applied sciences

MDPI

Article

Infrared Dim and Small Target Detection Based on Local–Global Feature Fusion

Xiao Ling, Chuan Zhang, Zhijun Yan, Bo Wang *, Qinghong Sheng and Jun Li

College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; xlingsky@nuaa.edu.cn (X.L.); 2115003@nuaa.edu.cn (C.Z.); yanzj@nuaa.edu.cn (Z.Y.); qhsheng@nuaa.edu.cn (Q.S.); jun.li@nuaa.edu.cn (J.L.)
* Correspondence: wangbo_nuaa@nuaa.edu.cn

Abstract: Infrared detection, known for its robust anti-interference capabilities, performs well in all weather conditions and various environments. Its applications include precision guidance, surveillance, and early warning systems. However, detecting infrared dim and small targets presents challenges, such as weak target features, blurred targets with small area percentages, missed detections, and false alarms. To address the issue of insufficient target feature information, this paper proposes a high-precision method for detecting dim and small infrared targets based on the YOLOv7 network model, which integrates both local and non-local bidirectional features. Additionally, a local feature extraction branch is introduced to enhance target information by applying local magnification at the feature extraction layer allowing for the capture of more detailed features. To address the challenge of target and background blending, we propose a strategy involving multi-scale fusion of the local branch and global feature extraction. Additionally, the use of a 1×1 convolution structure and concat operation reduces model computation. Compared to the baseline, our method shows a 2.9% improvement in mAP_{50} on a real infrared dataset, with the detection rate reaching 93.84%. These experimental results underscore the effectiveness of our method in extracting relevant features while suppressing background interference in infrared dim and small target detection (IDSTD), making it more robust.

Keywords: infrared dim and small target; local–non-local feature fusion; vision system; target detection

check for updates

Citation: Ling, X.; Zhang, C.; Yan, Z.; Wang, B.; Sheng, Q.; Li, J. Infrared Dim and Small Target Detection Based on Local–Global Feature Fusion. *Appl. Sci.* **2024**, *14*, 7878. https://doi.org/10.3390/app14177878

Academic Editor: Andrea Prati

Received: 30 July 2024
Revised: 25 August 2024
Accepted: 3 September 2024
Published: 4 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the field of infrared target detection, the imaging system captures images based on the thermal radiation signals emitted by the target within the 8–12 μm wavelength range. Since infrared imaging depends on the target's temperature and material properties, it demonstrates strong resistance to interference and can be used effectively in a wide range of conditions. This makes it indispensable for precision guidance, surveillance, and tracking [1–3]. According to the definition of dim and small targets provided by the International Society for Optical Engineering (SPIE), a target is considered dim and small if it occupies no more than 0.15% of the total image, has a size of fewer than 9×9 pixels (typically ranging from 2×2 to 10×10 pixels), and has a local signal-to-noise ratio (LSNR) of no greater than 5 dB [4,5]. Compared to general target detection, infrared dim and small target detection has distinct characteristics. First, the target is small and has an inconspicuous structure in long-distance detection and early warning guidance under infrared backgrounds, so it lacks obvious shape or texture features that are easily extracted. Second, the target has a low LSNR, making it more susceptible to being obscured by complex backgrounds compared to visible light images. As a result, frequent false detections and alarms occur, significantly reducing the accuracy of target detection in intricate environments.

The fundamental concept behind infrared dim and small target detection algorithms involves background suppression and target energy enhancement. Currently, research

methods in this field can be divided into two categories: detect before tracking (DBT) and tracking before detect (TBD). DBT, also known as the detection method based on measured data from the current frame, enhances the target and suppresses background clutter in the image containing the target, then sets a threshold to segment the target from the processed image. Clutter refers to background information or interference unrelated to the target, including areas of the background and noise that are similar to the target. Traditional single-frame detection methods include wavelet transform [6], top-hat [7] and two-dimensional least mean square filtering [8]. Lei et al. [9] proposed a denoising threshold function suitable for rotating multi-element infrared detection systems, which better retains target features, though this method is not robust against varying background clutters. Li et al. [10] employed the Gaussian Laplace filter to obtain target response maps with both large and small pixel values, extracting targets through adaptive threshold segmentation. While this approach achieved good detection results, the algorithm's high complexity limits its use for real-time detection. Deng et al. [11] utilize a continuous adaptive structure combined with quantum genetics to optimize the overall algorithm, but this approach does not align well with the geometric characteristics of image sequences, and it may also distort the shape of the target due to structural elements that are either too large or too small.

TBD, or tracking before detect, refers to a detection method based on multi-frame images, utilizing spatio-temporal information from continuous frames, such as motion paths, brightness, and scale changes to detect the targets with consistent trajectory or velocity priors. It firstly tracks a target by considering hypothetical potential paths in the multi-frame data, continuously accumulates the target's energy, and finally screens candidate targets using a discriminant operator. Since TBD sets no threshold or merely a low threshold for single-frame data, it retains as much information about dim and small targets as is possible, helping to avoid the target loss problem faced by the DBT method. By accumulating insights from multiple frames, TBD aims to accurately pinpoint the target's position within the image. Typical methods include the three-dimensional matching filter [12], the optical flow method [13], the Kalman filter [14], and the particle filter [15]. Zhang et al. [16] proposed a coarse-to-fine three-dimensional directional search filter based on three-dimensional matching filtering to reduce computational demands. However, the computational complexity remains high, leading to significant detection time costs. Fu et al. [17] developed a 3D Hough transform (3D-HT) algorithm that uses two sets of coordinates and their errors from different times to construct a 3D pipeline, but it struggles to meet real-time performance requirements when the target's speed changes rapidly. Traditional algorithms rely heavily on manually designed features, which demand extensive debugging and struggle to accommodate the diverse detection needs across various scenarios.

In recent years, the rapid progress of deep learning has revolutionized image processing, particularly in optical target detection and recognition. The integration of deep learning techniques into classifier and algorithm design continues to be a major research focus in this field [18,19]. Deep learning-based object detection methods are generally categorized into two types. The first type uses convolutional layers to extract target features and generate numerous candidate regions where the target may exist. The target detection network then outputs information such as target position and category within regions, as seen in methods like region convolutional neural networks (R-CNN) [20], fast RCNN [21], and others. The second type directly predicts the location, category, and other relevant information for each small region of the input image, such as in the single shot multibox detector (SSD) [22]. Deep learning-based DBT methods like those of Feng et al. have modified the structure of the channel attention module (CAM) and spatial attention module (SAM) to achieve multi-scale feature fusion [23]. However, feature learning for dim and small targets with heavy background clutter remains challenging due to the lack of clear texture and shape features in infrared dim and small targets. Fan et al. [24] proposed an enhanced convolutional neural network (CNN) for infrared images based on single-frame target detection and recognition principles, but it also introduced more noise. Wang et al. [25]

incorporated the adaptive fusion attention module (AFAM) into the you only look once (YOLO)v5 model to weight features and improve detection accuracy, but its precision remains insufficient in complex scenes.

Deep learning-based TBD methods like those of Hare et al. [26] extract short-term and space-dimensional information from 15 consecutive infrared image sequences using a 3D convolutional kernel, combined with a convolutional long short-term memory network (LSTM) to capture long-term spatio-temporal information. However, issues with missed detections and false alarms persist, and the method does not meet real-time requirements. Lee [27] combined the LSTM prediction network with the duality of convolutional neural networks (Du-CNN) classification network to create the cross duality of neural network (CR-DUNN), which classifies targets from clutter by fully utilizing information from both the present and future moments. Li et al. [28] improved detection accuracy by around 5% through the bidirectional fusion of deep and shallow features, along with skip connections to capture richer contextual information. However, their network model is large and contains too many parameters, resulting in high computational resource consumption. These methods are still falling short in accurately detecting dim and small infrared targets in complex scenes, with continued occurrences of missed and false detections.

This paper introduces an algorithm for infrared dim and small target detection based on local–global feature fusion to address these challenges. We are inspired by the convolutional block attention module (CBAM) [29], which enhances target features in both spatial and channel dimensions. Specifically, we design a multi-scale module that integrates global and local features, establishing a fusion mechanism by leveraging dual information.

The magnification of the focus area enhances target features, improves target saliency, and achieves accurate positioning by fusing with global feature information. In summary, the main contributions of this work are as follows:

(1) To address the challenge of limited feature information in infrared dim and small targets, we propose a novel module that zooms in on the focus area. This focus area is initially randomly determined and then dynamically updated to match the target area based on the output of the algorithmic model. By enhancing the detailed information in this area while maintaining the overall characteristics of the background, our approach improves detection performance for these challenging targets.
(2) To tackle the low contrast between infrared dim and small targets and their background, this study optimizes the multi-scale fusion structure within a dual-path feature extraction network, which captures positional and semantic information at different stages, enabling precise detection.

2. Methods

We present the proposed local–global feature fusion from dual input for dim and small target detection in detail. Shown in Figure 1 is the basic network architecture.

YOLOv7 strikes a good balance between speed and accuracy, making it ideal for applications that require real-time detection. The structure of YOLOv7 [30] is based on the cross stage partial network (CSPNet), which consists of a backbone network and a head network. The backbone network primarily comprises conv-batchnorm-SiLU (CBS), max-pooling (MP), and efficient layer attention network (ELAN) modules. ELAN effectively combines shallow and deep features through a multi-channel branching structure, which captures detailed information of the target at different scales without significantly increasing computational cost. This is especially important for detecting small targets and handling complex backgrounds. The head network consists of three target detectors that use a grid-based anchor mechanism to detect targets across feature maps of different scales. With the infrared dim and small target size of 0.15%, which falls within the small range, using multiple inspection heads to detect targets on feature maps of different resolutions can improve both the comprehensiveness and accuracy of detection. Building upon the YOLOv7 architecture, our model strategically divides the input image into four quadrants, carefully balancing the trade-off between computational speed and detection

accuracy. Specifically, we enhance the feature information related to dim and small targets by magnifying the local areas within the image. This localized information is then fused with the global features extracted from the entire infrared image, resulting in improved details for specific objects of interest. Additionally, our approach increases the receptive field, expanding the spatio-temporal context of the target. This is achieved through a dual-branch fusion mechanism that combines multi-scale feature information, allowing the weight distribution of the feature map to be dynamically adjusted. Further elaboration and specific scenarios are detailed in the remainder of this paper.

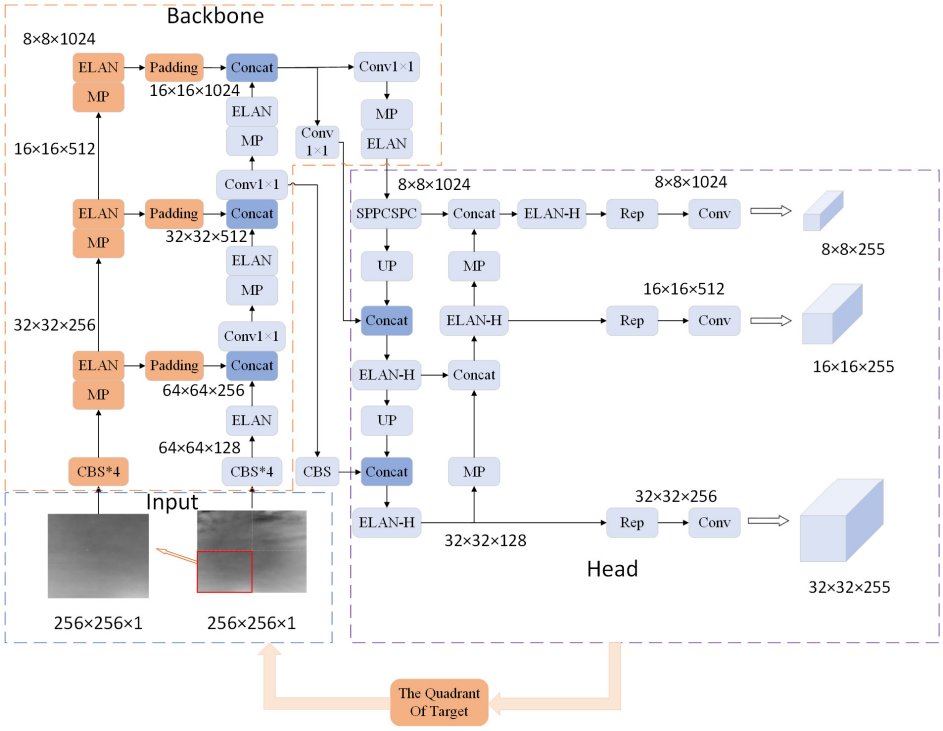


Figure 1. Improved YOLOv7 network structure diagram. The numbers below represent the size of each feature map. The CBS module is the basic component of the model, consisting of the conv module and the BN module.

2.1. Local Zoom Input

The fundamental principle in object detection is based on enhancing target features while simultaneously suppressing background noise. In deep learning, this involves strengthening the feature representation of the target and reducing the focus on background features, all within the constraints of limited computational resources. The vision focus mechanism, which mimics the selective focus used by organisms when processing external information, plays a crucial role in this process. By isolating key information relevant to the area of interest or specific objectives from the global context, vision focus mechanisms improve the model's discriminative power. The core of this method lies in extracting essential features related to the task at hand from a large pool of data and amplifying their importance by adjusting the dimensions of the original network input, thereby improving the model's performance.

As shown in Figure 2, the local branch input divides the image into four quadrants, with each quadrant representing a potential area for applying focus magnification, simulating the biological vision focus mechanism. This process enhances the capture of detailed information from dim and small targets more effectively. The vision process is divided into two stages at the feature extraction layer. In the initial stage, the system acquires various features in an unbiased manner, without incorporating vision focus mechanisms. In the next stage, a series of integrations are performed on the features obtained from

the first stage to form a cognitive representation, which then becomes the focus of visual attention mechanisms. This two-stage process ensures efficient feature extraction without compromising the pixel definition of the original image.

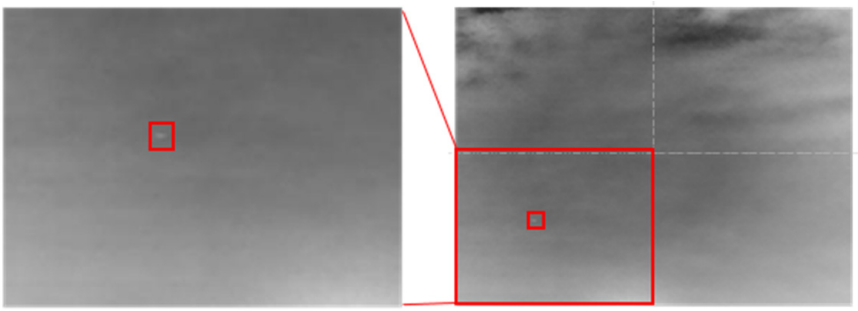


Figure 2. Vision focus simulation with local input. Dim and small targets are marked with small red rectangle.

The selected local area for applying vision focus is randomly determined in the initial frame of the input image. This selection is subsequently updated based on the output of the previous frame for the next frame. The spatio-temporal information derived from the image is used to update the magnified region. Given the weak characteristics of the target, we employ the double cubic interpolation (BiCubic) algorithm to zoom in on the selected region which is then input into the model. Equation (1) shows the above process:

$$ZF(x, y, z) = zoom(F(x, y, z), C) \tag{1}$$

where $zoom(\cdot)$ represents the bicubic interpolation magnification function applied to a selected quadrant C, and $F(x, y, z)$ denotes the processed frame. The target's coordinates (x, y) are used to determine the relocation of quadrant C, where the vision focus is applied. This design, which significantly enhances the local area, acts as an extension of the vision focus mechanism. It magnifies both the overall and detailed information of the target object of interest while enhancing its spatial location information, making the model more sensitive to dim and small targets.

2.2. Feature Fusion of Multi-Scale

Most methods primarily rely on advanced features for prediction to expedite detection. The feature map after multi-layer convolution has an expansive receptive field for the entire image, making global perception more prominent with rich semantic information. However, it is difficult to capture the spatial information of the target simultaneously, making systems that rely solely on high-level features less effective for detecting small-sized targets, particularly dim and small ones. In contrast, low-level features contain precise spatial information. Therefore, we fuse high-level features with low-level features to provide the final detection system with both ample semantic and spatial location information, improving the detection and precise positioning of dim and small targets in infrared imaging.

As shown in Figure 3, the input from the local region is fused with the convolutional layer of the global region input, facilitated by the backbone's convolutional layer. Essentially, the output channels are interconnected through a concat (channel connection) layer, which links the channels together. Afterward, the feature channels are processed through a 1×1 convolutional layer, which updates the focus weights for the dim and small target regions while eliminating superfluous information.

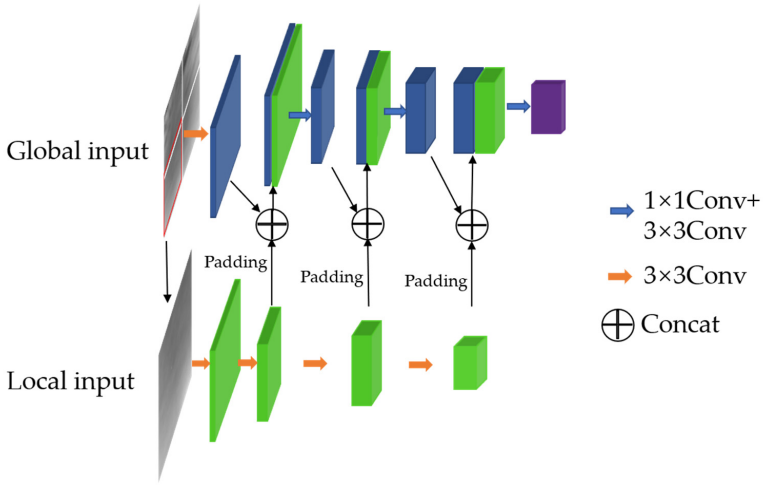


Figure 3. Dual-path feature fusion network.

As shown in Figure 4, the architecture of the branch input closely resembles that of the global input. The feature map from each local input is fused with the global feature map after passing through each convolution module, facilitated by the concat layer. The concat layer achieves feature fusion by increasing the number of channels rather than adding new feature maps. Due to the concat layer's output concatenating the input channel data, we apply a 1×1 convolutional layer after the concat layer to integrate the features of both the target and the background. Each convolution is followed by a SiLU activation. During the feature extraction stage at each branch scale, the feature map is downsampled using a 3×3 convolution with a stride of 2 and is then passed to the concat layer along with the other input to achieve multi-scale feature fusion. The final feature map in the backbone contains the largest receptive field, enriched with both semantic and spatial information. Equation (2) represents the process:

$$F = SiLU\left(f_{1\times1}\left(concat\left(F_l; F_g\right)\right)\right) \tag{2}$$

where $f_{1\times1}(\cdot)$ denotes the convolutional kernel of size 1×1, and $SiLU(\cdot)$ denotes the sigmoid linear unit operation. F_l and F_g correspond to the features of the local input and global input after the convolutional layer, respectively.

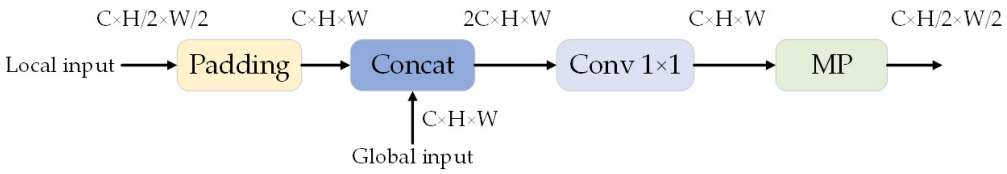


Figure 4. The structure of feature fusion.

3. Experiments

3.1. Data and Experiment Setup

The complexity of infrared dim and small target detection is measured by the local signal-to-noise ratio (LSNR). The smaller the LSNR, the smaller the intensity difference between the target and the background, making it more difficult to detect the target. The local signal-to-noise ratio (LSNR) is calculated as follows:

$$LSNR = 10\log_{10}(E_r - E_B)/\delta_B \tag{3}$$

where E_r is the mean intensity value of the target area, E_B is the mean intensity value of the background region surrounding the target, and δ_B is the intensity standard deviation of this background region, which is three times the size of the target.

To evaluate the performance and robustness of the proposed method, infrared images from four different scenes were used. The dataset includes a total of 12,500 images, including aerial surveillance, complex cloud background, and simple background, with each image containing one infrared dim and small target. One of these datasets was taken with a drone, five datasets were generated through simulation based on real infrared backgrounds and targets with varying LSNR, and one additional dataset was taken on the ground. Note that all simulated data were generated using a similar method [31] in the scene. Data1 and data2 depict sky scenes characterized by pronounced background edge clutter, with data1 captured by drones, while data3 and data4 show complex cloud scenes, where the target can easily blend into the clouds. Data5 and data6 consist of simpler smoother cloud scenes. Lastly, data7 include real images captured at night by a camera on the ground. These images are representative of the infrared dim and small targets in most sky backgrounds. All data were labeled by experts in the field of infrared detection and machine vision using rectangular box annotations. To ensure labeling quality, each infrared image was reviewed by three experts, with one assigned as the responsible expert. After the initial annotation by one expert, a second expert verifies it, and the responsible expert gives final confirmation.

Details of the dataset are shown in Table 1, and Figure 5 exhibits representative images under varying backgrounds and signal-to-noise ratio conditions. The targets within the images are marked with infrared rectangles, and all images are of the same size. To ensure accurate evaluation, we partitioned the dataset into training, validation, and testing sets using a 7:2:1 ratio, respectively. The proposed model was implemented in Python 3.8, running on Windows 11, with CUDA 11.4 using the PyTorch 1.11.0 backend and cuDNN 8.2.0. The system was set up on a workstation with an Intel Core i5-12400F@2.50 GHz processor with 64 GB of RAM and an NVIDIA GTX 3080Ti GPU with 12 GB of memory. The entire training process used a batch size of 64, and it ended after 300 epochs.

Table 1. Detailed information regarding data.

Dataset Number	Number of Images	Size	LSNR/dB
1	2000	256 × 256	2.22
2	2000	256 × 256	3.3
3	2000	256 × 256	0.49
4	2000	256 × 256	2.19
5	2000	256 × 256	−0.28
6	2000	256 × 256	2.15
7	500	256 × 256	−2.8

We utilized mAP_{50} and $mAP_{50:95}$ as comparative metrics for evaluating the model's performance. Mean average precision (mAP) reflects the average accuracy across the dataset, with mAP_{50} serving as an indicator of the algorithm's overall classification ability across different target types. $mAP_{50:95}$ measures the average detection accuracy across all 10 intersection over union (IoU) thresholds, ranging from 0.5 to 0.95, with a step size of 0.05. Generally, the higher the IoU threshold, the higher the demand on the model's regression ability, resulting in detection outcomes that align more closely with the actual target. In addition, we introduce the true positive rate (P_d) and false positive rate (F_a) [32] to further evaluate detection performance. These metrics are defined as follows:

$$P_d = \frac{\text{number of detected true detections}}{\text{numer of actual targets}} \tag{4}$$

$$F_a = \frac{\text{number of pixels for false detections}}{\text{number of total pixels in images}} \tag{5}$$

The P_d serves as a metric for evaluating the model's performance in terms of correct detection, representing the ratio of targets accurately predicted as targets to the total number of targets. Conversely, F_a reflects the model's precision in target detection, representing

the ratio of instances where the background is mistakenly predicted as a target to the total amount of the background. The P_d is plotted on the ordinate and the F_a on the abscissa to generate the receiver operating characteristic curve (ROC) for real data.

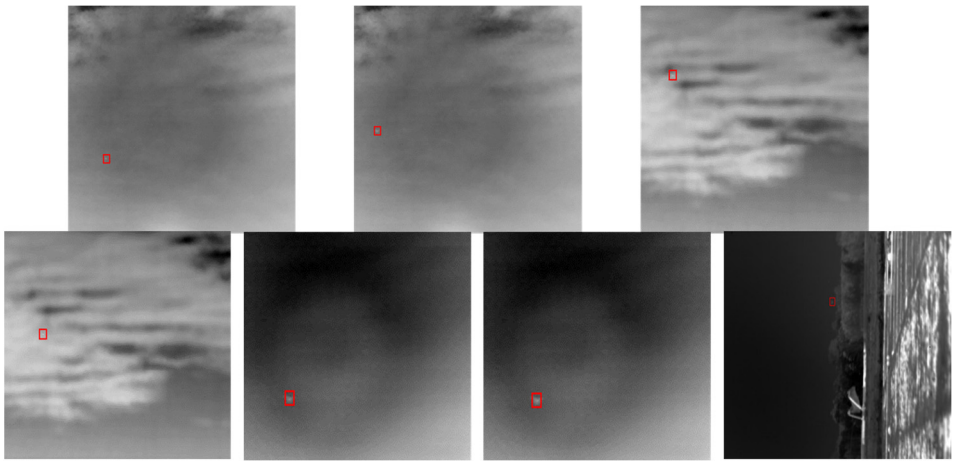


Figure 5. A frame of different data sequence images. Dim and small targets are marked with a red rectangle, and all images are the same size.

3.2. Experiments and Analysis

To demonstrate the effectiveness of the algorithm we proposed for infrared dim and small target detection, we conducted ablation experiments based on YOLOv7, as well as comparative experiments with existing advanced algorithms on datasets derived from a variety of scenarios.

3.2.1. Ablation Study

To comprehensively validate the sensitivity of the dual-path feature fusion extraction network to variation in infrared dim and small targets, as well as its detection efficacy across diverse scenes, we conducted experiments on datasets with different local signal-to-noise ratios in various scenes. Both the YOLOv7 model (baseline model) and the proposed model were employed. The outcomes of these experiments are shown in Table 2.

Table 2. Experimental results under different scenario data.

Dataset	Scene	mAP_{50}(Ours/YOLOV7)	$mAP_{50:95}$(Ours/YOLOV7)
1	Bright, thin cloud	**0.8942**/0.8463	**0.5758**/0.559
2		**0.9964**/0.97	**0.8064**/0.7535
3	Heavy cloud cover	**0.7793**/0.7522	**0.5537**/0.4984
4		**0.9917**/0.9567	**0.8408**/0.7924
5	Dim, thin cloud	**0.75**/0.7083	**0.5557**/0.4881
6		**0.9779**/0.9549	**0.8138**/0.7884

As is evident from the experimental results shown in Table 2, under various complex backgrounds, the algorithm employing local–global dual-path feature fusion demonstrates notable improvements in both mAP_{50} and $mAP_{50:95}$ compared to the original YOLOv7. Specifically, mAP_{50} exhibits an average increase of over 2%, while $mAP_{50:95}$ experiences an average increase of more than 3.6%. These findings underscore that the utilization of local magnification enhances the network's ability to extract both spatial and semantic feature information from the target. In the same scene, the experiment conducted by varying the local signal-to-noise ratio of the target reveals that even under conditions characterized by an extremely low local signal-to-noise ratio, the performance of the dual-path feature fusion algorithm shows an increase of more than 2.4% compared to the original YOLOv7 network,

despite a slight decline in detection accuracy. This underscores the efficacy of the dual-path fusion strategy in effectively enhancing the spatial and semantic feature information of dim and small targets within complex scenes, while simultaneously reducing the interference of noise.

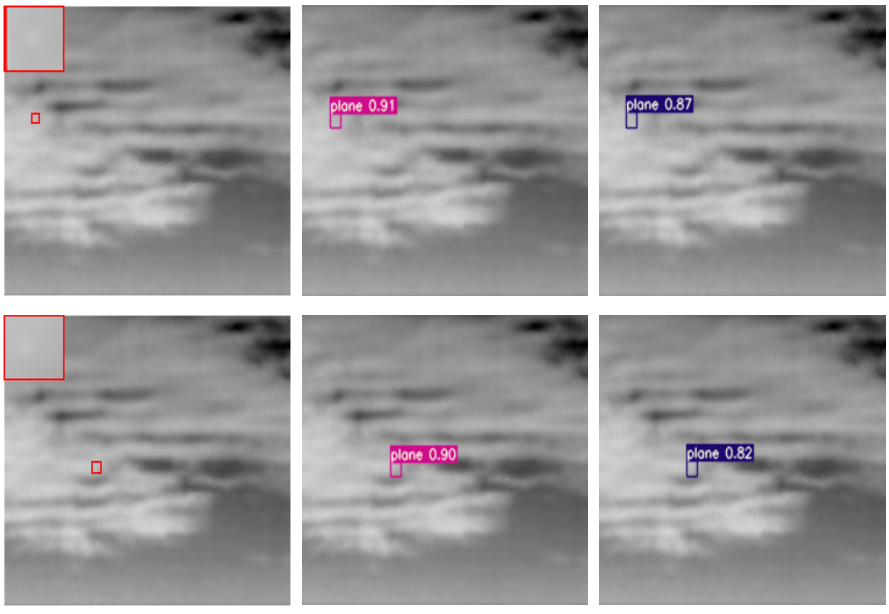
To verify the effectiveness of the network module proposed in this chapter for infrared dim and small target detection, we compared the mAP_{50} and $mAP_{50:95}$ results from ablation experiments conducted on data7. The experimental results are shown in Table 3.

Table 3. Ablation comparison on different combinations of modules.

	mAP_{50}	**$mAP_{50:95}$**
YOLOv7	0.9075	0.3721
YOLOv7 + Local input	0.9232	0.3769
YOLOv7 + Local input + Multiscale feature fusion	0.9384	0.4036

As shown in Table 3, scheme one refers to the original YOLOv7 network. Scheme two enhances the network by incorporating a local magnification branch for feature extraction. Scheme three, building on scheme two, performs feature fusion at the end of each feature extraction stage, integrating features from the main path. Notably, all schemes utilize the same pre-trained weights. As observed from the detection results of data7 in Table 3, the inclusion of a local magnification branch for infrared small targets has a certain inhibitory effect on the background. When the main network incorporates a local branch and performs feature fusion, both mAP_{50} and $mAP_{50:95}$ improve compared to the YOLOv7 model. This improvement is due to the fact that during feature extraction, the feature information of dim and small targets tends to dilute or even disappear as the network deepens. The local branch, however, has the ability to enhance the feature information of the dim and small targets during this process. Additionally, the multi-scale feature fusion helps retain this information, significantly boosting detection capabilities.

The following figure presents several detection result images of the YOLOv7 model and our proposed algorithm on the dataset. The post-detection images are displayed in Figure 6, where it is clear that our model demonstrates superior performance, as evidenced by a higher confidence score.



(**a**) (**b**) (**c**)

Figure 6. Object detection results: (**a**) original infrared image; (**b**) partial results of our method; (**c**) partial results of YOLOv7.

3.2.2. Comparison Experiments on Drone Vehicle

We compared the proposed infrared dim and small target detection algorithm with the following representative algorithms in this field: infrared patch-image (IPI) [31]; non-convex rank approximation minimization (NRAM) [33]; non-convex optimization with Lp-norm constraint (NOLC) [34]; partial sum of the tensor nuclear norm (PTSNN) [35], based on low-rank and sparse matrices; tri-layer local contrast measure (TLLCM) [36], based on the human vision system, which transforms target salience into a local optimization problem; new top-hat, based on mathematical morphology; local intensity and gradient properties (LIG) [37] based on vector fields; SSD; receptive field block net (RFBnet) [38]; and infrared small-target detection U-net (ISTRU-net) [39], based on deep learning. The parameters for each method are set according to their optimal experimental results, with detailed parameter settings shown in Table 4.

Table 4. Parameter settings of different methods.

Method	Parameter
IPI	Patch size: 45, step: 10
NRAM	Patch size: 50, step: 10
TLLCM	GS: [1 2 1;2 4 2; 1 2 1]/16, local window size: 3
PTSNN	Patch size: 20, step: 20, λ_L = 0.5
New Top-Hat	Inner scale: 4, outer scale: 9
LIG	k = 0.2, Patch size: N = 19
NOLC	Patch size: 30, step: 10, p:0.6
SSD	Iter: 10,000, batch size: 64, LR = 0.0001
RFBnet	Epoch: 300, batch size: 32, LR: 0.0004
ISTRU-net	Epoch: 500, batch size: 48, LR: 0.0001
Proposed	Epoch: 300, batch size: 64, LR = 0.001

The visualization results of different methods for various scenes and different local signal-to-noise ratio sequences are shown in Figure 7. The detected targets are marked with red rectangles, while false alarms are unmarked.

To evaluate the detection efficiency of various algorithms, Table 5 shows the detection time of each algorithm on different data sequences, providing an intuitive demonstration of the real-time performance of these algorithms. NOLC, NRAM, PSTNN, and IPI, which belong to low-rank sparse algorithms, generally exhibit moderate speeds. Additionally, IPI employs global fixed weights, which contributes to an increase in computation time. TLLCM requires a lengthy computation time due to its reliance on calculating locally related features. Although the new top-hat has a fast detection speed, it is characterized by a low detection rate and a high false alarm rate. LIG faces challenges in enhancing algorithm speed due to the time required for calculating multi-scale direction diversity. With sufficient computational resources, our proposed algorithm outperforms other deep learning algorithms, boasting the shortest average detection time across all datasets. It also shows an improvement of more than 16.3% in detection time compared to RFBnet, demonstrating superior real-time processing capability.

In the experiment, the model's inference speed on the PC is approximately 57 frames per second (FPS). In scenarios that require real-time processing, such as drone target detection and aerial monitoring, the algorithm can be deployed on field programmable gate arrays (FPGA) to achieve real-time processing capabilities, though several challenges remain. This is part of our ongoing research, and we aim to further optimize the model and increase the processing speed to meet real-time demands.

Table 5. Average time consumption (sc) of data1–7 obtained by 11 methods.

Methods	Data1	Data2	Data3	Data4	Data5	Data6	Data7
IPI	0.322	0.294	0.312	0.291	0.309	0.296	0.315
NRAM	0.145	0.146	0.151	0.153	0.162	0.164	0.157
TLLCM	0.183	0.185	0.188	0.193	0.185	0.186	0.221
PTSNN	0.082	0.073	0.072	0.071	0.081	0.079	0.0734
New Top-Hat	0.026	0.027	0.023	0.026	0.024	0.022	0.028
LIG	0.131	0.132	0.136	0.137	0.125	0.124	0.236
NOLC	0.137	0.134	0.154	0.153	0.145	0.143	0.162
SSD	0.066	0.069	0.056	0.058	0.076	0.079	0.080
RFBnet	0.019	0.021	0.022	0.023	0.021	0.020	0.021
ISTRU-net	0.285	0.287	0.276	0.277	0.289	0.278	0.351
Proposed	**0.017**	**0.018**	**0.017**	**0.018**	**0.018**	**0.017**	**0.018**

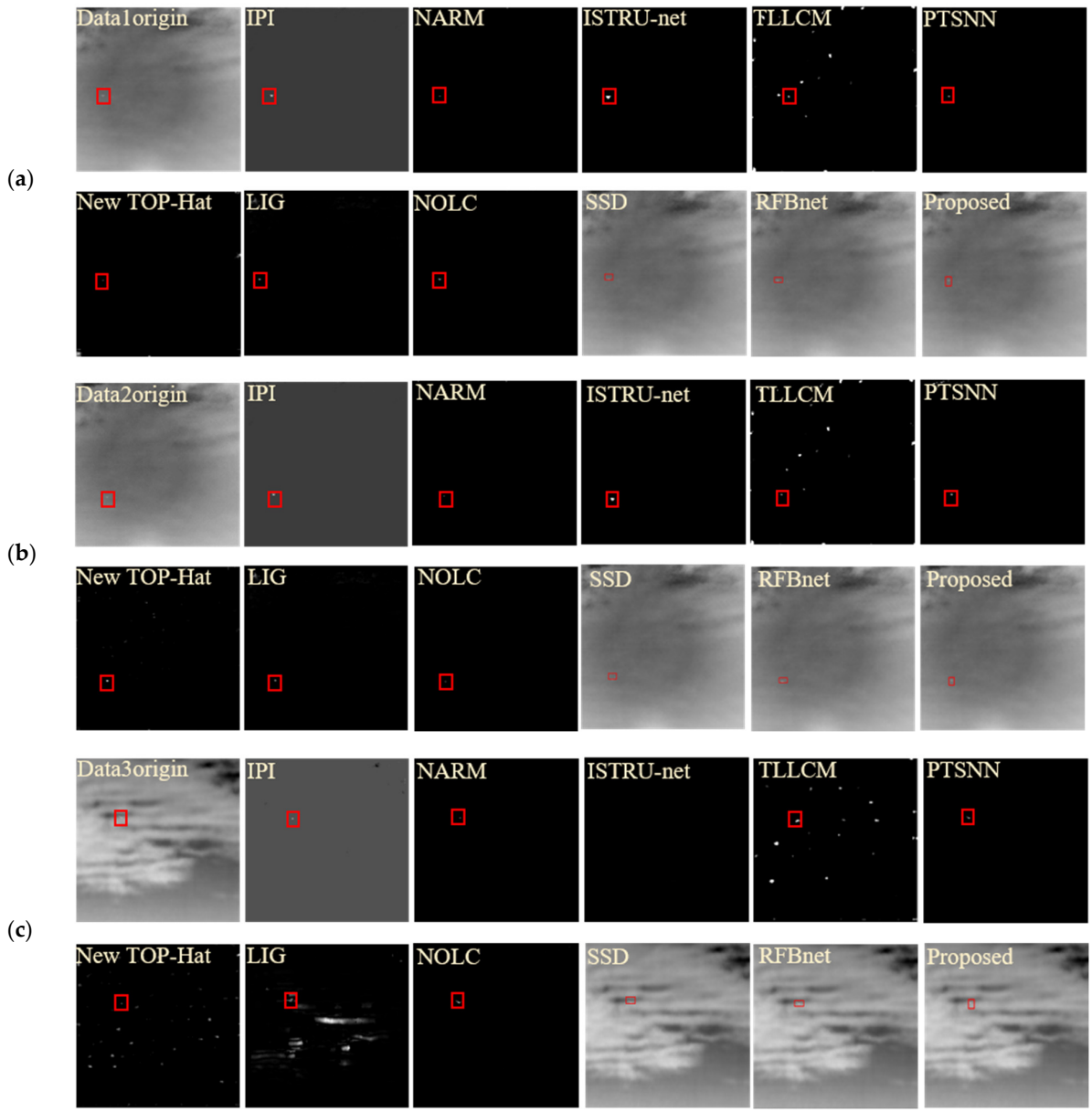


Figure 7. *Cont.*

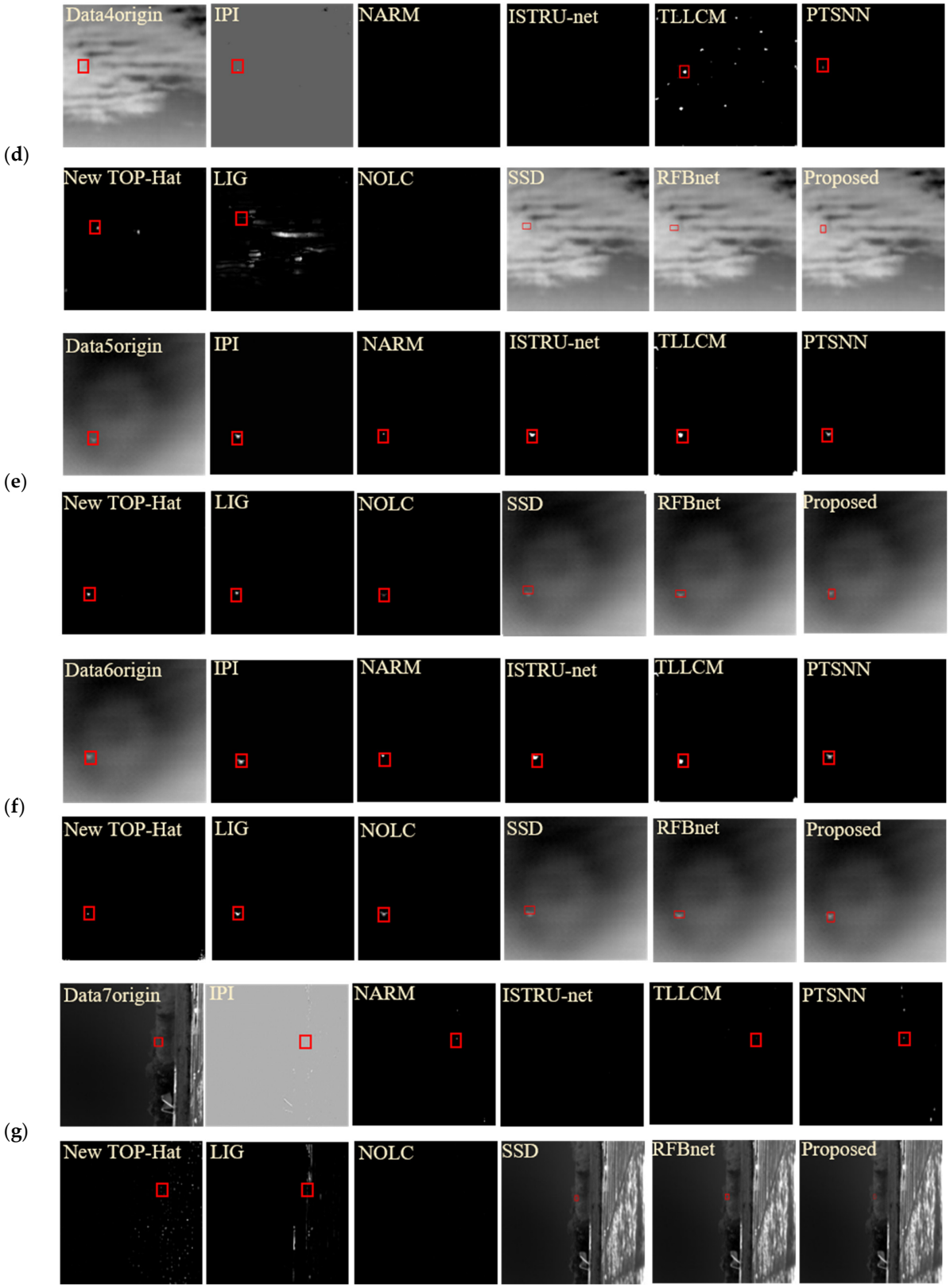


Figure 7. The detection results of each algorithm for typical images in data1to data7 by 11 methods: (**a**) the detection results of data1; (**b**) the detection results of data2; (**c**) the detection results of data3; (**d**) the detection results of data4; (**e**) the detection results of data5; (**f**) the detection results of data6; (**g**) the detection results of data7. Specifically, the red rectangles indicate the targets in the images.

To further demonstrate the detection performance of our method, the ROC curves for the real data are shown in Figure 8. Under the same false alarm rate, a higher detection rate indicates better performance and robustness of the detection method. Our method achieved a higher P_d and lower F_a than other methods when applied to real data. Furthermore, the ROC curve of our proposed method is closer to the upper left, indicative of its superior performance. Additionally, our method maintains a satisfactory detection rate even under conditions with an extremely low false alarm rate, demonstrating that it is well suited to meet the stringent requirements of dim and small target detection in infrared imaging.

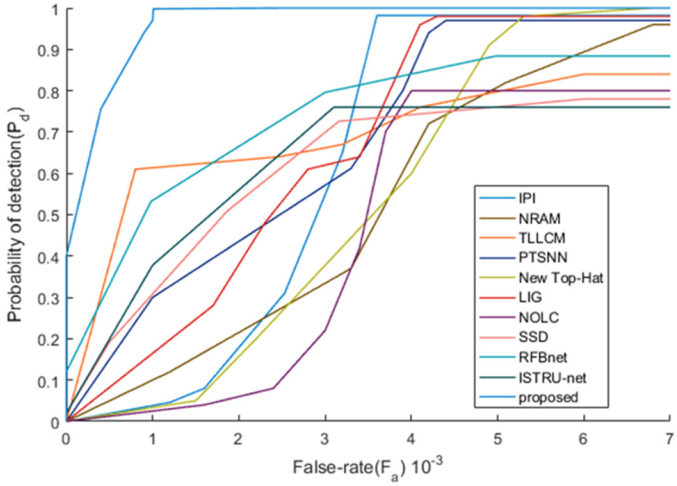


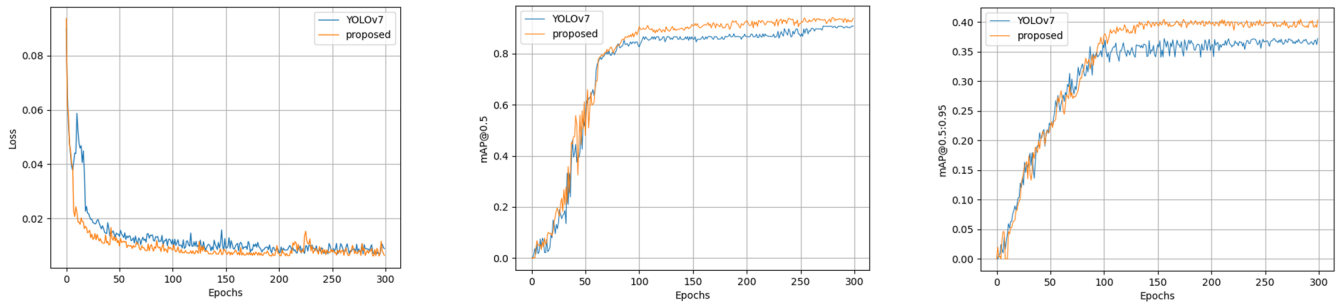
Figure 8. ROC curve of data7.

3.2.3. Analysis

In Figure 7, it becomes evident that the algorithm we proposed can effectively identify targets within infrared dim and small target datasets, which exhibit varying local signal-to-noise ratios across diverse scenes. However, other algorithms, except for RFBnet, encounter several issues. Although TLLCM detects the target, it produces many false alarms and cannot effectively filter out background noise. Similarly, LIG, IPI, PTSNN, and new top-hat are not immune to sporadic false alarms and struggle to suppress local high-light clutter. Both NARM and NOLC also remove valid signals around the target during the detection process, resulting in missed targets. There are many continuous and uniform cloud layer results in the results of the LIG, and its ability to suppress long strip noise is limited. The new top-hat shows commendable detection performance in simple backgrounds, but its performance in cluttered backgrounds requires improvement. Although the SSD and ISTRU-net algorithms detect most of the targets in the dataset, they fail to detect the target in the data3 sequence, indicating a deficiency in handling interference from high-intensity cloud backgrounds. It is important to note that when the difference between the target and the background is too small, specifically when the LSNR is less than -4.4, the algorithm's detection performance drops sharply, with the mAP_{50} falling below 0.38, or the target becoming completely undetectable.

To further analyze the improvement of the algorithm we proposed compared to YOLOv7, we visualize the curves of losses, mAP_{50}, and $mAP_{50:95}$, as shown in Figure 9.

It is clear that, compared to the YOLOv7 model, our network model shows faster loss reduction during the initial training stages. The decline slows down around 50 epochs before eventually stabilizing and converging. Throughout the training process, our model minimizes jitter and enhances stability, demonstrating superior performance in detecting dim and small infrared targets.



(**a**) (**b**) (**c**)

Figure 9. The changes in the YOLOv7 model and our model with the training epochs: (**a**) box loss; (**b**) mAP_{50}; (**c**) $mAP_{50:95}$.

4. Conclusions

In long-distance infrared detection, targets are characterized by their small size and sparse feature information within complex scenes, posing significant challenges for the accurate and efficient detection of dim and small targets. To address challenges related to dim and small target feature information, low signal-to-noise ratio, and detection difficulties in infrared images, we propose a dim and small target detection algorithm that mimics vision focus via dual-input information and fuses locally extracted feature information with global features in a multi-scale manner. Our approach utilizes local components to suppress complex background input and amplify target features. Simultaneously, the fusion design incorporates global and local input features at multiple scales, enabling the extraction of both spatial and semantic information during dim and small target detection. Furthermore, incorporating a 1×1 convolution module and using the concat operation effectively reduces the computational load of the model during the feature extraction and fusion process. Experimental results show that the mAP_{50} and $mAP_{50:95}$ of the proposed algorithm improve by more than 2% and 3.6%, respectively, on datasets with different scenes and noise compared to the baseline algorithm, achieving an average accuracy of 75% even in complex scenes. In the future, we will continue refining the vision focus mechanism and further enhance the model's generalization ability. Our goal is to achieve accurate detection of dim and small targets in various infrared scenes while better adapting to the analysis and processing requirements of infrared environments.

Author Contributions: Conceptualization, X.L. and Z.Y.; methodology, Q.S. and J.L.; writing—original draft preparation, C.Z.; writing—review and editing, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based on research/work supported by the National Science and Technology Foundation of China under Grant 2023-JCJQ-JJ-0976.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments, which helped improve this paper.

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

References

1. Fang, H.; Xia, M.; Zhou, G.; Chang, Y.; Yan, L. Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 7002305. [CrossRef]
2. Deng, H.; Sun, X.; Liu, M.; Ye, C.; Zhou, X. Small infrared target detection based on weighted local difference measure. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4204–4214. [CrossRef]

3. Sun, Y.; Yang, J.; An, W. Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3737–3752. [CrossRef]
4. Liu, S.; Chen, P.; Woźniak, M. Image Enhancement-Based Detection with Small Infrared Targets. *Remote Sens.* **2022**, *14*, 3232. [CrossRef]
5. Du, J.; Lu, H.; Zhang, L.; Hu, M.; Chen, S.; Deng, Y.; Shen, X.; Zhang, Y. A spatial-temporal feature-based detection framework for infrared dim small target. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 3000412. [CrossRef]
6. Boccignone, G.; Chianese, A.; Picariello, A. Small target detection using wavelets. In Proceedings of the Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170), Brisbane, Australia, 16–20 August 1998; pp. 1776–1778.
7. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2010**, *43*, 2145–2156. [CrossRef]
8. Zhao, Y.; Pan, H.; Du, C.; Peng, Y.; Zheng, Y. Bilateral two-dimensional least mean square filter for infrared small target detection. *Infrared Phys. Technol.* **2014**, *65*, 17–23. [CrossRef]
9. Lei, B.; Hao, W.; Yan, K.; Li, J. Signal denoising of multi element infrared signal based on wavelet transform. In Proceedings of the 2020 International Conference on Electrical Technology and Automatic Control ICETAC, Anhui, China, 7–9 August 2020; p. 012102.
10. Li, Q.; Nie, J.; Qu, S. A small target detection algorithm in infrared image by combining multi-response fusion and local contrast enhancement. *Optik* **2021**, *241*, 166919. [CrossRef]
11. Deng, L.; Zhu, H.; Zhou, Q.; Li, Y. Adaptive top-hat filter based on quantum genetic algorithm for infrared small target detection. *Multimed. Tools Appl.* **2018**, *77*, 10539–10551. [CrossRef]
12. Chen, Y. On suboptimal detection of 3-dimensional moving targets. *IEEE Trans. Aerosp. Electron. Syst.* **1989**, *25*, 343–350. [CrossRef]
13. Pang, D.; Shan, T.; Ma, P.; Li, W.; Liu, S.; Tao, R. A novel spatiotemporal saliency method for low-altitude slow small infrared target detection. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 7000705. [CrossRef]
14. Deng, L.; Zhu, H.; Tao, C.; Wei, Y. Infrared moving point target detection based on spatial-temporal local contrast filter. *Infrared Phys. Technol.* **2016**, *76*, 168–173. [CrossRef]
15. Yang, C.; Duraiswami, R.; Davis, L. Fast multiple object tracking via a hierarchical particle filter. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; Volume 1, pp. 212–219.
16. Zhang, T.; Li, M.; Zuo, Z.; Yang, W.; Sun, X. Moving dim point target detection with three-dimensional wide-to-exact search directional filtering. *Pattern Recognit. Lett.* **2007**, *28*, 246–253. [CrossRef]
17. Fu, J.; Wei, H.; Zhang, H.; Gao, X. Three-dimensional pipeline Hough transform for small target detection. *Opt. Eng.* **2021**, *60*, 023102. [CrossRef]
18. Mo, Y.; Wang, L.; Hong, W.; Chu, C.; Li, P.; Xia, H. Small-Scale Foreign Object Debris Detection Using Deep Learning and Dual Light Modes. *Appl. Sci.* **2024**, *14*, 2162. [CrossRef]
19. Lu, D.; Tan, J.; Wang, M.; Teng, L.; Wang, L.; Gu, G. Infrared Ship Target Detection Based on Dual Channel Segmentation Combined with Multiple Features. *Appl. Sci.* **2023**, *13*, 12247. [CrossRef]
20. Braham, M.; Van Droogenbroeck, M. Deep background subtraction with scene-specific convolutional neural networks. In Proceedings of the 2016 International Conference on Systems, Signals and Image Processing (IWSSIP), Bratislava, Slovakia, 23–25 May 2016; pp. 1–4.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the NIPS'15: 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28, pp. 91–99.
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
23. Feng, Z.; Xie, Z.; Bao, Z.; Chen, K. Real-time dense small object detection algorithm for UAV based on improved YOLOv5. *Acta Aeronaut. Astronaut. Sin.* **2023**, *44*, 327106.
24. Fan, Z.; Bi, D.; Xiong, L.; Ma, S.; He, L.; Ding, W. Dim infrared image enhancement based on convolutional neural network. *Neurocomputing* **2018**, *272*, 396–404.
25. Wang, Y.; Zhao, L.; Ma, Y.; Shi, Y.; Tian, J. Multiscale YOLOv5-AFAM-Based Infrared Dim-Small-Target Detection. *Appl. Sci.* **2023**, *13*, 7779. [CrossRef]
26. Hare, S.; Golodetz, S.; Saffari, A.; Vineet, V.; Cheng, M.-M.; Hicks, S.L.; Torr, P.H. Struck: Structured output tracking with kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2096–2109.
27. Lee, J.-Y. A Study of CR-DuNN based on the LSTM and Du-CNN to Predict Infrared Target Feature and Classify Targets from the Clutters. *Trans. Korean Inst. Electr. Eng.* **2019**, *68*, 153–158.
28. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-layer attention network for small object detection in remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 2148–2161. [CrossRef]
29. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

30. 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.
31. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [CrossRef] [PubMed]
32. Wu, L.; Fang, S.; Ma, Y.; Fan, F.; Huang, J. Infrared small target detection based on gray intensity descent and local gradient watershed. *Infrared Phys. Technol.* **2022**, *123*, 104171. [CrossRef]
33. Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; Peng, Z. Infrared small target detection via non-convex rank approximation minimization joint $l_{2,1}$ norm. *Remote Sens.* **2018**, *10*, 1821. [CrossRef]
34. 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]
35. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
36. 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.* **2019**, *17*, 1822–1826. [CrossRef]
37. Zhang, H.; Zhang, L.; Yuan, D.; Chen, H. Infrared small target detection based on local intensity and gradient properties. *Infrared Phys. Technol.* **2018**, *89*, 88–96. [CrossRef]
38. Wang, Z.; Cheng, Z.; Huang, H.; Zhao, J. ShuDA-RFBNet for Real-time multi-task traffic scene perception. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 305–310.
39. Hou, Q.; Zhang, L.; Tan, F.; Xi, Y.; Zheng, H.; Li, N. ISTDU-Net: Infrared Small-Target Detection U-Net. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7506205. [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.