remote sensing

Article

CSAN-UNet: Channel Spatial Attention Nested UNet for Infrared Small Target Detection

Yuhan Zhong, Zhiguang Shi *, Yan Zhang, Yong Zhang and Hanyu Li

School of Electronic Science, National University of Defense Technology, Changsha 410073, China; zhongyuhan@nudt.edu.cn (Y.Z.); zyyan@nudt.edu.cn (Y.Z.); yongz@nudt.edu.cn (Y.Z.); lihy@nudt.edu.cn (H.L.)

* Correspondence: szg@nudt.edu.cn

Abstract: Segmenting small infrared targets presents a significant challenge for traditional image processing architectures due to the inherent lack of texture, minimal shape information, and their sparse pixel representation within images. The conventional UNet architecture, while proficient in general segmentation tasks, inadequately addresses the nuances of small infrared target segmentation due to its reliance on downsampling operations, such as pooling, which often results in the loss of critical target information. This paper introduces the Channel Spatial Attention Nested UNet (CSAN-UNet), an innovative architecture designed specifically to enhance the detection and segmentation of small infrared targets. Central to CSAN-UNet's design is the Cascaded Channel and Spatial Convolutional Attention Module (CSCAM), a novel component that adaptively enhances multi-level features and mitigates the loss of target information attributable to downsampling processes. Additionally, the Channel-priority and Spatial Attention Cascade Module (CPSAM) represents another pivotal advancement within CSAN-UNet, prioritizing channel-level adjustments alongside spatial attention mechanisms to efficiently extract deep semantic information pertinent to small infrared targets. Empirical validation conducted on two public datasets confirms that CSAN-UNet surpasses existing state-of-the-art algorithms in segmentation performance, while simultaneously reducing computational overhead.

Keywords: infrared small target segmentation; UNet; channel and spatial attention

Citation: Zhong, Y.; Shi, Z.; Zhang, Y.; Zhang, Y.; Li, H. CSAN-UNet: Channel Spatial Attention Nested UNet for Infrared Small Target Detection. *Remote Sens.* **2024**, *16*, 1894. https://doi.org/10.3390/rs16111894

Academic Editor: Pedro Melo-Pinto

Received: 15 April 2024
Revised: 17 May 2024
Accepted: 20 May 2024
Published: 24 May 2024

1. Introduction

The detection of small infrared targets is an essential capability in critical applications such as maritime surveillance [1,2] and precision guidance systems [3,4]. These systems rely on the ability to accurately identify and track objects of interest under varying conditions, including during nighttime operations or through visual obstructions such as rain and fog. The unique challenges posed by the fast movement of sensor platforms, combined with the inherently small size of targets that often lack discernible texture and shape information, underscore the complexity of this task and lead to a decrease in the detection performance of spatiotemporal methods. Such targets are not only difficult to detect due to their minimal pixel presence but also because they frequently blend into their backgrounds or are obscured by environmental factors. Therefore, single-frame infrared small target (SIRST) detection becomes particularly important [5].

In the realm of single-frame infrared small target detection, traditional methods have been the cornerstone, employing model-driven approaches grounded in assumptions about the target's physical and imaging characteristics. These methods fall into three primary categories: filter-based methods [6–8], which aim to enhance the visibility of targets by predicting and subtracting the background; local information-based methods [9–15], which detect targets based on the premise that the gray value and intensity of objects and their surroundings undergo sudden changes; and data structure-based methods [16–21], which approach the detection task as a problem of sparse low-rank tensor decomposition, identifying targets through the analysis of structural data. However, these traditional methods exhibit significant limitations, notably their reliance on handcrafted features designed based on prior knowledge of target appearances, and the need for extensive parameter tuning to adapt to the expected characteristics of the targets. Furthermore, their poor adaptability to scene variability—manifesting as changes in target size, shape, signal-to-clutter ratio, and clutter background—severely impedes their performance. The fixed parameter settings inherent in these methods often result in suboptimal detection outcomes, as they cannot adequately adjust to the dynamic conditions encountered in practical applications.

The shift from traditional model-driven methods to Convolutional Neural Network (CNN)-based approaches in the field of single-frame infrared small target detection marks a pivotal evolution towards data-driven methodologies. This transition enables the extraction and learning of features directly from data, free from the constraints of predefined assumptions and handcrafted features. A significant milestone in this shift was achieved by Wang et al. [22], who released the first substantial dataset for single-frame infrared small target segmentation, spurring further research and innovation. Following this, a series of innovative CNN-based approaches emerged, each contributing novel solutions to enhance detection capabilities. Among these, Dai et al. [23] introduced the Asymmetric Context Modulation (ACM) to refine the Unet architecture by modifying its skip connections, accompanied by the launch of a high-quality SIRST dataset. The Attentional Local Contrast Network (ALC-Net) [24] innovated with a tensor shift operation aimed at more effectively extracting local target information. The Attention-Guided Pyramid Context Networks (AGPC-Net) [25] leveraged the ResNet34 backbone for robust feature extraction and incorporated attention blocks to integrate contextual information seamlessly. Zhang et al. [26] proposed Taylor Finite Difference (TFD)-inspired techniques, integrating an edge detection block and a two-orientation attention aggregation block to enhance target feature extraction amidst clutter and establishing the IRSTD-1k dataset as a benchmark. Furthermore, DNANet [27] introduced densely nested interaction modules (DNIM), fostering essential interactions between high-level and low-level features to preserve target details against cluttered backgrounds. These advancements collectively represent significant strides in improving the precision and effectiveness of small infrared target detection, demonstrating the potent capabilities of innovative architectural enhancements, attention mechanisms, and feature interaction modules in advancing detection technology.

The Channel Spatial Attention Nested UNet (CSAN-UNet) introduces an innovative architecture tailored for the detection of small infrared targets, drawing on the success of U-Net in medical imaging [28–31] and the application of hybrid attention mechanisms [32,33] in object detection. Central to CSAN-UNet are two innovative components: the Cascaded Channel and Spatial Convolutional Attention Module (CSCAM) and the Channel-Prior and Spatial Attention Cascade Module (CPSAM). CSCAM is designed to adaptively enhance target features at various levels, effectively pinpointing target locations while minimizing background clutter through meticulous channel and spatial attention. Concurrently, CPSAM leverages depth-wise convolution with strip convolution kernels of assorted scales to efficiently map spatial relationships, ensuring the precise extraction of target information with a lower computational footprint. By integrating CSCAM and CPSAM, CSAN-UNet adeptly balances computational resource efficiency with the preservation of detailed target features, adeptly navigating the challenge of maintaining small target details without imposing excessive computational demands. This architecture is further distinguished by a comparative analysis with existing methods, where the exclusive use of the Squeeze-and-Excitation (SE) module is critiqued for its lack of spatial attention—a critical element for effectively emphasizing target regions. Moreover, dependence solely on CSCAM or CPSAM introduces its own challenges; CSCAM alone raises computational demands significantly due to its elaborate processing, while exclusive reliance on CPSAM could result in the loss of small object features during downsampling operations. This underscores the necessity for a balanced approach, which CSAN-UNet successfully provides, offering a comprehensive solution for the nuanced task of infrared small target detection.

The main contributions of this work can be summarized as follows:

(1) We propose CSAN-UNet, an innovative architecture tailored for precise segmentation of small infrared targets, optimizing computational efficiency.
(2) We introduce a Cascaded Channel and Spatial Convolutional Attention Module (CSCAM) for improved feature enhancement during downsampling, preserving crucial target details.
(3) We develop a lightweight Channel-Priority and Spatial Attention Cascade Module (CPSAM) for efficient extraction of target semantic information with minimal computational demand.
(4) We demonstrate CSAN-UNet's superior performance and efficiency through rigorous testing on two public datasets, surpassing existing state-of-the-art solutions.

2. Related Work

2.1. Single-Frame Infrared Small Target Detection

Traditional model-driven methods for infrared small target detection, including local-information-based, filtering-based, and data structure-based approaches, have established the foundation for addressing this intricate task. However, these methods are heavily reliant on manually crafted features, predefined based on theoretical assumptions about target characteristics and environmental conditions. This reliance significantly limits their adaptability and effectiveness, particularly in complex real-world scenarios with prevalent background clutter and noise. Consequently, despite their foundational role, traditional model-driven methods face considerable challenges in detecting infrared small targets due to their inability to dynamically adapt to the complexities of natural environments.

CNN-based methods have significantly outperformed traditional model-driven approaches in detecting infrared small targets due to their superior feature representation capabilities, which are honed by learning directly from data. The advent of high-quality public datasets has further propelled the effectiveness of CNN-based detection methods by enabling more robust training and validation. Among the innovations in this space, AFFPN [34] stands out for its use of dilated convolution and adaptive pooling layers to generate multi-scale feature maps. Concurrently, Wang et al.'s employment of generative adversarial networks (GANs) [22] aims to reduce missed detections and false positives, enhancing model accuracy in infrared object detection. Similarly, Zhao et al. [35] have crafted a UNet-based GAN architecture, incorporating the generative adversarial principle within the established UNet framework to refine detection capabilities. Hou et al. [36] have blended CNN methodologies with traditional handcrafted features, creating a potent detection network for infrared small targets. Li et al. [27] introduced DNANet, featuring dense nested interaction networks that facilitate the progressive interaction between high-level and low-level features for comprehensive feature extraction. Wu et al. [37] incorporates a hybrid Vision Transformer and Convolutional Neural Network encoder to effectively handle the unique challenges of infrared small target detection by extracting multi-level features and capturing long-distance dependencies. Ying et al. [38] proposed the Label Evolution with Single Point Supervision (LESPS) framework for infrared small target detection, which reduces the need for extensive per-pixel annotations by using point-level supervision. Wu et al. [39] introduced the U-Net in U-Net (UIU-Net) framework to address the challenges of tiny object loss and feature distinguishability in deep neural networks for infrared small object detection.

Despite achieving commendable detection results, these methods fail to find an optimal balance between detection performance and the computational resources required, indicating a need for more efficient architectures or optimization techniques to enhance overall efficacy.

2.2. Attention Mechanisms and Feature Fusion

Attention mechanisms: Attention mechanisms have emerged as a pivotal innovation in deep learning, significantly enhancing network performance across various domains by enabling models to selectively focus on important data portions. This adaptive focusing capability has found extensive application in areas ranging from object detection and semantic segmentation to natural language processing. Particularly notable is the adoption of self-attention within Transformer architectures, which has become popular for its ability to process different segments of input data regardless of their positional relationship, thereby greatly improving the model's interpretative and processing efficiency. Wu et al.'s ViT-based Multi-level Feature Extraction (MVTM) module [37] further exemplifies the versatility of attention mechanisms, showing how attention can be extracted from image feature maps using CNN networks, with the Vision Transformer architecture facilitating nuanced multi-level feature attention. Meanwhile, the SENet architecture, as introduced by Hu et al. [40], demonstrates the potential of channel attention by automatically assessing the relevance of each channel to highlight key object features, thus refining the model's focus. Li et al. [41] proposed the adaptive local block aggregation (ALBA) module to enhance features. By constructing the relationship between a single block and its local blocks, it identifies similarities and creates feature blocks, aggregating them to mine local contextual information. Additionally, spatial attention methods [42] emphasize the importance of focusing on critical areas within the spatial expanse of feature maps, allowing networks to better emphasize and process essential regions within the input data. Collectively, these developments underscore the transformative impact of attention mechanisms on deep learning, allowing models to process information with unprecedented selectivity and depth.

Feature fusion: The emergence of U-Net [28] and Feature Pyramid Network (FPN) [43] has significantly advanced remote sensing image processing. Using the summation or concatenation between different levels, it is possible to achieve better results between shallow and deep networks. Li et al. [27] proposed DNIM to fuse deep semantic features with shallow detail features, incorporating an attention mechanism between different layers to enhance fusion performance. Li et al. [44] introduced the multi-scale information interaction (MII) unit, modeled between the two architectures through the direction-aware spatial context aggregation (DSCA) module. The learnable feedback compensation correction (LFCC) is constructed by combining the content generated by the multi-scale factor model to form the interaction feedback joint optimization mode, achieving better multi-scale feature fusion optimization. To enhance the effectiveness and efficiency of feature fusion, Zhang et al. [34] proposed AFM to consolidate low-level and deep-level semantics.

3. Proposed Method

In this section, we introduce the overall structure and main component modules of our CSAN-UNet in detail.

3.1. Network Architecture

As shown in Figure 1, the Channel Spatial Attention Nested UNet (CSAN-UNet) is meticulously designed for detecting small infrared targets within single-frame images, leveraging a robust architecture split into two main components: feature extraction and feature fusion. Initially, feature maps are generated from the input image using a backbone network, typically a ResNet18, ensuring robust initial feature extraction. This phase is augmented by attention-based modules such as the Cascaded Channel and Spatial Convolutional Attention Module (CSCAM), which enhances features post-downsampling to preserve critical small target information that may otherwise be lost. Additionally, the Channel-Prior and Spatial Attention Cascade Module (CPSAM) enriches the semantic information of targets during the fusion stage, effectively bridging the semantic gaps between different feature scales. The feature fusion process involves employing skip connections to a decoder where a sophisticated multi-layer fusion of features occurs. This process includes upsampling to align feature map sizes and the integration of shallow features, which are rich in spatial details, with deep features, which contain dense semantic information, through concatenation and convolution operations. Furthermore, the Feature Pyramid Fusion Module performs multi-scale fusion, creating robust feature maps that integrate both detailed and semantic information effectively. Through these advanced attention mechanisms and fusion strategies, CSAN-UNet optimizes the detection and segmentation of small infrared targets, handling the challenges of information loss and enhancing semantic clarity with high efficiency.

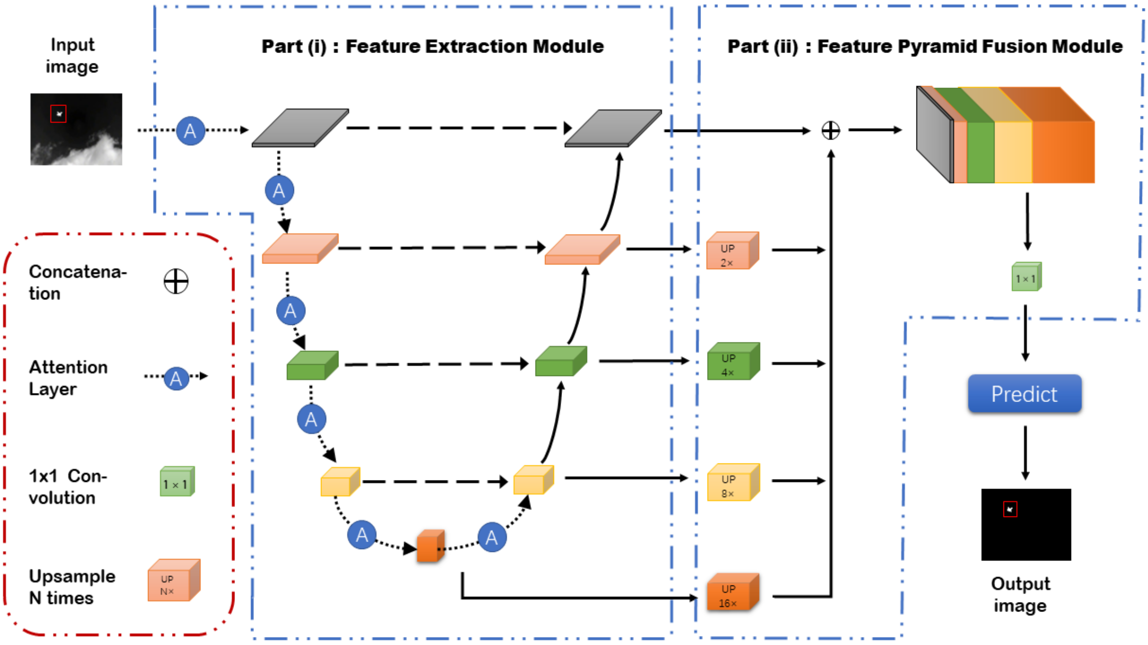


Figure 1. The diagram illustrates the structured process of the Channel Spatial Attention Nested UNet (CSAN-UNet), a deep learning model engineered for precise segmentation of small infrared targets. Starting with the initial input, images are first processed through a residual module, which performs crucial downsampling and feature extraction, preparing the image for subsequent enhancements. Following this, the image is further refined through spatial pyramid pooling and attention-based modules. These components are adept at selectively enhancing features at various levels, focusing on areas that require intricate attention to detail. The enhanced features are then seamlessly integrated using a feature fusion module, which upsamples and concatenates the multi-layer features, creating a unified and comprehensive feature set. The final step in the process is carried out by the prediction module, which utilizes the fused features to generate the segmentation results, accurately identifying and delineating the target areas within the image. This diagram encapsulates the entire workflow of CSAN-UNet, highlighting its capability to handle the complex task of segmenting small infrared targets with high precision.

3.2. The Attention-Based Feature Enhancement Layer

The traditional UNet architecture, often employing a ResNet encoder for multi-scale feature extraction, faces significant challenges in detecting small infrared targets. These targets, sometimes as minuscule as a single pixel, are particularly vulnerable to information loss or being obscured by background clutter during pooling operations, an issue that is exacerbated as the receptive field expands during the feature extraction process. Additionally, the traditional UNet approach encounters a semantic gap in feature fusion, where low-level features extracted by the encoder are passed through skip connections to the decoder and fused with high-level features. This fusion can create discrepancies due to the differing detail and abstraction levels of the features being integrated. To overcome these issues, Li et al. introduced DNANet, which features dense nested interaction networks that facilitate the progressive interaction between high-level and low-level features for comprehensive feature extraction. This approach allows for a more dynamic integration of features across the network, potentially reducing the semantic gaps identified in traditional UNet models. However, despite achieving commendable detection results, DNANet and similar enhancements still struggle to find an optimal balance between detection performance and the computational resources required. To combat these issues, a proposed solution incorporates an Attention-based Feature Enhancement Layer (Figure 2), which includes components such as the CPCAM and CSCAM. CPCAM is designed to preserve critical information throughout the processing pipeline, especially after pooling operations, while CSCAM aims to bridge the semantic gap between low and high-level features. This ensures that details pertinent to small targets are not lost during upsampling and fusion, refining the traditional UNet framework to enhance its efficacy in handling small infrared target detection by maintaining important features and ensuring their effective integration across the network's different levels.

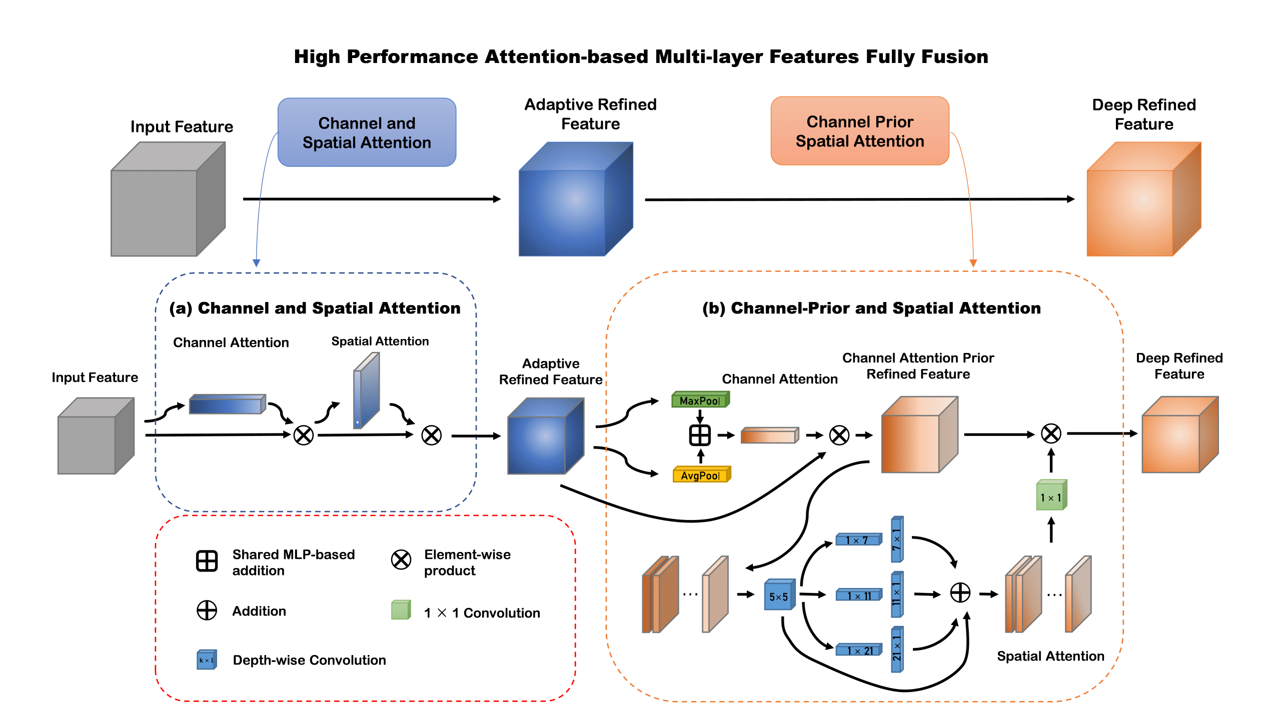


Figure 2. The Attention-based Feature Enhancement Layer. CSCAM is used to compensate for the loss of small target position information caused by downsampling operations. CPSAM is used to reduce the semantic gap at the multi-layer feature fusion stage.

3.2.1. CPSAM

Li et al. [27] developed a Densely Nested Interaction Module (DNIM), utilizing both channel and spatial attention mechanisms, to foster progressive interactions between high-level and low-level features. This innovative approach aims to bridge the semantic gap and preserve the integrity of features associated with small infrared targets. The effectiveness of this attention mechanism in enhancing infrared small target detection capabilities has been demonstrated, underscoring its potential in advanced imaging applications. However, the complex, densely nested nature of these attention mechanisms results in a high computational load, making the model quite resource-intensive. Consequently, there is a clear necessity for a more lightweight, yet equally effective, attention module to reduce the computational demands. Additionally, the uniform distribution of spatial attention weights across channels in DNIM poses a challenge, as it can introduce noise that potentially degrades detection performance. This highlights the need for further refinement in attention weight distribution to optimize the detection of small infrared targets without compromising on computational efficiency.

Therefore, we propose a lightweight Channel-Prior and Spatial Attention Cascade Module (CPSAM, Figure 2b). CPSAM allows for dynamic allocation of attention weights in both channel and spatial dimensions. By incorporating a multi-scale depth-wise convolutional module, CPSAM effectively captures spatial relationships while preserving the channel prior. Given the Feature map $M_{enhan} \in \mathbb{R}^{C\times H\times W}$ that has been adaptively enhanced by CSCAM, the CPSAM processed it by a channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ and a spatial attention map $M_s \in \mathbb{R}^{1\times H\times W}$. The overall process can be summarized as:

$$M'_{refin} = M_c \otimes M_{refin}. \tag{1}$$

$$M''_{refin} = M_s \otimes M'_{refin}. \tag{2}$$

where $\otimes$ denotes the element-wise multiplication.

The channel attention map is inferred through the channel attention module. The Channal attention module first performs maximum pooling and mean pooling operations on the input Feature map to generate two spatial context descriptors. These two descriptors containing spatial context information are fed into a shared multi-layer perceptron, and the results of the two are then added to obtain a channel attention map that incorporates contextual information. The computation of the channel attention can be summarized as:

$$M_c = \sigma(MLP(AvgPool(M_{refin}) + MLP(MaxPool(M_{refin})))). \tag{3}$$

where σ denotes the sigmoid function.

The spatial attention module integrates spatial relationship information to generate a spatial attention map. To avoid introducing noise into the final enhanced feature map due to consistent spatial attention weight distribution between channels, we use depth-wise convolution to dynamically allocate attention weights in both channel and spatial dimensions and reduce computational complexity. The computation of the spatial attention can be summarized as:

$$M_s = Conv_{1\times 1}(\sum_{i=0}^{3} Bran_i(DepWConv(M'_{refin}))). \tag{4}$$

where $DepWConv$ denotes depth-wise convolution and $Bran_i$, $i \in \{0,\ 1,\ 2,\ 3\}$ represents the i-th branch.

3.2.2. CSCAM

Although CPSAM can efficiently extract deep semantic features of small infrared targets, experimental results show that it cannot well compensate for the loss of small target position information caused by downsampling operations. Our ablation study for the CSCAM in Section 4.4 demonstrates this well. We believe that this is because the multi-scale characteristics of CPSAM make it easier for unenhanced small targets to be submerged in the clutter background. The long convolution kernel of the multi-scale convolution kernel (the longest is 21 × 1) allows small target features to interact with farther locations and has a larger receptive field. However, in the downsampled feature map, small target features are not strong, and large-scale interactions often make it easier for small target features to be submerged in clutter background features. Therefore, we propose a cascaded channel and spatial Convolutional Attention Module (CSCAM, Figure 2a), which uses the classic residual convolution structure and cascades the channel attention module and the spatial attention module to adaptively enhance infrared small target information. Its overall process and the calculation process of channel attention map M_c are the same as CPSAM and follow the formulas (1)–(3). The difference is the calculation process of the spatial attention map M_S, which can be summarized as:

$$Ms = \sigma\{f^{7\times 7}[AvgPool(M^{'}_{refin}),\ MaxPool(M^{'}_{refin})]\}. \tag{5}$$

where $f^{7\times 7}$ represents a convolutional operation with the filter size of 7 × 7 and [A, B] represents the concatenation of A and B.

3.3. The Feature Fusion Module

The Feature Fusion Module addresses the challenges associated with detecting small infrared targets, which range from single-pixel point targets to larger entities covering dozens of pixels. Traditional neural network architectures such as UNet often lose critical information during the downsampling process, especially as the network depth increases and feature maps are downscaled, leading to a loss of details pertinent to small targets. To overcome these limitations, the Pyramid Feature Fusion Module (Figure 1ii) is introduced. This module adeptly handles multi-scale representations, which significantly enhances the network's contextual understanding and detection accuracy by retaining important cross-scale information. The module employs an upsampling operation to resize feature maps from various network layers to a common size, improving the capture of semantic depth from deeper layers and fine details from shallower ones. By concatenating deep and shallow features, the fusion module creates rich, comprehensive feature maps that preserve critical features and spatial details essential for detecting small targets. To efficiently enhance the information content of feature maps across different layers of our neural network, we employ an upsampling operation that standardizes the dimensions of these feature maps to a common size, designated as $M_{up}^{i,\,J} \in R^{C_i \times H_0 \times W_0}$, $i \in \{0,\ 1,\ \ldots,\ I\}$, where i span from 0 to I, C_i represents the number of channels, and H_0 and W_0 are the height and width of the resized feature maps, respectively. This resizing ensures that both the deep-layer semantic information and the shallow-layer fine detail are captured effectively. By concatenating deep and shallow features, the network forms comprehensive feature maps that help prevent the loss of crucial features and ensure the effective utilization of spatial location details necessary for detecting small targets. For integrating these features into a coherent output for multi-feature fusion, denoted as $M_i^{'}$, the operation is defined mathematically by the equation:

$$M_i^{'} = UpSample^{2^i}(\theta(Normal(Conv_{3\times3}(M_i)))),\ i \in \{1,\ 2,\ 3\}. \tag{6}$$

where $UpSample(\cdot)$ denotes the upsampling operation using bilinear interpolation, with the exponent 2^i indicating the degree of upsampling which varies with the layer index. $Conv_{3\times3}(\cdot)$ refers to a convolutional operation utilizing a 3 × 3 kernel, $Normal(\cdot)$stands for group normalization to stabilize the activations across the maps, and $\theta(\cdot)$ represents the ReLU function, introducing non-linearity that enables the modeling of complex data patterns. This intricate fusion process enhances the network's capability to capture and process multi-scale and multi-level details, crucial for the effective detection of small and intricately varying targets.

4. Result of Evaluational Experiment

4.1. Evaluation Metrics

(1) Intersection over union (IoU): The Intersection over Union (IoU) metric is a fundamental tool used to evaluate the effectiveness of algorithms in semantic segmentation, focusing particularly on how accurately these algorithms can delineate object contours within an image. IoU is calculated by determining the ratio of the intersection area between the predicted labels from the segmentation algorithm and the ground truth labels, which represent the actual labels, to the union of these two areas. This calculation provides crucial insights into the precision with which a segmentation algorithm can outline and overlap with the actual boundaries of objects in an image. As such, IoU serves as an essential measure for assessing the accuracy of segmentation models, indicating the degree to which the predicted segmentation corresponds to the ground truth. IoU is defined as follows:

$$IoU = \frac{TP}{TP + FN + FP}. \tag{7}$$

where TP represents the True Positives, FN represents the False Negatives, and FP represents the False Positives.

(2) Probability of Detection: The Probability of Detection (P_d) is a metric used to evaluate the effectiveness of detection algorithms by measuring their ability to accurately identify target instances within a dataset. It is calculated as the ratio of the number of correctly predicted target instances $I_{correct}$ to the total number of actual target instances I_{all}, assessing the algorithm's accuracy in detecting present targets. This metric is crucial for gauging the reliability and effectiveness of detection systems in real-world applications, highlighting their practical utility in operational settings. Pd is calculated as follows:

$$P_d = \frac{I_{correct}}{I_{all}}. \tag{8}$$

In this paper, we determine whether a target is correctly predicted by comparing its centroid deviation to a predefined deviation threshold, Dthresh. If the centroid deviation of the target is lower than D_{thresh}, we classify it as a correctly predicted target. For this study, we have set the predefined deviation threshold to be 3.

(3) False-Alarm Rate: The False-Alarm Rate (Fa) is a critical metric used to measure the precision of a detection algorithm, quantifying how often the algorithm incorrectly identifies targets within an image. It focuses on the error aspect of detection by calculating the ratio of the number of falsely predicted pixels (N_{false}) to the total number of pixels in the image (N_{all}). This calculation helps assess the proportion of pixels incorrectly marked as targets, providing insights into the algorithm's precision and its ability to minimize false positives. The False-Alarm Rate is essential for evaluating the robustness and reliability of detection systems, emphasizing their effectiveness in avoiding incorrect target identifications. The definition of Fa is as follows:

$$F_a = \frac{N_{false}}{N_{all}}. \tag{9}$$

(4) Receiver Operation Characteristics: The Receiver Operating Characteristics (ROC) curve is a crucial tool that graphically illustrates the trade-offs between the Probability of Detection (Pd) and the False Alarm Rate (Fa) for detection algorithms. By plotting Pd against Fa, the ROC curve helps assess how effectively a detection system can identify true positives while minimizing false alarms. This visualization is critical for assessing how effectively a detection system performs in challenging environments, where small infrared targets need to be identified against complex backgrounds.

(5) Floating Point Operations: Floating Point Operations ($FLOPs$) represent the number of floating-point arithmetic operations performed by a model during its execution. In our work, $FLOPs$ are used to assess the efficiency and computational requirements of different models.

(6) Parameters: Parameters represent the learnable variables that enable the model to capture and represent patterns in the data. By analyzing the number of parameters in a model, we can gain insights into its size, memory requirements, and computational efficiency. The parameter count provides a quick and simple way to compare different models and understand their relative complexities.

(7) Frames Per Second: In our study, Frames Per Second (FPS) is utilized as an evaluation indicator to measure the speed of model inference. FPS quantifies the number of images processed by the model within a single second (N_{sec}). It provides a valuable metric to assess the efficiency and real-time performance of a model during inference. The definition of FPS is as follows:

$$FPS = \frac{1}{N_{sec}}. \tag{10}$$

4.2. Implementation Details

To thoroughly evaluate the effectiveness of our detection method for small infrared targets, we utilized two distinct datasets: SIRST [23] and IRSTD-1K [26]. The datasets are split evenly in a 1:1 ratio between training and testing sets. The preprocessing steps include normalization, random flipping, random cropping, and resizing images to 256 × 256 pixels to prepare them for input into the network.

For training, our network used the Soft-IoU loss function, optimized via the Adagrad method with momentum and weight decay set at 0.9 and 0.0004, respectively. An initial learning rate of 0.015 was applied, adjusted according to a poly decay strategy, with training conducted over 1500 epochs and 16 batches per iteration. The computational framework for our experiments was based on the PyTorch platform, executed on a high-performance computer equipped with an Intel (R) Core (TM) i9-10980XE CPU @ 3.00 GHz and an Nvidia GeForce 3080 GPU. This setup ensured robust testing and performance evaluation of the network under rigorous conditions.

4.3. Comparison with State-of-the-Art Methods

To demonstrate the superiority of CSAN-UNet over existing methods, a comprehensive evaluation was conducted using both qualitative and quantitative analyses against multiple state-of-the-art techniques. The comparison encompassed a wide spectrum of methodologies to ensure a robust assessment. Traditional Model-Driven Methods employed included filter-based approaches, such as top-hat [7] and max-median [8], along with methods inspired by the human visual system, such as WSLCM [14] and TLLSM [15], as well as data structure-based methods, such as IPI [20] and RIPI [21]. Classic Data-Driven Methods featured included general-purpose deep learning architectures, such as FPN [45], U-Net [28], GCN [46], and Exfuse [47], which, though not specifically designed for infrared target detection, are recognized for their effectiveness in various image processing tasks. Additionally, Specific Data-Driven Methods for Infrared Targets, such as MDvsFA [22], ALC [24], AGPC [25], and DNA [27], were chosen for their specialized capabilities in enhancing the detection of infrared small targets. This selection of methods provided a diverse range of traditional to modern deep learning techniques, allowing for a thorough evaluation of CSAN-UNet's performance in various scenarios and against different benchmarks.

4.3.1. Comparison with Traditional Model-Driven Methods

(1) Quantitative Results. For the traditional algorithms we compared, we followed a consistent procedure. First, we generated predictions using each method, and then we applied noise suppression by setting a threshold to eliminate low-response areas. In particular, We calculated the adaptive threshold ($T_{adaptive}$) using the following formula:

$$T_{adaptive} = Max[Max(O) \times 0.7,\ 0.5 \times \tau(O) + avg(O)]. \tag{11}$$

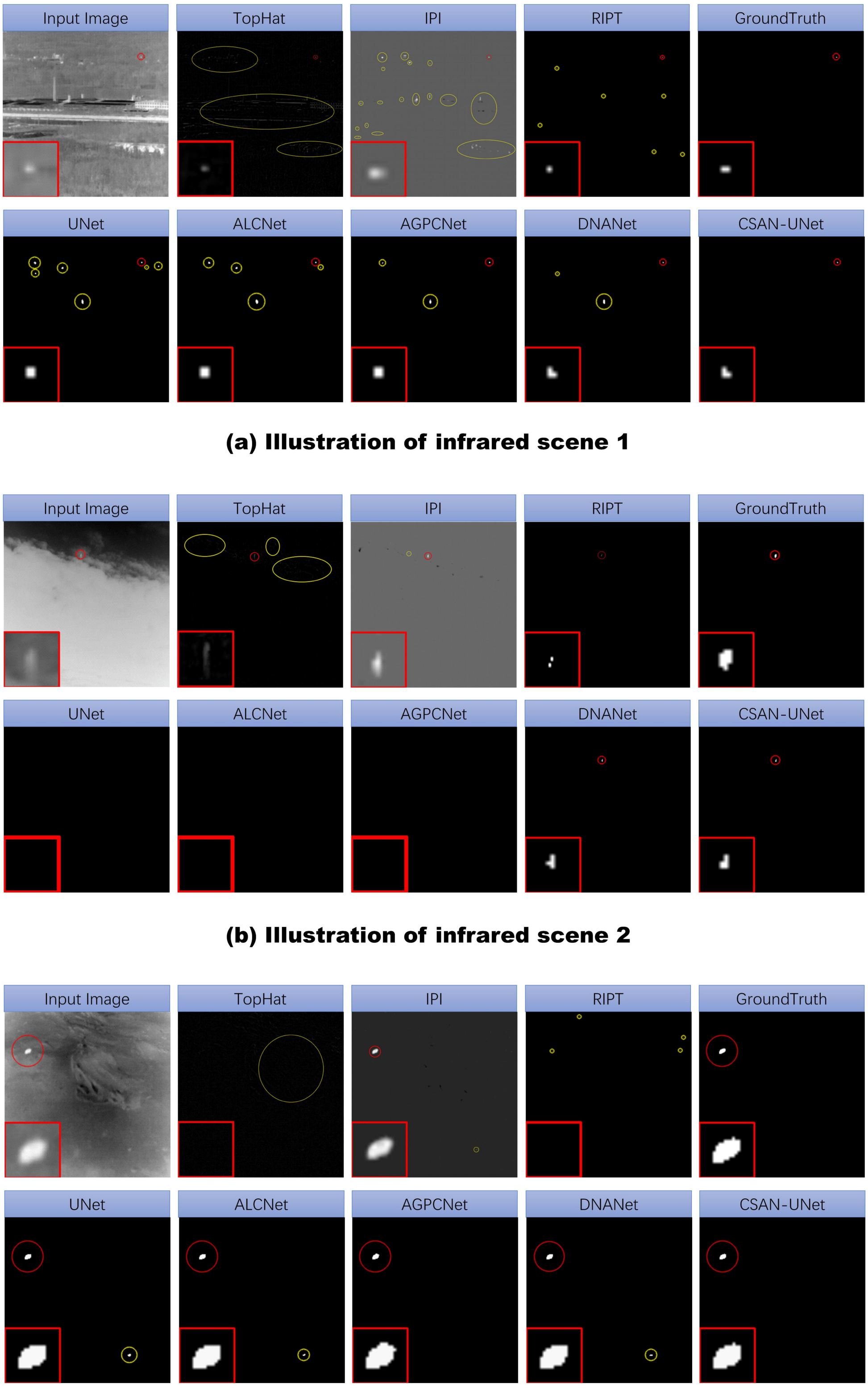
where $Max(O)$ refers to the maximum output value generated by the methods, often representing the most intense prediction, such as the highest probability or the strongest activation within the output layer, $T_{adaptive}$ represents an adaptive threshold that is dynamically determined by the outputs to optimize binary classifications, $\tau(O)$ represents the standard deviation, which measures how spread out the values are from the mean, and $avg(O)$ calculates the average value of the output.

The results from Table 1 demonstrate that CSAN-UNet significantly outperforms traditional model-driven methods in detecting and segmenting objects within the SIRST and IRSTD-1k datasets, which are known for their complex background clutter and varied object sizes and shapes. These challenging characteristics often confound traditional methods that depend on specific scenarios or target features, leading to limited adaptability due to their reliance on prior knowledge and manual parameter selection. In contrast, CSAN-UNet adopts a data-driven approach, directly learning from data, which equips it to more effectively handle the diversity of challenges presented by these datasets. This approach not only improves its performance in object detection but also in comprehensive segmentation tasks, distinguishing CSAN-UNet as a more robust and adaptable solution for intricate imaging tasks across dynamic and varied environments.

(2) Qualitative comparison (Figure 3). The comparison involves eight different detection methods, including CSAN-UNet and three traditional methods, focusing on three typical scenes of infrared small targets. Each method's results are distinctly labeled in the upper left corner of each image to facilitate identification. To improve visibility and allow for detailed examination of segmentation capabilities, the target areas within each scene are enlarged and positioned in the lower right corner of the image display. Results are color-coded for clarity: red circles indicate correctly detected targets, highlighting successful identifications, while yellow circles denote false positives, showing where methods incorrectly identified targets. The absence of red circles signifies a missed detection, indicating the method's failure to recognize an actual target present in the scene. This visualization strategy is designed to clearly demonstrate the effectiveness of each detection method, enabling a direct visual comparison that showcases their precision and accuracy in detecting and segmenting small infrared targets, thereby highlighting the strengths and weaknesses of each method.

As shown in Figure 3, the qualitative results of various detection methods on two datasets, NUAA-SIRST and IRSTD-1k, highlights significant limitations of traditional methods and the advantages of CSAN-UNet. Traditional methods struggle with adaptability to changes in target size and scene complexity, often resulting in a high incidence of false positives. This is exacerbated by issues like the sensitivity of top-hat filter-based methods to noise and the tendency of local ranking-based methods to incorrectly estimate target locations using shapeless points, leading to increased false positives (Figure 3b) and a sharp decline in performance with changes in target shape (Figure 3a). In contrast, CSAN-UNet excels in accurate localization and shape segmentation, maintaining a low rate of false positives even in complex backgrounds. Its robustness is evident as it effectively handles diverse contextual changes, unlike traditional model-driven methods that rely on handcrafted features and assumptions unsuited for dynamic environments. Overall, CSAN-UNet's advanced learning capabilities and adaptive approach make it a superior solution for detecting small targets across varied scenarios and conditions, significantly outperforming traditional methods in terms of flexibility and effectiveness.

The effectiveness of eight different detection methods in handling small infrared targets is visually presented in Figures 4–6, showcasing varied performance across methods. The top-hat method, for instance, exhibits notably high false positive rates and a low signal-to-noise ratio, as depicted in Figures 4 and 5, respectively. This method struggles particularly with cluttered backgrounds, often failing to effectively distinguish targets from surrounding interference. Other traditional methods, such as the IPI method, produce a significant number of false detections and demonstrate poor robustness in complex scenarios, while the RIPT method, which uses points to describe targets, lacks the capability to accurately determine target shapes, thereby increasing the likelihood of false detections. In stark contrast, CSAN-UNet displays superior adaptability and effectiveness in scenarios involving variable target shapes and complex backgrounds. Its strength lies in its data-driven learning approach, which, unlike traditional methods that rely on preset assumptions and manual hyperparameter settings, allows it to dynamically learn from data, thereby naturally adapting to and excelling in detecting small infrared targets. The overall assessment reveals that CSAN-UNet significantly outperforms traditional model-driven methods, particularly in challenging environments. This underscores the importance of adopting advanced machine learning techniques that can dynamically adapt to the data characteristics instead of relying on fixed, predefined models, highlighting CSAN-UNet's advantages in scenarios requiring flexible and effective detection solutions for small infrared targets.



(a) Illustration of infrared scene 1

(b) Illustration of infrared scene 2

(c) Illustration of infrared scene 3

Figure 3. The results of various infrared target detection methods are visually presented using three different scenes, where target areas are enlarged and placed in the top-left corner for clarity. Correct detections are marked with red dotted circles, and false alarms with yellow dotted circles, allowing for a quick assessment of each method’s accuracy.

Table 1. Quantitative comparison with state-of-the-art model-driven methods on the SIRST and IRSTD-1k dataset. The best results in each column are highlighted in bold.

Method	NUAA-SIRST			IRSTD-1K		
	IoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	Fa ($\times 10^{-6}$)	IoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	Fa ($\times 10^{-6}$)
Top-Hat [7]	7.143	79.84	1012	10.06	75.11	1432
Max-Median [8]	4.172	69.20	55.33	6.998	65.21	59.73
IPI [20]	25.67	85.55	11.47	27.92	81.37	16.18
RIPI [21]	11.05	79.08	22.61	14.11	77.55	28.31
WSLCM [14]	1.158	77.95	5446	3.452	72.44	6619
TLLCM [15]	1.029	79.09	5899	3.311	77.39	6738
CSAN	**75.89**	**97.72**	**8.1285**	**69.28**	**91.50**	**7.2877**

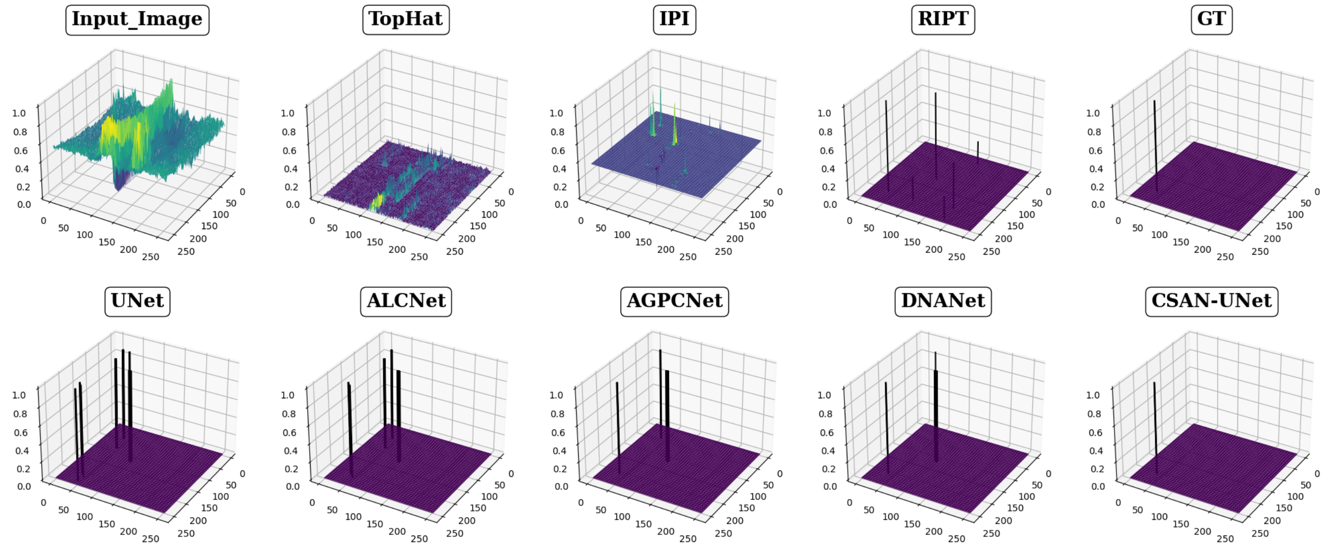


Figure 4. Visualization results in 3D of various methods for Scene 1. Different colors in the figure represent different grayscale values. Since RIPT and data-based methods have a high signal-to-noise ratio, the color of the target position predicted by these methods directly uses black.

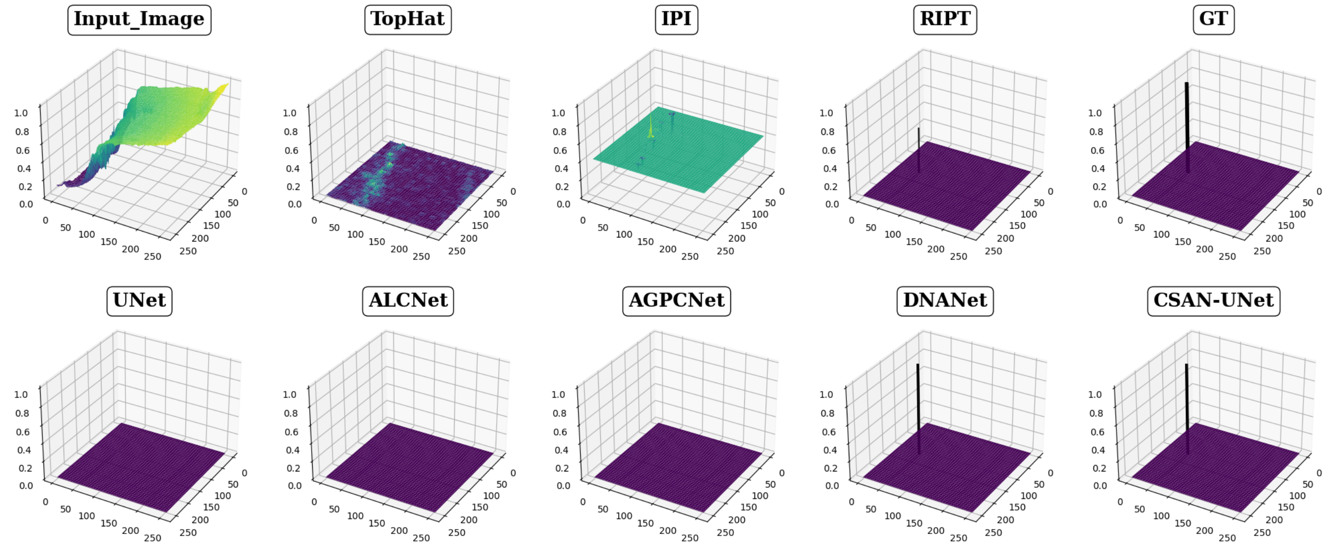


Figure 5. Visualization results in 3D of various methods for Scene 2. The meaning of the colors in the figure is similar to Figure 4.

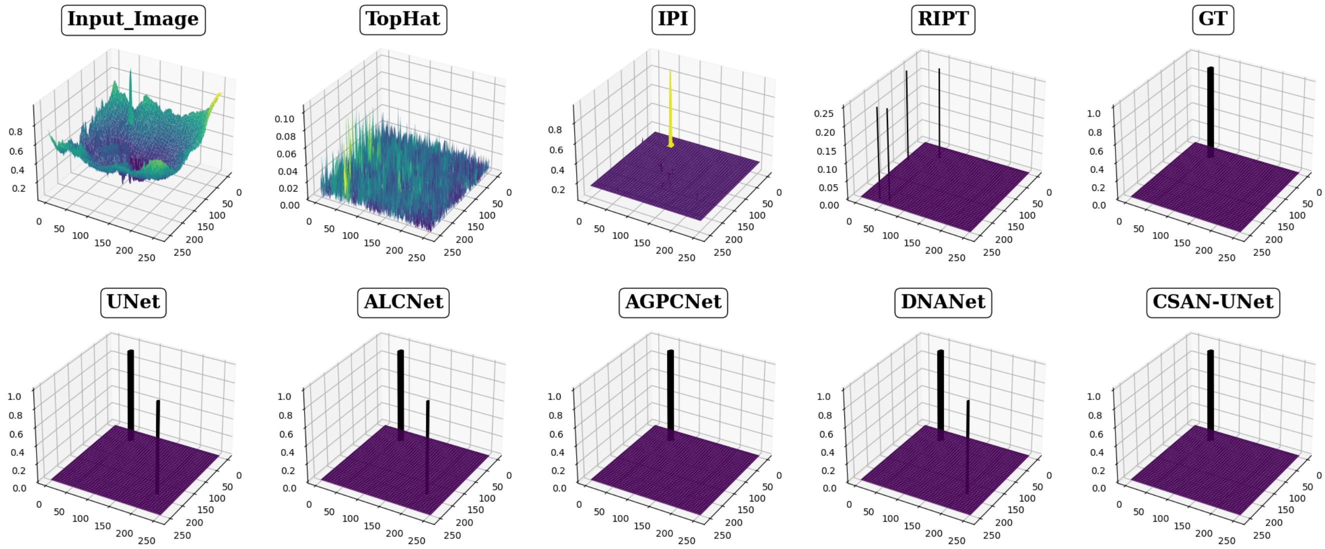


Figure 6. Visualization results in 3D of various methods for Scene 3. The meaning of the colors in the figure is similar to Figure 4. To more clearly illustrate the limitations of the traditional method with regard to its low signal-to-noise ratio, the z-axis is not constrained to a fixed range of 0 to 1. Instead, it dynamically adjusts to accommodate the range predicted by different methods.

4.3.2. Comparison with Data-Driven Methods

(1) Quantitative Results. To ensure a fair comparison, we retrained all data-driven methods using the same training datasets as our CSAN-UNet. We followed the original papers of these methods and used their specified fixed thresholds. All other parameters were kept consistent with the values stated in their original papers. It is important to highlight that we implemented these methods ourselves to ensure fairness in the comparison.

In Table 2, CSAN-UNet stands out for its superior performance over traditional data-driven methods, such as UNet, FCN, GCN, and Exfuse. This is attributed to its specially tailored backbone network, which preserves the intrinsic characteristics of small infrared targets deep within the network through modules such as CSCAM and CPSAM, enhancing the network's ability to represent small target features and thereby significantly improving detection performance. Compared to other small infrared target detection methods, such as DNANet, CSAN-UNet has the advantage of requiring fewer FLOPs and parameters, making it more suitable for deployment on embedded platforms. Additionally, CSAN-UNet has a faster inference speed, ranking second among other small infrared target detection algorithms with an FPS of approximately 53.

Table 2. Quantitative comparison with state-of-the-art methods on the NUAA-SIRST and IRSTD-1k dataset. The two highest scores are displayed in red and blue colors, respectively.

Method	FLOPs (G)	Params (M)	FPS	NUAA-SIRST			IRSTD-1K		
				IoU ($\times10^{-2}$)	Pd ($\times10^{-2}$)	Fa ($\times10^{-6}$)	IoU ($\times10^{-2}$)	Pd ($\times10^{-2}$)	Fa ($\times10^{-6}$)
FCN [45]	21.6534	20.8615	68.99	60.80	90.11	50.69	49.15	61.22	80.46
UNet [28]	47.2597	30.6029	74.9	70.00	95.81	63.45	55.35	90.51	42.96
GCN [46]	14.1150	55.4502	15.7	0.00	0.00	0.00	0.00	0.00	0.00
Exfuse [47]	50.7750	120.5935	20.8	0.00	0.00	0.00	0.00	0.00	0.00
MDvsFA [22]	230.1406	3.5937	5.9	60.30	89.35	56.35	49.50	82.11	80.33
ALC [24]	3.4458	0.3553	59.10	71.17	96.57	35.79	61.82	88.77	18.5231
AGPC [25]	40.2152	11.7878	15.6	72.01	95.81	35.29	62.30	89.45	19.58
DNA [27]	13.3013	4.4793	31.9	75.38	96.19	14.26	68.34	91.15	9.10
CSAN	5.4723	2.2795	52.6	75.89	97.72	8.12	69.28	91.50	7.28

DNANet uses a dense attention mechanism to compensate for the loss of infrared features during downsampling, adaptively enhancing the features of small infrared targets. However, the repeated interactions of the dense attention mechanism have a limited effect on maintaining deep infrared target characteristics but result in a significant increase in network parameters and FLOPs. Moreover, the dense attention mechanism combines channel and spatial attention but exhibits a uniform distribution of spatial attention weights across channels. This consistent weight distribution introduces noise and severely impacts the network's segmentation performance. AGPCNet, by using Global Contextual Attention (GCA) that considers correlations between pixels and global correlations between blocks, can more easily capture target information and suppress background noise. However, these attention mechanisms exhibit considerable computational complexity, making them computationally burdensome. In contrast, CSAN-UNet uses CPSAM with depth-wise convolution to dynamically allocate attention weights across spatial and channel dimensions, ensuring the preservation of channel relationships while reducing computational complexity, achieving a balance between detection accuracy and computational efficiency.

The effectiveness of CSAN-UNet is rigorously validated using ROC (Receiver Operating Characteristics) curves, which are pivotal in evaluating the performance of detection algorithms. The ROC curves presented in Figure 7 demonstrate that CSAN-UNet achieves better overall performance among the compared methods. Its Pd (Probability of Detection) can be dynamically adjusted according to changes in Fa (False-Alarm Rate), showcasing exceptional responsiveness to non-target disturbances within images. This adaptability, while maintaining high detection accuracy, emphasizes the robustness of CSAN-UNet across diverse operational conditions.

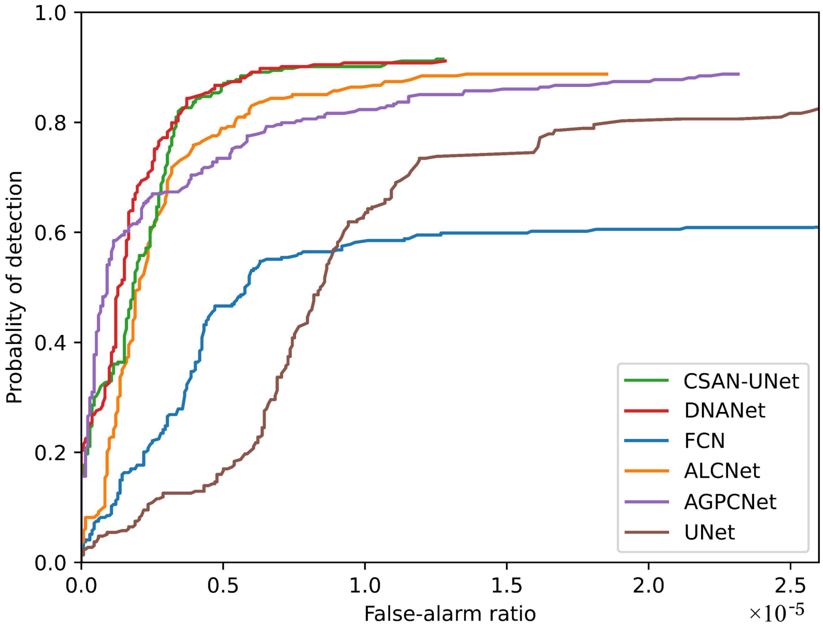


Figure 7. ROC results of CSAN-UNet and state-of-the-art methods.

(2) Qualitative comparison. In this part we use the same settings as the corresponding part in Section 4.3.1 and show the results in Figure 3.

In addressing the challenges of detecting small infrared targets, the U-Net architecture exhibits significant shortcomings due to its insufficient integration of feature layers and global context information, often leading to an ineffective synthesis of features that results in false positives, as evidenced in scene 1. A similar deficit is seen in ALCNet, which also fails to prevent false positives due to the absence of a feature attention fusion module, undermining its capacity to effectively assimilate pertinent features. AGPCNet, despite its potential, encounters limitations such as decreased spatial resolution from pooling operations and an inability to adeptly handle variations in object scales and aspect ratios, causing it to miss detections in more complex scenarios such as scenario 2. DNANet, although equipped with both channel and spatial attention mechanisms, suffers from a uniform distribution of spatial attention weights across all channels, which introduces noise and drastically undermines its segmentation accuracy, leading to observable false positives in scenes 1 and 3. In contrast, CSAN-UNet exhibits a robust capacity to adapt to scene changes and excels in shape segmentation, outperforming both ALCNet, AGPCNet, and DNANet.

4.4. Ablation Study

In this subsection, we analyze and compare our CSAN-UNet with various variants to examine the advantages brought by our network modules and design decisions.

(1) Ablation study for the CSCAM: CSCAM is used to adaptively enhance small infrared targets after downsampling, making up for the information loss after downsampling and maintaining deep target features, and enhancing the network's ability to extract contextual information. We compared CSAN-UNet with four variants to verify the benefits brought by this module.

- CSAN-UNet w/o attention: We use the classic residual connection layer instead of the attention-based feature enhancement layer
- CSAN-UNet-CSCAM-ResNet: We use the classic residual connection instead of CPSAM to fully evaluate the effectiveness of CSCAM in maintaining deep target features.
- CSAN-UNet-ResNet-CPSAM: We use the classic residual connection instead of CSCAM to evaluate the contribution of CBCAM to network performance.
- CSAN-UNet with CPSAM: We use CPSAM to replace the feature-enhanced CSCAM to explore the limitations of CPSAM in compensating for the information loss caused by downsampling.

Table 3 presents the results of ablation experiments for different variants, where higher values of IoU and Pd indicate better performance, while Fa follows the opposite trend. The best results in each column are highlighted in bold font. When comparing the performance of CSAN-UNet with CPSAM on the NUAA-SIRST dataset, it is observed that the IoU and Pd values decrease by 3.84% and 4.19%, while the Fa value increases by 2.289×10^{-5}. These findings demonstrate the limitations of CPSAM in compensating for information loss caused by downsampling operations.

Table 3. CSCAM and its main variants results for IoU, Pd and Fa on the NUAA-SIRST and IRSTD-1k datasets. The best results in each column are highlighted in bold.

Model	**NUAA-SIRST**			**IRSTD-1K**		
	IoU ($\times 10^{-2}$)	**Pd ($\times 10^{-2}$)**	**Fa ($\times 10^{-6}$)**	**IoU ($\times 10^{-2}$)**	**Pd ($\times 10^{-2}$)**	**Fa ($\times 10^{-6}$)**
CSAN-UNet w/o attention	73.50	95.43	47.48	65.84	88.09	40.84
CSAN-UNet-CSCAM-ResNet	73.75	95.81	16.11	66.10	89.45	38.83
CSAN-UNet-ResNet-CPSAM	74.25	96.57	21.46	66.94	89.79	40.46
CSAN-UNet with CPSAM	72.05	93.53	31.01	63.92	87.75	56.78
CSAN-UNet (our)	**75.89**	**97.72**	**8.12**	**69.28**	**91.50**	**7.28**

The results presented in Table 3 reveal that on the NUAA-SIRST dataset, the CSAN-UNet-ResNet-CPSAM variant experienced a decrease of 2.34% in IoU, a decrease of 1.71% in Pd, and an increase of 3.318×10^{-5} in Fa. Similar trends were observed on the IRSTD-1k dataset. These changes can be attributed to the fact that CSCAM performs better in compensating for the loss of target information during downsampling, thereby preserving deep target features and achieving improved performance.

(2) Ablation study for the CPSAM: CPSAM is used to efficiently extract deep semantic features of small targets and fully combine contextual information to improve the

modeling capabilities of the network. It serves as a lightweight, high-performance attention module that reduces the computational burden. In addition, CPSAM overcomes the limitation of uniform distribution of spatial attention weights between channels, reduces noise and improves the accuracy of infrared small target detection. We compared CSCAN-UNet with four variants to demonstrate the effectiveness of this module.

- CSAN-UNet w/o attention: In this variant, we excluded both the channel attention and spatial attention modules from the network architecture, in order to specifically study the individual contributions of these modules to the overall network performance.
- CSAN-UNet w/o CA: In this variant, we removed the Spatial attention (SA) module and retained only the Channel attention (CA) module to examine the contribution of SA to the network performance.
- CSAN-UNet w/o SA: In this variant, we removed the Channel attention (CA) module and retained only the Spatial attention (SA) module to examine the contribution of CA to the network performance.
- CSAN-UNet with CSCAM: In this variant, we utilized the CSCAM instead of the CPSAM to comprehensively evaluate the contribution of CPSAM to both reducing network computational complexity and improving network performance.

As shown in Table 4, removing all attention modules leads to a significant decrease in performance, with IoU dropping by 2.39%, Pd decreasing by 2.29%, and a slight increase in Fa on the NUAA-SIRST dataset, a similar trend is observed in the IRSTD-1k dataset. This emphasizes the critical role of channel and spatial attention modules in enhancing the capability of feature representation, which is particularly important for accurately detecting small infrared targets.

Table 4. CPSAM and its main variants results for FLOPs, Params, IoU, Pd, and Fa on the NUAA-SIRST and IRSTD-1k datasets. The best results in each column are highlighted in bold.

Model	FLOPs (M)	Params (M)	NUAA-SIRST IoU ($\times10^{-2}$)	 Pd ($\times10^{-2}$)	 Fa ($\times10^{-6}$)	IRSTD-1K IoU ($\times10^{-2}$)	 Pd ($\times10^{-2}$)	 Fa ($\times10^{-6}$)
CSAN-UNet w/o attention	6.5424	3.8696	73.50	95.43	47.48	65.84	88.09	40.84
CSAN-UNet w/o CA	6.5487	3.8952	74.09	95.81	20.46	66.78	90.13	35.90
CSAN-UNet w/o SA	6.5487	3.8952	74.40	96.57	9.91	67.26	90.81	10.40
CSAN-UNet with CSCAM	6.5487	3.8952	75.26	96.95	19.18	67.60	91.15	10.05
CSAN-UNet (our)	**5.4723**	**2.2795**	**75.89**	**97.72**	**8.12**	**69.28**	**91.50**	**7.28**

When channel attention is removed from the network, the intersection over union (IoU) on the NUAA-SIRST dataset decreases by 1.81%, the probability of detection (Pd) drops by 1.91%, and the false alarm rate increases by 1.234×10^{-5}, with a similar trend observed in the IRSTD-1k dataset. These changes highlight the critical function of channel attention in enhancing feature representation.

Removing spatial attention from CSAN-UNet significantly impacts its performance. Specifically, the Intersection over Union (IoU) decreases by 1.5%, the probability of detection (Pd) decreases by 1.15%, and the false alarm rate increases by 1.79×10^{-6}, with a similar trend observed in the IRSTD-1k dataset. This indicates the crucial role of spatial attention in CSAN-UNet, especially given the inherent difficulties of detecting small infrared targets, such as those obscured by thick clouds and environmental noise.

It is evident from Table 4 that replacing CSCAM with CPSAM in CSAN-UNet results in improvements in IoU, Pd, and Fa. This is because, unlike CSCAM's uniform spatial weight distribution, CPSAM enhances these metrics by dynamically allocating attention weights across both channel and spatial dimensions. This approach allows the network to more effectively focus on relevant features, reduce noise, and enhance segmentation performance. Moreover, transitioning to CPSAM significantly reduces computational complexity, with a reduction of 41.5% in parameters and a 19.67% decrease in floating-point operations (FLOP).

This efficiency stems from CPSAM's use of deep convolutional modules to create spatial attention maps.

5. Conclusions

In this research, we introduce a novel architecture aimed at enhancing the detection of small infrared targets, incorporating sophisticated components tailored for this purpose. At the feature extraction stage, the Cascaded Channel and Spatial Convolutional Attention Module (CSCAM) is employed to mitigate the loss of contextual information in deeper network layers by leveraging global contextual priors associated with small targets. During the feature fusion stage, the Channel-Prior and Spatial Attention Cascade Module (CPSAM) is utilized to minimize the semantic gap encountered in the multi-layer feature fusion process. These components are enhanced further by the addition of a Multiscale Feature Fusion Module, designed to improve the integration and utilization of information across different scales, thus ensuring the unique characteristics of small targets are effectively prioritized and preserved. The effectiveness of CSAN-UNet is rigorously validated through comparative analysis with state-of-the-art methods and comprehensive ablation studies to pinpoint the contributions of its various components. Demonstrating superior performance on publicly available datasets, such as NUAA-SIRST and IRSTD-1k, and CSAN-UNet, has proven its robustness and efficacy in detecting small infrared targets, marking a significant advancement in the field.

Author Contributions: Conceptualization, Y.Z. (Yuhan Zhong) and Z.S.; methodology, Y.Z. (Yuhan Zhong) and Z.S.; software, Y.Z. (Yuhan Zhong) and Y.Z. (Yong Zhang); supervision, Y.Z. (Yan Zhang) and H.L.; validation, Z.S., H.L. and Y.Z. (Yan Zhang); formal analysis, Y.Z. (Yuhan Zhong) and Z.S.; data curation, Z.S.; writing—original draft preparation, Y.Z. (Yuhan Zhong) and Z.S.; writing—review and editing, Y.Z. (Yan Zhang) and Y.Z. (Yong Zhang); project administration, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 51961302145).

Data Availability Statement: The SIRST are available at: https://github.com/YimianDai/sirst and The IRSTD-1k are available at: https://github.com/RuiZhang97/ISNet.

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

References

1. Teutsch, M.; Krüger, W. Classification of small boats in infrared images for maritime surveillance. In Proceedings of the 2010 International WaterSide Security Conference, Carrara, Italy, 3–5 November 2010; pp. 1–7.
2. Kou, R.; Wang, H.; Zhao, Z.; Wang, F. Optimum selection of detection point and threshold noise ratio of airborne infrared search and track systems. *Appl. Opt.* **2017**, *56*, 5268–5273. [CrossRef]
3. Rawat, S.S.; Verma, S.K.; Kumar, Y. Review on recent development in infrared small target detection algorithms. *Procedia Comput. Sci.* **2020**, *167*, 2496–2505. [CrossRef]
4. Huang, S.; Liu, Y.; He, Y.; Zhang, T.; Peng, Z. Structure-adaptive clutter suppression for infrared small target detection: Chain-growth filtering. *Remote Sens.* **2019**, *12*, 47. [CrossRef]
5. 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*, 87–119. [CrossRef]
6. Azimi-Sadjadi, M.R.; Pan, H. Two-dimensional block diagonal LMS adaptive filtering. *IEEE Trans. Signal Process.* **1994**, *42*, 2420–2429. [CrossRef]
7. Rivest, J.F.; Fortin, R. Detection of dim targets in digital infrared imagery by morphological image processing. *Opt. Eng.* **1996**, *35*, 1886–1893. [CrossRef]
8. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. Max-Mean and Max-Median Filters for Detection of Small-Targets. In Proceedings of the Signal and Data Processing of Small Targets 1999, Denver, CO, USA, 19–23 July 1999.
9. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]
10. Shi, Y.; Wei, Y.; Yao, H.; Pan, D.; Xiao, G. High-boost-based multiscale local contrast measure for infrared small target detection. *IEEE Geosci. Remote Sens. Lett.* **2017**, *15*, 33–37. [CrossRef]
11. 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]

12. Chen, Y.; Song, B.; Wang, D.; Guo, L. An effective infrared small target detection method based on the human visual attention. *Infrared Phys. Technol.* **2018**, *95*, 128–135. [CrossRef]
13. Kou, R.; Wang, C.; Fu, Q.; Yu, Y.; Zhang, D. Infrared small target detection based on the improved density peak global search and human visual local contrast mechanism. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 6144–6157. [CrossRef]
14. Han, J.; Moradi, S.; Faramarzi, I.; Zhang, H.; Zhao, Q.; Zhang, X.; Li, N. Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1670–1674. [CrossRef]
15. Chen, C.P.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A Local Contrast Method for Small Infrared Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 574–581. [CrossRef]
16. Recht, B.; Fazel, M.; Parrilo, P.A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **2010**, *52*, 471–501. [CrossRef]
17. Wang, X.; Peng, Z.; Kong, D.; Zhang, P.; He, Y. Infrared dim target detection based on total variation regularization and principal component pursuit. *Image Vis. Comput.* **2017**, *63*, 1–9. [CrossRef]
18. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
19. Zhou, F.; Wu, Y.; Dai, Y.; Wang, P. Detection of small target using schatten 1/2 quasi-norm regularization with reweighted sparse enhancement in complex infrared scenes. *Remote Sens.* **2019**, *11*, 2058. [CrossRef]
20. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [CrossRef]
21. Dai, Y.; Wu, Y. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3752–3767. [CrossRef]
22. Wang, H.; Zhou, L.; Wang, L. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8509–8518.
23. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric contextual modulation for infrared small target detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 950–959.
24. 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]
25. 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]
26. 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, 18 June–24 June 2022; pp. 877–886.
27. 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]
28. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
29. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
30. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
31. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
32. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
33. Huang, H.; Chen, Z.; Zou, Y.; Lu, M.; Chen, C. Channel prior convolutional attention for medical image segmentation. *arXiv* **2023**, arXiv:2306.05196.
34. Zuo, Z.; Tong, X.; Wei, J.; Su, S.; Wu, P.; Guo, R.; Sun, B. AFFPN: Attention fusion feature pyramid network for small infrared target detection. *Remote Sens.* **2022**, *14*, 3412. [CrossRef]
35. Zhao, B.; Wang, C.; Fu, Q.; Han, Z. A novel pattern for infrared small target detection with generative adversarial network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4481–4492. [CrossRef]
36. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust infrared small target detection network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 7000805. [CrossRef]
37. Wu, T.; Li, B.; Luo, Y.; Wang, Y.; Xiao, C.; Liu, T.; Yang, J.; An, W.; Guo, Y. MTU-Net: Multilevel TransUNet for Space-Based Infrared Tiny Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5601015. [CrossRef]

38. Ying, X.; Liu, L.; Wang, Y.; Li, R.; Chen, N.; Lin, Z.; Sheng, W.; Zhou, S. Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 15528–15538.
39. Wu, X.; Hong, D.; Chanussot, J. UIU-Net: U-Net in U-Net for Infrared Small Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 364–376. [CrossRef]
40. 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, 18–22 June 2018; pp. 7132–7141.
41. Li, Q.; Gong, M.; Yuan, Y.; Wang, Q. Symmetrical Feature Propagation Network for Hyperspectral Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5536912. [CrossRef]
42. Guo, C.; Szemenyei, M.; Yi, Y.; Wang, W.; Chen, B.; Fan, C. Sa-unet: Spatial attention u-net for retinal vessel segmentation. In Proceedings of the 2020 25th international conference on pattern recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1236–1242.
43. 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.
44. Li, Q.; Yuan, Y.; Wang, Q. Multiscale Factor Joint Learning for Hyperspectral Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5523110. [CrossRef]
45. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
46. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
47. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–284.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.