ISNet: Shape Matters for Infrared Small Target Detection

Mingjin Zhang[1,4], Rui Zhang[1*], Yuxiang Yang[2], Haichen Bai[1], Jing Zhang[3*], Jie Guo[1]
[1] State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China, [2] Hangzhou Dianzi University, Hangzhou 310018, China
[3] The University of Sydney, NSW 2006, Australia, [4] JD Explore Academy, China
mjinzhang@xidian.edu.cn, woshizhangrui@stu.xidian.edu.cn, yyx@hdu.edu.cn,
hcbai@stu.xidian.edu.cn, jing.zhang1@sydney.edu.au, jguo@mail.xidian.edu.cn

Abstract

Infrared small target detection (IRSTD) refers to extracting small and dim targets from blurred backgrounds, which has a wide range of applications such as traffic management and marine rescue. Due to the low signal-to-noise ratio and low contrast, infrared targets are easily submerged in the background of heavy noise and clutter. How to detect the precise shape information of infrared targets remains challenging. In this paper, we propose a novel infrared shape network (ISNet), where Taylor finite difference (TFD)-inspired edge block and two-orientation attention aggregation (TOAA) block are devised to address this problem. Specifically, TFD-inspired edge block aggregates and enhances the comprehensive edge information from different levels, in order to improve the contrast between target and background and also lay a foundation for extracting shape information with mathematical interpretation. TOAA block calculates the low-level information with attention mechanism in both row and column directions and fuses it with the high-level information to capture the shape characteristic of targets and suppress noises. In addition, we construct a new benchmark consisting of 1,000 realistic images in various target shapes, different target sizes, and rich clutter backgrounds with accurate pixel-level annotations, called IRSTD-1k. Experiments on public datasets and IRSTD-1k demonstrate the superiority of our approach over representative state-of-the-art IRSTD methods. The dataset and code are available at github.com/RuiZhang97/ISNet.

1. Introduction

Infrared small target detection (IRSTD) has a wide range of important applications such as traffic management and marine rescue [8, 33, 37]. Misdetections in these fields may cause significant damage to multiple aspects of the real world. Therefore, the improvement of IRSTD is one of the priorities in both academic research and industrial division.

Compared to general object detection targets, infrared small targets have the following characteristics: *1) Dim*: Infrared images have lots of noises and clutter in the background, and the targets are easily submerged in the background, resulting in low contrast and low signal-to-clutter ratio (SCR). *2) Small*: Due to the long camera to object distance, infrared targets usually occupy only about one to ten pixels in the images. *3) Varying shape*: The shape and size of the target varies in different scenes and situations according to different target types.

To detect infrared small targets, researchers have developed several pioneering works based on image processing and machine learning techniques including filtering, human visual system (HVS), and low-rank representation. However, these traditional methods have some limitations. Filtering-based methods, such as top-hat filter [2] and max-median/max-mean filter [9], can only suppress uniform background clutters but cannot suppress complex background noises, resulting in high false alarm rates and unstable performance. As for the methods based on HVS, the spectral residuals-based method [16] can not efficiently suppress the clutters in the background. Local-contrast-based methods [4, 12] are only suitable for high contrast targets instead of dim targets. Low-rank representation-based methods [5, 10, 38, 39] can adapt to low SCR infrared images but still suffer from a high false alarm rate on images with small and varying-shape targets in complex backgrounds. In addition to the issues above, most traditional methods heavily rely on hand-crafted features, which is suboptimal and ineffective in dealing with challenging cases. Besides, the design of handcraft features and tuning of hyper-parameters require expert knowledge and a lot of engineering efforts.

With the success of deep learning in many fields, it offers novel solutions to the above problems. Convolution neural network (CNN) can efficiently extract features from

*Corresponding author.

infrared small targets owing to the data-driven and end-to-end learning paradigm. Liu *et al.* [19] applied multi-layer perception (MLP) and constructed a five-layer network for IRSTD. With a conditional generative adversarial network, Wang *et al.* [34] proposed MDvsFA for IRTD and achieved the balance between two metrics, i.e., miss detection v.s. false alarm. To extract contextual features from different layers, Dai *et al.* [6] proposed an asymmetric contextual modulation (ACM) feature fusion method (ACMNet). Although existing CNN-based IRSTD methods have yielded good results, they can only detect the presence of the small target in infrared images, while the contour of the detected targets is very blurred. In fact, the edge and shape information of infrared targets is not only critical for target classification tasks but also extremely important for practical applications such as marine rescue by providing useful clues to help recognize their types. Due to the low contrast and low SCR between infrared small target and background, it is difficult to extract useful edge and shape features of the target, especially from multiple feature levels, where deep layers may have clear semantics but lack fine details of edge and shape. How to obtain precise edge and shape of infrared small targets remains challenging and unexplored.

In this paper, we make an attempt to address this problem by exploring a new idea that incorporates the reconstruction of target shape into the detection of small infrared targets. Specifically, we devise a novel infrared shape network (ISNet) with two key components for IRSTD. First, we devise a Taylor finite difference (TFD)-inspired edge block to aggregate the edge features by drawing inspiration from the neural ordinary differential equation (Neural ODE) area, where the ODE is interpreted as a second-order Taylor finite difference equation. Then, we devise a two-orientation attention aggregation (TOAA) block to extract cross-level features by exacting the low-level features from both row and column directions and integrating them with high-level features. After that, the cross-level features are fed to the TFD-inspired edge block to reconstruct the target edges. By stacking multiple TFD-inspired edge blocks and TOAA blocks in a sequence, the long-range contextual information of the target can also be captured. Thus, the network can better locate the target and obtain the accurate shape of the targets. Besides, we apply a bottleneck structure to remove high-frequency noise in infrared images and enable a more informative flow through the network. In addition, we also construct a new benchmark consisting of 1,000 realistic images in various target shapes, different target sizes, and rich clutter backgrounds with accurate pixel-level annotations, called IRSTD-1k. Experimental results on the popular NUAA-SIRST dataset and IRSTD-1k demonstrate that the proposed ISNet outperforms state-of-the-art (SOTA) IRSTD methods in terms of false-alarm rate, probability detection rate, intersection over union (IoU) ratio, and normalized intersection over union (nIoU) ratio.

The contributions of this study can be summarized as:

- We propose a novel idea to address the challenges in IRSTD, i.e., incorporating the reconstruction of target shape into the detection of small infrared targets.
- We devise two key components named TFD-inspired edge block and TOAA block to efficiently extract edge features and aggregate cross-level features from noisy, low contrast and SCR infrared images.
- We establish a new large benchmark called IRSTD-1k to facilitate the research in the area of IRSTD, which consists of 1,000 manually labeled realistic images with various target shapes, different target sizes, and rich clutter backgrounds from diverse scenes.

2. Related work

2.1. Infrared Small Target Detection

Traditional IRSTD methods rely on image processing techniques or handcrafted features. Representative methods include HVS based methods, such as tri-layer local contrast measure (TLLCM) [4] and weighted strengthened local contrast measure (WSLCM) [12], filtering based methods [2, 9], as well as low-rank based methods, such as reweighted infrared patch-tensor (RIPT) [5], partial sum of the tensor nuclear norm [39], Infrared patch-image (IPI) [10], and non-convex rank approximation minimization [38]. However, these methods based on image processing, filtering or handcrafted features are ineffective in dealing with challenging cases including targets with varying shapes and sizes and backgrounds with clutter and noise.

Deep neural networks, on the contrary, can learn features automatically from a large amount of data covering complex scenes, owing to the end-to-end learning paradigm [13, 21, 36, 40]. As a result, CNN-based methods usually deliver better performance for IRSTD than traditional methods. Liu *et al.* [19] proposed the pioneer IRSTD method based on an MLP network. On the basis of Faster-RCNN [27] and Yolo-v3 [26], McIntosh *et al.* [24] designed a target to clutter network. Then, Wang *et al.* [34] proposed MDvsFA for SIRSTD, which achieved a trade-off between false alarm and miss detection. Researchers also explored the denoising idea for IRSTD [30], which treats small targets as noise and subtracts the denoised output from the input image to obtain small targets. Although these methods can detect the small target in infrared images, they are incapable of getting clear shapes of the detected targets since they pay little attention to modeling target shapes. By contrast, we explore a new idea of incorporating the reconstruction of target shape into the detection of small infrared targets and devise a novel infrared shape network. It can detect small infrared targets with clear contours which is beneficial for many subsequent tasks, e.g., recognizing the target type.

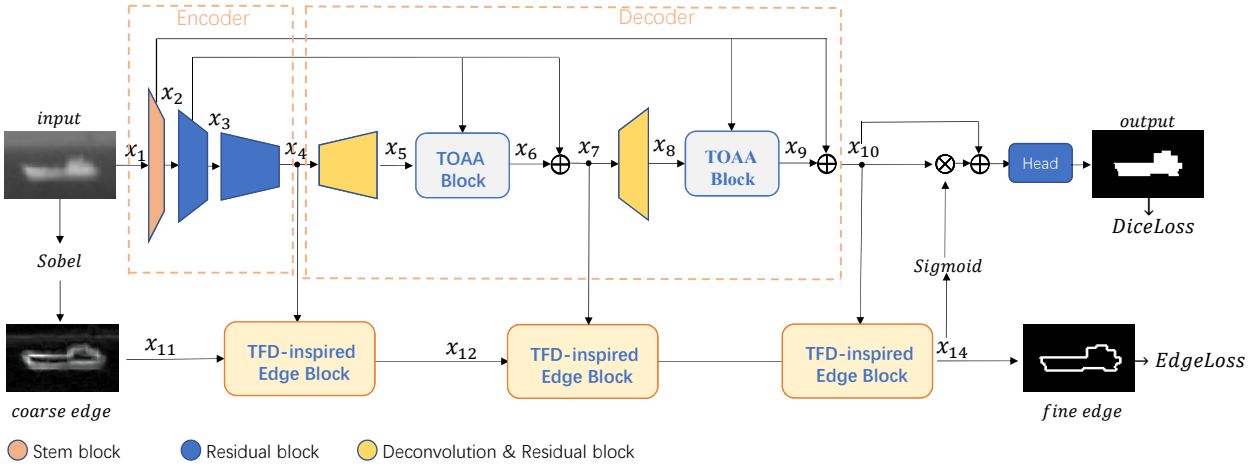

Figure 1. Overview of the proposed ISNet, which has a U-Net structure with TOAA blocks and TFD-inspired edge blocks.

2.2. Cross-Layer Feature Fusion

Typical cross-layer feature fusion methods include U-Net [28], PAN-net [20], and attention-based methods [17, 18, 23, 41]. U-Net was first designed to solve the medical image segmentation problem, which has been widely used in many other tasks. Redmon *et al.* [25] leveraged cross-layer feature fusion in object detection and improved the accuracy. Subsequently, contextual information was leveraged in IRSTD methods. Dai *et al.* designed an ACM-Net [6] and ALCNet [7] to extract contextual features from different layers. However, infrared small targets are usually dim and have varying shapes, making it difficult to extract and fuse useful shape features from multiple feature levels, where deep layers may have clear semantics but lack fine details of infrared targets. In contrast to these above methods, we design a two-orientation attention aggregation block, which can be incorporated into the U-Net structure to efficiently aggregate features from different levels.

2.3. ODE Inspired Network

Researchers have found an interesting link between ODE and neural networks. Weinan [35] firstly discovered the similarity and built a link between discrete ODE and ResNet [14]. Then, Chang *et al.* [3] analyzed the similarity of different neural networks and ODE. On the basis of these similarities, researchers devised specific networks based on ODE and achieved better performance in different fields. For example, He *et al.* [15] proposed a single image super-resolution method based on ODE, achieving SOTA performance. It is noteworthy that most existing ODE-based networks are designed based on the Euler method [29], although the numerical solutions of ODE based on the Taylor method can always deliver better accuracy [11]. Based on this observation, we apply the Taylor formula to get the numerical solution of ODE and design a novel edge block accordingly to extract useful edge features of infrared targets.

2.4. Datasets for IRSTD

Traditional methods train their networks on self-built datasets with targets in limited diversity. Only a few of them are publicly available, such as NUAA-SIRST [6] and MFIRST [34]. Although the two datasets have facilitated the research on IRSTD, they have some limitations. First, most images in MFIRST are synthetic and NUAA-SIRST only has a limited number of images. Second, both datasets pay less attention to the annotation of target shape, which can provide informative supervisory signals and is important for many downstream tasks. In this paper, we establish a new dataset named IRSTD-1k by collecting 1,000 realistic images with different targets in great diversity and annotating them with accurate pixel-level masks.

3. Method

In this section, we first introduce the overall architecture of our ISNet. Then, we present the details of the TFD-inspired edge block (Sec. 3.2) and the TOAA block with U-Net structure (Sec. 3.3), followed by the loss function in Sec. 3.4 as well as the IRSTD-1k dataset in Sec. 3.5.

3.1. Overall Architecture

As shown in Fig. 1, a single infrared image is fed into the encoder part of the U-Net structure. Then, in the decoder part of the U-Net structure, the proposed TOAA blocks are inserted to aggregate the cross-level features. By connecting the TOAA blocks with the TFD-inspired edge blocks step-by-step, we can obtain the coarse target shape and fine edges. Finally, we further refine the coarse shape with the help of fine edges via a convolutional segmentation head.

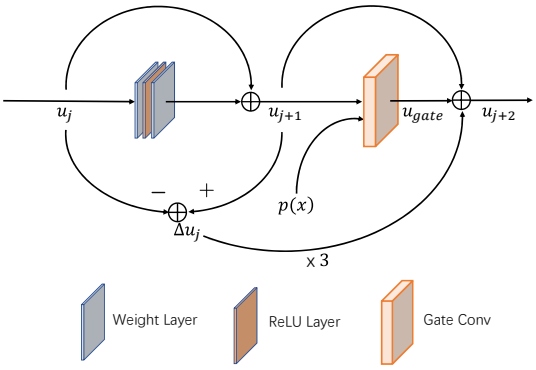

Figure 2. Structure of the TFD-inspired edge block (Sec. 3.2).

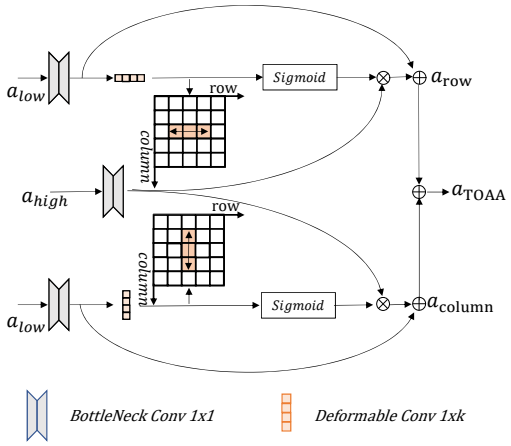

Figure 3. Structure of the TOAA block (Sec. 3.3).

3.2. TFD-inspired Edge Block

The infrared targets are usually small and dim while the infrared images always contain lots of noise and clutter. The low low contrast and SCR make it difficult to extract complete edge information of the targets. To address this issue, we revisit the similarity between the residual network structure used in existing methods and the Euler method [35], and devise a new TFD-inspired edge block based on the second-order Taylor finite difference equation, which enables to aggregate edge information from different levels and help obtain fine target edges.

Specifically, we leverage finite difference equations to discretize the ODE, where the partial derivatives can be replaced with a set of approximate differences. Since the Taylor finite difference method can deliver better accuracy than the Euler method [1], we adopt it to devise a novel TFD-inspired edge block. Mathematically, the second-order TFD equation can be formulated as:

$$\frac{\partial u}{\partial x} = \frac{-\frac{1}{2}u_{j+2} + 2u_{j+1} - \frac{3}{2}u_j}{\Delta x}. \tag{1}$$

Then, we rewrite it in the additive form:

$$u_{j+2} = -2\frac{\partial u}{\partial x}\Delta x + u_{j+1} + 3u_{j+1} - 3u_j. \tag{2}$$

To ease the training of deep neural networks, we adopt the residual learning idea and transform the direct mapping $H(x) = F(x) + x$ into the residual form $F(x) = H(x) - x$, where $H(x)$ and $F(x)$ denote the target output and the learned residual, respectively. We re-write Eq. (2) as:

$$-2\frac{\partial u}{\partial x}\Delta x = u_{j+2} - u_{j+1} - 3(u_{j+1} - u_j). \tag{3}$$

In this paper, we leverage several convolutional layers to implement the transformation from u_j to u_{j+1}, and adopt a gated convolution u_{gate} to get $-2\frac{\partial u}{\partial x}\Delta x$, as shown in Fig. 2. Thus, Eq. (2) can be expressed as:

$$u_{j+2} = u_{gate} + u_{j+1} - 3\Delta u_j, \tag{4}$$

where Δu_j denote the residual between u_{j+1} and u_j. In this way, the TFD-inspired edge block can extract the edge features in a residual learning manner. It is noteworthy that gated convolution can be considered as a partially learnable convolution, where a soft gating mechanism is used to better learn the edge information of the target while suppressing the background information. Specifically, the input of u_{gate} is a sum of u_{j+1} and the corresponding features (denoted as $p(x)$) from the the U-Net, *e.g.*, x_4, x_7, and x_{10} in Fig. 1.

3.3. TOAA Block

Since low-level features usually contain fine details of targets, which are absent in high-level features, we devise a TOAA block to refine the high-level features to facilitate the reconstruction of target shape and edge. As shown in Fig. 3, the TOAA block consists of two parallel attention modules, where each of them generates an attention map along one direction, *i.e.*, the row or column direction, and uses it to modulate the high-level features, respectively. Finally, the attentive features are summed together as the output of the block. This process can be expressed as:

$$\begin{aligned} a_{TOAA} &= TOAA(a_{low}, a_{high}) \\ &= a_{row} + a_{column}, \end{aligned} \tag{5}$$

where $TOAA(\cdot)$ denotes the mapping function learned by the TOAA block. a_{low} and a_{high} represent the low-level and high-level features from the U-Net encoder and decoder, respectively. a_{row} and a_{column} are the attentive features in row and column directions, and can be obtained as:

$$a_{row} = S(F_r(F_b(a_{low})))F_b(a_{high}) + F_b(a_{low}), \tag{6}$$

$$a_{column} = S(F_c(F_b(a_{low})))F_b(a_{high}) + F_b(a_{low}). \tag{7}$$

Here $S(\cdot)$ denotes the sigmoid function. $F_b(\cdot)$ stands for a bottleneck structure including two 1×1 convolutional layers to constrain high-frequency noise. The bottleneck structure is similar to the role of Nonnegative Matrix Factorization (NMF), which can retain useful features while filtering out redundant high-frequency noise [13, 22]. F_r denotes a $1\times$k deformable convolution in the row direction while F_c represents a k$\times1$ deformable convolution in the column direction. This two-orientation attention mechanism in the TOAA block promotes extracting shape information from low-level features in two directions and accordingly guide the refinement of high-level features. TOAA blocks are inserted into the U-Net decoder to perform cross-level feature fusion, as shown in Fig. 1.

We next briefly describe the flow of features in our ISNet. First, the input infrared image x_1 is first processed by the stem block in the encoder, which consists of a convolutional layer and a max-pooling layer with a stride of 2 each to downsample the image. The output x_2 is defined as:

$$x_2 = F_{max}(conv(x_1)), \tag{8}$$

where $conv(\cdot)$ and $F_{max}(\cdot)$ denote convolutional and max-pooling layers, respectively. Then, we perform a nonlinear transformation through two residual blocks to obtain features x_3 and x_4 with less noise and clutter.

For the decoder, we perform deconvolution with a stride of 2 on x_4 to double the image size and obtain the high-level feature x_5. Then, we fuse x_5 and the low-level feature x_3 with the same size via the TOAA block to obtain the refined feature x_6, *i.e.*,

$$x_6 = TOAA(x_3, x_5). \tag{9}$$

Similarly, we apply the TOAA block on the low-level feature x_2 and high-level feature x_8 to get x_9, *i.e.*,

$$x_9 = TOAA(x_2, x_8). \tag{10}$$

By stacking TOAA blocks sequentially in the U-Net decoder, our ISNet can efficiently extract cross-level features of infrared targets which embed both semantics and fine details, thereby facilitating the reconstruction of the target shape.

On the bottom path, the coarse edge x_{11} obtained by applying the Sobel operator on the input image together with the feature x_4 from the U-Net encoder are fed into the TFD-inspired edge block to extract the edge feature. Similarly, two extra such blocks are used to further refine the edge feature with the high-level features obtained from the TOAA blocks in the U-Net decoder. Finally, the edge feature is fed into a convolutional layer to get the fine edge prediction. It is also used to generate attention to refine the output feature of the U-Net decoder, which is further used by the segmentation head to predict the final target mask.

3.4. Loss Function

Dice Loss: Dice loss [31] is a common measure used to evaluate the difference between a mask prediction and the ground truth, which is defined as:

$$L_{Dice} = 1 - 2|Y' \cap Y|/(|Y'| + |Y|), \tag{11}$$

where $|Y' \cap Y|$ is the intersection of the prediction Y' and the ground truth Y. $|\cdot|$ is the number of pixels in the mask.
Edge loss: Binary cross-entropy (BCE) Loss is also used to measure the difference between the predicted mask and the ground truth. We leverage both Dice loss L_{Dice} and BCE loss L_{BCE} to supervise the edge prediction:

$$L_{Edge} = L_{Dice}^{Edge} + \lambda L_{BCE}^{Edge}, \tag{12}$$

where λ is a hyper-parameter to balance the two losses and set to 10 empirically. The final training objective is a combination of L_{Edge} and the dice loss on the mask prediction:

$$L = L_{Edge} + L_{Dice}^{Mask}. \tag{13}$$

3.5. IRSTD-1k Dataset

We construct a new benchmark called IRSTD-1k, consisting of 1,000 infrared images captured by an infrared camera in the real world. We annotate the targets at pixel level manually. The images are in the size of 512×512. IRSTD-1k contains different kinds of small targets, such as drones, creatures, vessels and vehicles, captured at different positions from a long imaging distance. The dataset covers lots of different scenes and the background contains the sea, river, field, mountain area, city, and cloud with heavy clutters and noises. IRSTD-1k can be used to comprehensively evaluate IRSTD methods.

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets: We conduct experiments on the *IRSTD-1k* dataset and *NUAA-SIRST* dataset [6]. NUAA-SIRST includes 427 infrared images while IRSTD-1k contains 1,000 infrared images. For each dataset, we split it into the training set, validation set, and test set at a ratio of 50:30:20.

Evaluation Metrics: We compare the proposed ISNet with SOTA methods using several common metrics. *Intersection over Union (IoU):* IoU is defined as:

$$IoU = A_i/A_u, \tag{14}$$

where A_i and A_u denote the size of intersection region and union region, respectively. *Normalized Intersection over Union (nIoU):* nIoU is the normalization of IoU, *i.e.*,

$$nIoU = \frac{1}{N}\sum_{i=1}^{N}(TP[i]/(T[i] + P[i] - TP[i])), \tag{15}$$

Table 1. Comparisons with SOTA methods on NUAA-SIRST and IRSTD-1k in $IoU(\%)$, $nIoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$.

Method	NUAA-SIRST (Tr=50%)				IRSTD-1k (Tr=50%)			
	Pixel-Level		Object-Level		Pixel-Level		Object-Level	
	IoU	nIoU	Pd	Fa	IoU	nIoU	Pd	Fa
Top-Hat [2]	7.143	5.201	79.84	1012	10.06	7.438	75.11	1432
Max-Median [9]	4.172	2.15	69.20	55.33	6.998	3.051	65.21	59.73
WSLCM [12]	1.158	0.849	77.95	5446	3.452	0.678	72.44	6619
TLLCM [4]	1.029	0.905	79.09	5899	3.311	0.784	77.39	6738
IPI [10]	25.67	24.57	85.55	11.47	27.92	20.46	81.37	16.18
NRAM [38]	12.16	10.22	74.52	13.85	15.25	9.899	70.68	16.93
RIPT [5]	11.05	10.15	79.08	22.61	14.11	8.093	77.55	28.31
PSTNN [39]	22.40	22.35	77.95	29.11	24.57	17.93	71.99	35.26
MSLSTIPT [32]	10.30	9.58	82.13	1131	11.43	5.932	79.03	1524
MDvsFA [34]	60.30	58.26	89.35	56.35	49.50	47.41	82.11	80.33
ACM [6]	72.33	71.43	96.33	9.325	60.97	58.02	90.58	21.78
ALCNet [7]	74.31	73.12	97.34	20.21	62.05	59.58	92.19	31.56
ISNet	**80.02**	**78.12**	**99.18**	**4.924**	**68.77**	**64.84**	**95.56**	**15.39**

where N is the total number of samples, $TP[\cdot]$ denotes the number of true positive pixels, $T[\cdot]$ and $P[\cdot]$ denotes the number of ground truth and predicted positive pixels, respectively. *Probability of Detection (P_d):* P_d is the ratio of correctly predicted targets N_{pred} and all targets N_{all}:

$$P_d = N_{pred}/N_{all}. \tag{16}$$

False-Alarm Rate (F_a): F_a is the ratio of false predicted target pixels N_{false} and all the pixels in the image N_{all}:

$$F_a = N_{false}/N_{all}. \tag{17}$$

4.2. Implementation Details

We adopt AdaGrad as the optimizer with a learning rate of 0.04. The training process lasts a total of 500 epochs with a weight decay of 10^{-4} and a batch size of 8. We select ALCNet [7], ACMNet [6], and MDvsFA [34] as the representative CNN-based IRSTD methods. For traditional methods, we choose Top-Hat [2], Max-Median [9], WSLCM [12], TLLCM [4], IPI [10], NRAM [38], RIPT [5], PSTNN [39], and MSLSTIPT [32].

4.3. Quantitative Results

As shown in Table 1, the proposed ISNet achieves the best performance in terms of all the evaluation metrics compared with SOTA methods on both NUAA-SIRST and IRSTD-1k datasets. For example, the P_d of our method on NUAA-SIRST reaches as high as 99.18%. Traditional methods based on hand-crafted features perform poorly in challenging cases, thereby having much worse scores than CNN-based methods. Nevertheless, CNN-based methods

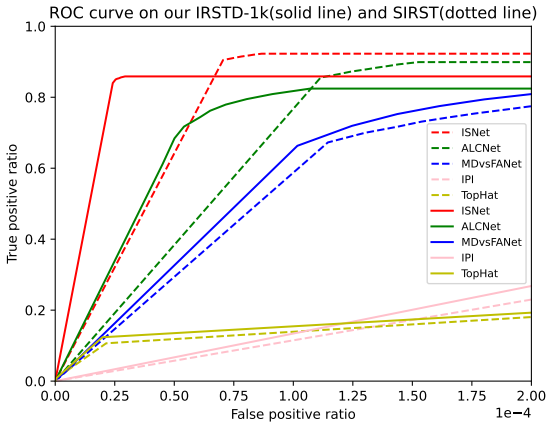

Figure 4. ROC curves of different methods on the NUAA-SIRST dataset (dotted line) and IRSTD-1k dataset (solid line).

Table 2. Ablation study of the TOAA block and TFD-inspired edge block in $IoU(\%)$, $nIoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$.

Method	IoU	nIoU	Pd	Fa
U-Net	68.31	67.85	92.95	60.16
U-Net+TOAA	75.65	74.81	98.93	3.573
U-Net+TFD	78.05	76.49	99.13	6.465
U-Net+TOAA+TFD	**80.02**	**78.12**	**99.18**	**4.924**

pay less attention to target edge and shape information, sufferring from inaccurate mask predictions, *e.g.*, lower IoU and nIoU. The performance of our ISNet is better on NUAA-SIRST over IRSTD-1k. It is because the IRSTD-1k dataset contains more challenging cases for IRSTD, including varying-shape targets and low contrast and low SCR background with clutters and noises. Nevertheless, our ISNet can still deliver promising results owing to the designed TOAA block to effectively aggregate cross-level features and the TFD-inspired edge block to extract edge features.

We also plot the ROC curves of different methods on the NUAA-SIRST dataset in Fig. 4. As can be seen, the performance of our ISNet is significantly better than other methods, where the area under the ROC curve (AUC) of our ISNet is larger than those of both the traditional methods and CNN-based methods, *e.g.*, 0.9612 AUC of ISNet v.s. 0.9495 AUC of ALCNet [7] on the NUAA-SIRST dataset.

4.4. Visual Results

Some visual results obtained by different methods on the NUAA-SIRST dataset are shown in Fig. 5. As can be seen, even in low contrast and low SCR situations, our ISNet can not only locate the target accurately but also obtain a complete and precise target shape. This is because the proposed TOAA block can model contextual information of the tar-

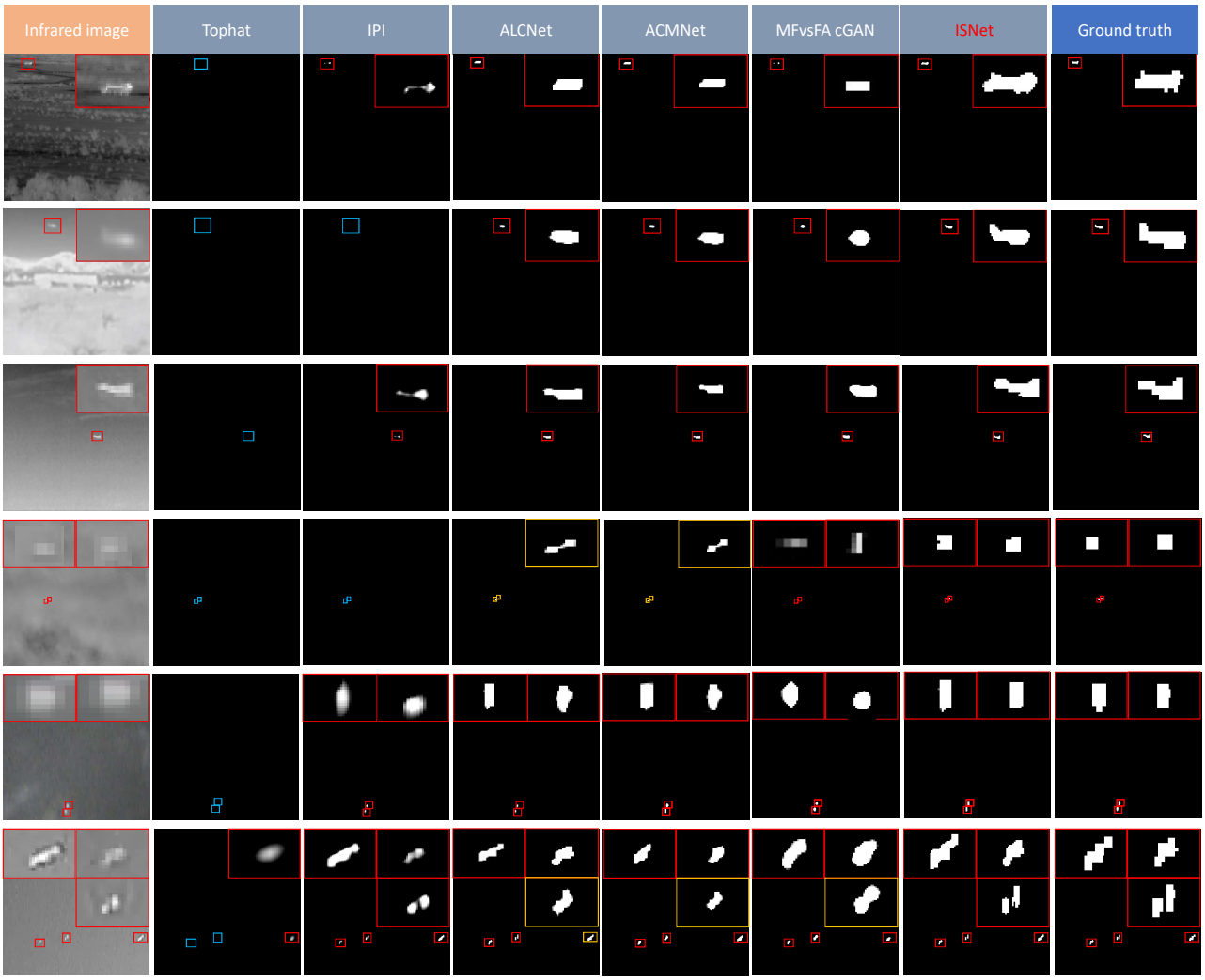

Figure 5. Visual results obtained by different IRSTD methods on the NUAA-SIRST dataset. Closed-up views are shown in the right top corner. Boxes in red, blue and yellow represent correctly detected targets, miss detected targets and false detected targets, respectively.

Table 3. Ablation study of the TOAA block and SOTA cross-layer feature fusion methods in $IoU(\%)$, $nIoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$.

Method	FPN Based			U-Net Based				Our U-Net
	SK [18]	GAU [17]	ACM [6]	SK	GAU	TBCNet [41]	ALCNet [7]	TOAA
IoU	70.21	70.15	73.18	70.81	71.82	72.19	74.31	**75.65**
nIoU	69.53	70.16	72.13	69.93	69.74	70.57	73.12	**74.81**
Pd	93.78	94.02	96.91	93.69	94.53	98.29	97.34	**98.93**
Fa	40.26	35.68	9.325	31.23	37.68	10.21	20.21	**3.573**

get via effective cross-level feature fusion, while the TFD-inspired edge block can extract useful edge features to get fine target edges and help reconstruct the target shape. Traditional IRSTD methods are prone to produce missed detections and false detections when SCR is low, and produce false-alarm detections when local-contrast is high. CNN-based methods generally perform better than traditional methods, but cannot predict accurate target shapes.

4.5. Ablation Study

To investigate the effectiveness of each component in our ISNet, we perform several ablation studies on the NUAA-SIRST dataset. The ablation study results of TOAA and TFD edge blocks are shown in Table 2. As can be seen, each of them improves the performance of the U-Net baseline and using both of them delivers the best results, implying their complementarity.

Impact of TOAA Block: As shown in Table 3, compared to other cross-layer feature fusion methods based on either FPN or U-Net, our TOAA outperforms them by a large margin, showing its superiority in fusing features from

Table 4. Ablation study on the different number of TOAA blocks in $IoU(\%)$, $nIoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$)

TOAA Blocks	IoU	nIoU	Pd	Fa
0	68.31	67.85	92.95	60.16
1	73.04	71.33	97.69	9.447
2	75.65	74.81	98.93	3.573
3	75.61	74.85	98.99	3.798

Table 5. Ablation study of the TFD-inspired edge block in $IoU(\%)$, $nIoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$.

Equation Type	IoU	nIoU	Pd	Fa
Gated Conv(only)	75.38	74.55	98.22	18.486
Gated Conv+Bottle Neck	76.32	75.29	98.73	9.823
Gated Conv+ResBlock	77.23	76.01	99.02	14.377
TFD	**78.05**	**76.49**	**99.13**	**6.465**

Table 6. Ablation study on the different number of TFD-inspired edge blocks in $IoU(\%)$, $nIoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$.

Edge Blocks	IoU	nIoU	Pd	Fa
0	68.31	67.85	92.95	60.16
1	74.35	73.21	97.89	30.21
2	76.56	74.98	98.59	13.61
3	78.05	76.49	99.13	6.465
4	78.15	76.15	99.27	9.062

both low and high levels. The specifically designed two-orientation attention mechanism promotes to extract shape information from low-level features and guides the refinement of high-level features. We also investigate the influence of using different numbers of TOAA blocks. As shown in Table 4, without using the TOAA blocks, the U-Net baseline produces lots of false predictions. With the help of the TOAA block, its performance can be improved significantly, especially when two TOAA blocks are used, which delivers the best results and is the default setting.

Impact of TFD-inspired Edge Block: We also ablate the design of the proposed TFD-inspired edge block. As shown in Table 7, if we only use gated convolutions to reconstruct edges, the targets are easily submerged in noises. Introducing residual blocks or bottleneck structures can improve the performance. Combining them together, the proposed TFD-inspired edge block achieves the best performance. We also carry out the ablation study on different numbers of TFD-inspired edge blocks. From the first two rows in Table 6, we can find that our TFD-inspired edge block improves the shape segmentation performance of the baseline U-Net significantly. Using more blocks generally delivers better results but increases the model complexity.

Table 7. Ablation study on the different feature extraction methods during pre-processing in $IoU(\%)$, $nIoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$.

Method	IoU	nIoU	Pd	Fa
Sobel+TFD	80.02	78.12	99.18	4.924
ResBlock+TFD	79.97	78.20	99.13	5.24
Gated Conv+TFD	79.85	77.95	99.01	4.26

We choose three blocks as the default setting.

Impact of sobel operator: In the data pre-processing stage, we use the Sobel operator to extract the coarse edge of the target from the input image. The Sobel operator can be replaced by other edge features extraction methods, such as gated convolution and residual blocks. As shown in Table 7, using either the Sobel operator or other learnable alternatives deliver similar results. For simplicity, we choose the Sobel operator as the default setting.

5. Conclusion

We propose a novel ISNet to handle the challenging IRSTD task under low contrast and low SCR situations. Specifically, we introduce two novel components, *i.e.*, the two-orientation attention aggregation block and the TFD-inspired edge block, where the former promotes cross-level feature fusion to enhance the shape representation capacity of high-level features and the latter extracts useful edge features to help predict accurate target mask with precise shape. Moreover, we establish a new large IRSTD dataset named IRSTD-1k, which could serve as a testbed for the evaluation of IRSTD methods and facilitate future research. Extensive experiments on both public dataset and our IRSTD-1k dataset validate the effectiveness of the proposed idea that incorporates the reconstruction of target shape into the detection of small infrared targets, and the superiority of the ISNet over representative methods.

Broader Impacts Detecting objects from infrared images benefits many real-world applications, such as traffic management, marine rescue, and wildlife conservation. Although it still has the potential to be used for military purposes, strict registration is expected to limit the misuse of the IRSTD methods, as well as other AI technologies.

Acknowledgement This work was supported in part by the National Natural Science Foundation of China under Grants 61902293, 62036007, the Equipment Advance Research Field Fund Project under Grant 80913010601, the Shaanxi Province Key Research and Development Program Project under Grant 2021GY-034, the Youth Talent Promotion Project of China Association for Science and Technology, the Youth Talent Promotion Project of Shaanxi University Science and Technology Association under Grant 20200103, the Fundamental Research Funds for the Central Universities under Grant XJS200112.

References

[1] John Anderson. Computational fluid dynamics: the basics with applications. *Multidisciplinary Digital Publishing Institute*, 1995. 4

[2] Xiangzhi Bai and Fugen Zhou. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6):2145–2156, 2010. 1, 2, 6

[3] Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, and Elliot Holtham. Reversible architectures for arbitrarily deep residual neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 1, 2018. 3

[4] CL Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. A local contrast method for small infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):574–581, 2013. 1, 2, 6

[5] Yimian Dai and Yiquan 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*, 10(8):3752–3767, 2017. 1, 2, 6

[6] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 950–959, 2021. 2, 3, 5, 6, 7

[7] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 3, 6, 7

[8] He Deng, Xianping Sun, Maili Liu, Chaohui Ye, and Xin Zhou. Small infrared target detection based on weighted local difference measure. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):4204–4214, 2016. 1

[9] Suyog D Deshpande, Meng Hwa Er, Ronda Venkateswarlu, and Philip Chan. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets 1999*, volume 3809, pages 74–83. International Society for Optics and Photonics, 1999. 1, 2, 6

[10] Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G Hauptmann. Infrared patch-image model for small target detection in a single image. *IEEE Transactions on Image Processing*, 22(12):4996–5009, 2013. 1, 2, 6

[11] David Francis Griffiths and Desmond J Higham. *Numerical methods for ordinary differential equations: initial value problems*. Springer, 2010. 3

[12] Jinhui Han, Saed Moradi, Iman Faramarzi, Honghui Zhang, Qian Zhao, Xiaojian Zhang, and Nan Li. Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Geoscience and Remote Sensing Letters*, 18(9):1670–1674, 2020. 1, 2, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[15] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019. 3

[16] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Pecognition*, pages 1–8. IEEE, 2007. 1

[17] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 3, 7

[18] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019. 3, 7

[19] Ming Liu, Hao-yuan Du, Yue-jin Zhao, Li-quan Dong, and Mei Hui. Image small target detection based on deep learning with snr controlled sample generation. In *Current Trends in Computer Science and Mechanical Automation Vol. 1*, pages 211–220. De Gruyter Open Poland, 2018. 2

[20] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 3

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[22] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. Face model compression by distilling knowledge from neurons. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 3560–3566, 2016. 5

[23] Benteng Ma, Jing Zhang, Yong Xia, and Dacheng Tao. Auto learning attention. *Advances in neural information processing systems*, 33:1488–1500, 2020. 3

[24] Bruce McIntosh, Shashanka Venkataramanan, and Abhijit Mahalanobis. Infrared target detection in cluttered environments by maximization of a target to clutter ratio (tcr) metric using a convolutional neural network. *IEEE Transactions on Aerospace and Electronic Systems*, 57(1):485–496, 2020. 2

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 3

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 2

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation.

In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 3

[29] Jiawei Shen, Zhuoyan Li, Lei Yu, Gui-Song Xia, and Wen Yang. Implicit euler ode networks for single-image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 218–219. 3

[30] Manshu Shi and Huan Wang. Infrared dim and small target detection based on denoising autoencoder network. *Mobile Networks and Applications*, 25(4):1469–1483, 2020. 2

[31] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. 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*, pages 240–248. 2017. 5

[32] Yang Sun, Jungang Yang, and Wei An. Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):3737–3752, 2020. 6

[33] Michael Teutsch and Wolfgang Krüger. Classification of small boats in infrared images for maritime surveillance. In *International WaterSide Security Conference*, pages 1–7. IEEE, 2010. 1

[34] Huan Wang, Luping Zhou, and Lei 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*, pages 8509–8518, 2019. 2, 3, 6

[35] E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017. 3, 4

[36] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[37] Jing Zhang and Dacheng Tao. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 8(10):7789–7817, 2020. 1

[38] Landan Zhang, Lingbing Peng, Tianfang Zhang, Siying Cao, and Zhenming Peng. Infrared small target detection via non-convex rank approximation minimization joint l2, 1 norm. *Remote Sensing*, 10(11):1821, 2018. 1, 2, 6

[39] Landan Zhang and Zhenming Peng. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sensing*, 11(4):382, 2019. 1, 2, 6

[40] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022. 2

[41] Mingxin Zhao, Li Cheng, Xu Yang, Peng Feng, Liyuan Liu, and Nanjian Wu. Tbc-net: A real-time detector for infrared small target detection using semantic constraint. *arXiv preprint arXiv:2001.05852*, 2019. 3, 7