Complex & Intelligent Systems
https://doi.org/10.1007/s40747-024-01410-6

ORIGINAL ARTICLE

A single-frame infrared small target detection method based on joint feature guidance

Xiaoyu Xu[1] · Weida Zhan[1] · Yichun Jiang[1] · Depeng Zhu[1] · Yu Chen[1] · Jinxin Guo[1] · Jin Li[1] · Yanyan Liu[1]

Received: 15 November 2023 / Accepted: 9 March 2024

Abstract

Single-frame infrared small target detection is affected by the low image resolution and small target size, and is prone to the problems of small target feature loss and positional offset during continuous downsampling; at the same time, the sparse features of the small targets do not correlate well with the global-local linkage of the background features. To solve the above problems, this paper proposes an efficient infrared small target detection method. First, this paper incorporates BlurPool in the feature extraction part, which reduces the loss and positional offset of small target features in the process of convolution and pooling. Second, this paper designs an interactive attention deep feature fusion module, which acquires the correlation information between the target and the background from a global perspective, and designs a compression mechanism based on deep a priori knowledge, which reduces the computational difficulty of the self-attention mechanism. Then, this paper designs the context local feature enhancement and fusion module, which uses deep semantic features to dynamically guide shallow local features to realize enhancement and fusion. Finally, this paper proposes an edge feature extraction module for shallow features, which utilizes the complete texture and location information in the shallow features to assist the network to initially locate the target position and edge shape. Numerous experiments show that the method in this paper significantly improves nIoU, F1-Measure and AUC on IRSTD-1k Datasets and NUAA-SIRST Datasets.

Keywords Deep learning · Infrared images · Small target detection · Global feature · Local feature

Introduction

In the field of marine rescue [1, 2] and traffic management, the detection of small targets such as drones and floating objects plays an important role. Restricted by the small size of the target, imaging distance, covert and other characteristics, in the use of visible light camera shooting and detection of small targets, there are leakage of false detection, etc.; and infrared imaging technology has the advantages of all-weather work, the face of the complexity of the background robustness, etc., the use of infrared imaging to assist in the detection of small targets, you can accurately capture the long-distance, small-sized target image. Therefore, the use of infrared imaging to detect small-sized targets is the focus of research by scholars today.

After decades of development, infrared small target detection methods have been transformed from model-driven to data-driven based. Among them, model-driven detection methods include filter-based methods [3, 4], human visual feature-based methods [5–7] and low-rank sparse decomposition-based methods [8–10]. Due to the over-

✉ Weida Zhan
zhanweida@cust.edu.cn

Xiaoyu Xu
cust-xxy@mails.cust.edu.cn

Yichun Jiang
jiangyichun@mails.cust.edu.cn

Depeng Zhu
zhudepeng@mails.cust.edu.cn

Yu Chen
chenyu@mails.cust.edu.cn

Jinxin Guo
guojinxin@mails.cust.edu.cn

Jin Li
jl11269@buaa.edu.cn

Yanyan Liu
liuyy306@163.com

[1] National Demonstration Center for Experimental Electrical, Changchun University of Science and Technology, street Weixing, Changchun 130022, Jilin, China

Published online: 24 April 2024

reliance on expert knowledge and feature templates, when the target size, shape and local signal-to-noise ratio change drastically, the above three methods are difficult to adapt to the changing scenarios, which can easily lead to false alarms and missed detections, resulting in poor robustness of the detection algorithms.

With the continuous development of deep learning technology, reaction diffusion neural network technology (RDNN), innovative event triggering strategies [11], point controllers [12], fuzzy model reconstruction and filtering methods [13], which provide new ideas for designing the image feature compression mechanism in this paper.

Meanwhile, data-driven small target detection methods are categorized into frame labeling-based detection methods [14–19] and pixel-by-pixel labeling segmentation methods [20–22], among which the pixel-by-pixel labeling segmentation detection methods are more effective in detecting single-frame target images.

Ren et al. [23] designed a dense nested interaction module to realize the progressive interaction of deep and shallow features, however, repeating the nested module many times can not further eliminate the background clutter and other information, and the key target information is lost in the downsampling process. information in the downsampling process. Wu et al. [24] nested Unet structures of different sizes layer by layer, which increased the depth of the network without decreasing the target resolution and avoided the loss of information in the downsampling process, but there is a problem of poor connection between global and local contextual information of weak target features. Li et al. [21] realized infrared small target pinpointing by anti-aliasing feature fusion module with ViT, however, the low-pass filtering led to the filtering of some targets with high response values, and the existence of ViT caused the model computation is too large.

Aiming at the problems in the above methods, this paper proposes an efficient infrared small target detection method. This method jointly utilizes BlurPool with multilevel dense connectivity to reduce the feature offset and loss of high-frequency features during continuous downsampling; this method proposes a deep feature mutual attention module and a shallow feature enhancement fusion module, which realizes global-local feature association and shallow feature enhancement, and reduces the computational difficulty of self-attention; this method proposes an edge feature module, which helps to recover the target shape and spatial location.

Specifically, the main contributions of the work in this paper are as follows:

1. In this paper, a feature extraction module based on BlurPool with multilevel dense0 connection is designed to mitigate the effect of downsampling operation on feature offset by blurring; at the same time, the lost high-frequency target features are supplemented by multilevel densities, which improves the consistency of the feature positions and the localization accuracy, and reduces the loss of infrared small-target features due to continuous downsampling.
2. In this paper, the mutual attention deep feature fusion module is designed to automatically learn the association expression between the background global features and the target localization, and model the feature associations in the global scope; at the same time, a compression mechanism based on the a priori knowledge and global attention is designed, which both reduces the computational difficulty of the model and improves the ability to perceive the global features through the cosine similarity indexing and the low-latitude remapping.
3. In this paper, the context local feature enhancement and fusion module is designed to extract the spatial information of the target from the target proximity region and the background region, and the deep semantic features are used to guide the enhancement and fusion of the shallow feature maps, which enhances the ability of adaptive characterization of the positional information of infrared small targets.
4. In this paper, an edge feature extraction module is designed to learn the edge response of shallow features through distance transform function, Sobel module and spatial channel weights to extract accurate edge feature representations, which assist the network in initial localization and segmentation of small target location and shape.

Related work

Model-driven infrared small target detection method

Early single-frame infrared small target detection methods mainly use null domain filtering methods, such as Max-Median filter [3, 4], Bilateral filter [25, 26], Top-Hat filter [27], etc. This type of method assumes that the background changes slowly and the neighboring pixels are highly correlated. To better explicitly construct a detection method that can reflect the characteristics of small targets and be more discriminative, Chen et al. [28] proposed a local contrast metric model, which utilizes the difference between an image block and its neighbors to construct a detection method based on local contrast. Han et al. [29] proposed a multi-scale local contrast algorithm (RLCM), a multi-scale three-layer local contrast detection framework (TTLCM). Xia et al. [30] proposed a detection method (LEF) that measures local contrast from local phase anisotropy and local luminance difference. Since the above two methods are less robust when facing complex backgrounds, some scholars have introduced the low-rank sparse decomposition into the field of infrared small target detection. Gao et al. [31] proposed the Infrared Patch-Image (IPI) model, by transforming the infrared small target detection problem into a low-rank sparse matrix decomposition problem. Dai et al. [32] proposed the column-weighted IPI model (WIPI) and the reweighted infrared patch tensor model (RIPT) on the basis of the IPI model.

Data-driven infrared small target detection method

At present, with the continuous improvement of the increasing maturity of deep learning technology, more and more scholars introduce deep learning technology into the field of infrared small target detection. Wang et al. [33] proposed to design three sub-networks in the adversarial network for image segmentation so as to balance the leakage rate and false alarm rate. Zhao et al. [34] proposed TBC-Net, a lightweight detection network, which uses the target extraction module and semantic constraints module to realize a model for real-time detection of infrared small targets by combining the semantic constraints. Ju et al. [35] proposed an efficient end-to-end detection network consisting of an image filtering module and an infrared small target detection module. detection module to form an efficient end-to-end network.

Infrared small target detection method based on global and local features

Dai et al. [36] regarded infrared small target detection as a semantic segmentation problem and designed an asymmetric background modulation module to aggregate deep and shallow features, however, infrared small targets have only a few effective pixels leading to poor detection. Li et al. [23] realized interaction between high and low level features as well as adaptive enhancement by means of densely nested interaction modules and a kind of cascading channel and spatial attention module. Multi-level features. Wang et al. [37] combined multi-stage, multi-scale localized features with a multi-stage feature pyramid to form the final detection result. Zhang et al. [38] sensed the pixel correlations within and between blocks at a specific scale, and then utilize the context pyramid module to fuse contextual information from multiple scales for better feature representation.

Infrared small target detection method based on attention mechanism

Tong et al. [39] proposed to improve the feature expression capability by enhancing the asymmetric attention module, same-layer feature information exchange and cross-layer feature fusion. Li et al. [21] proposed a proposed vision transformer branch that eliminates the interference of local flash elements by introducing non-local correlated features.

Advanced applications of deep learning techniques in the internet of things

Ali et al. [40] used a deep neural network architecture and data augmentation strategy to automatically learn a hierarchical representation of sensor data and efficiently capture temporal and spatial dependencies. Monem et al. [41] used stacked LSTM networks to capture complex patterns and temporal dependencies in selected features through temporal modeling to ensure the integrity and resilience of IoT networks. Abdelhafeez et al. [42] utilized a machine learning algorithm to extract sentiment-related insights from Twitter data to accurately classify the sentiment categories of tweets through an environmentally friendly preprocessing step and sentiment classification algorithm.

Advanced applications of deep learning techniques in Metaheuristics and Machine learning

Bacanin et al. [43] solved the overfitting problem by means of an explicit exploration mechanism and a chaotic local search strategy, and achieves superior results to other methods in image processing tasks. Malakar et al. [44] designed a hierarchical feature selection model based on genetic algorithm to improve the performance of handwritten text recognition technique by optimizing local and global features. Bacanin et al. [45] proposed an automated framework for solving the overfitting problem that employs a swarm intelligence approach to improve the model performance by selecting appropriate regularization parameters. Zivkovic et al. [46] proposed an automated image analysis framework based on CNN and XGBoost for identifying COVID-19 infected chest X-ray data and optimized the hyperparameters of XGBoost using hybrid AOA.

In summary, traditional infrared small target detection methods do not rely on a large amount of labeled data, but have poor adaptability in the face of complex backgrounds. Deep learning-based infrared small target detection methods can automatically learn and extract key features, but multiple downsampling operations can lead to feature offset and loss. At the same time, due to the lack of large-scale computationally powerful equipment and high-quality infrared small target data, this type of method has a large space for development. Therefore, it is of great significance for the field of infrared small target detection to attenuate feature offset and loss, enhance the degree of long-range feature correlation, strengthen shallow feature fusion, and improve the method operation speed.

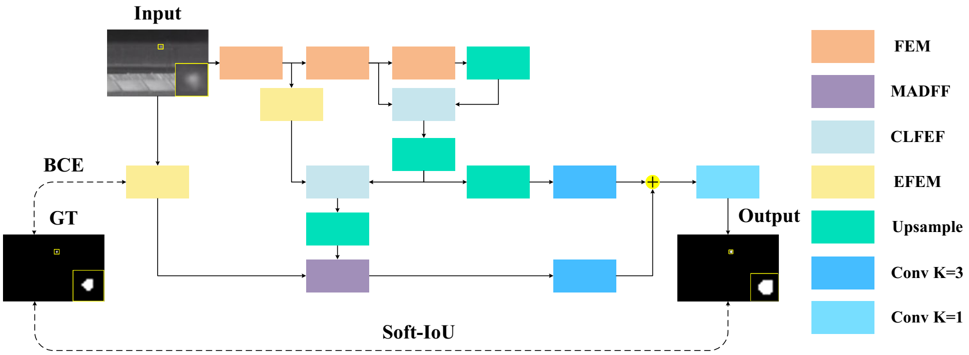


Fig. 1 Network framework of this paper

Methods

The framework diagram of the method proposed in this paper is shown in Fig. 1, which consists of four parts: feature extraction module(FEM), mutual attention deep feature fusion module(MADFF), context local feature enhancement and fusion module(CLFEF), and edge feature extraction module(EFEM). The working principle, design idea and specific implementation of each part will be introduced in detail in "Feature extraction module", "Mutual attention deep feature fusion module", "Contextual local feature enhancement and fusion module" and "Edge feature extraction module and loss function".

Feature extraction module

In classical u-shaped segmentation networks [20] or YOLO series detection networks [14–19], feature extraction networks are able to adequately extract different scales and different semantic levels of feature mapping through multiple convolutions and their cascading downsampling operations. However, due to the small size and irregular shape of infrared small targets and the extreme sensitivity of the targets to slight deviations in position [21, 47]. As downsampling continues in the feature extraction network, deviations in the spatial location of the target accumulate, leading to misalignment of the feature scale during deep feature fusion. Although, BlurPool is able to mitigate high frequency signal aliasing using a low-pass filter, it tends to attenuate small target regions at higher energy levels.

Meanwhile, multiple pooling operations are able to cause loss of important features in small-size targets during the process of reducing image feature resolution. Although, more jump-connected structures although more leap-connected structures allow more high-frequency interactions between features, the overly dense superposition operation can bring too many parameters and computations.

To address the above problems, this paper proposes an efficient feature extraction module (FEM) consisting of six BCRBs connected in series and a BPM.The BCRB reduces feature bias due to convolution operations by adding a BlurPool filter before each convolution operation and provides higher energy level target area while simplifying the structure by streamlining the number of hopping connections Features.By adding the BlurPool filter before the MaxPool pooling operation, the BPM is able to adaptively select the corresponding filling method according to the filter size, thus avoiding the loss of sparse features for small targets in the infrared due to multiple downsampling pooling operations and making the output features independent of the input offsets to improve the spatial consistency of the features.

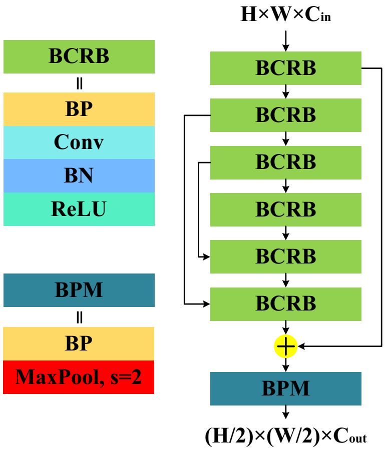


Fig. 2 Feature extraction module (FEM)

In FEM, BlurPool can be represented by Eqs. (1), (2):

$$\mathrm{BP} = \mathrm{LPF}(f) \tag{1}$$

$$\mathrm{LPF} = \int [-\infty, \infty] x(\tau) h(t - \tau) \mathrm{d}\tau, \tag{2}$$

where (BP) denotes (BlurPool), (LPF) denotes low-pass filtering, $x(\tau)$ denotes the input signal of (LPF), $h(\tau)$ denotes the frequency response of (LPF), and $y(\tau)$ denotes the filtered output signal.

Mutual attention deep feature fusion module

In the existing infrared small target detection techniques based on the Transformer attention mechanism [48], the following 2 main problems exist:

Although the Transformer attention mechanism is able to encode long distance dependencies in image features, a large number of complex local background and edge features are required if the location information of small target scarce features [35] is to be constructed, and the use of local features on a single scale is prone to feature learning bias.

Meanwhile, to achieve the effect of reducing the number of parameters and computation of the Transformer attention mechanism, pooling and other [49] methods are usually used to compress large-size features. However, pooling-based methods lead to redundancy of useless information; at the same time, it is difficult for linear attention-based compression methods to strike a balance between complexity and expressive power values [50, 51].

To solve the above problems, as shown in Fig. 3, this paper starts from the perspective of global feature association learning, adopts heterogeneous operators to compress the key information of shallow features and embed the deep features, introduces dilated convolution to filter the redundant high-frequency noises, utilizes the Transformer attention mechanism [48] to realize the interactive fusion of contextual information, and improves the network's attention to the location and shape of small targets.

Specifically, four dilation convolutions with dilation rates of 1, 2, 4, and 8 are used in this paper for local feature encoding, which capture the details and contextual information of the target at different scales, filter out the high-frequency background noise, and introduce complete and pure local features for the self-attention module.

Meanwhile, inspired by the bottleneck structure [52] in which a low-dimensional representation of image features is made, this paper proposed a compression mechanism based on a priori knowledge and global attention (EGF). As shown in Fig. 3, this paper sets up a set of key-value vectors LP_K and knowledge vector LP_V based on a priori knowledge and can be learned, and two kinds of vectors of size $\mathrm{HW} \times N$, where N can be set according to the actual situation of the number of dimensions by themselves. Firstly, the input image features are split into vector form $\mathrm{HW} \times C$ by doing Reshape operation, and then the transpose vector of each query vector LP_K is cosine similarity indexed (CSI) with the split feature vectors, and the useful features related to the target are widely queried in the whole feature map, and finally, the similarity measurements are done as low-latitude remapping (LLR) with the transpose vector of each knowledge vectors LP_V with Norm normalization to get the updated effective features in the shallow features after updating.

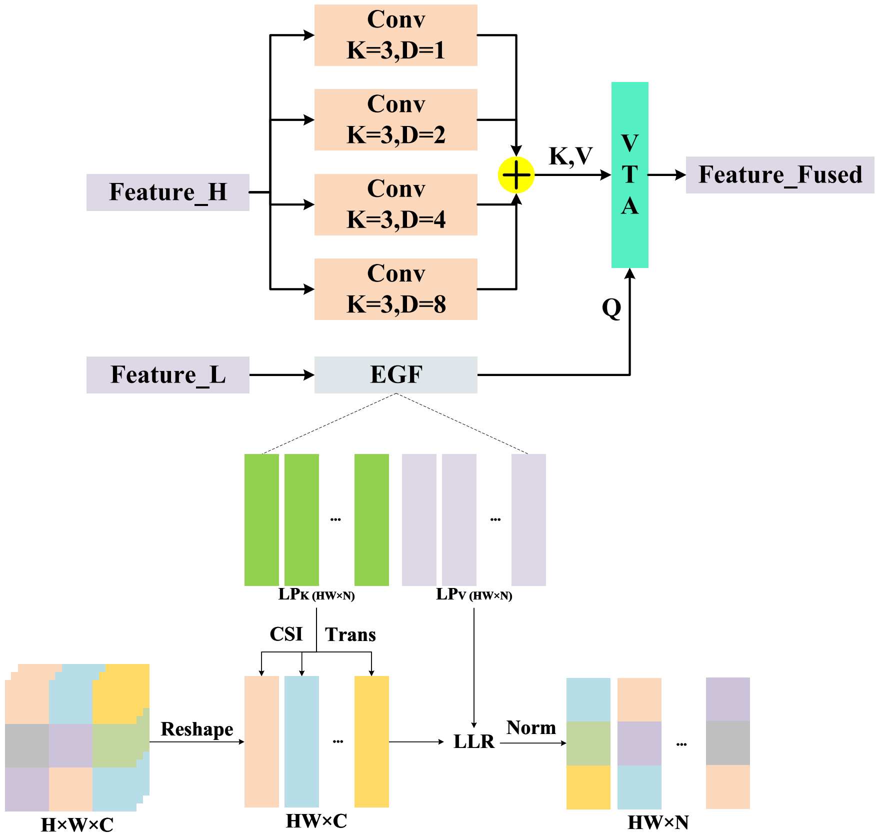


Fig. 3 Mutual attention deep feature fusion module (MADFF), VTA represents to vision transformer attention and EGF represents to efficient generalization of features

Among them, the learnable parameters LP_K and LP_V set in this paper can be regarded as multiple local information clustering centers based on a priori knowledge, and the local information represented by each clustering center can supplement the missing detail information in the deep semantic features; at the same time, the limited number of clustering centers can reduce the number of model parameters and computation amount, and realize efficient feature compression. After obtaining the learnable parameters LP_K and LP_V of the compressed large-size effective shallow features, they are fed into the Transformer module together with the query vectors Q in the deep features, so as to achieve the effect of recovering semantic information and fusing deep features.

Contextual local feature enhancement and fusion module

In existing infrared small target detection methods [22] the problem of determining the kind of internal pixels is only determined by using the semantic context information of all the pixels in the target's local area. However, infrared small targets are often in complex backgrounds [32], disturbed by natural and man-made environments as well as clutter noise, and lack physical features such as color and texture. Therefore, deep target semantic features and shallow target location features play an important role in target detection and recognition.

To solve the above problems and utilize deep semantic features H and shallow location features L more effectively. As shown in Fig. 4.

The input of the shallow feature map L is aggregated using global average pooling for global feature information, so that the network can better learn the relationship between infrared mini-targets and the background, and then the positional information of each channel is aggregated by point-by-point convolution, which enhances the distinction between the background and the target, and finally, the background and the target are enhanced by using the Sigmoid function to generate the bottom feature enhancement weights; the deep feature H mapping uses one-dimensional global average pooling to obtain the global features in the horizontal and vertical directions, respectively, and encodes the position coordinates using the horizontal and vertical directions to further deep semantic processing of the position features, and then generates the deep feature enhancement weights using the Sigmoid function. After multiplying and combining the shallow feature enhancement weights with the deep feature enhancement weights, the enhanced global features are fused into the underlying local features by elemental summation, and finally the enhanced fusion feature map with local and global features is generated.

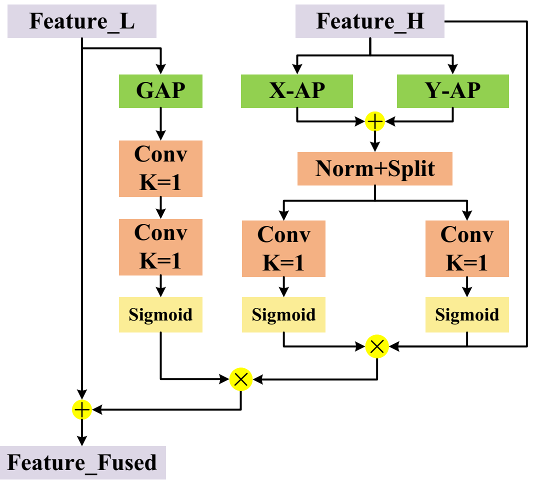


Fig. 4 Contextual local feature enhancement and fusion module (CLFEF)

As shown in Eqs. (3), (4), (5), the calculation of the enhancement and fusion module is as follows:

$$f(L) = \delta(\beta(\mathrm{pwc}(\lambda(\beta(\mathrm{pwc}(\mathrm{pool}(L))))))) \quad (3)$$

$$f(H) = H \cdot \delta(\omega(\mathrm{spl}(\omega([\mathrm{pool}^h(H) \cdot \mathrm{pool}^w(H)])))) \quad (4)$$

$$\mathrm{Fuse}(L, H) = L + f(L) \cdot f(H), \quad (5)$$

where $f(L)$ represents the output of the shallow features, $f(H)$ the output of the deep features, (pool) represents global average pooling, pool^h represents 1D average pooling in the x-axis; pool^w represents 1D average pooling in the y-axis; β represents BN normalization layer; λ represents ReLU activation function; δ represents Sigmoid function; ω represents 2D convolution operation; (pwc) represents PointWise point-by-point convolution operation; (spl) represents segmentation operation; $[\cdot]$ represents splicing operation; $\cdot$ represents convolution-by-convolution multiplication; X represents shallow feature maps; $\mathrm{Fuse}(L, H)$ represents the fused and enhanced features.

Edge feature extraction module and loss function

Infrared small target detection is characterized by low contrast and small size, if the shallow features can be extracted efficiently [53–55], it can retain clear detail information. At the same time, it provides the subsequent process by providing information about the shape and contour of the target. Therefore, fuzzy edges are analyzed and detected in shallow features, which help the network to accurately locate the segmented target area by learning shallow edge features.

As shown in Fig. 5, the module first uses Sobel with Gauss–Laplace operator to obtain the labeled map of the edge details of the infrared small target from the shallow features; then, the internal map of the infrared small target is obtained by using mean filter [56] and distance transform [57], and

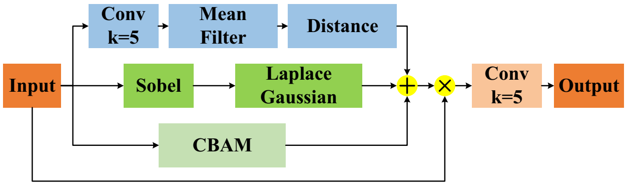


Fig. 5 Edge feature extraction module (EFEM)

the value of each pixel is replaced with the average of the pixel values of its surrounding neighborhood through mean filtering to reduce the influence of noise and make the edge details of the infrared small target more Clear and continuous, through the distance transform function to the internal pixels of the infrared small target will be assigned to the distance value closest to the edge, so that the internal region of the infrared small target has a higher gray value in the image, so that it is easier to be attended to and recognized by the attention mechanism; and then use the spatial-channel attention mechanism to obtain the dependence of the infrared small target on the whole image, and finally obtain the preliminary prediction map of the infrared small target, and with the real target mask to do BCE loss operation. In this paper, the above operations are used to help the network to supervise the training of feature maps, so as to retain more target high-frequency boundary information, which is used to compensate for the texture information lost in the convolution process. EFEM can be represented by Eqs. (6), (7), (8):

$$f_{\text{labeled}} = \text{LG}(\text{Sobel}(f_{\text{in}})) \tag{6}$$

$$f_{\text{inter}} = \text{Dis}(\text{Mean}(\text{Conv}(f_{\text{in}}))) \tag{7}$$

$$f_{\text{out}} = \text{Conv}((f_{\text{labeled}} + f_{\text{inter}} + CBAM(f_{\text{in}})) \times f_{\text{in}}), \tag{8}$$

where f_{in} represents input image features, f_{out} represents output image features, f_{labeled} represents detailed label image, f_{inter} represents internal map of features, (LG) represents Gauss–Laplace operator, (Dis) represents distance transform function, (Mean) represents mean filtering.

Mean filtering can be represented by Eq. (9):

$$\text{Mean}(x, y) = \frac{1}{N^2} \sum_{j=(-\frac{N}{2}\sim\frac{N}{2})}^{i=(-\frac{N}{2}\sim\frac{N}{2})} I(x+i, y+j), \tag{9}$$

where x, y represents the position of pixel point, N represents the neighborhood window size.

Distance transform function can be represented by Eq. (10):

$$\text{Dis}((x, y), (i, j)) = \sqrt{(x-i)^2 + (y-j)^2}, \tag{10}$$

where x, y represents the position of pixel point, i, j represents the coordinates of neighboring pixel points, $\text{Dis}((x, y), (i, j))$ represents the Euclidean distance between pixel (x, y) and neighboring pixel (i, j).

To train the training method proposed in this paper, the similarity between the prediction results and Ground Truth (GT) needs to be measured to maximize the prediction results to approximate the location and target pixels of the real infrared small target image, and this paper employs a hybrid loss function including the edge prediction loss and the detection result loss. Among them, the edge prediction loss uses the BCE-Loss loss function to construct accurate primary target edge features, and the detection result loss uses the Soft-IoU loss function to measure the gap between the prediction result and the real IR small target image. The two losses are shown in Eqs. (11), (12), (13):

$$L_{\text{bce}} = -\left(1 - \sum_{i,j} x_{i,j}\right) \log\left(1 - \sum_{i,j} s_{i,j}\right) - \sum_{i,j} x_{i,j} \log\left(\sum_{i,j} s_{i,j}\right) \tag{11}$$

$$L_{\text{soft}}(x, s) = \frac{\sum_{i,j} x_{i,j} \cdot s_{i,j}}{\sum_{i,j} s_{i,j} + x_{i,j} - x_{i,j} - s_{i,j}} \tag{12}$$

$$\text{Loss} = \alpha_1 L_{\text{bce}} + \alpha_2 L_{\text{soft}}, \tag{13}$$

where $s_{i,j} \in {}^{H\times W}$ represents the predicted score map, $x_{i,j} \in {}^{H\times W}$ represents the true mask map corresponding to the infrared small target image, $\alpha_1 = 0.3$ and $\alpha_2 = 0.7$.

Experiments

Datasets and evaluation metrics

Datasets

The IRSTD-1k Datasets [58] consists of 1000 infrared images with an image size of 512 × 512. These images were captured with an infrared camera in the real world, covering real targets such as drones, bright spots, boats and vehicles. The backgrounds include a variety of scenes such as oceans, rivers, fields, mountains, cities, and clouds.The NUAA-SIRST Datasets [23] contains a total of 427 real and independent high-quality infrared images covering real targets such as drones and boats, as well as highlights in a variety of scenes, such as oceans, rivers, fields, and mountains, cities, and clouds.

In this paper, the IRSTD-1k Datasets is divided into training dataset and validation dataset according to the ratio of 8:2, of which 80% is used for parameter training and tuning of the detection method, and 20% is used for evaluating the generalization ability of the detection method, respectively.

The above data division method ensures that different data samples are used in the training and validation process to accurately assess the effectiveness of the model.

To further evaluate the performance of the model, this paper selects the NUAA-SIRST Datasets as our test dataset. 50 images are randomly selected from the NUAA-SIRST Datasets as the test set, which have diverse scenes and lighting conditions, and are able to evaluate the robustness and accuracy of the detection model more comprehensively.

In addition to dataset segmentation, this paper also adopts a cross-validation approach to validate the performance of the detection method. Specifically, this paper uses K-fold cross-validation, where the training dataset is divided into K mutually exclusive subsets, and $K-1$ of them are used for training each time, and then the remaining one is used for validation. Such a repetitive process is performed K times to ensure that each subset is used for validation. Ultimately, in this paper, the results of the K validations are averaged to obtain a final performance evaluation.

Analysis of datasets

The above two datasets are applicable to the field of small target detection for binary classification, with only "target" and "background", and no classification of small target types.

To better design the infrared small target detection network, this paper analyzes the IRSTD-1k Datasets and NUAA-SIRST Datasets appearing in the paper for the physical characteristics of the small targets in infrared images using three metrics, namely, the number of targets, the size of the targets, and the brightness of the targets.

As shown in Fig. 6a, the two datasets have similar metrics in terms of the number of targets contained in a single image, with more than 80% of the images containing only 1 target.

As shown in Fig. 6b, the proportion of target size in the IRSTD-1k Datasets to the whole image is mostly concentrated in 0.03–0.15%, while the proportion of target size in the NUAA-SIRST Datasets is mostly concentrated in 0.01–0.03%.The image sizes of the NUAA-SIRST Datasets (300 × 300) are all smaller than the image size of the IRSTD-1k Datasets (512 × 512), which further indicates that the target sizes in the NUAA-SIRST Datasets are smaller.

As shown in Fig. 6c, the average brightness of small targets in the IRSTD-1k Datasets is much higher than that in the NUAA-SIRST Datasets.Meanwhile, the above two datasets are suitable for the target detection task of binary classification (the labeled images are binarized PNG images) and do not classify the specific kinds of targets. In summary, the targets and backgrounds in the two datasets are in an unbalanced state.

Evaluation metrics

In this paper, the infrared small target detection problem is regarded as an image segmentation problem, so the normalized intersection ratio (nIoU) [36], subject operating characteristic curve (ROC) [49], Precision, Recall, F-Measure, Cohen Kappa [59] are used as the evaluation metrics for assessing the detection methods.

(1) Cohen Kappa

For infrared small target detection data, there is a common problem of unbalanced target size and brightness distribution, this paper uses Cohen Kappa coefficient to evaluate the consistency of the prediction results with the real labeled images.

The Cohen Kappa coefficient can be represented by the Fig. 7 and Eqs. (14), (15), (16):

$$\text{Kappa} = \frac{p_o - p_e}{1 - p_e} \tag{14}$$

$$p_o = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \tag{15}$$

$$p_e = \frac{(\text{TP} + \text{FN})(\text{TP} + \text{FP})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})^2}, \tag{16}$$

where p_o denotes the sum of the number of correctly categorized samples in each category divided by the total number of samples; p_e denotes the result of multiplying the number of true samples in each category by the number of samples in the corresponding category, divided by the square of the total number of samples.

(2) nIoU

When combining deep learning techniques with infrared small target detection [36], the output of the detection method is a binary mask, and the values of these metrics are usually infinite. Meanwhile, existing segmentation-based detection methods model infrared small target detection as a segmentation process [60], which often sacrifices the integrity of the segmented target for a higher detection rate [61]. To obtain more scientific and accurate detection results, this paper adopts standardized Intersection over Union (nIoU) instead of IoU.

nIoU is defined as:

$$\text{nIoU} = \frac{1}{N} \sum_{i}^{N} \frac{\text{TP}[i]}{T[i] + P[i] - \text{TP}[i]} \tag{17}$$

where N represents the total number of samples in the dataset; $\frac{\text{TP}[i]}{T[i]+P[i]-\text{TP}[i]}$ represents the IoU value of the ith sample; $T[i]$ represents the number of target true mask pixels of the ith sample; $P[i]$ represents the number of pixels of the predicted segmentation result of the ith sample; $TP[i]$

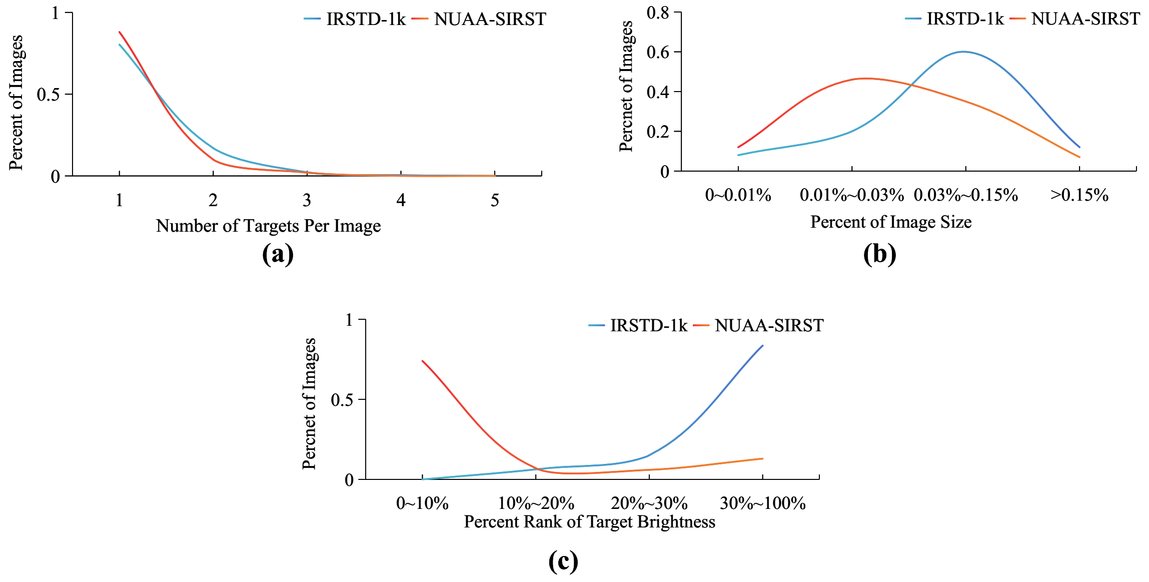
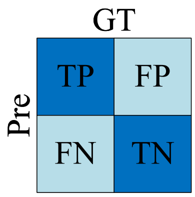


Fig. 6 Number of targets, target size, target brightness in IRSTD-1k datasets and NUAA-SIRST datasets

Fig. 7 Confusion matrix



represents the number of true mask pixels in the predicted segmentation result of the (ith) sample.

(3) ROC, AUC

The ROC curve is a common tool for evaluating the performance of binary classification algorithms, and the method in this paper is suitable for segmenting the target to be tested from an infrared image containing only one class of target and background. The curve is able to present a moving trend between false positive rate (FPR) and true positive rate (TPR), which helps the detection method to achieve the best detection performance under a sliding threshold.

TPR (True Positive Rate) indicates the true positive rate, the proportion of samples with true prediction and correct prediction to all true samples, as shown in Eq. 18:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} \tag{18}$$

FPR (False Positive Rate) represents the false positive rate, the proportion of samples with true predictions and incorrect predictions to all non-true samples, as shown in Eq. (19):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{P} \tag{19}$$

AUC is a comprehensive metric that does not depend on a specific threshold and is able to measure the accuracy of detection methods. At the same time, it can visually compare the performance differences between different detection methods on the ROC curve. In infrared small target detection, AUC is able to compare the detection capabilities of different methods in different scenarios.

(4) P-R, F1-Measure

F1-Measure is a metric for evaluating the performance of classifiers and is commonly used in binary classification problems. In infrared small target detection, targets and non-targets are usually categorized as two classes and then F1-Measure is used to evaluate the detection performance.

The calculation of Precision, Recall and F1-Measure are shown in Eqs. (20) to (22):

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{20}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{21}$$

$$F1 = \frac{2 \times (\mathrm{Precision} \times \mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}} \tag{22}$$

Experimental details

In this paper, to validate the effectiveness of the proposed method, we configured the following hardware and software environments: Intel(R) Core(TM) i9-10850K, 32 GB of RAM and NVIDIA Geforce RTX3090 GPU, Windows 11 system environment, Pycharm 2022.3.2 and Pytorch 1.10.1. In this paper, the input image size is adjusted to 512×512, SGD is used as the optimizer, the initial learning rate is set to 1e-3 for 150 epochs, the batch-size is 8, and the learning rate is adjusted downward to the initial 0.1 every 50 epochs.

For the training data, to ensure that the scarce features of infrared small targets are not lost, this paper does not resize the input image, ensures that the size of the original image is unchanged and input into the network, and sets the batch-size to 10 according to the storage capacity limit of the test platform.This paper selects ALCNet [62], ACM-Unet [36], TBC-Net [34], IAAANet [22], IST-TransNet [21] and MTU-Net [49] as data-driven IR small target detection methods, and LCM [28], RLCM [29], RIPT [63] and IPI [32] are selected as model-driven IR small target detection methods. To make a fair comparison, the hyper-parameters of all methods are used as provided by the original authors, and they are all trained and tested on the same dataset, and the test results will be taken as the mean value of all test images.

Comparative experiments

Quantitative analysis of IRSTD-1k Datasets

In this paper, a comprehensive comparison is made with the methods of this paper using the comparison methods on the IRSTD-1k Datasets.

The quantitative analysis data of each comparison method on the IRSTD-1k Datasets. As is shown in Table 1, in the model-driven detection based methods, the LCM, RLCM and RIPT cannot exclude the influence of background and clutter noise, which leads to the positive samples not being detected correctly and a large number of negative samples being mistaken as positive samples, resulting in a lower Precision and higher Recall situation; limited by the fixed target shape and size in the method, the IPI method is less robust to target diversity, resulting in lower F1-Measure and AUC indexes.

Among the data-driven detection methods, the method in this paper achieves 0.8125, 0.8588, 0.8916, and 0.8351 on multiple metrics of nIoU, F1-Measure, AUC and Cohen Kappa, respectively, which are 3.21%, 6.13%, 5.40%, and 6.42% better than the sub-optimal values, and achieves the optimal values among the various methods The optimal value among the multiple methods is reached. Finally, as shown in Fig. 8, this paper visualizes the three evaluation metrics of nIoU, F1-Measure and AUC by the method of box plot, and the method of this paper has no outliers and has high robustness.

Meanwhile, after comparative experiments and tests, compared with IAAANet, IST-Transformer and MTU-Net, which use the self-attention mechanism, this paper's method achieves a better balance between the number of parameters and the computation speed (on GPU) of a single image; compared with TBC-Net, this paper, with the addition of the self-attention mechanism and a large number of cascade operations, is still can realize processing a 512×512 image in 0.1 s, proving the effectiveness of the EGF method in MADFF.

Quantitative analysis of IRSTD-1k datasets

As shown in Fig. 11, this paper demonstrates the detection effect of some classical detection methods and SOTA methods on IRSTD-1k Datasets, and the model-driven detection methods based on model-driven detection have more problems of false alarms and missed detections.

The LCM and RLCM methods based on the principle of local contrast maximization can detect some small targets, but when facing a complex background or a small difference between the background contrast and the target, there is still the problem of false alarms due to the existence of the blocking effect itself. The IPI method based on local block tensor analysis is able to predict multiple targets with different sizes more accurately, but suffers from the problem of blurred edge features when facing targets with diverse shapes and sizes.

Although data-driven detection-based methods generally obtain better detection results and robustness, they still have more problems.

AlCNet, ACM-Uent methods are more stable in different backgrounds, but due to the fact that they are only based on multiple levels of concurrency in the local feature area, they do not associate global multi-scale features with contextual information, which makes it impossible to distinguish complex backgrounds such as cloud layers.

As the Transformer attention mechanism in IST-TransNet, IAAANet and MTU-Net methods extracts a large number of long range global features such as useless background and noise, and ignores the problem of the scarcity of local features of small targets, which fails to provide the Transformer with a sufficient amount of effective target data, resulting in a large number of false alarms when the target is obstructed by grass and trees. A large number of false alarm situations.

Due to the use of lightweight strategies such as reducing jump connections and channel compression, the edge texture and detail features cannot be recovered, which leads to the loss of the smallest size target when the TBC-Net method performs multi-target detection.

In this paper, when facing backgrounds such as clouds, grasses and complex textured road surfaces, the method benefits from the BlurPool and cascade method in FEM, which reduces the target position offset and feature loss; MADFF can fully correlate the global features of the background with the local features of the target, which excludes the influence of a large amount of background and clutter information on the later target feature identification, and reduces the case of false alarm of the target; and then, the CLFEF utilizes the deep global features in MADFF to strengthen the shallow local features of the target, which further reduces the loss

Table 1 Comparison of multiple evaluation metrics for different comparison methods on the IRSTD-1k dataset

Method	nIoU ↑	Recall ↑	Precision ↓	F1-Measure ↑	AUC ↑	Cohen Kappa↑	Params↓	Time on GPU/s ↓
LCM	0.1936	0.8692	0.0406	0.0776	0.3996	0.2552	-	-
RLCM	0.3028	**0.8761**	0.0254	0.0494	0.4519	0.2746	-	-
RIPT	0.3240	0.8362	0.0356	0.0683	0.4119	0.1986	-	-
IPI	0.6469	0.5562	0.7255	0.6297	0.4671	0.3627	-	-
ALCNet	0.6343	0.6845	0.4910	0.5718	0.6924	0.5872	**3.5M**	0.255
ACM-Unet	0.7287	0.7829	0.7378	0.7597	0.7296	0.6583	5M	0.159
TBC-Net	0.7385	**0.8752**	0.6217	0.7270	0.7862	0.6829	**5M**	**0.049**
IAAANet	0.7642	0.8143	0.5904	0.6845	0.8125	0.7456	18.24M	0.12
IST-TransNet	0.7528	0.7861	0.6845	0.7312	0.7577	0.7084	7.0M	**0.085**
MTU-Net	**0.8259**	0.8058	**0.8222**	**0.8139**	**0.8459**	**0.7849**	7.5M	0.108
Ours	**0.8525**	0.8723	**0.8553**	**0.8637**	**0.8916**	**0.8351**	8.1M	0.096

The optimal and sub-optimal values of each metric are indicated in bold blue and red, respectively, and ↑ indicates that the larger the value of the metric, the better the detection performance. The method in this paper achieved better results in nIoU, F1-Measure, AUC and Cohen Kappa

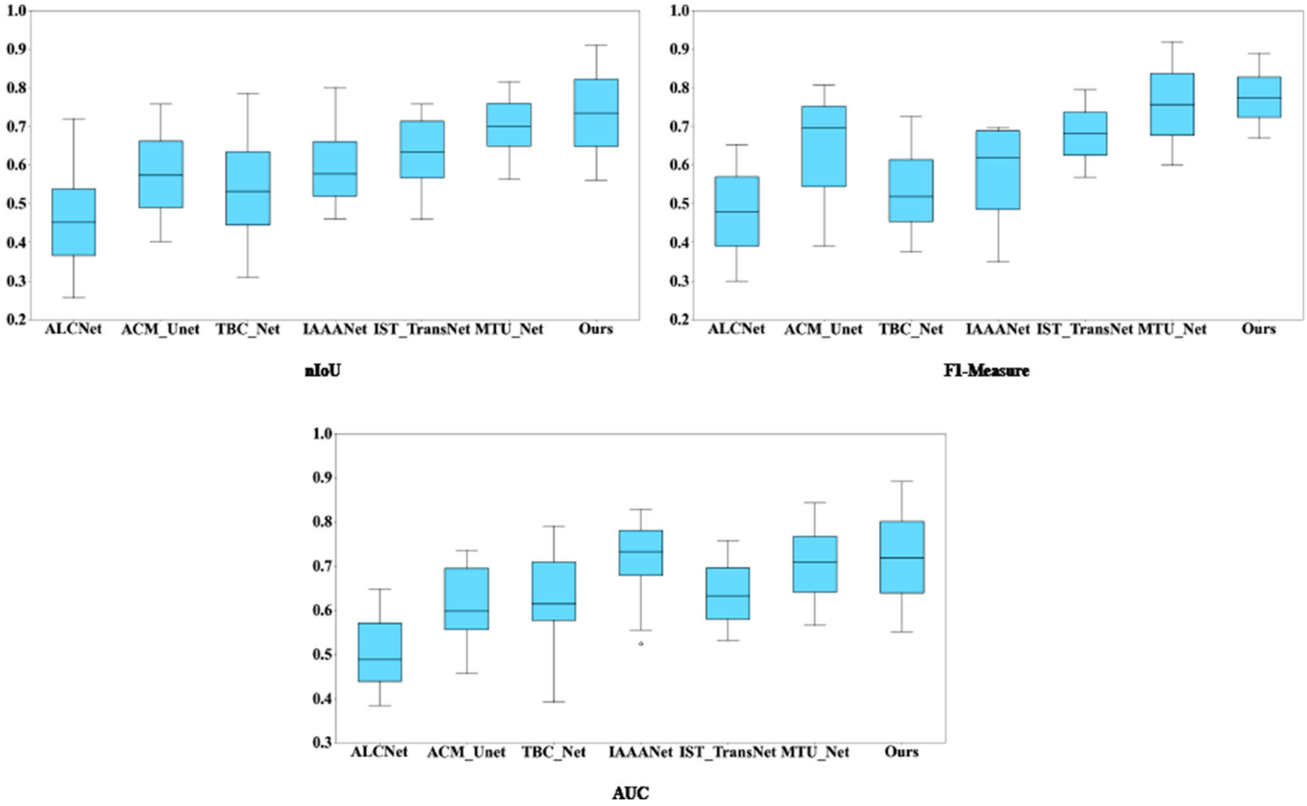


Fig. 8 nIoU, F1-Measure, AUC of different comparison methods in IRSTD-1k Datasets. As can be seen from the box plots, the method of this paper has high robustness with no outliers during the testing process, while obtaining competitive results

of internal target features; finally, EFEM utilizes the shallow contour and position information to ensure the integrity of the target shape and accurate localization.

Meanwhile, as shown in Figs. 9 and 10, this paper also plots the ROC curves of different methods on the IRSTD-1k Datasets with 3D effects, which proves that the proposed methods in this paper are obviously better than other methods.

NUAA-SIRST datasets

To fully reflect the validity of the methods in this paper, the same experimental setup and evaluation metrics as the comparison experiments on the IRSTD-1k Datasets are used in this paper, and the validation is carried out on the NUAA-SIRST Datasets (Fig. 11).

As shown in Table 2, LCM, RLCM and RIPT still appear to have generally lower Precision and higher Recall; compared with the above methods, the IPI method has better

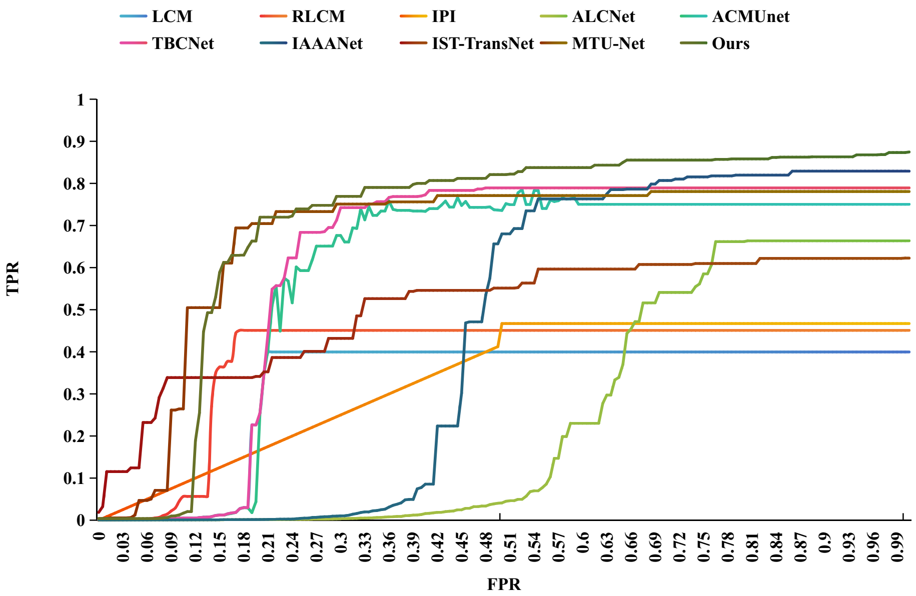


Fig. 9 ROC curves for different comparison methods on IRSTD-1k Datasets. The method in this paper demonstrates better classification and detection performance on the ROC curve

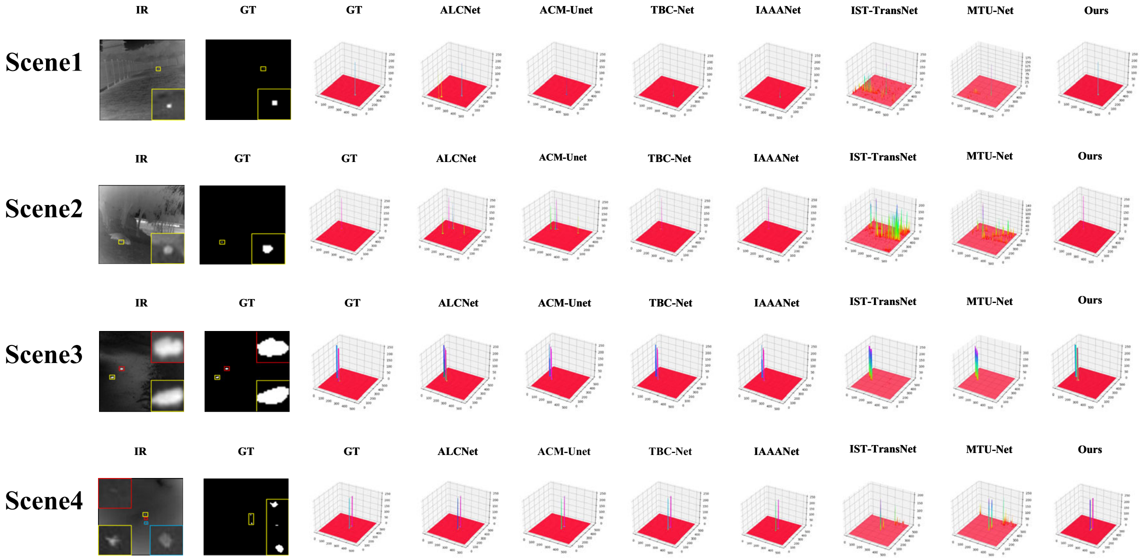


Fig. 10 3D visualization results of ALCNet, ACM-Unet, TBC-Net, IAAANet, IST-TransNet, MTU-Net and the proposed method in this paper on IRSTD-1k Datasets

Precision and Recall metrics, but worse F1-Measure and AUC metrics. The above detection methods rely heavily on a priori knowledge and manual parameterization, and in complex scenes, with the imaging distance and environmental factors changing, the shape and size of the small target will change accordingly, resulting in poor detection results. Meanwhile, among the data-driven detection-based methods, the method in this paper achieved 0.8057, 0.8111, 0.7981 and 0.7737 in nIoU, F1-Measure, AUC and Cohen Kappa metrics, respectively, which were 2.68%, 2.37%, 6.02%, and 6.77% higher than the suboptimal values, and reached the multiple methods' optimal value.

As shown in Fig. 12, this paper demonstrates the detection effect of some classical detection methods with SOTA methods on the NUAA-SIRST Datasets. The problems of false alarms, missed detections, loss of target center features and shape distortion when facing single target and multi-targets, which expose that the other comparison methods cannot exclude the background interference and lose a large number of target features in the detection process. In contrast, in this paper, through MADFF, CLFEF fully correlates the global and local features, weakens the influence of clutter noise, and preserves the internal features of the target; at the same time, EFEM preserves the position and edge information, and accurately reconstructs the shape of the target.

The analysis results of the comparison methods on the two datasets are consistent, and the quantitative and qualitative

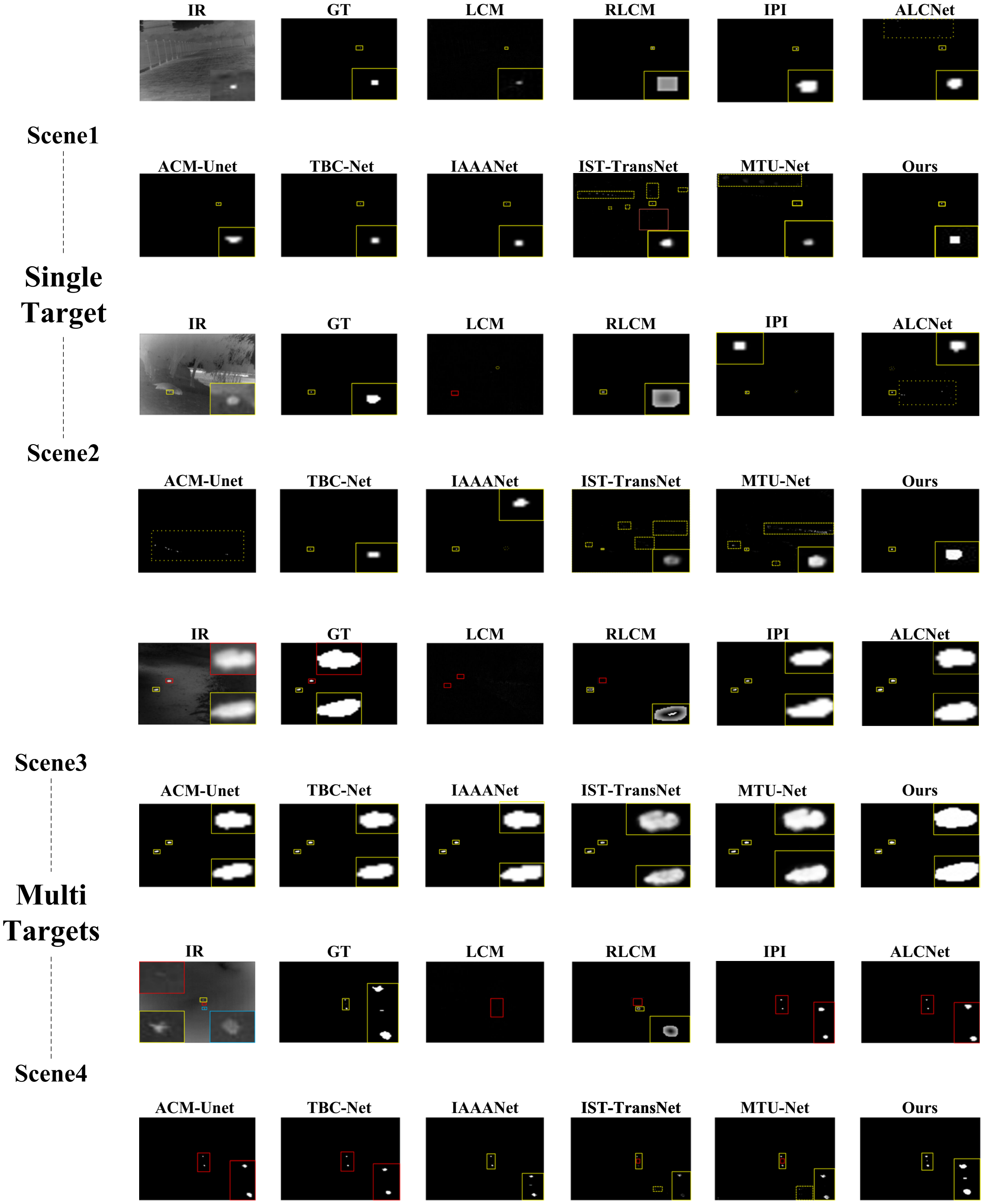


Fig. 11 Detection results of different detection methods on the IRSTD-1k Datasets. In the figure, the location of the small target is zoomed in, and the solid yellow box indicates "Target Successfully Detected", the dashed yellow box indicates "Target False Alarm", and the red box indicates "Target Miss Detection". The method in this paper does not have false alarms and missed detections, and effectively retains the target features and achieves better detection results

analysis results are consistent, which proves the effectiveness and advancement of this paper's method.

Ablation experiments

In this section, the paper uses convolutional layers and jump connections to form a Baseline containing 3 downsamplings. In this paper, ablation experiments are conducted on the proposed FEM, MADFF, CLFEF, and EFEM modules using the NUAA-SIRST Datasets, and the validity of each design is evaluated by adding different modules sequentially. To ensure the fairness of the ablation experiments, the same parameter settings are used in this paper for the ablation experiments of each type of module.

Table 2 Comparison of multiple evaluation metrics of different comparison methods on NUAA-SIRST Datasets

Method	nIoU ↑	Recall ↑	Precision ↓	F1-Measure ↑	AUC ↑	Cohen Kappa↑	Params ↓	Time on GPU/s ↓
LCM	0.2123	0.8507	0.0384	0.7352	0.3239	0.1896	-	-
RLCM	0.4296	0.8512	0.0301	0.5817	0.3958	0.2074	-	-
RIPT	0.2999	0.8129	0.0287	0.5543	0.4542	0.2358	-	-
IPI	0.6233	0.5714	0.7481	0.6479	0.4215	0.3681	-	-
ALCNet	0.6657	0.6581	0.4867	0.5596	0.6296	0.5428	**3.5M**	0.255
ACM-Unet	0.7748	0.7635	0.7996	0.7811	0.6730	0.5639	5M	0.159
TBC-Net	0.7230	**0.8571**	0.379	0.7325	0.7248	0.6294	**5M**	**0.049**
IAAANet	0.7863	0.8072	0.6114	0.6958	0.7523	0.6758	18.24M	0.12
IST-TransNet	0.7920	0.7935	0.7614	0.7771	0.7385	0.6441	7.0M	**0.085**
MTU-Net	**0.7847**	0.8023	**0.7825**	**0.7923**	**0.7528**	**0.7246**	7.5M	0.108
Ours	**0.8057**	**0.8185**	**0.8038**	**0.8111**	**0.7981**	**0.7737**	8.1M	0.096

The optimal and sub-optimal values of each index are shown in bold blue and red, respectively, and ↑ indicates that the larger the index value, the better the detection performance. The method in this paper obtains better results on nIoU, Precision, F1-Measure, AUC metrics, and Cohen Kappa

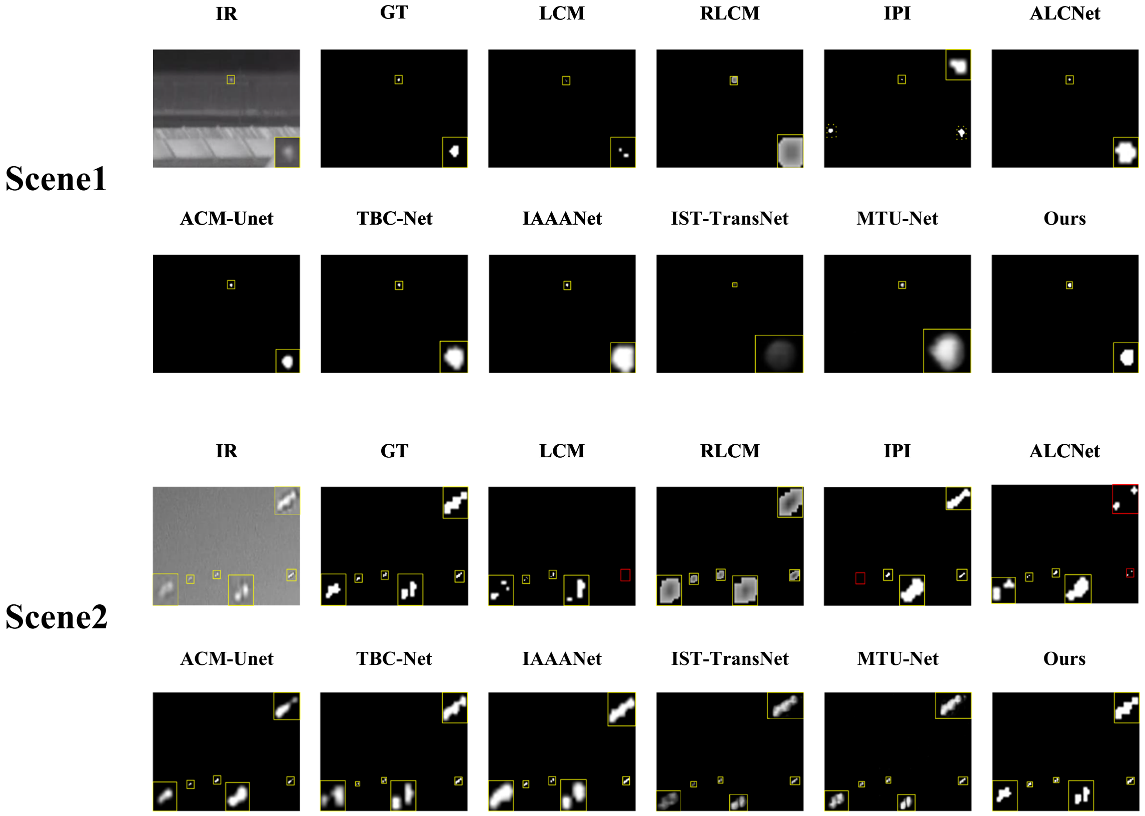


Fig. 12 Detection results of different detection methods on the NUAA-SIRST Datasets. In the figure, the location of the small target is zoomed in, and the solid yellow box indicates "Target Successfully Detected", the dashed yellow box indicates "Target False Alarm", and the red box indicates "Target Missed Detection". The method in this paper does not have false alarms and missed detections, and effectively retains the target features and achieves better detection results

MADFF

In this paper, to set the effectiveness of the dilation convolution and to explore the combination of dilation rates with the best detection performance, the dilation convolution with different dilation rates is included in Baseline for comparison experiments. As shown in Table 3, the best detection performance is achieved when using the combination of side-by-side convolutions with expansion rates of 1, 2, 4, and 8 (containing expansion rates 1, 2, 4, and 8), and set to Baseline-D.

In this paper, to verify the effectiveness of mutual attention deep feature fusion module and compression operation, this paper adds CBAM [64] attention module in Baseline-D and sets it as Method A, replaces CBAM attention module in Method A with Transformer module and sets it as Method B, and adds compression operation in Method B and sets it as Method C.

Table 3 Detection performance of Baseline-D under different dilation rate convolution combinations

Expansion	nIoU↑	F-Measure↑	AUC↑
1	0.5374	0.4717	0.4263
1,2	0.5426	0.4934	0.4696
1,2,4	0.5704	0.5195	0.4783
1,2,4,8	0.5892	0.5347	0.5064

As shown in Table 4, due to the introduction of the mutual attention deep feature fusion module and the use of the Transformer attention mechanism to enhance the understanding of the global regional semantic relationship between the target and the background, Method B improves the nIoU, F-Measure and AUC metrics in the nIoU, F-Measure and AUC metrics by 19.43%, 8.13% and 6.68%, respectively, compared with Method A. As a result of the incorporation of the a priori knowledge-based with the global attention compression mechanism, the nIoU, F-Measure and AUC metrics are improved by 3.01%, 6.69% and 4.59% respectively, and the Params metric is reduced by 40.82%, which indicates that the mutual attention deep feature fusion module proposed in this paper is able to enhance the degree of association of global features at the pixel level under the premise of reducing the number of model parameters, and enhance the network to captures long-range associations between pixels, thus retaining more small target key detail information.

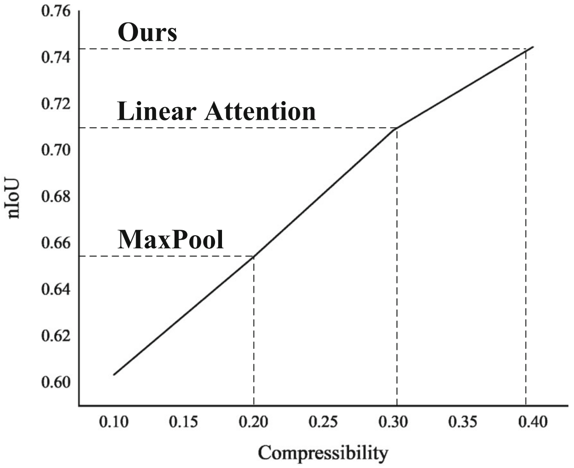


Fig. 13 nIoU for different compression methods at different compression rates

MaxPool compression method can only compress the input feature map to a fixed proportion of the size, and will lose some important but small values of the feature values; Linear Attention [65] compression method in the attention map is too smooth, it is difficult to focus on the local effective features, and at the same time, due to the low rank of the attention map, it restricts the selfattention module output feature diversity. As shown in Fig. 13, compared with MaxPool and Linear Attention, the compression mechanism based on prior knowledge and global attention proposed in this paper is able to customize the setting of the compression ratio and reduce the information loss of the input image features in the compression process; at the same time, in the face of the high computational complexity, the need for a large amount of training data, and the sensitivity to hyper-parameters, the maximum compression ratio is achieved while achieve the best detection effect.

In this paper, to verify the effectiveness of the CLFEF module and to form a comparison method, this paper uses a simple jump connection to fuse the feature maps of the same size in the feature extraction part and the decoder part on the basis of Method C and sets it as Method D; at the same time, this paper adds the context fusion module designed in this paper on the basis of Method C and sets it as Method E. As shown in Table 5, thanks to the context fusion module can dynamically guide the shallow local feature enhancement and fusion according to the deep pre-proposed features, so as to strengthen the characterization ability of the fused features on the position information of small targets, Method D improves 3.17%, 6.72% and 2.87% in the three indexes of nIoU, F-Measure and AUC, respectively, compared with Method C without feature fusion; meanwhile, Method E improves 3.17%, 6.72% and 2.87% in the three indexes of nIoU, F Measure and AUC metrics improved by 3.37%, 4.04% and 1.95%, respectively. The asymmetric local contextual feature fusion module proposed in this paper not only introduces a nonlinear activation method, but also is able to perform dynamic adaptive, context-aware feature refinement of the output features, suppressing irrelevant features and emphasizing relevant features at an early stage in the

Table 4 Results of the ablation experiments of MADFF, ↑ indicates that the larger the value of this indicator, the better the detection performance

Method	Module	nIoU↑	F-Measure↑	AUC↑	Params
Baseline	–	0.5497	0.4982	0.4726	–
Baseline-D	w/DConv	0.5892	0.5347	0.5064	–
Method A	w/CBAM	0.6052	0.6291	0.6616	–
Method B	w/MADFF	0.7228	0.6803	0.7058	9.8 M
Method C	w/EGF	0.7446	0.7258	0.7382	5.9 M

Table 5 Results of the ablation experiments of CLFEF, ↑ indicates that the larger the value of this indicator, the better the detection performance

Method	Module	nIoU ↑	F-Measure ↑	AUC ↑
Method D	w/o CLFEF	0.7682	0.7746	0.7594
Method E	w/ CLFEF	0.7941	0.8059	0.7742

Table 6 Results of the ablation of EFEM, ↑ indicates that the larger the value of the index, the better the detection performance

Method	Module	nIoU ↑	F-Measure ↑	AUC ↑
Method F	w/o EFEM	0.8011	0.8187	0.7984
Method G	w/ EFEM	0.8125	0.8302	0.8068

low-layer network, which makes the network able to encode high-level semantics more effectively.

EFEM

In this paper, to verify the EFEM and to form a comparative method, this paper adds the Sobel edge feature extraction operation on shallow image features on the basis of Method E and sets it as Method F. At the same time, the edge feature module is added on the basis of Method E and sets it as Method G.

The Sobel and Laplace operators in the EFEM are able to obtain primary edge features, the distance transform function is able to suppress the background noise, and the spatio-temporal attention mechanism is able to obtain the dependency of rough small targets on the whole image, which helps the network to supervise the training of the subsequent feature maps. As shown in Table 6, the edge feature module with the addition of single-layer Sobel operation only improves 0.88%, 1.59%, and 1.05% in the three metrics of nIoU, F-Measure, and AUC compared to Method E. However, Method G with the addition of the edge feature module improves 2.31%, 3.01%, and 3.01% in the three metrics of nIoU, F-Measure, and AUC compared to Method E. The edge feature module is a good example of a method that can be used to improve the performance of the network in the three metrics of the nIoU and AUC. 2.31%, 3.01% and 4.21%. The edge feature module proposed in this paper can effectively solve the loss and blurring of key information in the convolution process of small targets during the training process of the network due to the scarcity of features, which leads to inaccurate predicted location and structure information. As shown in Fig. 14, Method G with the added edge feature module can continuously optimize the edge and shape features of the small targets as the training process proceeds.

FEM

In this paper, to verify the effectiveness of FEM and to form a comparison method, this paper uses the structure of the feature extraction module but removes the BlurPool filter on the basis of Method G. MaxPool2D is still used as the downsampling layer and is set up as Method I. At the same time, the BlurPool filter is used as the downsampling layer on the basis of Method I.

Due to the streamlining of cascade jump connections in the feature extraction module and the use of dilation convolution to expand the network sensory field, multi-scale contextual features are extracted, as shown in Table 7, method I improves 3.06%, 1.39% and 2.41% in the three metrics of nIoU, F-Measure and AUC, respectively, compared to method G. Meanwhile, due to the incorporation of BlurPool filter, it improves network Meanwhile, due to the addition of the BlurPool filter, which improves the network translation isotropy and reduces the loss of sparse features, method J improves another 1.80%, 2.03% and 7.90% compared with method I, respectively. The feature extraction module proposed in this paper is able to gradually extract different levels of feature information while reducing the target diagnostic bias and retaining more effective and critical small target features.

To further demonstrate the effectiveness of adding BlurPool, as shown in Fig. 15, the BlurPool-based FEM designed in this paper can effectively reduce the loss of target features,b and attenuate the effect of feature offset (the red dots indicate the center-of-mass position of the target). Meanwhile, to overcome the blurring problem, this paper adds a multilevel cascade operation to the FEM, which can compensate for the lost high-frequency target features and eliminate the effect of low-pass filtering.

Hyper-parameter validation

To further prove the effectiveness of the method designed in this paper, the validity of the important hyperparameters in the method is verified in this paper, including the number of downsampling, the dilated convolution in MADFF, α_1 and α_2 in the loss function.

To verify the effect of downsampling times (number of FEM modules) on the detection effect in this paper's method, this paper evaluates the downsampling from 1 to 5 times respectively. The detection effect of different downsampling times in this paper is shown in Fig. 16a, as the number of downsampling times increases, deeper semantic information can be obtained, which helps to establish the global connection between the small target and the background and filter out most of the clutter noise. However, when the downsampling times are 4 and 5, the minimum feature map size in the network is only 32×32, and the small target size in the

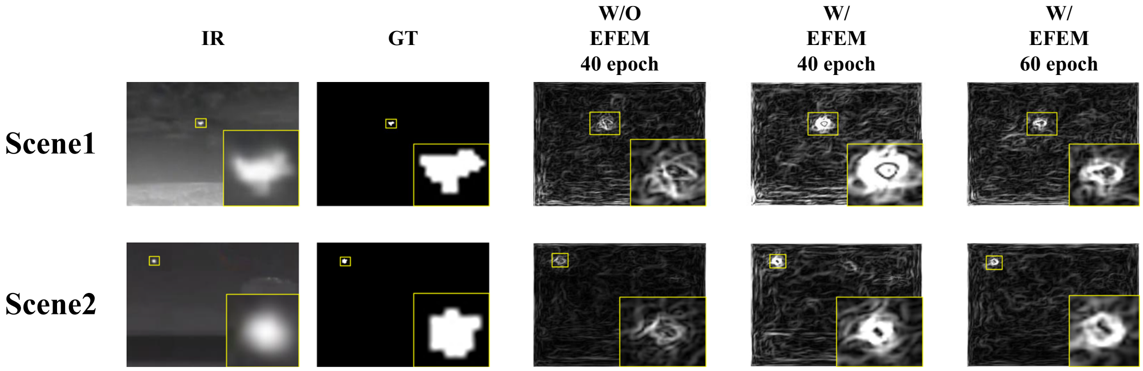


Fig. 14 Ablation experiments of EFEM

Table 7 Results of ablation experiments for the FEM, ↑ indicates that the larger the value of this metric, the better the detection performance

Method	Module	nIoU ↑	F-Measure ↑	AUC ↑
Method I	w/o BlurPool	0.8374	0.8417	0.8263
Method J	w/ BlurPool	0.8525	0.8588	0.8916

original image is smaller than 32×32, which will lead to the loss of target edges and detail information.

Different from the ablation experiments, to further verify the influence of different dilatation rate combinations on the MADFF effect, this paper sets the dilatation rate combinations of [1], [1,2,4] and [1,2,4,8]. The detection effect of different dilution rate combinations is shown in Fig. 16b. The [1,2,4,8] dilution convolution combination used in this paper can effectively capture multi-scale information, combine different dilution rates to extract wider contextual information within the perceptual field, and reduce the interference of small target size change and complex background.

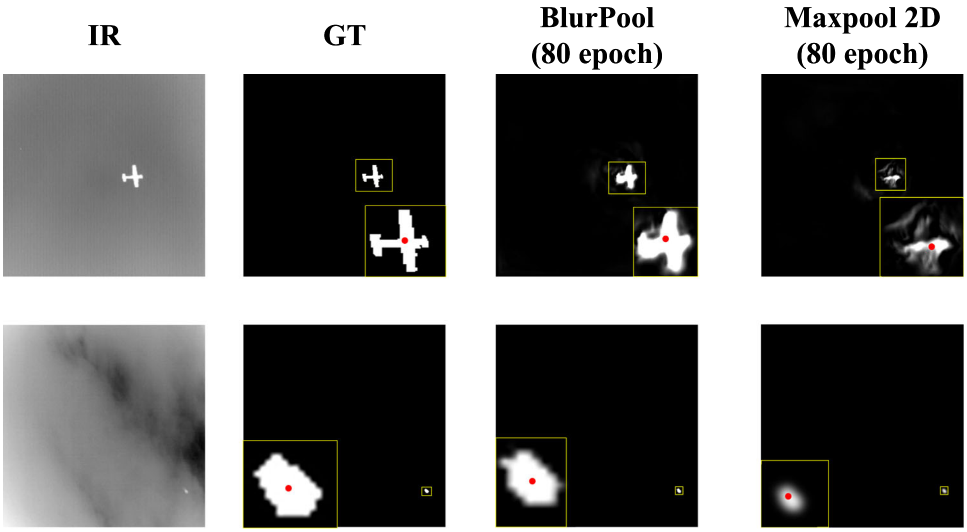
To verify the influence of the weighting coefficients of edge loss and detection result loss in the loss function on the detection effect, this paper has done several sets of comparison experiments on different combinations of weighting coefficients. As shown in Fig. 16c, the shallow edge features can be more fully utilized when $\alpha_1 = 0.3$ and $\alpha_2 = 0.7.$, and the shape and position information of small targets can be better recovered.

Limitations

The proposed method in this paper achieves good results on IRSTD-1k Datasets and NUAA-SIRST Datasets, but the detection method is still problematic for practical applications.

The processing speed is still close to 0.1 s per image, and the main reason for this problem is that the self-attentive mechanism needs to introduce additional parameters to learn the correlation between each location and other locations, which leads to a significant increase in computational com-

Fig. 15 The effect of BlurPool on feature offset and feature loss. The method in this paper can effectively prevent thephenomenon of feature offset and retain the important features of the target



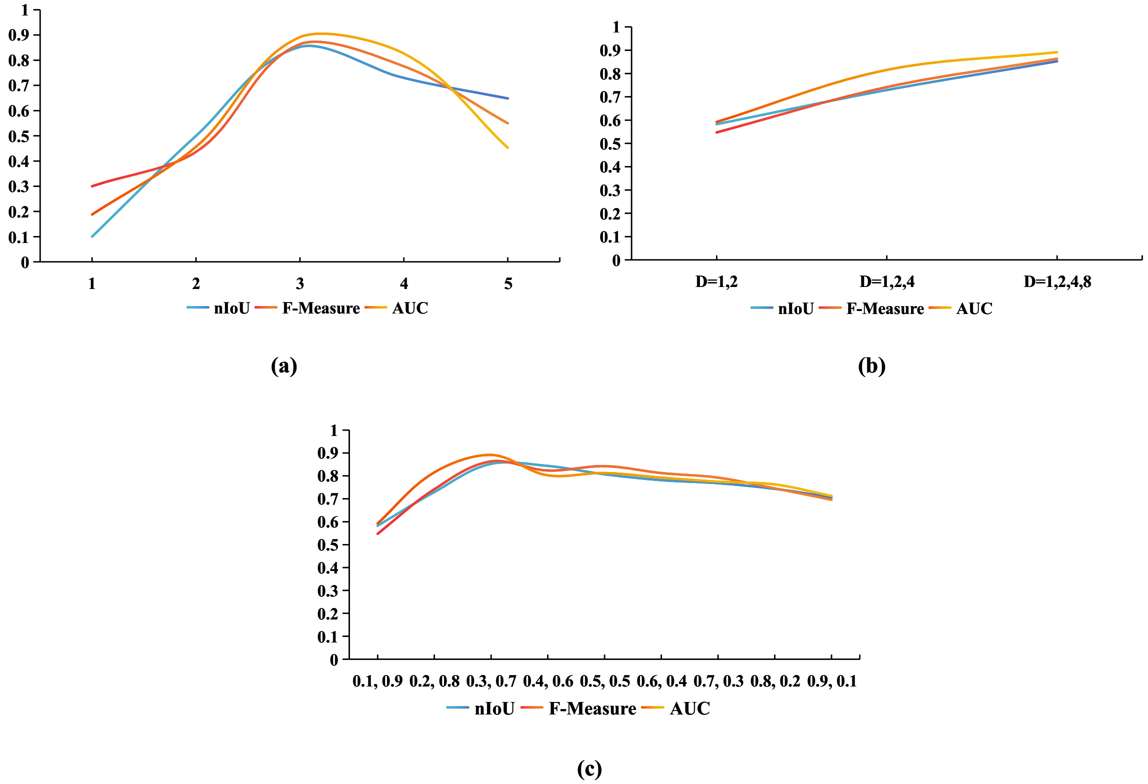


Fig. 16 Detection effect of different hyperparameter choices on IRSTD-1k Datasets. Among the combinations of multiple hyperparameters, the above three sets of hyperparameters set in this paper achieved the best detection results

plexity. Meanwhile, the number of publicly available infrared small target detection datasets is limited due to institutional constraints in the security domain. This scarcity of datasets leads to problems such as unbalanced dataset distribution and limited generalization ability during network training. Future studies will be carried out in the above-mentioned issues.

Conclusions

This paper proposes “A Single-Frame Infrared Small Target Detection Method based on Joint Feature Guidance”, which exhibits excellent detection performance in infrared small target detection.

First, we successfully mitigate the feature offset problem by introducing BlurPool in feature extraction. In addition, the high-frequency target features lost by blur processing are supplemented by multi-level dense connectivity. Then, MADFF, based on the self-attention mechanism, is able to automatically model the long-range relationship between background global features and target localization. Meanwhile, the cosine similarity index and low-dimensional remapping are utilized to improve the computational efficiency. Finally, we introduce CLFEF and EFEM for shallow features, which utilize deep semantic features to guide the enhancement and fusion of spatial information and edge texture features in shallow features, and enhance the ability of adaptive characterization of shape and localization information.

The future work is organized in the following two aspects. First, to construct an infrared small target detection dataset with high image quality, large data volume, and diverse backgrounds, so that the detection method can learn diverse target and background knowledge. Second, explore the feature compression method for the self-attention mechanism, in addition to fully obtaining the dependence relationship between the target and the complex background, to further reduce the number of model parameters and computational speed, to lay the foundation for subsequent practical engineering projects.

Funding This study was supported by the Chongqing Natural Science Foundation (CSTB2022NSCQ-MSX1071).

Data availability The codes used during the study are available from the corresponding author by request.

Declarations

Conflict of interest We declare that we have no financial or personal relationships that can influence our work.

References

1. Deng H, Sun X, Liu M, Ye C, Zhou X (2016) Small infrared target detection based on weighted local difference measure. IEEE Trans Geosci Remote Sens 54(7):4204–4214. https://doi.org/10.1109/TGRS.2016.2538295
2. Teutsch M, Krüger W (2010) Classification of small boats in infrared images for maritime surveillance. In: 2010 international WaterSide security conference. pp 1–7. https://doi.org/10.1109/WSSC.2010.5730289
3. Deshpande SD, Er MH, Venkateswarlu R, Chan P (1999) Max-mean and max-median filters for detection of small targets. In: Signal and data processing of small targets 1999, vol 3809. SPIE, pp 74–83
4. Arce G, McLoughlin M (1987) Theoretical analysis of the max/median filter. IEEE Trans Acoust Speech Signal Process 35(1):60–69. https://doi.org/10.1109/TASSP.1987.1165036
5. Kim S (2011) Min-local-log filter for detecting small targets in cluttered background. Electron Lett 47(2):1
6. Kim S, Yang Y, Lee J, Park Y (2009) Small target detection utilizing robust methods of the human visual system for irst. J Infrared Millim Terahertz Waves 30(9):994–1011. https://doi.org/10.1007/s10762-009-9518-2
7. Shao X, Fan H, Lu G, Xu J (2012) An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system. Infrared Phys Technol 55(5):403–408. https://doi.org/10.1016/j.infrared.2012.06.001
8. Dai Y, Wu Y, Song Y, Guo J (2017) Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. Infrared Phys Technol 81:182–194. https://doi.org/10.1016/j.infrared.2017.01.009
9. Zhang T, Wu H, Liu Y, Peng L, Yang C, Peng Z (2019) Infrared small target detection based on non-convex optimization with lp-norm constraint. Remote Sens. https://doi.org/10.3390/rs11050559
10. Zhang L, Peng L, Zhang T, Cao S, Peng Z (2018) Infrared small target detection via non-convex rank approximation minimization joint l2,1 norm. Remote Sens. https://doi.org/10.3390/rs10111821
11. Song X, Wu N, Song S, Stojanovic V (2023) Switching-like event-triggered state estimation for reaction–diffusion neural networks against dos attacks. Neural Process Lett 55(7):8997–9018. https://doi.org/10.1007/s11063-023-11189-1
12. Peng Z, Song X, Song S, Stojanovic V (2023) Hysteresis quantified control for switched reaction–diffusion systems and its application. Complex Intell Syst 9(6):7451–7460. https://doi.org/10.1007/s40747-023-01135-y
13. Zhang Z, Song X, Sun X, Stojanovic V (2023) Hybrid-driven-based fuzzy secure filtering for nonlinear parabolic partial differential equation systems with cyber attacks. Int J Adapt Control Signal Process 37(2):380–398
14. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 779–788
15. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 7263–7271
16. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. CoRR. arXiv:1804.02767
17. Bochkovskiy A, Wang C, Liao HM (2020) Yolov4: optimal speed and accuracy of object detection. CoRR arXiv:2004.10934
18. Wang C-Y, Bochkovskiy A, Liao H-YM (2023) 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. pp 7464–7475
19. Hussain M (2023) Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. Machines. https://doi.org/10.3390/machines11070677
20. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention-MICCAI 2015. Springer, Cham, pp 234–241
21. Li C, Huang Z, Xie X, Li W (2023) Ist-transnet: infrared small target detection based on transformer network. Infrared Phys Technol 132:104723. https://doi.org/10.1016/j.infrared.2023.104723
22. Wang K, Du S, Liu C, Cao Z (2022) Interior attention-aware network for infrared small target detection. IEEE Trans Geosci Remote Sens 60:1–13. https://doi.org/10.1109/TGRS.2022.3163410
23. Li B, Xiao C, Wang L, Wang Y, Lin Z, Li M, An W, Guo Y (2023) Dense nested attention network for infrared small target detection. IEEE Trans Image Process 32:1745–1758
24. Wu X, Hong D, Chanussot J (2023) Uiu-net: U-net in u-net for infrared small object detection. IEEE Trans Image Process 32:364–376. https://doi.org/10.1109/TIP.2022.3228497
25. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Sixth international conference on computer vision (IEEE Cat. No.98CH36271). pp 839–846. https://doi.org/10.1109/ICCV.1998.710815
26. Bae T-W, Sohng K-I (2010) Small target detection using bilateral filter based on edge component. J Infrared Millim Terahertz waves 31:735–743
27. Bai X, Zhou F (2010) Analysis of new top-hat transformation and the application for infrared dim small target detection. Pattern Recognit 43(6):2145–2156. https://doi.org/10.1016/j.patcog.2009.12.023
28. Chen CLP, Li H, Wei Y, Xia T, Tang YY (2014) A local contrast method for small infrared target detection. IEEE Trans Geosci Remote Sens 52(1):574–581
29. Han J, Liang K, Zhou B, Zhu X, Zhao J, Zhao L (2018) Infrared small target detection utilizing the multiscale relative local contrast measure. IEEE Geosci Remote Sens Lett 15(4):612–616
30. Xia C, Li X, Zhao L, Shu R (2020) Infrared small target detection based on multiscale local contrast measure using local energy factor. IEEE Geosci Remote Sens Lett 17(1):157–161
31. Gao C, Meng D, Yang Y, Wang Y, Zhou X, Hauptmann AG (2013) Infrared patch-image model for small target detection in a single image. IEEE Trans Image Process 22(12):4996–5009
32. Dai Y, Wu Y, Song Y (2016) Infrared small target and background separation via column-wise weighted robust principal component analysis. Infrared Phys Technol 77:421–430. https://doi.org/10.1016/j.infrared.2016.06.021
33. Wang, H., Zhou, L., Wang, L.: 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. pp 8509–8518 (2019)

34. Zhao M, Cheng L, Yang X, Feng P, Liu L, Wu N (2020) Tbc-net: a real-time detector for infrared small target detection using semantic constraint. CoRR arXiv:2001.05852
35. Ju M, Luo J, Liu G, Luo H (2021) Istdet: an efficient end-to-end neural network for infrared small target detection. Infrared Phys Technol 114:103659. https://doi.org/10.1016/j.infrared.2021.103659
36. Dai Y, Wu Y, Zhou F, Barnard K (2021) Asymmetric contextual modulation for infrared small target detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 950–959
37. Wang Y, Tian Y, Liu J, Xu Y (2023) Multi-stage multi-scale local feature fusion for infrared small target detection. Remote Sens. https://doi.org/10.3390/rs15184506
38. Zhang T, Li L, Cao S, Pu T, Peng Z (2023) Attention-guided pyramid context networks for detecting infrared small target under complex background. IEEE Trans Aerosp Electron Syst 59(4):4250–4261. https://doi.org/10.1109/TAES.2023.3238703
39. Tong X, Sun B, Wei J, Zuo Z, Su S (2021) Eaau-net: enhanced asymmetric attention u-net for infrared small target detection. Remote Sens. https://doi.org/10.3390/rs13163200
40. Ali A, Abdelhafeez A (2022) Deephar-net: a novel machine intelligence approach for human activity recognition from inertial sensors. Sustain Mach Intell J. https://doi.org/10.61185/SMIJ.2022.8463
41. Abdel-Monem A, Abouhawwash M (2022) A machine learning solution for securing the internet of things infrastructures. Sustain Mach Intell J. https://doi.org/10.61185/SMIJ.HPAO9103
42. Abdelhafeez A, Aziz A, Khalil N (2022) Building a sustainable social feedback loop: a machine intelligence approach for twitter opinion mining. Sustain Mach Intell J. https://doi.org/10.61185/SMIJ.2022.2315
43. Bacanin N, Stoean R, Zivkovic M, Petrovic A, Rashid TA, Bezdan T (2021) Performance of a novel chaotic firefly algorithm with enhanced exploration for tackling global optimization problems: application for dropout regularization. Mathematics. https://doi.org/10.3390/math9212705
44. Malakar S, Ghosh M, Bhowmik S, Sarkar R, Nasipuri M (2020) A GA based hierarchical feature selection approach for handwritten word recognition. Neural Comput Appl 32(7):2533–2552. https://doi.org/10.1007/s00521-018-3937-8
45. Bacanin N, Zivkovic M, Al-Turjman F, Venkatachalam K, Trojovský P, Strumberger I, Bezdan T (2022) Hybridized sine cosine algorithm with convolutional neural networks dropout regularization application. Sci Rep 12(1):6302. https://doi.org/10.1038/s41598-022-09744-2
46. Zivkovic M, Bacanin N, Antonijevic M, Nikolic B, Kvascev G, Marjanovic M, Savanovic N (2022) Hybrid cnn and xgboost model tuned by modified arithmetic optimization algorithm for COVID-19 early diagnostics from x-ray images. Electronics. https://doi.org/10.3390/electronics11223798
47. Zhang R (2019) Making convolutional networks shift-invariant again. In: International conference on machine learning. PMLR, pp 7324–7334
48. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (eds) Computer vision-ECCV 2020. Springer, Cham, pp 213–229
49. Wu T, Li B, Luo Y, Wang Y, Xiao C, Liu T, Yang J, An W, Guo Y (2023) Mtu-net: multilevel transunet for space-based infrared tiny ship detection. IEEE Trans Geosci Remote Sens 61:1–15. https://doi.org/10.1109/TGRS.2023.3235002
50. Choromanski K, Likhosherstov V, Dohan D, Song X, Gane A, Sarlós T, Hawkins P, Davis J, Mohiuddin A, Kaiser L, Belanger D, Colwell LJ, Weller A (2020) Rethinking attention with performers. CoRR arXiv:2009.14794
51. Shen Z, Zhang M, Zhao H, Yi S, Li H (2021) Efficient attention: attention with linear complexities. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 3531–3539
52. Park J, Woo S, Lee J, Kweon IS (2018) BAM: bottleneck attention module. CoRR arXiv:1807.06514
53. Maini D, Aggarwal AK (2018) Camera position estimation using 2d image dataset. Int J Innov Eng Technol 10:199–203
54. Brar DS, Aggarwal AK, Nanda V, Saxena S, Gautam S (2024) Ai and cv based 2d-cnn algorithm: botanical authentication of Indian honey. Sustain Food Technol, Royal Society of Chemistry, 2:373-385. https://doi.org/10.1039/D3FB00170A
55. Aggarwal AK A review on genomics data analysis using machine learning. WSEAS TRANSACTIONS ON BIOLOGY AND BIOMEDICINE 20:119-131. https://doi.org/10.37394/23208.2023.20.12
56. Justusson BI (1981) Median filtering: statistical properties. Springer, Berlin, pp 161–196. https://doi.org/10.1007/BFb0057597
57. Ma J, Wei Z, Zhang Y, Wang Y, Lv R, Zhu C, Gaoxiang C, Liu J, Peng C, Wang L et al (2020) How distance transform maps boost segmentation cnns: an empirical study. In: Medical imaging with deep learning. PMLR, pp 479–492
58. Zhang M, Zhang R, Yang Y, Bai H, Zhang J, Guo J (2022) Isnet: shape matters for infrared small target detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 877–886
59. Schober P, Mascha EJ, Vetter TR (2021) Statistics from a (agreement) to z (z score): a guide to interpreting common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research. Anesth Analg 133(6):1633–1641
60. Gao C, Meng D, Yang Y, Wang Y, Zhou X, Hauptmann AG (2013) Infrared patch-image model for small target detection in a single image. IEEE Trans Image Process 22(12):4996–5009. https://doi.org/10.1109/TIP.2013.2281420
61. Dai Y, Wu Y (2017) Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. IEEE J Sel Top Appl Earth Obs Remote Sens 10(8):3752–3767. https://doi.org/10.1109/JSTARS.2017.2700023
62. Dai Y, Wu Y, Zhou F, Barnard K (2021) Attentional local contrast networks for infrared small target detection. IEEE Trans Geosci Remote Sens 59(11):9813–9824. https://doi.org/10.1109/TGRS.2020.3044958
63. Dai Y, Wu Y (2017) Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. IEEE J Sel Top Appl Earth Obs Remote Sens 10(8):3752–3767. https://doi.org/10.1109/JSTARS.2017.2700023
64. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp 3–19
65. Li R, Su J, Duan C, Zheng S (2020) Linear attention mechanism: an efficient attention for semantic segmentation. CoRR arXiv:2007.14902

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.