Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 21(1): 1712–1737.
DOI: 10.3934/mbe.2024074
Received: 25 September 2023
Revised: 05 December 2023
Accepted: 21 December 2023
Published: 02 January 2024

Research article

IMD-Net: Interpretable multi-scale detection network for infrared dim and small objects

Dawei Li[1,]*, Suzhen Lin[2] and Xiaofei Lu[3], Xingwang Zhang[1], Chenhui Cui[2] and Boran Yang[2]

[1] College of Electricity and Control Engineering, North University of China, Taiyuan 030051, China
[2] College of Data Science and Technology, North University of China, Taiyuan 030051, China
[3] Jiuquan Satellite Launch Center, Dongfeng Frame, Jiuquan 735000, China

* **Correspondence:** Email: lidawei@nuc.edu.cn; Tel: +8613834522163.

Abstract: This study proposed an interpretable multi-scale infrared small object detection network (IMD-Net) design method to improve the precision of infrared small object detection and contour segmentation in complex backgrounds. To this end, a multi-scale object enhancement module was constructed, which converted artificially designed features into network structures. The network structure was used to enhance actual objects and extract shallow detail and deep semantic features of images. Next, a global object response, channel attention, and multilayer feature fusion modules were introduced, combining context and channel information and aggregated information, selected data, and decoded objects. Finally, the multiple loss constraint module was constructed, which effectively constrained the network output using multiple losses and solved the problems of high false alarms and high missed detections. Experimental results showed that the proposed network model outperformed local energy factor (LEF), self-regularized weighted sparse model (SRWS), asymmetric contextual modulation (ACM), and other state of the art methods in the intersection-over-union (IoU) and $F_{measure}$ values by 10.8% and 11.3%, respectively. The proposed method performed best on the currently available datasets, achieving accurate detection and effective segmentation of dim and small objects in various infrared complex background images.

Keywords: multi-scale object enhancement module; global object response module; multilayer feature fusion module; multiple loss constraint module; dim and small object detection

1. Introduction

The infrared search and track system (IRST) is widely used in many fields, such as aerospace precision guidance, military early warning, and sea rescue [1,2]. Dim-small targets detection and recognition is one of the bottleneck problems in the intelligent process of various early warning systems, precision guidance systems, security systems, and unmanned aerial vehicle (UAV) inspection systems [3–5]. Infrared imaging has the advantages of long imaging distance and strong anti-interference ability over visible light imaging. However, due to the detection distance and imaging of grayscale images, the actual object in the source image is displayed as a Gaussian distribution of gray spots. The manifestation of dim-small targets includes tiny size, variable (object size 2 × 2 ~ 9 × 9), and low signal-to-noise ratio (less than 5.0). Such targets are very common in deep space and sea surface exploration. Additionally, cloud edges in the complex background, the corners of the natural scenery and artificial buildings, ocean clutter and deep-space noise pose a great challenge to detecting infrared small and weak objects.

At the present stage of multi-mode/multi-band imaging used to detect dim targets, the main purpose is to make comprehensive use of the advantages of different mode imaging to achieve a more accurate and comprehensive interpretation of the scene and target. Typical applications in Europe and the United States integrate the near-ultraviolet to far-infrared multi-band image information for deep space detection and security monitoring. China's satellites use visible light, infrared short wave, infrared medium wave and infrared long wave four bands to implement all-weather monitoring of the same target simultaneously. With the intensification of global competition and the expansion of applications, the intelligence of high-precision detection systems has now become one of the new commanding heights of high-tech competition worldwide. Intelligent dim-small targets detection and recognition means that the algorithm model can quickly extract dim targets from one or more images and automatically classify the extracted targets according to the changes of detected image features, just like the human brain.

Infrared dim-mall object detection methods are mainly classified into model-driven mathematical modeling methods [6] and data-driven deep learning (DL) [7] methods. Among them, mathematical modeling methods design handcrafted features for infrared weak objects' physical and mathematical characteristics and construct mathematical models using a priori knowledge to extract objects and suppress background. This class of methods is further subdivided into three categories: 1) The background subtraction method [8,9], which constructs a mathematical model to predict the background and obtains the actual object by the difference between the source image and the predicted background; 2) The local contrast method based on the human vision system (HVS) [10,11], which extracts the actual object by using the grayscale difference between the object and the local neighborhood and 3) the paradigm-constrained optimization method [12], which exploits the sparse features of the actual object and the low-rank properties of the background matrix [13] to perform paradigm constraints and optimal solutions for the object and background. Model-driven mathematical modeling-based methods are fast in detection, good in specific types of backgrounds, and require no training, and the computational process and output results are controllable. However, this type of method has the following main drawbacks. First, the method is less robust, due to the excessive reliance on hand-designed features, the detection accuracy cannot be guaranteed for different types of complex backgrounds, and it isn't easy to apply to practical engineering. Second, this method is less descriptive of the contour and can only obtain the object's center of mass and the coordinates of the surrounding pixels, providing less a priori knowledge for the next recognition or tracking operations. In modern local wars and conflicts, real-time tracking of UAV is a challenge for the downstream task of target detection. This type of target is not easy to be captured, detected and tracked at high speed, which is a hot spot in the field of dim target detection. Han et al. proposed a generic framework for a correlation filter (CF) based tracker, which jointly considers the discrimination and reliability information in the filter learning stage. Context patches are employed into the filter training stage to better distinguish the target from backgrounds [14].

Data-driven DL methods excavate the grayscale distribution of actual infrared objects, various types of complex backgrounds, and clutter noise from a large amount of data and then use the powerful feature extraction ability and nonlinear data fitting ability of neural networks to extract features and pixel classification of infrared source images to obtain actual objects. DL-based detection methods are mainly classified into two categories. The first is the object regression network, which adopts the regression of the object's minimum outer rectangular box, such as Faster-RCNN (faster-region convolutional neural networks), YOLO (you only look once), SSD (single shot multibox detector) [15,16] and other series of networks. These networks achieve weak object detection by fine-tuning the detection structure. Still, because the objects are too small and need more detailed features such as texture and color, it is very easy to lose actual objects in the stage of extracting image features, so this type of method leads to a high leakage rate. The second category is the object segmentation network, which achieves object detection by classifying each pixel in the source image. The object segmentation network structure not only locates the position of the actual object accurately but also effectively describes the contour features of the infrared object at different scales.

A series of infrared object segmentation networks have been proposed in recent years. For example, the context-based network ACM model by Dai et al. [17] used an asymmetric contextual module to aggregate shallow and deep features, then introduced an expanded local contrast to achieve a trainable local contrast metric based on the introduction of expanded local contrast to achieve a trainable local contrast metric. In the follow-up study of the same research team, the ALC-Net (attentional local contrast networks) was constructed [18]. Alternatively, Wang et al. decomposed the infrared object detection problem into two relative subproblems. First, they used the generative adversarial network MDvsFA-CGAN (miss detection vs. false alarm: Conditional GAN) [19] to compromise between missed detection and false alarms in infrared small object detection, then they adopted the dense nested interaction structure of the DNA-Net (dense nested attention network) model [20], combining different information in the deep and shallow layers of the neural network for redundancy to ensure that the object information could be maintained at high intensity to the decoder side for object decoding to achieve weak small object detection. Alternatively, the LSPM (local similarity pyramid module) segmentation network proposed in [21] simulated the multi-scale features of infrared weak objects by designing a local similarity pyramid. AGPC (attention-guided pyramid context) [22] segmentation network empowered the modeling capability of infrared weak small targets with multi-scale features by incorporating a well-designed attention-guided pyramid context module. It aggregated shallow and deep features using an asymmetric feature fusion module. Furthermore, a segmentation network called FC3Net (feature compensation and cross-level correlation) was proposed to segment weak infrared small targets using feature compensation structures and attention mechanisms [23]. Multiple well-designed feature compensation modules facilitated the transmission of detailed information on infrared weak small targets to deeper layers of the network. The above network models could segment infrared objects in different complex backgrounds. Still, there is the problem of completely discarding the a priori knowledge of infrared weak objects in the source image, failing to enhance the weaker objects effectively. The computational process of DL is not interpretable, making the model training and parameter tuning extremely difficult [24]. The interpretability of network models refers to the ability to explain to users via understandable logic rules, i.e., the ability to use symbols or words to describe the model structure rationally and ensure that the theoretical design is consistent with the actual output. Although the robustness and generalization of network models based on DL are better than those based on mathematical modeling methods, DL results in a less reliable detection system due to poor interpretability and black box characteristics.

Given the above issues, this study proposes an interpretable multi-scale infrared weak object detection network (IMD-Net) to address the problems of object regression and segmentation networks, improving the network model's interpretability while enhancing the object using hand-designed features. The detection network first transforms the hand-designed features into a network structure to enhance the actual object and extracts the shallow and deep features of the source image at the same time. It calculates the pixel global correlation to capture the object's long-range dependence and obtains the object response; then it fuses the different levels of features with redundancy to decode the object. Finally, it unites multiple loss functions to constrain the network output effectively and obtains the network output. While using the neural network's powerful feature extraction and data fitting ability, the hand-designed features are combined to enhance the weak objects and achieve the robust segmentation of multi-scale infrared weak objects under various complex backgrounds. In this paper, "network" refers to the structure of the network model and 'method' refers to the way to solve the problem, which has different meanings.

2. Materials and Methods

2.1. Related works

Object detection using convolutional networks for infrared source images faces the following problems: 1) The actual object has a low and weak gray level, and the process of extracting image features by cascading convolutional blocks is prone to result in the loss of the object, preventing its maintenance at the decoder side; and 2) the convolutional network suffers from a constrained sensory field, preventing its aggregation of the global information for classifying the object and background pixels. This section introduces the proposed network to solve the above problems and improve the detection segmentation accuracy and interpretability of the detection model.

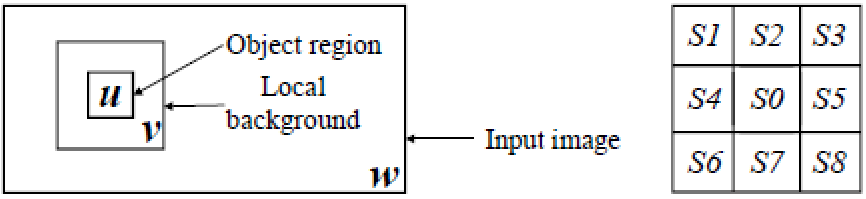
2.1.1. Infrared weak object enhancement

In the infrared source image, the infrared weak object has a small size and low signal-to-noise ratio. However, it differs from the neighborhood background, where (i) the infrared weak object gray value is larger than the neighborhood background gray value, implying that it is brighter than the local neighborhood background, and (ii) the infrared weak object gray level is a Gaussian distribution, which is the same as the neighborhood background gray difference in eight directions. Therefore, a mathematical model can be constructed based on these physical properties to enhance the infrared ray (IR) weak object and suppress the background clutter.

Given an infrared source image as w, a sliding window v is used to compute the local grayscale contrast information. In Figure 1, u is the central object region, while v is presented in Figure 1(b) and is equally divided into nine cells, and u corresponds to S0 in Figure 1(b). The sliding window slides from left to right and top to bottom on the source image. The mean value of the pixels in the sliding window is calculated as follows:

$$m_{i,k} = \frac{1}{N_i}\sum_{j=1}^{N_i} I_j^{i,k}, \tag{1}$$

where N_i is the number of pixels in the ith cell; $I_j^{i,k}$ is the gray level of the jth pixel in the ith cell (cell size$k \times k(k = 3,5,7,9)$); $m_{i,k}(i = 1,2, \dots ,8)$ is the mean gray value of the ith cell (S1, S2,..., S8) when the cell size is $k \times k$.



(a) Grayscale contrast calculation (b) Sliding window

Figure 1. Calculating local gray contrast.

The contrast between the center cell and the neighboring cells is defined as

$$c_i^{n,k} = \frac{L_{n,k}}{m_{i,k}} , \tag{2}$$

where $L_{n,k}$ is the average gray value of the nth sliding window center cell S0 in window v, and $c_i^{n,k}$ is the contrast between the nth sliding window center cell and the ith neighborhood cell. From Eq (1), the object pixel is $c_i^{n,k} \geq 1$ and the background pixel is $c_i^{n,k} \leq 1$. From Eq (2), the object pixel is $\min_i(c_i^{n,k}) \geq 1(i = 1,2, \dots ,8)$ and strong edges and clutter are $\min_i(c_i^{n,k}) \leq 1(i = 1,2, \dots ,8)$.

Therefore, object enhancement and background suppression can be performed using the grayscale contrast of the object neighborhood at different scales to ensure that the actual object can be maintained up to the deeper layers of the network and the decoder side. This study implemented this mathematical theory into a network structure to introduce a multi-scale object enhancement module. The object and neighborhood grayscale contrast is defined as follows:

$$C_n = \max_k(\min_i(C_{i,k}^n)) , \tag{3}$$

2.1.2. Global context module

Inspired by nonlocal mean filtering, the global context (GC) block structure [25] shown in Figure 2 breaks the fixed sense field limitation of the convolution module by calculating the correlation between pixels in the feature map, capturing the long-distance dependency between pixels, and responding to the actual object after aggregating the contextual information. This paper refers to the GC block as the global object corresponding module. As seen in Figure 2, The GC block structure mainly comprises three parts: 1) A global attention structure for context modeling, 2) a bottleneck transform structure for capturing channel dependencies and 3) a fusion structure for feature fusion after broadcasting operations on pixel values. The specific operation can be expressed as follows:

$$z_i = x_i + W_{v2} R\,e\,l\,u\left(LN\left(W_{v1}\sum_{j=1}^{N_p}\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} x_j\right)\right), \tag{4}$$

where x_i, $\mathrm{x_j}$ and x_m are the ith, jth and mth$\left(1 \le i, j, m \le N_p\right)$ pixels of the input feature map, respectively; z_i is the output pixel of the output feature map at the position corresponding to x_i; N_p is the total number of pixels of the feature map; $\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}}$ is the attention weight after global pooling; $W_{v2} R\,e\,l\,u\left(LN\left(W_{v1}(g)\right)\right)$ is the bottleneck transform structure; W_k, W_{v1}, and W_{v2} are the nonlinearly varying convolutional kernel parameters; $LN(g), Relu(g)$ is the normalization operation and activation function. A normalization layer is added to the bottleneck transform structure to simplify the optimization and act as a regularizer to enhance the network's generalization ability.

In summary, Section 1.1 can use the differences between actual objects and their neighborhood background to enhance the object and prevent object loss in the feature extraction stage. This paper converts this mathematical model into a network structure for model construction and enhances the object in a multi-scale range. The GC block structure can compute the interpixel correlation to capture the long-distance dependency. This paper introduces this network structure to enhance the image with contextual information aggregation to make the actual object responsive and solve the problem of a limited sensory field.

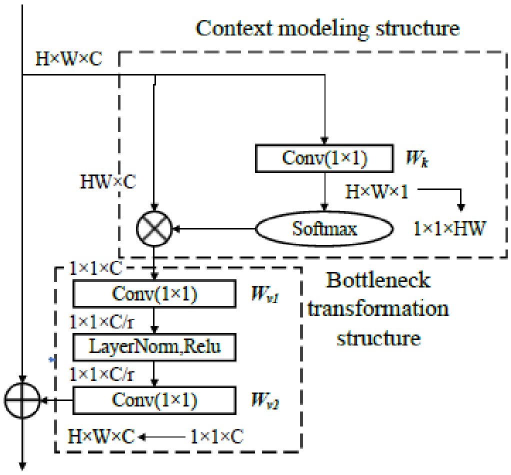


Figure 2. GC block.

2.1.3 Datasets for single frame infrared small target (SIRST) detection

Currently, there is a significant scarcity of publicly available datasets for detecting infrared dim and small targets on the internet. Some researchers have published datasets they used in their papers [17,19,20], such as Wang et al. [19], who constructed a detection dataset with 10,000 frames of images. They created this dataset by cropping high-resolution images to form backgrounds, then superimposed real targets or synthetic objects onto these backgrounds. However, the synthetic nature of the dataset results in noticeable artifacts, and the annotations are only partially accurate. Dai et al. [17] created a dataset with real captured detection images comprising 427 frames. While this dataset contained true data, it was relatively small and might not meet the training requirements of neural networks. Li et al. [20] constructed a synthetic-detection dataset with 1327 frames. Compared to the synthetic dataset in [19], it featured more realistic and subdued targets, smoother boundaries, and more lifelike generated images. However, this dataset had fewer frames, particularly for multiple target images, and might not adequately satisfy the training and generalization needs of detection models.

This paper employs an approach that combines augmentation and semi-simulation to create a dataset of infrared dim and small targets with varying quantities, sizes, backgrounds, and precise annotations, while requiring relatively less manpower and resources. A detailed description of the dataset is presented in Section 3.1.

2.2. The proposed method

In this section, the overall structure of the network is first introduced, then the multi-scale object enhancement (MTE) module, global object response (GTR) module, channel attention (Ch_atte) module, multilayer feature fusion (MFF) module, and multiple loss constraint (MLC) module are introduced.

2.2.1. Overall network architecture

Given an infrared source image, the input image is classified pixel-by-pixel by a fully convolutional network to segment the actual infrared weak object, and the final output is a detection result image of the same size as the input image, which is thresholded to obtain the actual object. The network structure proposed in this paper is shown in Figure 3.

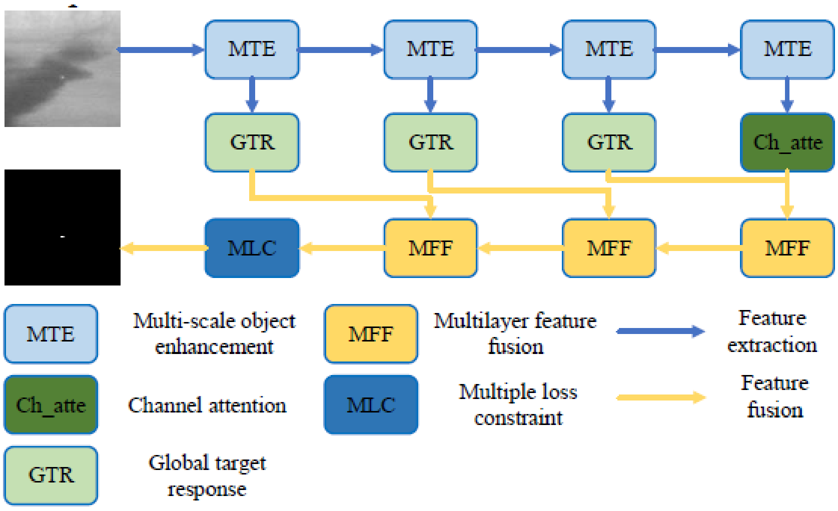


Figure 3. Overall network structure.

First, the infrared source image is fed into the MTE module for feature extraction and object enhancement; The GTR module calculates the global response of the object features in the feature map and establishes the long distance dependence between the object pixels, the Ch_atte module assigns different weights to the deep semantic features of the network to pay attention to the useful information while ignoring the useless information, the MFF module fuses the feature information of different levels. It decodes the object information, and the MFF module fuses the feature information of different levels and decodes the object information. The MLC module combines the multiclass loss to effectively constrain the output results and obtain more accurate object location and pixel classification.

2.2.2. Multi-scale object enhancement module

As mentioned in Section 1, the reasonable use of prior knowledge can effectively enhance weak infrared objects and improve object detection accuracy and network module interpretability. However, because the network model in this paper is an end-to-end fully convolutional network, the constructed mathematical model needs to be transformed into a neural network structure to meet the model's end-to-end learning requirements. The MTE module constructed in this paper is depicted in Figure 4.

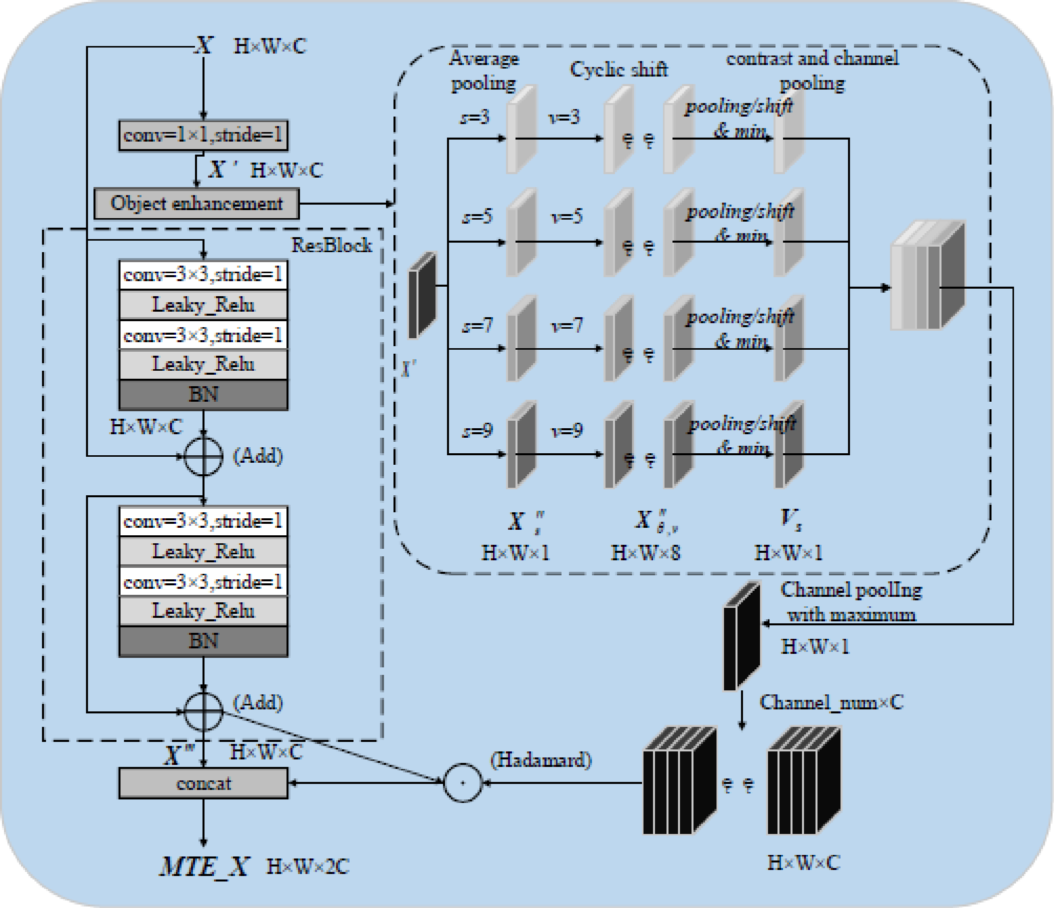


Figure 4. Multi-scale object enhancement module.

A. Object enhancement

Let the input feature map be $X \in R^{H\times W\times C}$. To reduce the amount of parameter computation, use the 1×1 convolution operation to reduce the number of feature map channels to one to obtain the reduced dimensionality of feature map X'. First, to compute the grayscale contrast between the object region and the neighborhood background region, carry out the mean pooling operation on X' to obtain the pooled feature map X''_s, with a pooling diameter of $s(s = 3,5,7,9)$. The step length of one does not change the size of the feature map, and the pooling operation corresponds to Eq (1) to find the sliding mean value of each cell in the window. Next, the cyclic shift operation is performed on X''_s, and the operation process is shown in Figure 6, with shift direction $\theta(\theta = 0, \frac{\pi}{4}, \frac{\pi}{2}, \ldots, \frac{7\pi}{4})$ and shift distance $v(v = 3,5,7,9)$, to obtain the shifted feature map $X''_{\theta,v}$. Different pooling diameters and shift distances are used to compute the contrast between the actual object region and the local neighborhood background at different scales. The contrast feature map group obtained by the same shift distance is subjected to a channel minimum pooling operation to obtain the weight feature map corresponding to this shift distance P, as shown in Eq (5).

$$V_s = \min_{\theta}\left(\frac{X''_s}{X''_{\theta,v}+\alpha}\right); \quad s = v, \theta = 0, \frac{\pi}{4}, \ldots, \frac{3\pi}{2}, \frac{7\pi}{4} \ , \tag{5}$$

where the ratio of pooled-to-shifted feature maps corresponds to comparing the gray mean of the object region with its neighbors in different directions via Eq (2); α is a very small positive value used to avoid zero divisors.

B. Feature extractions

The input feature map X is subjected to ResBlock for feature extraction, in which the convolution kernel is selected to be 3 × 3 in size, with a step size of one. The padding method is selected as “same pixel padding at the edges” so that the output feature map maintains the same size as the input feature map. The activation function is Leaky_rectified linear unit (Leaky_ReLU), which transforms the linear mapping into a nonlinear transformation to fit a more realistic data distribution, and the batch normalization (BN) layer represents the data normalized layer, which improves the training speed of the network to learn the data distribution and enhances the generalization ability of the network. The short join operation avoids the vanishing gradient problem caused by too deep a network. In the infrared source image, the size of the weak object is 2 × 2–9 × 9. Given the feature extraction process, downsampling easily leads to the loss of the weak object. This paper’s fully convolutional network without downsampling uses only the convolution operation to extract fixed sensory field features. The sensory field is limited to the problem that will be solved in Section 3.3. The feature extraction process can be expressed as follows:

$$X''' = Fe(Fe(X)) \tag{6}$$

$$Fe(X) = X + BN(\sigma(W_2\sigma(W_1X))) \ , \tag{7}$$

where W_1, W_2, σ, and BN are the first and second layer convolution kernel parameters, i.e., the activation function Leaky_ReLU and the data normalized operation, respectively; $Fe()$ is the feature extraction operation; and X''' is the intermediate features extracted after the ResBlock.

Perform channel maximum pooling and channel expansion operations on the weight feature map $V_S = \{V_3, V5, V_7, V_9\} \in R^{H\times W\times 4}$ with the expansion multiplier C. The point multiplication operation is with X''', then perform the $concat$ operation with X''' to obtain the output of MTE_X

from the MTE module. The MTE object enhancement module performs object enhancement and background suppression according to the grayscale comparison between the actual object and the neighborhood to speed up the training speed of the model and improve the *MTE_X* network's detection performance. The following equation describes the calculation process:

$$MTE\_X = concat(X''', (\underset{s}{max}(V_s))^C \odot X'''), \tag{8}$$

where the superscript C is the channel expansion by a factor of C; $\odot$ is the matrix multiplication operation pixel by pixel; $concat(g)$ is the channel merge operation; *MTE_X* is the output of the MTE module.

2.2.3. Global object response module

In the network proposed in this paper, no downsampling operation is performed in the encoding stage to prevent object loss, so each pixel in the feature map *MTE_X* can only aggregate features with limited sensory fields. However for intensive detection tasks such as weak object detection in infrared images, relying on local information alone cannot accurately classify the object and the background pixels, so it is necessary to aggregate global contextual features to each pixel to provide a long-range dependency and then judge the pixel category based on the global information. Therefore, this paper introduces the GC block model to complete the aggregation of context information. The input feature map of this module is *MTE_X* and the output is the global object response feature map *GTR_X*.

2.3. Ch_atte module

Multiple cascaded MTE modules convert shallow detail features of infrared source images into deep semantic features, and it is easy to classify the actual object in multichannel high-dimensional data with a large difference between the actual object and clutter background features. However, the output of the convolution layer in the feature extraction process of the MTE module does not consider the dependence on each channel. The features of each channel are independent of each other or even mutually exclusive in the multichannel deep semantic features, so it is necessary to let the network selectively enhance the informative features and suppress the useless features to facilitate the subsequent fusion of the useful features for the decoding of the object. Therefore, this paper introduces the channel attention mechanism [26,27] as the Ch_atte module to selectively extract the data of each channel, where the input feature map is MTE_X and the output feature map is ChAtte_X. Notably, this study uses L2 regularization to enhance the sparsity of the semantic feature maps after the fully connected layer of the channel attention mechanism, which not only speeds up the training speed of the model but also makes the network pay more attention to useful features.

2.4. Multilayer feature fusion module

In the object decoding stage, this study adopts the multilayer feature fusion module to effectively fuse the deep semantic features and shallow detail features, to achieve progressive interaction between high-level features and low-level features, to make good use of the contextual information of the small objects through repeated fusion and enhancement, to ensure that the weak objects always remain in the feature layer and to achieve the purpose of accurate classification of the object and background pixels.

We adopt the residual-in-residual dense block network (RDBNet) [28] as a multilayer feature fusion module, whose infrastructure consists of dense nested networks to achieve redundancy in combining multilayer information. The input feature map of this module comprises *ChAtte_X* and *GTR_X*, representing deep semantic features (local features) and shallow detail features (global object response features), respectively. These two feature layers are augmented by the channel attention and spatial attention structures, respectively, so both ensure that the actual object is not lost, the weights of their input feature maps are 0.5 and 0.5, respectively, and *MFF_X* is the module's output feature map.

2.5. Multi-loss constraint module

The following problems exist in the loss calculation in the infrared weak object detection network: *Problem 1*. The number of infrared weak object pixels accounts for a small percentage of the infrared source image, and there is a serious imbalance between positive and negative ratios in the loss calculation. The loss value is mainly composed of the loss of negative samples. Hence, the network mainly considers the correctness of the classification of the background pixels in the training, which leads to a low accuracy rate of the object detection results.

Problem 2. Since the grayscale of infrared weak objects conforms to a Gaussian distribution, the grayscale of the actual object edge pixels is close to that of the background pixels in the neighborhood, and the grayscale of the background strong edges, corners, and strong noise is close to the grayscale of the actual object center. The detection network can easily classify these types of pixels incorrectly, which leads to poor network profile description performance and a high false alarm rate.

To address the above two problems, this paper constructs the MLC module to combine multiple loss functions to effectively constrain the network output, as shown in Figure 5, where the loss functions are selected as $SoftIoU\_Loss$ and $Focal\_Loss$ [29,30].

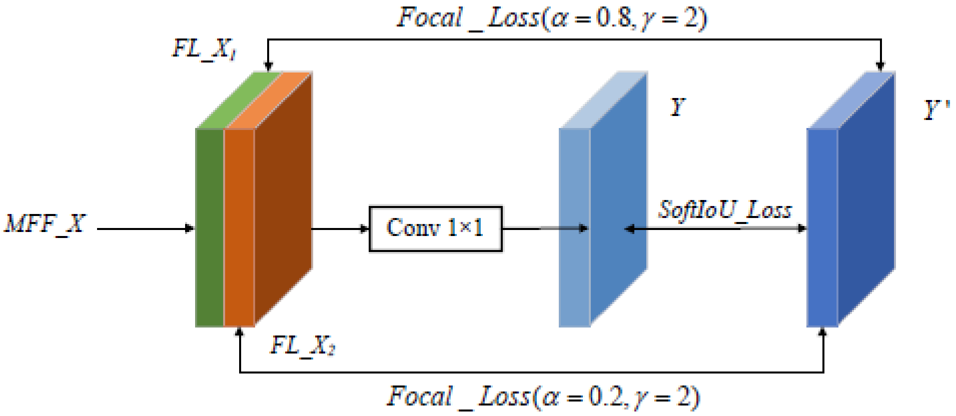


Figure 5. MLC module.

$SoftIoU\_Loss$ is calculated as follows:

$$L_{sf} = 1 - \frac{y \times y'}{y + y' - y \times y'} \ , \tag{9}$$

where y' is the labeled image, y is the network output image, $y \times y'$ is the intersection of the labeled image and the output segmented image, $y + y' - y \times y'$ is the concatenation of the labeled image and the output segmented image. The $SoftIoU_Loss$loss value reacts to the network model's ability to segment the contour of the actual object, and a larger value represents that the network output object is closer to the actual object in terms of shape and contour.

$Focal\_Loss$ is derived via Eq (10):

$$= -y\alpha(1-y')^{\gamma}\log y' - (1-y)(1-\alpha)y'^{\gamma}\log(1-y')$$
$$= \begin{cases} -\alpha(1-y')^{\gamma}\log y', & y=1 \\ -(1-\alpha)y'^{\gamma}\log(1-y'), & y=0 \end{cases} \tag{10}$$

where y' is the label, with values of zero and one representing the background and object pixels, respectively; y is the network output located within [0,1] range; α is the balancing factor, which balances the importance of the positive and negative samples and is generally taken to be 0.25 to solve the above problem; γ is the adjustment factor for the weights of simple and easy-to-classify samples; $\gamma > 0$ reduces the loss of easy-to-classify samples so that the network focuses on the difficult and wrongly classified samples and is generally taken to have a value of two to solve the above problem.

In the $Focal\_Loss$ loss function, the choice of the α value determines the correctness of the network output in favor of which type of pixel classification: The closer the α value is to one, the more attention the network pays to detecting the actual object, but it is easy to introduce false alarms such as corner points and strong edges, resulting in a high false alarm rate; the closer the α value is to zero, the more attention the network pays to removing the background clutter, but it is easy to remove weak gray pixels in the actual object, resulting in a high leakage detection rate. Therefore, the α value of 0.25 directly cannot achieve the best detection results for the A value of different selections of the network output caused by the impact of this paper to build a multi-loss constraint module MLC, in which for the *MFF_X* for the module's input feature map, the output results for the Y. Convolution operation on *MFF_X* to obtain the feature map $[FL\_X_1, FL\_X_2] \in R^{H\times W\times 2}$, in which the FL_X_1 and FL_X_2, respectively, through the activation function δ and the labeled image Y' for the $Focal\_Loss$ loss function, the parameters are $(\alpha = 0.8, \gamma = 2)$ and $(\alpha = 0.2, \gamma = 2)$ to ensure that FL_X_1 has a high detection rate and FL_X_2 has a low false alarm rate. A 1×1 convolution is used to fuse the features of FL_X_1 and FL_X_2 to integrate the advantages of the two feature maps effectively. The fusion result is Y after the activation function δ, and the $SoftIoU\_Loss$ loss function is calculated between Y and the labeled image Y'. This ensures that the classification of individual pixels in the network output is correct. At the same time, the shape and contour of the whole object area are as close to the actual object as possible.

In summary, the final loss function is defined as shown in Eq (11).

$$Loss = \varepsilon_1 \times L_{fl}(\delta(FL\_X_1), Y', \alpha = 0.8, \gamma = 2) + \varepsilon_2 \times L_{fl}(\delta(FL\_X_2), Y', \alpha = 0.2, \gamma = 2) + \varepsilon_3 \times L_{sf}(\delta(Y), Y') \quad , \tag{11}$$

where δ is the activation function $tanh/2 + 0.5$; $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are the corresponding weights of each loss function, intending to adjust the values of each loss function to the same order of one, one, and 1000, respectively; $L_{fl}(g)$ and $L_{sf}(g)$ are $Focal\_Loss$ and $SoftIoU\_Loss$, respectively; Y and Y' are the final outputs of the network and the labeled image, respectively.

3. Results

3.1. Datasets

In this study, the Nanjing university of aeronautics and astronautics-SIRST(NUAA-SIRST) dataset serves as the foundational dataset, which was created by Dai et al. using short-wave, mid-wave and long-wave infrared cameras. This dataset comprises 427 frames of images, each with a size of 300 × 300 pixels and containing 480 infrared targets. Among these images, 55% of feature targets occupy only 0.02% of the image size, with dimensions approximately 3 × 3 pixels. This scenario necessitates detection models to capture more contextual information, and the models need to possess stronger feature extraction capabilities to deal with dim targets against cluttered backgrounds. Additionally, 10% of the images contain two or more infrared targets, breaking away from the single-target scenario where detection models can only detect the most sparse or prominent influences. In 35% of the images, the grayscale values of the targets are higher than those of the entire image, with most targets exhibiting minimal differences from the background. This aspect effectively enhances the model's ability to improve target saliency detection.

Due to the limited number of frames in the NUAA-SIRST dataset, it can pose challenges for large-parameter network models, leading to issues such as unstable training, model convergence difficulties, and overfitting when dealing with a small amount of data. Therefore, it is necessary to augment the base data. Augmentation techniques employed include rotation, cropping, the addition of random noise, and introducing weak small targets following a Gaussian distribution. Through data augmentation, the dataset size is expanded to a total of 5000 images.

To enhance the robustness of the network models, the authors augmented the NUAA-SIRST dataset by capturing real infrared scenes under different backgrounds using existing long-wave infrared cameras, thereby creating the north university of China-SIRST (NUC-SIRST) dataset. The infrared backgrounds in the NUC-SIRST dataset encompass various scenarios, including sky/cloud backgrounds, artificial architectural backgrounds, pedestrian interference backgrounds, and natural scenery backgrounds.

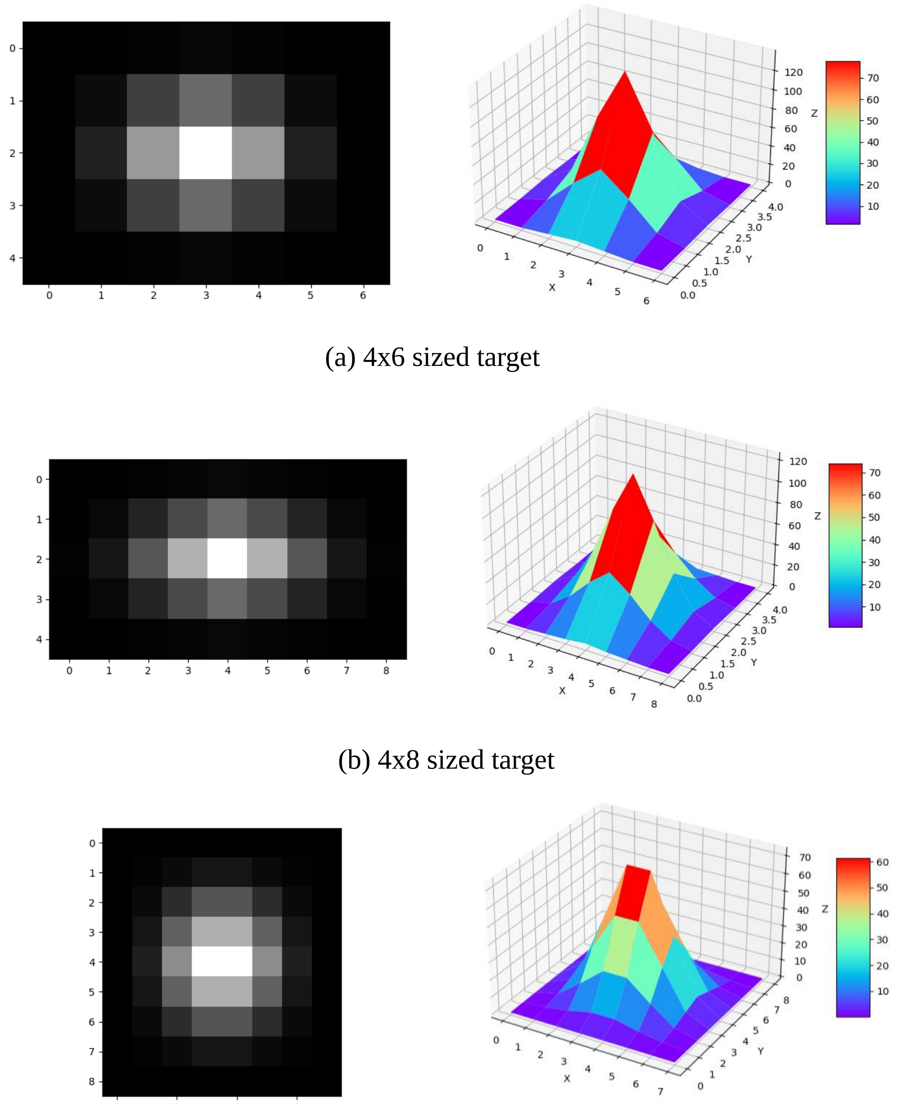
The targets in the NUC-SIRST dataset are manually added virtual targets, and the target generation function is as follows:

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})=\frac{1}{2\pi\sigma}\exp(-(\frac{(\boldsymbol{x}-\boldsymbol{x}_0)^2}{2\sigma^2}+\frac{(\boldsymbol{y}-\boldsymbol{y}_0)^2}{2\sigma^2})) \quad , \tag{12}$$

where σ represents the variance, which is set to one, and x and y are the pixel coordinates within the target image, while $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ denote the coordinates of the target image's center point. Since $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ takes values in the range of 0–1, they are scaled to the range of 0–255 to conform to the distribution of grayscale values in infrared weak small targets. The scaling formula is as follows:

$$\boldsymbol{f}'(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\times\frac{\boldsymbol{G}_{set}}{\boldsymbol{G}_{\max}}, \tag{13}$$

where $\boldsymbol{f}'(\boldsymbol{x},\boldsymbol{y})$ is the grayscale value at coordinates x and y within the target image, $\boldsymbol{G}_{set}$ is the average grayscale value of the background image with the addition of a random value between 20 and 40 and $\boldsymbol{G}_{\max}$ is the maximum value in the entire target image. Finally, the three-dimensional grayscale distribution map of the target and the actual image are generated, as shown in Figure 6.



(a) 4x6 sized target

(b) 4x8 sized target

(c) 8x6 sized target

Figure 6. Three-dimensional grayscale distribution and actual images of targets at different scales.

To achieve a smoother transition between the targets and the background, generating a more realistic dataset, this paper introduced an Alpha channel for the generated targets, representing the image's transparency. In this setup, the targets have lower transparency at their central positions and higher transparency at their edges. This approach makes integrating targets into the background image more reasonable and lifelike. The final NUC-SIRST dataset is constructed to encompass various infrared weak small targets against different complex backgrounds, with target sizes ranging from 2 × 2 to 9 × 9 pixels and signal-to-noise ratios (SNR) below 5.0. The distribution of the NUC-SIRST dataset is as presented in the table below:

Table 1. The number of image frames in this paper's dataset varies depending on different backgrounds and the number of targets.

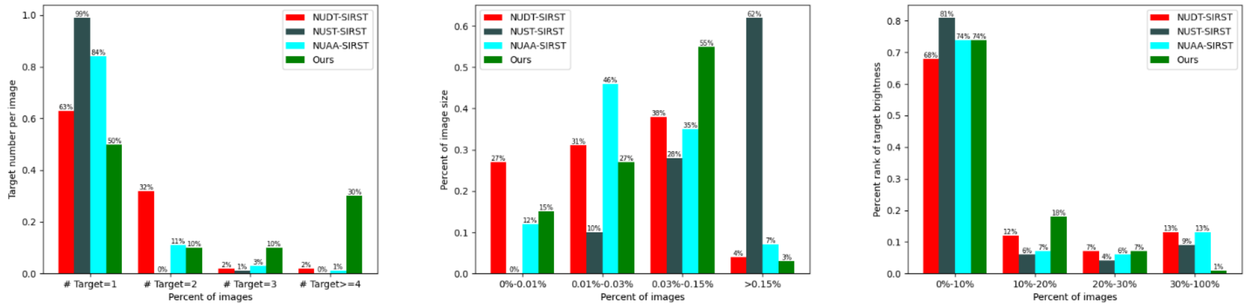
Image background	Number of images with 2 targets	Number of images with 3 targets	Number of images with 4 targets	Number of images with 5 targets	Number of images with 6 targets
Sky/cloudy background	250	250	250	250	250
Artificial architectural background	250	250	250	250	250
Pedestrian interference background	250	250	250	250	250
Natural scenery background	250	250	250	250	250

The merging of these two datasets caters to the training and testing of DL networks, making them suitable for research in areas such as feature extraction and detection of weak infrared small targets. Therefore, the combined dataset utilized in this paper consists of a total of 10,000 images. The dataset is partitioned into training, validation and test sets, with proportions of 70, 20, and 10%, respectively. To enhance the network's operational efficiency, all images in the dataset have been resized to 128 × 128 pixels.

To validate the fairness and representativeness of the dataset proposed in this paper, a comparative analysis is conducted with the NUAA-SIRST, Nanjing university of science and technology-SIRST (NUST-SIRST) and national university of defense technology-SIRST (NUDT-SIRST) datasets, as illustrated in the detailed comparisons presented in Table 2. Additionally, in the comparative analysis process, three metrics are employed for evaluation: Target quantity, target size, and target brightness ranking, as depicted in Figure 7.

Table 2. The primary characteristics of various popular SIRST datasets.

Datasets	Image type	Background scene	#Image	Label type	Target type
NUAA-SIRST	Real	Cloud/city/sea	427	Manual coarse label	Point/spot/ extended
NUST-SIRST	Synthetic	Cloud/city/river/road	10000	Manual coarse label	Point/spot
NUDT-SIRST	Synthetic	Could/city/sea/field/ highlight	1327	Ground truth	Point/spot/ extended
Ours	Real+Synthetic	Cloud/city/sea/filed/ person/noise	10000	Ground truth	Point/spot/ extended



(a) the number of targets (b) target size (c) target brightness

Figure 7. Comparison of existing public SIRST datasets.

According to Figure 7, concerning the comparison of target quantities, this paper's dataset comprises 50% of its data with no fewer than two infrared weak small targets, and 30% of the data features four or more infrared weak small targets, representing a higher number of targets compared to other datasets. Regarding the comparison of target sizes, 97% of the dataset's data have targets smaller than 0.15% of the image area, with most targets aligning with the definition of infrared weak small targets (i.e., targets smaller than 0.15% of the whole image area). They are larger than those in other datasets. Regarding target brightness ranking, 74% of the dataset's targets exhibit very low brightness, posing greater challenges for detection methods.

3.2. Experimental design

This study used the NUAA-SIRST dataset (with 427 sheets) as the base dataset for model training. For network models with many parameters, a small number of datasets can easily lead to problems such as unstable network training, failure of the model to converge, and overfitting, so it was necessary to expand the base data. Expansion methods included rotating, cropping, adding random noise, adding weak objects conforming to Gaussian distribution, etc. By expanding the base data, the amount of base data was increased to 5000 sheets.

Aiming to strengthen the robustness of the network model, infrared long-wave cameras were used to capture actual infrared scenes in different backgrounds. Next, we constructed the NUC-SIRST dataset to expand the NUAA-SIRST dataset, which had 5000 sheets, with the background of actual infrared backgrounds (including the background of the sky and clouds, the background of the man-made buildings, the background of the pedestrian interference, and the background of the natural scenery), and the objects were manually added. The virtual object's random number varied from one to three; the object position conformed to a uniform distribution, the object grayscale conformed to a 2D Gaussian distribution, the grayscale maximum value was 180~255, the SNR ratio ranged from 2.0 to 5.0 and the object size varied from 2 × 2 to 9 × 9). Therefore, the dataset used in this paper contained 10,000 images, of which the training set, validation set, and test set accounted for 70, 20, and 10%, respectively, and the image size was 128 × 128.

In this study, six evaluation indices, namely, signal-clutter ratio gain (SCRG), background suppression factor (BSF), intersection of union (IoU), precision (Pr), recall (Recall, Re) and F_{measure}, were used to evaluate the detection results of different methods. Among them, SCRG was used to evaluate the enhancement degree of the method to the object, BSF to evaluate the ability of the method to suppress the background, and IoU to assess the ability of the method to describe the contour of the actual object. Meanwhile, Pr and Re parameters were used to evaluate the ability of the method to remove false alarms and omission of detection, respectively. Finally, $F_{measure}$ reflected the combined performance of Pr and Re via Eq (14):

$$F_{\text{measure}}=\frac{2\times Pr\times Re}{Pr+Re}\ , \tag{14}$$

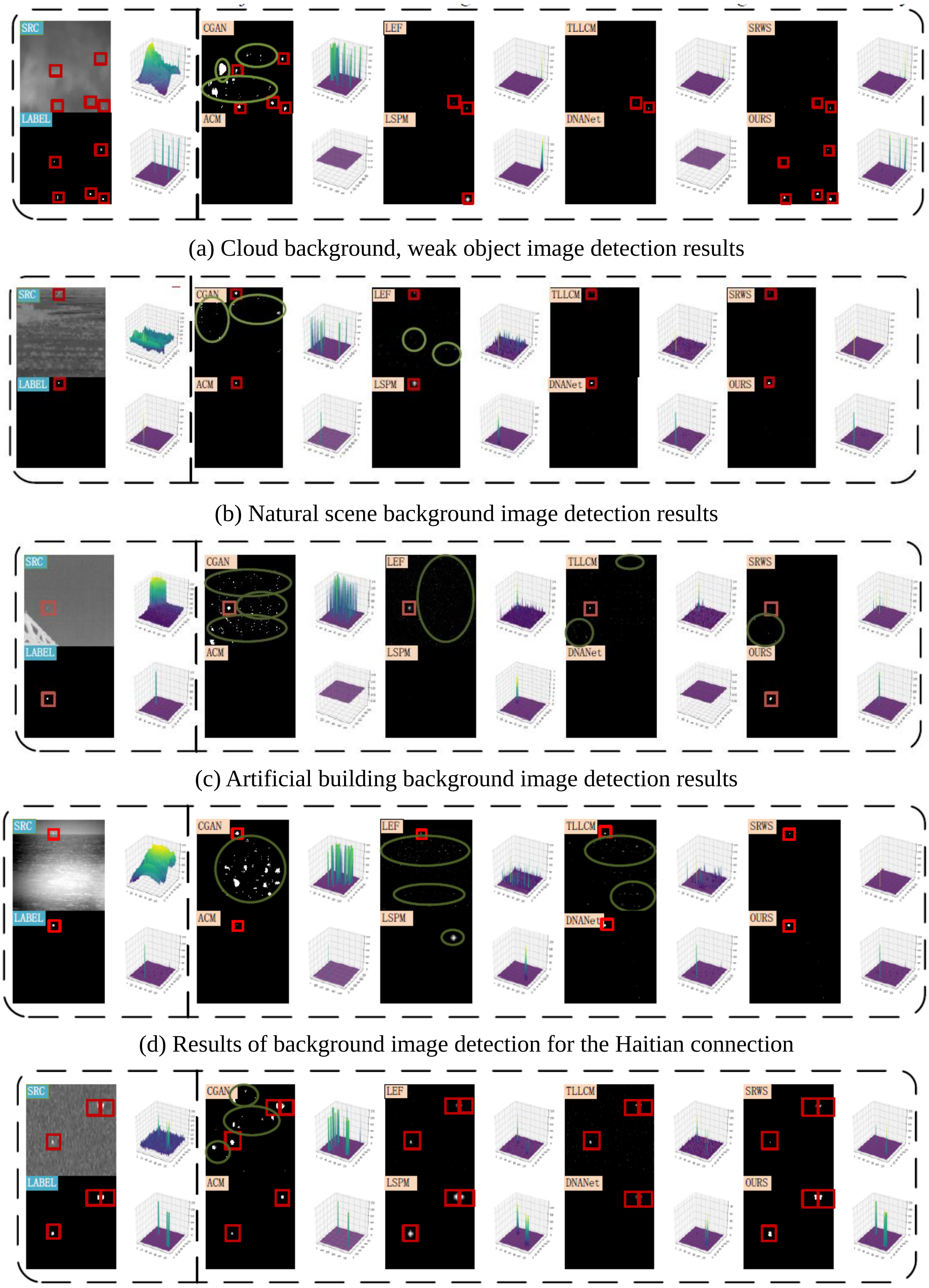
Larger values of the six indices indicated methods with stronger detection capability. In addition, this study used a receiver operating characteristic (ROC) curve to validate the proposed method's feasibility. The ROC curves were plotted with the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis. The closer the ROC curve of the particular method to the upper left corner, the better the method's performance.

To test the effectiveness of the proposed infrared small object detection method, this study used a large amount of actual and simulated data containing small objects for experimental verification. All codes in this study were run on an Ubuntu server with a Tesla M40 graphics card and 12 GB of video memory, using PyCharm 2019.3 as the test software. The numbers of network training, validation, and test data points were 7000, 2000, and 1000, respectively, with an image size of 128 × 128 and a total number of training epochs of 20. The learning rate was initialized at 0.001 and decreased by 30% every five epochs, the optimization function was selected as the Adam optimizer and the network framework was TensorFlow 2.4.

3.3. Comparative numerical tests

3.3.1. Qualitative comparative experiment and analysis

To verify the effectiveness of the proposed method, nine representative infrared weak object detection methods were selected as reference methods, including LEF [31], TLLCM (tri-layer local contrast measure) [32], SRWS, MDvsFA-CGAN, ACM, LSPM, DNANet, AGPC and FC3Net. Notably, the choice of different backbone networks in DNANet can significantly impact detection performance. To ensure a fairer and more representative comparison, we selected the DNANet detection method with the best performance for comparison, which incorporates the ResNet-18 backbone. Figure 8 shows the experimental results, in which the infrared source image was selected from the dataset with an obvious contrast effect, and the actual object area is shown with a red rectangular box. The detection results of the proposed and reference methods are visualized via a 3D salient map. The detection results of the methods on the actual object are shown in the rectangular box calibration part of the detection salient map, and the elliptical box indicates the detection of false alarms. If there is no rectangular box calibration, the actual object remains undetected or was hidden in the background clutter and could not be segmented accurately.



(a) Cloud background, weak object image detection results

(b) Natural scene background image detection results

(c) Artificial building background image detection results

(d) Results of background image detection for the Haitian connection

(e) Strong noise background, large-scale object image detection results

Figure 8. Infrared small object detection with different methods against various complex backgrounds.

As seen in Figure 8, for infrared source images with different types of complex backgrounds, the network proposed in this paper detected the obvious object of the salient map, located the object accurately, and the contour segmentation was close to the actual object. The reference methods used in this study had the following drawbacks:

1) The HVS class methods lead to a high leakage rate when the weak object and the neighborhood background have low contrast, such as in Figure 8(a), where the complex cloud layer is close to the gray level of the weak object, and the object is hidden in the cloud layer, resulting in a low gray level contrast, which is not easy to detect.

2) The paradigm-constrained optimization class methods lead to a high false alarm rate due to the sparse characteristics of the corner and the strong edges of the background, such as in Figure 8(c), where the network detects the strong edge of an artificial building and corner points. The strong edges and corner points of the artificial building in Figure 8(c) are mistakenly detected as objects.

3) The comparative DL methods do not perform object enhancement, so the weaker objects are easily lost in the feature extraction process. The loss function does not consider the grayscale distribution of the difficult-to-classify pixels (object edge pixels, corner points, strong edge pixels, etc.), so there is still room for improvement in the performance of the method's silhouette description, such as the method complexity. As shown in Figure 6(b),(d),(e), the background is interfered with by natural scenery, marine clutter, and strong noise; it is not easy to distinguish the actual object from the interfering objects, and the enhancement of the actual object does not result in false alarms and missed detection.

In contrast to the above reference methods, the proposed method ensured object enhancement in the process of feature extraction so that the actual object could be maintained in the deep layer of the network, and at the same time, the output layer of the network used the multi-loss constraint module to fully integrate the feature maps of low false alarms and low leakage detection, effectively extracting the object while reducing the impact of false alarms so that it could effectively detect multi-scale weak objects in all kinds of actual scenes, and the method was more robust.

3.3.2. Quantitative comparison experimental results and analysis

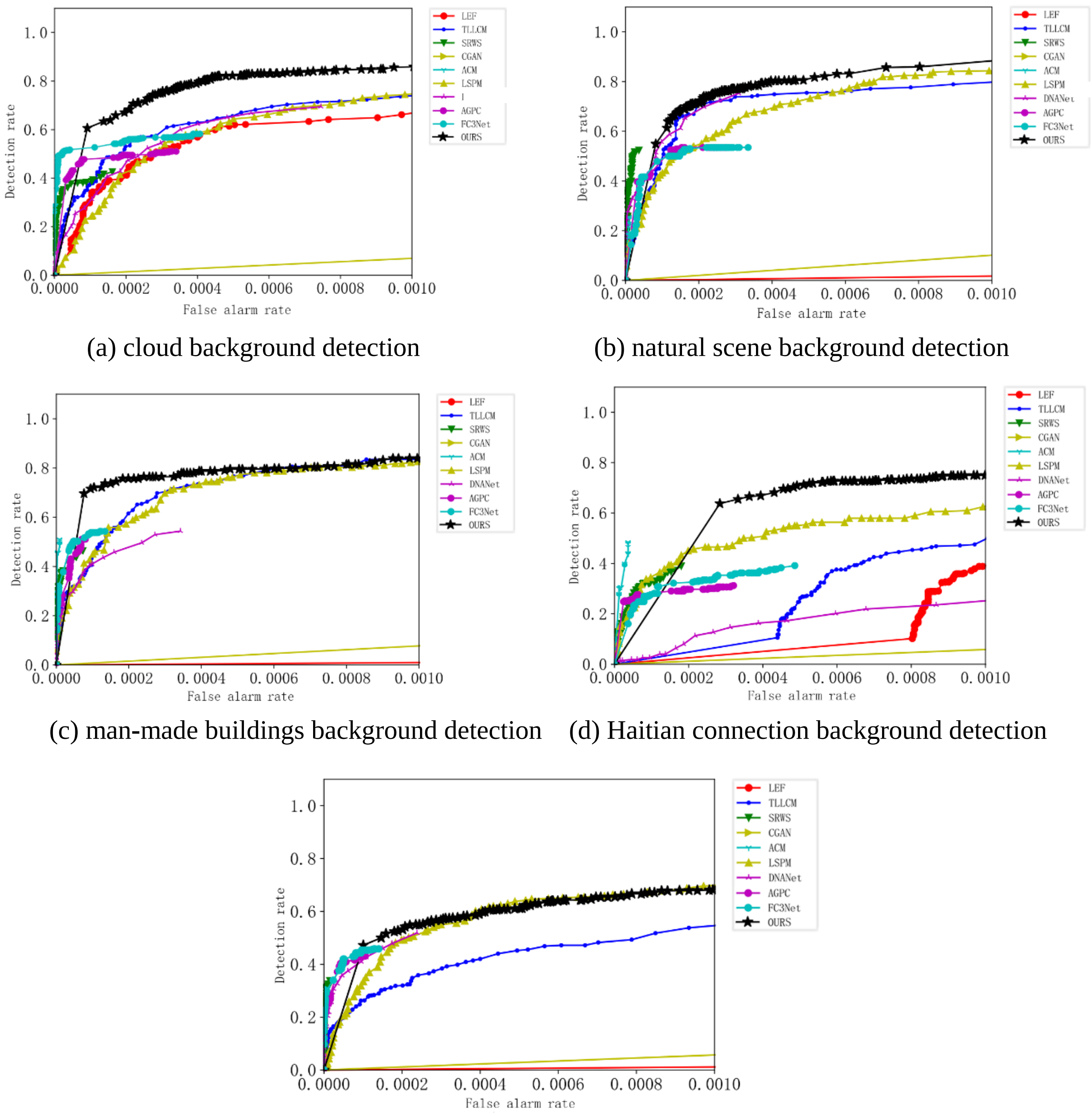
The metrics (indices) introduced in subsection 3.3.2 were selected to compare the performances of the proposed and reference methods, in which the test data was chosen to be 1,000 test sets of infrared source images with different types of complex backgrounds, different scales of objects, and different numbers of objects. To ensure that the metrics of all types of methods were assessed fairly, the DL methods were retrained using the dataset of this paper, while the mathematical modeling methods were tested using the best parameters designed in the original paper.

Table 3. Comparison of different methods' indices for small object detection results using various infrared images with complex backgrounds.

Index	LEF	TLLCM	SRWS	CGAN	ACM	LSPM	DNANet	OURS
SCRG	128.861	128.117	**734.878****	17.320	34.010	69.857	107.034	**1167.166***
BSF	23.191	22.548	**196.904****	1.121	24.697	44.976	**182.056***	63.233
IoU	0.303	0.334	0.191	0.117	0.382	0.356	**0.584****	**0.647***
Pr	0.829	**0.855****	**0.886***	0.123	0.556	0.436	0.680	0.719
Re	0.367	0.370	0.192	**0.872****	0.590	0.718	0.634	**0.895***
$F_{measure}$	0.426	0.459	0.298	0.192	0.572	0.493	**0.688****	**0.766***

Note: * is the optimal value for each row, and ** is the suboptimal value for each row.

In compliance with [33], the respective ROC curves were plotted in Figure 9.



(a) cloud background detection

(b) natural scene background detection

(c) man-made buildings background detection

(d) Haitian connection background detection

(e) strong noise background detection

Figure 9. ROC curves of detection results of different complex background types.

According to the data analysis of Table 1, the proposed method performed optimally in IoU, SCRG, Re, and $F_{measure}$ metrics, However, it poorly performed in BSF and Pr metrics. This can be attributed to the following reasons: The BSF value is related to the standard deviation of the background of the detected image. The SRWS method is based on optimizing the paradigm constraints and solves for the center of mass position of the actual object directly, with the position of the center of mass being defined as one and the background region suppressed as zero. The center of mass position was defined as one, and the background region was suppressed to zero. Therefore, this method achieved the highest BSF value, while the rest had a slightly lower BSF value because the background could not be completely suppressed.

Meanwhile, the Re and Pr metrics of the DL-based methods were lower than those of the mathematical modeling-based methods because the latter were more biased toward detection, i.e., they focused on obtaining the center-of-mass coordinates of the actual object and, therefore, had a higher value of Pr and stronger defalse-alarm ability. In contrast, the DL-based methods were more biased toward segmentation, i.e., they classified the object background category for each pixel. The BSF value was slightly lower for each pixel to classify the object background category; thus, they had higher Re values and better defalse alarm ability. Although the Pr index of the proposed method was lower than that of the mathematical modeling method, it had the best performance among the DL methods, outperforming all reference methods by Re index and implying its defalse-alarm and deleakage-detection abilities. IoU and $F_{measure}$ are the key indices for infrared weak object detection methods. The proposed method achieved the optimal performance in these two indices, where IoU measures the degree of similarity between the network output and the actual object. Larger IoU values indicate that the method learns the data distribution of infrared images better, improving the segmentation capability of the actual object contour. $\mathrm{F}_{\mathrm{measure}}$ is a comprehensive combination of Pr and Re indicators, which measures the detection accuracy of the method, and the larger the value, the easier it is to discriminate the actual object from the background clutter. Compared with the reference methods, the proposed one improved the IoU and $\mathrm{F}_{\mathrm{measure}}$ metrics by 10.8 and 11.3%, respectively. Therefore, combining the above six metrics and their significance, this method had better detection performance than the baseline one.

From the data analysis in Figure 9, it can be seen that the detection performance of this method is better than that of the comparison method under different complex background interferences. The detection rate can be higher simultaneously with the lower false alarm rate, and the detection rate is the first to reach the peak. In contrast, the comparison method reaches a certain threshold in the false alarm rate before the detection rate reaches the peak, so it can be seen that the method of this paper has a higher localization ability for the actual object and detects the location accurately.

3.4. Ablation numerical tests

This section describes the ablation tests conducted on the MTE module, the GTR module, and the MLC module to assess their effectiveness. The ResNet-18 module replaced the MTE module, the direct connection module replaced the GTR module and the MLC module $Focal\_Loss$ was directly applied to the network output layer with the parameter selection $(\alpha = 0.25, \gamma = 2)$. The controlled variable method was used to conduct the above tests.

Table 4. Index comparison of test results of different module combinations.

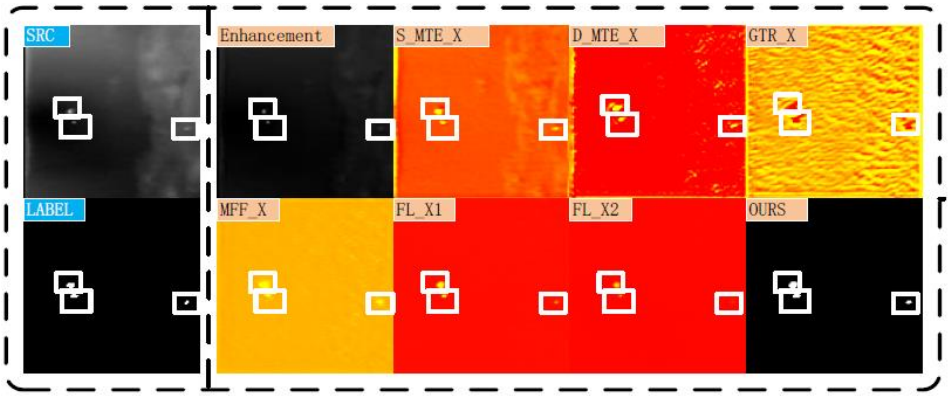
Module	SCRG	BSF	IoU	Pr	Re	$\mathrm{F}_{\text{-measure}}$
No	214.050	31.128	0.507	0.588	0.855	0.643
MTE	351.168	**68.591***	0.550	0.670	0.827	0.683
GTR	263.462	57.404	0.512	0.647	0.790	0.647
MLC	358.131	23.406	0.563	0.642	0.880	0.699
MTE+GTR	277.260	32.559	0.570	0.658	0.866	0.703
MTE+MLC	384.784	29.147	**0.596****	0.673	**0.895***	**0.726****
GTR+MLC	**568.131****	43.762	0.592	**0.680****	0.871	0.722
MTE+GTR+MLC	**1167.166***	**63.233****	**0.647***	**0.719***	**0.895****	**0.766***

Note: * is the optimal value for each row, and ** is the suboptimal value for each row.

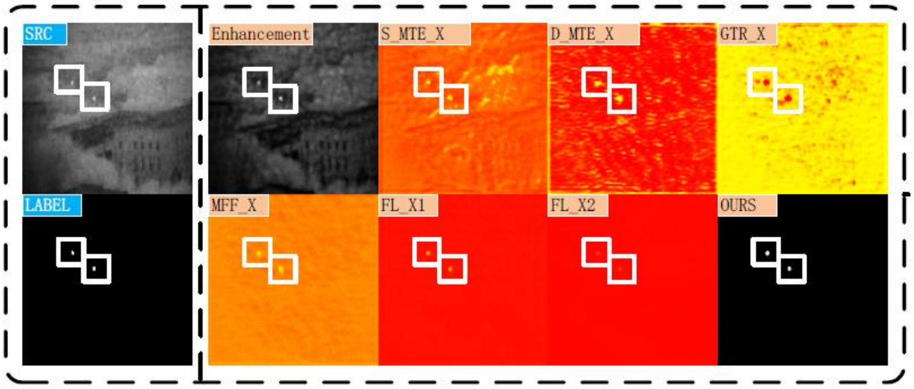
The results including the key indices IoU and $F_{measure}$ for various combinations of MTE, GTR, and

MLC modules and those excluding some of these are listed in Table 4. Their analysis proved that the GTR, MTE and MLC modules enhanced the detection performance of the network. Comparing the MTE, GTR and MLC module combinations, it can be seen that the importance of the GTR, MTE and MLC modules for detecting weak IR objects increases in turn. Comparing the MTE, GTR and MLC module combinations with the MTE + GTR, MTE + MLC and GTR + MLC module combinations, there was no conflict among the modules. The importance of the GTR, MTE, and MLC modules for infrared weak object detection increased. Their combination can further improve the performance of the network. An effective combination of the three modules can make the network model achieve the best detection effect, which also proves the practicality of the three network modules.

3.5. Network model interpretability tests



(a) Outputs based on the modules of the sky background



(b) Outputs based on the modules of the sky background

Figure 10. Output of different modules of the proposed network.

To verify the interpretability of the network proposed in this paper, infrared source images with different backgrounds were subjected to weak object detection. The outputs of each module of the network were converted and displayed in pseudo-color, and the actual objects were labeled using white rectangular boxes, as shown in Figure 9, where the left side of the dotted line is the source image and the corresponding labeled image. The right side of the dotted line, from left to right and from top to bottom: The image of the object enhancement using the mathematical model, the output of the shallow MTE module of the network, the output of the deep MTE module of the network, the output of the deep GTR module of the network, the output of the deep MFF module of the network, the output of the MLC module of the network $FL_X_1$and FL_X_2 and the final output of the network.

Comparing the object enhancement and source images, the mathematical model constructed in this study could enhance the object and suppress the background, proving that the mathematical model was effective and had interpretability. Comparing the object enhancement image and the shallow MTE_X and deep MTE_X images, it can be seen that the MTE module effectively enhanced the actual object. It was maintained from the shallow to the deep layer in extracting the image features, proving that the interpretation of the MTE module in subsection 2.2 was correct, and possessed interpretability. According to the GTR_X image, the object pixels considered the global information and established the long distance dependence between objects, and the actual object was responded to, proving that the GTR module was correctly explained in Section 2.3 and possessed interpretability. Comparing the deep MTE_X, GTR_X and MFF_X images, it can be seen that the MFF module fully integrated the features of MTE_X and GTR_X, decoded the actual object, and suppressed the background, proving that the explanation of the MFF module in subsection 2.4 was correct, and the module had interpretability. Comparing FL_X_1 and FL_X_2 of the MLC module and the final output image of the network, it can be seen that FL_X_1focused on the correct classification of object pixels, while FL_X_2 focused on the correct classification of background pixels. The final output of the network incorporated the characteristics of the two feature maps mentioned above. It effectively constrained the object contour using $SoftIoU\_Loss$, proving that the MLC module was correctly explained in subsection 2.5 and had interpretability.

In summary, the actual output of each network module was consistent with the theoretical design, and all of them had interpretability. When all the submodules of the network model can be interpreted, the whole multi-scale infrared weak object detection network formed by the combination of the modules is also interpretable.

3.6. Calculation of the complexity of detection methods

For the method proposed in this paper, we calculated its complexity, which mainly includes floating point operations (FLOPs), a measure of the computational complexity of neural networks and a measure of the number of model parameters. We compared it with other DL methods, as shown in Table 5.

Table 5. Complexity computation for different deep learning methods.

Index	MDvsFA	LSPM	DNANet	AGPC	FC3Net	This study
Parameter	3.9M	16.9M	4.7M	12.4M	7.0M	5M
FLOPs	61.7G	15.4G	3.51G	10.8G	648.7M	2.5G

From the experimental results, it can be seen that the method proposed in this paper is slightly higher than the DNANet and MDvsFA algorithms in terms of the number of parameters, but lower than other algorithms in terms of the FLOPs index. After testing, the detection frame rate of the algorithm proposed in this paper is about 20 frames, satisfying real-time demand.

4. Conclusions

This study proposed an interpretable multi-scale infrared weak object detection network design method for mitigating the problem of inaccurate object localization and contour segmentation in infrared weak object detection in a complex background. The proposed network model first performed object enhancement and shallow detail feature extraction on the input infrared source image and obtained high-level semantic features after cascading processing of multiple multi-scale object enhancement modules. Next, the low-level detail features and high-level semantic features were fused repetitively after calculating the global object response and completing the pixel classification of the actual object and the background noise in high-dimensional data. Finally, multiple loss joint constraint network outputs completed the pixel classification of the actual object and the background noise to make it close to the actual object distribution. Numerous comparative and ablation tests were conducted, proving the robustness of the proposed method and the effectiveness of each network module. For various types of infrared weak object detection in different types of complex backgrounds, the proposed method exhibited strong object detection and object contour description performances, and the designed detection system had high reliability. The follow-up study will focus on the infrared weak object detection method under the dual-drive mode, aiming at improving the coupling degree of the mathematical modeling method and DL method and enhancing the object detection capability of the refined method.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Natural Science Foundation of Shanxi Province of China (Grant No. 20210302123026), Fundamental Research Program of Shanxi Province (2023030212111 47) and Shanxi Province patent transformation project (2302001).

Conflict of interest

The authors declare no conflicts of interest.

References

1. S Wu, K Zhang, S Li, J Yan, Joint feature embedding learning and correlation filters for aircraft tracking with infrared imagery, *Neurocomputing*, **450** (2021), 104–118. https://doi.org/10.1016/j.neucom.2021.04.018
2. N. Zou, J. W. Tian, Research on multi feature fusion infrared ship wake detection method, *Comput. Sci.*, **45** (2018), 172–175.
3. C. Deng, S. He, Y. Han, B. Zhao, Learning dynamic spatial-temporal regularization for UAV object tracking, *IEEE Signal Process. Lett.*, **6** (2021), 1230–1234. https://doi.org/10.1109/LSP. 2021. 3086675
4. Y. Han, H. Liu, Y. Wang, C. Liu, A comprehensive review for typical applications based upon unmanned aerial vehicle platform, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*., **15** (2022), 9654–9666. https://doi.org/10.1109/JSTARS.2022.3216564

5. Y. Han, H. Wang, Z. Zhang, W. Wang, Boundary-aware vehicle tracking upon UAV, *Electron. Lett.*, **8** (2020), 873–876. https://doi.org/10.1049/el.2020.1170
6. X. Liang, L. Liu, M. Luo, Z. Yan, Y. Xin, Robust infrared small object detection using hough line suppression and rank-hierarchy in complex backgrounds, *Infrared Phys. Technol.*, **120** (2022),103893. https://doi.org/10.1016/j.infrared.2021.103893
7. L. Zhong, Y. He, J. W. Zhang, Small object detection algorithm based on the fusion of context and semantic features, *Comput. Appl.*, **42** (2022), 6.
8. L. Deng, J. Zhang, G. Xu, H. Zhu, Infrared small object detection via adaptive M-estimator ring top-hat transformation, *Patt. Recognit.*, **112** (2021), 107729. https://doi.org/10.1016/j.patcog.2020. 107729
9. Y. Li, Z. Li, C. Zhang, Z. Luo, Y. Zhu, Z. Ding, Infrared maritime dim small object detection based on spatiotemporal cues and directional morphological filtering, *Infrared Phys. Technol.*, **115** (2021), 103657. https://doi.org/10.1016/j.infrared.2021.103657
10. Y. Lu, S. Huang, W. Zhao, Sparse representation based infrared small object detection via an online-learned double sparse background dictionary, *Infrared Phys. Technol.*, **99** (2019), 14–27. https://doi.org/10.1016/j.infrared.2019.04.001
11. C. Chen, H. Li, Y. Wei, T. Xia, Y. Tang, A local contrast method for small infrared object detection, *IEEE Trans. Geosci. Remote Sens.*, **52** (2013), 574–581. https://doi.org/10.1109/TGRS.2013.2242477
12. L. Zhang, Z. Peng, Infrared small object detection based on partial sum of the tensor nuclear norm, *Remote Sens.*, **11** (2019), 382. https://doi.org/10.3390/rs11040382
13. T. Zhang, Z. Peng, H. Wu, Y. He, C. Li, C. Yang, Infrared small object detection via self-regularized weighted sparse model, *Neurocomputing*, **420** (2021), 124–148. https://doi.org/10.1016/j.neucom.2020.08.065
14. Y. Han, C. Deng, B. Zhao, D. Tao, State-aware anti-drift object tracking, *IEEE Trans. Image Process*., **5** (2019), 4075–4086. https://doi.org/10.1109/TIP.2019.2905984
15. I. V. Pustokhina, D. A. Pustokhin, T. Vaiyapuri, D. Gupta, S. Kumar, K. Shankar, An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety, *Saf. Sci.*, **142** (2021), 105356. https://doi.org/10.1016/j.ssci.2021.105356
16. S. H. Xie, W. Z. Zhang, P. Cheng, YOLOv4 fire and smoke detection model with embedded channel attention, *Chin. J. Liquid Crystal Displ.*, **36** (2021), 1445–1453.
17. Y. Dai, Y. Wu, F. Zhou, K. Barnard, Asymmetric contextual modulation for infrared small object detection, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2021). https://doi.org/10.1109/ WACV48630.2021.00099
18. Y. Dai, Y. Wu, F. Zhou, K. Barnard, Attentional local contrast networks for infrared small object detection, *IEEE Trans. Geosci. Remote Sens.*, **59** (2021), 9813–9823. https://doi.org/10.1109/TGRS. 2020.3044958
19. H. Wang, L. Zhou, L. Wang, Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 8509–8518. https://doi.org/10.1109/ICCV.2019.00860
20. B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, Dense nested attention network for infrared small target detection, *IEEE Trans. Image Process*., **32** (2022), 1745–1758. https://doi.org/10.1109/TIP.2022. 3199107

21. L. Huang, S. Dai, T. Huang, X. Huang, H. Wang, Infrared small object segmentation with multiscale feature representation, *Infrared Phys. Technol.*, **116** (2021), 103755. https://doi.org/10.1016/j.infrared.2021. 103755
22. T. Zhang, S. Cao, T. Pu, Z. Peng, AGPCNet: Attention-guided pyramid context networks for infrared small target detection, *IEEE Trans. Aerosp. Electron. Syst.*, **59** (2023), 4250–4261. https://doi.org/10.1109/TAES.2023.3238703
23. M. Zhang, K. Yue, J. Zhang, Y. Li, X. Gao, Exploring feature compensation and cross-level correlation for infrared small target detection, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 1857–1865. https://doi.org/10.1145/3503161.3548264
24. Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, *IEEE Trans. Emerg. Top Comput. Intell.*, (2021), 1–17. https://doi.org/10.1109/TETCI.2021.3100641
25. Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. https://doi.org/10.48550/arXiv.1904.11492
26. S. K. Ghosh, A. Ghosh, ENResNet: A novel residual neural network for chest X-ray enhancement based COVID-19 detection, *Biomed. Signal Process. Control*, **72** (2022), 103286. https://doi.org/10.1016/j.bspc.2021.103286
27. W. Li, J. Li, J. Li, Z. Huang, D. Zhou, A lightweight multi-scale channel attention network for image super-resolution, *Neurocomputing*, **456** (2021), 327–337. https://doi.org/10.1016/j.neucom.2021.05.090
28. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Esrgan: Enhanced super-resolution generative adversarial networks, in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018. https://doi.org/10.48550/arXiv.1809.00219
29. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in *Proceedings of the IEEE international conference on computer vision*, 2017. https://doi.org/10.48550/ arXiv.1708.02002
30. G. Chen, W. Wang, X. Li, Designing and learning a lightweight network for infrared small target detection via dilated pyramid and semantic distillation, *Infrared Phys. Technol.*, **131** (2023), 104671. https://doi.org/10.1016/j.infrared.2023.104671
31. C. Xia, X. Li, L. Zhao, R. Shu, Infrared small object detection based on multiscale local contrast measure using local energy factor, *IEEE Trans. Geosci. Remote Sens.*, **17** (2019), 157–161. https://doi.org/10.1109/LGRS.2019.2914432
32. J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, Q. Zhao, A local contrast method for infrared small-target detection utilizing a tri-layer window, *IEEE Trans. Geosci. Remote Sens.*, **17** (2019), 1822–1826. https://doi.org/10.1109/LGRS.2019.2954578
33. S. Huang, Y. Liu, Y. He, T. Zhang, Z. Peng, Structure-adaptive clutter suppression for infrared small object detection: Chain-growth filtering, *Remote Sens.*, **12** (2020), 47. https://doi.org/10.3390/rs12010047