

Article

SCUA-Net: Selective Contextual Uplift and Attention Network for Robust Infrared Small Target Detection in Complex Clutter

Jiawei Lin ¹, Xiaoyan Wang ¹, Songjie Luo ^{1,*}, Ziyang Chen ¹ , Xiaoyan Wu ^{2,*} and Jixiong Pu ¹

¹ Fujian Provincial Key Laboratory of Light Propagation and Transformation, College of Information Science & Engineering, Huaqiao University, Jimei Avenue No. 668, Xiamen 361021, China; jixiong@hqu.edu.cn (J.P.)

² Institute of Fluid Physics, China Academy of Engineering Physics, Mianyang 621900, China

* Correspondence: songjie@hqu.edu.cn (S.L.); wuxiaoyan1219@sina.cn (X.W.)

Abstract

Infrared small target detection (ISTD) remains challenging in complex cluttered environments because targets usually occupy only a few pixels and exhibit weak thermal radiation with limited texture information. The problem becomes more severe in high-resolution infrared imaging systems, where sliding-window inference is commonly adopted under memory and computational constraints. However, the truncated field of view may lead to contextual information loss and increased false alarms in cluttered regions. To address these issues, we propose the Selective Contextual Uplift and Attention Network (SCUA-Net). The proposed network adopts a U-Net++-style densely nested encoder–decoder architecture to enhance multi-scale feature interaction and preserve fine-grained weak-target features. In addition, a Global-Context Calibration Coordinate Attention (GCC-CA) module is introduced to inject window-level contextual statistics into coordinate attention, thereby improving clutter suppression and localization robustness under sliding-window inference. During training, a joint optimization strategy combining Online Hard Example Mining (OHEM) and Dice Loss is employed to alleviate severe foreground–background imbalance. During inference, Gaussian-weighted fusion is adopted to reduce stitching artifacts between adjacent windows. Experimental results on NUDT-SIRST and IRSTD-1k validate the effectiveness of the proposed method. SCUA-Net achieves 99.15% Pd, 0.558×10^{-6} Fa, and 0.9570 IoU on NUDT-SIRST, while maintaining competitive performance on IRSTD-1k at 161.6 FPS on an NVIDIA RTX 4090 platform, demonstrating favorable accuracy, robustness, and real-time performance in complex infrared scenarios.

Keywords: infrared small target detection; densely nested encoder–decoder; coordinate attention; contextual calibration; sliding-window inference; false alarm suppression



Received: 10 June 2026

Revised: 30 June 2026

Accepted: 6 July 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Infrared small target detection (ISTD) is fundamental and critical in infrared search and tracking, long-range surveillance, remote sensing, and nighttime perception systems [1,2]. Compared with conventional object detection tasks, infrared small targets usually occupy only a very limited number of pixels and are characterized by weak radiation intensity and scarce texture information. Moreover, they are easily submerged in complex cluttered backgrounds such as cloud edges, sea-surface ripples, and high-reflective ground objects. As a result, ISTD frequently encounters target vanishing, high false alarm rates from background clutter, and significant shape distortion in real-world conditions. For such tasks, detection models are required not only to possess strong weak-target representation

capability but also to maintain robust background suppression ability and low false alarm rates under complex cluttered environments.

Traditional ISTD approaches mainly include filtering-based methods, local contrast-based methods, and low-rank decomposition-based methods. Filtering-based methods rely on the assumption that the background is relatively smooth and enhance target responses through specifically designed filters while suppressing background interference [3,4]. Local contrast-based methods exploit grayscale differences between targets and surrounding regions to highlight salient responses. Low-rank decomposition-based methods model infrared images as the superposition of low-rank backgrounds and sparse targets for target separation [5,6]. Although these methods exhibit good interpretability, they heavily depend on handcrafted priors and often suffer from limited robustness in scenarios containing rapidly changing backgrounds, strong local high-brightness interference, or complex texture structures, thereby producing numerous pseudo-responses.

The success of convolutional neural networks in dense prediction tasks provides potential in ISTD methods. Different from conventional object detection methods based on bounding boxes, infrared small targets usually occupy only a few pixels and lack explicit structural features. Therefore, formulating ISTD as a pixel-wise binary semantic segmentation task has become the dominant paradigm in recent years, where precise target localization is achieved through pixel-level prediction. Under this segmentation-based framework, encoder–decoder architectures represented by U-Net [7] provide an effective feature fusion structure for dense prediction. Many subsequent studies further introduce contextual modulation, local contrast enhancement, densely nested connections, and attention mechanisms to improve the discriminability of weak targets at the pixel level. For example, ACM [8] enhances cross-layer feature interaction through asymmetric contextual modulation, ALCNet [9] embeds local contrast priors into the network to strengthen target responses, DNA-Net [10] utilizes densely nested connections to repeatedly aggregate and reconstruct multi-level features, and UIU-Net [11] further improves feature representation capability through multi-resolution nested structures.

Recent years have witnessed rapid progress in ISTD, with increasing attention devoted to richer contextual modeling and more discriminative feature representation. Beyond conventional CNN-based feature enhancement, recent studies have explored transformer-based global dependency modeling, lightweight detection architectures, and hybrid representation learning to further improve robustness in complex cluttered scenes. In parallel, tensor-based spatiotemporal methods—such as LRTA-SP [12] (Low-Rank Tensor Approximation with Saliency Prior), FDMDEA-STTM [13] (Facet Derivative-Based Multidirectional Edge Awareness and Spatial–Temporal Tensor Model), and STTM-SFR [14] (Spatial–Temporal Tensor Modeling with Saliency Filter Regularization)—have demonstrated strong performance by leveraging holistic spatial–temporal priors, edge awareness, and saliency regularization for target–background separation in infrared sequences. These advances indicate that currentIRSTD research is gradually evolving from local contrast enhancement toward more comprehensive contextual modeling and adaptive feature learning.

Although these methods have substantially improved feature representation and clutter suppression, they are developed under the full-image processing paradigm, where sufficient global contextual information is available during inference. However, this assumption does not always hold in practical high-resolution infrared applications, where sliding-window or patch-wise inference is frequently adopted because of GPU memory and computational constraints. Under such settings, the model can only access local regions during inference and therefore cannot obtain complete background statistical information from the entire scene. Consequently, contextual information loss caused by patch-wise in-

ference has become an increasingly important practical issue that is insufficiently addressed by existing full-image-oriented detection frameworks.

To mitigate these issues, we propose SCUA-Net, a densely nested encoder–decoder framework equipped with Global-Context Calibration Coordinate Attention (GCC-CA) for contextual calibration under sliding-window inference. It strengthens multi-scale feature fusion for weak targets by employing a U-Net++-style [15] densely nested architecture as its backbone. We introduce the GCC-CA module to inject local window statistics into the coordinate attention [16] mechanism, thereby enhancing discriminative power in cluttered scenes. During training, Online Hard Example Mining (OHEM) [17] and Dice Loss [18] are jointly employed to strengthen hard-sample learning. During inference, a sliding-window strategy combined with Gaussian-weighted fusion is adopted to alleviate stitching artifacts near window boundaries. These strategies improve the recognition efficiency of ISTD.

2. Materials and Methods

The difficulty of ISTD in complex backgrounds arises not only from the weak radiation, low texture, and extremely small size of the target itself, but also from the local inference protocol commonly used in practical deployment. In particular, when processing high-resolution panoramic infrared images, sliding-window inference can alleviate memory pressure, but it also truncates the global background statistics and deprives the model of the overall contextual reference within each local window, which in turn leads to more pronounced false detections in regions such as bright cloud edges and sea-surface ripples. To address this issue, this paper proposes SCUA-Net, which adopts a densely nested encoder–decoder structure as its backbone to preserve weak-target details while enhancing multi-scale feature interaction. Meanwhile, a GCC-CA module is introduced to inject global contextual statistics within local windows into the coordinate attention mechanism, thereby improving discriminative stability under complex backgrounds.

2.1. Overall Dense Nested Architecture

As shown in Figure 1, SCUA-Net adopts a U-Net++-style [15] densely nested encoder–decoder topology as the overall framework. The network consists of five levels. Multi-scale semantic information is obtained through progressive downsampling. Meanwhile, cross-level skip connections are introduced between the encoder and decoder. In addition, dense nested feature aggregation is performed among nodes at the same scale, so as to enable repeated interaction between shallow spatial details and deep semantic features. Compared with the traditional U-Net, which relies on a single skip connection for feature transfer, the dense nested structure allows a node at the same scale to receive multiple feature inputs from adjacent levels, which helps mitigate the gradual attenuation of infrared small targets in deep representations.

The node feature is denoted as $x^{i,j}$, where i indicates the i -th downsampling stage and j indicates the j -th nested level. The feature of this node is determined by the feature from the previous node at the same level, the upsampled feature from the next encoder stage, and the progressively aggregated features generated at the same level. This process can be formulated as follows:

$$x^{i,0} = \mathcal{H}^{i,0} \left(\mathcal{D} \left(x^{i-1,0} \right) \right), \quad (1)$$

$$x^{i,j} = \mathcal{H}^{i,j} \left(\left[x^{i,0}, x^{i,1}, \dots, x^{i,j-1}, U \left(x^{i+1,j-1} \right) \right] \right), j > 0, \quad (2)$$

where $\mathcal{H}^{i,j}(\cdot)$ denotes the nonlinear convolutional transformation block, $[\cdot]$ denotes channel-wise concatenation, and $\mathcal{D}(\cdot)$ and $U(\cdot)$ denote downsampling and upsampling operations, respectively. In this way, the network can repeatedly fuse local edges, weak bright spots, and contextual structural information across multiple scales, thereby preserving shallow

high-resolution features as fully as possible in the subsequent decoding process. Compared with structures that rely only on a single skip connection, this dense nested form is more suitable for ISTD tasks, in which the target occupies very few pixels but requires highly accurate localization.

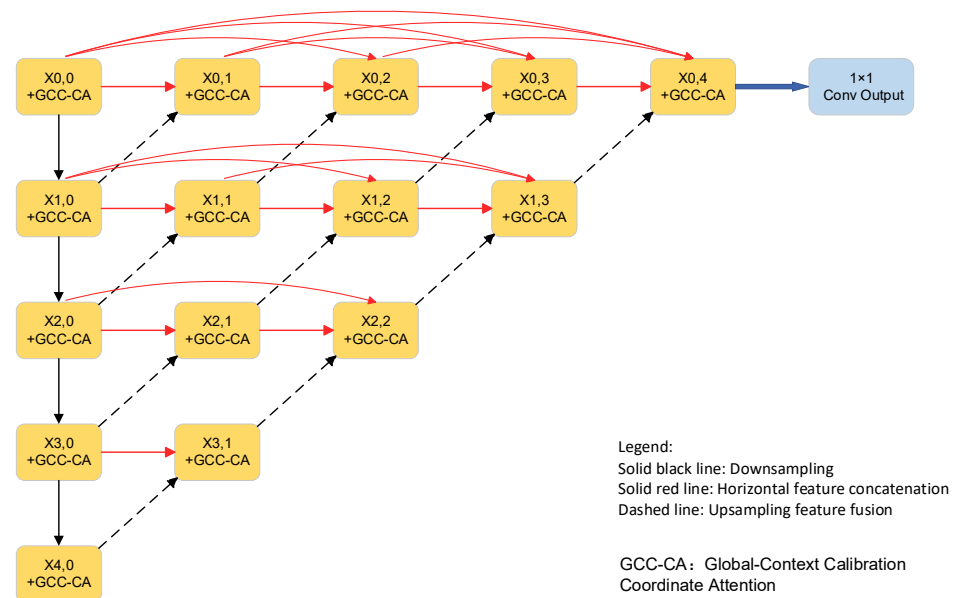


Figure 1. SCUA-Net overall architecture.

In practice, we apply a double convolution block at each node for feature extraction and insert the GCC-CA module afterward for calibration, which deepens the network and allows recalibration at every scale, thereby curbing the propagation of background clutter through the dense connections. Overall, the proposed architecture balances detail preservation, semantic aggregation, and background suppression, making it a suitable backbone for small-target segmentation in complex infrared scenes.

2.2. GCC-CA: Global-Context Calibration Coordinate Attention

Although sliding-window inference enables real-time processing of high-resolution images, the resulting loss of global context is a primary source of elevated false alarms. When a window falls entirely within a bright clutter region, conventional attention mechanisms may misclassify strong background responses as targets due to the absence of global statistical references. To address this problem, this paper proposes the GCC-CA module, whose micro-level mechanism is shown in Figure 2.

As illustrated in Figure 2, the GCC-CA module consists of three parallel branches: a vertical pooling branch, a horizontal pooling branch, and a global-context calibration (GCC) branch. The first two branches, which are inherited from Coordinate Attention (CA) [16], are used to capture the precise spatial position of the target. The global-context branch extracts channel-wise statistical information of the current local window through global average pooling, providing the necessary background reference for local inference.

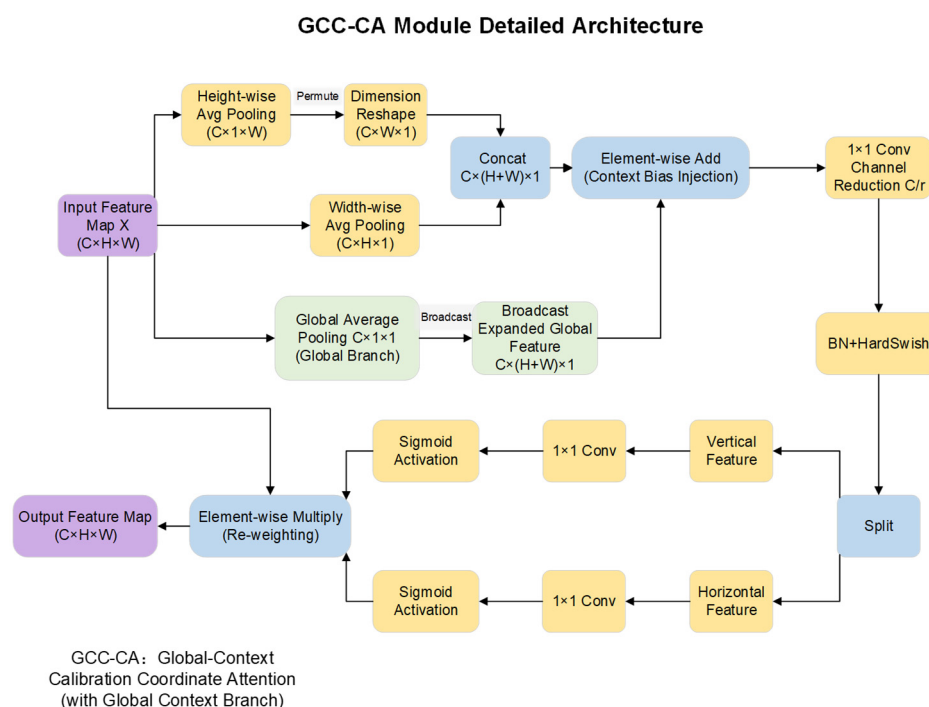


Figure 2. Micro-level schematic of the GCC-CA module.

Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, one-dimensional average pooling is first performed along the horizontal and vertical directions, yielding the vertical coordinate feature $X_h \in \mathbb{R}^{C \times H \times 1}$ and the horizontal coordinate feature $X_w \in \mathbb{R}^{C \times 1 \times W}$. Meanwhile, the global-context branch applies global average pooling [19] to the input feature to obtain the channel-wise global descriptor X_g , as follows:

$$X_g = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W X(:, h, w) \in \mathbb{R}^{C \times 1 \times 1}. \quad (3)$$

To effectively inject global context, X_w is first permuted into $C \times W \times 1$ and concatenated with X_h along the spatial dimension to obtain the joint coordinate feature. Then, the broadcasted global-context vector $\tilde{g} \in \mathbb{R}^{C \times 1 \times 1}$ is injected into the joint coordinate feature by element-wise additive fusion to compensate for environmental bias. This feature aggregation and nonlinear transformation process can be formulated as follows:

$$f = \delta \left(\mathcal{B} \left(\left[X_h, X_w^T \right] + \tilde{g} \right) \right), \quad (4)$$

where $[\cdot]$ denotes concatenation along the spatial dimension, $f \in \mathbb{R}^{\frac{C}{r} \times (H+W) \times 1}$ is the intermediate feature map containing multi-directional and global context information, $\mathcal{B}(\cdot)$ denotes a 1×1 convolution operation with reduction ratio r followed by batch normalization [20], and $\delta(\cdot)$ denotes the HardSwish activation function [21].

The intermediate feature f is then split along the spatial dimension into two separate tensors, $f_h \in \mathbb{R}^{\frac{C}{r} \times H \times 1}$ and $f_w \in \mathbb{R}^{\frac{C}{r} \times W \times 1}$. These two tensors are independently restored to the original channel dimension C through separate 1×1 convolutions, and the final vertical and horizontal attention maps are generated by the Sigmoid function σ , as follows:

$$A_h = \sigma(\mathcal{C}_h(f_h)), \quad (5)$$

$$A_w = \sigma(\mathcal{C}_w(f_w)). \quad (6)$$

Finally, the output of GCC-CA is obtained by element-wise multiplication of the input feature X with the two directional attention maps.

$$Y = X \odot A_h \odot A_w. \quad (7)$$

Unlike conventional attention mechanisms that rely solely on multiplicative reweighting, GCC-CA introduces an explicit contextual compensation by additively injecting global statistics before computing attention weights. This enables robust suppression of complex clutter, such as bright cloud edges and sea-surface ripples, without unduly weakening the already faint target signals. The experimental section further verifies the significant contribution of this module to false-alarm reduction.

2.3. Training Objective and Optimization Strategy

The ISTD training faces severe class imbalance, with target pixels vastly outnumbering background pixels. To address this issue, we adopt a joint loss combining OHEM-based BCE [17] and Dice Loss [18], to simultaneously take into account the ability to mine hard examples in complex backgrounds and to reconstruct the target area.

2.3.1. OHEM with Dice Loss

The joint loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{Dice} + \lambda \mathcal{L}_{OHEM-BCE}, \quad (8)$$

The coefficient λ controls the balance between structural target preservation (Dice Loss) and hard clutter suppression (OHEM-BCE Loss). A smaller λ weakens hard-example mining, increasing false alarms; conversely, a larger λ overly dominates the optimization with background gradients, suppressing faint target responses. In our framework, because the OHEM strategy dynamically filters out 95% of easy background pixels, the extreme class imbalance is alleviated at the sample level. This ensures the numerical scales of both loss components are naturally aligned. Therefore, we set $\lambda = 1.0$ to provide a balanced optimization objective without introducing artificial gradient skewing.

The mathematical definitions of the loss terms are as follows:

$$\mathcal{L}_{OHEM-BCE} = -\frac{1}{|\Omega_{pos}| + |\Omega_{hard}|} \left(\sum_{i \in \Omega_{pos}} y_i \log p_i + \sum_{j \in \Omega_{hard}} (1 - y_j) \log(1 - p_j) \right), \quad (9)$$

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_{i=1}^N p_i y_i}{\sum_{i=1}^N p_i + \sum_{i=1}^N y_i + \epsilon}. \quad (10)$$

Here, p_i and y_i denote the predicted probability and ground-truth label of the i -th pixel in the full segmentation map, respectively. Ω_{pos} denotes the set of positive target pixels, and Ω_{hard} denotes the set of hard negative pixels selected during training. The Dice loss measures the overlap between the predicted segmentation and the ground truth [22], which helps maintain stable optimization under severe class imbalance. The OHEM-based BCE term is computed on Ω_{pos} and Ω_{hard} , so that the loss is dominated neither by the abundant easy background pixels nor by trivial negatives. In this paper, only the top 5% hardest background pixels are retained for the BCE term, while the remaining easy background pixels are ignored. In this way, the network focuses more on ambiguous background regions that are more likely to cause false alarms.

Compared with using BCE or Dice alone, this joint strategy is more conducive to suppressing random responses in complex backgrounds and improving the response stability of low-contrast targets. It should be noted that OHEM is not intended to replace

all background supervision; rather, it emphasizes difficult examples that are closer to the decision boundary, thereby improving the model's adaptability to complex scenes.

2.3.2. Weight Decay Regularization

Introducing OHEM significantly weakens the gradients of a large number of easy background samples. Consequently, the network weights responsible for background representations may suffer from insufficient updates or response drift. To suppress this phenomenon, a weight decay regularization term is introduced into the optimizer to provide a stability constraint for parameter updates, thereby reducing the random fluctuations caused by sparse background samples during training. This setting is particularly important for datasets such asIRSTD-1k, where the background is more complex and the interference is stronger.

The OHEM focuses on learning hard negative examples, Dice Loss preserves target regions, and weight decay ensures training stability. The combination of these three strategies enforces the model to focus more on the discriminative regions that truly affect performance in complex backgrounds, rather than being dominated by a large number of easy background pixels. This strategy is consistent with the contextual calibration design of GCC-CA; the former reduces background interference at the optimization level, while the latter suppresses false responses at the feature level.

2.4. Sliding-Window Inference with Gaussian Weighted Fusion

Considering the high-resolution panoramic images commonly encountered in practical infrared warning scenarios, sliding-window inference is adopted during testing to satisfy memory and speed constraints. Unlike the training stage, which directly processes fixed-size cropped images, the inference stage requires block-wise scanning of the entire image, which introduces a certain truncation effect at the window boundaries. To alleviate this issue, a Gaussian weighted fusion strategy is used to reconstruct the prediction results of adjacent windows.

The final pixel-wise output can be expressed as follows:

$$P_{final}(u, v) = \frac{\sum_{k=1}^K G_k(u, v) P_k(u, v)}{\sum_{k=1}^K G_k(u, v) + \epsilon}, \quad (11)$$

where the predicted probability of the k -th window is denoted as P_k , and the corresponding Gaussian weight is denoted as G_k . ϵ is a stabilizing term introduced to avoid division by zero. This strategy is complementary to the proposed GCC-CA. GCC-CA focuses on compensating for the lack of local context using statistics within the window, whereas Gaussian fusion reduces discontinuities caused by block-wise stitching from the output level. Therefore, they constitute a complete inference scheme suitable for high-resolution infrared scenes. Figure 3 illustrates the overall difference between the training and inference pipelines.

The SCUA-Net combines multi-scale detail preservation through dense nested connections, contextual calibration via GCC-CA, effective hard-sample learning with the joint loss, and continuous high-resolution output through Gaussian-weighted fusion. Through the synergy of these modules, SCUA-Net achieves strong robustness and engineering adaptability in complex backgrounds.

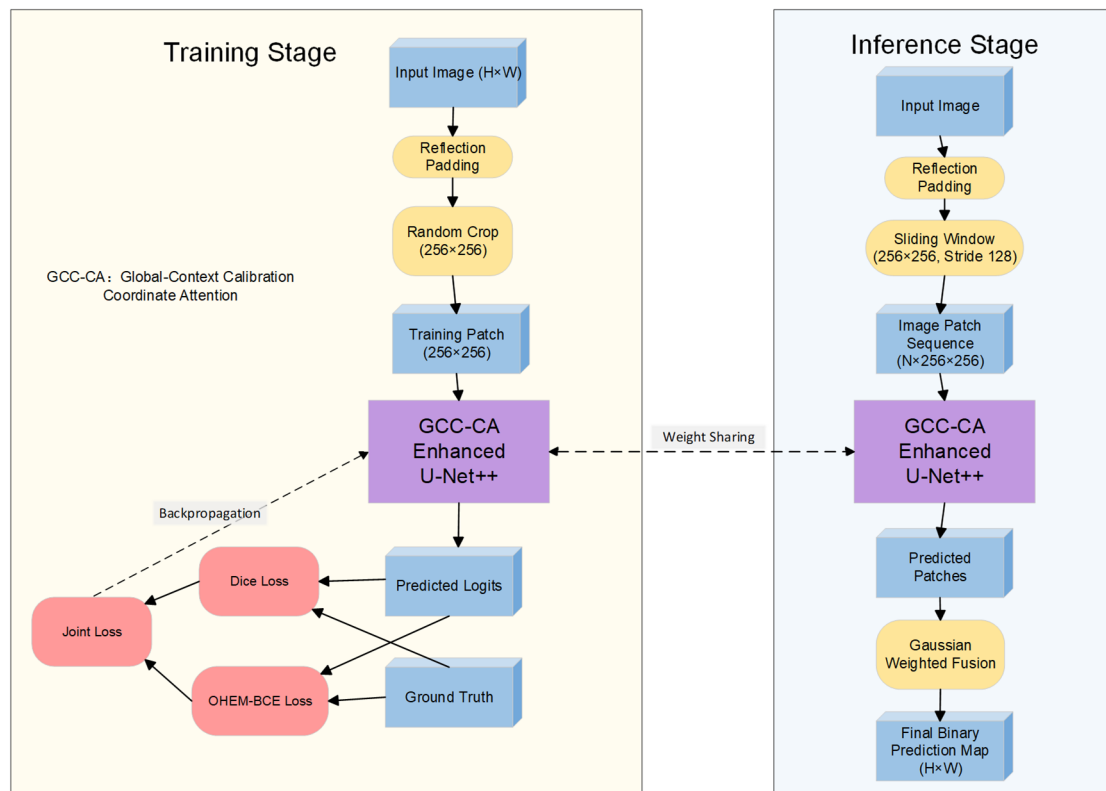


Figure 3. Overall pipeline of the proposed SCUA-Net in training and inference stages.

3. Results

3.1. Datasets and Evaluation Metrics

3.1.1. Datasets

The experiments were conducted on the NUDT-SIRST [10] and IRSTD-1k [23] datasets. NUDT-SIRST contains relatively rich mixed backgrounds composed of synthetic and real scenes, with extremely small targets and diverse shapes, making it suitable for evaluating the localization and reconstruction capability of a model under complex backgrounds. IRSTD-1k consists of a large number of highly challenging scenes such as bright cloud layers and sea-surface ripples, where targets are often obscured by local high-frequency clutter, and is therefore more suitable for assessing background suppression ability and robustness in highly complex scenes.

3.1.2. Evaluation Metrics

Probability of Detection (P_d), False Alarm Rate (F_a), Intersection over Union (IoU), and F-measure (F_1) are adopted as evaluation metrics. The definitions are as follows:

$$P_d = \frac{N_{det}}{N_{gt}}, \quad (12)$$

$$F_a = \frac{N_{FP}}{N_{all}}, \quad (13)$$

$$IoU = \frac{TP}{TP + FP + FN}. \quad (14)$$

where N_{gt} denotes the total number of ground-truth targets, N_{det} denotes the number of correctly detected targets, N_{FP} denotes the number of false-detected pixels, and N_{all} denotes the total number of pixels in the image. TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

Furthermore, as ISTD is formulated as a pixel-wise semantic segmentation task in this framework, we adopt the pixel-level F-measure (F_1) to comprehensively assess the harmonic mean of precision and recall. At the pixel level, the F-measure is mathematically equivalent to the Dice coefficient and can be equivalently expressed using IoU.

$$F_1 = \frac{2\text{IoU}}{1 + \text{IoU}} \quad (15)$$

In the tables, the F_a values are presented in units of 10^{-6} for better readability under extremely low false-alarm settings.

3.2. Implementation Details

To ensure comparability, all models were trained from scratch under the same software and hardware environment. The experiments were carried out on a cloud server provided by AutoDL (SeetaCloud Technology Co., Ltd., Nanjing, China), equipped with a single NVIDIA RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) with 24 GB VRAM, running on Ubuntu 20.04 (Canonical Ltd., London, UK), Python 3.8 (Python Software Foundation, Wilmington, DE, USA), and PyTorch 2.0.1 (Meta AI, Menlo Park, CA, USA) with CUDA 11.8. Following common practices in ISTD, a target-centric cropping strategy was adopted for all compared models during training. Specifically, training samples were generated around the target region with a probability of 60–70%, and the crop size was fixed to 256×256 . This standard strategy helps alleviate the extreme positive–negative imbalance typical in ISTD, improves the learning efficiency on sparse targets, and ensures identical data sampling conditions for a fair comparison.

The proposed SCUA-Net was optimized using the AdamW optimizer [24] with a batch size of 16. The initial learning rate (η_0) was explicitly set to 0.05, and the training was performed for 160 epochs using a cosine annealing schedule with a minimum learning rate of 1×10^{-5} . Additionally, the reduction ratio (r) in the GCC-CA module was set to 16 to balance the computational overhead and channel-wise feature interaction. The loss function adopted a joint optimization paradigm combining OHEM and Dice Loss, where the truncation ratio of OHEM was set to 5%, meaning that only the top 5% hardest background samples with the largest prediction errors were retained for gradient updates, while the remaining easy background samples were suppressed. To further alleviate parameter drift caused by sparse background gradients after OHEM, a stronger weight decay regularization of 1×10^{-2} was used on IRSTD-1k, whereas a standard weight decay of 1×10^{-4} was used on NUDT-SIRST.

During testing, to simulate the processing mode of real infrared warning systems for high-resolution panoramic images, a sliding-window inference strategy with a stride of 128 and a window size of 256×256 was adopted. The outputs of adjacent overlapping windows were seamlessly reconstructed through a 2D Gaussian-weighted fusion matrix, where the standard deviation (σ) of the Gaussian kernel was empirically set to 40 to effectively eliminate contextual truncation artifacts at the window boundaries. This reduces the stitching artifacts introduced by contextual truncation at window boundaries and makes the final output more spatially continuous and more stable in localization.

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

SCUA-Net was compared with four representative methods, namely ACM [8], ALC-Net [9], ISTDU-Net [25], and DNA-Net [10]. These methods respectively represent typical paradigms of contextual modulation, local contrast enhancement, bidirectional feature fusion, and densely nested architecture in ISTD. Table 1 summarizes their quantitative results under the standard binary segmentation threshold of 0.5 on NUDT-SIRST and

IRSTD-1k. While this fixed threshold provides a direct pixel-level evaluation, a comprehensive performance comparison across dynamic thresholds is further presented via ROC and PR curves.

Table 1. Quantitative comparison on NUDT-SIRST and IIRSTD-1k datasets. The best results are highlighted in bold.

Method	NUDT-SIRST				IIRSTD-1k			
	Pd (%) ↑	Fa ($\times 10^{-6}$) ↓	IoU ↑	F_1 ↑	Pd (%) ↑	Fa ($\times 10^{-6}$) ↓	IoU ↑	F_1 ↑
ACM	97.74	1.360	0.8572	0.9231	95.96	113.018	0.467	0.637
ALCNet	96.46	0.784	0.8605	0.9250	92.26	86.447	0.492	0.660
ISTDU-Net	99.06	2.370	0.8842	0.9385	92.59	63.635	0.623	0.768
DNA-Net	96.62	0.830	0.8993	0.9470	93.94	71.644	0.627	0.771
SCUA-Net	99.15	0.558	0.9570	0.9780	91.87	52.210	0.628	0.772

3.3.1. Quantitative Results

On NUDT-SIRST, SCUA-Net outperforms other methods across Pd, Fa, and IoU. In particular, it achieves a substantially lower false alarm rate of 0.558×10^{-6} , highlighting the strong clutter suppression capability enabled by the densely nested architecture and GCC-CA module. Moreover, the high IoU of 0.9570 indicates superior target contour reconstruction and pixel-level consistency, effectively preserving weak target details while reducing background interference.

On the more challenging IIRSTD-1k dataset, false alarm rates of all methods increase markedly due to the presence of strong clutter such as bright cloud layers, sea-surface ripples, and local high-frequency structures. Under such conditions, methods that emphasize detection sensitivity tend to generate more positive responses, which can improve Pd but often leads to a substantial increase in false alarms.

In contrast, SCUA-Net is specifically designed to suppress clutter-induced pseudo-responses through the proposed GCC-CA module. By introducing contextual calibration into the attention generation process, the network can more effectively distinguish true targets from strong background interference. Consequently, some extremely weak targets located in highly ambiguous regions may be suppressed together with clutter responses, leading to a slightly lower Pd compared with several competing methods.

Nevertheless, this conservative decision strategy substantially reduces false positives, resulting in the lowest Fa and the highest IoU on IIRSTD-1k. Since IoU jointly reflects the balance among true positives, false positives, and false negatives, the significant reduction in false alarms outweighs the minor decrease in detected targets. These results suggest that SCUA-Net favors detection reliability and background robustness, which are particularly important in practical infrared early-warning systems where excessive false alarms can severely affect operational effectiveness.

Figure 4 further presents the precision–recall (PR) curves and receiver operating characteristic (ROC) curves of all methods. SCUA-Net outperforms the comparison methods over most threshold ranges and achieves the largest AUC, indicating stronger robustness under dynamic thresholds.

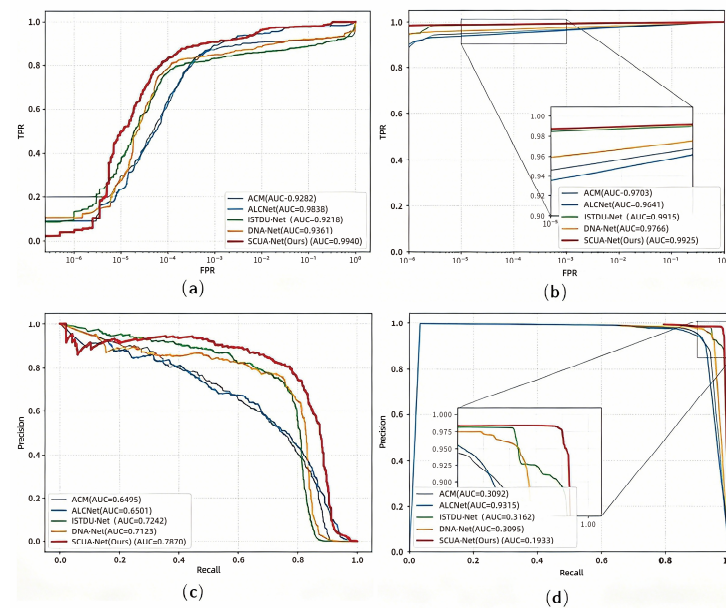


Figure 4. Comparison of ROC and PR curves on two benchmark datasets. (a) ROC curve on IRSTD-1k; (b) ROC curve on NUDT-SIRST; (c) PR curve on IRSTD-1k; (d) PR curve on NUDT-SIRST.

3.3.2. Qualitative Results

To more intuitively demonstrate the detection differences among methods in complex scenes, qualitative visual comparisons are further provided.

Figure 5 shows the qualitative results of different methods on IRSTD-1k and NUDT-SIRST. The first two rows correspond to complex scenes from IRSTD-1k, including strong clutter such as cloud edges and sea-surface ripples, while the last two rows correspond to typical scenes from NUDT-SIRST. From left to right, the figure shows (a) the original infrared image with an enlarged target region; (b) the ground truth (GT); (c) ACM; (d) ALCNet; (e) ISTDU-Net; (f) DNA-Net; and (g) the proposed SCUA-Net. The red box indicates the target region. Whereas ACM and ALCNet often generate extra false alarms in bright backgrounds or suffer from target misses and distorted boundaries, SCUA-Net yields cleaner maps with better shape recovery and effectively suppresses false positives caused by strong clutter.

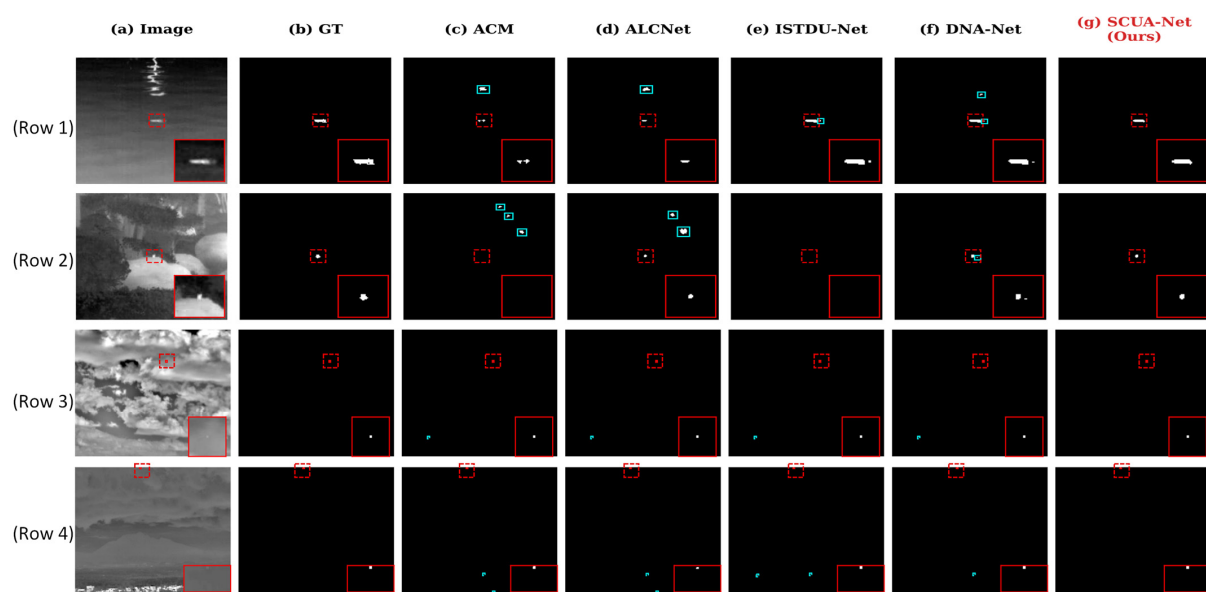


Figure 5. Visual qualitative results of different methods on the IRSTD-1k and NUDT-SIRST datasets. Red dashed boxes indicate the target regions and their zoomed-in views.

3.4. Complexity and Efficiency Analysis

In addition to detection accuracy, the parameter count, computational cost, and inference speed of the models are also important indicators of practical value. Table 2 summarizes the number of parameters, floating-point operations, and inference FPS of each method under the same testing setting. SCUA-Net has 9.22 M parameters, 34.87 G FLOPs, and an inference speed of 161.6 FPS. Although its parameter count and computational cost are higher than those of lightweight methods, it still maintains a real-time inference speed of 161.6 FPS, which meets the real-time processing requirements of practical infrared early-warning systems.

Table 2. Comparison of model complexity.

Method	Params (M) ↓	FLOPs (G) ↓	FPS ↑
ACM	0.15	2.26	2405.0
ALCNet	0.22	9.25	1642.1
ISTDU-Net	8.29	15.04	606.0
DNA-Net	0.56	5.22	772.2
SCUA-Net (Ours)	9.22	34.87	161.6

The SCUA-Net prioritizes detection robustness and false-alarm suppression over pursuing an extreme lightweight design. While lightweight models often struggle to suppress pseudo-responses in highly cluttered backgrounds, the proposed method achieves significantly lower false alarm rates through its densely nested architecture and contextual calibration. Consequently, this leads to an increase in parameter count and computational overhead. Nevertheless, this trade-off is acceptable for practical systems. With an inference speed of 161.6 FPS on the RTX 4090 platform, SCUA-Net successfully meets real-time processing constraints while delivering superior anti-noise capability and detection reliability.

4. Discussion

4.1. Ablation Study

To verify the effectiveness of the core modules in SCUA-Net, ablation experiments were conducted on both NUDT-SIRST and IRSTD-1k. The ablation settings include using only the basic U-Net++ backbone, adding the GCC module, adding the CA module, and the complete model with both GCC and CA. The corresponding results are listed in Table 3.

Table 3. Ablation study of different components in SCUA-Net. Baseline denotes the original U-Net++ without attention mechanisms. The best results are highlighted in bold.

Datasets	Baseline (U-Net++)	GCC Module	CA Module	Pd (%) ↑	Fa ($\times 10^{-6}$) ↓	IoU ↑
NUDT-SIRST	✓			98.63	1.90	0.891
	✓	✓		96.99	0.92	0.926
	✓		✓	99.12	1.56	0.922
	✓	✓	✓	99.15	0.558	0.957
IRSTD-1k	✓			92.26	113.018	0.577
	✓	✓		93.27	67.981	0.625
	✓		✓	92.26	53.160	0.604
	✓	✓	✓	91.87	52.210	0.628

On NUDT-SIRST, introducing either GCC or CA can improve the model performance to different extents, where GCC more effectively reduces the false alarm rate, while CA is more beneficial for preserving target contours and boundary information. When these two

modules are used together, the model achieves the best overall performance, indicating that global-context calibration and coordinate localization are complementary.

On IRSTD-1k, the baseline model exhibits a clearly higher Fa, indicating that the densely nested structure alone is insufficient to suppress false alarms in complex bright backgrounds. After adding GCC, the Fa decreases significantly. After adding CA, the model's capability for position modeling is enhanced. The complete model achieves the best Fa and IoU, indicating that the synergistic design of GCC-CA is particularly important for background suppression in complex scenes.

SCUA-Net demonstrates stronger target preservation and a lower false alarm rate on NUDT-SIRST, while showing stronger complex-background suppression and more stable performance on IRSTD-1k. Its advantages mainly stem from three aspects: the preservation of weak target details by the densely nested structure, the calibration of local-window context by GCC-CA, and the focused learning of hard negative samples by the joint OHEM + Dice optimization. The experimental results indicate that the proposed method achieves a relatively balanced trade-off among accuracy, robustness, and real-time performance.

4.2. Noise Robustness Analysis

In practical infrared sensing systems, imaging quality is frequently degraded by thermal noise and environmental perturbations. To further evaluate the generalization capability of the proposed method under degraded imaging conditions, additive zero-mean Gaussian noise with varying standard deviations ($\sigma_n = 1, 2$, and 3) was injected into the test images. The pre-trained SCUA-Net was directly evaluated without any noise-specific retraining or fine-tuning. The quantitative results are presented in Table 4.

Table 4. Evaluation of noise robustness on the NUDT-SIRST dataset under different Gaussian noise standard deviations (σ_n).

Noise Std. (σ_n)	Pd (%) $\uparrow$	Fa ($\times 10^{-6}$) $\downarrow$	IoU $\uparrow$
Clean Baseline	98.66	0.598	0.9552
Gaussian ($\sigma_n = 1$)	98.71	0.693	0.9508
Gaussian ($\sigma_n = 2$)	97.43	0.957	0.9230
Gaussian ($\sigma_n = 3$)	89.28	0.797	0.8203

As shown in Table 4, SCUA-Net maintains stable detection performance under mild Gaussian noise ($\sigma_n = 1, 2$), with only marginal performance degradation observed. When the noise level further increases to $\sigma_n = 3$, the detection probability (Pd) decreases more noticeably because the contrast between extremely small infrared targets and the cluttered background is substantially weakened, causing faint targets to be physically submerged by noise.

However, it is noteworthy that the false alarm rate (Fa) does not exhibit a severe surge at $\sigma_n = 3$, but rather remains effectively bounded at a low level (0.797×10^{-6}). This robustness can be attributed to the proposed GCC-CA module, which generates attention weights based on aggregated contextual statistics rather than isolated local pixel responses, thereby alleviating the perturbation of random noise spikes. Furthermore, the joint OHEM-based optimization enforces a strict decision boundary that conservatively suppresses highly ambiguous noisy inputs rather than generating massive false responses. Although the overall performance gradually decreases under severe noise, the proposed method still preserves satisfactory target detection capability and high operational reliability without any additional denoising preprocessing, demonstrating reasonable robustness in practical infrared imaging scenarios.

4.3. Interpretability and Feature Visualization

To provide a more intuitive understanding of the internal representation learned by SCUA-Net and to demystify its decision-making process, we further visualize the deep semantic feature activations immediately before the final segmentation head. Specifically, a forward hook was employed to extract the high-dimensional feature tensor, which was subsequently projected into a two-dimensional activation map through channel-wise average pooling. The normalized activation map was then superimposed on the corresponding infrared image using a standard pseudo-color rendering pipeline to illustrate the spatial distribution of discriminative features.

Figure 6 presents a representative example of the learned feature responses in a complex cluttered scene. In conventional convolutional networks, sharp cloud edges typically induce massive diffuse activations, leading to extensive false-positive regions. In contrast, SCUA-Net produces highly concentrated activations precisely around the true infrared target (represented by the warm/red core) while effectively suppressing the vast background clutter into a uniform, low-activation state (represented by the cold/dark blue regions).

Such interpretable behavior is consistent with the design objective of the proposed Global-Context Calibration Coordinate Attention (GCC-CA) module. By systematically integrating global contextual statistical bias to calibrate local responses, the network autonomously penalizes the energy coefficients of non-target high-brightness fluctuations. Although this visualization is qualitative, it provides intuitive evidence supporting the effectiveness of the proposed clutter-clearing mechanism and perfectly complements the quantitative experimental results presented in the previous sections.

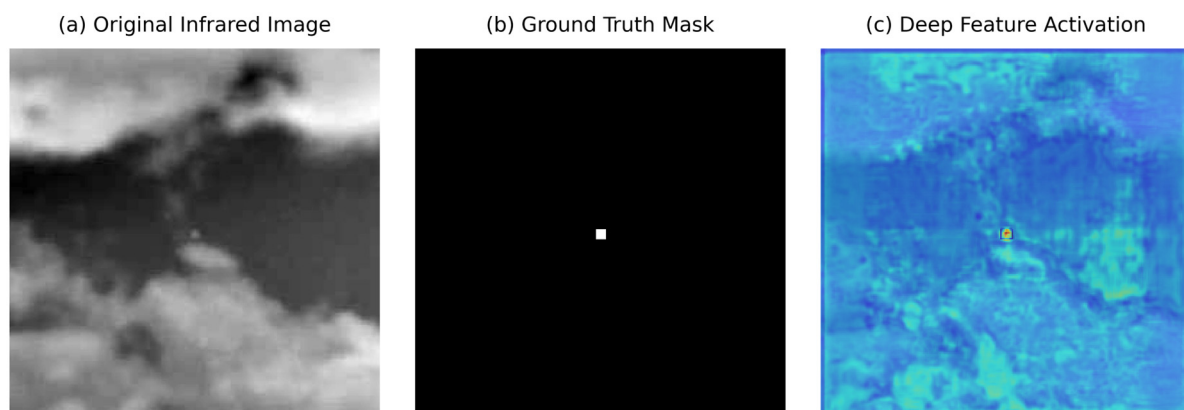


Figure 6. Visualization of deep semantic feature activations learned by the proposed SCUA-Net. (a) Original infrared image containing severe cloud clutter. (b) Ground-truth annotation. (c) Deep feature activation map superimposed on the original image. The network generates highly concentrated feature responses (warm colors) around the true infrared target while effectively suppressing complex background clutter (cold colors).

5. Conclusions

This paper proposed SCUA-Net to address target submergence, a high false alarm rate, and the loss of local context issues in ISTD under complex cluttered backgrounds. The method adopts a U-Net++-style densely nested encoder–decoder architecture to enhance multi-scale interaction between shallow spatial details and deep semantic features. A GCC-CA module is further introduced to inject contextual statistics within local windows into the coordinate attention mechanism, thereby alleviating background false detections under local inference conditions. Meanwhile, a joint optimization strategy combining OHEM-BCE and Dice Loss is employed, and sliding-window inference with Gaussian weighted

fusion is used during testing to improve output continuity and localization stability for high-resolution infrared images.

Experimental results show that SCUA-Net achieves superior performance on both the NUDT-SIRST and IRSTD-1k datasets. On NUDT-SIRST, the model obtains a detection probability of 99.15%, a false alarm rate of 0.558×10^{-6} , and an IoU of 0.9570, demonstrating that the proposed network can preserve weak target details while effectively suppressing complex background interference. On IRSTD-1k, the false alarm rate is further reduced to 52.210×10^{-6} , with an IoU of 0.628, indicating that the model still maintains strong background suppression and structural restoration capability in more demanding real-world complex scenes. Meanwhile, the inference speed on the RTX 4090 platform reaches 161.6 FPS, showing that the proposed method also satisfies real-time requirements to a certain extent. Notably, SCUA-Net's primary strength lies in substantially reducing false alarms while preserving stable performance—a priority well suited to practical infrared early-warning systems. SCUA-Net offers an effective approach to contextual calibration for ISTD in challenging environments.

Author Contributions: Conceptualization, S.L. and X.W. (Xiaoyan Wu); methodology, J.L. and X.W. (Xiaoyan Wang); software, J.L. and X.W. (Xiaoyan Wang); validation, J.L. and X.W. (Xiaoyan Wang); investigation, J.L., Z.C. and X.W. (Xiaoyan Wang); data curation, J.L., S.L. and X.W. (Xiaoyan Wang); writing—original draft preparation, J.L., Z.C. and S.L.; writing—review and editing, X.W. (Xiaoyan Wu), and J.P.; visualization, J.L.; supervision, S.L. and X.W. (Xiaoyan Wu); project administration, J.P.; funding acquisition, S.L., X.W. (Xiaoyan Wu) and Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Presidential Foundation of CAEP, grant number YZJJZQ 2022001; National Natural Science Foundation of China, grant number 62375092; Natural Science Foundation of Fujian Province, grant number 2025J01155.

Data Availability Statement: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

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

References

1. Cheng, Y.; Lai, X.; Xia, Y.; Zhou, J. Infrared dim small target detection networks: A review. *Sensors* **2024**, *24*, 3885. [[CrossRef](#)] [[PubMed](#)]
2. 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](#)]
3. Zeng, M.; Li, J.; Peng, Z. The design of top-hat morphological filter and application to infrared target detection. *Infrared Phys. Technol.* **2006**, *48*, 67–76. [[CrossRef](#)]
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. Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; Peng, Z. Infrared small target detection via non-convex rank approximation minimization joint $l_{2,1}$ norm. *Remote Sens.* **2018**, *10*, 1821. [[CrossRef](#)]
6. Zhang, T.; Wu, H.; Liu, Y.; Peng, L.; Yang, C.; Peng, Z. Infrared small target detection based on non-convex optimization with L_p -norm constraint. *Remote Sens.* **2019**, *11*, 559. [[CrossRef](#)]
7. 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, Munich, Germany, 5–9 October 2015; pp. 234–241. [[CrossRef](#)]
8. 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 (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 949–958. [[CrossRef](#)]
9. 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](#)]
10. 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.* **2023**, *32*, 1745–1758. [[CrossRef](#)] [[PubMed](#)]

11. 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](#)] [[PubMed](#)]
12. Pang, D.; Shan, T.; Ma, Y.; Ma, P.; Hu, T.; Tao, R. LRTA-SP: Low-Rank Tensor Approximation with Saliency Prior for Small Target Detection in Infrared Videos. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 2644–2658. [[CrossRef](#)]
13. Pang, D.; Shan, T.; Li, W.; Ma, P.; Tao, R.; Ma, Y. Facet Derivative-Based Multidirectional Edge Awareness and Spatial-Temporal Tensor Model for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5001015. [[CrossRef](#)]
14. Pang, D.; Ma, P.; Shan, T.; Li, W.; Tao, R.; Ma, Y.; Wang, T. STTM-SFR: Spatial-Temporal Tensor Modeling with Saliency Filter Regularization for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623418. [[CrossRef](#)]
15. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 3–11. [[CrossRef](#)] [[PubMed](#)]
16. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
17. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
18. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2017; pp. 240–248. [[CrossRef](#)] [[PubMed](#)]
19. Lin, M.; Chen, Q.; Yan, S. Network in network. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
20. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
21. Howard, A.G.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [[CrossRef](#)]
22. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [[CrossRef](#)]
23. 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 (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 877–886. [[CrossRef](#)]
24. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
25. Hou, Q.; Zhang, L.; Tan, F.; Xi, Y.; Zheng, H.; Li, N. ISTDU-Net: Infrared small-target detection U-Net. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7506205. [[CrossRef](#)]

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.