Unmanned Aerial Vehicle Target Detection Method Based on Combined Infrared and Visible Light

XU Song[1,2], *JI Guipeng*[3], *ZHAO Hongsheng*[4], *YU Tengli*[3], *WANG Ershen*[1,5*], *QU Pingping*[1], *CHEN Yunhao*[6,7], *ZHANG Hongxuan*[1]

1. College of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, P. R. China; 2. State Key Laboratory of Extreme Environment Optoelectronic Dynamic Testing Technology and Instrument, North University of China, Taiyuan 030051, P. R. China; 3. College of Aerospace Engineering, Shenyang Aerospace University, Shenyang 110136, P. R. China; 4. Chinese Aeronautical Radio Electronics Research Institute, Aviation Industry Corporation of China, Shanghai 200233, P. R. China; 5. College of Civil Aviation, Shenyang Aerospace University, Shenyang 110136, P. R. China; 6. College of Electrical and Information Engineering, Yunnan Minzu University, Kunming 650500, P. R. China; 7. Yunnan Key Laboratory of Unmanned Autonomous Systems, Yunnan Minzu University, Kunming 650500, P. R. China

(Received 30 May 2025; revised 11 December 2025; accepted 16 January 2026)

Abstract: Visible light cameras are excellent at capturing subtle features of moving targets in well-wit and stable scenes. However, such cameras may not be able to accurately detect targets when encountering occlusion, fluctuating light intensity, or shadow effects, leading to the occurrence of missed or false alarms. Aiming at the problem of poor anti-interference ability of visible light images in complex scenes, a target detection method based on the combination of visible light and infrared images is proposed. The Canny algorithm is used to preprocess the unmanned aerial vehicle (UAV) infrared image, the seed points are obtained through the Sobel operator, and the image segmentation is performed using the maximum inter-class variance value of the image as the growth criterion in order to locate the UAV region in the infrared image. Then the corresponding region of the visible image is cropped, and UAV target detection is performed in this region. The joint detection method narrows the scope of detection and effectively reduces the interference of factors such as illumination changes and interfering objects on the detection results. Experimental results show that the proposed method achieves 87.6% precision and 75.9% recall, with $\mathrm{mAP}_{0.5}$ and $\mathrm{mAP}_{0.5:0.95}$ values of 83.9% and 52.9%, respectively.

Key words: UAV target detection; infrared image; visible image; image conjunction; YOLOv5 algorithm

CLC number: TP751 **Document code:** A **Article ID:** 1005-1120(2026)03-0400-12

0 Introduction

In the last two decades, infrared and visible light image fusion technology has attracted extensive attention worldwide because this technology can precisely and efficiently achieve target detection in many application scenarios. Su et al.[1] considered the limitations of single vision, proposed a simple and efficient river embankment leakage detection method based on unmanned aerial vehicle (UAV) infrared and visible light image fusion. The application of the proposed methods was demonstrated on a river embankment, and the results showed that the proposed strategy was feasible to identify embankment leakage and it had a great application prospect for embankment leakage detection in flood season. Rao et al.[2] proposed an infrared and visible image fusion algorithm which combines the transformer model with adversarial learning, employed the transformer technique to capture effective global fusion relationships. Additionally, adversarial learning

*Corresponding author, E-mail address: wes2016@sau.edu.cn.

How to cite this article: XU Song, JI Guipeng, ZHAO Hongsheng, et al. Unmanned aerial vehicle target detection method based on combined infrared and visible light[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2026, 43(3):400-411.

http://dx.doi.org/10.16356/j.1005-1120.2026.03.006

was incorporated into the training process to enhance output discrimination by enforcing competitive consistency from the inputs, reflecting the specific features in infrared and visible images. Experimental results demonstrated the effectiveness and usability of the proposed algorithm. Yuan et al.[3] proposed the infrared small drone detection YOLOv5 (IRSDD-YOLOv5) model and designed an infrared small target detection module (IRSTDM) suitable for infrared recognition of small UAVs in the feature extraction stage. In the target prediction stage, a small target prediction head (PH) is also used to complete the prediction of the priori information output through IRSTDM. The loss function is optimized by calculating the distance between the real box and the predicted box to improve the detection performance of the algorithm. Han et al.[4] proposed a progressive feature fusion network (PFF-NET) for detecting small drone targets in infrared images, addressing issues of feature loss in deep convolutional neural network (CNN) layers and insufficient utilization of multi-scale encoding maps. The algorithm integrates a pooling pyramid fusion (PPF) to enhance global contextual priors for high-level features. A feature selection module (FSM) is used to improve the interaction between shallow and deep semantic information, and a lightweight segmentation head is adopted for progressive multi-scale fusion. Experiments on public datasets (SIRST Aug and IRSTD-1k) demonstrated that it has effectively balancing accuracy and efficiency in complex scenarios. Xu et al.[5] proposed a new RetinaNet with an asymmetric attentional fusion mechanism for the detection of weak UAVs in infrared images in response to the problem of texture loss and coarse resolution in the detection of weak UAV targets. The algorithm adopted an asymmetric attention fusion mechanism which constitutes an asymmetric attention fusion pyramid network for cross-layer bi-directional feature fusion, and employed a global average pooling layer to capture global spatially sensitive information, which effectively identified the features and achieved classification. Li et al.[6] introduced an infrared and visible light image fusion method, which was implemented on two platforms equipped with widely used hardware accelerators for embedded vision applications: Zedboard (Advanced reduced instruction set computer(RISC) machine(ARM) and field programmable gate array(FPGA)) and NVIDIA TX1 (ARM and graphics processing unit(GPU)). And their performances were compared. Additionally, extensive experiments were conducted to validate that image fusion enhanced the target detection capability of UAV in various scenarios. Tang et al.[7] proposed a darkness-free infrared and visible image fusion method which enhanced visibility, facilitated complementary information aggregation and designed a scene-illumination disentangled network to remove illumination degradation in nighttime visible images while preserving key features. Additionally, a texture-contrast enhancement fusion network was developed to integrate complementary information and improve contrast and texture details. Hao et al.[8] tackled the issues of energy loss and edge blur in traditional infrared and visible image fusion methods by proposing an enhanced approach based on visual differentiation feature extraction. Experimental results demonstrated that the proposed algorithm not only offered improved visual quality, as assessed both subjectively and objectively, but also preserved the rich brightness information from the source image more effectively. Qiu et al.[9] used a combination of visible and infrared images in a combined strategy for accurate target extraction, where a rough entropy model based on inter-frame differences was used to localize moving targets in infrared images, while a local binary pattern model was used in visible images and a Gaussian mixture model was used in background modeling. Although these improve the accuracy and generality of the algorithm, the information on the visible and infrared images is not fully utilized. Mo et al.[10] addressed the challenges of uneven feature distribution, low repeatability and ambiguous feature matching in infrared and visible image registration for dual-sensor UAV systems by proposing a coarse-to-fine registration framework. Experimental results demonstrated superior performance, with the proposed method achieving low root-mean-square error and maintaining computational efficiency. Li et al.[11] proposed a hybrid method combining non-subsampled contourlet transform (NSCT) and GoogLeNet. The method decomposed infrared and visible images into low-frequency and high-frequency sub-bands. Compared with classical methods, the proposed method had high computational efficiency, balancing visual quality and real-time applicability. Liu et al.[12] proposed a multi-modal image fusion method combining a lightweight modal quality-aware network and an interactive enhancement mechanism. The method utilized infrared and visible images to generate fused images by enhancing high-quality modalities and suppressing low-quality information through spatial and channel attention mechanisms. Compared with traditional single-modality approaches, the proposed method significantly reduced positioning errors and improved feature extraction robustness. This paper addresses challenges in global positioning system (GPS) signal dependency and environmental adaptability for agricultural UAV. In this paper, a method based on combined infrared and visible light for UAV target detection is proposed.

1 Infrared Image Target Candidate Region Extraction

To further improve the efficiency of UAV target detection, this paper investigates a target region extraction method for infrared images. Firstly, noise reduction is performed on the UAV infrared image by median filtering and Gaussian filtering, and the image information is increased or decreased using extreme contrast adaptive histogram equalization to enhance the contrast of the processed image. Then the edge detection is performed using Canny edge detection algorithm. Secondly, seed points are obtained by the Sobel operator. The image segmentation is performed by using the maximum inter-class variance value of the image as a growth criterion to localize the UAV region in the infrared image. Selecting the appropriate target frame size is an important part for accurately locating the UAV region. The size of the target frame significantly influences both the recognition effect and processing speed of the image. The appropriate target frame can be found through experiments and analysis. Finally, images which meet conditions are retained for further processing, while those that do not meet the requirements are discarded. The flow of the target region extraction method of the infrared image in this paper is shown in Fig.1.

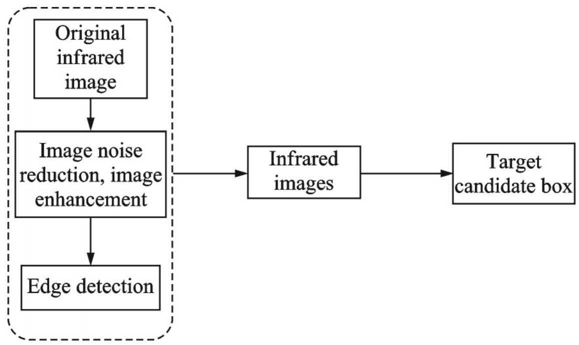

Fig.1 Candidate region extraction of target in infrared image

1.1 Image preprocessing

In this paper, the Gaussian filter described in the literature[13] is chosen to preprocess the image. In image processing, a Gaussian filter is commonly used to deal with images disturbed by Gaussian noise. Gaussian filter is a linear smoothing filter with a normal distribution of probability density function. Gaussian filtering is a process of weighted averaging of the entire image, where the average value of each pixel point is obtained from a weighted average of its own and other pixel point values in the field. The Gaussian filter function is shown as

$$G(u,v)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{-\frac{u^2+v^2}{2\sigma^2}} \tag{1}$$

where (u, v) denotes the pixel coordinates which are integers in image processing, σ the standard deviation. The 2D Gaussian filter is discretized and obtained function values are used as weight coefficients of the Gaussian kernel. For a convolution window of size $(2k+1)\times(2k+1)$, the values of the elements in the Gaussian kernel at position (i, j) are calculated as

$$f(i,j)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{-\frac{(i-k-1)^2+(j-k-1)^2}{2\sigma^2}} \tag{2}$$

1.2 Threshold selection

Typically, the intensity distribution of complex background images varies greatly, and traditional fixed threshold segmentation algorithms are not generalized and universal. Therefore, accurate foreground extraction using effective adaptive threshold segmentation is an important part of target detection algorithms. This paper uses Otsu's method from the literature[14], which is an adaptive threshold selection technique in image processing that is mainly used to segment gray-scale images into foreground and background regions.

Otsu's method was proposed by Japanese scholar Otsu in 1979, and its core idea is to automatically determine the optimal global threshold by calculating the inter-class variance under different thresholds. Setting the threshold value as T, the image is segmented into two parts, the target image C1 and the background image C2, and traversing the whole grayscale interval to determine the appropriate threshold T. This can maximize the gray scale variance of the two parts C1 and C2.

For a given image $(x, y)^{\mathrm{T}}$, assume that the threshold value for segmenting the target and background images is T. The number of pixel points belonging to the target graphic as a proportion of the whole image is ω_0, and its average grayscale is μ_0. The number of pixel points belonging to the background image as a proportion of the whole image is ω_1, and its average grayscale is μ_1. Assume that the size of the image (x, y) is $M \times N$ and that the grayscale value of the image is less than the threshold T. The number of pixel points in the image is denoted as N_0, and the number of pixel points with a gray value which is greater than the threshold T is denoted as N_1.

$$\omega_0 = \frac{N_0}{M \times N} \tag{3}$$

$$\omega_1 = \frac{N_1}{M \times N} \tag{4}$$

$$N_0 + N_1 = M \times N \tag{5}$$

$$\omega_0 + \omega_1 = 1 \tag{6}$$

$$\mu = \omega_0 \mu_0 + \omega_1 \mu_1 \tag{7}$$

$$g = \omega_0 (\mu_0 - \mu)^2 + \omega_1 (\mu_1 - \mu)^2 \tag{8}$$

Substituting μ into g gives

$$g = \omega_0 \omega_1 (\mu_0 - \mu_1)^2 \tag{9}$$

By traversing the entire grayscale interval, find the threshold value that maximizes the inter-class variance g. This is also the threshold value required to segment the image.

1.3 Seed point detection

In this paper, we determine the seed point[15] by calculating the gradient direction and performing non-maximum suppression (NMS) as follows.

Step 1: Calculate the gradient component G in horizontal and vertical directions respectively using Sobel's algorithm[16] and calculate the magnitude and direction of the gradient based on these two components. Convert the gradient angle from radians to degrees and adjust the negative angle values to ensure that all the angle values are between 0 and 180 degrees for subsequent processing with following equations

$$G = \sqrt{G_x^2 + G_y^2} \tag{10}$$

$$\theta = \arctan \frac{G_x}{G_y} \times \frac{180}{\pi} \tag{11}$$

where $G_x = \mathrm{Sobel}_x(I)$ is the horizontal component of Sobel, $G_y = \mathrm{Sobel}_y(I)$ the vertical component of Sobel.

Step 2: Iterate over all the pixels inside the image, determine their two neighboring pixels in the edge direction according to the gradient direction. Compare the gradient value of the current pixel with the gradient values of these two neighboring pixels. If the gradient value of the current pixel is a local maximum, save its gradient value, otherwise set it to zero.

Step 3: Compare the gradient value saved in Step 2 with the maximum inter-class difference value in Section 1.2, and those greater than the threshold are seed points.

2 Infrared and Visible Light Image Joint Detection Algorithm Flow

2.1 Visible light detection

In this paper, we used YOLOv5 as a target detection model for visible images. YOLOv5 stands out for its higher flexibility, faster speed and easier rapid deployment. The overall architecture of the model can be divided into four core parts: The input layer (Input), the backbone network layer (Backbone), the feature fusion layer (Neck), and the prediction output layer (Prediction).

In the input stage, YOLOv5 employs mosaic data enhancement to enrich the diversity of training samples, introduces an adaptive image scaling mechanism to optimize the processing capability for inputs of different sizes, and utilizes adaptive anchor frame computation to dynamically adjust the size and ratio of the detection frames, thus improving the adaptability to targets of different scales. The backbone part of the backbone network integrates the cross stage partial darknet (CSP-Darknet) structure and the spatial pyramid pooling (SPP) module[17], which effectively enhances the feature expression capability through efficient convolution operations such as layer-specific depth-separable convolution and global information extraction. In the design of the feature fusion layer neck, YOLOv5 adopts the combined structure of path aggregation network (PANET) and feature pyramid network (FPN)[18], which realizes the effective fusion and propagation of multi-scale features, and improves the detector's response accuracy to targets of various sizes. The prediction output layer is responsible for generating the final bounding box prediction results, including the location, category probability and other possible attribute predictions for each candidate box, and is optimally trained with a well-designed loss function. Finally, a NMS algorithm is applied to remove redundant predictions to ensure that the output is the most confident detection result.

In general, the innovative YOLOv5 will learn the correlation between the overall and partial features to highlight the UAV features, which can have better results. Swin transformer, which integrates layer normalization (LN) and multi-layer perceptron (MLP) within its blocks, is used as the backbone network of YOLOv5 to extract features. The multi-head self-attention mechanism based on shiftable window can model the dependence between different spatial location features, effectively capture global context information, and have better feature extraction ability, so that the model can better extract image features. The backbone network structure diagram is shown in Fig.2.

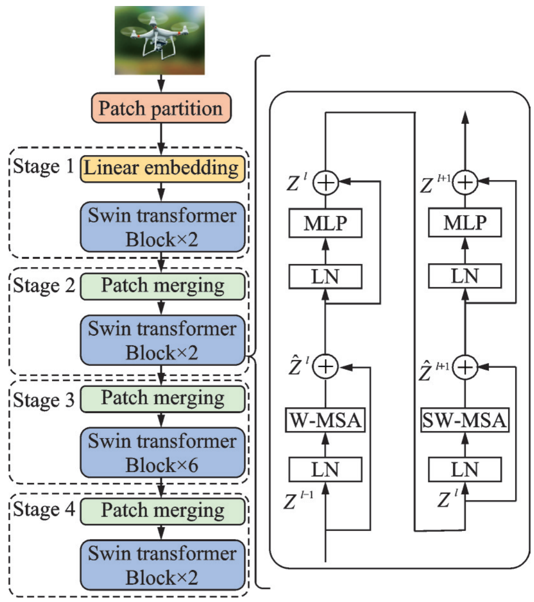

Fig.2 Backbone network structure diagram

2.2 Joint detection methods

Traditional fusion methods mainly consider designing appropriate principles to keep useful information in the source images[19]. Traditional infrared and visible image fusion detection methods are mainly categorized into pixel-level, feature-level and decision-level. These three methods have their advantages and disadvantages. Pixel-level fusion methods are relatively simple, but may lose some subtle features or high-frequency information of the original image, which is more likely to occur in complex scenes. Feature-level fusion methods are more complicated in the process of feature extraction and selection. Incorrect feature selection or matching may reduce the fusion effect, or even introduce misleading information. At the same time, the number of parameters in this method is larger, which may require more computational resources and time. In this study, a joint detection method for infrared and visible images is proposed to address problems of traditional fusion detection methods, which have a large amount of computation and small UAV infrared image datasets.

Infrared images have characteristics of thermal radiation, which can effectively detect the temperature distribution and thermal characteristics of objects, and thus have unique advantages in target identification and localization[20]. The main algorithm flow in this chapter is as follows: Extract the target possible area of the infrared image using the above method, crop the target area of the visible image according to the target area of the infrared image correspondingly, carry out the drone target detection in the cropped visible image, and restore the detected drone image to the visible image according to the cropped coordinates to do the final result display. This can not only reduce the feature extraction area but also avoid the interference of interfering targets and improve the detection accuracy. The specific process schematic is shown in Fig.3. Fig.3 sequentially illustrates the steps performed using infrared and visible light image datasets: Edge detection, candidate region cropping, visible light image cropping, object detection and result restoration. Fig.3 sequentially displays the results obtained by capturing infrared and visible light image datasets, followed by edge detection, candidate region cropping, visible light image cropping, target detection and result restoration.

The infrared image coordinates to the visible light image coordinates via affine transformation. An affine transformation is a single spatial linear transformation, or a combination of multiple distinct spatial linear transformations. The definition of affine transformation is as follows

$$\begin{pmatrix} x_{\mathrm{rgb}} \\ y_{\mathrm{rgb}} \end{pmatrix} = \boldsymbol{A} \cdot \begin{pmatrix} x_{\mathrm{ir}} \\ y_{\mathrm{ir}} \end{pmatrix} + \boldsymbol{B} \tag{12}$$

where $(x_{\mathrm{rgb}}, y_{\mathrm{rgb}})^{\mathrm{T}}$ is visible light image coordinates, $(x_{\mathrm{ir}}, y_{\mathrm{ir}})^{\mathrm{T}}$ infrared image coordinates, Eq.(12) is used for the affine transformation of infrared image coordinates to visible light image coordinates. Matrix $\boldsymbol{A}$ is the linear transformation matrix, responsible for spatial correction between the two images. The two-dimensional vector $\boldsymbol{B}$ is the translation offset, used to compensate for the overall pixel offset of the two images. We selected three corresponding points from annotation boxes of infrared and visible light images as feature points to determine the affine transformation matrix. This approach maximizes the coverage area of points on the image plane while minimizing numerical errors.

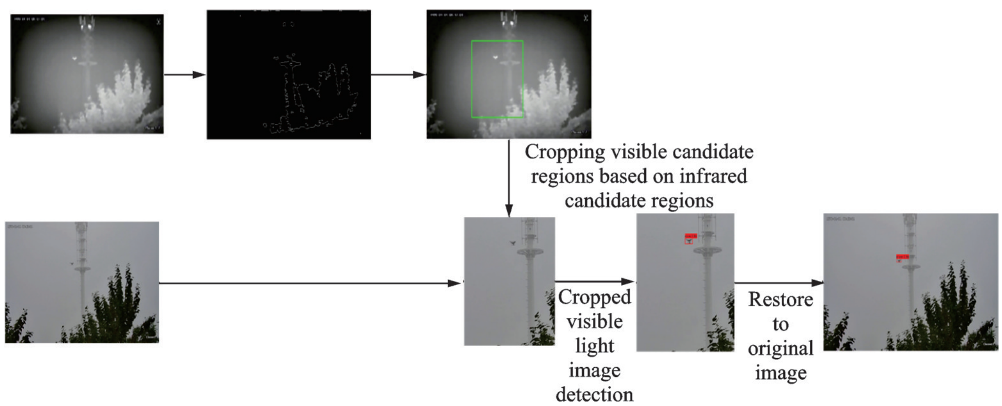

Fig.3 Schematic diagram of the joint detection method of infrared and visible light images

3 Experimental Results and Analysis

The method is trained and tested on a computer equipped with Win11 operating system, Python 3.8, NVIDIA GeForce RTX 3090 Ti GPU, 12th Generation Intel(R) Core(TM) i9-12900KF 3.20 GHz processor and CUDA11.1. To demonstrate the algorithm's advantages in drone detection accuracy, we employed a three-band electro-optical turret to simultaneously capture available infrared and visible light drone datasets, as illustrated in Fig.4. The dataset includes drone images captured during rainy, snowy, foggy and indoor conditions, as illustrated in Fig.5. Table 1 shows that the proposed method achieves an $\mathrm{mAP}_{0.5}$ (This value represents the average precision of all categories when the intersection over union (IoU) threshold is set to 0.5, and is used to measure the overall detection performance of the target detection model) of 83.9% when combined with the YOLOv5 algorithm. Experimental results demonstrate that the proposed joint visible light and infrared detection method outperforms single-modal approaches, facilitating drone detection in complex environments while maintaining high detection accuracy.

Fig.4 Tri-band electro-optical turret

Fig.5 Drone images captured during rainy, snowy, foggy weather and indoors

Table 1 Comparison of UAV detection performances among different YOLO variants and the proposed method

Model	Modality	P/%	R/%	$mAP_{0.5}$/%	$mAP_{0.5:0.95}$/%
YOLOv5n	RGB	74.3	69.5	70.5	41.4
	Infrared	76.2	70.6	71.6	42.9
YOLOv8n	RGB	76.5	70.7	74.3	44.9
	Infrared	75.3	70.1	73.9	44.7
YOLOv10n	RGB	77.8	72.3	76.2	47.1
	Infrared	77.5	72.3	76.1	46.9
YOLOv11n	RGB	81.3	74.5	78.3	49.4
	Infrared	82.4	75.2	79.5	50.2
YOLOv13n	RGB	80.7	73.4	78.1	49.4
	Infrared	82.6	75.0	80.0	50.9
Ours	RGB+Infrared	87.6	75.9	83.9	52.9

To better analyze the effect of the improved methods in this chapter on detection results, this experiment is conducted to detect UAVs under three conditions: Foggy weather, low light, and containing interference targets. Due to the small size of the drone target and the accompanying small detection box, which makes observation difficult, we place a localized enlargement of the detection results in the lower right corner of the image.

The experimental results are presented in Table 1. As shown by the data, the proposed object detection method based on visible and infrared image fusion achieves significant improvements across all key metrics, outperforming algorithms such as YOLOv5, the latest versions of YOLOv11 and YOLOv13. For both red, green, blue (RGB) and infrared images, our method outperforms the aforementioned state-of-the-art algorithms in terms of precision (P), recall (R), $mAP_{0.5}$ and $mAP_{0.5:0.95}$ metrics. $mAP_{0.5:0.95}$ indicates that for the IoU threshold ranging from 0.5 to 0.95 with an interval of 0.05, a total of 10 levels (0.5, 0.55, ⋯, 0.95), the mAP is calculated for each level and then the average is taken to comprehensively evaluate the model's target localization accuracy and detection capability. Specifically, compared to YOLOv5, our method achieves improvements of 13.3%, 6.4%, 13.4% and 11.5% in P, R, $mAP_{0.5}$ and $mAP_{0.5:0.95}$, respectively. Compared to YOLOv11n on RGB images, our method achieves improvements of 6.3%, 1.4%, 5.6%, and 3.5% in P, R, $mAP_{0.5}$ and $mAP_{0.5:0.95}$, respectively. Compared to YOLOv13, our proposed method achieves improvements of 6.9%, 2.5%, 5.8%, and 3.5% in P, R, $mAP_{0.5}$ and $mAP_{0.5:0.95}$, respectively. This demonstrates that by fusing visible and infrared images, the proposed method more effectively leverages the strengths of both modalities. It overcomes limitations inherent to single-modality images in complex scenes, such as lighting variations and occluding objects, significantly enhancing detection accuracy and robustness. Furthermore, despite continuous updates to the YOLO series, the proposed method maintains a substantial performance advantage, fully showcasing its effectiveness and advanced capabilities in multimodal image fusion and object detection.

To verify the improvement effect of the proposed modules on the model's detection performance, ablation experiments are conducted in this study to evaluate the detection performance of different modules. With YOLOv5n as the baseline model and infrared images as detection objects, various improved methods mentioned in this paper are gradually integrated to measure their individual and combined impacts on the UAV target detection performance. The experimental results are shown in Table 2. When no improved modules are adopted, the model exhibits relatively average performance, with P of 76.2%, R of 70.6%, $mAP_{0.5}$ of 71.6% and $mAP_{0.5:0.95}$ of 42.9%. After integrating the IR-guided cropping module (Experiment 2), $mAP_{0.5}$ and $mAP_{0.5:0.95}$ increase to 75.8% and 45.7%, respectively, with slight improvements in P and R. This indicates that the infrared-guided cropping module can effectively enhance the basic detection capability of the model. The introduction of NMS (Experiment 3), infrared pre-processing (Experiment 4), and swin backbone replacement (Experiment 5) all contribute to performance improvement, among which the swin backbone achieves the most significant enhancement ($mAP_{0.5}$ reaches 79.5%). When all four modules are integrated (Experiment 6), the model's performance indicators reach their peak values, with P of 87.6%, R of 75.9%, $mAP_{0.5}$ of 83.9% and $mAP_{0.5:0.95}$ of 52.9%. Compared with the baseline model, $mAP_{0.5}$ increases by 12.3% and $mAP_{0.5:0.95}$ increases by 10.0%, achieving the optimal detection effect.

Table 2 Ablation study experiment

Experimental group	Infrared-guided cropping	NMS	Infrared pre-processing	Swin backbone change	P/%	R/%	$mAP_{0.5}$/%	$mAP_{0.5:0.95}$/%
1	×	×	×	×	76.2	70.6	71.6	42.9
2	√	×	×	×	80.3	72.1	75.8	45.7
3	×	√	×	×	78.5	73.4	74.2	44.2
4	×	×	√	×	79.1	72.8	75.1	45.2
5	×	×	×	√	82.7	74.0	79.5	49.5
6	√	√	√	√	87.6	75.9	83.9	52.9

Note: "×" indicates that the experimental step is not performed, while "√" indicates that the experimental step is carried out.

To accurately evaluate the operational efficiency of the algorithm proposed in this study, a runtime comparison experiment is conducted, with the runtime and related performance indicators presented in Table 3. The experimental data demonstrates that the frame rate of the proposed algorithm reaches 53 FPS, which not only far exceeds the basic real-time requirement for target detection in complex environments (usually requiring ⩾30 FPS) but also exhibits significant advantages in operational efficiency. Moreover, while maintaining the optimal detection accuracy, the proposed algorithm achieves frame rate improvements of 55.9%, 82.8%, and 12.8% respectively compared with three mainstream comparison algorithms, namely BoT-YOLOv7 (34 FPS), cross-feature transformer (CFT) (29 FPS) and iterative cross-attention (ICA) fusion (47 FPS), further highlighting the performance advantages of the proposed algorithm.

Table 3 Runtime comparison

Model	$mAP_{0.5}$/%	$mAP_{0.5:0.95}$/%	Frame rate/ FPS
BoT-YOLOv7	78.4	47.0	34
CFT	79.3	47.8	29
ICA fusion	80.1	48.7	47
Ours	83.9	52.9	53

Fig.6 shows the visible image, infrared image, and processed infrared image of the UAV under foggy weather. Detection results under foggy weather using the method proposed in this chapter are compared with the YOLOv5 algorithm, and the detection results are shown in Fig.7.

In the long-distance flight experiment under foggy weather, the UAV presents a small size in the image, and the fog in the air reduces the clarity of the image, which affects the detection of the tar-

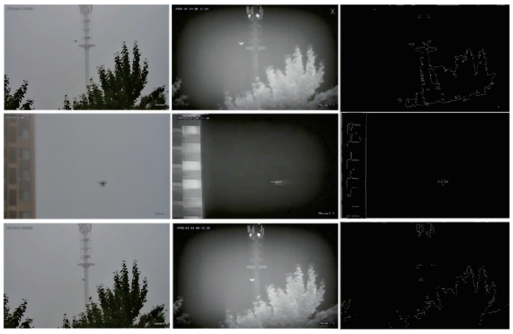

Fig.6 Maps of three states of UAV images in foggy weather

Fig.7 Comparison of YOLOv5 experiments in foggy weather

get greatly, together with the presence of complex interfering factors such as the surrounding buildings. The performance of the UAV detection algorithm is degraded, which is manifested in low detection probability or missed detection phenomenon. In the same experimental scenario, algorithms in this chapter can narrow the detection range through infrared images, which in turn reduce the interference factors during detection. In the same experimental scenario, the detection probability of this chapter's method combined with the YOLOv5 algorithm is improved, when the original algorithm occurs leakage detection, by combining the method in this chapter can accurately detect the drone and reduce the occurrence of leakage detection phenomenon.

Fig.8 shows the UAV visible image, infrared image, and processed infrared image under a low illumination environment. From the figure, it can be seen that UAV visible light images under low-light environments usually contain a lot of noise and low contrast, which directly affects the clarity of target detection.

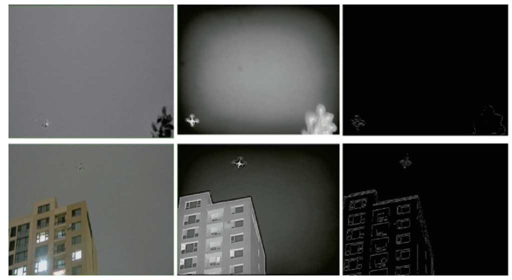

Fig.8 Maps of three states of UAV images in low-light conditions

Fig.9 illustrates the experimental results of UAV object detection under low-light conditions using the original YOLOv5 and the method proposed in this chapter. The experimental results demonstrate that the detection performance of YOLOv5 is somewhat limited, and it even exhibits missed detections under challenging scenarios. In contrast, the proposed algorithm demonstrates stronger robustness. It not only increases the confidence score by 0.01 in scenarios where both models successfully detect the target, but also achieves a confidence score of 0.68 on images where YOLOv5 failed to detect the UAV.

Fig.9 Comparison of YOLOv5 experiments under low-light conditions

Fig.10 shows the comparative effect of the visible image, infrared image and infrared image processed containing interfering targets in the room, by observing it can be found that interfering targets on the poster in the visible image are clear and the surrounding environment is complex, while interfering targets in the infrared image are filtered out.

Fig.10 Three state diagrams of indoor containing interference targets

Fig.11 shows the experimental comparison of YOLOv5 for drone detection in indoor complex environments. As can be seen from Fig.11, the traditional YOLOv5 algorithm incorrectly recognizes interfering targets on posters as drones, whereas the combination of methods in this chapter effectively filters out interfering targets on posters, thus accurately identifies real drones.

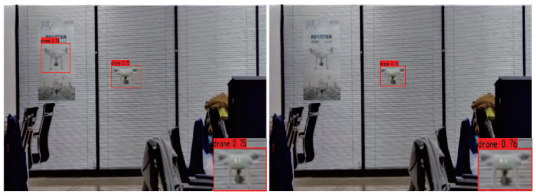

Fig.11 Comparison of YOLOv5 indoor experiments containing interfering targets

Similarly, when there are interference targets such as kites around the UAV when it is flying, it is usually the case that the algorithm mistakenly detects the kites as UAVs, and the method in this chapter can filter out these interference targets through the processing of infrared images, making features of the UAV in the visible image more obvious and identifying real UAVs more accurately.

4 Conclusions

In this paper, a joint detection method of infrared and visible light is proposed for the problem of poor anti-interference ability of visible light images in complex environments. Firstly, advantages of infrared images in UAV detection in complex environments are analyzed. The purpose of narrowing the detection range of UAVs is achieved by extracting the target region of the infrared image. Then, the corresponding region of the visible image is cropped accordingly to match the target candidate region of the infrared image in order to detect the region. By filtering out part of the redundant environmental background, the interference of background noise on target detection is reduced, and the detection accuracy and detection probability are improved. The method in this chapter does not require separate model training for infrared images, which means that the algorithm's demand for computational resources and time cost is reduced, which is very favorable for scenarios with limited resources. Through the joint processing of infrared and visible images, the method can maintain high detection performance under different environmental conditions, especially under bad weather, showing good environmental adaptability.

References

[1] SU H Z, MA J J, ZHOU R L, et al. Detect and identify earth rock embankment leakage based on UAV visible and infrared images[J]. Infrared Physics & Technology, 2022, 122: 104105.

[2] RAO D Y, XU T Y, WU X J. TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network[J]. IEEE Transactions on Image Processing, 2023. DOI: 10.1109/TIP.2023.3273451.

[3] YUAN S D, SUN B, ZUO Z, et al. IRSDD-YOLOv5: Focusing on the infrared detection of small drones[J]. Drones, 2023, 7(6): 393.

[4] HAN Z Q, ZHANG C, FENG H Z, et al. PFF-NET: A fast progressive feature fusion network for detecting drones in infrared images[J]. Drones, 2023, 7(7): 424.

[5] XU Z J, SU J J, HUANG K. A-RetinaNet: A novel RetinaNet with an asymmetric attention fusion mechanism for dim and small drone detection in infrared images[J]. Mathematical Biosciences and Engineering, 2023, 20(4): 6630-6651.

[6] LI J, PENG Y X, JIANG T. Embedded real-time infrared and visible image fusion for UAV surveillance[J]. Journal of Real-Time Image Processing, 2021, 18(6): 2331-2345.

[7] TANG L F, XIANG X Y, ZHANG H, et al. DIVFusion: Darkness-free infrared and visible image fusion[J]. Information Fusion, 2023, 91(C): 477-493.

[8] HAO S, HE T, AN B Y, et al. VDFEFuse: A novel fusion approach to infrared and visible images[J]. Infrared Physics & Technology, 2022, 121: 104048.

[9] QIU S, LUO J S, YANG S, et al. A moving target extraction algorithm based on the fusion of infrared and visible images[J]. Infrared Physics & Technology, 2019, 98: 285-291.

[10] MO Y, KANG X D, ZHANG S, et al. A robust infrared and visible image registration method for dual-sensor UAV system[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5004113.

[11] LI Y Y, WANG C Y, YAO C. Multi-sensors image

fusion via NSCT and GoogLeNet[J]. Transactions of Nanjing University of Aeronautics & Astronautics, 2020, 37(S): 88-94.

[12] LIU X D, LV M B, MA C, et al. Multi-modal image fusion of visible and infrared for precise positioning of UAVs in agricultural fields[J]. Computers and Electronics in Agriculture, 2025, 232: 110024.

[13] BHAGEERATH K E, MARNDI A, HARINI D N D. Performance assessment of Gaussian filter-based image fusion algorithm[C]//Proceedings of the Fourth Congress on Intelligent Systems. Singapore: Springer, 2024: 41-50.

[14] SHAIKH A, SHARMILA B, KRISHNA M, et al. Otsu-based differential evolution method for image segmentation[M]//Intelligent Computing and Communication. Singapore: Springer Nature Singapore, 2023: 693-704.

[15] ZHANG H, FAN Q M, ZHANG Z, et al. Study on blade zoning measurement point planning method[J]. Measurement, 2024, 232: 114670.

[16] KONG W B, CHEN J Z, SONG Y B, et al. Sobel edge detection algorithm with adaptive threshold based on improved genetic algorithm for image processing[J]. International Journal of Advanced Computer Science and Applications, 2023. DOI: 10.14569/IJACSA.2023.0140266.

[17] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[18] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 936-944.

[19] MA J Y, MA Y, LI C. Infrared and visible image fusion methods and applications: A survey[J]. Information Fusion, 2019, 45: 153-178.

[20] WEI H, GU J, REN F, et al. Smart materials for dynamic thermal radiation regulation[J]. Small, 2021, 17(35): 2100446.

Acknowledgements This work was supported by the Open Fund of State key Laboratory of Extreme Environment Optoelectronic Dynamic Testing Technology and Instrument, North University of China (No. 2023-SYSJJ-04), the Aeronautical Science Foundation of China (No.20240055054001), the Open Fund of Key Laboratory of Spatio-temporal Sensing and Intelligent Processing, Ministry of Natural Resources of the People's Republic of China (No.232203), the Open Fund of Key Laboratory of Technology and Equipment of Tianjin Urban Air Transportation System (No. TJKL-UAM-202305), the Open Fund of Key Laboratory of Civil Aviation Flights Wide Area Surveillance and Safety Control technology of Civil Aviation University of China (No. 202105), and the Open Fund of Key Laboratory of Flight Techniques and Flight Safety CAAC (No.FZ2021KF15).

Authors

The first author Mr. **XU Song** received the B.S. degree in electronic information engineering from Dalian Maritime University, China, in 2009, and the Master's degree in information and communication engineering from Shenyang Aerospace University, in 2016. He is currently a laboratory technician in Shenyang Aerospace University. His research interests include computer vision, image processing, target tracking, artificial intelligence and applications to unmanned systems.

The corresponding author Prof. **WANG Ershen** received the Ph.D. degree in communication and information system from Dalian Maritime University, China, in 2009. From January 2014 to December 2016, he worked as a postdoc in Beihang University, China. He is currently a professor in Shenyang Aerospace University, Shenyang, China. His research interests include BeiDou navigation satellite system (BDS)/global navigation satellite system (GNSS) and signal processing, integrity monitoring, integrated navigation, target tracking, artificial intelligence and applications to unmanned systems.

Author contributions Mr. **XU Song** designed the study, conducted the experiments, analyzed the data, and wrote the manuscript. Mr. **JI Guipeng** contributed to data collection and preprocessing of infrared and visible light images. Mr. **ZHAO Hongsheng** provided technical support and validation of the experimental results. Ms. **YU Tengli** assisted in algorithm implementation and performance evaluation. Prof. **WANG Ershen** supervised the research, interpreted the results, and revised the manuscript. Ms. **QU Pingping** contributed to the literature review and data annotation. Mr. **CHEN Yunhao** provided resources and guidance for the multimodal image fusion method. Mr. **ZHANG Hongxuan** assisted in model training and experimental analysis. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: WANG Jie)

基于红外与可见光结合的无人机目标检测方法

徐 嵩[1,2]，纪贵鹏[3]，赵鸿盛[4]，于腾丽[3]，王尔申[1,5]，
曲萍萍[1]，陈云浩[6,7]，张宏轩[1]

（1.沈阳航空航天大学电子信息工程学院,沈阳 110136,中国；2.中北大学极限环境光电动态测试技术与仪器全国重点实验室,太原 030051,中国；3.沈阳航空航天大学航空宇航学院,沈阳 110136,中国；4.中国航空工业集团有限公司中国航空无线电电子研究所,上海 200233,中国；5.沈阳航空航天大学民用航空学院,沈阳 110136,中国；6.云南民族大学电气信息工程学院,昆明 650500,中国；7.云南民族大学云南省无人自主系统重点实验室,昆明 650500,中国）

摘要：针对可见光图像在复杂场景下抗干扰能力不足的问题,如在遮挡、光照强度波动、阴影效应、雾天、低光照及干扰目标等条件下易产生漏检和误检,本文提出一种基于红外与可见光结合的无人机目标检测方法,旨在缩小检测范围、减少背景干扰,提高复杂环境中目标检测的准确率和鲁棒性。该方法利用红外图像可有效反映目标热辐射分布及热特征的优势,对无人机红外图像进行中值滤波、Gaussian滤波和对比度增强等预处理,并采用Canny边缘检测获得边缘信息;通过Sobel算子计算梯度方向并进行非极大值抑制获得种子点,以图像最大类间方差值作为生长准则进行图像分割,定位红外图像中的无人机目标候选区域;进而依据红外图像与可见光图像的对应关系裁剪可见光图像相应区域,使用YOLOv5等目标检测模型在裁剪区域内进行无人机检测,并按照裁剪坐标将检测结果恢复至可见光图像中显示。为验证上述方法的有效性,采用三波段光电吊舱同步获取雨、雪、雾和室内等复杂条件下的红外与可见光无人机数据。测试结果表明,本文方法的Precision、Recall、$\mathrm{mAP}_{0.5}$和$\mathrm{mAP}_{0.5:0.95}$分别达到87.6%、75.9%、83.9%和52.9%,均高于单模态可见光图像和单模态红外图像的检测结果;消融实验表明,集成红外引导裁剪、非极大值抑制(Non-maximum suppression, NMS)算法、红外预处理和Swin骨干网络替换后,$\mathrm{mAP}_{0.5}$较基线提高12.3%,$\mathrm{mAP}_{0.5:0.95}$提高10.0%;运行效率对比显示,该算法帧率达到53 FPS,在保持较高检测精度的同时满足复杂环境目标检测的实时性要求。研究结果表明,红外与可见光联合处理方法能够充分发挥两种模态的互补优势,有效抑制光照变化、遮挡和干扰物等因素对检测结果的影响,降低背景环境干扰,显著提高无人机目标检测的精度和检测概率;同时,该方法无需针对红外图像单独训练模型,减少了计算资源和时间成本,在恶劣天气和资源受限场景下表现出良好的环境适应性。

关键词：无人机目标检测;红外图像;可见光图像;图像结合;YOLOv5算法

研究亮点：

1.提出一种基于红外与可见光图像联合的无人机目标检测方法,通过红外图像提取目标候选区域并裁剪可见光图像对应区域,缩小了检测范围,降低了光照变化和干扰物对检测结果的影响。

2.采用Canny算法对红外图像进行预处理,通过Sobel算子获得种子点,以图像最大类间方差值作为生长准则进行图像分割,实现红外图像中无人机目标候选区域的准确定位。

3.将Swin Transformer作为YOLOv5的骨干网络,利用其基于可移位窗口的多头自注意力机制提取图像全局上下文特征,增强了模型对无人机目标的特征表达能力。

4.该方法在自建多场景无人机数据集上的$\mathrm{mAP}_{0.5}$达到83.9%,较单模态可见光图像和红外图像均有显著提升,且无需对红外图像单独训练模型,降低了计算资源需求。