD can be ranked second with 48.91 FPS, but, when the target was submerged in the background, the target was not detected. The detection speed of the RetinaNet model was ranked third at 35.34 FPS; it achieved real-time detection performance, but the detection performance was poor. Since Faster R-CNN is a two-stage detector, the detection process is complicated and the detection speed was only 11.73 FPS, which cannot meet the demands of real-time detection.

In Figure 12, we show example plots of the detection results of our model and six other models in different scenarios; our proposed method had better detection and a high confidence level. Specifically, the CenterNet and SSD networks could not detect the target when it was submerged in the ground background; alternatively, CenterNet only showed partial miss detection and the SSD network used shallow features to detect small targets with inadequate feature extraction, resulting in difficulty in detecting small targets. In some complex scenes, Faster R-CNN could successfully detect the drone target, but it also mistakenly detected the highlighted noise information in the background as the target, which is due to Faster R-CNN's multiple downsampling and resulted in the loss of some information about small targets; and, the detection performance was not very good. Since mosaic enhancement is used in YOLOv5 to improve the detection performance for small targets, YOLOv5 and A-RetinaNet could successfully detect targets in any scenario, but, in terms of confidence, the performance of YOLOv5 was second only to our model. When two targets were very close in the image, both our model and YOLOX could only partially detect the targets. The main reason is that the targets are too small, occupying a few pixels in a 256 × 256 image, and that the two targets are too close together to remove the box with the low confidence score from the subsequent filter box. The detector is not good at detecting multiple dim-small drone targets in infrared images that are close in distance. In addition, although YOLOv7 performed well in terms of both confidence scores, recall and detection speed, it sometimes detected the background information of drone targets. Our proposed method can avoid the false detection.

Figure 13 shows a comparison of the precision-recall curves for our model and several other models. The area under the precision-recall curve for the A-RetinaNet model was always larger than that of the other models, indicating a better overall performance of our approach. The possible reasons are summarized as follows: 1) Unlike the RetinaNet model, we fundamentally solve the deficiency of the feature pyramid in the fusion stage. We fuse the feature layers of different depths in the network via the AAFM, as well as fuse the contextual semantic and position information for each level of the feature layers; thus, the feature pyramid contains as much feature information as possible and can focus on the infrared detection of dim-small drone targets. 2) Unlike the RetinaNet model, we generate adaptive anchor boxes, which effectively solves the problem of mismatching the anchor boxes size with the bounding boxes and improves the model's detection performance. 3) Unlike other models, the SRTE module reduces the effects of background noise and increases the texture content information of the target through content, texture extractor and sub-pixel convolution, allowing the network to better distinguish the background from the target.

Finally, we compare the Receiver Operating Characteristic (ROC) curves among four selected methods in Figure 14. It can be seen that the proposed A-RetinaNet model performed the best, showing the effectiveness of the proposed modules. Another interesting point is that, although YOLOv7 performed worse than CenterNet in terms of AP and precision (Tables 3 and 4), it performed better than CenterNet in terms of the ROC (Figure 14). To the best of our understanding, the reason behind this is that precision reflects the accuracy of the classifier in terms of its ability to predict positive samples without taking into account whether there are missed positive samples; and, the AP reflects the ability to cover positive samples, while the ROC reflects the trade-off between the classifier's ability to cover positive and negative samples. It shows that YOLOv7 balances the detection ability and the target integrity.

Table 3. Comparison of the AP of different detection models.

Model	Faster R-CNN	SSD	YOLOv5	RetinaNet	CenterNet	YOLOX	YOLOv7	A-RetinaNet
AP	49.64%	91.85%	92.96%	79.11%	92.95%	93.25%	86.30%	**95.43%**

Table 4. Comparison of precision, recall and detection speed of different detection models.

Model	Precision	Recall	F1	FPS
Faster R-CNN	35.19%	61.96%	0.44	11.73
SSD	94.46%	64.49%	0.76	48.91
YOLOv5	92.88%	**85.23%**	0.88	21.06
RetinaNet	98.22%	31.26%	0.47	35.34
CenterNet	**98.40%**	62.73%	0.76	**54.40**
YOLOX	97.79%	70.24%	0.82	29.91
YOLOv7	84.51%	83.24%	0.84	30.52
A-RetinaNet	97.27%	80.60%	**0.88**	24.58

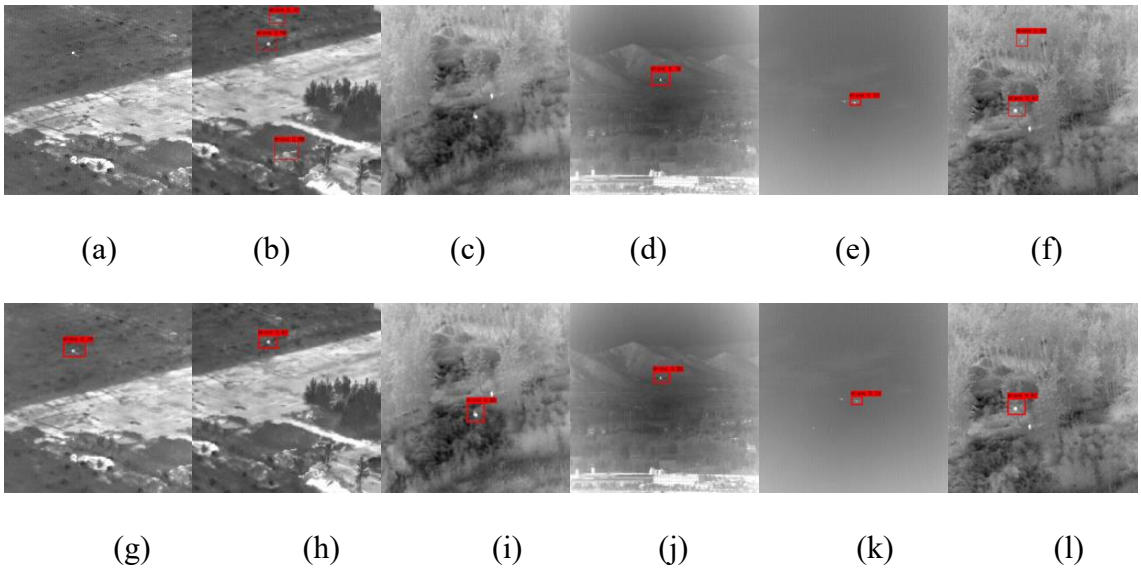


Figure 12. Detection performance in different scenes. (a)–(f) CenterNet, Faster-RCNN, SSD, YOLOv5, YOLOX and YOLOv7; (g)–(l) A-RetinaNet.

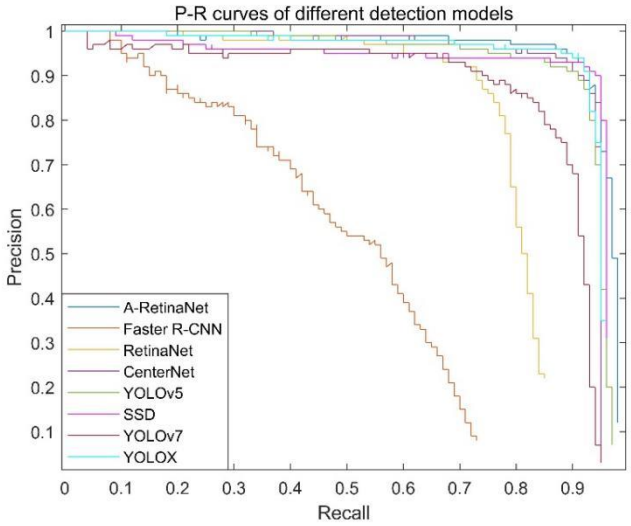


Figure 13. Precision-recall curves for different detection models.

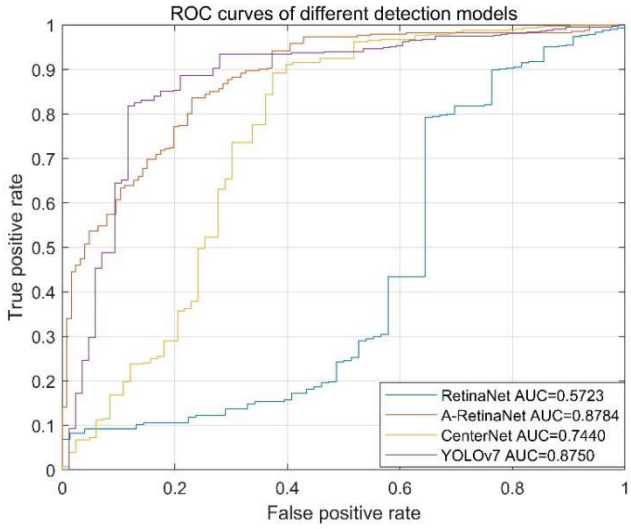


Figure 14. ROC curves for different detection models.

5. Conclusions

In this paper, we have proposed a new infrared-purposed dim and small drone target detection method, called A-RetinaNet, to counter the lack of texture information and poor resolution of dim and small targets in infrared images. A-RetinaNet is composed of SRTE, AAFM, AD and adaptive anchor box modules. For small targets lacking texture and poor feature resolution problems, SRTE is used to extract the target texture information and generate super-resolution images at the same time. For the problem of difficulty in detecting targets submerged in the background, we use AAFM to improve the feature representation of infrared images by integrating contextual information in a bidirectional manner while taking into account semantic information in the deep network and position information in the shallow network. Finally, the global average pooling layer is used to effectively identify small targets and improve detection accuracy. In addition, the adaptive anchor box with a k-means clustering algorithm has been introduced to significantly improve the target recall and detection accuracy. Our model was trained on an infrared-based drone dataset and compared with other state-of-the-art models.

Numerous experiments show that our proposed improvements yielded good results, especially for targets in complex scenes. The proposed method achieves a good balance between detection accuracy and detection speed. Moreover, our method outperformed other existing mainstream methods in terms of detection accuracy, and it can provide a good solution not only for infrared-based dim and small target detection, but also for other small target detection tasks.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant no. 62271303) and Shanghai Pujiang Program (Grant no. 22PJD029).

Conflicts of interest

The authors declare that there is no conflict of interest.

References

1. X. Shao, H. Fan, G. Lu, J. Xu, An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system, *Infrared Phys. Technol.*, **55** (2012), 403–408. https://doi.org/10.1016/j.infrared.2012.06.001
2. M. Fan, S. Tian, K. Liu, J. Zhao, Y. Li, Infrared small target detection based on region proposal and CNN classifier, *Signal, Image Video Process.*, **15** (2021), 1927–1936. https://doi.org/10.1007/s11760-021-01936-z
3. Y. Zhang, L. Zheng, Y. Zhang, Small infrared target detection via a Mexican-Hat distribution, *Appl. Sci.*, **9** (2019). https://doi.org/10.3390/app9245570
4. X. Cao, C. Rong, X. Bai, Infrared small target detection based on derivative dissimilarity measure, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **12** (2019), 3101–3116. https://doi.org/10.1109/JSTARS.2019.2920327
5. C. Chen, H. Li, Y. Wei, T. Xia, Y. Tang, A local contrast method for small infrared target detection, *IEEE Trans. Geosci. Remote Sens.*, **52** (2014), 574–581. https://doi.org/10.1109/TGRS.2013.2242477
6. J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, Y. Fang, A robust infrared small target detection algorithm based on human visual system, *IEEE Geosci. Remote Sens. Lett.*, **11** (2014), 2168–2172. https://doi.org/10.1109/LGRS.2014.2323236
7. Y. Qin, B. Li, Effective infrared small target detection utilizing a novel local contrast method, *IEEE Geosci. Remote Sens. Lett.*, **13** (2016), 1890–1894. https://doi.org/10.1109/LGRS.2016.2616416
8. L. Wu, Y. Ma, F. Fan, M. Wu, J. Huang, A double-neighborhood gradient method for infrared small target detection, *IEEE Geosci. Remote Sens. Lett.*, **18** (2021), 1476–1480. https://doi.org/10.1109/LGRS.2020.3003267
9. Y. Lu, L. Dong, T. Zhang, W. Xu, A robust detection algorithm for infrared maritime small and dim targets, *Sensors*, **20** (2020). https://doi.org/10.3390/s20041237
10. X. Wang, Z. Peng, D. Kong, Y. He, Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene, *IEEE Trans. Geosci. Remote Sens.*, **55** (2017), 5481–5493. https://doi.org/10.1109/TGRS.2017.2709250

11. L. Dong, B. Wang, M. Zhao, W. Xu, Robust infrared maritime target detection based on visual attention and spatiotemporal filtering, *IEEE Trans. Geosci. Remote Sens.*, **55** (2017), 3037–3050. https://doi.org/10.1109/TGRS.2017.2660879
12. L. Ding, X. Xu, Y. Cao, G. Zhai, F. Yang, L. Qian, Detection and tracking of infrared small target by jointly using SSD and pipeline filter, *Digital Signal Process.*, **110** (2021). https://doi.org/10.1016/j.dsp. 2020.102949
13. Q. Guo, Z. Li, W. Song, W. Fu, Parallel computing based dynamic programming algorithm of track-before-detect, *Symmetry*, **11** (2019). https://doi.org/10.3390/sym11010029
14. Q. Hu, J. Yang, P. Qin, S. Fong, Towards a context-free machine universal grammar (CF-MUG) in natural language processing, *IEEE Access*, **8** (2020). https://doi.org/10.1109/ACCESS.2020.3022674
15. A. Ilioudi, A. Dabiri, B. J. Wolf, B. De Schutter, Deep learning for object detection and segmentation in videos: toward an integration with domain knowledge, *IEEE Access*, **10** (2022), 34562–34576. https://doi.org/10.1109/ACCESS.2022.3162827
16. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 580–587. https://doi.org/10.1109/CVPR.2014.81
17. R. Girshick, Fast R-CNN, in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 7–13. https://doi.org/10.1109/ICCV.2015.169
18. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
19. W. Liu, A. Dragomir, E. Dumitru, S. Christian, R. Scott, C. Fu, et al., SSD: Single shot multibox detector, in *Lecture Notes in Computer Science*, (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
20. J. Redmon, A. Farhadi, YOLOV3: An incremental improvement, *arXiv preprints*, (2018), arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767
21. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in 2*017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2999–3007. https://doi.org/10.1109/ICCV.2017.324
22. M. Ju, J. Luo, P. Zhang, M. He, H. Luo, A simple and efficient network for small target detection, *IEEE Access*, **7** (2019), 85771–85781. https://doi.org/10.1109/ACCESS.2019.2924960
23. H. Liang, J. Yang, M. Shao, FE-RetinaNet: small target detection with parallel multi-scale feature enhancement, *Symmetry*, **13** (2021). https://doi.org/10.3390/sym13060950
24. Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, W. Zhang, RISTDnet: Robust infrared small target detection network, *IEEE Geosci. Remote Sens. Lett.*, **19** (2021), 1–5. https://doi.org/10.1109/LGRS.2021.3050828
25. Y. Li, S. Li, H. Du, L. Chen, D. Zhang, Y. Li. YOLO-ACN: Focusing on small target and occluded object detection, *IEEE Access*, **8** (2020), 227288–227303. https://doi.org/10.1109/ACCESS.2020.3046515
26. Z. Song, Y. Zhang, Y. Liu, K. Yang, M. Sun, MSFYOLO: Feature fusion-based detection for small objects, *IEEE Lat. Am. Trans.*, **20** (2022), 823–830. https://doi.org/10.1109/TLA.2022.9693567
27. K. Wang, S. Li, S. Niu, K. Zhang, Detection of infrared small targets using feature fusion convolutional network, *IEEE Access*, **7** (2019), 146081–146092. https://doi.org/10.1109/ACCESS.2019.2944661

28. M. Ju, J. Luo, G. Liu, H. Luo, ISTDet: An efficient end-to-end neural network for infrared small target detection, *Infrared Phys. Technol.*, **114** (2021). https://doi.org/10.1016/j.infrared.2021.103659
29. S. Yao, Q. Zhu, T. Zhang, W. Cui, P. Yan, Infrared image small-target detection based on improved FCOS and spatio-temporal features, *Electronics*, **11** (2022). https://doi.org/10.3390/electronics11060933
30. X. Tong, B. Sun, J. Wei, Z. Zuo, S. Su, EAAU-Net: Enhanced asymmetric attention U-Net for infrared small target detection, *Remote Sens.*, **13** (2021). https://doi.org/10.3390 /rs13163200
31. Q. Xu, H. Huang, C. Zhou, X. Zhang, Research on real-time infrared image fault detection of substation high-voltage lead connectors based on improved YOLOv3 network, *Electronics*, **10** (2021). https://doi.org/10.3390/electronics10050544
32. C. Deng, M. Wang, L. Liu, Y. Liu, Y. Jiang, Extended feature pyramid network for small object detection, *IEEE Trans. Multimedia*, **24** (2021), 1968–1979. https://doi.org/10.1109/TMM.2021.3074273
33. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243
34. Y. Dai, Y. Wu, F. Zhou, K. Barnard, Asymmetric contextual modulation for infrared small target detection, in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2021), 949–958. https://doi.org/10.1109/WACV48630.2021.00099
35. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11531–11539. https://doi.org/10.1109/CVPR42600. 2020.01155
36. M. Pastorino, G. Moser, S. B. Serpico, J. Zerubia, Fully convolutional and feedforward networks for the semantic segmentation of remotely sensed images, in *2022 IEEE International Conference on Image Processing (ICIP)*, (2022), 1876–1880. https://doi.org/10.1109/ICIP46 576.2022.9897336
37. M. Lin, Q. Chen, S. Yan, Network in network, *arXiv preprints*, (2013), arXiv:1312.4400. https://doi.org/10.48550/arXiv.1312.4400
38. B. Hui, Z. Song, H. Fan, P. Zhong, W. Hu, X. Zhang, et al., A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background, Science Data Bank, (2019). https://doi.org/10.11922/sciencedb
39. X. Zhou, D. Wang, P. Krähenbähl, Objects as points, *arXiv preprint*, (2019), arXiv:1904.07850. http://arxiv.org/abs/1904.07850

AIMS Press

©2023 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)