U2TSG-Net: Infrared Small Target Detection via Nested U-Structure with Target-Sensitive Attention and Gated Interaction

1st Zewen Wang
School of Computer Science and Technology, Xinjiang University
Urumqi, China
107552404017@stu.xju.edu.cn

2nd Boyuan Li
School of Computer Science and Technology, Xinjiang University
Urumqi, China
liboyuan@stu.xju.edu.cn

3rd Shuai Li
School of Computer Science and Technology, Xinjiang University
Urumqi, China
18712599865@163.com

4th Shengbin Hao
School of Computer Science and Technology, Xinjiang University
Urumqi, China
107552403792@stu.xju.edu.cn

5th Alimjan Aysa*
School of Computer Science and Technology, Xinjiang University
Urumqi, China
alim@xju.edu.cn

6th Kurban Ubul*
School of Computer Science and Technology, Xinjiang University
Urumqi, China
kurbanubul@sina.com

***Abstract*—Infrared small target detection (IRSTD) plays a pivotal role in precision guidance and surveillance systems. The submersion of weak infrared signals by complex background clutter results in extremely low signal-to-noise ratios (SNR), thereby making the accurate detection of dim and small targets a significant challenge. Existing U-shaped networks often suffer from deep feature dilution during continuous downsampling, causing faint targets to be submerged. Furthermore, direct feature fusion via skip connections often introduces background clutter from shallow layers into the decoder, leading to false alarms. To address these dual challenges, we propose U2TSG-Net, a novel nested U-structure network. Specifically, in the encoding stage, we propose a target sensitive selective attention (TSSA) module. By incorporating intensity priors with sparse attention, TSSA enforces target-sensitive feature preservation, effectively preventing small targets from being lost in deep layers. In the decoding stage, we propose a gated cross-level feature interaction (GCFI) module. It generates a spatial gate under the guidance of deep semantics, thereby selectively injecting target context into shallow features and suppressing activations caused by clutter. Extensive experiments on public benchmarks, including NUAA-SIRST, NUDT-SIRST, and IRSTD-1k, demonstrate that U2TSG-Net surpasses state-of-the-art methods, fully validating the superior robustness and practical value of the proposed architecture in complex backgrounds.**

***Index Terms*—Infrared small target detection, feature interaction, attention mechanism, prior guidance**

I. Introduction

Infrared detection systems, benefiting from passive sensing, strong anti-interference capability, and all-weather operability [1]–[3], play an indispensable role in precision guidance, wide-area surveillance, and coastal defense early warning [4]–[8]. However, constrained by atmospheric scattering in long-range imaging and sensor thermal noise, infrared small targets in images often exhibit the "dim and small" characteristics: the target typically occupies fewer than 9×9 pixels within the local field of view and lacks salient shape and texture cues. More critically, bright cloud edges, sea-surface clutter, and man-made thermal sources can easily submerge faint target signals, resulting in an extremely low signal-to-noise ratio (SNR) between the target and the background. Therefore, achieving robust infrared small target detection (IRSTD) in complex scenes remains a major challenge in the computer vision community.

To tackle the above challenges, researchers have developed a diverse range of detection algorithms. Early infrared small target detection methods were predominantly traditional, model-driven single-frame approaches, which can be broadly categorized into filtering-based background suppression methods such as Top-hat [9] and Butterworth filtering [10], local contrast-based methods inspired by the human visual system such as LCM [11] and its variants [12]–[15], and low-rank representation methods such as IPI [16], RIPT [17], and PSTNN [18]. Although these methods achieved good performance in simple backgrounds, they rely heavily on manually designed features, leading to insufficient adaptability and stability when facing varying target scales and environmental changes.

With the rapid advancement of deep learning, CNN-based methods have gradually become the dominant paradigm [19]–[22]. Wang et al. [23] exploit adversarial learning to balance missed detections and false alarms. DNANet, proposed by Li et al. [24], addresses deep-layer target disappearance via dense nested interactions. AGPCNet, introduced by Zhang et

*Corresponding authors.

This study was supported by the National Natural Science Foundation of China (62061045), the Xinjiang Uygur Autonomous Region Tianshan Ying Talents Leading Talents Program for Scientific and Technological Innovation (2023TSYCLJ0025), and the National Natural Science Foundation of China (62266044), and the project "Research on Infrared Small Target Detection Methods under Resource-Constrained Conditions" (XJ2026G044).

al. [25], couples an attention mechanism with pyramid context aggregation. To alleviate the difficulty of extracting accurate target contours under low contrast, Zhang et al. [26] propose ISNet, which innovatively integrates a shape reconstruction task into the detection framework. SCTransNet, presented by Yuan et al. [27], employs a spatial–channel cross-attention mechanism to achieve full-level feature interaction, substantially strengthening global context modeling. Shen et al. [28] propose GCLNet, which fuses local feature learning with a multiscale graph reasoning mechanism, enabling effective capture of cross-scale global context while preserving fine-grained details.

Despite the substantial progress achieved by existing deep learning methods, several issues remain: 1) Infrared small targets are highly susceptible to losing spatial details under successive downsampling operations, causing their responses to be ”diluted” or even completely vanish in deep semantic feature maps; consequently, the network struggles to reconstruct target locations during decoding. 2) Conventional skip connections directly concatenate shallow features into the decoder. While this supplements spatial information, it also indiscriminately introduces a large amount of target-like background clutter, such as single-pixel noise spikes and bright cloud edges, thereby severely degrading segmentation accuracy. To address the above issues, we propose U2TSG-Net, a nested U-structure network featuring target sensitive attention and gated interaction. We build a nested residual encoder–decoder architecture, aiming to capture targets finely within multi-scale spatial representations. Specifically, during encoding, we design the target sensitive selective attention (TSSA) module, which strengthens local responses to counteract feature attenuation caused by downsampling, thereby ensuring that tiny targets are preserved in deep feature representations. During cross-level connections, we propose the gated cross-level feature interaction (GCFI) module, which abandons the conventional direct concatenation strategy and instead employs a gating mechanism to adaptively suppress background clutter in shallow features, allowing only effective information that facilitates target recovery to flow into the decoder.

The main contributions of this work are summarized as follows:

- We propose a TSSA, which leverages a selective attention mechanism to dynamically enhance feature responses in target regions, substantially reducing the risk of missed detections caused by successive downsampling.
- We propose a GCFI, establishing a cross level feature filtering mechanism. This module effectively distinguishes target details from background clutter in shallow features, thereby providing clean feature inputs for high-precision segmentation.
- We propose the U2TSG-Net, a nested U-shaped detection framework tailored to the characteristics of infrared small targets. Compared to other state-of-the-art algorithms, our algorithm achieved better detection results on different datasets.

II. PROPOSED METHOD

A. Overview

As illustrated in Fig. 1, to address the issues where target cues are prone to being submerged during successive downsampling and complex background clutter causes severe interference, we construct U2TSG-Net as a nested U-shaped architecture built upon residual nested blocks (ReSidual U-blocks, RSU) [29]. Unlike plain convolution blocks, the RSU employs an intra-stage nested U-structure to extract multi-scale context while preserving high-resolution local details via residual connections. This design enables the network to capture robust semantic features without sacrificing the spatial information essential for small targets. Specifically, as shown in Fig. 2, we employ standard RSU-L blocks, where L denotes the number of layers in the encoder, in the shallow-to-mid stages to extract local details, while deploying the dilated RSU-4F block at the bottleneck to capture global context without reducing resolution.

Based on this backbone, the proposed U2TSG-Net processes the infrared image through a multi-scale encoding-decoding stream. In the encoding stage, a TSSA module is embedded at the end of each level. By exploiting intensity priors, TSSA dynamically highlights target regions prior to downsampling to counteract feature loss. Subsequently, in the decoding path, where spatial resolution is progressively restored, the GCFI module replaces standard skip connections. It leverages deep semantics to gate shallow features, effectively suppressing clutter. Finally, deep supervision is applied to multiscale predictions to accelerate convergence.

B. Target Sensitive Selective Attention Module

Preventing faint target cues from being submerged by background clutter is critical. Standard channel attention relies on Global Average Pooling (GAP), which indiscriminately averages small targets with vast backgrounds, leading to severe feature dilution. To address this issue, we propose the TSSA module. As illustrated in Fig. 3, unlike spatial ranking approaches, TSSA introduces a channel-wise intensity prior via spatial Top-k pooling. Specifically, for an input feature map X, we compute a channel saliency score S by averaging the k highest-intensity spatial responses within each channel. This peak-preserving statistic highlights target-responsive channels. Crucially, S is used to modulate the Query and Key projections before affinity computation, encouraging the interaction modeling to prioritize target-relevant cues.

To further strengthen feature representation, we construct a multi-granularity sparse attention mechanism comprising four parallel branches. We argue that a single fixed sparsity level is insufficient to capture the variations of small target signatures. To address this, as illustrated in the dashed box, we deploy four distinct branches operating with specific selection thresholds labeled as k_1 through k_4, where k_1 corresponds to a 50% retention ratio. This quad-branch design enables the network to simultaneously explore target correlations at different concentration levels. By fusing outputs from these

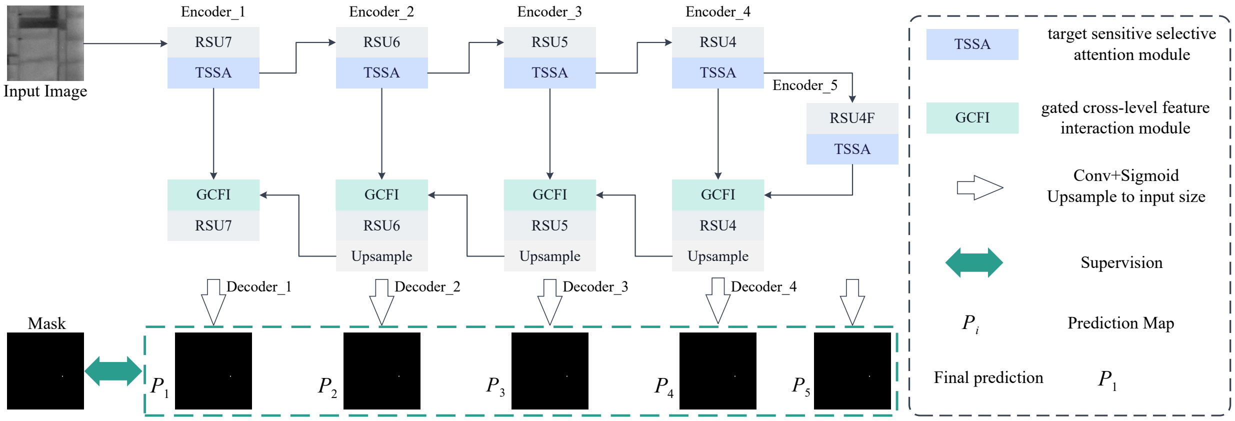


Fig. 1: Architecture of the proposed U2TSG-Net

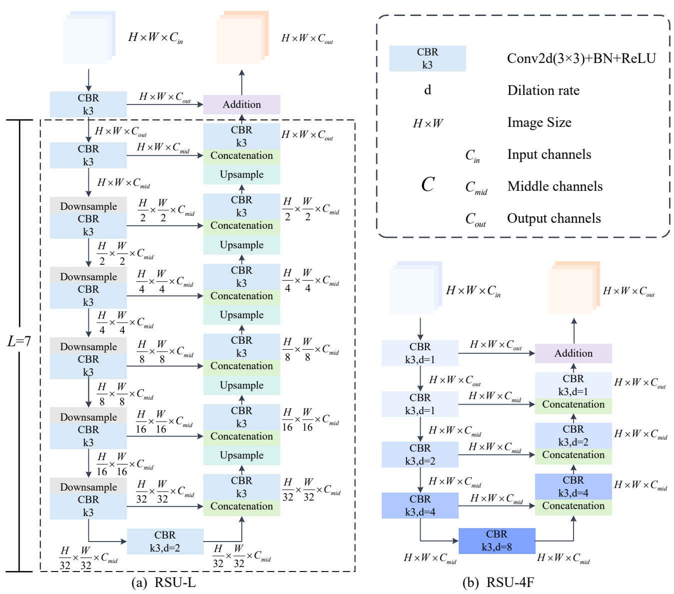


Fig. 2: Illustration of the ReSidual U-blocks (RSU). (a) RSU-L. (b) RSU-4F.

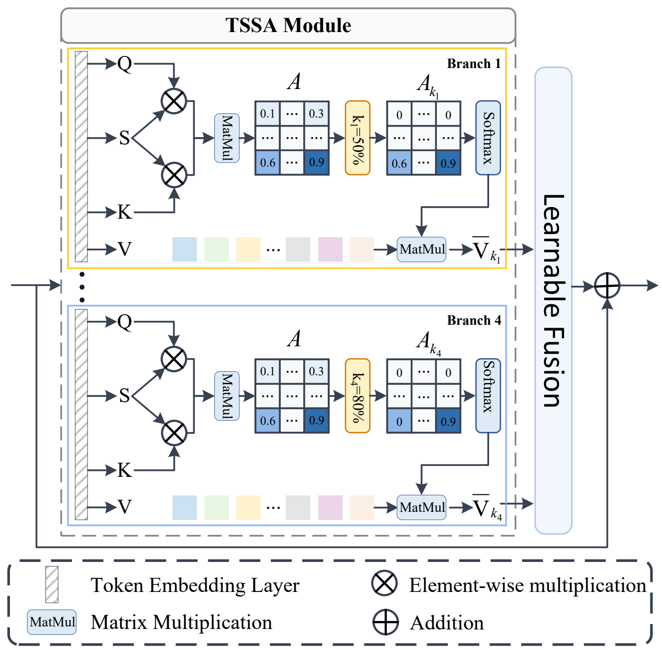


Fig. 3: Illustration of the TSSA module. Here, S denotes the channel saliency score computed via Top-k average pooling, which is used to modulate the Query and Key projections to highlight target-dominant channels. The dashed box depicts the generation and application procedure of the multi-granularity sparse masks.

diverse granularities, the module explicitly blocks interference pathways from noise-dominated channels while aggregating comprehensive target cues from highly informative ones.

Formally, given the input X, we generate the Query (Q), Key (K), and Value (V) projections via learnable 1×1 convolutional projections. To enforce target sensitivity, we compute the intensity prior S via Top-k_{pool} pooling to modulate Q and K:

$$\begin{cases} S = \text{Avg}(\mathcal{P}_{pool}(|X|)) \\ \hat{Q} = Q \odot S \\ \hat{K} = K \odot S \end{cases} \tag{1}$$

where $\mathcal{P}_{pool}$ extracts the largest spatial elements, and $\odot$ denotes channel-wise multiplication.

Subsequently, we perform multi-granularity interaction via four parallel branches. For the i-th branch with a specific selection threshold k_i , the sparse attention map A_{k_i} is derived as:

$$A_{k_i} = \text{SparseSoftmax}(\hat{Q}\hat{K}^T, k_i) \tag{2}$$

Here, $\text{SparseSoftmax}(\cdot, k_i)$ retains only the top-k_i strongest correlations per row and masks others to negative infinity. Finally, the outputs from all branches are aggregated to yield the final refined feature Y:

$$\mathbf{Y} = \mathbf{X} + \mathcal{F}_{\text{fuse}}\left(\text{Concat}\left[\mathbf{A}_{k_1}\mathbf{V}, \mathbf{A}_{k_2}\mathbf{V}, \mathbf{A}_{k_3}\mathbf{V}, \mathbf{A}_{k_4}\mathbf{V}\right]\right) \tag{3}$$

where $\text{Concat}[\cdot]$ denotes channel-wise concatenation and $\mathcal{F}_{fuse}$ represents the fusion convolution.

C. Gated Cross-Level Feature Interaction Module

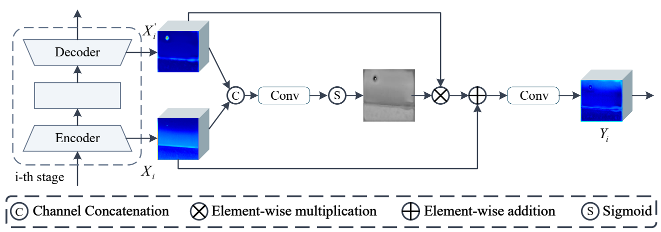


Fig. 4: Illustration of the GCFI module.

Cross-level skip connections are vital for recovering spatial details but suffer from the semantic discrepancy between shallow (X_i) and deep (X_i') features. Naive concatenation indiscriminately propagates high-frequency clutter from shallow layers, leading to false alarms. To bridge this gap, we propose the GCFI module, as illustrated in Fig. 4. Unlike direct fusion, GCFI employs a residual gating mechanism, leveraging deep semantics as a spatial filter to suppress background noise while preserving target-relevant details.

Specifically, we first upsample the deep feature X_i' to match the resolution of X_i, yielding X_i'. To assess pixel-wise semantic consistency, a spatial gating map $\mathcal{G}$ is generated by concatenating the features followed by a 1×1 convolution:

$$\mathcal{G} = \sigma\left(\mathcal{F}_{gate}([X_i', X_i])\right) \in [0,1]^{C\times H\times W} \tag{4}$$

where $[\cdot]$ denotes concatenation and σ is the Sigmoid function. $\mathcal{G}$ serves as a soft mask, highlighting potential targets while suppressing clutter.

Finally, we adopt a selective injection strategy, adding the gated deep semantics to the shallow features via a residual connection:

$$Y_i = \mathcal{F}_{fuse}\left(X_i + \mathcal{G} \odot X_i'\right) \tag{5}$$

where $\odot$ denotes element-wise multiplication. This design effectively supplements high-resolution spatial information while blocking the cascading diffusion of background noise.

III. EXPERIMENTAL RESULTS

A. Experimental Settings

1) Datasets: To evaluate the effectiveness of U2TSG-Net, we utilize three widely-adopted public datasets: NUAA-SIRST [30], NUDT-SIRST [24], and IRSTD-1k [26]. These datasets encompass diverse and challenging scenarios, offering high-quality annotations and rich scene variability to enable a persuasive assessment of the method's performance across different conditions.

2) Evaluation Metrics: Following the standard protocol in infrared small target detection, the detection performance is quantitatively evaluated from both pixel-level and target-level perspectives. Specifically, we adopt pixel-level metrics including intersection over union (IoU), normalized IoU (nIoU) and F-measure (F1), as well as target-level metrics including the probability of detection (P_d), false-alarm rate (F_a), and the receiver operating characteristic (ROC) curve.

3) Implementation Details: The proposed method is implemented using the PyTorch deep learning framework. Regarding dataset partitioning, NUAA-SIRST and IRSTD-1k follow an 8:2 split, while NUDT-SIRST adheres to its official protocol with 663 training images and 664 testing images. Prior to training, all input images are resized to 256×256 pixels and augmented via random flipping and rotation. The network parameters are optimized using the AdamW optimizer with a batch size of 8. The initial learning rate is set to 1×10^{-3} and adjusted via the CosineAnnealingLR scheduler over 1,500 training epochs. Finally, all experiments were conducted on a workstation equipped with an Intel(R) Core(TM) i9-7980XE CPU @ 2.60GHz and a single NVIDIA GeForce RTX 3090 Ti GPU.

B. Ablation Study

To validate the effectiveness of the core components in our method, we conduct extensive ablation studies on the NUAA-SIRST dataset. We adopt U2-Net as the baseline model and progressively integrate the TSSA module and GCFI module to quantify their individual contributions. The detailed quantitative comparisons are summarized in Table I.

TABLE I: Ablation results of different core modules on the NUAA-SIRST dataset.

Method	TSSA	GCFI	$IoU\uparrow$	$nIoU\uparrow$	$F1\uparrow$	$P_d\uparrow$	$F_a\downarrow$
U2-Net			78.32	77.75	87.84	98.17	17.41
U2-Net	✓		82.36	82.26	90.32	100.00	16.89
U2-Net		✓	81.74	80.79	89.95	99.08	14.85
U2-Net	✓	✓	**83.65**	**83.73**	**91.09**	**100.00**	**8.02**

1) Effectiveness of TSSA module: As reported in the second row of Table I, introducing TSSA at the encoder stage yields a notable improvement of 4.04% and 4.51% in IoU and nIoU, respectively. This substantial gain validates the critical role of TSSA in mitigating feature dilution. Unlike standard operations that risk discarding faint signals, TSSA employs a multi-granularity sparse attention mechanism. By aggregating features from multiple branches with distinct retention ratios, this design allows the network to capture both local dominant cues and global contextual dependencies simultaneously. Consequently, this strategy not only prioritizes target-dominant features but also explicitly blocks the propagation of background noise, effectively preventing faint target cues from being submerged by complex clutter in deep feature spaces.

2) Effectiveness of GCFI module: Integrating GCFI boosts IoU by 3.42% and lowers the F_a from 17.41×10^{-6} to 14.85×10^{-6} validate its strong background suppression capability. By employing deep semantics to perform pixel-wise semantic purification on shallow features, GCFI effectively bridges the cross-level semantic gap. This process filters out high-frequency clutter while preserving fine-grained target details for precise boundary recovery.

TABLE II: Quantitative comparison results of different methods on the three datasets. Here, IoU, $nIoU$, $F1$, and P_d are reported in percentage (%), and F_a is reported in units of 10^{-6}. The symbols ↑ and ↓ indicate that higher and lower values are better, respectively. In the table, the best and second-best results in each column are highlighted in **bold** and underlined font, respectively.

Method	NUAA-SIRST					NUDT-SIRST					IRSTD-1k					#Params.	FLOPs
	$IoU\uparrow$	$nIoU\uparrow$	$F1\uparrow$	$P_d\uparrow$	$F_a\downarrow$	$IoU\uparrow$	$nIoU\uparrow$	$F1\uparrow$	$P_d\uparrow$	$F_a\downarrow$	$IoU\uparrow$	$nIoU\uparrow$	$F1\uparrow$	$P_d\uparrow$	$F_a\downarrow$		
TopHat [9]	8.45	15.19	15.59	86.24	4682.42	20.00	23.30	33.34	87.83	1444.05	9.35	9.71	17.10	78.45	1419.62	-	-
IPI [16]	34.93	58.10	51.78	95.41	456.50	36.05	54.41	52.99	90.16	560.83	11.81	27.77	21.12	86.53	845.08	-	-
PSTNN [18]	24.99	33.60	39.99	71.56	25.26	21.55	31.37	35.46	68.36	232.21	16.93	25.07	28.94	61.28	92.37	-	-
DNANet [24]	79.55	80.33	88.61	99.08	26.96	94.92	95.74	97.39	**99.37**	8.46	66.78	64.34	80.08	88.55	18.50	10.43	44.00
ISNet [26]	76.77	75.69	86.86	95.41	50.68	73.56	76.43	84.77	97.88	9.95	68.27	61.08	81.14	91.92	20.61	1.09	30.64
UIUNet [31]	80.04	80.53	88.91	98.17	14.51	93.57	92.52	96.68	98.94	3.01	69.93	64.14	82.30	91.92	17.35	50.54	54.43
RDIAN [32]	77.93	80.98	87.60	98.17	8.53	83.96	85.11	91.28	98.31	11.03	62.80	62.44	77.14	89.90	26.91	**0.22**	**3.72**
SCTransNet [27]	79.62	81.75	88.65	98.17	18.43	95.70	95.91	97.80	99.26	**1.49**	66.12	66.45	79.59	92.93	16.74	11.33	10.12
MSHNet [33]	74.54	80.56	85.41	95.41	9.56	94.46	95.04	97.15	99.15	3.77	65.15	66.00	78.88	93.27	17.84	4.07	6.10
GCLNet [28]	82.86	82.04	90.63	**100.00**	10.58	93.01	92.57	96.38	98.62	2.32	70.10	67.03	82.42	**93.94**	30.80	19.04	22.53
ILNet [34]	76.81	78.20	86.88	98.17	17.41	90.38	90.22	94.95	98.20	6.09	68.06	63.81	80.99	91.92	29.26	1.58	21.04
U2TSG-Net(Ours)	**83.65**	**83.73**	**91.09**	**100.00**	**8.02**	**96.10**	**95.95**	**98.01**	99.05	1.91	**72.78**	**67.19**	**84.24**	92.93	**14.10**	19.54	23.74

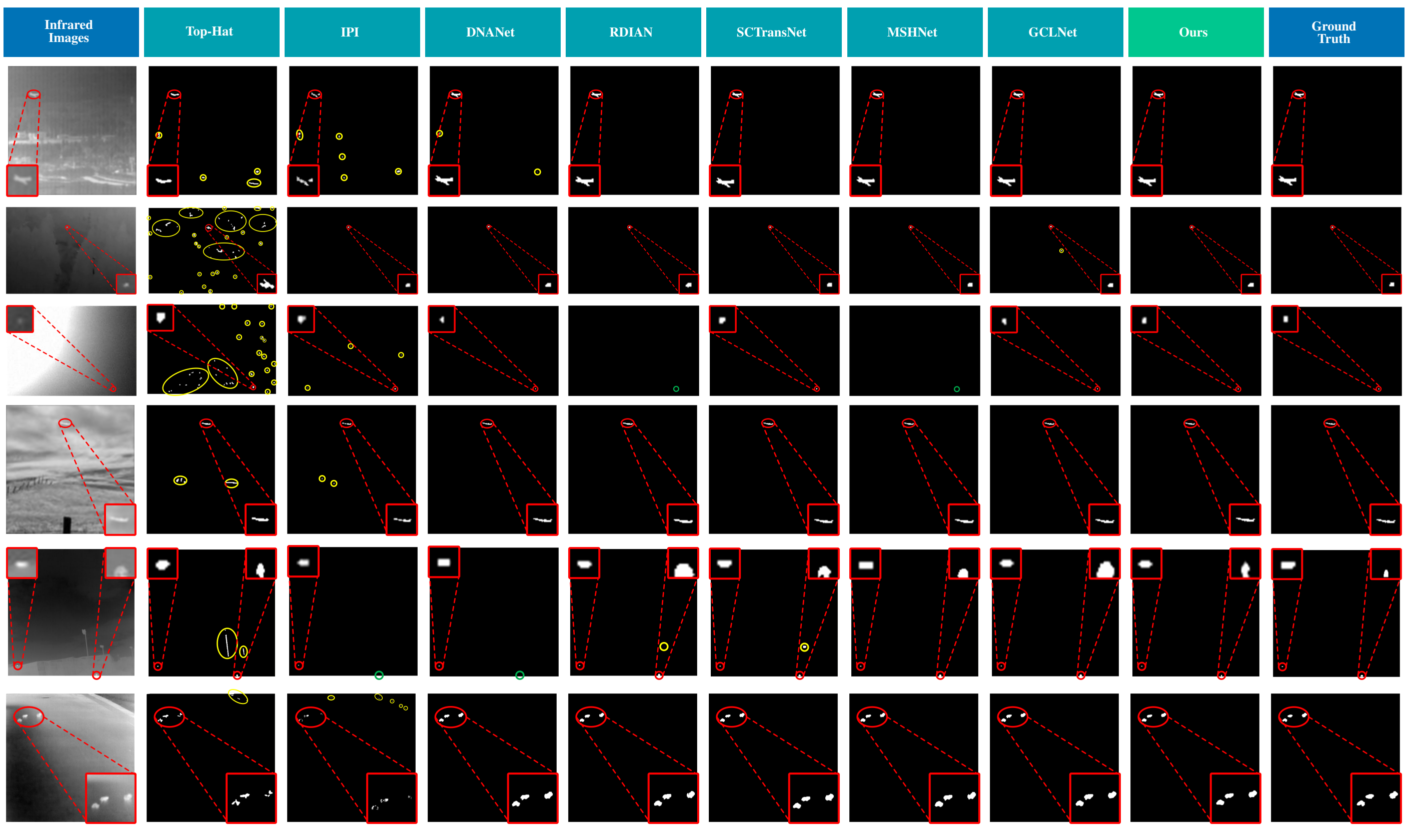
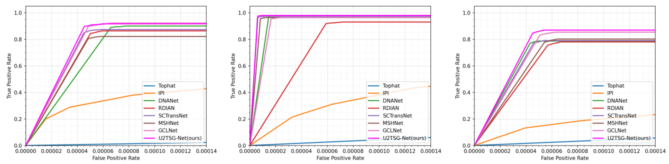


Fig. 5: Qualitative visualizations of different methods on three datasets. Correct detections, false alarms, and missed targets are highlighted with red, yellow, and green circles, respectively. For clearer visualization, a zoomed-in view of the target region is shown in the corner of each image.

C. Comparison to the State-of-The-Art Methods

1) Visual Result Analysis: Fig. 5 illustrates the visual segmentation results of different methods on representative images from the NUAA-SIRST, NUDT-SIRST, and IRSTD-1k datasets. As observed, competing methods suffer from missed detections and false alarms and perform poorly in preserving fine details at target edges. In contrast, our method demonstrates significant superiority across various highly challenging scenarios, encompassing complex background clutter (Rows 1 and 2), low SNR and extreme illumination conditions (Rows 3 and 4), as well as multi-target and tiny point-like target scenarios (Rows 5 and 6). Notably, as shown in Row 2, although GCLNet successfully detects the target under an inhomogeneous thick cloud background, it misclassifies background noise as targets. Conversely, U2TSG-Net accurately distinguishes targets from the background, effectively suppressing false alarms while preserving target responses.



(a) (b) (c)

Fig. 6: ROC curves of different methods on (a) NUAA-SIRST, (b) NUDT-SIRST, and (c) IRSTD-1k.

This benefits from the proposed TSSA and GCFI modules, which synergistically enhance faint target cues while filtering out background interference.

2) Numerical Result Analysis: To comprehensively evaluate the effectiveness of U2TSG-Net, we conduct extensive quantitative comparisons on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1k datasets. The detailed results are summarized in Table II. In addition, to further assess the performance under varying thresholds, Fig. 6 presents the receiver operating characteristic (ROC) curves of all compared methods.

As detailed in Table II, U2TSG-Net achieves leading performance across most metrics. In terms of segmentation, it exhibits clear superiority, particularly on the challenging IRSTD-1k dataset where it attains an IoU of 72.78%, surpassing the runner-up GCLNet score of 70.10% by 2.68%. Regarding detection capability, U2TSG-Net achieves the most favorable P_d–F_a trade-off. On NUAA-SIRST, it maintains 100% P_d with a minimal F_a of 8.02×10^{-6}. Notably, on IRSTD-1k, our F_a of 14.10×10^{-6} is less than half the rate of 30.80×10^{-6} reported by GCLNet. Collectively, these metrics demonstrate that U2TSG-Net effectively resolves the trade-off between detail preservation and noise suppression, ensuring robust performance in complex scenes. Fig. 6 compares the receiver operating characteristic (ROC) curves of different methods on the three benchmark datasets. As can be observed, the steeper slope and larger area under the curve (AUC) further confirm the effectiveness of our methods. These results strongly demonstrate that our approach achieves higher detection accuracy while maintaining a low false-alarm rate, indicating a clear performance advantage.

IV. CONCLUSION

In this paper, we propose a novel deep learning framework called U2TSG-Net to address the critical challenges of deep feature dilution and cross-level semantic discrepancy in the field of IRSTD. Specifically, the TSSA module leverages a multi-granularity sparse attention mechanism to enforce target-sensitive feature preservation, effectively preventing faint signals from being submerged during downsampling. Complementarily, the GCFI module utilizes deep semantics to perform spatial semantic purification on shallow features, bridging the semantic gap to suppress background clutter while recovering fine-grained details. Extensive experiments demonstrate that U2TSG-Net achieves state-of-the-art performance, providing a robust solution for accurate small target detection in complex scenarios.

REFERENCES

[1] H. Lv, B. Deng, and X. Li, "Research on image fusion technology of infrared and visible image based on mst and cnn," in *2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*. IEEE, 2022, pp. 1395–1399.

[2] Z. Xu, K.-S. Chen, X. Liu, and H. Guo, "Spatiotemporal variation of imaging swath in earth observing lunar-based sar under orbital perturbations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2023.

[3] B. Li, X. Li, S. Li, Y. Zhang, and K. Liu, "Adaptive feature fusion network for infrared small target detection," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2024, pp. 1–6.

[4] B. Li, Z. Ren, S. Hao, and K. Ubul, "Pgnet: Position guided infrared small target detection," *Knowledge-Based Systems*, vol. 332, p. 114669, Jan. 2026.

[5] H. Feng, H. Xuran, L. Bin, W. Haipeng, and Z. Decai, "Infrared image recognition technology based on visual processing and deep learning," in *2020 Chinese Automation Congress (CAC)*. IEEE, 2020, pp. 641–645.

[6] P. Qin, Y. Liu, G. Zhou, C. Tang, and J. Zhang, "Dual-stream infrared object detection network combined with saliency images," in *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*. IEEE, 2022, pp. 428–432.

[7] B. Li, Z. Ren, X. Li, and K. Ubul, "Edge-aware network with confidence feature fusion for infrared small target detection," in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, Jun. 2025, pp. 752–759.

[8] J. Yin, Z. Li, L. Cui, W. Zhang, Q. Wang, and G. Si, "Cyclegan-based visible-infrared image enhancement method for infrared power equipment object detection," in *2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS)*. IEEE, 2023, pp. 411–416.

[9] M. Zeng, J. Li, and Z. Peng, "The design of top-hat morphological filter and application to infrared target detection," *Infrared physics & technology*, vol. 48, no. 1, pp. 67–76, 2006.

[10] T. Zhao, T. Wang, Y. Cao, and Y. Cai, "Infrared dim and small target detection and tracking based on single multi-frame algorithm under sea clutter background," in *2019 Chinese Control Conference (CCC)*. IEEE, 2019, pp. 7912–7917.

[11] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 1, pp. 574–581, 2013.

[12] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4204–4214, 2016.

[13] J. Liu, Z. He, Z. Chen, and L. Shao, "Tiny and dim infrared target detection based on weighted local contrast," *IEEE geoscience and remote sensing letters*, vol. 15, no. 11, pp. 1780–1784, 2018.

[14] J. Nie, S. Qu, Y. Wei, L. Zhang, and L. Deng, "An infrared small target detection method based on multiscale local homogeneity measure," *Infrared Physics & Technology*, vol. 90, pp. 186–194, 2018.

[15] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern recognition*, vol. 58, pp. 216–226, 2016.

[16] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 4996–5009, 2013.

[17] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 10, no. 8, pp. 3752–3767, 2017.

[18] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sensing*, vol. 11, no. 4, p. 382, 2019.

[19] B. Li, X. Li, S. Li, Y. Zhang, K. Liu, J. Ma, and D. Wu, "Cross-layer feature guided multiscale infrared small target detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[20] S. Chen, L. Ji, J. Zhu, M. Ye, and X. Yao, "Sstnet: Sliced spatio-temporal network with cross-slice convlstm for moving infrared dim-small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[21] B. Li, Z. Ren, T. Baidi, S. Ma, and K. Ubul, "A lightweight multi-domain collaborative representation network for infrared small target detection," *Expert Systems with Applications*, p. 132334, 2026.

[22] ShengjieCheng, P. Zhou, YuLiu, HongjiMa, A. Aysa, and K. Ubul, "Research on knowledge distillation algorithm based on yolov5 attention mechanism," *Expert Systems with Applications*, vol. 240, p. 122553, Apr. 2024.

[23] H. Wang, L. Zhou, and L. Wang, "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*, 2019, pp. 8509–8518.

[24] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "Dense nested attention network for infrared small target detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 1745–1758, 2022.

[25] T. Zhang, S. Cao, T. Pu, and Z. Peng, "Agpcnet: Attention-guided pyramid context networks for infrared small target detection," *arXiv preprint arXiv:2111.03580*, 2021.

[26] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "Isnet: Shape matters for infrared small target detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 877–886.

[27] S. Yuan, H. Qin, X. Yan, N. Akhtar, and A. Mian, "Sctransnet: Spatial-channel cross transformer network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[28] Y. Shen, Q. Li, C. Xu, C. Chang, and Q. Yin, "Graph-based context learning network for infrared small target detection," *Neurocomputing*, vol. 616, p. 128949, 2025.

[29] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.

[30] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 950–959.

[31] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2022.

[32] H. Sun, J. Bai, F. Yang, and X. Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset irdst," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[33] Q. Liu, R. Liu, B. Zheng, H. Wang, and Y. Fu, "Infrared small target detection with scale and location sensitivity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 490–17 499.

[34] H. Li, J. Yang, R. Wang, and Y. Xu, "Ilnet: Low-level matters for salient infrared small target detection," *IEEE Transactions on Aerospace and Electronic Systems*, 2025.