

Article

UAV Detection in Low-Altitude Scenarios Based on the Fusion of Unaligned Dual-Spectrum Images

Zishuo Huang ¹, Guhao Zhao ^{1,*}, Yarong Wu ¹ and Chuanjin Dai ²

¹ Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an 710051, China; kgdtthzs@163.com (Z.H.)

² Information and Navigation College, Air Force Engineering University, Xi'an 710082, China

* Correspondence: zghlupin@163.com; Tel.: +86-177-9124-2979

Highlights

What are the main findings?

- The SURF algorithm, improved based on the grayscale centroid method, can effectively match the feature points in visible light–infrared paired images.
- An enhanced YOLOv11 model with a modified loss function can achieve robust UAV detection in the fused images.

What are the implications of the main findings?

- The plane-adaptive pixel remapping algorithm can be used to get dual-spectrum fusion image based on the feature point matching result.
- The “Fusion and Detection” methodology enables stable image monitoring of UAV targets in complex environments.

Abstract

The threat posed by unauthorized drones to public airspace has become increasingly critical. To address the challenge of UAV detection in unaligned visible–infrared dual-spectral images, we present a novel framework that comprises two sequential stages: image alignment and object detection. The Speeded-Up Robust Features (SURF) algorithm is applied for feature matching, combined with the gray centroid method to remove mismatched feature points. A plane-adaptive pixel remapping algorithm is further developed to achieve images fusion. In addition, an enhanced YOLOv11 model with a modified loss function is employed to achieve robust object detection in the fused images. Experimental results demonstrate that the proposed method enables precise pixel-level dual-spectrum fusion and reliable UAV detection under diverse and complex conditions.

Keywords: dual-spectrum; feature matching; pixel reconstruction; object detection



Academic Editor: Anastasios Dimou

Received: 17 November 2025

Revised: 25 December 2025

Accepted: 5 January 2026

Published: 7 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the rapid increase in drone usage, unauthorized operations have become more frequent, posing substantial risks to public safety. In airport environments, such activities may interfere with normal aircraft takeoff and landing. For example, in 2024, flight delays at Tianjin Airport were attributed to disruptions caused by unauthorized drones. In sensitive areas, drones equipped with cameras can be exploited for illicit imaging and real-time online video transmission, raising concerns regarding potential espionage [1].

Air surveillance technologies are essential for countering unmanned aerial vehicles (UAV) [2]. Among them, radar-based surveillance remains the most widely employed approach [3]. In 2019, the U.S. Department of Defense introduced the short-range “Spyglass” radar system, which integrates 3D radar with artificial intelligence (AI) to enhance the detection and identification of UAV swarms. However, drones are characterized as “low, slow, and small” targets, making detection technically challenging due to limited radar cross-sections, low-altitude flight, and frequent occlusion [4].

Electro-optical detection is commonly applied as a complementary surveillance technique [5,6]. The “Raven” counter-UAV system, developed by Indra Defense Systems in Spain, combines high-resolution imaging radar with an electro-optical detection subsystem to enable rapid UAV identification and tracking. In 2024, the AI-powered “Sky Eye” surveillance system was deployed in Washington, DC, USA, integrating advanced computer vision algorithms and capable of simultaneously monitoring up to 5000 aerial targets.

During the synchronous acquisition of visible and infrared images by photoelectric equipment, discrepancies in pixel alignment often occur. These misalignments result from several factors, including the difficulty in precisely synchronizing the imaging times of different sensors, challenges in adjusting the lens focal lengths to achieve consistent image scales, and relative angular deviations between optical centers caused by camera shake and mechanical installation errors. As illustrated in Figure 1, the position and size of the unmanned aerial vehicle differ in the dual-band images captured simultaneously.

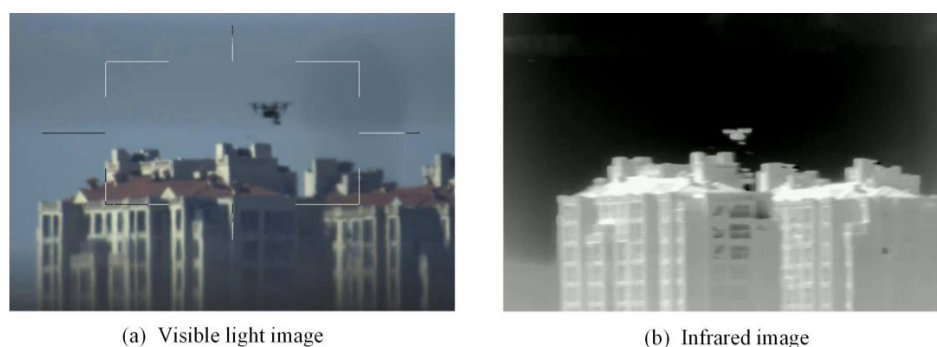


Figure 1. Misaligned visible and infrared image pair.

This study introduces a novel UAV detection framework that employs visible–infrared image fusion under non-aligned conditions in various perspectives and scales. The main contributions are summarized as follows:

- (a) A feature matching and fusion method for non-aligned visible and infrared dual-spectrum images is proposed, applicable to the output processing of diverse dual-spectrum surveillance systems.
- (b) An enhanced YOLOv11-based detection algorithm is developed, designed to operate jointly with the proposed fusion method to achieve accurate UAV identification in fused imagery.

2. Research Status of UAV Detection

2.1. Visible Light-Based Method

This type of research primarily focuses on exploring the extension methods for object detection models. Liu et al. (2024) [7] provided a comprehensive review of the main challenges in drone detection, summarized existing solutions, and evaluated and compared commonly used datasets. Samadzadegan et al. (2022) [8] utilized the YOLOv4 model to detect two types of aerial targets—drones and birds—and implemented strategies to reduce bird-related interference. Yasmine et al. (2023) [9] enhanced the YOLOv7 architecture

by incorporating the CSPResNeXt module, integrating the Transformer module with the C3TR attention mechanism, and adopting a decoupled head structure to improve detection performance. Hu et al. (2023) [10] proposed a modified YOLOv7-tiny-based detection model that optimized anchor box allocation based on the aspect ratios of actual bounding boxes and introduced the hard sample mining loss function (HSM Loss) to enhance the network's focus on challenging samples.

2.2. Infrared-Based Method

The majority of research in this field is concentrated on the detection of infrared dim and small targets. Lu et al. (2022) [11] proposed a UAV detection model based on an enhanced adaptive feature pyramid network (EAFPN) to address the issue of weakened feature representation in cross-scale infrared feature fusion. Zhao et al. (2023) [12] introduced a method using the Isolation Forest (iForest) algorithm. By constructing a global iForest they leveraged the inherent isolation tendency of UAVs, effectively mitigating detection difficulties caused by low signal-to-clutter ratios (SCRs). Xu et al. (2024) [13] developed the RMT-YOLOv9s model, which incorporated an improved multi-scale feature fusion network, RMTELAN-PANet, to effectively capture semantic information across feature maps. Additionally, Fang et al. (2024) [14] proposed the Spatial and Contrast Interactive Super-Resolution Network (SCINet), consisting of two branches: a spatial enhancement branch (SEB) and a contrast enhancement branch (CEB). Yang et al. (2023) [15] drew upon the visual processing mechanism of the human brain to propose an effective framework for UAV detection in low signal-to-noise ratio infrared images. In addition, Pan et al. (2024) [16] proposed the AIMED-Net, which incorporates a variety of feature enhancement techniques based on YOLOv7. Fang et al. (2022) [17] redefined infrared UAV detection as a residual image prediction task, developing an end-to-end U-shaped network that learns to map input images to residual images.

Some other studies focused on the pursuit of lightweight models. Cao et al. (2024) [18] introduced YOLO-TSL, which embeds the triplet attention mechanism into the YOLOv8n backbone and adopts the Slim-Neck architecture, achieving reduced computational complexity while maintaining high detection accuracy. Wang et al. (2024) [19] developed the PHSI-RTDETR model, combining a lightweight feature extraction module (RPCConv-Block), a refined neck structure (SSFF), the HiLo attention mechanism, and the Inner-GIoU loss function to improve detection performance.

2.3. Dual-Spectrum Fusion-Based Method

This type of research usually simultaneously proposes methods for the fusion of dual-spectrum image features and object detection. Wang et al. (2025) [20] proposed RF-YOLO, an object detection method designed for both visible light and infrared images, which addresses the challenges of noise and uncertainty across different modalities by incorporating a dual feature fusion (DFF) module and a feature fusion corrector (FFC) module into the YOLOv5 framework. Gao et al. (2024) [21] proposed EfficientFuse, a multi-level cross-modal fusion network that captures both local dependencies and global contextual information at shallow and deep network levels. They further developed the AFI-YOLO detection framework to effectively reduce background interference in fused imagery. Chang et al. (2025) [22] introduced MPIFusion, an image fusion network that integrates a dual-channel shallow detail fusion module (DC-SDFM) and a deep feature fusion block (DFFB), effectively merging hierarchical features to produce high-quality fused images. Liu et al. (2025) [23] introduced a visible light and infrared image matching method based on a joint local attention mechanism, enabling precise positioning for UAVs. Jiang et al. (2024) [24] converted the cross-modal task into a single-modal framework by

extracting and fusing features from infrared and visible light images through deep learning networks, thereby enabling target alignment for small UAVs. Zhu et al. (2025) [25] proposed the YOLOv9-CAG, an enhanced algorithm that demonstrated strong performance on both infrared and visible light datasets. Suo et al. (2025) [26] proposed a novel detection model, YOLOv9-AAG. The model's detection head incorporates GSConv to efficiently capture edge and contour information, and integrates an Attention-based Feature Fusion (AFF) mechanism to enhance shape feature representation, thereby improving discrimination between birds and drones.

Most existing studies primarily concentrate on the design of image fusion and object detection network architectures, while overlooking the pixel-level coordinate discrepancies between visible light and infrared images, and have not yet incorporated more sophisticated object detection frameworks to address complex real-world scenarios.

3. Research Methods

The methodology flowchart of this paper is illustrated in Figure 2, which outlines the research methodology of “Fusion and Detection”, it also reflecting the thought of “Clarification—Recognition”. The workflow illustrated in Figure 2 can be summarized as follows: Firstly, the input visible-light and infrared images undergo a sharpening process. An improved speeded-up robust features (SURFs) algorithm is then employed to extract, match, and screen feature points from both modalities. Subsequently, pixel remapping and pixel reconstruction are performed to achieve feature-level image fusion. Finally, target detection is conducted on the fused image using an enhanced YOLOv11 model.

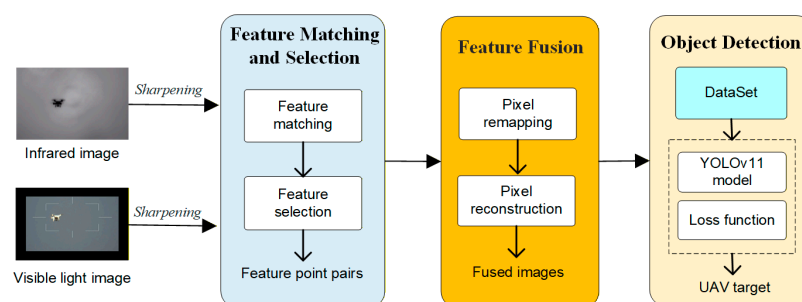


Figure 2. Methodology flowchart.

3.1. Feature Matching and Selection

3.1.1. Image Sharpening

Considering the distinct pixel distribution characteristics between visible light and infrared images, along with their similar edge features, the Laplacian operator is applied to perform differential computations within the pixel neighborhood. This method effectively captures abrupt changes in gray-level intensity and enhances the image's edge features. The calculation is as follows:

$$I_{\text{sharpened}} = I - k \cdot (I * \nabla^2) \quad (1)$$

Here, I denotes the original image, k is the scaling coefficient that controls the sharpening intensity, and ∇^2 represents the Laplacian convolution kernel.

3.1.2. Feature Matching

The SURF algorithm [27] was used to detect and match feature points between visible and infrared images. This approach enables the rapid detection of multi-scale feature points by constructing the Hessian matrix. Meanwhile, it improves computational efficiency through the use of integral images. To generate rotation-invariant descriptors, Haar wavelet

responses are statistically analyzed within the local neighborhoods of the detected feature points. Finally, feature correspondences are established by matching these descriptors via the nearest-neighbor method.

3.1.3. Feature Selection

Due to the inherent differences in the color spaces of visible light and infrared images, mismatches are inevitable. After removing feature point pairs with excessively large spatial distances, the grayscale centroid method is used to compute the grayscale dissimilarity between corresponding feature points. The key principle of this method involves calculating an angular difference. This is done between the centroid vectors of the image regions that surround each feature point:

$$\theta = \arctan2(m_{01}, m_{10}) \quad (2)$$

Here, $\arctan2(\cdot)$ denotes the four-quadrant inverse tangent function, while m_{01} and m_{10} correspond to the first-order moments of the image block. Feature point pairs exhibiting significant grayscale differences are excluded, and the remaining pairs are retained as the final matching results.

3.2. Feature Fusion

3.2.1. Pixel Remapping

The spatial variation in objects in low-altitude surveillance imagery, particularly the differences in scales between visible light and infrared images, presents a challenge for accurate pixel-level mapping. Traditional pixel coordinate mapping relies on homography transformation, which necessitates a large number of control points. To overcome this limitation, this study proposes an alternative approach that assigns each pixel to a distinct plane. By selecting the two closest sets of feature points: A and B , identify their corresponding matching points in the infrared image: A' and B' , as the control points for pixel C . The coordinate calculation method for point C' corresponding to point C in the infrared image is as follows:

$$\begin{cases} C'_x = A'_x + a(C_x - A_x) - b(C_y - A_y) \\ C'_y = A'_y + a(C_y - A_y) + b(C_x - A_x) \end{cases} \quad (3)$$

$$\begin{cases} a = \frac{(B'_x - A'_x)(B_x - A_x) + (B'_y - A'_y)(B_y - A_y)}{(B_x - A_x)^2 + (B_y - A_y)^2} \\ b = \frac{(B'_y - A'_y)(B_x - A_x) + (B'_x - A'_x)(B_y - A_y)}{(B_x - A_x)^2 + (B_y - A_y)^2} \end{cases} \quad (4)$$

The above formulas indicate that the two corresponding pixels in dual-spectrum images share the same relative position as the control points. The control points in planes at varying depth are illustrated in Figure 3, demonstrating the feasibility of precise inter-image pixel matching.

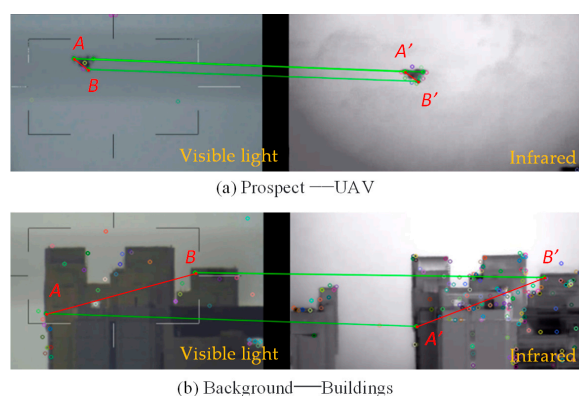


Figure 3. Control points distributed across multiple depth planes.

3.2.2. Pixel Reconstruction

Based on the mapping relationship between visible light and infrared pixel coordinates, and leveraging the “black hot” feature of the drone target in the infrared image, the surrounding region of the target pixel is reconstructed. The calculation procedure is as follows:

$$\text{RGB_Enhanced}(x, y) = [\text{RGB}(x, y) + \varepsilon] \times \text{Infrared}(x', y') / 255 \quad (5)$$

Here, ε denote the regulation parameter. Thus, the thermal values from the infrared image can be mapped onto the visible-light image. Each reconstructed pixel thereby carries combined information from both the visible and infrared spectral bands.

3.3. Object Detection

3.3.1. Model Structure

YOLOv11 [28], released in 2024, improves upon YOLOv10 [29] by addressing issues related to multi-scale object detection, small-object omissions, and bounding box overlap in complex scenes. The overall architecture is illustrated in Figure 4, which consists of three parts: Backbone, Neck, and Head. Its main innovation lies in the incorporation of the Bottleneck Block, SPPF, and C2PSA modules, which jointly strengthen feature extraction. We employ YOLOv11, incorporating an optimized loss function, as the drone target detection model.

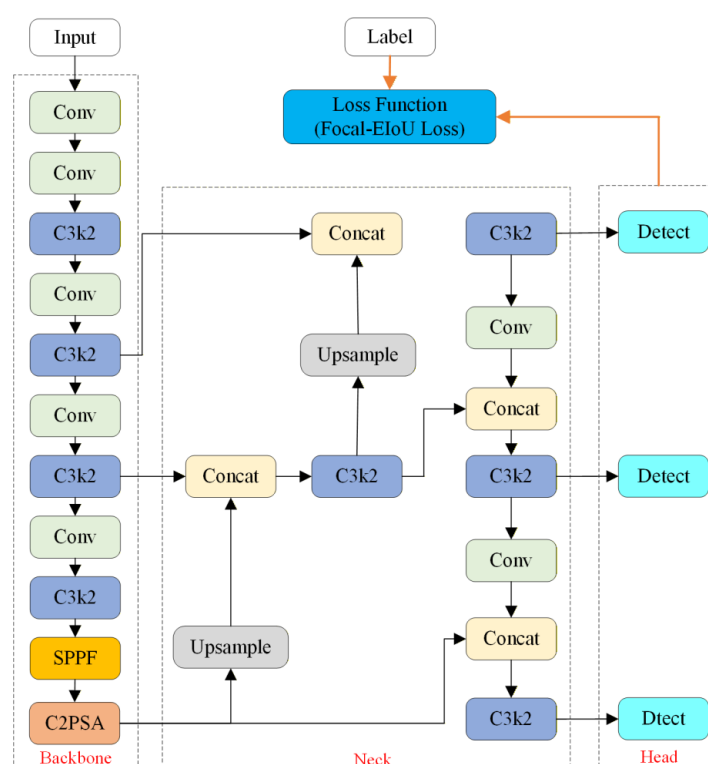


Figure 4. Structure of the UAV detection model.

3.3.2. Loss Function

The EIoU Loss [30] further decomposes the aspect ratio penalty term of the original CIoU Loss in YOLOv11, independently constraining the width–height disparity to enhance regression efficiency. The loss function is calculated as follows:

$$\text{EIoU_Loss} = 1 - \text{IoU} + \frac{p^2(b, b^{gt})}{c^2} + \frac{p^2(w, w^{gt})}{c_w^2} + \frac{p^2(h, h^{gt})}{c_h^2} \quad (6)$$

Here, $p^2(b, b^{gt})$, $p^2(w, w^{gt})$ and $p^2(h, h^{gt})$ represent the square of the distance between the center points, the width and height differences between the predicted bounding box and the ground truth. c , c_w and c_h correspond to the diagonal length, width and height of the smallest enclosing rectangles, which are used to normalize the penalty term.

Based on Equation (5), focal loss [31] is introduced, which is specially designed to address the imbalance between positive and negative samples. It improves bounding box regression accuracy by dynamically adjusting the weights of samples with different IoU values, enabling the model to focus more effectively on hard-to-regress samples during training. The Focal-EIoU loss function is obtained as:

$$Loss = 1 - IoU + \alpha \times \left[\frac{p^2(b, b^{gt})}{c^2} + \frac{p^2(w, w^{gt})}{c_w^2} + \frac{p^2(h, h^{gt})}{c_h^2} \right] + \beta \times \left[\frac{Area_{intersection}^n}{Area_{union} + \sigma} \right] - \gamma \quad (7)$$

Here, $Area_{intersection}$ and $Area_{union}$ refer to the intersection and union, respectively, of the predicted bounding box and the ground truth. σ is a minimum value, $n = 0.05$. Through a series of experimental evaluations conducted in this study, the optimal value of the parameters α , β and γ are determined as 0.9, 0.1 and 0.065, respectively.

4. Dataset

This study utilizes the Anti-UAV drone object detection dataset [32] released by the University of Chinese Academy of Sciences, which includes over 300 visible–infrared image pairs featuring a variety of commonly available drone models currently on the market. The drones in the dataset exhibit a wide range of flight distances, orientations, lighting conditions (including backlighting), and background environments, with additional inclusion of nighttime scenarios. A key feature of the dataset is its ability to simulate misaligned images captured by a dual-spectrum camera under typical operational conditions. This characteristic makes the dataset particularly suitable for evaluating the performance of the proposed visible–infrared image fusion method.

A challenging subset of images was selected from the visible and infrared image database, covering multi-scale objects and complex architectural backgrounds. A total of 1570 image pairs along with the corresponding annotation files (accurately marking the UAV target locations) were collected to form the final dataset and will be used for dual-light fusion and object detection evaluations.

5. Experiment and Analysis

5.1. Dual-Spectrum Fusion

We selected three groups of dual-spectrum images from the dataset as the test images, which were sourced from representative scenarios. Among them, the sizes of the drones in the first group of images vary significantly, while the drones in the second and third groups are smaller and have different backgrounds.

According to Section 3.1, feature matching and selection were conducted on three representative images obtained under scenes of diverse background and drone distance, the results were presented in Figure 5, and further detailed with statistical data in Table 1. It is evident that the proposed feature matching method effectively identifies feature points in both foreground and background regions, while eliminating incorrect feature correspondences.

Building upon the results of feature matching, further feature fusion was performed. This process is illustrated in Figure 6. As can be observed in the final visible light test image, the effective pixel regions—defined as those containing no fewer than two corresponding control points—are enhanced by the texture and brightness characteristics of the infrared image. Demonstrate that feature fusion significantly enhances the contrast of

the foreground regions in the original image, thereby facilitating more effective feature extraction for the object detection network.

Table 1. Statistics of feature matching and selection results.

Scene	Before Selection			After Selection		
	All Pairs	Matched	Precision	All Pairs	Matched	Precision
Scene 1	12	8	66.7%	6	6	100.0%
Scene 2	25	22	88.0%	14	13	92.9%
Scene 3	14	6	42.9%	6	6	100.0%

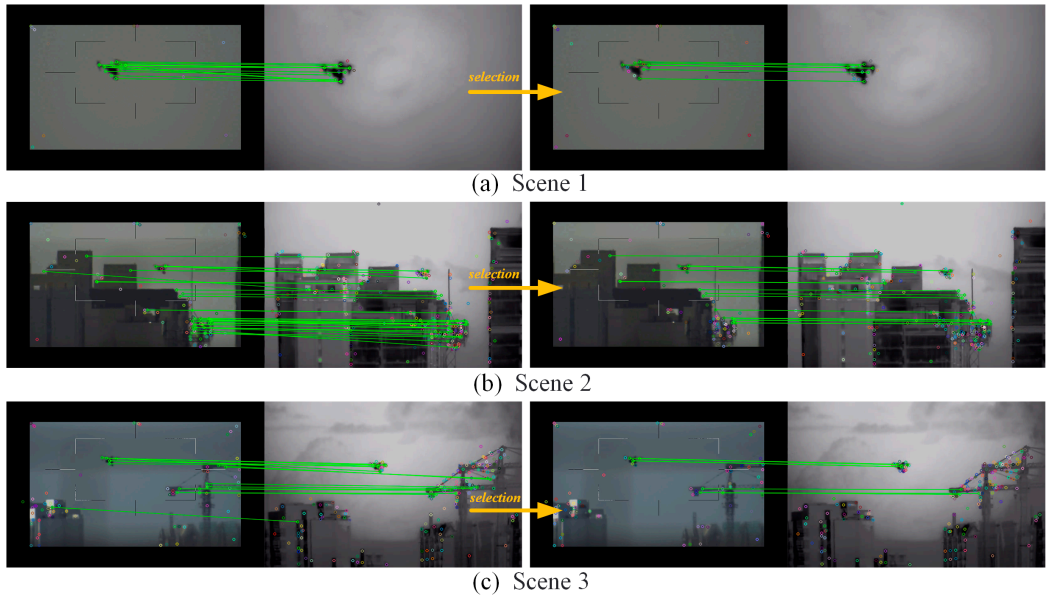


Figure 5. Results of feature matching and selection.

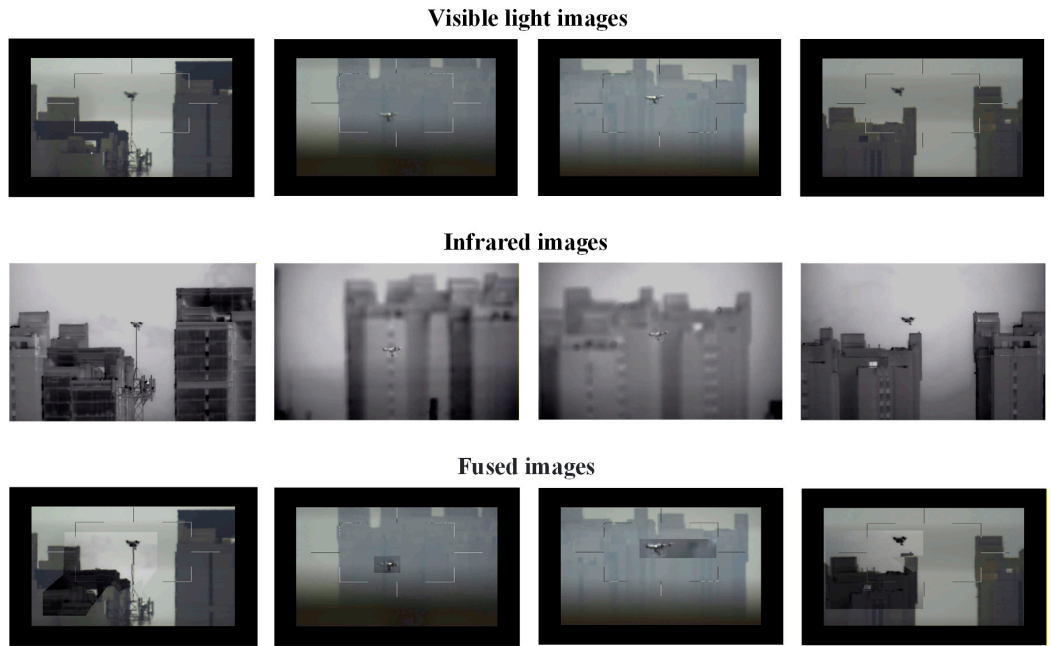


Figure 6. Schematic diagram of dual-spectrum images fusion.

5.2. Object Detection

We initially conduct data preparation. For each pair in the dataset, the corresponding feature-fused image was generated using the proposed Dual-spectrum fusion method, with

original labels preserved. These image pairs were then partitioned into training, validation, and test sets at a ratio of 7:2:1, respectively.

The model training was conducted using the Python 3.8 programming language. The experimental platform is a Lenovo brand AI workstation purchased from Shanghai, China, with a CPU (Intel Core i9-14900HX, 24 cores), 32 GB of RAM, and a GPU (NVIDIA GeForce RTX 4060). The model was trained with an initial learning rate of 0.01, a momentum of 0.937, a batch size of 6, and utilized 8 CPU workers. The training utilized the dataset described in Section 4.

To evaluate the model's detection capability, an ablation experiment was conducted. The results are presented in Table 2. As shown in the table, the YOLOv11-EIoU model demonstrates significantly improved detection performance compared to the original YOLOv11 model, but the proposed method achieves the highest performance. Compared with state-of-the-art studies in this field, the YOLOv9-CAG architecture [25] and the YOLOv9-AAG architecture [26] achieved mean Average Precision (mAP) scores of 0.842 and 0.816, respectively, on their respective combined dataset of visible and infrared images. In contrast, the proposed method attained mAP of 0.995, demonstrating its superior learning capability.

Table 2. Results of the ablation experiment.

Model	Precision	Recall	mAP50	mAP50-95
YOLOv11	0.990	0.990	0.994	0.675
YOLOv11-EIoU	0.995	0.995	0.994	0.671
Proposed method	0.995	0.995	0.995	0.673

The model was directly applied to UAV detection in fused dual-spectral images. Detection experiments were conducted using multispectral image data collected from diverse scenarios, encompassing variations in scale and background complexity. The results, presented in Figure 7, show that the detection confidence levels are all above 0.7, demonstrating that the detection model is capable of accurately localizing UAV targets across a range of challenging conditions.

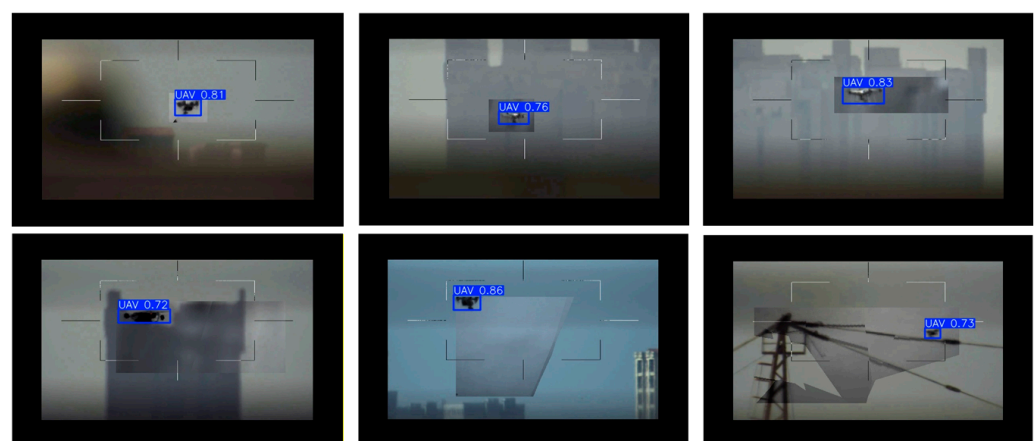


Figure 7. Schematic diagram of dual-spectrum fusion-based UAV detection.

The measured FPS (Frames Per Second) is 213.4, indicating that the model can achieve a relatively high detection speed under the existing computational resources. Furthermore, given that the proposed image fusion method is a traditional image processing approach, which consumes negligible computational resource, the “Fusion—Detection” methodology has good usability in real-time drone detection tasks.

6. Conclusions

This paper presents a spatially adaptive feature matching and fusion method for visible–infrared images, specifically designed to address the misalignment characteristics of dual-spectrum Low-altitude aerial monitoring images. The proposed method significantly enhances the feature representation capabilities of key regions in single-spectrum images. Additionally, the loss function of the YOLOv11 model is optimized and applied to improve the detection performance of drone targets in the fused images.

The method drawing on the widely adopted “Fusion–Detection” paradigm in related studies, this method proposes an image fusion approach that is computationally more efficient than conventional deep learning-based fusion networks. Through model lightweight optimization and format conversion, the detection framework can be deployed on lower-end systems or real-time applications, enabling work in conjunction with fixed surveillance cameras as well as airborne or vehicle-mounted mobile cameras.

In this study, the detection model was trained solely on the Anti-UAV dataset. Provided that sufficient image samples and scene diversity are available, the trained model can be effectively transferred to other drone platforms, including fixed-wing, hybrid drones and so on.

The proposed pixel reconstruction strategy relies heavily on the feature point combinations. Under extreme conditions such as low illumination or severe object occlusion, it may be difficult to obtain valid feature point groups to serve as control points, leading to potential failure of the feature fusion process. Future work will focus on further optimizing the algorithm to enhance its robustness in feature extraction under varied scales, lighting conditions, and complex environments.

Author Contributions: Z.H.: Writing—original draft, Investigation, Methodology, Validation, Visualization, Data curation, and Software. G.Z.: Writing—review & editing, Formal analysis, and Project administration. Y.W.: Supervision, Funding acquisition, and Resources. C.D.: Writing—review & editing and Conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number “52074309”.

Data Availability Statement: The publicly available Anti-UAV dataset used in this study is from the web site <https://github.com/ucas-vg/Anti-UAV> (accessed on 8 May 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Al-Room, K.; Iqbal, F.; Baker, T.; Shah, B.; Yankson, B.; MacDermott, A.; Hung, P. Drone forensics: A case study of digital forensic investigations conducted on common drone models. *Int. J. Digit. Crime Forensics* **2001**, *13*, 1–25. [\[CrossRef\]](#)
2. Shi, X.; Yang, C.; Xie, W.; Liang, C.; Shi, Z.; Chen, J. Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges. *IEEE Commun. Mag.* **2018**, *56*, 68–74. [\[CrossRef\]](#)
3. Tong, F.; Zhang, Y.; Xu, J.; Li, L.; Liao, G. A low side lobe anti-UAV radar beamforming method based on hybrid basic beams superposition. *Electron. Lett.* **2025**, *61*, e70304. [\[CrossRef\]](#)
4. Guo, D.; Qu, Y.; Zhou, X.; Sun, J.; Yin, S.; Lu, J.; Liu, F. Research on automatic tracking and size estimation algorithm of “low, slow and small” targets based on Gm-APD single-photon LIDAR. *Drones* **2025**, *9*, 85. [\[CrossRef\]](#)
5. Cui, Y.; Zhu, C.; Wei, Y.; Zhang, A.; Bai, D. An automatic image tracking system for urban low-altitude anti-UAV fusion using the DSST–KCF algorithm. *Exp. Technol. Manag.* **2024**, *41*, 11.
6. Du, L.; Gao, C.; Feng, Q.; Wang, C.; Liu, J. Small UAV detection in videos from a single moving camera. In *CCF Chinese Conference on Computer Vision*; Springer: Singapore, 2017; pp. 187–197.
7. Liu, Z.; An, P.; Yang, Y.; Qiu, S.; Liu, Q.; Xu, X. Vision-based drone detection in complex environments: A survey. *Drones* **2024**, *8*, 643. [\[CrossRef\]](#)

8. Samadzadegan, F.; Dadrass Javan, F.; Ashtari Mahini, F.; Gholamshahi, M. Detection and recognition of drones based on a deep convolutional neural network using visible imagery. *Aerospace* **2022**, *9*, 31. [\[CrossRef\]](#)
9. Yasmine, G.; Maha, G.; Hicham, M. Anti-drone systems: An attention based improved YOLOv7 model for a real-time detection and identification of multi-airborne target. *Intell. Syst. Appl.* **2023**, *20*, 200296. [\[CrossRef\]](#)
10. Hu, S.; Zhao, F.; Lu, H.; Deng, Y.; Du, J.; Shen, X. Improving YOLOv7-tiny for infrared and visible light image object detection on drones. *Remote Sens.* **2023**, *15*, 3214. [\[CrossRef\]](#)
11. Lu, Z.; Yueping, P.; Zecong, Y.; Rongqi, J.; Tongtong, Z. Infrared small UAV object detection algorithm based on enhanced adaptive feature pyramid networks. *IEEE Access* **2022**, *10*, 115988–115995. [\[CrossRef\]](#)
12. Zhao, M.; Li, W.; Li, L.; Wang, A.; Hu, J.; Tao, R. Infrared small UAV object detection via isolation forest. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5004316. [\[CrossRef\]](#)
13. Xu, K.; Song, C.; Xie, Y.; Pan, L.; Gan, X.; Huang, G. RMT-YOLOv9s: An infrared small object detection method based on UAV remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 7002205. [\[CrossRef\]](#)
14. Fang, H.; Ding, L.; Wang, X.; Chang, Y.; Yan, L.; Liu, L.; Fang, J. SCINet: Spatial and contrast interactive super-resolution assisted infrared UAV object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5006722. [\[CrossRef\]](#)
15. Yang, Z.; Lian, J.; Liu, J. Infrared UAV target detection based on continuous-coupled neural network. *Micromachines* **2023**, *14*, 2113. [\[CrossRef\]](#)
16. Pan, L.; Liu, T.; Cheng, J.; Cheng, B.; Cai, Y. AIMED-Net: An enhancing infrared small object detection net in UAVs with multi-layer feature enhancement for edge computing. *Remote Sens.* **2024**, *16*, 1776. [\[CrossRef\]](#)
17. Fang, H.; Ding, L.; Wang, L.; Chang, Y.; Yan, L.; Han, J. Infrared small UAV object detection based on depthwise separable residual dense network and multiscale feature fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5019120. [\[CrossRef\]](#)
18. Cao, L.; Wang, Q.; Luo, Y.; Hou, Y.; Cao, J.; Zheng, W. YOLO-TSL: A lightweight object detection algorithm for UAV infrared images based on triplet attention and slim-neck. *Infrared Phys. Technol.* **2024**, *141*, 105487. [\[CrossRef\]](#)
19. Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. PHSI-RTDETR: A lightweight infrared small object detection algorithm based on UAV aerial photography. *Drones* **2024**, *8*, 240. [\[CrossRef\]](#)
20. Wang, C.; Yang, J.; Sun, D.; Gao, Q.; Liu, Q.; Wang, T.; Hu, A.; Wang, L. Air-to-ground object detection and tracking based on dual-stream fusion of unmanned aerial vehicle. *J. Field Robot.* **2025**, *42*, 3582–3599. [\[CrossRef\]](#)
21. Gao, H.; Wang, Y.; Sun, J.; Jiang, Y.; Gai, Y.; Yu, J. Efficient multi-level cross-modal fusion and detection network for infrared and visible image. *Alex. Eng. J.* **2024**, *108*, 306–318. [\[CrossRef\]](#)
22. Chang, K.; Huang, J.; Sun, X.; Luo, J.; Bao, S.; Huang, H. Infrared and visible image fusion network based on multistage progressive injection. *Complex Intell. Syst.* **2025**, *11*, 367. [\[CrossRef\]](#)
23. Liu, X.; Lv, M.; Ma, C.; Fu, Z.; Zhang, L. Multi-modal image fusion of visible and infrared for precise positioning of UAVs in agricultural fields. *Comput. Electron. Agric.* **2025**, *232*, 110024. [\[CrossRef\]](#)
24. Jiang, W.; Pan, H.; Wang, Y.; Li, Y.; Lin, Y.; Bi, F. A multi-level cross-attention image registration method for visible and infrared small unmanned aerial vehicle targets via image style transfer. *Remote Sens.* **2024**, *16*, 2880. [\[CrossRef\]](#)
25. Zhu, J.; Rong, J.; Kou, W.; Zhou, Q.; Suo, P. Accurate recognition of UAVs on multi-scenario perception with YOLOv9-CAG. *Sci. Rep.* **2025**, *15*, 27755. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Suo, P.; Zhu, J.; Zhou, Q.; Kou, W.; Wang, X.; Suo, W. YOLOv9-AAG: Distinguishing birds and drones in infrared and visible light scenarios. *IEEE Access* **2025**, *13*, 76609–76619. [\[CrossRef\]](#)
27. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
28. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [\[CrossRef\]](#)
29. Wang, C.Y.; Liao, H.Y.M. YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems. *APSIPA Trans. Signal Inf. Process.* **2024**, *13*, 1. [\[CrossRef\]](#)
30. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [\[CrossRef\]](#)
31. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
32. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Guo, G.; Ye, Q.; Jiao, J. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. *IEEE Trans. Multimed.* **2021**, *25*, 486–500. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.