drones

MDPI

Article

Infrared Target Detection Based on Image Enhancement and an Improved Feature Extraction Network

Peng Wu, Zhen Zuo *, Shaojing Su and Boyuan Zhao

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; pengwu@nudt.edu.cn (P.W.); ssjing@nudt.edu.cn (S.S.); zhaoboyuan24@nudt.edu.cn (B.Z.)

* Correspondence: z.zuo@nudt.edu.cn

Abstract

Small unmanned aerial vehicles (UAVs) pose significant security challenges due to their low detectability in infrared imagery, particularly when appearing as small, low-contrast targets against complex backgrounds. This paper presents a novel infrared target detection framework that addresses these challenges through two key innovations: an improved Gaussian filtering-based image enhancement module and a hierarchical feature extraction network. The proposed image enhancement module incorporates a vertical weight function to handle abnormal feature values while preserving edge information, effectively improving image contrast and reducing noise. The detection network introduces the SODMamba backbone with Deep Feature Perception Modules (DFPMs) that leverage high-frequency components to enhance small target features. Extensive experiments on the custom SIDD dataset demonstrate that our method achieves superior detection performance across diverse backgrounds (urban, mountain, sea, and sky), with mAP@0.5 reaching 96.0%, 74.1%, 92.0%, and 98.7%, respectively. Notably, our model maintains a lightweight profile with only 6.2M parameters and enables real-time inference, which is crucial for practical deployment. Real-world validation experiments confirm the effectiveness and efficiency of the proposed approach for practical UAV detection applications.

Keywords: infrared target detection; image enhancement; deep feature perception

check for updates

Academic Editor: Bo Li

Received: 23 July 2025

Revised: 7 September 2025

Accepted: 16 September 2025

Published: 10 October 2025

Citation: Wu, P.; Zuo, Z.; Su, S.; Zhao, B. Infrared Target Detection Based on Image Enhancement and an Improved Feature Extraction Network. *Drones* **2025**, *9*, 695. https://doi.org/10.3390/drones9100695

1. Introduction

The proliferation of small unmanned aerial vehicles (UAVs) has created unprecedented security challenges for critical infrastructure protection, border surveillance, and public safety [1]. While UAVs offer numerous beneficial applications in areas such as delivery services, agricultural monitoring, and search and rescue operations, their potential misuse for illicit activities, including smuggling, espionage, and attacks on sensitive facilities, necessitates robust detection and tracking systems [2].

Infrared imaging has emerged as a crucial technology for UAV detection due to its ability to operate in day/night conditions and various weather scenarios [3]. However, detecting small UAVs in infrared imagery presents significant technical challenges. UAVs typically appear as small targets with limited thermal signatures, occupying only a few pixels in the image [4]. The problem is further compounded by complex backgrounds, atmospheric interference, and the inherent limitations of infrared sensors, including low signal-to-noise ratio and poor contrast [5].

Traditional infrared image processing methods often struggle to balance noise reduction with feature preservation, leading to blurred target edges and reduced detection accuracy [6]. Moreover, conventional object detection algorithms designed for visible-light imagery fail to adequately address the unique characteristics of infrared small target detection, particularly the lack of texture information and the prevalence of false alarms from background clutter [7].

Recent advances in deep learning have shown promise for infrared target detection, with methods such as YOLO variants achieving reasonable performance in certain scenarios [8]. However, these architectures face fundamental challenges when applied to the specific problem of infrared small target (IST) detection, particularly for UAVs. The core issue stems from the inherent conflict between the design of standard CNNs and the nature of ISTs. Firstly, the extensive down-sampling (e.g., to 32 × 32 stride) in backbones like YOLO critically erodes the limited pixel-level evidence of ISTs, which may comprise only a few pixels. This irreversible information loss at the earliest stages makes subsequent detection exceedingly difficult. Secondly, the anchor-based proposal mechanisms in many YOLO versions are poorly suited for small infrared targets, as predefined anchors often vastly exceed the target size, leading to poor matching and learning inefficiency. Furthermore, the loss functions (e.g., IoU) and feature fusion strategies optimized for larger objects in natural images fail to effectively leverage the faint, high-frequency cues that are paramount for distinguishing ISTs from cluttered backgrounds. Consequently, these approaches struggle with the extremely small and low-signature targets that characterize UAV detection scenarios [9]. While computational burden is a concern, these architectural limitations are a primary bottleneck. The need for real-time processing further constrains the complexity of viable solutions, as practical deployment requires processing speeds sufficient for responsive countermeasures [10].

This paper addresses these challenges through a comprehensive framework that combines advanced image preprocessing with specialized feature extraction tailored for infrared small target detection. Our main contributions are as follows:

(1) An improved Gaussian filtering algorithm that incorporates a vertical weight function within a robust estimation model to adaptively re-weight the contribution of each pixel during convolution. This core modification preserves edge information significantly better than conventional filtering, thereby providing higher-quality images for subsequent detection tasks.

(2) A novel SODMamba backbone architecture featuring Deep Feature Perception Modules (DFPMs) that exploit high-frequency components to enhance small target features while suppressing background noise.

(3) Comprehensive evaluation on a custom SIDD dataset containing 4737 infrared images across four distinct scenarios, demonstrating superior performance compared to state-of-the-art methods.

(4) Real-world validation using a complete UAV detection system, confirming the practical applicability of our approach with real-time processing capabilities.

The remainder of this paper is organized as follows: Section 2 reviews related work in infrared image enhancement and target detection. Section 3 presents our proposed method in detail. Sections 4 and 5 describe experimental results on benchmark and real-world datasets. Finally, Section 6 concludes the paper and discusses future directions.

2. Related Work

2.1. Infrared Image Enhancement

Infrared images inherently suffer from low contrast, high noise levels, and limited dynamic range due to sensor limitations and environmental factors [11]. Various enhancement techniques have been proposed to improve image quality for subsequent processing tasks.

Traditional enhancement methods have evolved significantly in recent years. Li et al. [12] proposed an infrared image enhancement method based on adaptive histogram partition and brightness correction, which effectively addresses over-enhancement issues in traditional histogram equalization. Wang et al. [13] introduced a detail-preserving image enhancement algorithm using guided filtering and multi-scale decomposition, demonstrating superior performance in preserving edge information while improving contrast. Deep learning approaches have revolutionized infrared image enhancement. Kuang et al. [14] developed a single infrared image enhancement method using deep convolutional neural networks, achieving significant improvements over traditional methods. More recently, Li et al. [15] proposed an infrared image enhancement network based on multi-scale feature extraction and attention mechanisms, which adaptively enhances different image regions based on their characteristics. Hybrid approaches combining traditional and learning-based methods have shown promise. Zhang et al. [16] introduced a method that integrates physical models with deep learning for infrared image enhancement, leveraging domain knowledge to improve network performance. Liu et al. [17] proposed a two-stage enhancement framework that combines Retinex theory with deep neural networks, achieving balanced enhancement across different scene types.

2.2. Infrared Small Target Detection

Infrared small target detection remains a challenging problem due to the limited information available from small targets and complex backgrounds. Recent advances have focused on both traditional and deep learning approaches.

Model-based Methods: Recent model-based methods have focused on exploiting local contrast and structural information. Han et al. [18] proposed a local contrast measure based on multi-scale patch-based contrast, achieving robust detection in complex backgrounds. Wei et al. [19] introduced a multi-scale gray difference weighted image entropy method that effectively suppresses background clutter while preserving small targets. The infrared patch-tensor (IPT) model has gained attention for its effectiveness. Zhang et al. [20] proposed a non-convex rank approximation minimization method for infrared small target detection, demonstrating superior performance in suppressing background edges. Dai et al. [21] extended this work with a weighted infrared patch-tensor model that adaptively adjusts to local image characteristics.

Deep Learning Methods: Deep learning has transformed infrared small target detection capabilities. Li et al. [22] proposed a dense nested attention network (DNANet) specifically designed for infrared small target detection, utilizing multi-level features with attention mechanisms. This work demonstrated significant improvements over traditional methods, particularly in complex scenes.

Transformer-based architectures have recently been applied to this domain. Wang et al. [23] introduced UIU-Net, combining transformer blocks with CNN features for infrared small target detection, achieving state-of-the-art performance. Zhang et al. [24] proposed ISNet, an infrared small target detection network with scale and location sensitivity, addressing the challenge of targets at different scales. The asymmetric contextual modulation (ACM) approach by Dai et al. [25] has been particularly influential, introducing a novel way to model the relationship between targets and backgrounds in infrared images. This method has inspired several subsequent works focusing on contextual information exploitation.

2.3. Infrared Drone Detection Methods

The detection of unmanned aerial vehicles (UAVs) using infrared imaging has emerged as a critical technology in counter-drone systems due to its all-weather capability and passive detection nature. Unlike traditional infrared small target detection, drone detection presents unique challenges that require specialized approaches.

Current infrared drone detection methods can be broadly categorized into traditional and deep learning-based approaches. Traditional methods typically rely on temporal or spatial filtering techniques. Background subtraction methods, such as those proposed by Zheng et al. [26], proposed a novel approach inspired by the human visual system (HVS), where block compressed sampling theory was used first, then small abnormal regions in the modulation map were detected by a high-speed local contrast method and defined as candidate targets. However, these methods struggle with dynamic backgrounds and require careful parameter tuning.

Deep learning-based approaches have recently dominated this field. Convolutional neural networks (CNNs), particularly YOLO variants, have been widely adopted due to their real-time performance. However, their performance is inherently limited by the convolutional operations' local receptive field, which struggles to capture the global context crucial for distinguishing small drones from noisy backgrounds. To address this, Transformer-based architectures have been explored for their superior global modeling capabilities through self-attention mechanisms. Yuan et al. [27] introduced a spatial-channel cross transformer network. Its core is a spatial-channel interactive transformer block (SCTB) designed to establish correlations between encoder and decoder features, predicting contextual differences between target and background in deeper layers. Liu et al. [28] proposed a parallel transformer–CNN hybrid (TCM) module for learned image compression, showcasing an effective design for combining local modeling from CNN and non-local modeling from transformer. Wang et al. [29] proposed a lightweight infrared small target detection method based on a linear transformer, which consists of two parts, namely a multi-scale linear transformer and a lightweight dual feature pyramid network, to achieve high accuracy and low delay in infrared small target detection.

Despite these advances, infrared drone detection methods face several fundamental limitations. Extreme Scale Variation: Drones can appear as small as 4 × 4 pixels at long distances, making feature extraction extremely challenging. This scale issue is exacerbated by the rapid changes in apparent size as drones approach or recede from the camera. Low Signal-to-Noise Ratio: Thermal signatures of drones are often weak and inconsistent due to varying operational conditions, weather effects, and different drone materials. This results in poor thermal contrast against backgrounds. Dataset Limitations: The lack of large-scale, diverse infrared drone datasets covering various scenarios, seasons, and weather conditions hinders the development of robust models.

Our work addresses these limitations by proposing a specialized image enhancement module that improves thermal contrast and an efficient network architecture that maintains high accuracy while meeting real-time requirements.

2.4. YOLO-Based Detection for Infrared Applications

The YOLO family of detectors has been extensively adopted for infrared applications due to its real-time capabilities. Recent versions have shown particular promise for small target detection tasks.YOLOv5 adaptations have been successful in infrared domains. Ding et al. [30] proposed an improved YOLOv5 for infrared small target detection, incorporating multi-scale feature fusion and attention mechanisms. Their method achieved significant improvements in detecting small UAVs while maintaining real-time performance. YOLOv7 and YOLOv8 have introduced architectural improvements beneficial for infrared detection. Wang et al. [31] adapted YOLOv7 for infrared maritime target detection, demonstrating robust performance in challenging maritime environments. The integration of the E-ELAN (Extended Efficient Layer Aggregation Network) structure proved particularly effective for small target feature extraction. Recent work has focused on loss function improvements for small targets. Xu et al. [32] proposed Normalized Wasserstein Distance (NWD) for tiny object detection, addressing the sensitivity of IoU-based metrics for small objects. This approach has been successfully applied to infrared UAV detection scenarios.

2.5. UAV Detection Systems

Comprehensive UAV detection systems combining multiple technologies have emerged. Shi et al. [33] developed an anti-UAV system integrating radar and electro-optical sensors, demonstrating the benefits of multi-modal fusion. Their work highlighted the importance of infrared imaging for all-weather operation. Real-time processing requirements have driven system design innovations. Liu et al. [34] proposed a lightweight neural network for real-time UAV detection in infrared images. This work demonstrated the feasibility of deploying deep learning models in resource-constrained environments. Despite these advances, existing methods still face challenges in detecting extremely small UAVs (less than 32 × 32 pixels) against complex backgrounds while maintaining real-time performance. Our work addresses these limitations through a comprehensive approach combining enhanced preprocessing with specialized feature extraction designed specifically for infrared small target characteristics.

3. Proposed Method

3.1. Overall Framework

The proposed architecture presents a comprehensive object detection framework that integrates advanced image preprocessing with a hierarchical feature extraction backbone. The overall structure consists of four main components: an image enhancement module, the SODMamba backbone, a path aggregation feature pyramid network (PAFPN), and a multi-scale detection head.

Figure 1 shows the diagram of the proposed infrared target detection.The input image first passes through an image enhancement module, which performs preprocessing operations. This module employs an improved Gaussian filtering technique that incorporates a vertical weight function to effectively handle abnormal feature values and achieve superior denoising performance. This enhancement step ensures more stable and reliable feature extraction in subsequent stages. The enhanced image is then fed into the SODMamba backbone, which adopts a hierarchical architecture with four distinct levels (L1–L4). The backbone begins with a Simple Stem containing an ODSSBlock for initial feature extraction. Each subsequent level incorporates a deep feature perception module (DFPM) paired with an ODSSBlock, enabling the network to capture both local and global feature representations effectively. The DFPM modules leverage high-frequency perception capabilities to enhance feature discrimination. At the final stage, an SPPF module aggregates multi-scale contextual information.

The extracted features from levels P3, P4, and P5 are then processed through the PAFPN, which facilitates bidirectional feature fusion. The PAFPN employs a combination of upsampling operations, concatenation mechanisms, and ODSSBlocks to merge features across different scales, ensuring rich semantic and spatial information is preserved throughout the network. Finally, the detection head processes the multi-scale features through three parallel branches corresponding to different spatial resolutions: 80 × 80 for small objects, 40 × 40 for medium objects, and 20 × 20 for large objects. This multi-scale design enables the network to effectively detect objects of varying sizes within the input image, producing the final detection output.

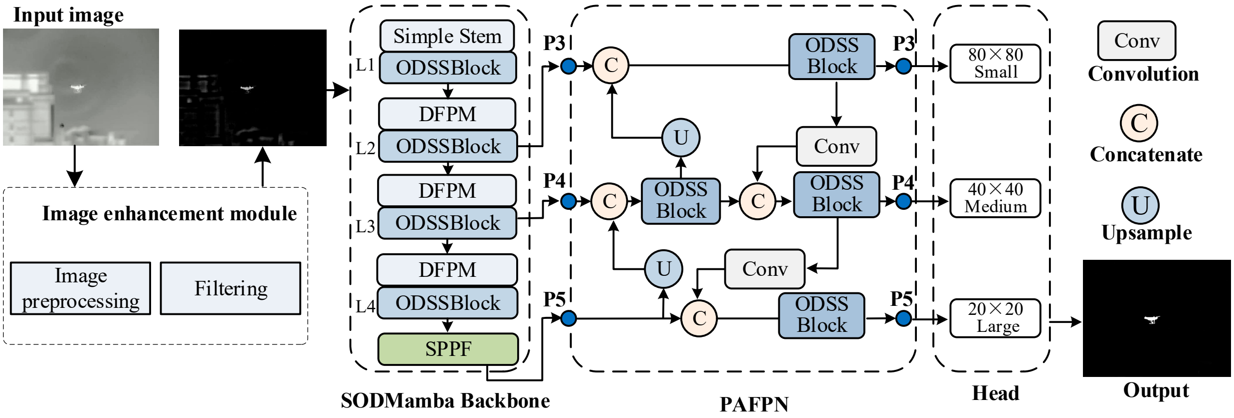

Figure 1. The diagram of the proposed infrared target detection.

3.2. Image Enhancement Module

Infrared images suffer from significant noise interference and blurred edge features, which are detrimental to feature extraction and learning-based detection. While Gaussian filtering demonstrates good performance in noise removal, the traditional Gaussian filtering method operates on the fundamental principle of convolution based on the Gaussian function, where the filter kernel weights are determined by the Gaussian function. During the filtering process, the gray value of each pixel in the image is replaced by the weighted average of pixels within its neighborhood, with weights calculated according to the Gaussian distribution; pixels closer to the central pixel receive higher weights. As a linear filtering method, Gaussian filtering inevitably loses some high-frequency information of the image while removing noise, such as edges and details, resulting in blurred image features that are unfavorable for subsequent object detection operations. Figure 2 shows examples of infrared images.

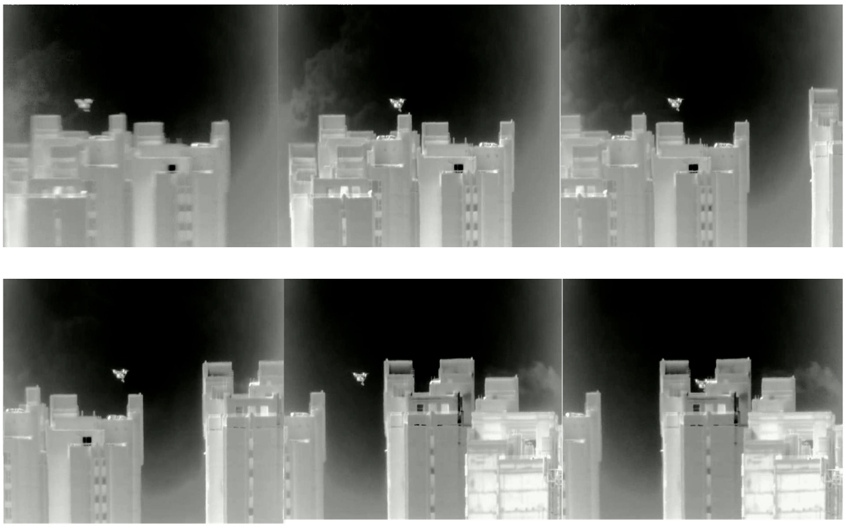

Figure 2. Examples of infrared images.

This paper presents an improved Gaussian filtering algorithm for infrared image enhancement. The proposed enhancement algorithm builds upon conventional Gaussian filtering by integrating grayscale transformation enhancement and maximum variance thresholding techniques while incorporating a vertical weight function to better preserve image features, thereby providing higher-quality images for subsequent target detection tasks.

To mitigate the feature blurring issues inherent in traditional Gaussian filtering approaches for image processing, grayscale transformation enhancement and maximum variance thresholding methods are employed as preprocessing techniques. The former adjusts the grayscale dynamic range to accentuate contrast between bright and dark regions while enhancing detailed features. The latter improves feature extraction accuracy and reduces noise interference by segmenting the image into background and foreground components, ultimately yielding infrared images with enhanced visual clarity and more prominent features that facilitate subsequent processing operations.

For tangent orientation determination, a grayscale means square deviation calculation method utilizing one-dimensional templates is adopted. By quantifying the local grayscale dispersion, this approach effectively reduces noise interference while maintaining consistency between tangent and texture directions, thereby further improving the efficacy of subsequent edge feature extraction.

Regarding the enhancement of Gaussian filtering, the introduction of a vertical weight function eliminates the influence of anomalous feature values, resulting in superior noise suppression and improved image processing stability.

3.2.1. Image Preprocessing

A series of preprocessing operations was applied to the image, including grayscale transformation, enhancement, and binarization processing. After linear stretching of the original image's grayscale values, this approach effectively addresses issues of underexposure and low contrast, resulting in clearer images with more prominent features. The subsequent operation involves binarization processing, which segments the image into background and target regions. The employed method is the maximum variance thresholding technique, with the specific procedure as follows:

Let T denote the total number of pixels in the image to be processed, with an average grayscale value $\overline{g}$. Let a represent the segmentation threshold between background and target. The proportion of pixels belonging to the target relative to the entire image is denoted as $E_0(a)$, with its average grayscale value $\chi_0(a)$. Similarly, the proportion of pixels belonging to the background relative to the entire image is denoted as $E_1(a)$, with its average grayscale value $\chi_1(a)$. The following equations are established:

$$\overline{g} = E_0(a) * \chi_0(a) + E_1(a) * \chi_1(a) \tag{1}$$

Subsequently, the inter-class variance between the background and foreground pixels is calculated using the following equation.

$$\delta = E_0(a)[\overline{g} - \chi_0(a)]^2 + E_1(a)[\overline{g} - \chi_1(a)]^2 \tag{2}$$

By substituting Equation (1) into Equation (2), the final formula for inter-class variance is obtained as

$$\delta = E_0(a)E_1(a)[\chi_0(a) - \chi_1(a)]^2 \tag{3}$$

Finally, through an iterative approach, an optimal threshold a is identified that maximizes the inter-class variance, which serves as the required threshold for binarization processing.

3.2.2. Improved Gaussian Filtering for Image Enhancement

Traditional Gaussian filtering often yields poor performance when dealing with abnormal feature values in images, struggling to eliminate their influence and consequently affecting subsequent object detection. To address this issue, we improved the traditional

Gaussian filtering algorithm by incorporating a vertical weight function. This modification enables better handling of abnormal feature values, achieves improved denoising performance, and enhances the stability of processing results. The robust Gaussian filter is expressed as follows:

$$\int_{a_2}^{a_1} [\varepsilon(\varphi) - \varphi(\alpha)]^2 \gamma(\alpha)\eta(\alpha - \varphi)D(\varphi) \Rightarrow \{\varphi(\alpha)\} \tag{4}$$

where α is an independent variable representing the position of the image contour. $\varepsilon(\varphi)$ represents the surface profile; $\varphi(\alpha)$ denotes the low-frequency baseline signal; $\eta(\alpha - \varphi)$ is the newly introduced vertical weight function; $D(\varphi)$ indicates the convolution operation of the Gaussian weight function; $\gamma(\alpha)$ represents the frequency response function; and a_1, a_2 are the integration limits. The appropriate vertical weight function $\gamma(\alpha)$ is obtained through iterative operations.

The robust Gaussian filtering centerline is then calculated using the following formula:

$$\varphi(\alpha)' = \frac{\int_{a_2}^{a_1} \varepsilon(\alpha - \varphi)\gamma(\alpha)\eta(\alpha - \varphi)D(\varphi)}{\int_{a_2}^{a_1} \eta(\varphi)\gamma(\alpha)D(\varphi)} \tag{5}$$

where $\eta(\alpha - \varphi)$ represents the convolution operation of the surface profile; $\eta(\varphi)$ denotes the frequency response function.

Discretizing the above results yields

$$\varphi(i) = \frac{\sum\limits_{x_1}^{x_2} \varepsilon(\alpha - \varphi)\gamma(\alpha)\eta(\alpha)}{\sum\limits_{x_1}^{x_2} \gamma(\alpha)\eta(\alpha)} \tag{6}$$

where $\eta(\alpha)$ is the Gaussian weight function.

During the iterative process of $\varphi(i)$, Gaussian filtering is applied to both $\gamma(\alpha)$ and $\varepsilon(\alpha - \varphi)\gamma(\alpha)$, while introducing a residual function, resulting in the following expression:

$$W_E = 5.6328 \times mid(|\lambda(\alpha) - \varphi(\alpha)|) \tag{7}$$

where *mid* denotes the median value, and $\lambda(\alpha)$ represents the integral function.

Finally, a robust estimation model is constructed. This section employs the *Tukey* biweight estimator as the vertical weight function, which achieves better model stability and effectively excludes the influence of abnormal feature values. The *Tukey* biweight estimator is expressed as

$$\gamma(\alpha) = \begin{cases} \left(1 - U^2\right)^2 & U \leq 1 \\ 0 & U > 1 \end{cases} \tag{8}$$

where U represents the ratio of the deviation-related parameter to the parameter median. The final robust estimation model is formulated as

$$\begin{cases} \begin{cases} \gamma(\alpha + 1) = \left[1 - \left(\frac{\Delta o(\alpha)}{W_E}\right)\right] & \left|\frac{\Delta o(\alpha)}{W_E}\right| < 1 \\ \gamma(\alpha + 1) = 0 & \left|\frac{\Delta o(\alpha)}{W_E}\right| \geq 1 \end{cases} \\ \Delta o(\alpha) = |\lambda(\alpha) - \varphi(\alpha)| \end{cases} \tag{9}$$

The improvement in traditional Gaussian filtering explained in this section is achieved by constructing a robust estimation model for Gaussian filtering. Through iterative processes, the influence of abnormal feature values is eliminated, resulting in better denoising performance and improved image quality.

3.3. SODMamba Backbone Network

The SODMamba backbone serves as the primary feature extraction engine of our proposed network, designed to hierarchically capture and refine features from the enhanced infrared image. As illustrated in Figure 1, the backbone adopts a multi-level architecture (L1–L4), progressively processing the input through a series of specialized modules. It begins with the Simple Stem for initial patch embedding and feature dimension adjustment. Subsequently, each level integrates ODSSBlocks for robust local feature modeling, interspersed with deep feature perception modules (DFPMs) to specifically amplify the high-frequency cues critical for small drone targets. This design ensures a balanced focus on both local details and global context, making it particularly adept at handling the challenges of infrared small target detection.

3.3.1. Deep Feature Perception Module (DFPM)

Figure 3 shows the structure of DFPM. DFPM leverages the characteristic that small target features primarily correspond to high-frequency components in images. It extracts high-frequency responses from input feature maps through a high-pass filter, filtering out low-frequency background information. Subsequently, the high-frequency responses are utilized to generate channel attention weights and spatial attention weights, respectively: channel weights highlight channels containing more small target features, while spatial weights focus on regions where small targets are located. Finally, weighted fusion enhances the representation of small targets in the original features.

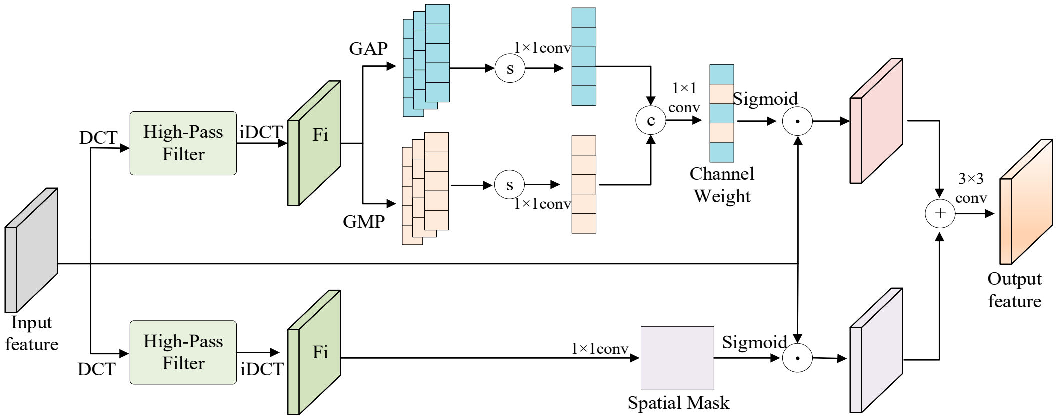

Figure 3. The structure of DFPM.

DFPM consists of three core components:

High-Frequency Feature Generator: Extracts high-frequency responses through discrete cosine transform (DCT) and a high-pass filter (controlled by the parameter α for filtering range). In our experiments, we employed a 5 × 5 high-pass filter kernel and set $\alpha = 0.25$. This value was empirically determined to optimally preserve the high-frequency cues of small targets while effectively suppressing low-frequency background noise. The parameter α determines the fraction of low-frequency DCT coefficients to be filtered out.

Channel Path (CP): Applies global average pooling (GAP) and global max pooling (GMP) to high-frequency responses, generates channel weights through convolutional layers, and dynamically allocates importance to each channel;

Spatial Path (SP): Aggregates channel information of high-frequency responses through 1 × 1 convolution to generate spatial masks, localizing pixel regions containing small targets. The outputs from both paths are element-wise multiplied and added, and then processed through a 3 × 3 convolution to obtain enhanced features.

When DFPM is integrated into the backbone, it further enhances high-frequency details of small targets in detection network output features while suppressing background noise. Through dynamic weight allocation, it enables the detection network to focus more on small target regions, improving detection accuracy for small targets in complex scenes and reducing missed detections and false positives.

3.3.2. Simple Stem

Figure 4 shows the structure of Simple Stem. Traditional sampling process, a convolution kernel with size 4 and stride 4 is typically used to segment the input image into non-overlapping patches for subsequent processing. However, this aggressive downsampling can lead to a significant loss of fine-grained spatial information and high-frequency details, which is particularly detrimental for detecting small targets that may only span a few pixels [35]. We propose a Simple Stem approach, redesigning the initial layers with two sequential convolutions of kernel size 3 and stride 2. This design generates overlapping patches, which helps to preserve more detailed information from the input image, especially crucial edge and texture features that are essential for identifying small infrared targets. While halving the hidden layer channels after each convolution to maintain computational efficiency, this overlapping patch strategy provides a richer feature representation for the subsequent network layers.

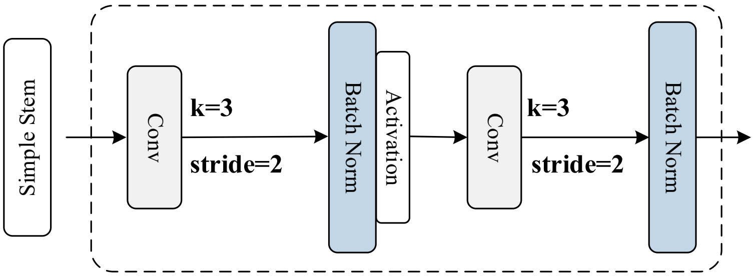

Figure 4. The structure of Simple Stem.

3.3.3. ODSSBlock

ODSSBlock is comprised of three sub-modules: LSBlock, RGBlock, and SS2D. Through sequential processing of input images, it preserves richer and deeper feature information, facilitating subsequent deep learning and establishing more reliable models. Meanwhile, batch normalization ensures efficient and stable training and inference processes. Figure 5 illustrates the overall structure of ODSSBlock.

Through batch normalization, layer normalization, and residual connections, ODSSBlock enables effective information flow when stacked in deep architectures. The overall formulation is as follows:

$$Z^{l-2} = \hat{\phi}\left(BN\left(Conv_{1\times1}\left(Z^{l-3}\right)\right)\right) \tag{10}$$

where $\hat{\phi}$ denotes the activation function, BN represents batch normalization, $Conv_{1\times1}$ indicates depthwise convolution, and Z^{l-3} is the input image.

$$Z^{l-1} = SS2D\left(LN\left(Z^{l-2}\right)\right) + Z^{l-2} \tag{11}$$

$$Z^{l} = RG\left(LN\left(Z^{l-1}\right)\right) + Z^{l-1} \tag{12}$$

where *SS2D* and *RG* represent the corresponding modules, and *LN* denotes layer normalization.

Figure 6 shows the structure of SS2D. SS2D consists of three main steps: Scan Expansion, S6 Block, and Scan Merge, with the algorithm workflow illustrated below:

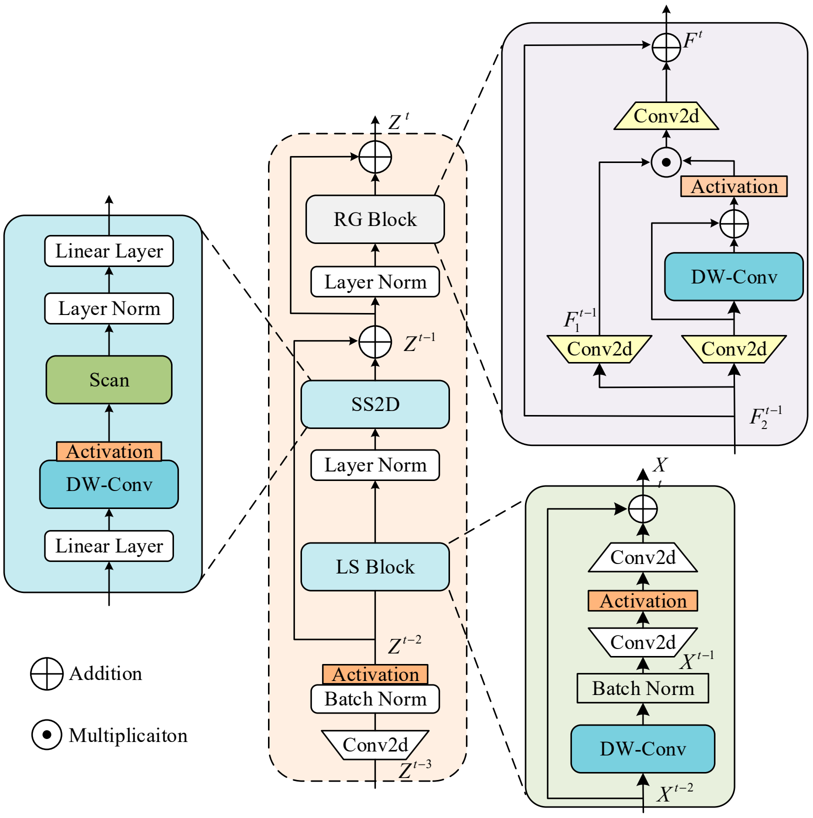

Figure 5. The structure of ODSSBlock.

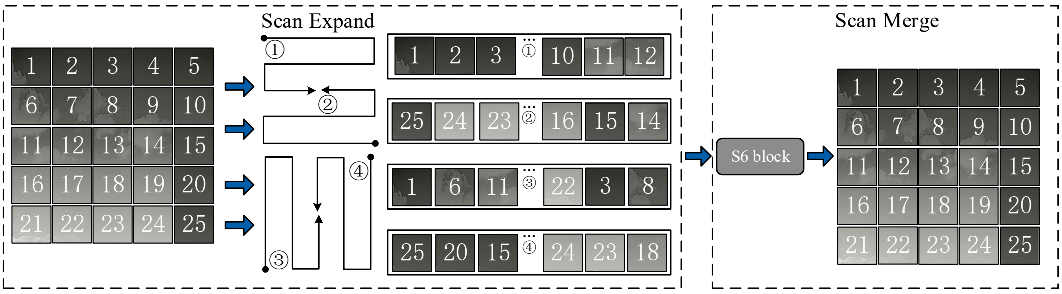

Figure 6. The structure of SS2D.

Scan Expansion expands the input image in different directions to generate four distinct sub-images. When viewed from the four corners of the image, the expansion directions of the four sub-images are symmetrical, originating from each corner and proceeding clockwise in an S-shaped pattern to complete image expansion. The advantage of multi-directional image expansion is comprehensive coverage of all image regions, enabling better feature extraction, providing richer information, and improving the efficiency and effectiveness of multi-dimensional image feature capture. After image expansion, the S6 Block processes the sub-images through feature extraction, and finally, the four sub-images are recombined through Scan Merge to output an image of the same size as the original.

Figure 7 shows the structure of S6 Block. The core of S6 Block is the State Space Model (SSM), which maps input vectors $x(t)$ to output vectors $y(t)$ through implicit intermediate states $h(t)$, enabling parameter updates and learning. The basic operational workflow is as follows:

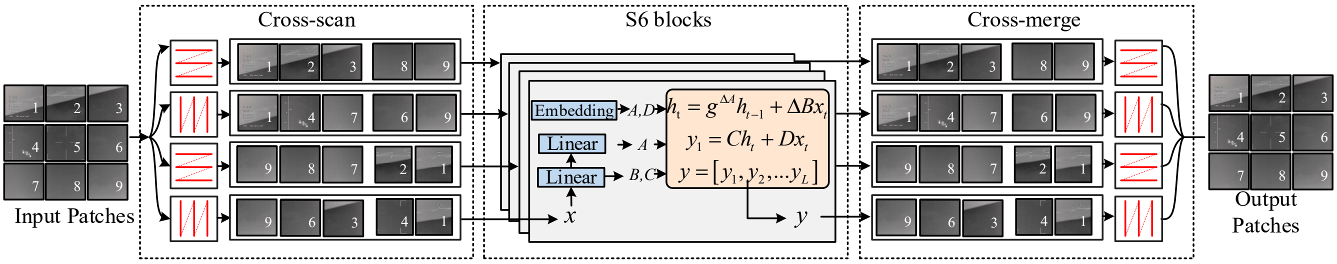

Figure 7. The structure of S6 Block.

3.4. PAFPN Module

Following the backbone, the path aggregation feature pyramid network (PAFPN) module acts as the neck of our detection architecture, responsible for effectively fusing the multi-scale features extracted from the backbone's different levels (P3, P4, and P5). The PAFPN facilitates bidirectional (top-down and bottom-up) information flow. It employs upsampling operations and concatenation to merge semantically strong features from higher levels with spatially rich features from lower levels. This process is further refined through ODSSBlocks, enhancing the representational power of the fused features. The output is a set of strengthened feature maps at multiple scales, each containing rich semantic information and precise spatial details, which are then fed into the detection head for precise target localization and classification.

4. Experiments on the Dataset

4.1. Dataset and Metrics

4.1.1. Dataset

To address the insufficient datasets in current drone target detection research, we selected publicly available infrared image datasets from ground/aerial backgrounds and partial data from the 1st CVPR Anti-UAV dataset to create a custom dataset. Targets in different scenes and scales are classified and annotated at the pixel level to construct the SIDD dataset. All images were meticulously annotated at the pixel level. The annotation was performed using the LabelMe annotation tool. The labeling criterion for a positive sample was defined as any visible drone within the image. For each drone instance, a polygon mask was drawn to tightly enclose all visible parts of the drone body, including the rotors, even in cases of partial occlusion or extreme blur. Images containing no drones were excluded from the dataset. To simulate realistic drone intrusion scenarios as closely as possible, the SIDD dataset is divided into four scenarios to investigate the impact of different backgrounds on low-altitude infrared drone target detection accuracy. Table 1. Shows the number of training and test sets in SIDD. The SIDD dataset contains 4737 infrared images with 640 × 512 pixel resolution with different backgrounds, including 1093 urban scene images, 2151 mountain scene images, 713 sea surface scene images, and 780 sky scene images. The dataset was split with 80% for algorithm training and 20% for testing.

Table 1. The number of training and test sets in SIDD.

Different Background	Training Set	Test Set	Total Number
Urban scene	874	219	1093
Mountain scene	1720	431	2151
Sea scene	570	143	713
Sky scene	624	156	780
Total number	3788	949	4737

Figure 8 shows sample images from the SIDD dataset, displaying urban, mountain, sea surface, and sky scenarios from top to bottom. In all four scenarios, a quadrotor drone serves as the detection target, with drone targets marked by red circles.

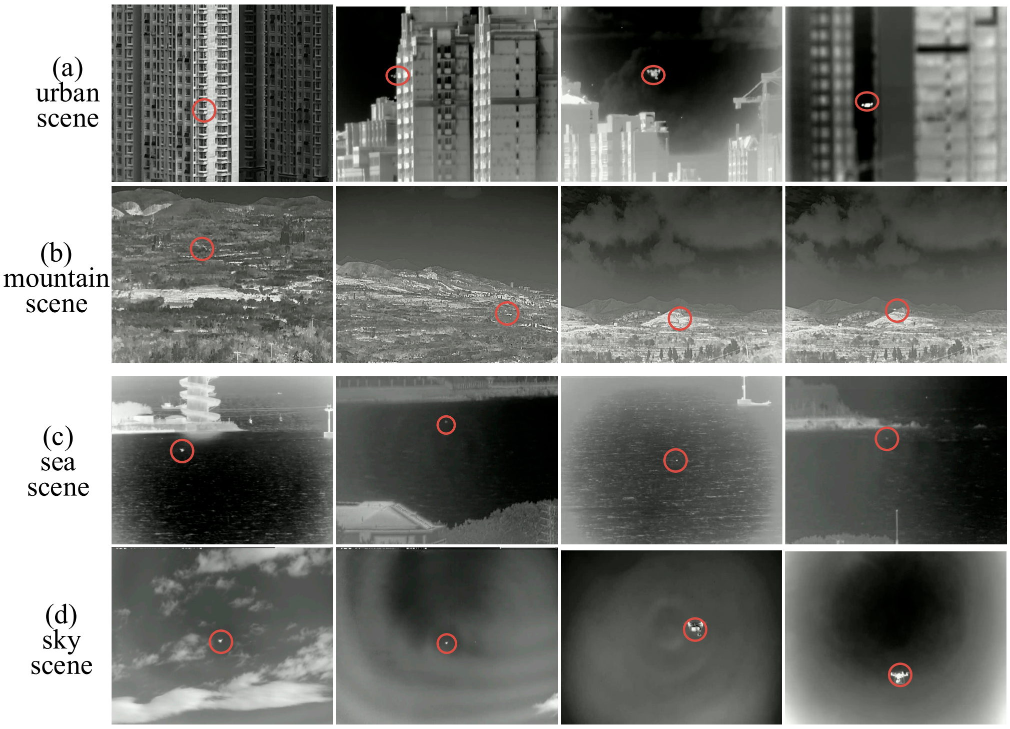

Figure 8. The examples of the SIDD dataset.

4.1.2. Metrics

This paper treats drone target detection as an instance segmentation task. Therefore, classic instance segmentation evaluation metrics were used to compare the performance of different algorithms, including average precision (AP) and model parameters, to assess the detection performance of various algorithms.

Average Precision: In instance segmentation tasks, IoU is typically used to determine whether prediction results are positive samples. IoU refers to the ratio of intersection to union between the predicted target mask and the ground truth region.

$$\text{IoU} = \frac{S_{overlop}}{S_{union}} \tag{13}$$

The definition of precision and recall is as follows:

$$\text{Precision} = \frac{TP}{TP+FP}\ \text{Recall} = \frac{TP}{TP+FN} \tag{14}$$

where *TP* (true positive) refers to the number of samples correctly predicted as positive by the model. *FP* (false positive) refers to the number of samples predicted as positive but actually negative. *FN* (false negative) refers to the number of samples predicted as negative but actually positive. The precision–recall (*P*-*R*) curve is a visualization method for evaluating detection algorithm performance, plotted with precision as the *x*-axis and recall as the *y*-axis. The mAP (mean average precision) value is obtained by calculating the area under the *P*-*R* curve to assess detection accuracy.

$$\text{mAP} = \int P(R)\mathrm{d}R \tag{15}$$

This study uses mAP$_s$ to evaluate the model's detection capability for small targets (targets smaller than 32 × 32 pixels in the image), mAP@0.5:0.95 to evaluate the average value when IoU thresholds are {0.5, 0.55, $\cdots$, 0.95}, and mAP@0.5 to evaluate the value when the IoU threshold is 0.5.

Model Parameters: The numerical values of parameters contained in the model, including weight matrices in convolutional and fully connected layers used in the model. The size of model parameters reflects the model complexity to some extent.

4.2. Implementation Details

The experiments in this chapter were conducted on the Ubuntu 18.04 operating system with the PyTorch 1.8.0 deep learning framework. All detection algorithms were trained for 50 epochs on a custom SIDD dataset. During training, 2 images were processed per iteration, with an initial learning rate of 0.0025, weight decay of 0.005, and AdaGrad as the training optimizer. The hardware setup consists of an NVIDIA GeForce RTX3060 GPU (Santa Clara, CA, USA) with 6 GB graphics memory, an AMD Ryzen 7-5800H CPU (Santa Clara, CA, USA), and 16 GB system memory.

4.3. Image Enhancement Results

We conducted experimental verification of the improved Gaussian filtering infrared enhancement algorithm on the aforementioned datasets. To more intuitively demonstrate the experimental results, this section presents a comparison between the original dataset images and the processed images. In each group of images, the left side shows the unprocessed image, while the right side displays the image after improved Gaussian filtering and infrared enhancement processing. The following four tables present the experimental contrast data, with the contrast formula defined as follows:

$$STD = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(I(i,j)-\mu)^2} \tag{16}$$

where STD represents contrast, and μ represents pixel values.

Figures 9–12 show the comparison results in the urban, mountain, sea and sky scene respectively. Tables 2–5 show the contrast data in the urban, mountain, sea and sky scene respectively. To provide a more comprehensive assessment following the reviewer's suggestion, we employed two additional standard image quality metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). The PSNR measures the ratio between the maximum possible power of a signal and the power of distorting noise, expressed in decibels (dB). A higher PSNR value typically indicates lower distortion and higher fidelity to the original image. The SSIM assesses the perceptual quality by measuring the similarity between two images based on their luminance, contrast, and structure. Its value ranges from −1 to 1, with a value closer to 1 indicating higher structural similarity.

Based on the analysis of the experimental data and images above, among the four scenarios, the urban scene exhibited the highest noise levels in the original images, presenting the greatest processing difficulty. This is reflected in the lower post-enhancement PSNR and SSIM values for this scenario, as the algorithm worked aggressively to suppress noise and enhance contrast, inevitably altering the image structure significantly. The sea and sky scenes demonstrated more favorable results in terms of absolute contrast improvement, with the sea scene also showing relatively higher PSNR values, indicating a better balance between enhancement and fidelity in these less cluttered environments. For the mountain scene, while the original image quality was relatively high and the processing results were satisfactory, the inherent similarity between the target and background textures limited the absolute contrast gain. The moderate PSNR and SSIM values here indicate a perceptible but less drastic transformation compared to the urban scene.

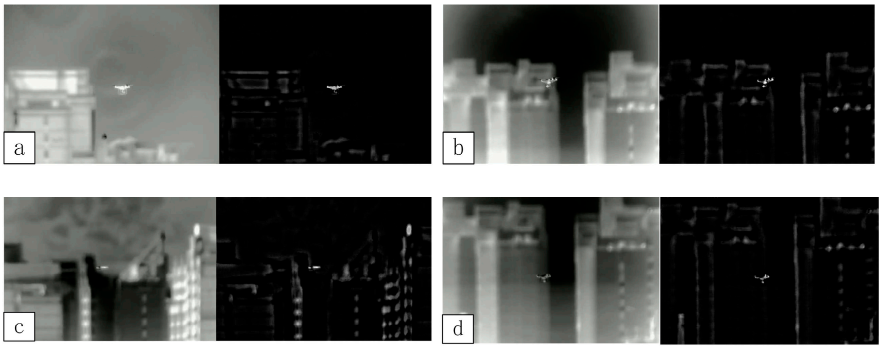

Figure 9. Comparison results in the urban scene. (**a**–**d**) shows four didderent images of original image and enhanced image.

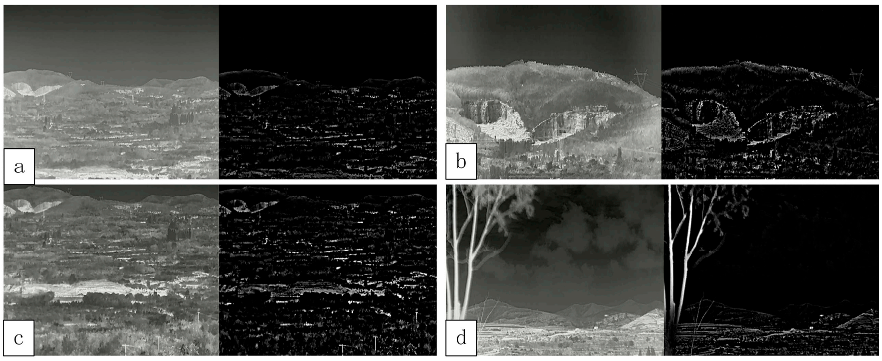

Figure 10. Comparison results in the mountain scene. (**a**–**d**) shows four didderent images of original image and enhanced image.

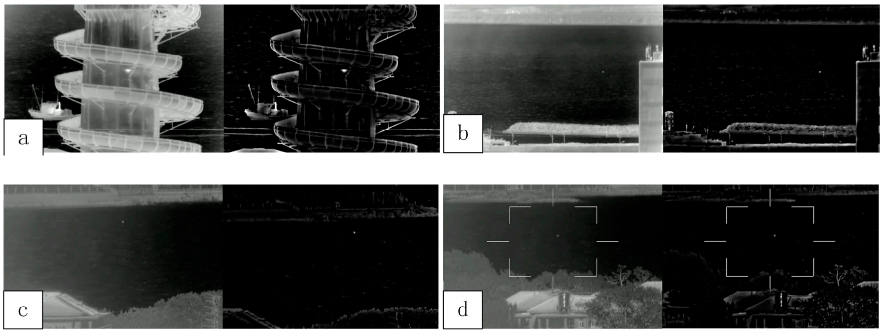

Figure 11. Comparison results in the sea scene. (**a**–**d**) shows four didderent images of original image and enhanced image.

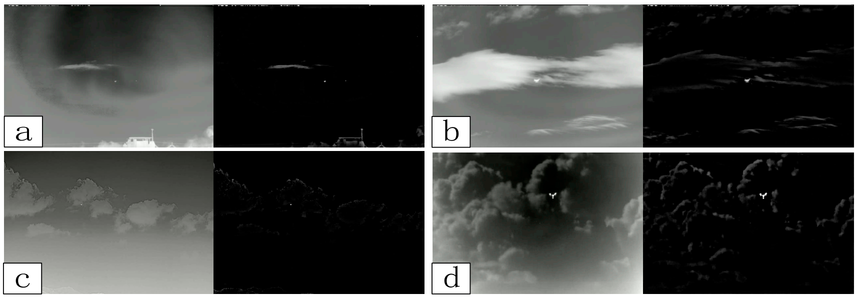

Figure 12. Comparison results in the sky scene. (**a**–**d**) shows four didderent images of original image and enhanced image.

Table 2. Contrast data in the urban scene.

Scene	Image Number	Original Contrast	Enhanced Contrast	PSNR	SSIM
Urban	a	27.60	74.27	5.24	0.06
	b	36.50	67.76	7.47	0.16
	c	35.87	47.53	7.84	0.13
	d	36.98	55.90	7.96	0.16

Table 3. Contrast data in the mountain scene.

Scene	Image Number	Original Contrast	Enhanced Contrast	PSNR	SSIM
mountain	a	36.82	43.02	5.58	0.09
	b	39.52	49.58	7.69	0.12
	c	35.87	40.60	7.92	0.18
	d	46.05	55.79	9.24	0.14

Table 4. Contrast data in the sea scene.

Scene	Image Number	Original Contrast	Enhanced Contrast	PSNR	SSIM
sea	a	46.57	63.70	8.23	0.22
	b	41.28	62.08	7.72	0.14
	c	40.73	48.19	10.38	0.12
	d	43.36	46.99	11.12	0.15

Table 5. Contrast data in the sky scene.

Scene	Image Number	Original Contrast	Enhanced Contrast	PSNR	SSIM
sky	a	33.00	44.13	7.61	0.04
	b	33.22	43.03	5.56	0.05
	c	31.69	45.32	5.65	0.04
	d	37.53	53.03	8.78	0.12

Crucially, the interpretation of the low PSNR and SSIM values requires an understanding of the enhancement objective. These metrics measure fidelity to the original image. However, the goal of our algorithm is not faithful reconstruction but rather strategic alteration to optimize the image for subsequent target detection. The significant reduction in these fidelity metrics is a direct consequence of the aggressive contrast stretching and noise suppression that define our approach. This is a justified trade-off, as the ultimate validation is the superior detection performance achieved using the enhanced images, as conclusively demonstrated in Section 4.4.

Overall, the improved Gaussian filtering-based infrared enhancement algorithm achieved its primary goal of producing images with higher contrast and reduced noise, providing more distinct features for the detection network, despite the inherent reduction in similarity to the original images as captured by PSNR and SSIM.

4.4. Target Detection Results Compared with Other Methods

The detection performance of various models in the urban scene is shown in Table 6. The three key metrics of the proposed method in this paper reached 66.1%, 66.8%, and 96%, respectively, all ranking first. With only 6.2 M parameters, significantly smaller than most models, it demonstrates substantial advantages in both detection performance and real-time capability.

Table 6. Detection performance of various models in the urban scene.

Different Methods	**Urban Scene**			
	mAP@0.5:0.95	**mAP@0.5**	**mAP$_s$**	**Paras (M)**
BoxInst	0.197	0.538	0.197	273.8
CondInst	0.565	0.936	0.564	272.6
SOLOv2	0.620	0.936	0.607	372.0
Mask-Rcnn	0.629	0.937	0.621	353.3
YOLOv5n-seg	0.473	0.936	0.465	4.1
YOLOv7	0.440	0.877	0.435	76.3
YOLOv8n-seg	0.477	0.927	0.474	9.6
Yolact++	0.423	0.902	--	199.0
Proposed method	0.668	0.960	0.661	6.2

The detection performance for the mountain scene is presented in Table 7. The three key metrics of the proposed method reached 42.6%, 42.6%, and 74.1%, respectively, with the first two metrics ranking first, demonstrating strong small target detection capability. The third metric also performed well, only slightly lower than Mask-RCNN.

Table 7. Detection performance of various models in the mountain scene.

Different Methods	**Mountain Scene**			
	mAP@0.5:0.95	**mAP@0.5**	**mAP$_s$**	**Paras (M)**
BoxInst	--	0.013	--	273.8
CondInst	0.284	0.731	0.284	272.6
SOLOv2	--	--	--	372.0
Mask-Rcnn	0.416	0.749	0.416	353.3
YOLOv5n-seg	0.245	0.689	0.245	4.1
YOLOv7	0.269	0.746	0.269	76.3
YOLOv8n-seg	0.253	0.664	0.253	9.6
Yolact++	0.177	0.625	--	199.0
Proposed method	0.426	0.741	0.426	6.2

Table 8 shows the detection performance for the sea scene. The proposed method's three key metrics reached 45.8%, 45.8%, and 92%, respectively. While performance remained good, compared to the previous two scenarios, the first two metrics ranked second, 0.5% lower than Mask-RCNN, still demonstrating high small target detection capability. The third metric was 1.3% lower than Mask-RCNN and 1% lower than YOLOv7.

The detection performance for the sky scene is shown in Table 9. The proposed method's three key metrics reached 73.4%, 73.4%, and 98.7%, respectively, all ranking first. Due to the simple sky background, all models showed high performance overall, but the proposed method still demonstrated exceptional target detection performance and real-time capability.

Based on the data from these four scenarios, it can be concluded that the proposed method shows significant improvement in small target detection, particularly in scenarios where other models perform poorly, while maintaining extremely high processing speed and high precision.

The qualitative comparison of experimental results: The qualitative comparison of experimental results under four different backgrounds is shown in Figure 13, arranged from top to bottom as urban, mountain, ocean, and sky backgrounds, and from left to right showing the experimental results of the various models. Red boxes indicate correct target detection results, yellow circles indicate incorrect detection results, and no display indicates undetected targets. For ease of overall comparison, the detected target portions are magnified in the upper right corner marked by red boxes.

Table 8. Detection performance of various models in the sea scene.

Different Methods	Sea Scene			
	mAP@0.5:0.95	**mAP@0.5**	**mAP_s**	**Paras (M)**
BoxInst	--	--	--	273.8
CondInst	0.292	0.819	0.292	272.6
SOLOv2	--	--	--	372.0
Mask-Rcnn	0.463	0.933	0.463	353.3
YOLOv5n-seg	0.342	0.886	0.342	4.1
YOLOv7	0.335	0.930	0.355	76.3
YOLOv8n-seg	0.337	0.829	0.337	9.6
Yolact++	0.163	0.445	--	199.0
Proposed method	0.458	0.920	0.458	6.2

Table 9. Detection performance of various models in the sky scene.

Different Methods	Sky Scene			
	mAP@0.5:0.95	**mAP@0.5**	**mAP_s**	**Paras (M)**
BoxInst	0.395	0.806	0.397	273.8
CondInst	0.673	0.977	0.648	272.6
SOLOv2	0.686	0.934	0.664	372.0
Mask-Rcnn	0.711	0.987	0.703	353.3
YOLOv5n-seg	0.572	0.966	0.533	4.1
YOLOv7	0.580	0.974	0.561	76.3
YOLOv8n-seg	0.591	0.957	0.565	9.6
Yolact++	0.561	0.958	--	199.0
Proposed method	0.734	0.987	0.734	6.2

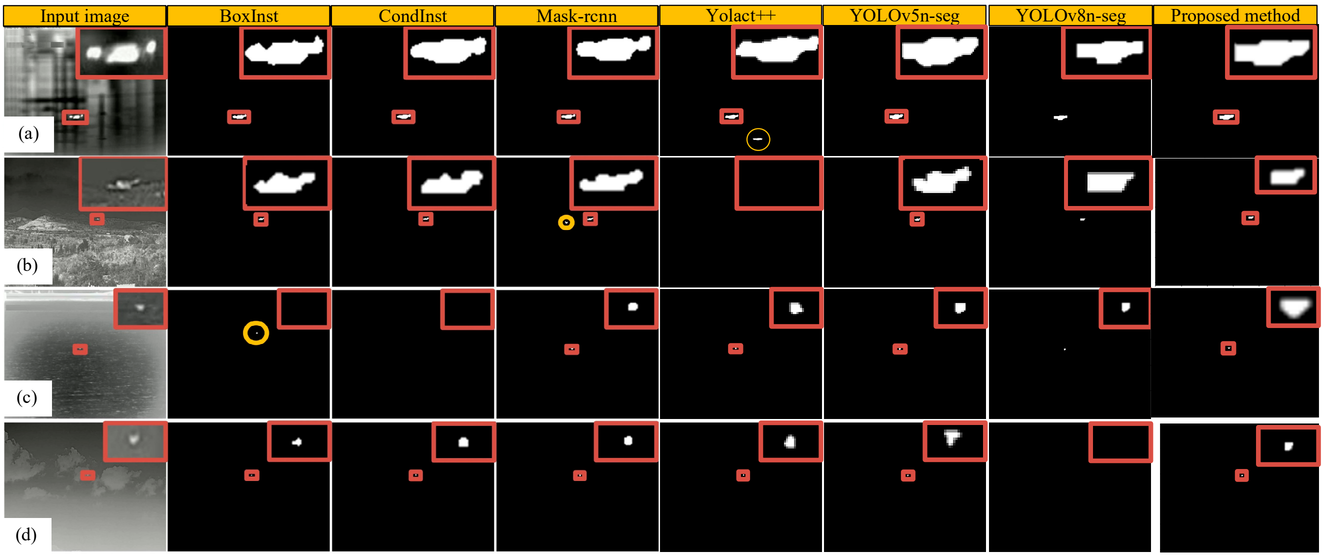

Figure 13. Original image and detection results. (**a**–**d**) shows four didderent images of original image and detection image.

The comparative experiments show that the BoxInst algorithm produced detection errors in ocean backgrounds, Mask-RCNN produced detection errors in mountain backgrounds, and YOLACT++ produced incorrect detection results in urban backgrounds. Overall, the proposed method demonstrated superior detection performance, achieving good detection results across all background types.

Combining quantitative and qualitative experimental results, the proposed method effectively completes low-altitude small UAV target detection tasks in complex scenarios while maintaining a small parameter size and high real-time performance, comprehensively improving target detection performance.

4.5. Ablation Studies

To justify the contribution of the proposed image enhancement method, we carried out an ablation study comparing the final detection performance (mAP) of the full model against versions using: (a) no image enhancement, (b) a simple baseline like contrast-limited adaptive histogram equalization (CLAHE), and (c) the proposed image enhancement. The results of the ablation study on image enhancement methods are presented in Table 10. The quantitative comparison unequivocally demonstrates the contribution of our proposed enhanced Gaussian filtering method to the final detection performance. Across all four challenging scenarios, our method consistently achieved the highest scores in all three metrics (mAP@0.5:0.95, mAP@0.5, and mAPs).

Table 10. Detection performance of various models in the sky scene.

Image Enhancement Methods	**Urban Scene**		
	mAP@0.5:0.95	**mAP@0.5**	**mAP$_s$**
no image enhancement	0.607	0.916	0.606
with CLAHE	0.612	0.940	0.639
with proposed image enhancement	0.668	0.960	0.661
Image enhancement methods	**Mountain scene**		
	mAP@0.5:0.95	**mAP@0.5**	**mAP$_s$**
no image enhancement	0.397	0.677	0.331
with CLAHE	0.415	0.723	0.414
with proposed image enhancement	0.426	0.741	0.426
Image enhancement methods	**Sea scene**		
	mAP@0.5:0.95	**mAP@0.5**	**mAP$_s$**
no image enhancement	0.401	0.904	0.430
with CLAHE	0.449	0.915	0.440
with proposed image enhancement	0.458	0.920	0.458
Image enhancement methods	**Sky scene**		
	mAP@0.5:0.95	**mAP@0.5**	**mAP$_s$**
no image enhancement	0.713	0.971	0.725
with CLAHE	0.729	0.972	0.734
with proposed image enhancement	0.734	0.987	0.734

Notably, the performance gain is most substantial in the most complex urban environment, where our method outperforms both "no enhancement" and CLAHE by a significant margin. This indicates that our method is particularly effective at handling severe noise and clutter, which are predominant in such scenes. While CLAHE also provides a noticeable improvement over no enhancement, its effect is less pronounced than our method, especially for the more stringent mAP@0.5:0.95 metric. This suggests that our proposed enhancement offers a more robust and targeted improvement, likely due to its integrated noise suppression and edge-preserving capabilities, which are crucial for preparing the image for small target detection. The results in the simpler sky scene, where all methods perform well, further confirm that our enhancement does not degrade performance in already favorable conditions. In conclusion, this ablation study validates that our proposed image enhancement module is a key contributor to the state-of-the-art performance of the overall detection framework, justifying its design complexity through superior results.

5. Experiments

5.1. Experimental System Setup

5.1.1. Data Acquisition Platform

The remote data acquisition platform consists of a UAV, payload, and airborne communication module, as shown in Figure 14.

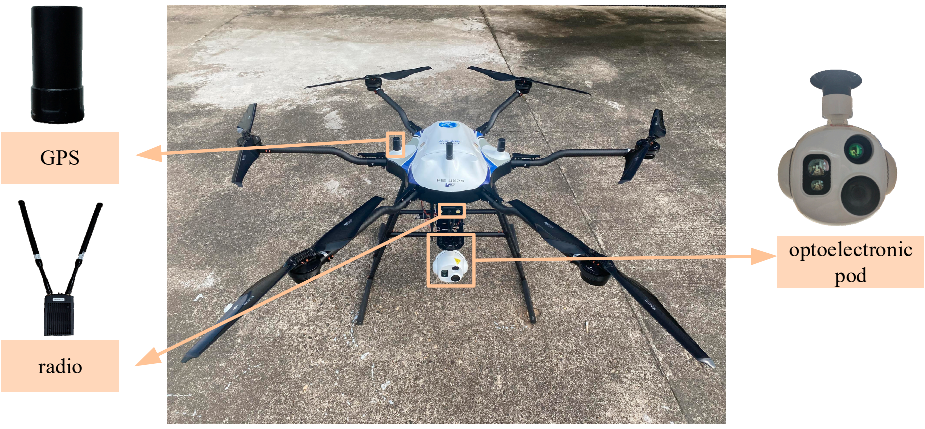

Figure 14. Schematic diagram of the remote data acquisition platform structure.

The optoelectronic pod employed is the Q30TIRM electro-optical reconnaissance pod (China), featuring high precision and stabilization capabilities. Its integrated shock absorption and gimbal design significantly enhances image acquisition stability and quality. Equipped with a 25mm lens supporting infrared imaging at 640 × 480 resolution, the payload offers picture-in-picture switching and electronic zoom functions, facilitating effective image acquisition for subsequent algorithm processing and validation.

During image acquisition, the payload was mounted on the UAV, which remained grounded throughout. The payload was connected to the onboard communication module via Ethernet, transmitting captured data through the terminal communication module to the processing platform for algorithm validation. The real-world validation dataset was captured under diverse but representative environmental conditions to test the robustness of our system. Data collection sessions were conducted during both day and night to cover varying illumination conditions. The weather during capture was clear and dry, with an ambient temperature ranging from approximately 15 °C to 25 °C. The terrain surrounding the stationary UAV included a mix of open fields, asphalt paving, and low vegetation, providing a variety of background thermal signatures.

5.1.2. Data Processing Platform

The data processing platform comprises a display control terminal, processing platform, communications antenna, and switch, as illustrated in Figure 15.

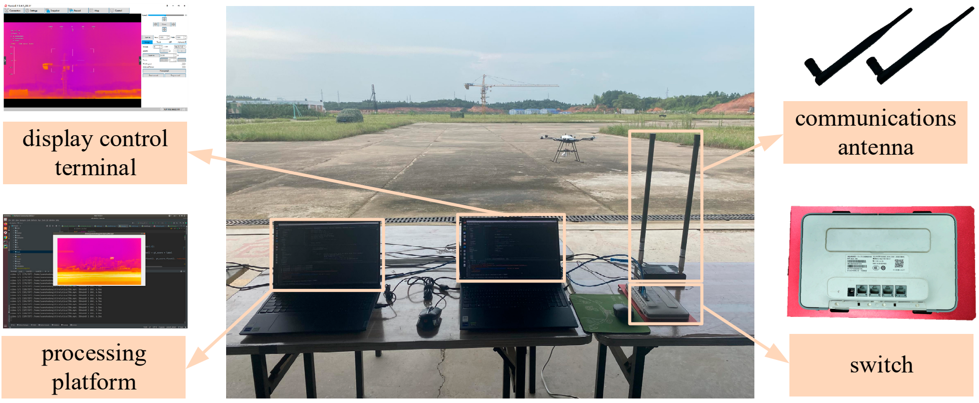

Figure 15. Schematic diagram of data processing platform structure.

The display control terminal visualizes infrared images transmitted from the terminal communication module along with angle and GPS parameters, while also controlling payload and shooting angles to ensure effective image capture. The backend processor hosts algorithms and performs real-time target detection on received images for algorithm validation. The terminal communication link receives images and connects to the switch, integrating all modules for coordinated command execution.

5.2. Dataset Construction

This section validates the improved algorithm's performance in practical applications. Instead of public datasets, we constructed a dataset using real-time infrared images captured by the payload, ensuring high timeliness and authenticity. Figure 16 shows sample images with UAV targets annotated in white boxes.

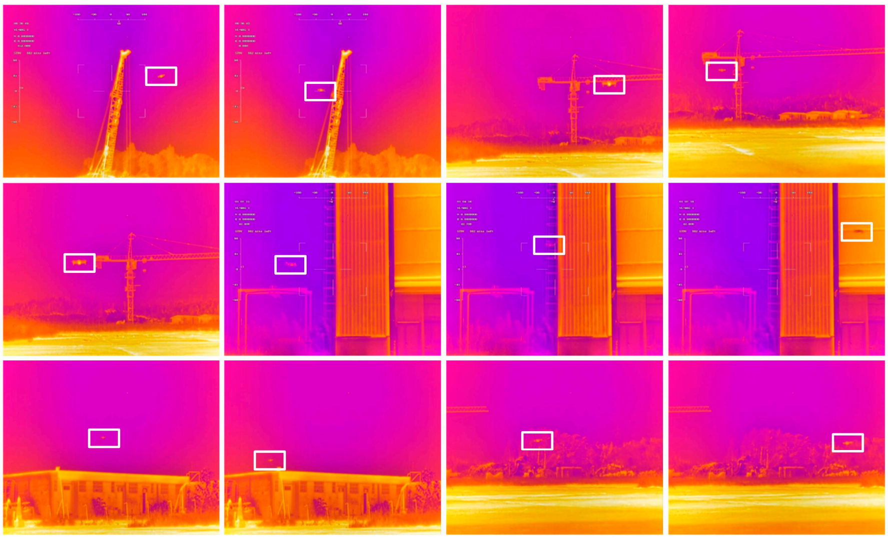

Figure 16. Sample images from the real-time dataset.

The images demonstrate that at larger pitch angles with sky backgrounds, interference is minimal and image quality is high; at smaller pitch angles with urban backgrounds, interference increases and image quality decreases.

5.3. Validation of Gaussian Filtering-Based Infrared Image Enhancement Algorithm

This section employs comparative experiments between original real-time images and enhanced images processed by the improved infrared enhancement algorithm. Results show significantly reduced noise interference and more prominent targets in enhanced images. Figure 17 illustrates comparisons between the original and enhanced images. Table 11 shows the UAV image contrast data.

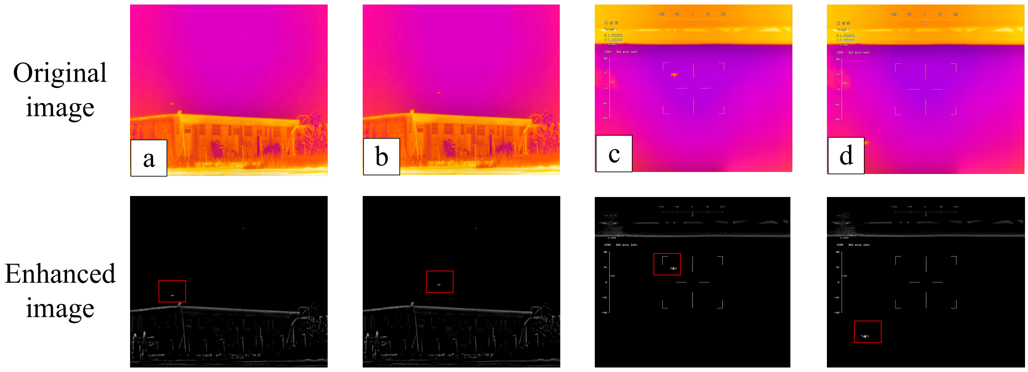

Figure 17. Comparison between enhanced and original images. (**a**–**d**) shows four didderent images of original image and enhanced image.

Table 11. UAV image contrast data.

Image Number	Original Contrast	Enhanced Contrast	PSNR	SSIM
a	31.04	38.73	7.12	0.05
b	31.03	38.77	7.19	0.05
c	15.29	41.79	8.10	0.04
d	31.17	44.48	8.05	0.04

Under real experimental conditions, overall image contrast is low. After enhancement processing, contrast improvements remain modest, with increases of only 7.69, 7.74, 26.5, and 13.31 in the examples shown. Enhanced images still exhibit contrast issues and suboptimal quality, showing some effectiveness compared to public dataset results, but with overall performance degradation.

5.4. *Validation of Proposed Method*

Experiments across different IoU thresholds and target types demonstrate that the proposed method has superior detection performance. For small targets, it achieves 54.7% average precision, outperforming YOLOv5n-seg by 17.5% and YOLOv8n-seg by 11%. Under stricter criteria (averaged IoU threshold 0.5–0.95), it maintains 51.2% average precision, exceeding YOLOv5n-seg and YOLOv8n-seg by 17.1% and 13.9%, respectively. Overall detection accuracy reaches 97.2%, surpassing YOLOv5n-seg and YOLOv8n-seg by 14% and 5%, respectively, confirming the proposed method's robust performance under real conditions. Detailed results are shown in Table 12.

Table 12. Target detection performance of algorithms on the real-time dataset.

Different Methods	$\mathrm{mAP_s}$	mAP@0.5	mAP@0.5:0.95
YOLOv5n-seg	0.372	0.832	0.341
YOLOv8n-seg	0.437	0.922	0.373
Proposed method	0.547	0.972	0.512

Figure 18 demonstrates the proposed method's detection results on the real-time dataset. Despite significant interference, small target size, and indistinct features in real-time images, the proposed improved algorithm accurately detects target positions, exhibiting excellent detection performance.

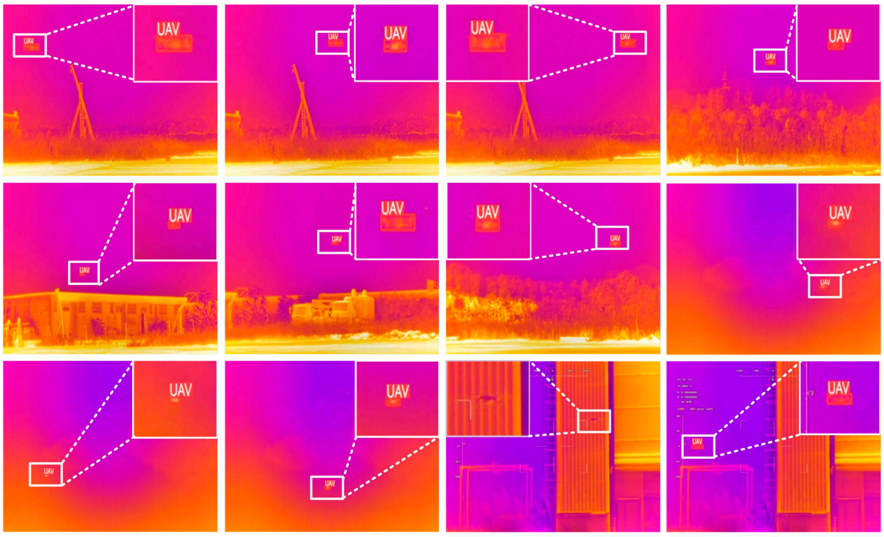

Figure 18. Sample detection results of the proposed method.

6. Conclusions

This paper presented a comprehensive framework for infrared UAV target detection that addresses the fundamental challenges of small target detection in complex backgrounds. Through the integration of an improved Gaussian filtering-based image enhancement module and a specialized deep learning architecture, we achieved significant improvements in both detection accuracy and processing efficiency. The proposed improved Gaussian filtering algorithm successfully addressed the limitations of traditional filtering approaches by incorporating a vertical weight function that effectively handles abnormal feature values. Experimental results demonstrated substantial contrast improvements across diverse scenarios, with the enhanced images showing clearer target features and reduced noise interference. This preprocessing step proved crucial for improving the performance of subsequent detection algorithms. Our SODMamba backbone, featuring deep feature perception modules (DFPMs) and ODSSBlocks, demonstrated exceptional capability in extracting and preserving small target features. The DFPM's focus on high-frequency components proved particularly effective for enhancing small UAV signatures while suppressing background clutter. The hierarchical architecture with multi-scale detection heads enabled robust detection across varying target sizes and distances. Comprehensive experiments on the custom SIDD dataset validated the superiority of our approach. The method achieved outstanding performance across all four scenarios (urban, mountain, sea, and sky), with particularly notable improvements in challenging urban environments where existing methods struggled. Real-world validation using a complete UAV detection system further confirmed the practical value of our approach. Despite the additional challenges of real-time data acquisition and processing, the system maintained robust detection performance, demonstrating its readiness for deployment in actual security applications. Despite these strengths, the current computational footprint of the model may present a challenge for deployment on extremely resource-constrained edge devices.

Future work will focus on several directions: (1) extending the method to handle even smaller targets and longer detection ranges, (2) incorporating temporal information for improved tracking capabilities, (3) optimizing the model for edge deployment through techniques such as model pruning, quantization, and knowledge distillation to significantly reduce its computational and memory requirements without substantial loss in performance, and (4) exploring the integration with other sensor modalities for more robust multi-sensor fusion systems.

Author Contributions: P.W. was the main author of the work in this paper. P.W. conceived the methodology of this paper, designed the experiments, collected and analyzed the data, and wrote the paper. Z.Z. was responsible for directing the writing of the paper and the design of the experiments. S.S. was responsible for managing the project and providing financial support. B.Z. was responsible for assisting in the completion of the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Provincial Natural Science Foundation of Hunan 2020JJ5672, the Hunan Provincial Innovation Foundation for Postgraduate CX20210020 and National University of Defense Technology Innovation Research Fund Program ZK25-51.

Data Availability Statement: Data available upon request due to reasons of sensitivity. The data presented in this study are available upon reasonable request from the corresponding author.

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

References

1. Seidaliyeva, U.; Akhmetov, D.; Ilipbayeva, L.; Matson, E.T. Real-time and accurate drone detection in a video with a static background. *Sensors* **2020**, *20*, 3856. [CrossRef] [PubMed]
2. Zheng, Y.; Chen, Z.; Lv, D.; Li, Z.; Lan, Z.; Zhao, S. Air-to-air visual detection of micro-UAVs: An experimental evaluation of deep learning. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1020–1027. [CrossRef]
3. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Zhao, J.; Guo, G.; Han, Z. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. *IEEE Trans. Multimed.* **2022**, *25*, 486–500. [CrossRef]
4. Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 91–124. [CrossRef]
5. Zhao, J.; Zhang, J.; Li, D.; Wang, D. Vision-based anti-UAV detection and tracking. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25323–25334. [CrossRef]
6. Liu, S.; Liu, D.; Srivastava, G.; Polap, D.; Woźniak, M. Overview and methods of correlation filter algorithms in object tracking. *Complex Intell. Syst.* **2021**, *7*, 1895–1917. [CrossRef]
7. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large selective kernel network for remote sensing object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 16794–16805.
8. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
9. Chen, C.; Zheng, Z.; Xu, T.; Guo, S.; Feng, S.; Yao, W.; Lan, Y. YOLO-based UAV technology: A review of the research and its applications. *Drones* **2023**, *7*, 190. [CrossRef]
10. Zhang, R.; Shao, Z.; Huang, X.; Wang, J.; Li, D. Object detection in UAV images via global density fused convolutional network. *Remote Sens.* **2020**, *12*, 3140. [CrossRef]
11. Wang, X.; Gao, L.; Song, J.; Shen, H. Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 510–514. [CrossRef]
12. Li, S.; Jin, W.; Li, L.; Li, Y. An improved contrast enhancement algorithm for infrared images based on adaptive double plateaus histogram equalization. *Infrared Phys. Technol.* **2018**, *90*, 164–174. [CrossRef]
13. Wang, B.J.; Liu, S.Q.; Li, Q.; Zhou, H.X. A real-time contrast enhancement algorithm for infrared images based on plateau histogram. *Infrared Phys. Technol.* **2006**, *48*, 77–82. [CrossRef]
14. Kuang, X.; Sui, X.; Liu, Y.; Chen, Q.; Gu, G. Single infrared image enhancement using a deep convolutional neural network. *Neurocomputing* **2019**, *332*, 119–128. [CrossRef]
15. Li, J.; Feng, X.; Hua, Z. Low-light image enhancement via progressive-recursive network. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4227–4240. [CrossRef]
16. Zhang, T.; Wu, H.; Liu, Y.; Peng, L.; Yang, C.; Peng, Z. Infrared small target detection based on non-convex optimization with Lp-norm constraint. *Remote Sens.* **2019**, *11*, 559. [CrossRef]
17. Liu, J.; He, Z.; Chen, Z.; Shao, L. Tiny and dim infrared target detection based on weighted local contrast. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1780–1784. [CrossRef]
18. Han, J.; Moradi, S.; Faramarzi, I.; Liu, C.; Zhang, H.; Zhao, Q. A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1822–1826. [CrossRef]
19. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]

20. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
21. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
22. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **2023**, *32*, 1745–1758. [CrossRef] [PubMed]
23. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior attention-aware network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002013.
24. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 877–886.
25. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric contextual modulation for infrared small target detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 950–959.
26. Cui, Z.; Yang, J.; Li, J.; Jiang, S. An infrared small target detection framework based on local contrast method. *Measurement* **2016**, *91*, 405–413. [CrossRef]
27. Yuan, S.; Qin, H.; Yan, X.; Yan, X. SCTransNet: Spatial-Channel Cross Transformer Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5002615. [CrossRef]
28. Liu, J.; Sun, H.; Katto, J. Learned Image Compression with Mixed Transformer-CNN Architectures. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canad, 17–24 June 2023.
29. Wang, B.; Wang, Y.; Mao, Q.; Cao, J.; Zhang, H.; Zhang, L. Lightweight Infrared Small Target Detection Method Based on Linear Transformer. *Remote Sens.* **2025**, *17*, 2016. [CrossRef]
30. Ding, L.; Xu, X.; Cao, Y.; Zhai, G.; Wang, F.; Qian, Z. Detection and tracking of infrared small target by jointly using SSD and pipeline filter. *Digit. Signal Process.* **2021**, *110*, 102949. [CrossRef]
31. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
32. Xu, C.; Wang, J.; Yang, W.; Yu, L. Normalized Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
33. 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]
34. Liu, Y.; Liao, L.; Wu, H.; Qin, J.; He, L.; Yang, G.; Zhang, H.; Zhang, J. Trajectory and image-based detection and identification of UAV. *Vis. Comput.* **2021**, *37*, 1769–1780. [CrossRef]
35. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.

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.