Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15333-w

SLBAF-Net: Super-Lightweight bimodal adaptive fusion network for UAV detection in low recognition environment

Xiaolong Cheng[1] · **Keke Geng**[1] · **Ziwei Wang**[1] · **Jinhu Wang**[1] · **Yuxiao Sun**[1] · **Pengbo Ding**[1]

Received: 30 November 2022 / Revised: 18 February 2023 / Accepted: 6 April 2023

Abstract
Unmanned aerial vehicle (UAV) detection has significant research value in the field of military and civilian applications. However, the traditional object detection algorithms commonly lack satisfying accuracy and robustness due to the intense illumination changes and extremely small size of UAVs on the remote sensor images with the sky background. This paper proposes a super-lightweight bimodal network SLBAF-Net with the adaptive fusion of visible light and infrared images for UAV detection under complex illumination and weather conditions. To handle complex illumination environments and meeting the low computing requirements of airborne computers, a super-lightweight bimodal UAV detection network inspired by YOLO's network structure is developed. In order to fuse bimodal features more effectively, the bimodal adaptive fusion module (BAFM) is proposed to perform an adaptive fusion of visible and infrared feature maps for the purpose of improving detection robustness in complex environments. To verify the superiority of our method, we build a complex dual-modal UAV dataset and conduct comprehensive comparison experiments with various state-of-art object detection networks. The experimental results show that the proposed SLBAF-Net outperforms other algorithms in terms of detection performance and robustness in harsh environments, with a precision rate of 0.909 and a recall rate of 0.912. Moreover, the SLBAF-Net can meet the real-time requirements of airborne computers, and the network size is only 5.6 MB.

Keywords UAV detection · Bi-modal network · Adaptive fusion · Infrared image · SLBAF-Net

1 Introduction

In recent years, UAVs have been widely used in military and civilian applications [31], such as detection [29], rescue [24], reconnaissance [32] and express delivery [3], due to their small size, flexibility, and low safety risk factor. UAV detection technology

✉ Keke Geng
jsgengke@seu.edu.cn

[1] Department of Mechanical Engineering, Southeast University, Nanjing, China

enables UAVs to perform more complex missions and facilitates area control of drones. Currently, methods for UAV detection have been developed using audio signal analysis, radar data analysis, Radio Frequency (RF) signal analysis, and visual data analysis [20]. A novel machine learning (ML) framework was proposed for drone detection using audio signals in noisy environments [16]. However, the use of audio signals to detect drones is not applicable in noisy environments. A radar detection and tracking method was proposed for detecting UAVs in clutter conditions [14]. But the detection range of the radar is short in snow and haze weather. UAV-YOLO is a small object detection method based on YOLOv3 from the perspective of a special UAV [21]. However, drone detection generally has a sky background and visible light cameras are more affected by luminous light. Using data fusion to detect and locate malicious drones from sound and image information [16], which method is not lightweight enough for use on airborne computers. In short, there are no detection algorithms suitable for detecting UAVs in low-identification environments, and this has major implications for the adoption of drone technology. Therefore, this paper focuses on UAV detection techniques for UAV detection tasks in low-identification environments with noise and complex lighting.

The main challenge of the vision-based UAV detection approach is the high sensitivity of visible light cameras to noise, bad weather, intensive illumination changes, etc. Aiming at solving these problems, we develop a dual-modal network named SLBAF-net. The inputs of the network are visible light images and infrared images. Because the characteristic of infrared image is that it can reduce the interference of the external environment such as sunlight and fog, and it can easily separate the target from the background. Infrared images are often used to detect and identify low-resolution objects [35, 39]. At the same time, visible-light images contain many visible light edges and object details and conforms to human visual characteristics and has better performance for object detection [35]; [5]; [37]. Therefore, the fusion of visible and infrared images is particularly suitable for UAV detection missions in complex environments.

This paper is aiming to improve the robustness and accuracy of UAV detection tasks in complex illumination and weather conditions. For this purpose, we propose a new bimodal network architecture that uses the fusion of visible and infrared features for UAV detection. We propose the bimodal adaptive fusion module to make more reasonable use of features. Finally, we perform abundant comparative experiments on the complex environment dual UAV dataset to illustrate the superiority of our approach.

The main contribution of our work are four-fold: A super-lightweight bimodal adaptive fusion network (SLBAF-Net) is proposed to cope with the UAV detection problem in low-identification environments; we propose a bimodal adaptive fusion module (BAFM) to a more effective fusion of visible-light and infrared features; the complex dual-UAV dataset is established, which include night-time, overexposure, interference target, occlusion complex, and normal environments; we have done extensive experiments on normal and complex dual-UAV datasets respectively to demonstrate the superiority of our network.

The rest arrangement of this paper is organized as follows. Section 2 reviews the relevant work of UAV detection and fusion strategy. In Sect. 3, we first introduce the structure overview of SLBAF-Net and then the structure of each network component is described in detail. The comprehensive experimental results of the SLBAF-Net and corresponding analysis are presented in Sect. 4. Finally, Sect. 5 concludes this work.

2 Related work

2.1 Classical detectors

Classical detectors are mainly divided into two-stage methods and single-stage methods. The two-stage methods first generate region proposals and then classify the samples by convolutional neural networks. Common two-stage object detection algorithms including R-CNN [11], SPP-Net [12], Fast R-CNN [8], Faster R-CNN [18] and R-FCN [6]. In the field of object detection, efficiency and real-time are particularly important, so one-stage approaches have emerged, which extract features directly through the convolutional neural network to predict target classification and location, including the YOLO series network [27], [26], [32], [33]. Although there are many object detection networks, there are less literature related to UAV detection networks for complex environments.

The CNN-SVM method [7] is used to detect small drones in a single moving camera, by first stabilizing the video to detect fast-moving drones. For small target detection with complex backgrounds, the CotYOLO-v3 [38], which adds attention module and replace the up-sampling method with sub-pixel convolution to YOLOv3, was proposed. The fused RetinaNet detector [1] was proposed to small targets in aerial images, and replace the FPN structure with a new fusion module to improve the semantic information of low-level graph features. The above-mentioned UAV detection algorithms are based on visible light cameras, which are influenced by light and are not suitable for UAV detection tasks with the sky as the background. Infrared cameras have huge advantages over visible light in nighttime, low-visibility urban environments, illustrating the feasibility of infrared cameras in the field of UAV detection [2]. Some scholars also used a UAV to collect visible-light and thermal infrared images, and then built a rice lodging recognition model based on hybrid image analysis. And the results showed that combining visible-light and thermal infrared image features could significantly improve the recognition accuracy of rice lodging [19]. The use of a single visible light camera makes it difficult to cope with UAV inspection tasks in complex environments. The application of infrared cameras can compensate for the disadvantage of visible light. Therefore, this paper revolves around a bimodal detection network with an adaptive fusion of visible-light and infrared features.

2.2 Fusion method

With the application field of object detection gradually expanding, traditional single visible-light camera recognition is unable to meet the demand. In recent years, object detection based on multi-sensor fusion has developed rapidly, which has high precision and robustness to face the complex environment. Dual-YOLACT [25] is a bimodal segmentation network, which use RGB images and dense depth maps of Lidar to improve segmentation accuracy and robustness based on YOLACT [4]. MAF-YOLO [37] use RGB images and infrared images for pedestrian detection, which is not affected by the light factor. Multi-sensor fusion can make the detection performance reach a new height. In this paper, we use a visible-light camera and an infrared camera to improve the accuracy of UAV detection in complex environments. Due to the complementarity of visible-light and infrared images, many fusion methods have been proposed. A depth learning framework was constructed to generate a single image containing all the features of infrared and visible-light images [17]. A generative adversative nets method, FusionGAN [23], is proposed to fuse these two types of information using a generative countermeasure network. However, the direct fusion of visible and infrared images can

cause a large amount of interference information, which affects the training effectiveness of the network. Our method uses an adaptive fusion of visible-light and infrared feature maps, and deep feature fusion enables the network to learn more effective features.

2.3 Adaptive weight

Attention is particularly important for improving network performance [15]. Attention mechanism was first applied to the channel dimension to increase the representativeness of the network [12]. The attention mechanism goes through a long development [36] [10], the widely applicable model CBAM is proposed by Woo et al. [34]. However, the above attention mechanisms are applied to unimodal networks. In a dual-modal network, it often happens that one input quality is good and the other is bad. For example, the visible-light image is pure black in the dark, and the infrared image has rich information. In another way, the weight of the infrared image should be as large as possible in the dark environment, which is conducive to object detection. The sharpening mixture of expert fusion [25] is proposed to solve this problem, which can learn the robust kernels from complementary modalities. The MAF-YOLO [37] introduces a dual attention module according to the brightness of the visible-light image to achieve higher accuracy. In this paper, we propose a bimodal adaptive fusion module that automatically assigns weights to dual inputs to improve detection performance in complex environments.

3 Method

3.1 System overview

The overall SLBAF-Net pipeline is shown in Fig. 1, which mainly includes four parts, bimodal adaptive fusion module, backbone, FPN structure, and detection head. The bimodal adaptive fusion module, consisting of an adaptive weighting block and a channel attention block, is proposed to integrate visible light and infrared features more efficiently. The backbone network is to extract deep features on the fused feature map. The FPN (feature pyramid networks) structure is able to pass deep semantic information to the underlying layer, thus obtaining high-resolution and strong semantic features. The feature maps are fed into the detection head to obtain detection results.

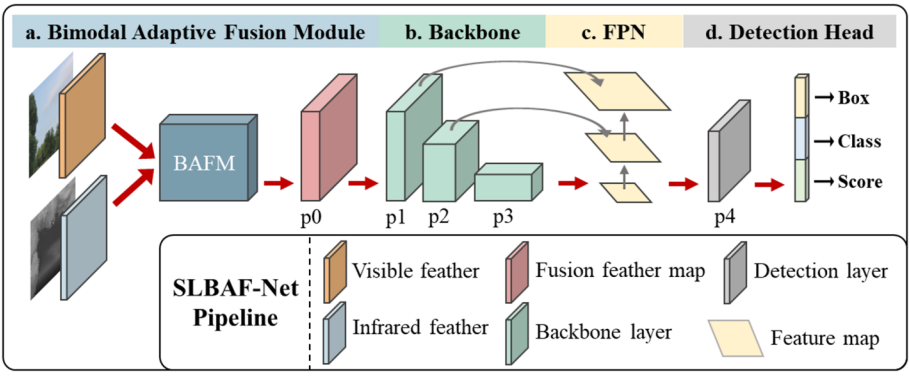

Fig. 1 The system overview of SLBAF-Net

The inputs to the network are two types of images, visible-light and infrared images with a width of 640 pixels, a height of 640 pixels, and three channels (red, green, and blue (RGB)). And the output is the classification target, detection box position, and confidence. A visible light image and an infrared image are first fused by the bimodal adaptive fusion module. Then the fused feature layer p_0 is sent to the backbone for down-sampling, and through the FPN structure for cross-layer fusion. Finally, the feature map p_4 is sent to the detection head, which is used to obtain the detection results.

3.2 Optimization for UAV detection

3.2.1 Fusion strategy

UAV detection with only a visible light camera often fails due to weather or lighting, so our method uses a visible light camera and an infrared camera at the same time. At present, the fusion strategy of dual-dataset can be divided into three kinds, including data-level fusion, result-level fusion, and feature-level fusion.

Data-level fusion is the simplest method, and it does not require modification of the network structure. A network with data-level fusion has one input and one output. It fuses the information of the dual data into one data before entering the network, but it may confuse the information if there is a huge difference between the dual data. The result-level fusion needs two networks, which have the largest amount of computation among the three strategies. Due to this key reason, it is not suitable to run on an onboard computer. The feature-level fusion takes advantage of both approaches. It can not only integrate the information of the dual-data but also ensure the light weight of the model. The network of feature-level fusion has two inputs and one output. The feature-level fusion strategy is most suitable for our purposes.

3.2.2 Network structure

Our method focuses on small UAV detection, the network structure shown in Fig. 2. The SLBAF-Net has a total of 17 layers, and to avoid the loss of small target information, the whole network adopts $8\times$ downsampling.

The bimodal adaptative fusion module includes an adaptative weight block and a channel attention block, which has a detailed introduction in later chapters. The Backbone network is a lightweight improvement based on the YOLOv5 backbone network, which includes the CONV module, C3 module, and SPPF module. The CONV module as the most basic module includes the convolution layer, batch normalization layer, and SILU activation function. The function of the C3 module is to learn the residual characteristics. The SPPF module is space pyramid pooling which can obtain features from different scales. The FPN structure fuses the features of the deep feature map and the shallow feature map when performing upsampling operations to obtain a strong semantic high-resolution feature map.

Regarding network structure, the visible light and infrared images go through the bimodal adaptive fusion module for fusion operations, which can extract more useful features. And through a large number of experiments, we have concluded that the earlier the fusion, the more effective the detection of small targets. Thus, the visible and infrared feature layers are adaptively fused by BAFM into a feature map with 640*640*16 size. Then, the feature maps are downsampled to 80*80*128 size by the backbone network. Finally, the feature map obtains semantic features and fusion features through the FPN structure. Focus on the detection of aerial UAVs, we use two detection feature maps with 160*160*64 size

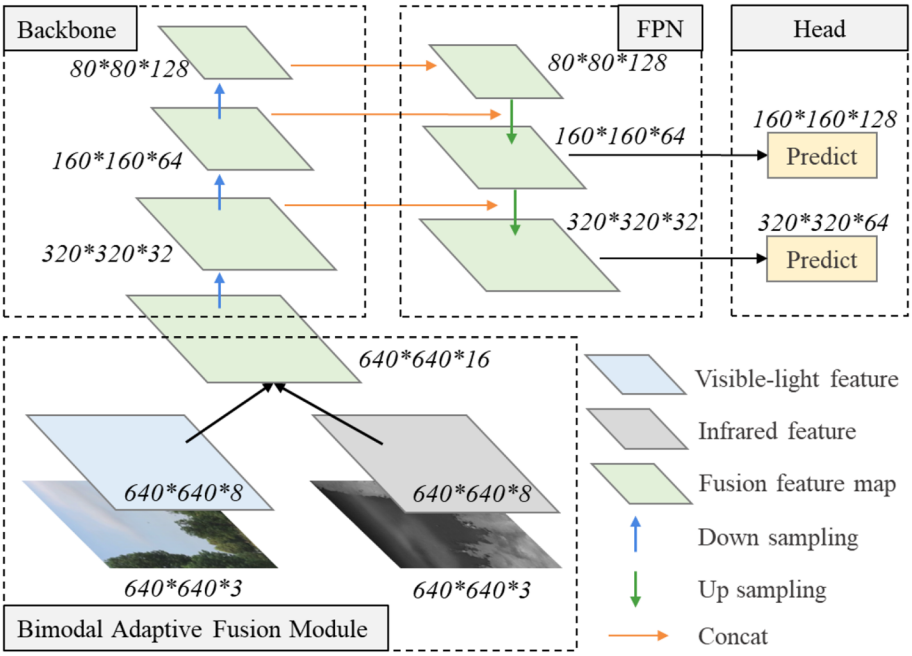

Fig. 2 The SLBAF-Net structure. The 640*640*3 means the feather map with a width of 640 pixels, a height of 640 pixels, and three channels

and 320*320*32 size respectively. Compared to the YOLO series network, our network structure is super-lightweight and more suitable for the detection of small objects.

3.2.3 Loss

The general detection network considers three errors, location error, confidence error, and classification error. Our method is used to detect one class so that we don't consider classification errors. The location error of YOLOv5 [9] is CIOU [42], [40] developed the EIOU on the CIOU. The EIOU loss function includes three parts, overlapping loss, center distance loss, and loss of width and height. The first two parts are the same as the CIOU, but the width and height loss directly minimizes the difference between the width and height of the target box and the anchor box, which makes the convergence speed faster. The loss function of our method is:

$$L_{EIOU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{C_w^2} + \frac{\rho^2(h, h^{gt})}{C_h^2} \tag{1}$$

$$L_{obj}(p_o, p_{iou}) = BCE_{obj}^{sig}(p_o, p_{iou}; w_{obj}) \tag{2}$$

$$L(x_p, x_{gt}) = \sum_{k=0}^{K} \left[\alpha_{box} \sum_{i=0}^{S^2} \sum_{j=0}^{B} M_{kij}^{obj} L_{EIOU} + \alpha_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} M_{kij}^{obj} L_{obj} \right] \tag{3}$$

where L_{EIOU} denotes an effective bounding box regression loss; ρ^2 denotes the Euclidean distance between the midpoint of the prediction box and the midpoint of the target box; b denotes the midpoint of the prediction box; b^{gt} indicates the midpoint of the target box; w indicates the width of the prediction box; w^{gt} indicates the width of the target box; h indicates the height of the prediction box; h^{gt} indicates the height of the target box; c indicates the diagonal distance covering the smallest box of the two detection frames; C_w indicates the width of the smallest box covering the two detection boxes; C_h indicates the height of the smallest box covering the two detection boxes; L_{obj} denotes loss of confidence; p_0 denotes the target confidence score in the prediction box; p_{iou} denotes the iou value of the prediction box and the corresponding target box; BCE^{sig}_{obj} denotes the binomial cross-entropy loss function; w_{obj} denotes the weight of the positive sample; K denotes output feature map; S^2 denotes the cell of output feature map; B denotes the prediction anchor boxes in each cell; M^{obj}_{kij} denotes whether the k th output feature map the i th cell and the j th anchor box is a positive sample; the x_p and x_{gt} denote prediction vector and ground-truth vector; the α_{box} and α_{obj} denote the weights of location error and confidence error.

3.3 Bimodal adaptive fusion module

In the process of the convolutional neural network feed-forward, all features pay attention to the same weight. For the dual-input network, sometimes there is one input of good quality and the other of poor quality. For example, visible-light cameras can't see any valid information in the dark, but rich information is still obtained by infrared cameras. Considering the impact of image quality on detection performance, we propose the BAFM (bimodal adaptative fusion module), which includes two blocks, an adaptative weight block and a channel attention block, shown in Fig. 3, and the BAFM operation process is as follows.

Given a visible-light feature map $F_v \in R^{C\times H\times W}$ and an infrared feature map $F_i \in R^{C\times H\times W}$ as inputs. Then, through an adaptive weight block to get the initial weights $w_v, w_i \in R^{1\times 1\times 1}$ and obtain the mixed feature map $F_m \in R^{2C\times H\times W}$. Finally, the mixed feature map through the channel attention block to get the channel weight $M_c = R^{2C\times 1\times 1}$ and the output $F_O \in R^{2C\times H\times W}$. The overall process can be summarized as:

$$\begin{aligned} F_m &= concat\left(w_v F_v, w_i F_i\right) \\ F_O &= M_c(F_m) \otimes F_m \end{aligned} \tag{4}$$

where $\otimes$ denotes element-wise multiplication. The C, H, and W denote the feature map's channels, height, and width.

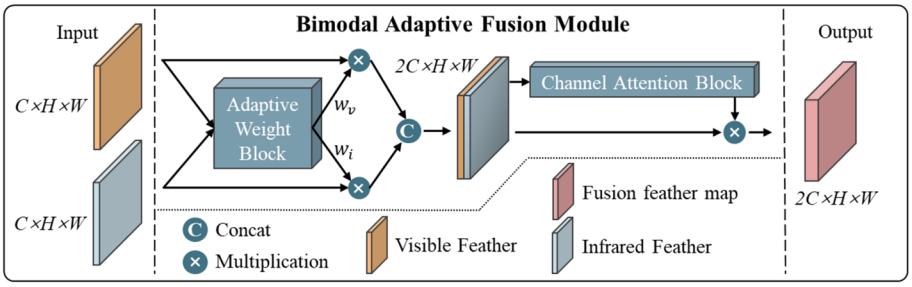

Fig. 3 Bimodal adaptive fusion module. The $C \times H \times W$ indicates the number of channels, the height of feature map, and the weight of feature map

3.3.1 Adaptive weight block

The AW (adaptive weight block) is proposed to obtain the visible-light and infrared weights w_v and w_i. The AW can be divided into three steps. The first step obtains two 2D weight maps $f_{avg} \in R^{1\times H\times W}$ and $f1\times H\times W_{max}$ by average-pooling and max-pooling, see the formula (5). About the expression of spatial information, Hu et al. [13] used average-pooling to compute spatial statistics and Woo et al. [34] argued that the max-pooling can provide as many feature cues as possible, so that our method both use average-pooling and max-pooling. Then, these 2D weight maps are concatenated to form a spatial weight map by concatenation layer and convolved by convolution layer to obtains an efficient feature descriptor $f_d \in R^{1\times H\times W}$. In the second step, a visible light feature score $w_1 \in R^{1\times 1\times 1}$ and an infrared feature score $w_2 \in R^{1\times 1\times 1}$ are obtained by the maximum and average in efficient feature descriptor, see the formula (6). In the end, using the formula (7) obtains a visible-light feature weight $w_v \in R^{1\times 1\times 1}$ and an infrared feature weight $w_i \in R^{1\times 1\times 1}$. The $f(x)$ is a variant of the sigmoid activation function. Through several experiments, when α takes the value of 5, β takes the value of 0.5, the effect is considerable.

$$f_d = Conv(Concat(MaxPool(F), AvgPool(F))) \tag{5}$$

$$w_{1,2} = Sig(Mean(f_d) + Max(f_d)) \tag{6}$$

$$\begin{cases} w_i = w_1/w_2 \\ w_v = f\left(w_i\right) \\ f(x) = \frac{1}{(1+exp(\alpha(x-1)))} + \beta \end{cases} \tag{7}$$

where Conv denotes the convolution layer and the Concat denotes the concatenation layer. The Sig denotes the sigmoid activation function. The overall process of the adaptive weight block is shown in Fig. 4.

3.3.2 Channel attention block

The mixed feature map $F_m \in R^{2C\times H\times W}$ is obtained through an adaptive weight block, which includes visible-light and infrared feature maps. Each feature map is considered a feature detector [41]. We need to give more weight to better detectors, no matter whether it is a visible-light feature or an infrared feature. We refer to the channel attention block in CBAM, which process is shown in Fig. 5.

The input is the mixed feature map $F_m \in R^{2C\times H\times W}$, then use average-pooling and max-pooling generate the different descriptors $d_{avg} \in R^{C\times 1\times 1}$ and $dC\times 1\times 1_{max}$, which forwarded to a shared network and through the sigmoid activation function to generate the channel attention weight $w_c \in R^{C\times 1\times 1}$. Finally, the output feature map $F_O \in R^{2C\times H\times W}$ is obtained:

$$\begin{gathered} w_c = Sig\left(MLP\left(AvgPool\left(F_m\right)\right) + MLP\left(MaxPool\left(F_m\right)\right)\right) \\ F_o = w_c \otimes F_m \end{gathered} \tag{8}$$

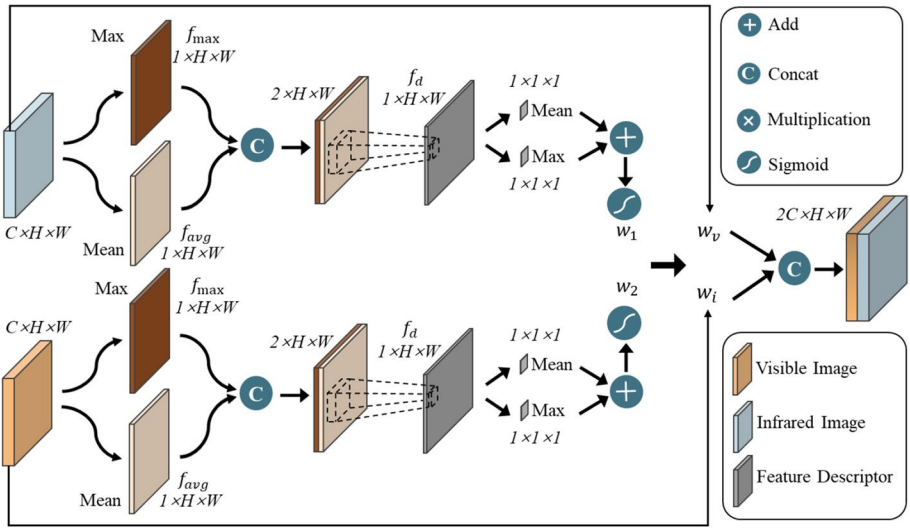

Fig. 4 Adaptive weight block

4 Experiment

4.1 Dataset

4.1.1 Dual-Modal Dataset

Infrared images can distinguish between targets and backgrounds based on differences in thermal radiation and are unaffected by weather and light. Visible images are widely used in the field of visual recognition because of their higher resolution, texture detail and semantic information. However, dual datasets on visible and infrared images are very scarce or not applicable to our study. Due to the dataset in public of dual-UAV is rare, we make the dual-UAV dataset by publicly available UAV dataset [30], which contains 2850 visible light images of different types of drones. We need to produce pseudo-infrared images to compose the dual-dataset. Specifically, the pseudo-infrared images are generated

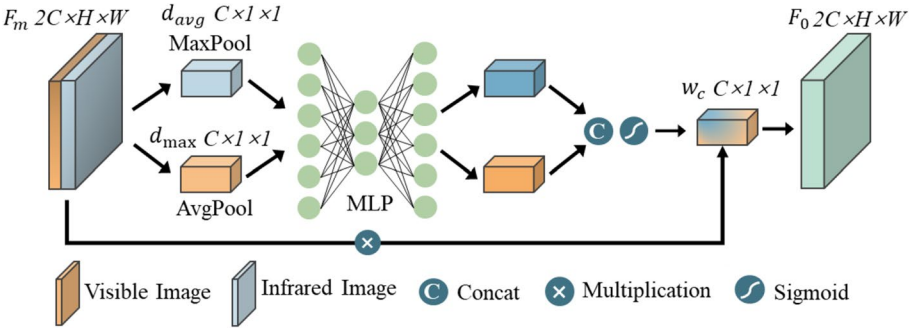

Fig. 5 Channel attention block

by the GAN network trained on visible images. The Pearl-GAN is a GAN network for generating a pseudo-infrared image, for more details about the Pearl-GAN can be found in the reference [22]. The generated dual-UAV dataset is shown in Fig. 6. Finally, we produced a dual modal dataset containing 2850 pairs of visible light and infrared images.

4.1.2 Complex dataset

The ordinary dual drone dataset cannot show the advantages of infrared images, so we also produced complex datasets. To demonstrate the benefits of infrared images, we process visible light images to simulate four complex environments, namely night-time, overexposure, interference targets, and occlusion, as shown in Fig. 7. It can be seen that the quality characteristics of the visible images are substantially reduced on the complex dataset. In the end, the complex dual-UAV dataset includes 2850 pairs of images, including 500 pairs of night-time images, 500 pairs of overexposed images, 250 pairs of interference target images, and 250 pairs of occlusion images.

4.2 Result on dual-dataset

In order to design a super-lightweight and high-performance network, five network structures are designed for comparison shown in Fig. 8a AF0-D16xF: Adaptive fusion at original image size, with images downsampled 16 times by the backbone network, with FPN structure only b AF0-D8xF: Adaptive fusion at original image size, with images downsampled 8 times by the backbone network, with FPN structure only. c AF0-D4xF: Adaptive fusion at original image size, with images downsampled 4 times by the backbone network, with FPN structure only. d AF2x-D8xF: Adaptive fusion is performed after 2 times downsampling, with images downsampled 8 times by the backbone network, with FPN

Fig. 6 Dual-UAV Dataset. **a** The first column is visible-light images from the public dataset. **b** The second column is the pseudo-infrared images generated by the Pearl-GAN

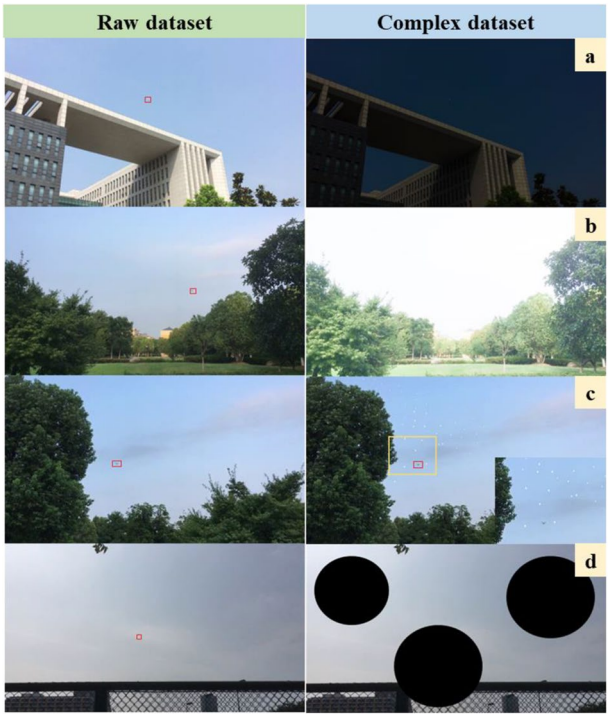

Fig. 7 Complex dual-UAV Dataset. **a** Simulating night-time images **b** Simulating overexposed images **c** Simulating interference target images **d** Simulating occlusion images

structure only. e AF4x-D8xF: Adaptive fusion is performed after 4 times downsampling, with images downsampled 8 times by the backbone network, with FPN structure only. f AF0-D8xFP: Adaptive fusion is performed at the original image size, with the image downsampled 8 times by the backbone network, with FPN&PAN structure.

The a, b, and c networks compare the influence of network depth on small object detection performance. The b, d, and e networks contrast the influence of fusion position on small object detection performance. The b and f networks compare the influence of network structure on small object detection performance. In order to illustrate the superiority of the model structure more forcefully, the six networks and Yolov5s are compared on the normal dual-UAV dataset. Especially, Yolov5s only use visible-light images for training. The training result on a normal dual-UAV dataset is shown in Table 1. The experimental results of the mAP0.5 rate are shown in Fig. 9a, and the loss rate is shown in Fig. 9b.

Through the experiment, the results show that increasing the network depth properly makes a better performance of detection, but the size of networks and the amount of computation is also multiplied. What's more, for small target detection, the earlier fusion, the better performance. Our inference is the down-sampling causes a loss of information on the small target. The detection performance of YOLOv5s for small UAV targets is still outstanding, with map0.5 reaching 0.918. But compared with our network structure, YOLOv5s has a larger network volume and a greater amount of computation. To select the network structure with the best overall performance, we define an evaluation metric λ, see Eq. (9), the higher its value, the better the overall performance of the network. It can be confirmed from the table, AF0-D8xF network structure has a maximum λ of 6.96. Finally,

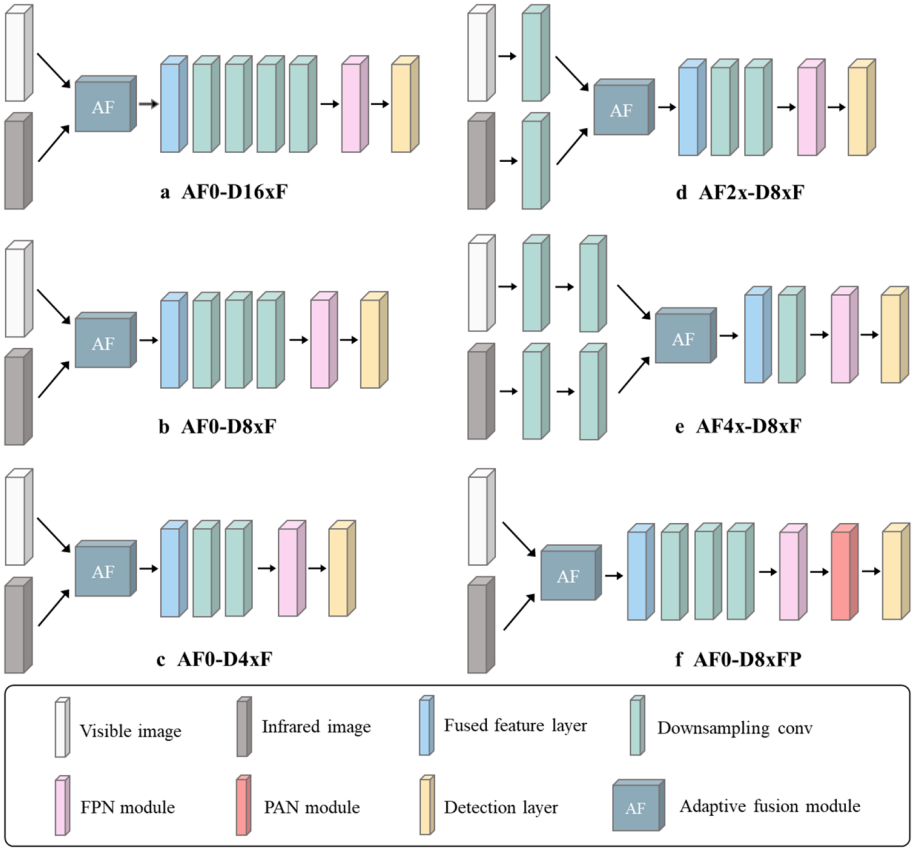

Fig. 8 An illustration of different network structures

we select the AF0-D8xF as the SLBAF-Net structure, which has good detection performance and a small computation volume.

$$\lambda = \frac{(mAP0.5 - min(mAP0.5)) * (max(size) - min(size))}{max(mAP0.5) - min(mAP0.5) * (size - min(size))} + \frac{2 * P * R}{P + R} \tag{9}$$

Table 1 The result of the normal dual-UAV dataset

Structure	map0.5	map0.95	Precision	Recall	Parameter	Size	λ
AF0-D16xF	0.94	0.384	0.939	0.936	11,075,090	22.8	1.94
AF0-D8xF	**0.921**	**0.364**	**0.927**	**0.919**	**2,474,980**	**5.5**	**6.96**
AF0-D4xF	0.821	0.294	0.873	0.798	508,370	2.7	0.83
AF2x-D8xF	0.914	0.364	0.909	0.912	2,608,868	5.8	5.98
AF4x-D8xF	0.854	0.334	0.885	0.87	2,017,042	4.7	3.66
AF0-D8xFP	0.882	0.336	0.892	0.887	2,918,266	6.5	6.87
YOLOv5s	0.918	0.366	0.915	0.918	7,012,822	14.4	2.32

The bold entries indicate our final choice of network structure, which is better adapted to our application scenario

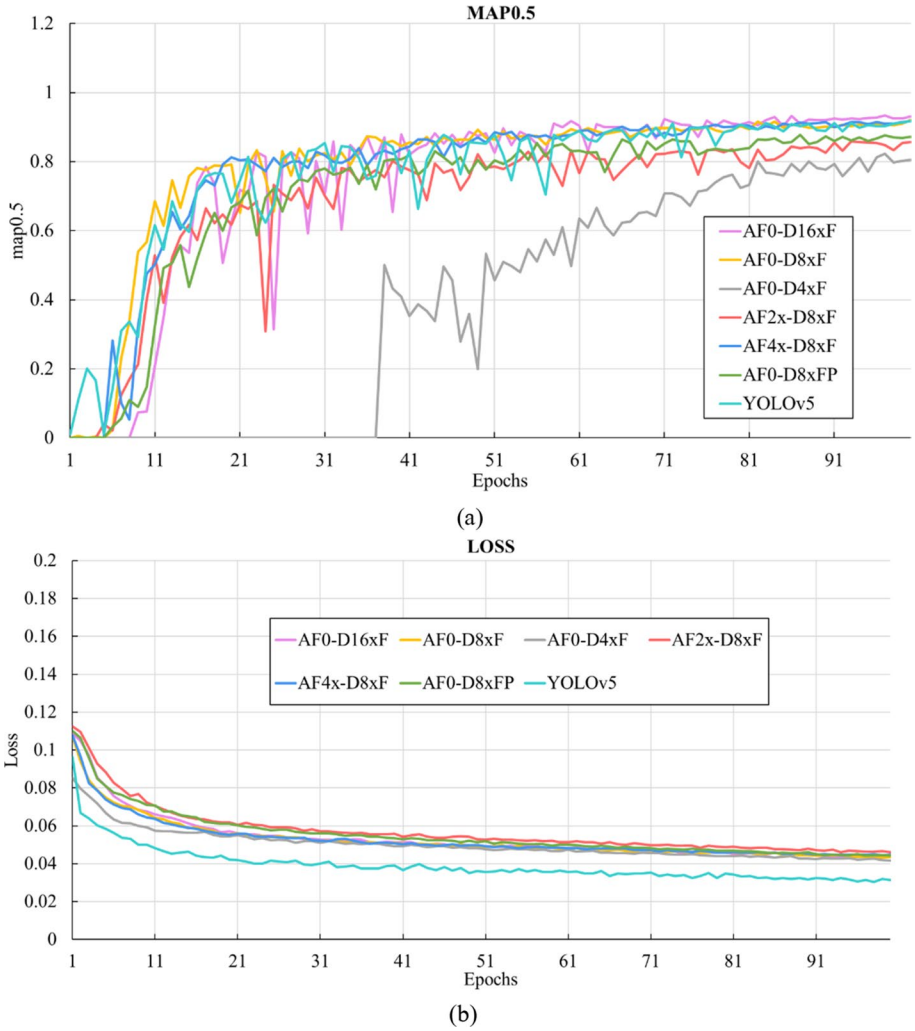

Fig. 9 Training process of mAP0.5 and loss of different networks in dual-UAV dataset

4.3 Result on a complex dataset

4.3.1 Comparison of fusion methods

In the field of UAV detection, stability of detection is particularly essential. For example, reconnaissance UAVs often perform their missions at night. At the same time, because the camera's viewpoint faces the sky, it is often exposed to direct sunlight and complex light transitions that render visible light cameras useless. Our proposed bimodal adaptive fusion module gives appropriate weights to the fused features according to the quality of the visible and infrared images to make the fused features effective. To highlight the effectiveness of our proposed bimodal adaptive fusion module, we conducted experiments on the complex dataset using different adaptive fusion methods. The experimental results are shown in Table 2.

Table 2 The result of the complex dual-UAV dataset

Fusion method	map0.5	map0.95	Precision	Recall	Parameter	Size
CBAM	0.892	0.345	0.905	0.893	2,492,988	5.6
CAM	0.902	0.36	0.906	0.899	2,492,882	5.6
SAM	0.878	0.343	0.882	0.878	2,482,548	5.6
AWM	0.893	0.348	0.902	0.909	2,492,661	5.6
BAFM	**0.907**	**0.35**	**0.915**	**0.901**	**2,492,978**	**5.6**
AWM-SAM	0.869	0.336	0.895	0.874	2,492,864	5.6
AWM-CBAM	0.884	0.329	0.915	0.881	2,549,190	5.8
Pure	0.893	0.355	0.904	0.894	2,472,658	5.5

The bold entries indicate our final choice of adaptive fusion module, which has a superior fusion effect

The CBAM denotes the convolutional block attention module. The CAM denotes the channel attention module. The SAM denotes the spatial attention module. The AWM denotes the adaptive weight module. The BAFM denotes the bimodal adaptive fusion module. The AWM-SAM denotes the adaptive weight module combined with the spatial attention module. The AWM-CBAM denotes the adaptive weight module combined with the convolutional block attention module. The Pure denotes the network has no adaptive convergence module. The experimental results of the mAP0.5 rate are shown in Fig. 10a, and the loss rate is shown in Fig. 10b.

From the experimental results, the SAM's map0.5 is 0.878, which is lower than the module without any weight assignment. And AWM-SAM's map0.5 is lower than AWM's map0.5. It can be seen that the spatial attention module is not suitable for detection on our dataset. We infer that spatial attention is more likely to interfere when detecting tiny objects. What's more, it can be obtained that CBAM plays a minimal role in the complex dataset. According to the experimental results, our proposed BAFM, which is a combination of adaptive weights block and channel attention block, map0.5 can reach 0.907, and the increase in computation is negligible, which is more suitable for bimodal adaptive fusion.

Gradient-weighted Class Activation Mapping (Grad-CAM) obtains the weights by finding the bias derivatives of the category confidence of the network output to the feature map, allowing attention visualization of various visual tasks without discrimination [28]. To illustrate the effectiveness of our method more visually, we used the Grad-CAM algorithm to draw the thermal map shown in Fig. 11. The leftmost column is the visible light images, and the red box in the image is the UAV. The thermal map generated by a suitable network model should be clearly highlighted in the red box area.

As seen by the thermal map, the YOLOv5 network model has a confusing focus in complex environments because only visible-light images are input. The SLBAF-pure, which much better attention to the network model due to the additional input of infrared images, but cannot focus effectively on small targets. Figure 11(d) shows that the SLBAF-pure model focuses more on buildings. By adding the CBAM attention mechanism, the focus of the model improves a little, but it is still difficult to focus on small goals. Using our method BAFM, from the thermal map, the SLBAF model attention has improved considerably, and focusing more on small goals. Contrasting state-of-the-art attention mechanisms CBAM, our method BAFM is applied to dual-input networks, and allowing the network to focus more on small objects.

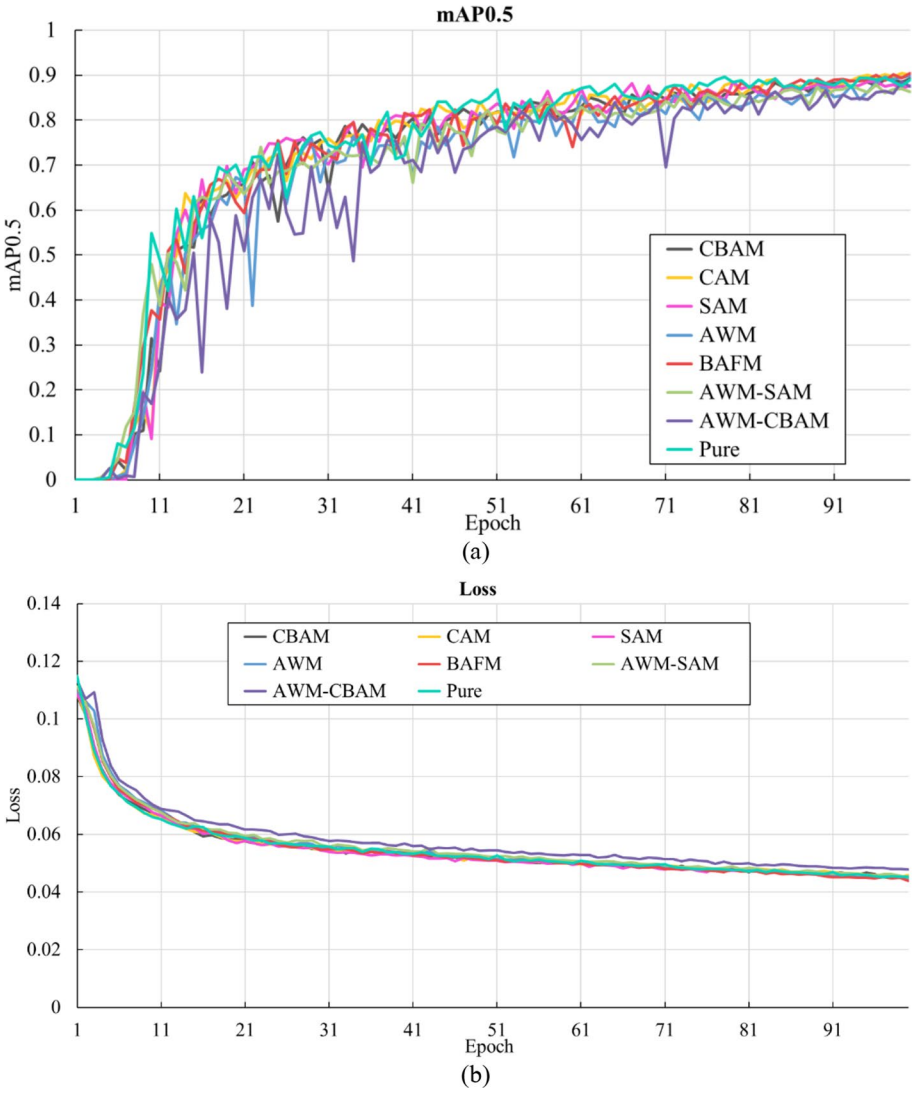

Fig. 10 Training process comparison of mAP0.5 and Loss in the complex dual-UAV dataset

4.3.2 Comparison of networks

To ensure the fairness of the experiment, we trained 100 epochs on the same computer, which is equipped with a Core i7-9700, GeForce GTX 1080Ti, and 16 GB DDR4, for the test experiment. To illustrate the necessity of fusing visible light images, we trained all single-modal networks again on infrared images.

We show in Table 3 the quantitative results of the performance of SLBAF with several other popular unimodal detection algorithms. The training averages of map0.5 and map0.95 for the SLBAF network are 0.907 and 0.35, reaching the highest level among these methods. YOLOv5m is the best-trained unimodal network, with training

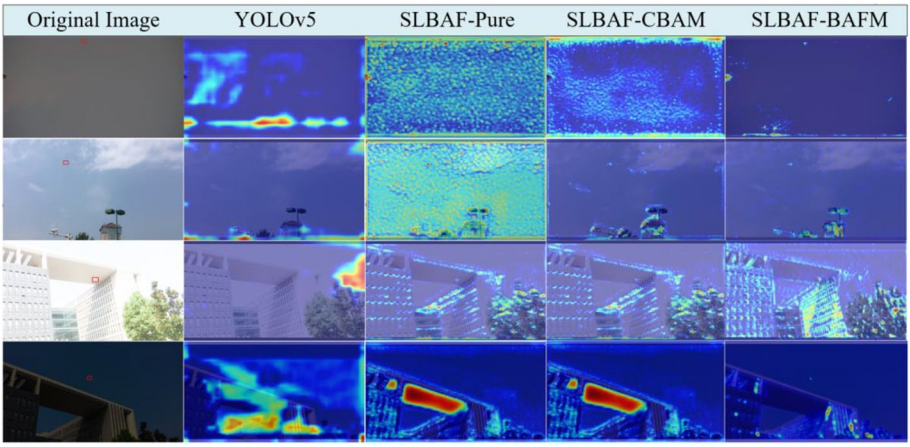

Fig. 11 Grad-CAM thermal Map

results map0.5 and map0.95 on IR images of 0.904 and 0.34 respectively, second only to SLBAF. But the network size of YOLOv5m is 7.5 times larger than our method. For running speed, since SLBAF is a bimodal network, its FPS is not outstanding, but 29.4HZ meets the real-time demand of most scenarios.

Some results from different networks are shown in Fig. 12. We selected five visible light images of different environments and the corresponding infrared images for testing. It can be visualized that our method can detect UAVs accurately in all environments with strong robustness.

Table 3 Results of different networks on the complex dual-UAV dataset

Network	Input	map0.5	map0.95	Precision	Recall	Size (Mb)	FPS
SLBAF	Vis, Inf	**0.907**	**0.35**	**0.915**	0.901	5.6	29.4
YOLOv3	Vis	0.8672	0.3314	0.8922	0.8902	117	45.5
	Inf	0.856	0.308	0.871	0.882		
YOLOv5n	Vis	0.797	0.28	0.852	0.813	**3.8**	**200.0**
	Inf	0.851	0.306	0.884	0.843		
YOLOv5s	Vis	0.798	0.278	0.858	0.766	14.4	166.7
	Inf	0.881	0.327	0.9	0.875		
YOLOv5m	Vis	0.871	0.338	0.9	0.871	42.2	142.9
	Inf	0.904	0.34	0.915	**0.902**		
YOLOv5l	Vis	0.8	0.299	0.883	0.797	92.8	83.3
	Inf	0.885	0.33	0.895	0.889		
YOLOv5x	Vis	0.832	0.309	0.89	0.819	173.1	43.5
	Inf	0.863	0.336	0.877	0.801		
YOLOv7	Vis	0.3745	0.1102	0.5117	0.3563	71.3	12.5
	Inf	0.518	0.155	0.56	0.541		
SSD	Vis	0.7898	0.236	0.878	0.799	95	3.0
	Inf	0.81	0.265	0.82	0.812		

The bold entries indicate the optimal result for each parameter

Fig. 12 Comparison of SLBAF-Net results with other approaches on complex dual-UAV dataset **a** Raw data **b** Night-time data **c** Overexposure data **d** Noise data **e** Occlusion data

5 Conclusion

In this paper, we propose the SLBAF-Net for UAV detection, which is a super-lightweight adaptive dual-modal network. The detection task of UAV usually faces small targets and restricted lighting environments. Our method is able to have a good detection performance

with a low computing resource. Firstly, we built and optimized the network structure, which has dual-input, more lightweight, and more suitable for tiny object detection. Secondly, for better UAV detection in complex illumination environments, we proposed the BAFM, which can obtain adaptive weight distribution according to the quality of visible-light and infrared feature information. Ultimately, we have done rich comparative experiments to prove the advantages of our network. The experimental results show that the SLBAF-Net has excellent accuracy and stable performance for UAV detection in low-recognition, the SLBAF-Net's MAP0.5 score is 0.915. The future research direction of this work is to improve the resolution of infrared images for escaping false detection of small targets.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant no. 52272414 and 51905095).

Data Availability All data, models, or code generated or used during this study are available from the corresponding author by request.

Declarations

Disclosure statement No potential conflict of interest was reported by the author(s).

References

1. Alsanad HR, Sadik AZ, Ucan ON (2022) YOLO-V3 based real-time drone detection algorithm. Multimed Tools Appl 81:26185–26198. https://doi.org/10.1007/s11042-022-12939-4
2. Andraši P, Radišić T, Muštra M, Ivošević J (2017) Night-time detection of uavs using thermal infrared camera. Transp Res Procedia 28(2017):183–190. https://doi.org/10.1016/j.trpro.2017.12.184
3. Bai Z, Feng Q, Qiu Y (2021) Design and Research of UAV For Campus Express Delivery. In 2021 2nd International Conference on Intelligent Design (ICID). 208–213. https://doi.org/10.1109/ICID54526.2021.00048
4. Bolya D, Zhou C, Xiao F, Lee Y J (2019) Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF international conference on computer vision. 9157–9166
5. Cui G, Feng H, Xu Z, Li Q, Chen Y (2015) Detail preserved fusion of visible-light and infrared images using regional saliency extraction and multi-scale image decomposition. Op Commun 341(2015):199–209. https://doi.org/10.1016/j.optcom.2014.12.032
6. Dai J, Li Y, He K, Sun J (2016) R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems, 29
7. Du L, Gao C, Feng Q, Wang C, Liu J (2017) Small UAV detection in videos from a single moving camera. In CCF Chinese Conference on Computer Vision. 187–197
8. Girshick R (2015) Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, 1440–1448
9. Guang J. YOLOv5 release v6.1. https://github.com/ ultralytics/yolov5/releases/tag/v6.1, 2022. 2, 7, 10
10. Guang J, Xi Z (2022) ECAENet: EfficientNet with efficient channel attention for plant species recognition. J Intell Fuzzy Syst 43(4):4023–4035. https://doi.org/10.3233/JIFS-213314
11. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. Proceedings of the IEEE International Conference on Computer Vision (ICCV), FAIR, 2961–2969
12. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
13. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018:7132–7141
14. Hu C, Wang Y, Wang R, Zhang T, Cai J, Liu M (2019) An improved radar detection and tracking method for small UAV under clutter environment. SCIENCE CHINA Inf Sci 62(2):1–3. https://doi.org/10.1007/s11432-018-9598-x

15. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259. https://doi.org/10.1109/34.730558
16. Jamil S, Rahman M, Ullah A, Badnava S, Forsat M, Mirjavadi SS (2020) Malicious UAV detection using integrated audio and visual features for public safety applications. Sensors 20(14):3923. https://doi.org/10.3390/s20143923
17. Li C, Song D, Tong R, Tang M (2019) Illumination-aware faster R-CNN for robust multispectral pedestrian detection. Pattern Recogn 85:161–171. https://doi.org/10.1016/j.patcog.2018.08.005
18. Li H, Wu X J, Kittler J (2018) Infrared and visible image fusion using a deep learning framework. In 2018 24th international conference on pattern recognition (ICPR). 2705–2710
19. Liu T, Li R, Zhong X, Jiang M, Jin X, Zhou P, Guo W (2018) Estimates of rice lodging using indices derived from UAV visible and thermal infrared images. Agric For Meteorol 252:144–154. https://doi.org/10.1016/j.agrformet.2018.01.021
20. Liu B, Luo H (2022) An Improved Yolov5 for Multi-Rotor UAV Detection. Electronics Electronics 11(15):2330. https://doi.org/10.3390/electronics11152330
21. Liu M, Wang X, Zhou A, Fu X, Piao YM (2020) Uav-yolo: Small object detection on unmanned aerial vehicle perspective. Sensors 20(8):2238. https://doi.org/10.3390/s20082238
22. Luo F, Li Y, Zeng G, Peng P, Wang G, Li Y (2022) Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. IEEE Trans Intell Transp Syst 23(9):15808–15823. https://doi.org/10.1109/TITS.2022.3145476
23. Ma J, Yu W, Liang P, Li C, Jiang J (2019) FusionGAN: A generative adversarial network for infrared and visible image fusion. Information fusion 48:11–26. https://doi.org/10.1016/j.inffus.2018.09.004
24. McKenna P, Erskine PD, Lechner AM, Phinn S (2017) Measuring fire severity using UAV imagery in semi-arid central Queensland. Australia Int J Remote Sens 38(14):4244–4264. https://doi.org/10.1080/01431161.2017.1317942
25. Peng P, Geng K, Yin G, Lu Y, Zhuang W, Liu S (2021) Adaptive Multi-modal Fusion Instance Segmentation for CAEVs in Complex Conditions: Dataset, Framework and Verifications. Chin J Mech Eng. 34(1):1–11. https://doi.org/10.1186/s10033-021-00602-2
26. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788
27. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767
28. Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proc IEEE international Conf Comp Vision. 618–626
29. Semsch E, Jakob M, Pavlicek D, Pechoucek M (2020) Autonomous UAV surveillance in complex urban environments. Int Joint Conf Web Intell Intell Agent Technol 2:15–18. https://doi.org/10.1109/WI-IAT.2009.132
30. Sun H, Yang J, Shen J, Liang D, Ning-Zhong L, Zhou H (2020) TIB-Net: Drone detection network with tiny iterative backbone. Ieee Access 8:130697–130707. https://doi.org/10.1109/ACCESS.2020.3009518
31. Thiel C, Schmullius C (2017) Comparison of UAV photograph-based and airborne lidar-based point clouds over forest from a forestry application perspective. Int J Remote Sens 38(8–10):2411–2426. https://doi.org/10.1080/01431161.2016.1225181
32. Verykokou S, Ioannidis C, Athanasiou G, Doulamis N, Amditis A (2018) 3D reconstruction of disaster scenes for urban search and rescue. Multimed Tools Appl 77(8):9691–9717. https://doi.org/10.1007/s11042-017-5450-y
33. Wang C Y, Bochkovskiy A, Liao H Y M (2022) YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696. https://doi.org/10.48550/arXiv.2207.02696
34. Woo S, Park J, Lee J Y, Kweon I S (2018) Cbam: Convolutional block attention module. In Eur Conf Comp Vision (ECCV). 3–19
35. Xin J, Qian J, Shaowen Y, Dongming Z, Rencan N, Jinjin H, Kangjian He (2017) A survey of infrared and visual image fusion methods. Infrared Phys Technol 85:478–501. https://doi.org/10.1016/j.infrared.2017.07.010
36. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In Int Conf Mach Learn 2048–2057
37. Xue Y, Ju Z, Li Y (2021) Zhang W (2021) MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection. Infrared Phys Technol 118:103906. https://doi.org/10.1016/j.infrared.2021.103906
38. Yang Y, Han J (2022) Real-Time object detector based MobileNetV3 for UAV applications. Multimed Tools Appl, 1–17. https://doi.org/10.1007/s11042-022-14196-x

39. Zhang B, Lu X, Pei H, Zhao Y (2015) A fusion algorithm for infrared and visible-light images based on saliency analysis and non-subsampled Shearlet transform. Infrared Phys Technol 73(2015):286–297. https://doi.org/10.1016/j.infrared.2015.10.004
40. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T (2022) Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing 506:146–157. https://doi.org/10.1016/j.neucom.2022.07.042
41. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In Eur Conf Comput Vis. 818–833
42. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020) Distance-IoU loss: Faster and better learning for bounding box regression. Proc Innov Appl Artif Intell 34(7):12993–13000. https://doi.org/10.1609/aaai.v34i07.6999