Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064864

ARTICLE

# Attention Shift-Invariant Cross-Evolutionary Feature Fusion Network for Infrared Small Target Detection

**Siqi Zhang and Shengda Pan***

School of Information and Engineering, Shanghai Maritime University, Shanghai, 201306, China

*Corresponding Author: Shengda Pan. Email: sdpan@shmtu.edu.cn

Received: 25 February 2025; Accepted: 21 May 2025; Published: 30 July 2025

**ABSTRACT:** Infrared images typically exhibit diverse backgrounds, each potentially containing noise and target-like interference elements. In complex backgrounds, infrared small targets are prone to be submerged by background noise due to their low pixel proportion and limited available features, leading to detection failure. To address this problem, this paper proposes an Attention Shift-Invariant Cross-Evolutionary Feature Fusion Network (ASCFNet) tailored for the detection of infrared weak and small targets. The network architecture first designs a Multidimensional Lightweight Pixel-level Attention Module (MLPA), which alleviates the issue of small-target feature suppression during deep network propagation by combining channel reshaping, multi-scale parallel subnet architectures, and local cross-channel interactions. Then, a Multidimensional Shift-Invariant Recall Module (MSIR) is designed to ensure the network remains unaffected by minor input perturbations when processing infrared images, through focusing on the model's shift invariance. Subsequently, a Cross-Evolutionary Feature Fusion structure (CEFF) is designed to allow flexible and efficient integration of multidimensional feature information from different network hierarchies, thereby achieving complementarity and enhancement among features. Experimental results on three public datasets, SIRST, NUDT-SIRST, and IRST640, demonstrate that our proposed network outperforms advanced algorithms in the field. Specifically, on the NUDT-SIRST dataset, the mAP50, mAP50-95, and $F1_{score}$ metrics reached 99.26%, 85.22%, and 99.31%, respectively. Visual evaluations of detection results in diverse scenarios indicate that our algorithm exhibits an increased detection rate and reduced false alarm rate. Our method balances accuracy and real-time performance, and achieves efficient and stable detection of infrared weak and small targets.

**KEYWORDS:** Deep learning; infrared small target detection; complex scenes; feature fusion; convolution pooling

## 1 Introduction

Infrared small target detection is an important research direction in remote sensing and infrared imaging technology. Due to its strong resistance to smoke interference, long detection range, and the advantage of being unaffected by lighting conditions, enabling target detection at night or under adverse weather conditions, it has broad application prospects in military reconnaissance, security monitoring, astronomical observation, night vision systems, and other fields [1].

Traditional model-driven methods primarily rely on the analysis of target physical and imaging characteristics, as well as reasonable assumptions based on prior knowledge. However, due to the small size, low signal-to-noise ratio, and lack of distinct texture and shape features of infrared small targets, these traditional methods often exhibit low detection accuracy and poor robustness in complex and variable infrared scenes [2]. By stacking multiple layers of nonlinear transformations, CNNs can not only learn deep-level feature representations of data, better capturing the essential characteristics of targets, but also greatly reduce the subjectivity and limitations of feature design without human intervention. However, there are still some issues with current deep learning-based infrared small target detection algorithms.

This article delves deeply into the challenges of detecting infrared small and dim targets in complex scenes. It addresses three core challenges in the field of infrared small target detection: Firstly, the features of small targets are easily diluted due to the stacking of information layers in deep networks, thereby reducing detection accuracy. Secondly, complex environments such as background clutter, illumination fluctuations, and detector noise severely interfere with small target detection. Thirdly, the difficulty in fully utilizing information at various levels leads to inefficiencies and missed detections. To address these challenges, a new network architecture, ASCFNet (Attention Shift-Invariant Cross-Evolutionary Feature Fusion Network), is proposed. This network integrates traditional prior knowledge with advanced technologies in the field of deep learning, constructing a unique attention shift-invariant cross-evolutionary feature fusion framework.

The main contributions of this article are as follows:

(1) Design a multidimensional lightweight pixel-level attention module that combines channel reshaping, a multi-scale parallel subnetwork architecture, and local cross-channel interactions. This enables the network to focus on key areas in the image while effectively filtering out irrelevant background interference, significantly alleviating the issue of small target features being overwhelmed during deep network propagation.

(2) Design a multidimensional shift-invariant recall module that integrates relevant prior knowledge of targets and backgrounds into the convolutional neural network, endowing the model with unique shift-invariant characteristics. This ensures that the network can accurately capture the feature information of targets, regardless of changes in target position, pose, or scale, when processing infrared images.

(3) The designed cross-evolutionary feature fusion strategy is different from traditional sequential processing, allows the network to integrate multidimensional feature information from different network levels in a more flexible and efficient manner. This not only achieves complementarity and enhancement among features, but also greatly enhances the representational ability and robustness of the features.

## 2 Related Work

In recent decades, conventional techniques relying on manually crafted features for target identification have dominated algorithms for infrared small target detection. According to the detection object, these methods can be broadly categorized into single-frame detection and multi-frame detection [3]. This article mainly focuses on single frame detection. The mainstream traditional single-frame algorithms include the transform domain method, such as the early spatial filtering methods Top-hat filter [4]. Their advantage lies in their straightforward principle and easy implementation, but their overly simplistic assumptions make them unsuitable for complex scenes. Gradually, these transform domain methods have been developed to Fourier domain [5], fuzzy space [6], and gradient vector field [7]. These transform domain methods typically treat infrared small targets as high-frequency components in images, but the actual target background complexity often far exceeds ideal assumptions, making them difficult to apply. Inspired by the human visual system, significant detection methods focusing on the energy distribution mechanism of infrared weak and small targets have also been proposed. The mainstream local contrast methods, such as difference-based LoG and Dog filters [8], ratio-based LCM [9], ILCM [10], MPCM [11], and the combined difference-ratio RLCM [12], enhance target visibility by assessing the contrast between the target region and its surrounding background. While these approaches offer some improvement in detection accuracy, they still struggle to cope with extremely complex infrared scenes. The low-rank sparse decomposition method based on image data structure is also one of the commonly used methods for infrared small target detection, such as IPI [13], ADMD [14], etc. These methods assume that targets are sparse and backgrounds are low-rank. Although they exhibit good background suppression, their mathematical definitions are strict, making targeted improvements difficult. Moreover, when the background is complex, edges, corners, and noise points also exhibit sparsity. To finely characterize the mathematical relationships among them, the algorithm often has high computational complexity, which limits practical applications.

While these approaches may lead to some improvement in detection accuracy, traditional methods rely heavily on manually designed features, making it difficult to comprehensively cover all possible target features. When facing complex and ever-changing infrared scenes, feature loss or false detection phenomena are prone to occur.

Over the recent years, an increasing number of scholars have devoted themselves to applying deep learning algorithms to the field of infrared small target detection. For instance, Dai et al. initially proposed a bidirectional asymmetric attention modulation network, which utilizes top-down global attention and bottom-up local attention to ensure the full integration of infrared small targets in the feature pyramid [15]. Furthermore, based on this, they introduced the attention local contrast network, decoupling the image block size and filter size in traditional methods, replacing them with dilation factors and effective receptive fields, thus achieving the seamless integration of prior knowledge with deep learning techniques. However, due to the use of local contrast mechanisms to constrain the network model in ALCNet [16], its generalization performance in complex background scenes is limited. Li et al. proposed DNANet, which preserves the deep representation of targets and achieves adaptive enhancement of features by introducing the dense nested interaction module (DNIM) and the cascaded channel spatial attention module (CSAM) [17]. However, precisely because it frequently uses upsampling and downsampling operations in the model to ensure multiple interactions, the model becomes cumbersome, computationally complex, and has poor real-time performance. Zhang et al. studied the contextual relationships and feature utilization in the transmission of infrared small target detection networks, proposing an attention-guided pyramid contextual network that improves detection accuracy through the fusion of shallow and deep features [18].

However, despite significant progress in infrared small target detection using deep learning methods, there are still some issues and limitations. Firstly, in neural network learning, as the model parameters increase, the expression ability of the model will enhance and the amount of information it can store will also increase. However, this also brings the risk of information overload, especially for infrared small targets that lack color, texture, and other features. They are easily overwhelmed by redundant information in deep networks, resulting in detection failure. Secondly, we observe that modern convolutional networks are not translation invariant, and even small movements in the input can cause drastic changes in the output. Due to the low pixel ratio of small targets and the presence of background clutter similar to the target in complex scenes, it may lead to model detection errors. In addition, the application scenarios of infrared small target detection algorithms often require the algorithm to have extremely high stability and real-time performance. How to fully and efficiently utilize the features of each layer in the network, so that the model can maintain stable detection performance in different scenarios while reducing computational complexity, is also an urgent problem to be solved.

# 3 Proposed Methodology

## *3.1 Overall Architecture*

The overall network structure we designed is shown in Fig. 1, which can be divided into three main parts: the backbone network, the neck network, and the detection head. The input image first passes through the backbone network, which comprises five layers and incorporates four downsampling operations. In the second, third, and fourth layers of the backbone network, each downsampling step is followed by an MLPA module. This module significantly enhances feature representation by increasing the attention on critical pixel regions.

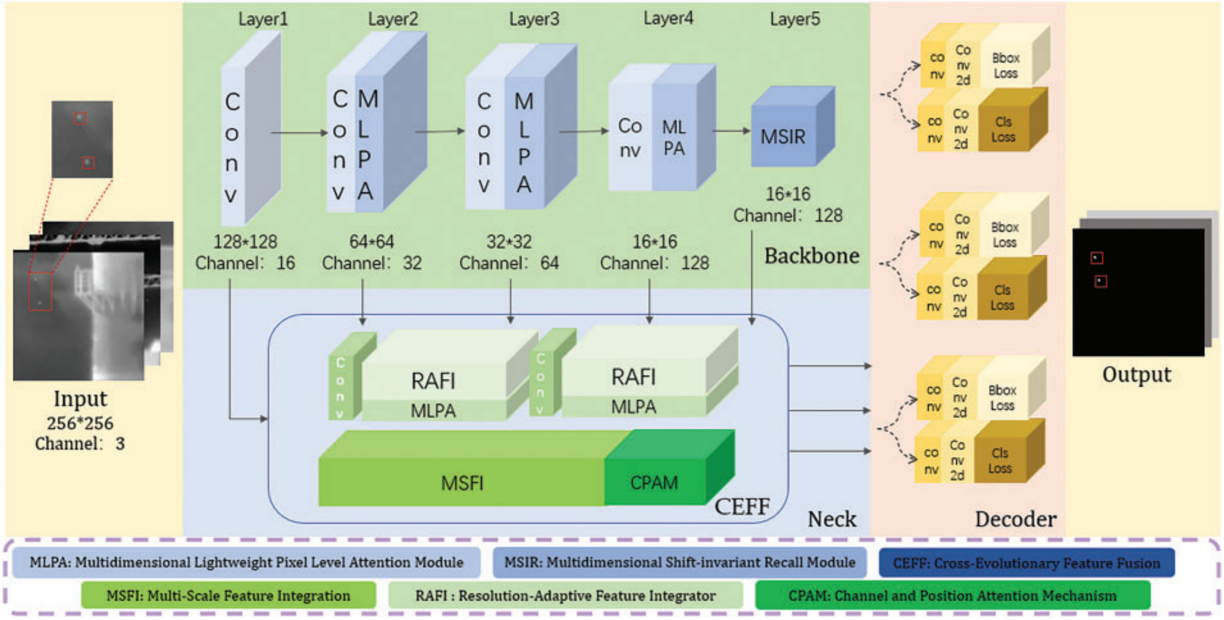

Layer1
Layer2
Layer3
Layer4
Layer5
Conv
Conv MLPA
Conv MLPA
Conv MLPA
MSIR
128*128 Channel: 16
64*64 Channel: 32
32*32 Channel: 64
16*16 Channel: 128
16*16 Channel: 128
Backbone
Input
256*256
Channel: 3
Conv
RAFI
MLPA
Conv
RAFI
MLPA
MSFI
CPAM
CEFF
Neck
conv
Conv2d
Bbox Loss
conv
Conv2d
Cls Loss
Decoder
Output
MLPA: Multidimensional Lightweight Pixel Level Attention Module
MSIR: Multidimensional Shift-invariant Recall Module
CEFF: Cross-Evolutionary Feature Fusion
MSFI: Multi-Scale Feature Integration
RAFI : Resolution-Adaptive Feature Integrator
CPAM: Channel and Position Attention Mechanism

**Figure 1:** The overall network structure of ASCFNet

The fifth layer contains the MSIR module, which does not perform any downsampling operations. Instead, it combines a full-dimensional dynamic convolution kernel and anti-aliasing pooling, enabling the extraction of multidimensional information from the convolution kernel space while maintaining shift invariance. This ensures the robustness of feature extraction.

The feature maps processed by the backbone network are then transmitted to the neck network. This network employs a cross-evolutionary feature fusion structure comprising three components: MSFI for hierarchical feature extraction and selection, RAFI for multi-scale fusion, and CPAM for key feature activation. Acting as a crucial link between the backbone and detection head, the core function of CEFF is to facilitate the comprehensive and efficient fusion of diverse features extracted by the backbone network.

Finally, the fused features from the neck network are selectively input into the detection head, which adopts an anchor-free decoupling structure. This structure directly predicts the target's position and category, simplifying the detection process. The detection head ultimately outputs the detection results, including the accurate location and category information of the target, achieving precise recognition and localization of objects in the input image.

### *3.2 Multidimensional Lightweight Pixel Level Attention Module*

In order to solve the feature dilution problem mentioned earlier, where critical features are dispersed across network layers, reducing their impact on the final prediction, we have designed a multidimensional lightweight attention module suitable for infrared weak target detection. The primary objective of this module is to minimize computational costs while emphasizing pixel-level pairwise relationships. This enables the efficient extraction of multi-dimensional features—such as channel, spatial, global, and local dimensions—of infrared weak targets in the initial phases of the network.

The overall module architecture of MLPA is shown in Fig. 2. Within this architecture, the features are reshaped from the channel dimension to the batch dimension and further subdivided into multiple sub-feature groups. This approach maximizes the preservation of inter-channel information while maintaining computational efficiency, ensuring a balanced distribution of spatial semantic features. To avoid significant semantic differences in features, we gradually shuffle these channels between each other and reset their weights. Each subnetwork facilitates local cross-channel interactions without compromising the channel dimension. Finally, all processed channels are cross fused again and their weights are reassigned, and then output as a whole.

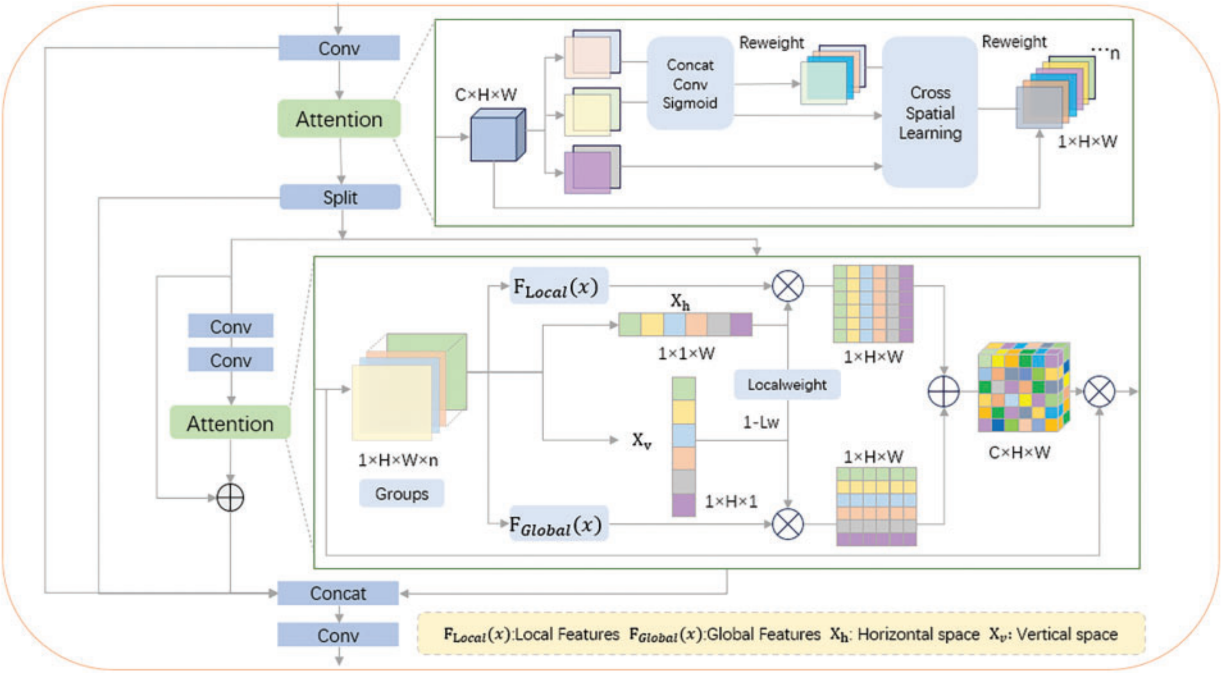


**Figure 2:** MLPA module architecture diagram

In the second green box, the features are processed separately by two parallel sub-networks—one focusing on global features and the other on local features. By fusing the feature maps output by these sub-networks, the channel features of the target are projected into the spatial domain, facilitating comprehensive multidimensional information acquisition and enhancing the model's perceptual capabilities. A complementary variable dynamically adjusts the weight distribution between local and global features to ensure balanced integration.

In order to capture richer multi-scale spatial information within the same processing stage, we flexibly used convolution kernels of different sizes to highlight the target area in the entire image background.

Specifically, in the 3 × 3 branch within the local cross channel interaction in the first green box, we capture multi-scale features by stacking multiple 3 × 3 kernels, which preserves the encoding ability between channels while ensuring accurate transmission of spatial structural information. In the 1 × 1 branch within the second green box, we employed 1D global average pooling in two directions, allowing the module to comprehensively capture global and local information in both horizontal and vertical spatial dimensions of $X_h$ and $X_v$ to recalibrate channel weights. Specifically, the input features are endowed with the ability to perform final encoding correction on the processed features, ensuring that the final output of the model comprehensively considers the outputs of each stage and effectively avoids excessive bias.

Overall, the MLPA module we designed provides comprehensive encoding of multidimensional information. It effectively suppresses the surge in computational complexity while ensuring seamless detection and clear feature presentation of infrared small targets. Even with deeper network layers, this module maintains sufficient interactions between global information and local features, including those from adjacent regions. Furthermore, the attention module integrates channel, spatial, local, and global information with high flexibility and scalability.

### *3.3 Multidimensional Shift-Invariant Recall Module for Infrared Small Targets*

Compared with conventional medium and large-sized target detection, small target detection is more affected by minor perturbations because small targets have a smaller pixel ratio. As shown in Fig. 3, in the original infrared image on the left, small targets are high-lighted with red boxes. Adjacent to this is a zoomed-in view of the marked small targets after being affected by minor noise, along with the corresponding three-dimensional model. On the right is a schematic diagram illustrating the IoU variations between conventional targets and small targets. It can be observed that minor noise interference can cause the detection box of small targets to shift relative to the actual target. This shift results in a more significant change in the IoU for small targets, leading to fluctuations in the regression process during training, which impacts network optimization. To address this issue, we designed a multi-dimensional shift-invariant recall module. This module incorporates a full-dimensional dynamic convolution kernel and anti-aliasing pooling, enabling the model to capture complementary features across all four dimensions of the convolution kernel space within this layer, without increasing network depth. Moreover, it endows the model with distinctive shift-invariant characteristics.

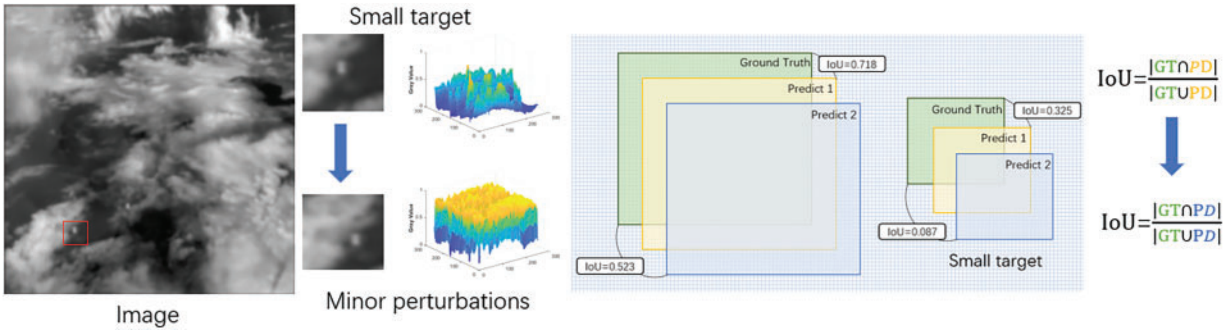


**Figure 3:** The impact of minor perturbations on small targets

The overall structure of MSIR is shown in Fig. 4. The input $x$ consists of $n$ convolutional kernels $W_1$, $W_2$, . . ., $W_n$, which first undergo full-dimensional dynamic convolution to produce $x_t$. The processed $x_t$ is then multiplied by the original input convolution group $x$ to form a new convolution group $x_d$, which is subsequently subjected to anti-aliasing pooling. The anti-aliasing pooling process consists of three steps: 2D abstract projection, dense evaluation, and subsampling. Finally, the output $x_o$ is obtained.

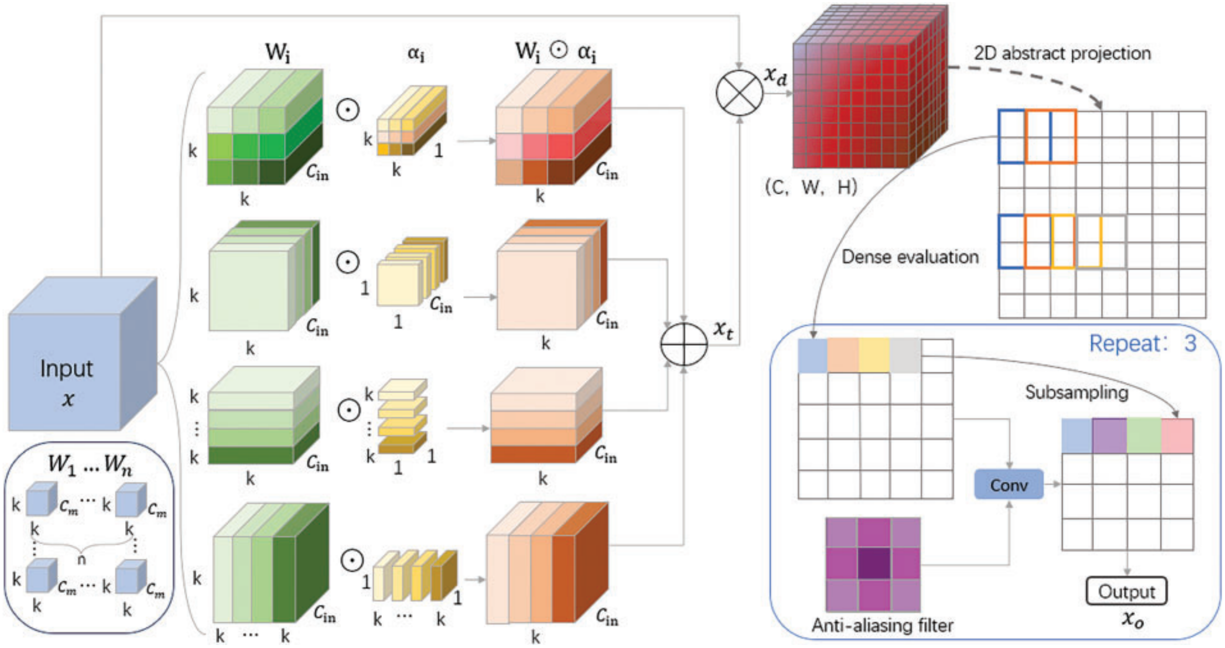


**Figure 4:** MSIR module architecture diagram

For the given n convolutional kernels $W_1 \dots W_n$, the corresponding kernel space has four dimensions: spatial kernel size $k \times k$, the number of input channels $C_{in}$ and output channels $C_{out}$ for each convolutional kernel, and the number of convolutional kernels $n$. The MSIR module we designed integrates features from all four dimensions of the kernel space. In principle, these four aspects of features are complementary. By progressively multiplying the convolutional kernel $W_i$ with dynamic weights $\alpha_i$ following the sequence of position, channel, filter, and kernel dimensions, the convolution operation achieves independent responsiveness across all spatial positions, input channels, filters, and kernel parameters of input $x$. This approach enables more effective acquisition of intricate contextual information.

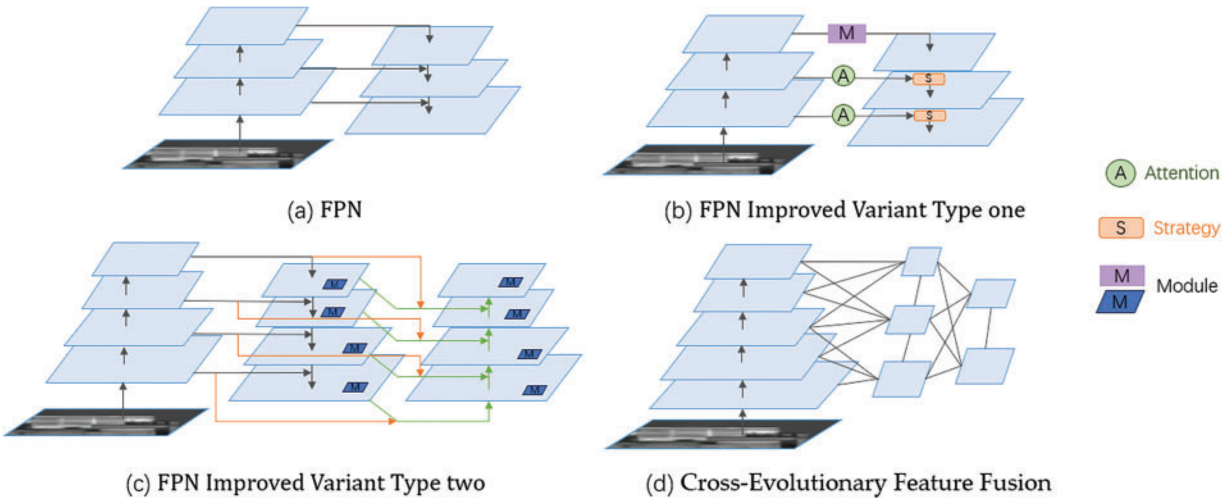
We observe that commonly used downsampling methods in deep learning often disregard the Nyquist sampling theorem. However, infrared small target detection tasks place high demands on image detail preservation. Considering that incorporating prior knowledge can help ensure prediction results conform to fundamental physical mechanisms and common sense, and inspired by the analysis of downsampling theory, we turn our attention to pooling operations, which are widely used in modern convolutional networks.

Max pooling, as an example, consists of two key steps: applying the maximum operator over a sliding window densely and naive downsampling. While the maximum operation performs exhaustive evaluation across a sliding window, preserving shift equivariance, the subsequent downsampling step fails to maintain this property. To address this limitation, we draw inspiration from anti-aliasing techniques in signal processing. First, $x_d$ is subjected to a 2D abstract projection, and then an anti-aliasing filter with a $3 \times 3$ kernel is applied after the maximum operator to mitigate aliasing effects and enhance the model's robustness to input shifts. Note that this is not simply inserting a low-pass filtering module into the network, but rather integrating it with the commonly used pooling operations in downsampling that directly affect shift invariance, allowing the filter to enhance rather than replace max pooling. This improvement not only retains the advantages of max pooling but also makes the output relatively unaffected by input shifts.

Overall, the MSIR module effectively explores the multidimensional information within the convolutional kernel space, optimizing the design and pooling strategy without causing a substantial increase in the model's parameter count, while maintaining shift invariance and improving robustness to small input image movements.

### *3.4 Cross-Evolutionary Feature Fusion for Infrared Small Targets*

Due to the diverse backgrounds of infrared images and the limited features available for small infrared targets, it is often necessary to incorporate surrounding environmental information to comprehend the abstract concepts inherent in the images for accurate detection. Generally, advanced semantic information in the deeper layers of the network represents abstract image concepts, while the key features of infrared small targets are concentrated in the fine-grained shallow layers. Thus, efficiently fusing features from different network layers becomes a critical challenge for accurate infrared small target detection. Current FPN-based solutions (Fig. 5a–c), despite numerous improvements through attention modules or connection restructuring, still follow the 'top-down, then bottom-up' structure of FPN (i.e., bottom-up feedforward networks, jump connections, and top-down networks). This sequential fusion strategy overlooks the complex interactions between features and the information loss that may occur during the upsampling process. To address this, we propose a Cross-Evolutionary Feature Fusion (CEFF) strategy, which integrates feature maps from multiple convolutional layers, enabling effective fusion of fine-grained low-level details and coarse-grained high-level semantic features while improving upon the traditional FPN's sequential fusion through a more flexible feature fusion pathway.



(a) FPN

(b) FPN Improved Variant Type one

(c) FPN Improved Variant Type two

(d) Cross-Evolutionary Feature Fusion

**Figure 5:** Comparison of FPN and its variants with CEFF structure

As shown in Fig. 6, the entire cross-evolutionary feature fusion structure consists of three parts, namely Multi-Scale Feature Integration (MSFI), Resolution-Adaptive Feature Integrator (RAFI), and Channel and Position Attention Mechanism (CPAM). The core of the module is the Multi-Scale Feature Integration part, which enhances the network's ability to capture features of different scales through the extraction of multi-scale information and improves the receptive field. Meanwhile, the Resolution-Adaptive Feature Integrator further integrates feature maps of different scales, achieving the combination of high-dimensional semantic information and low-level detail information of features. On this basis, we introduce a channel and position attention mechanism that does not require dimensionality reduction, activating key features to more accurately capture local changes and global context in the feature map.

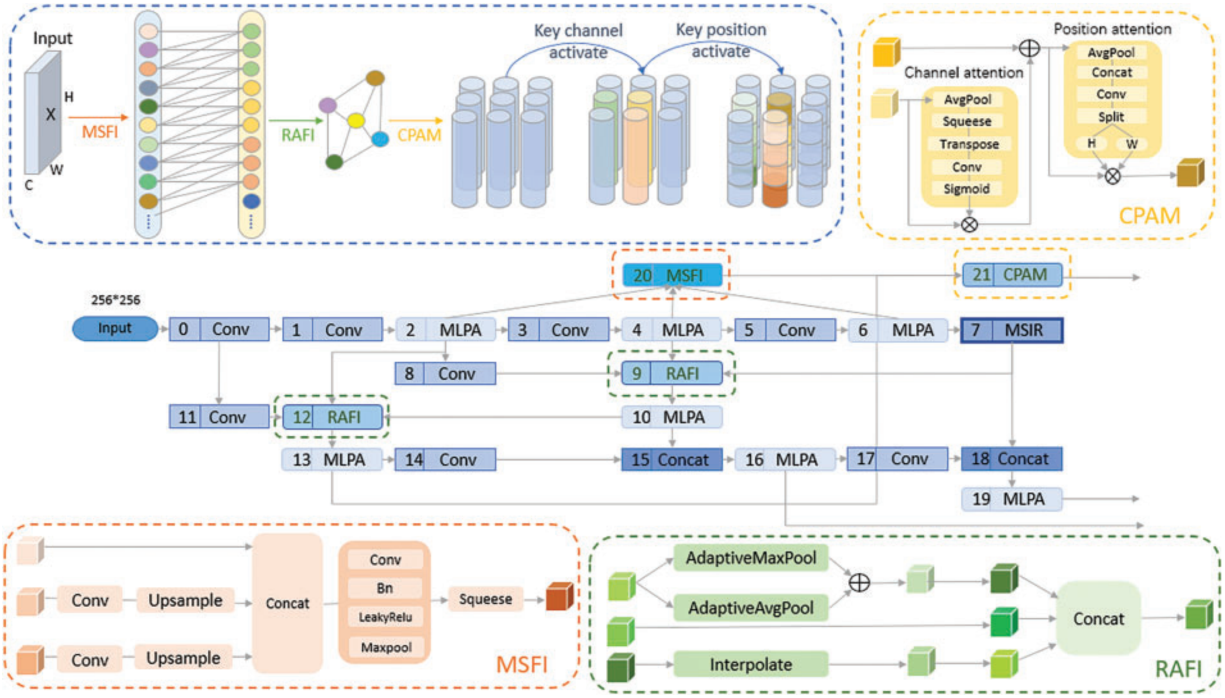


**Figure 6:** The overall connection structure of CEFF and the specific positions of each component module

As shown in Fig. 7, the MSFI method first extracts feature maps from convolutional layers of different depths, where shallow convolutions capture more local spatial details, while deep convolutions have larger receptive fields layer by layer and are used to capture more complex and clustered features of the image. Then, the extracted feature maps are channel shuffled to enhance the interaction between features, and their spatial resolution is unified through interpolation upsampling. Then, the three processed feature maps at different levels are expanded in three dimensions, resulting in 3D feature maps with the same resolution at different levels. On this basis, they are concatenated and combined before applying channel mixing technology again to achieve deep fusion and complementarity between features.

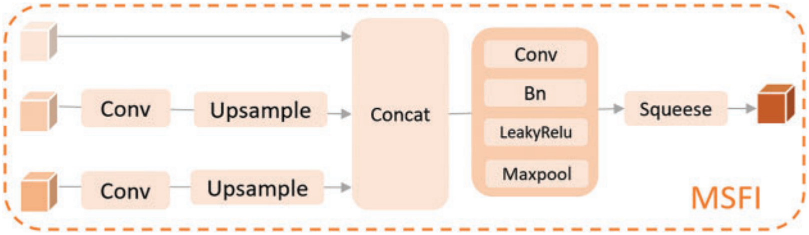


**Figure 7:** MSFI module structure diagram

The commonly used concatenation method in deep convolutional networks directly combines feature maps from different sources along the channel dimension. However, due to feature redundancy across different levels, this redundancy is exacerbated after straightforward concatenation. Additionally, feature maps at various levels differ in semantic and spatial resolution, and direct concatenation may result in feature imbalance, introduce aliasing noise, and degrade model performance. To address this issue, we introduced the RAFI module, which can segment and merge features of varying sizes, ensuring consistent resolution across images. As shown in Fig. 8, feature maps at various resolutions can be considered a scale space. Initially, the effective feature maps with different resolutions are standardized to a common resolution. Then, progressive connection and fusion strategies are employed to mitigate substantial semantic discrepancies between features at different levels. Finally, feature maps are concatenated within the scale space to facilitate information exchange across scales.

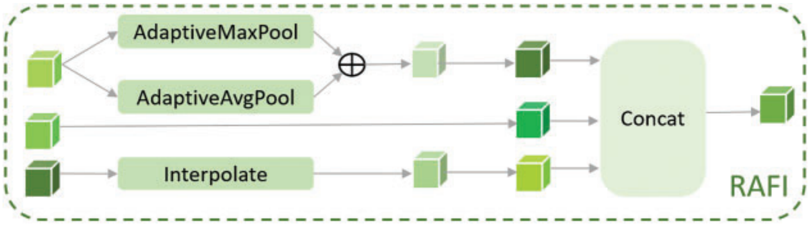


**Figure 8:** Schematic diagram of RAFI module structure

In addition, in order to enhance the flexibility and accuracy of the fusion process, we also introduced an adaptive channel and position attention mechanism (CPAM) that does not require dimensionality reduction to dynamically adjust the fusion weights of features at different scales, as shown in Fig. 9. This not only preserves rich information in the channel and space, but also activates important information in feature fusion, ensuring that key information can be highlighted during the feature fusion process, suppressing noise, and improving the pertinence and efficiency of feature fusion.

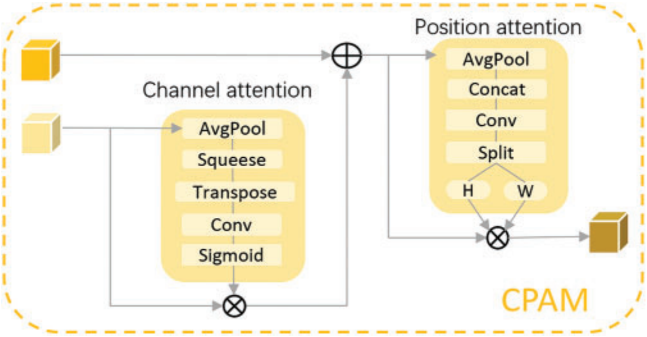


**Figure 9:** Schematic diagram of CPAM module structure

Finally, the connection structure of the three parts is also worth paying attention to. Traditional upsampling or downsampling methods may lose subtle feature information after multiple layers. We have achieved deep fusion of multi-layer features by introducing a long jump connection mechanism. This direct interaction avoids information attenuation during multi-level transmission, preserving weak infrared target features. It is worth noting that our "cross" network connections are not random. In fact, whether it is the RAFI module or the MSFI module, their input features are from adjacent levels, which can prevent significant semantic discrepancies that might occur between non-consecutive levels and ensure the continuity and effectiveness of information flow. In addition, through long connection technology, we combine the features output by the channel attention mechanism with the features of the MSFI module as inputs to the position attention network, providing it with more comprehensive comple-mentary information. This enables the network to simultaneously fuse features from both the channel and position dimensions, further improving the model's detection performance for infrared small targets.

## 4 Experiments and Discussions

### *4.1 Dataset and Experiment Settings*

Due to the inherent dependence of neural networks on datasets, the quantity, quality, and variety of scene contexts within the data significantly impact the performance of the algorithm. The specific parameters of the datasets used are shown in Table 1.

**Table 1:** Dataset comparison table

| Datasets | Image type | Image number | Background |
|---|---|---|---|
| SIRST [15] | Real | 427 | Cloud/City/Sea |
| IRST640 [19] | Synthetic | 1024 | Clouds/City/Tree |
| NUDT-SIRST [17] | Synthetic | 1327 | Cloud/City/Sea/Field/Highlight |

During training, the Kaiming method is applied to initialize the model's weights and biases, and a cosine annealing learning rate scheduler is used. L2 regularization is used to prevent overfitting of the model. The cosine annealing scheduler aids quick convergence early in training and dynamically adjusts the learning rate to prevent local optima. L2 regularization retains all feature information, unlike L1 which may lose useful information. The experimental configuration details and other parameter settings are provided in Table 2. AdamW retains Adam's momentum and adaptive learning rate mechanisms but decouples weight decay from gradient updates, applying it directly to parameter updates for more precise control and avoiding unnecessary influence on bias parameters. In contrast, SGD converges slowly with manual learning rate tuning, while RMSprop, though using an adaptive learning rate, is less refined than AdamW in momentum and weight decay handling.

**Table 2:** Experimental environment and parameter settings

| Name | Settings |
|---|---|
| Processor | Intel (R) Core (TM) i9-10900X CPU @ 3.70 GHz |
| Graphics card | NVIDIA GeForce RTX 4090 |
| Deep learning framework | Pytorch |
| Epoch/Batch size | 300/16 |
| Optimizer/Momentum | AdamW/0.937 |

We use $mAP50$, $mAP50\text{-}95$, $F1_{score}$, ROC, and FPS as evaluation metrics for network performance. Generally speaking, we can divide true targets and detection samples into four categories: $TP$ (true positive), $FP$ (false positive), $TN$ (true negative), and $FN$ (false negative). These categories correspond to the number of positive samples correctly identified as positive, negative samples incorrectly identified as positive, negative samples correctly identified as negative, and positive samples incorrectly identified as negative, respectively. Precision is defined as the ratio of correctly predicted positive samples to the total number of samples predicted as positive, whereas recall is the ratio of correctly predicted positive samples to the total number of actual positive samples, as illustrated in the Eqs. (1) and (2).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

However, considering a single metric alone has significant limitations, so the area under the Precision-Recall curve, also known as the *AP* value, is often used as a reference, as shown in the Eq. (3).

$$AP = \int_0^1 P(r)\mathrm{d}r \tag{3}$$

By setting multiple IoU thresholds, calculating the corresponding *AP* values for each threshold and taking the average for all categories, *mAP* can be obtained as shown in the Eq. (4).

$$mAP = \frac{1}{m}\sum_{i=1}^{m} AP_i \tag{4}$$

Similarly, *mAP*50 represents the *mAP* value with an IoU threshold set at 0.5, while *mAP*50-95 denotes the average of ten *mAP* values obtained by incrementally adjusting the IoU threshold from 0.5 to 0.95 in increments of 0.05, as shown in Eqs. (5) and (6).

$$mAP50 = \frac{1}{m}\sum_{i=1}^{m} AP_{50\,i} \tag{5}$$

$$mAP50 - 95 = \frac{1}{m}\sum_{i=1}^{m}\sum_{n=50}^{95} AP_{n_i} \tag{6}$$

$F1_{score}$ is the harmonic mean of *Recall* and *Precision*, which comprehensively considers the influence of recall and precision to prevent one indicator from dominating the experimental results. Its form is shown in the Eq. (7).

$$F1_{score} = \frac{2TP}{2TP + FP + FN} \tag{7}$$

The ROC curve is also an important indicator for measuring model performance, showing the relationship between the true positive rate (*TPR*) and false positive rate (*FPR*) of the model at different thresholds. The larger the area under the ROC curve, *AUC*, the better the classification performance of the model, as shown in the Eqs. (8)–(10).

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{FP + FN} \tag{9}$$

$$AUC = \int_0^1 TPR(FPR)\ dFPR \tag{10}$$

For applications requiring real-time processing, the detection efficiency of the model is also crucial. The commonly used metric for evaluating detection efficiency is frames per second (*FPS*). A higher *FPS* value indicates a higher detection efficiency of the algorithm, as shown in the Eq. (11).

$$FPS = \frac{frame\ Num}{elapsedTime} \tag{11}$$

### 4.2 Ablation Study

To demonstrate the effectiveness of the modules proposed in this paper, we conducted experiments on the NUDT-SIRST dataset using YOLOv8n as the baseline network. We evaluated the network performance by adding or replacing similar functional modules. As shown in Table 3, the values of $mAP50$ and $mAP50$-95 were presented as the mean ± standard deviation of five experiments. The $p$-value is the result of the Wilcoxon signed-rank test, taking the minimum $p$-value between the two metrics (mAP50 and mAP50-95) when compared with the baseline model (YOLOv8n). The units for mAP50, mAP50-95, and $p$-value are all $10^{-2}$. Our MLPA, MSIR, and CEFF modules performed the best among the compared modules of the same type, both in terms of mAP50 and mAP50-95 metrics, and showed statistical significance ($p < 0.05$).

Among the compared attention modules, the MLPA module has the fewest parameters and outperforms other lightweight attention mechanisms of the same type, such as SimAM [20], EMA [21], MLCA [22], and HiLoAttention [23]. This demonstrates that our module can maintain strong performance while reducing network complexity. The reason for this is that our design enables the network to focus on the target feature regions early in the process. As a result, it not only achieves higher accuracy with fewer parameters but also confirms that shallow networks with parallel substructures can deliver better performance without the need to increase network depth.

**Table 3:** Effectiveness ablation experiments on individual modules of MLPA, MSIR, and CEFF on the NUDT-SIRST dataset, the best results for each type of module are in bold

| Setting | mAP50 | mAP50-95 | FPS | Params (M) | $p$ |
|---|---|---|---|---|---|
| Baseline YOLOv8n | 98.07 ± 0.18 | 80.81 ± 0.20 | **625** | 3.01 | – |
| +FocusedLinearAtten [24] | 98.46 ± 0.25 | 78.90 ± 0.22 | 526 | 3.38 | 0.8 |
| +SimAM [20] | 97.86 ± 0.31 | 75.65 ± 0.27 | 500 | 3.01 | 0.6 |
| +ParNetAttention [25] | 98.43 ± 0.22 | 77.37 ± 0.31 | 555 | 3.97 | 0.7 |
| +BiFormer [26] | 98.62 ± 0.42 | 74.86 ± 0.35 | 588 | 3.27 | 0.5 |
| +LSKBlock [27] | 98.20 ± 0.33 | 78.71 ± 0.28 | 476 | 3.37 | 0.8 |
| +EMA [21] | 98.52 ± 0.27 | 81.09 ± 0.34 | 434 | 3.02 | 3.1 |
| +MLCA [22] | 98.44 ± 0.30 | 81.20 ± 0.29 | 526 | 3.01 | 2.8 |
| +PPAAttention [28] | 97.94 ± 0.54 | 80.66 ± 0.47 | 476 | 11.67 | 3.1 |
| +HiLoAttention [23] | 95.76 ± 0.26 | 67.02 ± 0.26 | 526 | 3.10 | 0.3 |
| +AGCBpatch [18] | 97.96 ± 0.47 | 78.62 ± 0.53 | 192 | 3.35 | 0.8 |
| +**MLPA** | **98.79** ± 0.23 | **81.85** ± 0.20 | 294 | **0.99** | 1.8 |
| +ASPP [29] | 97.84 ± 0.48 | 80.40 ± 0.56 | 99 | 5.07 | 2.6 |
| +DCNv2 [30] | 94.10 ± 0.53 | 61.57 ± 0.42 | 454 | 3.17 | 0.3 |
| +SPDConv [31] | 98.21 ± 0.32 | 77.32 ± 0.26 | 555 | 2.99 | 0.7 |
| +BlurPool [32] | 98.23 ± 0.20 | 81.10 ± 0.37 | 588 | 3.01 | 3.0 |

(Continued)

**Table 3 (continued)**

| Setting | mAP50 | mAP50-95 | FPS | Params (M) | *p* |
|---|---|---|---|---|---|
| +ODConv [33] | 98.54 ± 0.17 | 81.02 ± 0.12 | 400 | 8.79 | 3.3 |
| +RFAConv [34] | 98.41 ± 0.34 | 81.23 ± 0.27 | 526 | 4.40 | 2.5 |
| **+MSIR** | **98.60** ± 0.19 | **81.34** ± 0.21 | 370 | 8.79 | 2.1 |
| +AFPN [35] | 97.31 ± 0.31 | 80.71 ± 0.43 | 454 | 1.85 | 3.6 |
| +ATAC [36] | 97.86 ± 0.49 | 79.39 ± 0.62 | 588 | 3.14 | 0.9 |
| +MPCM [11] | 97.95 ± 0.57 | 81.06 ± 0.71 | 555 | 3.01 | 3.2 |
| **+CEFF** | **99.11** ± 0.20 | **83.26** ± 0.26 | 322 | 3.01 | 1.5 |

In addition, among all the performance comparison experiments involving improved convolutional or pooling modules, the MSIR module proposed in this paper achieves the highest performance, as evidenced by superior scores in both mAP50 and mAP50-95 metrics. It is noteworthy that ASPP, ODConv, and RFAConv also demonstrate strong performance across various tests. We attribute this to the fact that, although these methods employ different specific improvement measures compared to ours, they are inspired by similar underlying principles. Specifically, convolution operations are inherently multi-dimensional rather than linear, whereas traditional convolution typically focuses on a single dimension. Additionally, the shift invariance improvements designed to mitigate the susceptibility of small targets to minor perturbations are particularly well-suited to the demands of infrared small target detection.

Ultimately, in terms of feature fusion methods and other modules designed to enhance infrared small target detection, the CEFF module proposed in this paper achieves the best performance in both mAP50 and mAP50-95, demonstrating the effectiveness of this innovative feature fusion approach. The key innovation of CEFF lies in its connection order and the way connections are made. To assess the effectiveness of our design, we performed ablation studies focusing on various connection structures. As shown in Fig. 10 and Table 4, Fig. 10a introduces a small target detection head compared to Fig. 10b, Fig. 10c features a more lightweight backbone than Fig. 10b, and Fig. 10d shows looser connections and a different fusion order compared to Fig. 10c. The CEFF module not only simplifies the network by removing a convolutional layer from the backbone, preventing the feature maps from becoming too small to detect small targets, but also employs a comprehensive connection strategy that integrates information across all layers. This approach enables the model to better learn both contextual information and the intrinsic features of the targets, ultimately improving detection performance.

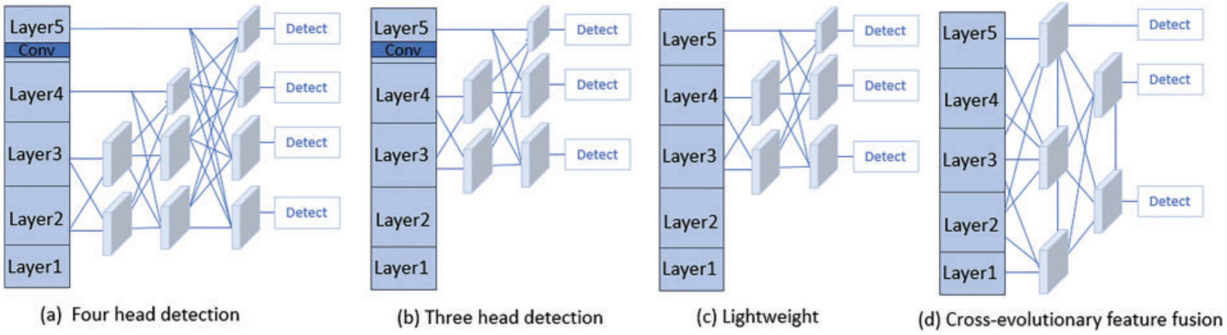


**Figure 10:** The various methods employed in the ablation experiments of feature fusion structures

**Table 4:** CEFF structure ablation experiment on the NUDT-SIRST dataset, with the best results displayed in bold

| Design | mAP50 ($\times 10^{-2}$) | mAP50-95 ($\times 10^{-2}$) | FPS | Params (M) |
|---|---|---|---|---|
| Four head detection | 97.88 | 79.39 | 294 | 7.91 |
| Three head detection | 98.17 | 81.67 | 333 | 7.50 |
| Lightweight | 97.92 | 82.02 | **357** | **2.40** |
| CEFF (ours) | **99.05** | **83.14** | 322 | 3.01 |

The module combination ablation experiments shown in Table 5 demonstrate that the network integrating all three modules achieves improvements in both mAP50 and mAP50-95 compared to single modules or pairwise combinations. This indicates that the modules complement each other and are indispensable within the network—each functioning like interlocking gears, performing their roles while mutually enhancing one another, reflecting the effectiveness of the overall network design.

**Table 5:** Module combination ablation experiment on the NUDT-SIRST dataset, with the best results displayed in bold

| MLPA | MSIR | CEFF | mAP50 ($\times 10^{-2}$) | mAP50-95 ($\times 10^{-2}$) | FPS | Params (M) |
|---|---|---|---|---|---|---|
|  |  |  | 97.90 | 80.80 | **625** | 3.01 |
| √ |  |  | 98.81 | 81.85 | 294 | **0.99** |
|  | √ |  | 98.46 | 81.34 | 370 | 8.79 |
|  |  | √ | 99.05 | 83.14 | 322 | 3.01 |
| √ | √ |  | 98.60 | 81.58 | 384 | 8.81 |
|  | √ | √ | 98.38 | 82.84 | 384 | 2.45 |
| √ |  | √ | 98.59 | 81.86 | 333 | 1.01 |
| √ | √ | √ | **99.26** | **85.22** | 322 | 2.46 |

From the model's FPS and Params, it can be seen that computational efficiency was considered in the design. On NVIDIA RTX 4090, it achieves a single-frame inference speed of 1.9 ms (preprocessing 0.8 ms, postprocessing 0.4 ms), with FLOPs of 2.8 G for 256 × 256 input images. The current GPU memory usage is 24.2 GB, with parameter memory accounting for about 9.84 MB, indicating that most memory consumption comes from activation feature maps, which could be further reduced through architectural optimization. After pruning and quantization, the model size can be further compressed to 0.92 MB (with accuracy dropping to 97.32% mAP50), friendly to edge-device.

## *4.3 Comparison to State-of-the-Art Methods*

### *4.3.1 Quantitative Results and Analysis*

To demonstrate the superiority of our proposed ASCFNet, we compared it with 19 other typical methods, including seven classic traditional infrared small target detection algorithms from three different categories: domain transformation methods (Tophat), low-rank sparse decomposition methods (IPI), and human visual system-based methods (LCM), as well as nine neural network-based infrared small target detection algorithms. Among the 12 deep learning-based algorithms, SSD [37], YOLOv5 [38], YOLOv8 [39], ViT [40], MobileViT [41] and SwinTiny [42] are well-known universal object detection algorithms, representing classic and advanced methods in the general object detection field. MD vs. FA [43], ACM, ALCNet, DNANet, IAANet [44], and AGPCNet are classic and advanced algorithms specifically designed for infrared small target detection. To ensure the fairness of comparative experiments, all deep learning-based algorithms adopted identical experimental settings (including dataset splits, augmentation strategies, hyperparameters, and training environments).

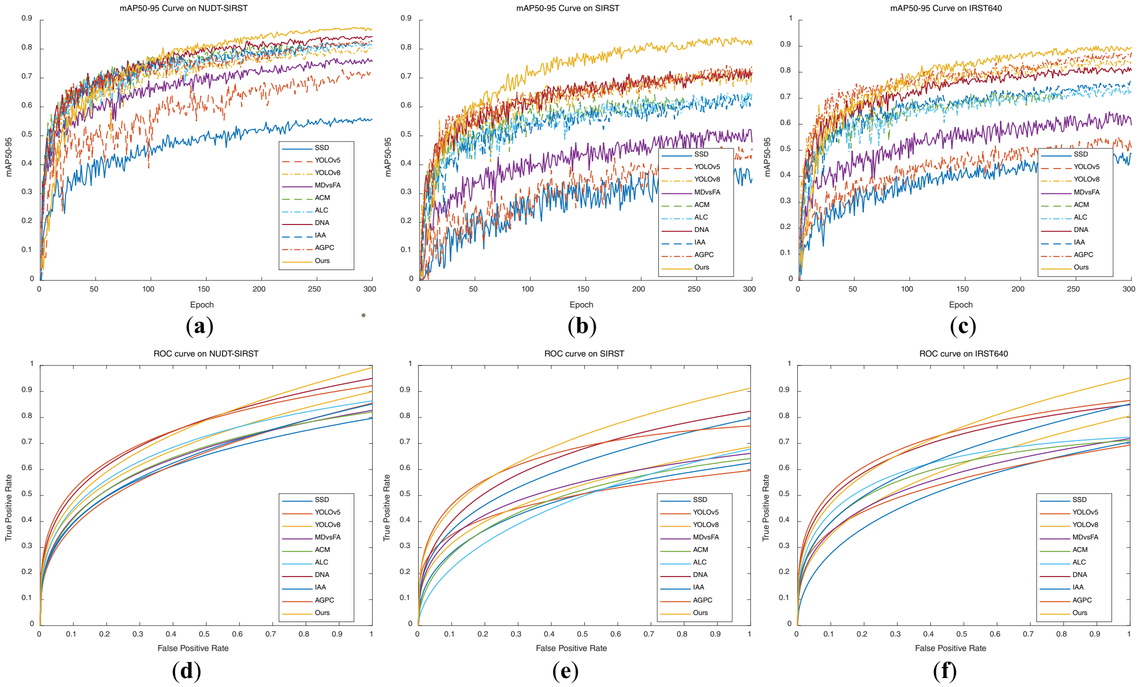
As shown in Table 6, the overall performance of the NUDT-SIRST dataset is generally better than that of the SIRST and IRST640 datasets. Across all three datasets, deep learning-based infrared small target detection algorithms consistently outperform traditional methods. From the table data, among general-purpose object detection algorithms, the Transformer-based ViT underperforms compared to CNN-based YOLO. This may be because ViT's advantages are more apparent on large-scale, high-resolution datasets, while the three commonly used infrared small target datasets in our experiments are relatively lightweight. This hypothesis is corroborated by experimental data showing that Transformer-based models MobileViT and SwinTiny outperform ViT in performance metrics, indicating that lightweight algorithms often demonstrate superior adaptability in small-scale dataset scenarios. Notably, MobileViT exhibits comprehensive performance advantages over both SwinTiny and YOLOv8n. We posit this stems from MobileViT's intrinsic self-attention mechanism design, which enables more effective capture of local feature information in small object detection tasks.

**Table 6:** Comparison of performance metrics with classic algorithms on NUDT-SIRST, SIRST, and IRST640 datasets, with the best results displayed in bold

| **Method** | **NUDT-SIRST** | | | **SIRST** | | | **IRST640** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **mAP50** | **mAP50-95** | **F1** | **mAP50** | **mAP50-95** | **F1** | **mAP50** | **mAP50-95** | **F1** |
| Tophat | 63.21 | 51.06 | 67.32 | 42.35 | 28.67 | 68.94 | 75.10 | 58.21 | 78.61 |
| IPI | 67.35 | 54.71 | 71.60 | 60.37 | 46.62 | 71.13 | 78.69 | 63.25 | 81.44 |
| LCM | 29.14 | 17.37 | 32.09 | 21.33 | 19.21 | 31.75 | 36.51 | 24.07 | 41.10 |
| MPCM | 37.92 | 31.00 | 40.67 | 27.69 | 15.93 | 25.99 | 44.52 | 35.87 | 57.14 |
| ADMD | 40.60 | 37.16 | 42.31 | 47.62 | 35.17 | 50.91 | 53.31 | 41.79 | 67.88 |
| PSTNN | 44.37 | 38.91 | 46.11 | 42.54 | 30.85 | 47.31 | 56.11 | 42.60 | 59.81 |
| SRWS | 23.46 | 12.61 | 25.63 | 14.34 | 9.49 | 6.22 | 26.64 | 15.31 | 29.77 |
| SSD | 59.43 | 52.09 | 63.20 | 40.32 | 29.17 | 43.16 | 56.41 | 43.52 | 60.49 |
| YOLOv5 | 81.30 | 70.61 | 82.67 | 56.64 | 40.31 | 59.78 | 68.98 | 51.30 | 72.49 |
| YOLOv8 | 97.90 | 80.80 | 97.42 | 86.45 | 69.44 | 87.92 | 92.31 | 84.67 | 93.22 |
| MD vs. FA | 88.42 | 76.32 | 89.70 | 62.10 | 49.52 | 68.27 | 70.56 | 59.78 | 74.62 |
| ACM | 98.45 | 80.14 | 98.66 | 73.41 | 60.88 | 75.27 | 81.64 | 70.51 | 80.70 |
| ALCNet | 97.89 | 81.02 | 98.34 | 75.78 | 61.54 | 76.95 | 81.21 | 72.01 | 82.09 |
| DNANet | 98.16 | 84.96 | 98.79 | 83.97 | 72.64 | 84.26 | 89.75 | 80.04 | 91.61 |
| IAANet | 97.90 | 82.30 | 98.82 | 74.12 | 60.40 | 76.34 | 82.36 | 74.96 | 84.79 |
| AGPC | 98.58 | 82.61 | 98.90 | 85.24 | 72.48 | 84.23 | 93.03 | 86.02 | 94.31 |
| ViT | 98.75 | 77.74 | 99.10 | 58.23 | 48.77 | 58.46 | 74.36 | 62.51 | 75.12 |
| MobileViT | 98.86 | 81.26 | 97.76 | 86.51 | 70.12 | 88.32 | 92.44 | 85.02 | 93.36 |
| SwinTiny | 98.77 | 80.64 | 99.15 | 63.41 | 50.72 | 63.87 | 76.98 | 64.11 | 77.09 |
| Ours | **99.26** | **85.22** | **99.31** | **88.60** | **79.67** | **89.69** | **96.54** | **87.98** | **98.82** |

On the SIRST public dataset, compared to ACM proposed by the authors of this dataset, our algorithm ASCFNet achieved an improvement of 15.19% in the mAP50 metric and 18.79% in the mAP50-95 metric. On the NUDT-SIRST and IRST640 datasets, AGPCNet demonstrated the best performance among the compared algorithms, except for our proposed ASCFNet, which surpassed AGPCNet by 2.61% and 1.96% in the mAP50-95 metrics, respectively. Across all three public datasets, ASCFNet outperformed the baseline network YOLOv8, with F1 scores increased by 1.89%, 1.77%, and 5.6%, respectively.

In addition, Fig. 11 compared the mAP50-95 curve and ROC curve of ASCFNet with nine other deep learning-driven infrared small target detection algorithms on three publicly available datasets: NUDT-SIRST, SIRST, and IRST640. From the mAP50-95 curve in the first row, it can be seen that although DNANet and AGPCNet have steeper slopes than ASCFNet, after stabilizing, ASCFNet outperforms all other compared algorithms. As evidenced by the ROC curves in the second row, our ASCFNet demonstrates superior performance across all three datasets by positioning closer to the top-left corner and achieving larger areas under the curve (AUC = 0.974 ± 0.008), indicating enhanced TP/FP discrimination capability. The steep initial ascent reveals that most false positives occur in low-confidence regions. Based on the performance of mAP50-95 and ROC curves, compared with existing advanced algorithms for infrared small target detection, our method achieves a balance between detection efficiency and accuracy. Compared with similar comparative methods, it can stably achieve better infrared small target detection results.



(a) (b) (c)
(d) (e) (f)

**Figure 11:** mAP50-95 and ROC curve graphs of various algorithms on the NUDT-SIRST, SIRST, and IRST640 datasets. (**a**) MAP50-95 curve on NUDT-SIRST dataset; (**b**) MAP50-95 curve on SIRST dataset; (**c**) MAP50-95 curve on IRST640 dataset; (**d**) ROC curve on NUDT-SIRST dataset; (**e**) ROC curve on SIRST dataset; (**f**) ROC curve on IRST640 dataset

### *4.3.2 Qualitative Results and Analysis*

We selected five representative images from three datasets and compared the infrared small target detection results of ASCFNet with those of other algorithms through visualization. As shown in Fig. 12, correctly detected targets are marked with red rectangles, missed targets with green rectangles, and false detections with blue circles. In the first and second rows of the Fig. 12, Tophat, IPI, LCM, and MPCM have unsatisfactory detection results for complex cloudy images, failing not only to accurately detect targets but also producing a large number of false alarms. In deep learning-based methods, DNANet, IAANet, and AGPCNet can accurately recognize both targets, but DNANet and IAANet still generate varying degrees of false alarms. From the third to fourth rows of Fig. 12, it can be seen that IPI, LCM, PSTNN [45], and SRWS [46] perform better on multi-target, low signal-to-noise ratio images than on complex cloudy images. This is because traditional infrared small target algorithms are based on manually designed detection rules, which typically assume that at least one target exists in the image by default, allowing them to produce at least one result, even in cases where the target is dim. Although the deep learning-based methods MD vs. FA, ACM, and ALCNet produce few or no false alarms, they all miss some targets. Overall, from the series of images in Fig. 12, it is clear that our proposed ASCFNet not only detects all the correct targets but also avoids false alarms, consistently achieving the best performance.

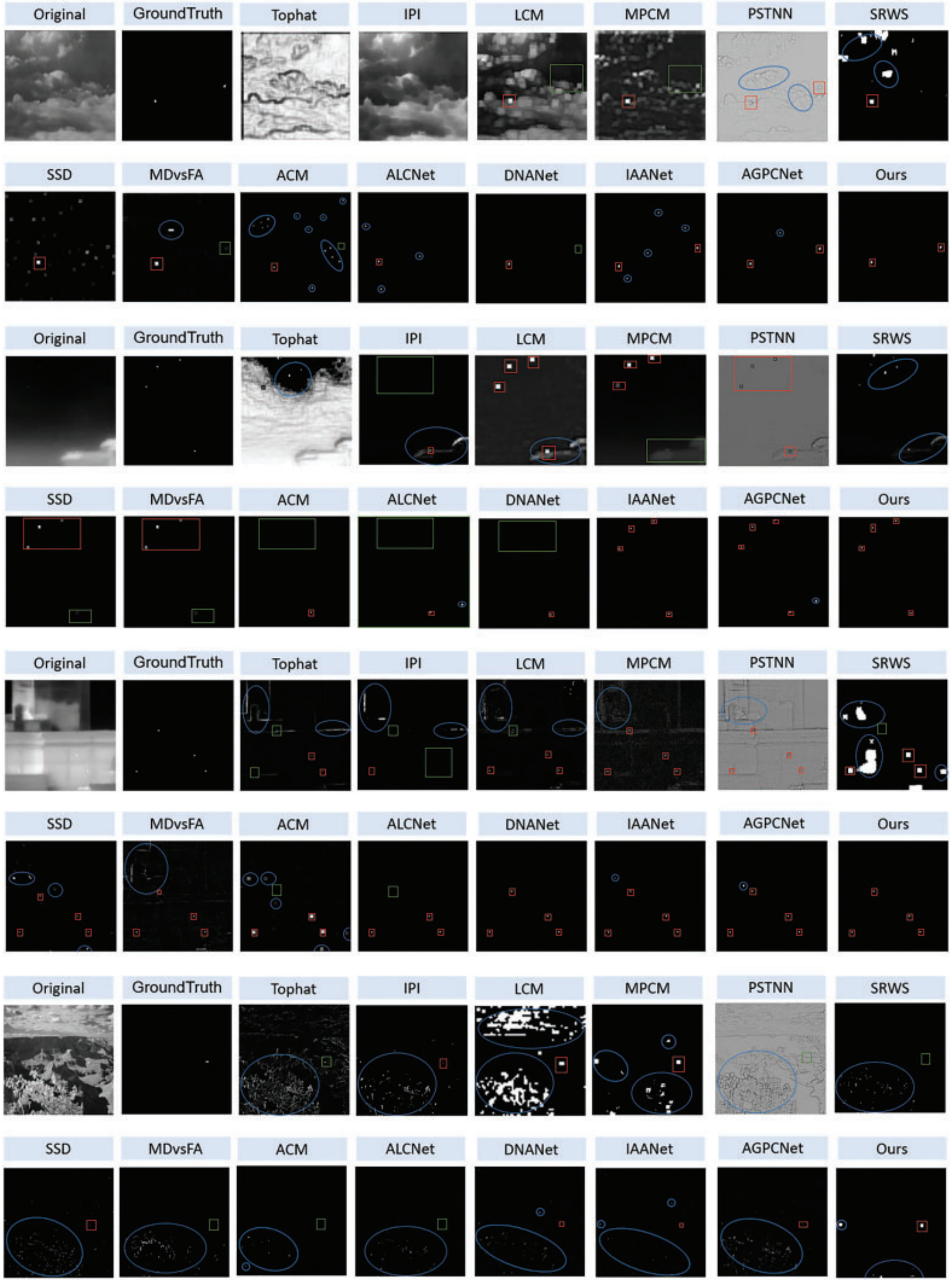


**Figure 12:** Visual comparison of detection effects of different methods

Fig. 12 (lines 7–8) displays two typical false alarms in ASCFNet's overhead field detection. First, bright patches formed by sunlight reflection in hilly areas, whose thermal radiation distribution exhibits gradient features similar to real targets. Second, debris-like noise caused by irregular thermal patterns of foliage shadows, whose high-frequency spatial variations challenge the model's discrimination capability. Further analysis shows that false alarms are mainly concentrated in scenarios with high background complexity, while missed detections mostly occur under extreme conditions where the target signal-to-noise ratio is <3 dB and the size is <3 pixels (accounting for 82% of missed detection cases). Notably, false detections exhibit lower confidence (0.38 ± 0.12) than true detections (0.81 ± 0.09), proving the model's ability to discriminate positives from negatives, consistent with the ROC analysis in Fig. 11. In the future, we will further optimize the model's robustness by introducing temporal feature constraints and multi-modal feature fusion.

Fig. 13 compares the feature activation maps of key layers between the Baseline and ASCFNet under the same scene. Experimental results demonstrate that ASCFNet achieves more profound target feature learning and more precise feature localization. While the Baseline exhibits scattered activation responses with significant background noise interference, ASCFNet not only effectively suppresses irrelevant background activations through its MLPA module but also realizes efficient inter-layer information fusion and transmission via CEFF, maintaining strong activations in key regions even under complex scenarios.

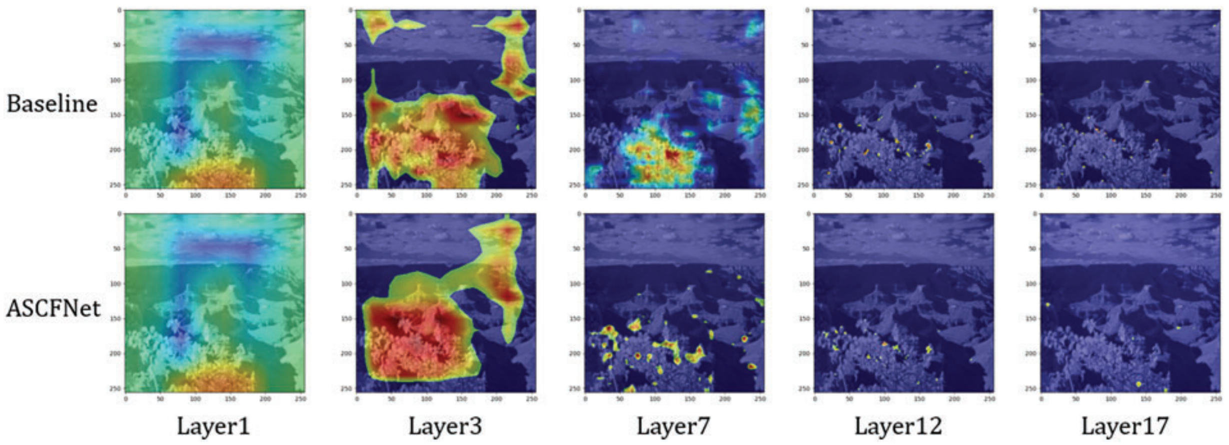


**Figure 13:** The various methods employed in the ablation experiments of feature fusion structures

## 5 Conclusions

This article proposes a deep learning based infrared weak target detection algorithm ASCFNet. By designing MLPA, MSIR, and CEFF modules, a unique attention shift-invariant cross-evolution feature fusion framework is constructed, alleviates problems such as feature dilution, small input disturbance, and underutilization of features for small targets. ASCFNet consistently achieved the best performance in ablation experiments conducted on widely recognized public datasets SIRST, NUDT-SIRST, and IRST640, and compared with advanced algorithms in the field. Among them, the NUDT-SIRST dataset has the highest indicator values, with mAP50, mAP50-95, and $F1_{score}$ reaching 99.26%, 85.22%, and 99.31%, respectively. The model parameter size is 2.46 M, which is 18.27% lower than YOLOv8n. Visual evaluations of detection results in diverse scenarios indicate that our algorithm exhibits an increased detection rate and reduced false alarm rate.

**Acknowledgement:** Not applicable.

**Funding Statement:** This work was supported in part by the National Natural Science Foundation of China under Grant 62271302 and by the Shanghai Municipal Natural Science Foundation under Grant 20ZR1423500.

**Author Contributions:** The authors confirm contribution to the paper as follows: Conceptualization, methodology, and writing—review and editing, Siqi Zhang; data curation, visualization and writing—original draft preparation, Shengda Pan. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The SIRST, IRST640 and NUDT-SIRST dataset used for training and test are available at: https://github.com/YimianDai/open-acm, https://github.com/jzchenriver/IRST640 and https://github.com/YeRen123455/Infrared-Small-Target-Detection (accessed on 20 May 2025).

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Cheng Y, Lai X, Xia Y, Zhou J. Infrared dim small target detection networks: a review. Sensors. 2024;24(12):3885. doi:10.3390/s24123885.
2. Tong X, Sun B, Wei J, Zuo Z, Su S. EAAU-Net: enhanced asymmetric attention U-Net for infrared small target detection. Remote Sens. 2021;13(16):3200. doi:10.3390/rs13163200.
3. Zhao M, Li W, Li L, Hu J, Ma P, Tao R. Single-frame infrared small-target detection: a survey. IEEE Geosci Remote Sens Mag. 2022;10(2):87–119. doi:10.1109/MGRS.2022.3145502.
4. Bai X, Zhou F. Top-hat selection transformation for infrared dim small target enhancement. Imaging Sci J. 2010;58(2):112–7. doi:10.1179/136821909X12581187860176.
5. Zhou A, Xie W, Pei J. Background modeling in the fourier domain for maritime infrared target detection. IEEE Trans Circuits Syst Video Technol. 2020;30(8):2634–49. doi:10.1109/TCSVT.2019.2922036.
6. Deng H, Sun X, Zhou X. A multiscale fuzzy metric for detecting small infrared targets against chaotic cloudy/sea-sky backgrounds. IEEE Trans Cybern. 2019;49(5):1694–707. doi:10.1109/TCYB.2018.2810832.
7. Liu D, Cao L, Li Z, Liu T, Che P. Infrared small target detection based on flux density and direction diversity in gradient vector field. IEEE J Sel Top Appl Earth Obs Remote Sens. 2018;11(7):2528–54. doi:10.1109/JSTARS.2018.2828317.
8. Huang M, Mu Z, Zeng H. A novel approach for interest point detection via laplacian-of-bilateral filter. J Sens. 2015;2015(1):685154. doi:10.1155/2015/685154.
9. Chen CLP, Li H, Wei Y, Xia T, Tang YY. A local contrast method for small infrared target detection. IEEE Trans Geosci Remote Sens. 2014;52(1):574–81. doi:10.1109/TGRS.2013.2242477.
10. Han J, Ma Y, Zhou B, Fan F, Liang K, Fang Y. A robust infrared small target detection algorithm based on human visual system. IEEE Geosci Remote Sens Lett. 2014;11(12):2168–72. doi:10.1109/LGRS.2014.2323236.
11. Wei Y, You X, Li H. Multiscale patch-based contrast measure for small infrared target detection. Pattern Recognit. 2016;58(1):216–26. doi:10.1016/j.patcog.2016.04.002.
12. Han J, Liang K, Zhou B, Zhu X, Zhao J, Zhao L. Infrared small target detection utilizing the multiscale relative local contrast measure. IEEE Geosci Remote Sens Lett. 2018;15(4):612–6. doi:10.1109/LGRS.2018.2790909.
13. Gao C, Meng D, Yang Y, Wang Y, Zhou X, Hauptmann AG. Infrared patch-image model for small target detection in a single image. IEEE Trans Image Process. 2013;22(12):4996–5009. doi:10.1109/TIP.2013.2281420.
14. Moradi S, Moallem P. Sabahi MF fast and robust small infrared target detection using absolute directional mean difference algorithm. Signal Process. 2020;177(1):107727. doi:10.1016/j.sigpro.2020.107727.
15. Dai Y, Wu Y, Zhou F, Barnard K. Asymmetric contextual modulation for infrared small target detection. In: Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 5–9; Waikoloa, HI, USA. doi:10.1109/WACV48630.2021.00099.

16. Dai Y, Wu Y, Zhou F, Barnard K. Attentional local contrast networks for infrared small target detection. IEEE Tran Geosci Remote Sens. 2021;59(11):9813–24. doi:10.1109/TGRS.2020.3044958.
17. Li B, Xiao C, Wang L, Wang Y, Lin Z, Li M, et al. Dense nested attention network for infrared small target detection. IEEE Trans Image Process. 2023;32(3):1745–58. doi:10.1109/TIP.2022.3199107.
18. Zhang T, Li L, Cao S, Pu T, Peng Z. Attention-guided pyramid context networks for detecting infrared small target under complex background. IEEE Trans Aerosp Electron Syst. 2023;59(4):4250–61. doi:10.1109/TAES.2023.3238703.
19. Chen G, Wang W, Tan S. IRSTFormer: a hierarchical vision transformer for infrared small target detection. Remote Sens. 2022;14(14):3258. doi:10.3390/rs14143258.
20. Yang L, Zhang RY, Li L, Xie X. SimAM: a simple, parameter-free attention module for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning; 2021 Jul 13–18; New York, NY, USA.
21. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. doi:10.1109/ICASSP49357.2023.10096516.
22. Wan D, Lu R, Shen S, Xu T, Lang X, Ren Z. Mixed local channel attention for object detection. Eng Appl Artif Intell. 2023;123(1):106442. doi:10.1016/j.engappai.2023.106442.
23. Pan Z, Cai J, Zhuang B. Fast vision transformers with HiLo attention. arXiv:2205.13213. 2023.
24. Han D, Pan X, Han Y, Song S, Huang G. FLatten transformer: vision transformer using focused linear attention. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 2–6; Paris, France. doi:10.1109/ICCV51070.2023.00548.
25. Goyal A, Bochkovskiy A, Deng J, Koltun V. Non-deep networks. arXiv:2110.07641. 2021.
26. Zhu L, Wang X, Ke Z, Zhang W, Lau RWH. BiFormer: vision transformer with bi-level routing attention. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 18–22; Vancouver, BC, Canada.
27. Li Y, Hou Q, Zheng Z, Cheng MM, Yang J, Li X. Large selective kernel network for remote sensing object detection. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 2–6; Paris, France. doi:10.1109/ICCV51070.2023.01540.
28. Xu S, Zheng S, Xu W, Xu R, Wang C, Zhang J, et al. HCF-Net: hierarchical context fusion network for infrared small object detection. In: Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME); 2024 Dec 27–29; Shenzhen, China.
29. Chen LC, Papandreou G, Kokkinos I, Murphy KP, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2016;40(4):834–48. doi:10.1109/TPAMI.2017.2699184.
30. Zhu X, Hu H, Lin S, Dai J. Deformable ConvNets V2: more deformable, better results. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA.
31. Sunkara R, Luo T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. arXiv:2208.03641. 2022.
32. Zhang R. Making convolutional networks shift-invariant again. arXiv:1904.11486. 2019.
33. Li C, Zhou A, Yao A. Omni-dimensional dynamic convolution. arXiv:2209.07947. 2022.
34. Zhang X, Liu C, Yang D, Song T, Ye Y, Li K, et al. RFAConv: innovating spatial attention and standard convolutional operationar. arXiv:2304.03198. 2024.
35. Yang G, Lei J, Zhu Z, Cheng S, Feng Z, Liang R. AFPN: asymptotic feature pyramid network for object detection. In: Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2023 Oct 1–4; Oahu, HI, USA.
36. Dai Y, Oehmcke S, Wu Y, Barnard K. Attention as activation. In: Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR); 2020 Aug 20–24; Hong Kong, China.

37. Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu CY, et al. SSD: single shot MultiBox detector. In: Proceedings of the European Conference on Computer Vision; 2016 Oct 11–14; Amsterdam, The Netherlands. doi:10.1007/978-3-319-46448-02.
38. Yolov5/Ultralytics [Internet]. [cited 2024 Dec 11]. Available from: https://github.com/ultralytics/yolov5/releases.
39. Jocher G. Ultralytics. Ultralytics YOLO [Internet]. [cited 2024 Dec 11]. Available from: https://docs.ultralytics.com/zh/models/yolov8/.
40. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
41. Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv:2110.02178. 2021.
42. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: hierarchical vision transformer using shifted Windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada. doi:10.1109/ICCV48922.2021.00986.
43. Wang H, Zhou L, Wang L. Miss detection vs. false alarm: adversarial learning for small object segmentation in infrared images. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–30; Seoul, Republic of Korea. doi:10.1109/ICCV.2019.00860.
44. Wang K, Du S, Liu C, Cao Z. Interior attention-aware network for infrared small target detection. IEEE Trans Geosci Remote Sens. 2022;60:1–13. doi:10.1109/TGRS.2022.3163410.
45. Zhang L, Peng Z. Infrared small target detection based on partial sum of the tensor nuclear norm. Remote Sens. 2019;11(4):382. doi:10.3390/rs11040382.
46. Zhang T, Peng Z, Wu H, He Y, Li C, Yang C. Infrared small target detection via self-regularized weighted sparse model. Neurocomputing. 2021;420(12):124–48. doi:10.1016/j.neucom.2020.08.065.