applied sciences

MDPI

Article

Infrared Multi-Scale Small-Target Detection Algorithm Based on Feature Pyramid Network

Sanxia Shi and Yinglei Song *

School of Automation, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 211110303103@stu.just.edu.cn
* Correspondence: syinglei2013@163.com

Abstract: Technologies for the detection of dim and small targets in infrared images play an increasingly important role in various applications, including military early warning, precise guidance, military reconnaissance, environmental monitoring, and aerospace applications. This paper proposes a new approach for the detection of infrared multi-scale small targets based on a feature pyramid network. Three pyramid segmentation–connection modules are incorporated into the proposed pyramid network to capture both local and global context information across various layers. Furthermore, a dual attention fusion module is proposed to fuse the feature maps containing context information and the deep features that have been upsampled twice through the attention mechanism of the dual attention fusion module to highlight important semantic information. Experimental results on two benchmark datasets show that the proposed method can generate results with good accuracy on both datasets and outperforms several other state-of-the-art methods for small-target detection in terms of accuracy and robustness.

Keywords: object detection; deep learning; feature pyramid network; attention mechanism fusion

Citation: Shi, S.; Song, Y. Infrared Multi-Scale Small-Target Detection Algorithm Based on Feature Pyramid Network. *Appl. Sci.* **2024**, *14*, 5587. https://doi.org/10.3390/app14135587

Academic Editor: Sungho Kim

Received: 18 May 2024
Revised: 16 June 2024
Accepted: 18 June 2024
Published: 27 June 2024

1. Introduction

In the field of computer vision, object detection plays a crucial role in tasks such as object tracking, image segmentation, and scene understanding. It is not only the core element for determining the presence or absence of an object in an image and establishing its specific location, but also the foundation for significant advancements in intelligent video surveillance, autonomous driving, medical image analysis, and other fields [1]. Through image processing technology, the location and spatial information of a target can be efficiently and accurately extracted from a large amount of digital image data. Such information can be subsequently processed by other applications for the accurate analysis of the target.

Traditional target detection methods mainly include three types. Detection methods based on a background estimation approach model the background information and treat the target as an abnormal object deviated from the background [2–6]. Detection methods based on human vision understand and perceive targets in ways that are similar to those of human vision [7–11]. Detection techniques based on the low-rank sparse decomposition model decompose a target into low-rank signals, while decomposing the background and noise into sparse signals [12–16]. Traditional object detection methods mainly rely on features manually extracted from an object to achieve object detection. However, these traditional methods often require a large amount of prior information on the background statistics. These methods thus are easily affected by factors such as noise and inhomogeneity and cannot adapt to changes in the target scale.

Recently, deep learning-based approaches have been employed to detect dim and small targets in infrared images. An important advantage of deep learning-based methods over other methods is their ability to extract features automatically. Most of the deep

learning-based methods can be divided into two types in the early stages. One type of method utilizes candidate boxes to perform detections in two stages [17], and the other type of method detects infrared small targets in one stage based on regression [18,19]. However, due to the scarcity of infrared dim small-target datasets, the challenges in training network models, and the degradation of infrared dim small-target features as the number of network layers increases, further improvements still need to be achieved in the performance of infrared dim small-target detection.

In the realm of target detection, infrared small-target detection can also be accomplished as an object segmentation task. Applying RGB image segmentation based on deep convolutional neural networks to the segmentation of small targets in infrared images remains a challenge that requires further exploration [20].

Recognizing objects of various sizes is a fundamental challenge in object detection. Pyramid networks have been a fundamental component of multi-scale object detection. However, the pyramid network involves a significant amount of computation, which could potentially slow down the entire detection process. Therefore, in order to enhance detection speed, many methods opt to avoid using pyramid networks and instead rely solely on high-level features for prediction. High-level features contain rich semantic information, but accurately capturing the position of the object is challenging due to the low resolution [21]. On the contrary, although low-level features contain less semantic information, they offer high resolution and rich details, enabling accurate representation of object locations. Therefore, the fusion of low-level and high-level features can achieve accurate recognition and localization in object detection systems. This fusion can combine the rich semantic information in high-level features with the accurate location information in low-level features to achieve more precise target detection and localization.

Lin et al. [22] proposed the feature pyramid network (FPN) in 2017, which has significantly improved the performance of small-object detection without increasing the original computational load. Ronneberger et al. [23] proposed the U-Net architecture in 2015. Although the U-Net architecture is not specifically designed for detecting infrared small targets, it has demonstrated excellent performance in infrared small-target detection. Zhao et al. [24] proposed the TBC-Net in 2021, which consists of a target extraction module (TEM) and a semantic constraint module (SCM). It can effectively reduce false alarms caused by complex backgrounds. Dai et al. [25] proposed ALC-Net, which contains a bottom-up attention modulation module to integrate detailed information from low-level features into deeper high-level features. It also utilizes multi-scale local contrast measures to address the issue of target scale variations. Li et al. [26] proposed DNA-Net, which implements progressive interaction between high-level and low-level features and can adaptively enhance multi-level features. In addition, some other methods based on deep learning and feature pyramid networks [27–29] have also yielded promising results. However, further improvements in detection accuracy are still required for applications where infrared multi-scale small targets need to be accurately recognized and analyzed.

Therefore, this paper proposes an infrared multi-scale small-target detection method based on a dual attention fusion mechanism of a feature pyramid network to improve the detection accuracy of infrared multi-scale small target images. The main contributions of the paper are as follows.

(1) Three pyramid segmentation–connection (PSC) modules are incorporated into the feature pyramid network to obtain both local and global context information from various shallow and deep features.
(2) A dual attention fusion (DAF) module is proposed to fuse the feature maps containing context information and the deep features that have been upsampled twice through the attention mechanism of the DAF module. The module is used to highlight important semantic information.

2. Materials and Methods

Although many current methods employ various techniques to enhance detection performance, there is still the issue of small targets disappearing in the deep network in many methods. Therefore, considering the limitations of the current detection algorithm, this paper proposes a dual attention fusion mechanism based on a pyramid network for infrared multi-scale small-target detection. It is mainly divided into three parts: pyramid segmentation–connection (PSC), context semantic expression (CSE), and dual attention fusion (DAF) modules. The context semantic representation module includes a global semantic relevance (GSR) module and a local semantic stitching (LSS) module.

2.1. Overall Algorithm Architecture

Figure 1 illustrates the overall architecture of the proposed infrared multi-scale small-target detection model. In the first step, the image that contains an infrared multi-scale small target is processed using a deep convolutional neural network, and feature maps of 1/2, 1/4, and 1/8 of the original image are obtained. The feature maps from various stages of feature extraction are processed by the PSC module. Then, the feature map generated by the PSC module in each stage is upsampled by a factor of 2 and fused with the feature map obtained in the previous stage with the PSC module to extract the portion that encapsulates information crucial for small-target identification. The fusion is performed with a DAF module. After two fusions, the final fused feature map is used to segment the small target, resulting in the acquisition of the small target in an infrared image.

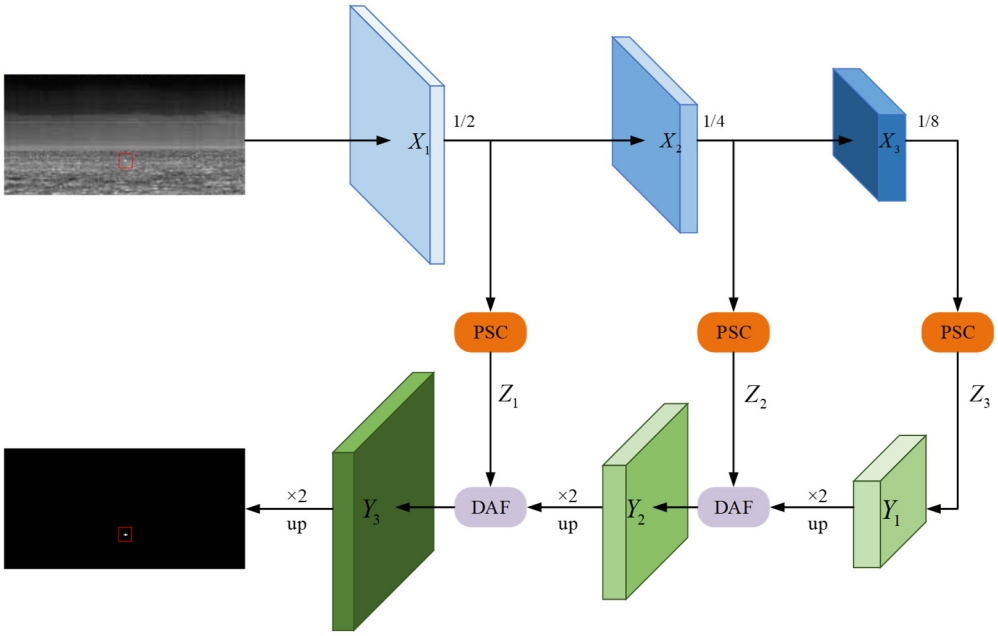

Figure 1. The overall model architecture.

In this paper, ResNet18 [30] is chosen as the backbone network for the feature extraction stage. The input infrared small-target image is downsampled to extract detailed and rich small-target features. After multiple operations of downsampling and feature extraction, the deep semantic features of small targets are obtained. The low-level feature map retains detailed information such as the location and specifics of the small target, while the high-level features can capture the semantic information of the small target. The feature map obtained through the fusion of low-level features and high-level features not only preserves the benefits of low-level feature maps but also maintains the advantages of high-level feature maps. The final feature map can detect the position and shape of small targets more accurately, with higher detection accuracy and a lower false-alarm rate.

2.2. Pyramid Segmentation–Connection Module

The pyramid segmentation–connection module takes the feature map X obtained from feature extraction as its input and processes it to produce the output feature map Z. The detailed structure of PSC is shown in Figure 2. PSC feeds the feature map X into multiple scales of CSE, and the scale is denoted as $b \in \{b_1, b_2, \ldots, b_n\}$. For each scale, CSE preserves the essential details of small targets by incorporating contextual information, as illustrated in Figure 3, depicting the structural diagram of the CSE model. It can be seen in the figure that CSE is divided into two parallel processing flows, namely, LSS and GSR. The feature maps processed by LSS and GSR are convolved, and then the convolved feature maps are added to X to obtain the final output M. Subsequently, M and X obtained at different scales are concatenated together. Finally, the information from multiple scales is integrated, and the feature map Z is obtained through 1×1 convolution. As can be seen from the model architecture of infrared multi-scale small-target detection in Figure 1, the PSC model has been processed three times throughout the entire process. Although the size of the input feature map X varies across different stages, the selected scale b remains consistent.

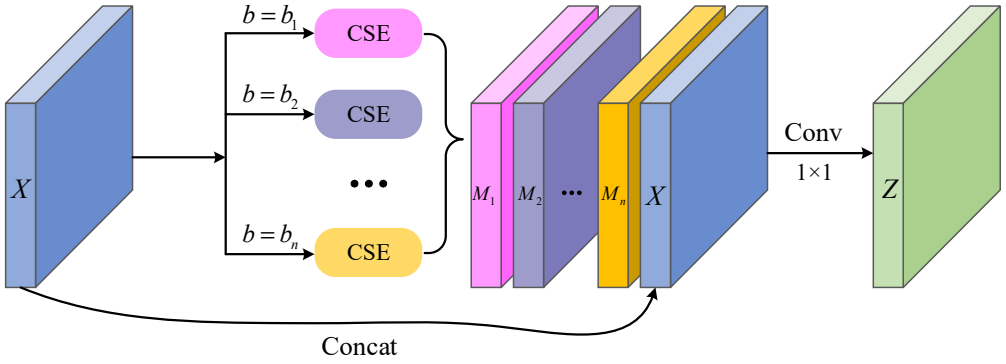

Figure 2. The PSC module.

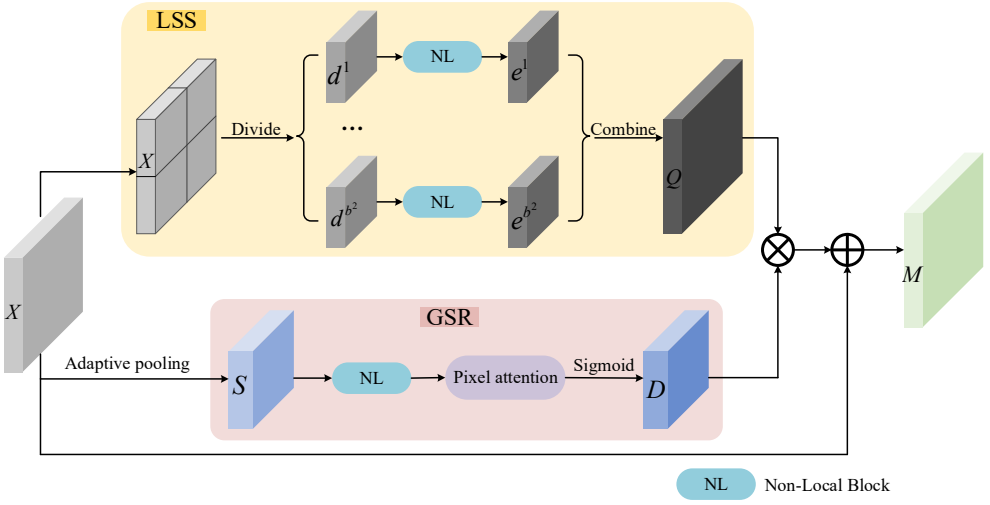

Figure 3. The CSE module.

2.3. Local Semantic Stitching Module

The structure of the local semantic stitching module is shown in the upper part of Figure 3, and is used to extract the features of the local block of the feature map, referred to as LSS. For the feature map, $X \in R^{C \times H \times W}$, obtained through feature extraction at a given scale, b, X is divided into $b \times b$ blocks to obtain b^2 feature maps $d^i \in R^{C \times \frac{H}{b} \times \frac{W}{b}}$ and $i \in \{1, 2, \ldots, b^2\}$. The b^2 feature map d^i is processed by non-local blocks to obtain the b^2 feature blocks $e^i \in R^{C \times \frac{H}{b} \times \frac{W}{b}}$ and $i \in \{1, 2, \ldots, b^2\}$. Subsequently, the b^2 feature blocks are merged based on the position of the feature map X to obtain the feature map $Q \in R^{C \times H \times W}$

with the semantic splicing information of the local blocks. The non-local block captures long-distance dependencies by calculating the similarity between each position in the feature graph d and all other positions. The operations in the process are described by Equations (1) and (2):

$$Q_k^i = \beta \sum_{j=1}^{HW/b^2} \omega_{kj}^i \varphi\left(X_k^i\right) + X_k^i \tag{1}$$

$$\omega_{kj}^i = \frac{\exp\left(\theta\left(X_k^i\right)^T \phi\left(X_j^i\right)\right)}{\sum_{j=1}^{HW/b^2} \exp\left(\theta\left(X_k^i\right)^T \phi\left(X_j^i\right)\right)} \tag{2}$$

where Q_k^i represents k elements, β represents the scalar that can be learned, ω_{kj}^i represents the element of the kth row and jth column of the coefficient matrix ω^i, and $\varphi(.)$, $\theta(.)$, and $\phi(.)$ are all 1×1 convolution operations.

2.4. Global Semantic Relevance Module

The global semantic correlation module is shown in the lower part of Figure 3. It is used to estimate the levels of dependency among blocks and will be referred to as GSR. For feature map $X \in R^{C \times H \times W}$, obtained through feature extraction at a given scale b, X is processed by adaptive pooling to obtain a feature map $S \in R^{C \times b \times b}$ of size $C \times b \times b$. Each point in set S corresponds to the feature block at the corresponding position in the LSS module. The non-local block then estimates the similarity between each position in S and all other positions, capturing long-range dependencies by estimating the correlation between each block in feature map d. Afterward, the obtained features are input into the pixel attention module. This module adjusts the weight of each pixel dynamically, highlighting important pixel regions and suppressing unimportant ones, thus enhancing the representation ability of pixel-level features. Finally, the feature map $D \in R^{c \times b \times b}$ is obtained by the sigmoid function, and the operations performed in the process are described by Equations (3)–(5):

$$D_m = \delta(PA(\beta \sum_{j=1}^{b^2} \omega_{mj} \psi(S_m) + S_m)) \tag{3}$$

$$\omega_{mj} = \frac{1}{Z_m} \exp\left(\theta(S_m)^T \phi(S_j)\right) \tag{4}$$

$$Z_m = \sum_{j=1}^{b^2} \exp\left(\theta(S_m)^T \phi(S_j)\right) \tag{5}$$

where D_m represents the mth element of D, PA represents the pixel-level attention module, $\delta(.)$ represents the sigmoid function, ω_{mj} represents the element in the mth row and jth column of the coefficient matrix, and $\psi(.)$ is the 1×1 convolution operation.

The feature map Q obtained after processing the feature map X with the LSS module represents the semantic associations at the local level, while the feature map D obtained after the GSR module represents the correlations between feature blocks. To integrate the two types of information, the obtained D was used as the convolution kernel, and the feature map Q was convolved. In order to preserve the information conveyed by the original feature map X, the final output m of the context semantic expression module is obtained by adding and fusing the result of the convolution operation with X. The operation in the process is shown in Equation (6).

$$M_i = Q_k^j * D_i + X_i \tag{6}$$

2.5. Dual Attention Fusion Module

The area of infrared small targets contains only a few pixels, and as the network deepens, the targets are likely to be lost in deeper networks. Shallow networks capture detailed information about small objects, such as their location, size, and edges. However, they lack a deep understanding of small objects. Deep networks capture various semantic

information about small targets, but may lose the details of these small targets. Feature representation is extremely challenging when the final feature map lacks information from the different network stages. Therefore, knowing how to integrate shallow features and deep features is crucial for accurately detecting infrared small targets. Previous studies have shown that the fusion of spatial attention and channel attention [31] can highlight the optimal semantic feature regions and reduce the computational complexity [32]. This fusion enables the retention and extraction of the original features of small targets in deep networks [33].

In order to enhance detection effects and improve detection accuracy, this paper introduces a dual attention fusion module, referred to as a DAF module. The module structure is shown in Figure 4. The module can not only extract the deep semantic information of the small target but also fuse detailed information, such as the location of the small target. It can also solve the problems of information redundancy and inadequate feature fusion.

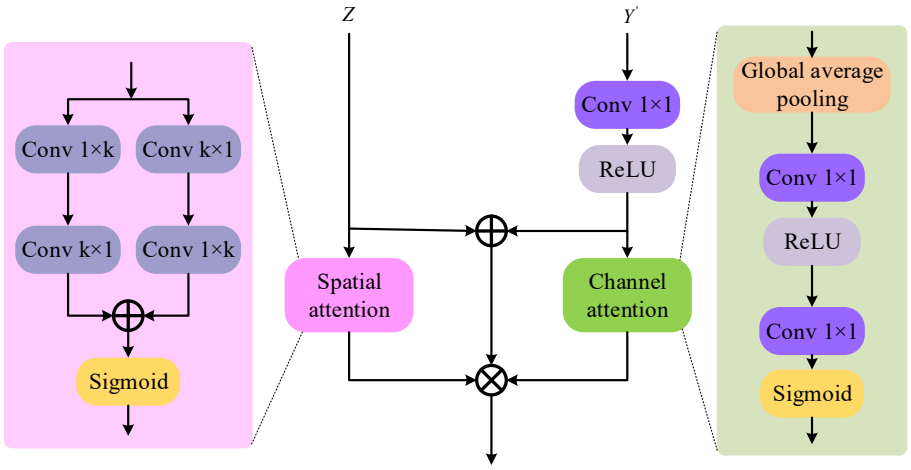

Figure 4. The DAF module.

The low-level features Z and high-level features $Y\prime$ represent different semantic information. The high-level features $Y\prime$ are obtained from Y in Figure 1 after two rounds of upsampling operations. The low-level features contain the location information of multiple small targets, and the spatial attention mechanism module can be used to process the local location information. The high-level features can utilize the channel attention mechanism module to process channel semantic information, prioritizing the semantic expression of different channels. The low-level features and high-level features are incorporated together and fused with the low-level features constrained by spatial attention and the high-level features constrained by channel attention, respectively, as shown in Equation (7):

$$DAF(Z, Y') = (Z + f(Y)) \bigotimes PA(Z) \bigodot CA(f(Y')) \quad (7)$$

where $PA(.)$ and $CA(.)$ represent the spatial and channel attention constraints, $f(.)$ represents the convolution operation, and $\bigotimes$ and $\bigodot$ represent the corresponding multiplication operations of the element and vector tensors, respectively.

During upsampling, the DAF module fuses low-level and deep-level semantic information of $1/4$ and $1/2$ spatial sizes respectively. They are all preceded by a bilinear interpolation operation. Finally, the segmentation network of fused features was used to predict the final detection result of infrared small targets.

Aiming at the characteristics of class imbalance and weak texture differences between infrared small targets and the infrared image background, the target detection network needs to comprehensively consider spatial and channel features, as well as cross-layer feature fusion. This implies that the network should not only capture the target features within a single image layer but also focus on the information transfer and fusion between different layers to improve the identification and localization of small infrared targets. This

mechanism of cross-layer feature fusion is helpful for resolving the contrast between the target and the background in the infrared image, enhancing the accuracy and robustness of target detection.

3. Results

To demonstrate the effectiveness of the proposed algorithm in infrared multi-scale small-target detection, experiments are conducted to validate its performance. Firstly, the evaluation index used is introduced, and then the implementation details of the proposed method are provided. Finally, ablation experiments and comparison experiments are conducted, and the experimental results are analyzed. The experiment uses Ubuntu 18.04 operating system, Pytorch 1.8.0 deep learning framework, 3.00 GHz Intel Core i9-13900KF CPU (Intel, Santa Clara, CA, USA), and CUDA version 11.1.1.

3.1. Evaluation Index

The approach proposed in this paper achieves the detection of infrared multi-scale small targets through semantic segmentation. Therefore, the F1 score, intersection over union (IoU), and normalized IoU($nIoU$) are used as quantitative evaluation metrics for semantic segmentation.

(1) F1 is the harmonic mean of precision and recall, which can be used to comprehensively evaluate the performance of the model. The formula for F1 is shown in Equation (8):

$$F_1 = (2 * Precision * Recall)/(Precision + Recall) \tag{8}$$

where $Precision$ refers to the proportion of samples predicted as positive by the model that are actually positive, while $Recall$ refers to the proportion of samples that are actually positive and are predicted as positive. Higher values of $Precision$ and $Recall$ indicate better performance of the network model, and their values are calculated with Equations (9) and (10), respectively:

$$Precision = TP/(TP + FP) \tag{9}$$

$$Recall = TP/(TP + FN) \tag{10}$$

where true positive (TP) refers to the true examples. This is the number of pixels where the predicted box intersects the true box. TP represents the number of samples that were correctly predicted as positive by the model. On the other hand, false positive (FP) is the number of false positives that the predicted box incorrectly contains outside the true box. FP indicates the number of examples that the model incorrectly predicts to be positive. False negative (FN) refers to the false-negative examples. It represents the number of pixels in the true box that are not covered by the predicted box or the number of samples that are incorrectly predicted as negative by the model.

(2) IoU is one of the pixel-level metrics used to evaluate the performance of image segmentation models. It is used to evaluate the ability of an approach to describe the contour. It represents the ratio of the intersection area to the union area between the predicted segmentation results and the true segmentation results. It is calculated by Equation (11).

$$IoU = TP/(TP + FP + FN) \tag{11}$$

(3) $nIoU$ is a normalized version of IoU and is commonly used to calculate the average degree of overlap between multiple predicted boxes and multiple true boxes. It is the index obtained by calculating the intersection over union (IoU) for each class and then averaging the values. nIoU is an evaluation index specifically designed for infrared dim- and small-target detection. It can be calculated based on Equation (12):

$$nIoU = \left(\sum_{i=1}^{N} IoU_i\right)/N \tag{12}$$

where N represents the number of classes, and IoU_i represents the intersection over union (IoU) value of the ith class.

3.2. Implementation Details

Two datasets, IRSTD-1k [34] and SIRST [25], are selected for experimental verification. IRSTD-1k is a single-frame infrared small-target dataset. The dataset contains 1000 samples with various backgrounds, distances, and types of small targets, as shown in Figure 5. SIRST contains 427 samples and 480 small-target instances, many of which are blurred and hidden in complex backgrounds such as sky or water, as shown in Figure 6. In real infrared scenes, the network's performance may be limited by differences in image sizes and the restricted amount of data available, making it susceptible to overfitting and convergence failure. In this case, it is necessary to design the network structure specifically to enhance the generalization ability and stability of the model. Therefore, according to the characteristics of the segmentation network designed in this paper, the size of the image in the dataset at the input of the network is fixed at 256 × 256. The training and test sets are divided based on a ratio of 8:2, as shown in Table 1.

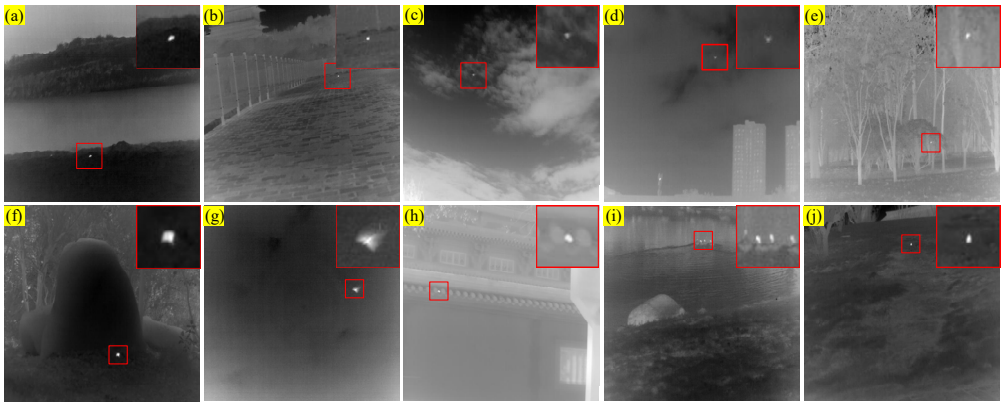

Figure 5. The IRSTD-1k [34] dataset; (**a**–**j**) are examples of images in the dataset.

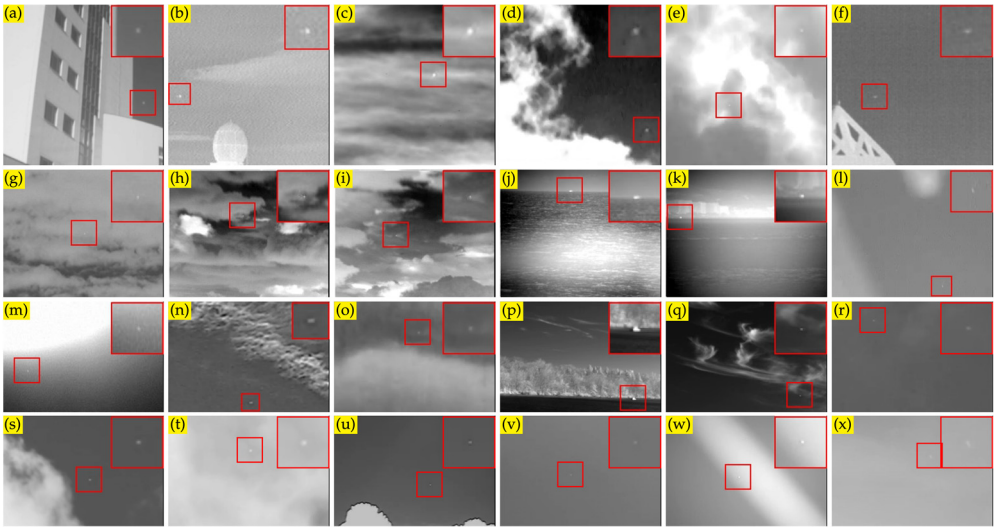

Figure 6. The SIRST [25] dataset; (**a**–**x**) are examples of images in the dataset.

Table 1. Dataset partitioning details.

	IRSTD1k	**SIRST**
Training set	800	341
Test set	201	86
Sum	1001	427

In the implementation process, the optimizer of the proposed method utilizes stochastic gradient descent (SGD) with a momentum of 0.9. The initial learning rate is set to 0.05, the weight decay coefficient is 0.0004, the batch size is 8, and the number of epochs is 300. The loss function employed is soft IoU, and its value is calculated as shown in Equation (13).

$$SoftIoU = 1 - IoU \tag{13}$$

The proposed model is compared with five data-driven methods, namely, FPN [22], U-Net [23], TBC-Net [24], ALC-Net [25], and DNA-Net [26]. The parameter settings are the same as those in the original paper. Table 2 shows the parameter settings involved in various methods.

Table 2. Parameter settings for data-driven methods.

Methods	Parameter Setting		
	Learning Rate	Batch Size	Epochs
FPN	0.02	-	-
U-Net	-	-	-
TBC-Net	0.005	128	130
ALC-Net	0.1	10	400
DNA-Net	0.05	16	1500

3.3. Ablation Experiment

In this section, the ablation test is designed to verify the rationality and effectiveness of each proposed or added module.

Compared with the original FPN [22], the PSC module and DAF module were added based on it, with both modules being added simultaneously. Experiments are carried out on two datasets: IRSTD-1k and SIRST. The nIoU and F1 scores obtained from the experiment are displayed in Figure 7. Here, "None" refers to the original FPN network, "+PSC" denotes the addition of the PSC module to the original FPN network, "+DAF" indicates the integration of the DAF module into the original FPN network, and "+PSC + DAF" represents the inclusion of both modules.

It can be seen from Figure 7 that both PSC and DAF modules have excellent performance on the two datasets. Compared with the original FPN network, the nIoU and F1 values of the PSC module and DAF module have significantly improved. This suggests that both modules contribute to enhancing detection performance. After adding two modules simultaneously, the values of nIoU and F1 are higher than when adding each module separately. This suggests that the best detection results are achieved only when both modules are used together.

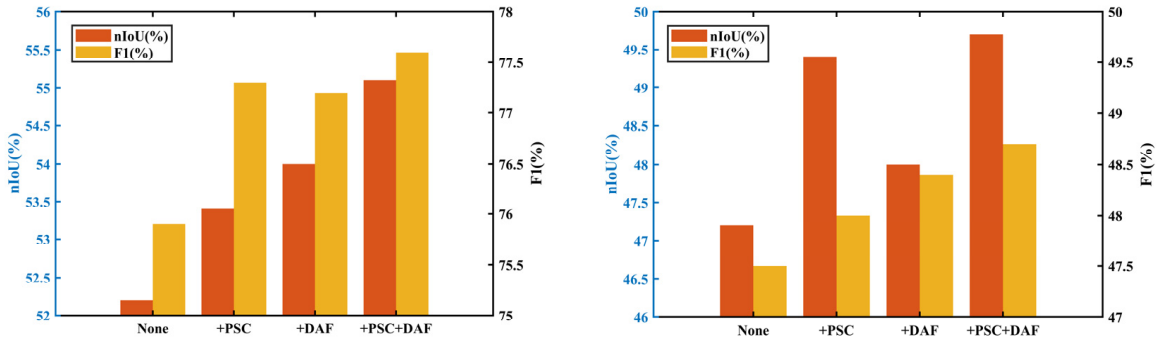

Figure 7. The performance of PSC and DAF modules on both datasets. (**a**) IRSTD-1k, (**b**) SIRST.

Secondly, the proposed algorithm is compared with AGPCNet as proposed in [28]. Compared with AGPCNet, the proposed approach incorporates three PSC modules into

the overall network. These modules are designed to repeatedly process shallow features, enabling a profound fusion of local and global features. This approach is more beneficial for feature representation. In addition, a DAF module is proposed to effectively integrate shallow features and deep features to emphasize important semantic information while downplaying or disregarding unimportant semantic information. The experimental results of the two models on the two datasets are shown in Figure 8. The notation "+PSC" denotes adding a single PSC module in the network, specifically in the module where the output is at position Z_3, as shown in Figure 1. In contrast, "+3PSC" indicates the inclusion of three PSC modules in the network.

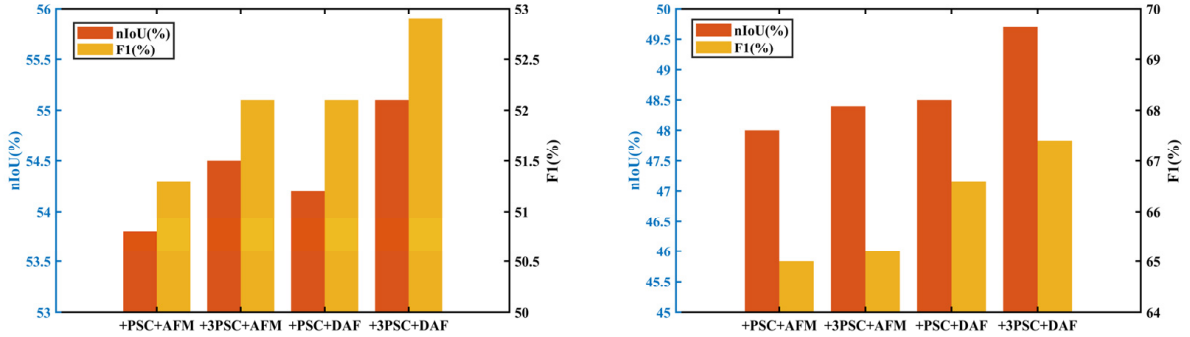

Figure 8. The AGPCNet model compared with the approach proposed in this paper. (**a**) IRSTD-1k, (**b**) SIRST.

As depicted in Figure 8, the quantity of PSC modules utilized has a specific influence on the experimental outcomes. The utilization of three PSC modules is superior to using just one, suggesting that employing PSC modules can preserve more intricate low-level information like position and texture without generating excessive redundant data. However, this performance does not imply that adding more PSC modules to the detection model will always result in better outcomes. It needs to be analyzed with respect to specific problems in practical applications. The processing effect of the DAF module proposed in this paper is better than that of AFM, which indicates the effectiveness of the DAF module proposed in this thesis.

3.4. Comparative Experiment

The performance of the proposed approach is compared with that of several other state-of-the-art methods for small-target detection. In order to accurately illustrate the effectiveness of the proposed method, its detection accuracy is quantitatively compared with that of the other state-of-the-art methods. Table 3 presents the values of the accuracy measures of the six methods on the two datasets. Among them, the best result for a measure is highlighted in bold, and the second place is marked with a horizontal line.

Table 3. The detection accuracy obtained with each tested method.

Method	IRSTD-1k				SIRST			
	IoU (%)	nIoU (%)	F1 (%)	AUC (%)	IoU (%)	nIoU (%)	F1 (%)	AUC (%)
FPN [22]	48.36	48.77	62.41	76.88	42.38	43.96	60.50	69.31
U-Net [23]	49.98	49.35	68.95	83.22	44.81	43.26	62.93	72.67
TBC-Net [24]	51.48	52.33	70.25	83.87	47.23	46.58	65.31	77.34
ALC-Net [25]	52.63	52.21	76.88	84.21	46.85	47.60	66.38	75.36
DNA-Net [26]	50.66	51.74	71.22	83.56	38.12	37.55	43.20	65.03
Proposed	54.63	55.26	77.69	84.33	48.32	49.71	67.42	76.48

As shown in Table 3, there is minimal difference in performance between the FPN method and the U-Net method on the two datasets. The performance of the U-Net method surpasses that of the FPN method, suggesting that while the introduction of the FPN method marks a milestone for infrared multi-scale small-target detection methods, the U-Net network still maintains a competitive edge in infrared small-target detection due to its encoder–decoder "U"-shaped network. The TBC-Net model incorporates semantic constraint information from high-level classification tasks, effectively addressing the imbalance issue between small targets and backgrounds. The ALC-Net model utilizes a bottom-up attention mechanism to effectively preserve the features of small targets. The DNA-Net model exhibited poor performance in this experiment. The method proposed in this paper demonstrates the best performance compared to all other methods. Although its AUC value on the SIRST dataset is not the highest, the method proposed in this paper achieves the best overall performance and is the most effective of all tested methods for infrared multi-scale small-target detection.

4. Conclusions

In this paper, an infrared multi-scale small-target detection method based on a feature pyramid network is proposed. The proposed approach incorporates two new components into the ResNet18 architecture. Firstly, three PSC modules are added to the feature pyramid network to obtain features that can enhance the detection ability. Secondly, the DAF module is utilized to improve the detection accuracy.

The infrared small-target detection architecture proposed in this paper is more accurate and robust compared to other methods. It is capable of retaining the location information of small targets while extracting deep semantic information, resulting in higher detection accuracy and lower false-alarm rates. However, it requires high GPU and memory resources and may not be suitable for resource-constrained environments. As the use of infrared small-target detection technology increases in detecting and tracking targets at long distances or under low-light conditions, there is a need to continuously improve the speed and accuracy of model detection to meet the requirements of real-world scenarios in the future.

Author Contributions: Conceptualization, S.S. and Y.S.; methodology, S.S.; software, S.S.; validation and formal analysis, S.S.; investigation, S.S.; resources, S.S.; data curation, S.S.; writing—original draft preparation, S.S.; writing—review and editing, Y.S.; visualization, S.S.; supervision, Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jiangsu University of Science and Technology grant 1132921208.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data in this research can be accessed at the following links: https://github.com/RuiZhang97/ISNet and https://github.com/YimianDai/sirst (accessed on 14 May 2024).

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

References

1. Zabłocki, M.; Gościewska, K.; Frejlichowski, D.; Hofman, R. Intelligent video surveillance systems for public spaces—A survey. *J. Theor. Appl. Comput. Sci.* **2014**, *8*, 13–27.
2. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Philip, C. Maxmean and max-median filters for detection of small targets. *Proc. SPIE* **1999**, *3809*, 74–83.
3. Tom, V.T.; Peli, T.; Leung, M.; Bondaryk, J.E. Morphology-based algorithm for point target detection in infrared backgrounds. In Proceedings of the Signal and Data Processing of Small Targets, Orlando, FL, USA, 22 October 1993; pp. 2–11.
4. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2010**, *43*, 2145–2156. [CrossRef]
5. Kingsbury, N.; Magarey, J. Wavelet transforms in image processing. In *Signal Analysis and Prediction. Applied and Numerical Harmonic Analysis*; Birkhäuser: Boston, MA, USA, 1998; pp. 27–46.

6. Long, I.L. Weak and small object detection based on wavelet multi-scale analysis and fisher algorithm. *J. Infrared Millim. Waves* **2003**, *22*, 353–356.
7. Zhang, X.; Li, L.; Xin, Y. Adaptive Multimode Infrared Dim and Small Target Detection Based on Wavelet Transform. *Laser Infrared* **2017**, *47*, 647–652.
8. Chen, C.L.P.; Li, H.; Wei, Y. A local contrast method for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 574–581. [CrossRef]
9. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for infrared small target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]
10. 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*, 612–616. [CrossRef]
11. Han, J.; Moradi, S.; Faramarzi, I.; Liu, C.; Zhang, H.; Zhao, Q. A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1822–1826. [CrossRef]
12. Han, J.; Moradi, S.; Faramarzi, I.; Honghui, Z.; Zhao, Q.; Zhang, X.; Nan, L. Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1670–1674. [CrossRef]
13. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Terahertz Sci. Technol.* **2013**, *22*, 4996–5009. [CrossRef] [PubMed]
14. Guo, J.; Wu, Y.; Dai, Y. Small target detection based on reweighted infrared patch-image model. *IET Image Process.* **2018**, *12*, 70–79. [CrossRef]
15. Zhang, L.; Peng, L.; Zhang, T. Infrared small target detection via non-convex rank approximation minimization joint l2, 1 norm. *Remote Sens.* **2018**, *10*, 1821. [CrossRef]
16. Zhang, T.; Peng, Z.; Wu, H. Infrared small target detection via self-regularized weighted sparse model. *Neurocomputing* **2021**, *420*, 124–148. [CrossRef]
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
19. Zuo, Z.; Tong, X.; Wei, J.; Su, S.; Wu, P.; Guo, R.; Sun, B. AFFPN: Attention fusion feature pyramid network for small infrared target. *Remote Sens.* **2022**, *14*, 3412. [CrossRef]
20. Huang, L.; Dai, S.; Huang, T.; Huang, X.; Wang, H. Infrared small target segmentation with multiscale feature representation. *Infrared Phys. Technol.* **2021**, *116*, 103755. [CrossRef]
21. Zhao, T.; Wu, X. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3080–3089.
22. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI, Munich, Germany, 5–9 October 2015; pp. 234–241.
24. Zhao, M.; Cheng, L.; Yang, X.; Feng, P.; Liu, L.; Wu, N. TBC-Net: A real-time detector for infrared small target detection using semantic constraint. *arXiv* **2019**, arXiv:2001.05852.
25. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *59*, 9813–9824. [CrossRef]
26. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **2022**, *32*, 1745–1758. [CrossRef]
27. Chen, Y.; Li, L.; Liu, X.; Su, X. A multi-task framework for infrared small target detection and segmentation. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 5003109. [CrossRef]
28. 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*, 4250–4261. [CrossRef]
29. Kou, R.; Wang, C.; Peng, Z.; Zhao, Z.; Chen, Y.; Han, J.; Huang, F.; Yu, Y.; Fu, Q. Infrared small target segmentation networks: A survey. *Pattern Recognit.* **2023**, *143*, 109788. [CrossRef]
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
32. Zhang, Q.L.; Yang, Y.B. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.

33. 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*, 3200. [CrossRef]
34. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape Matters for Infrared Small Target Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 877–886.