



E2SAM: Edge-Enhanced SAM with FFC Adapter for Few-Shot Infrared Small Target Detection

Ziyi Li¹, Yinghui Xing^{1(✉)}, Feng Sang^{2,3}, Shizhou Zhang¹, Lingyan Ran¹,
and Yanning Zhang¹

¹ School of Computer Science, Northwestern Polytechnical University, Xi'an, China
xyh7491@nwpu.edu.cn

² Northwestern Polytechnical University, Xi'an, China

³ Shanghai Institute of Satellite Engineering, Shanghai, China

Abstract. Infrared small target detection (IRSTD) is a key component of infrared target tracking and search. This task is quite challenging due to the small and dim infrared targets, complex backgrounds, and low target-background contrast. Current infrared remote sensing detection algorithms exhibit performance that is strongly dependent on data scale and quality, yet this performance is constrained by the scarcity of publicly available datasets. Moreover, most existing methods are trained only on a single style domain, resulting in poor generalization to cross-domain scenarios. To address these issues, we propose E2SAM, an Edge-Enhanced Segment Anything Model tailored for IRSTD. E2SAM adopts a dual-branch design: one branch leverages the Segment Anything Model (SAM) for mask prediction, while the other utilizes lightweight convolution for edge detection. We further introduce a fast Fourier convolution adapter to extract high-frequency features. By embedding physical priors such as high-frequency and edge information, E2SAM guides the base model to extract target features more accurately, mitigating style domain bias. To evaluate its few-shot learning capability, we conduct experiments with limited training data (using 30%, 50%, and 80% of the original datasets), and the results show that E2SAM maintains excellent detection performance under such conditions. It also generalizes well in cross-domain scenarios, validating the effectiveness of the proposed method.

Keywords: Infrared Small Target Detection · Segment Anything Model · Edge Enhancement · Frequency Information

1 Introduction

Infrared small target detection has become an essential task in various fields such as maritime rescue [1], urban security [2], tracking systems [3] and target warning [4]. Unlike visible light imaging, infrared imaging is less affected by lighting conditions and camouflage, making it a more robust modality for detecting objects in complex backgrounds. However, IRSTD remains highly challenging due to low signal-to-noise ratio (SNR), lack of semantic information, and cluttered backgrounds.

Traditional methods for infrared small target detection often rely on handcrafted features or background suppression techniques. They can be divided into three categories: filtering-based methods [5, 6], human visual system-based methods [7–12], and low rank based methods [13–18]. Although these methods are computationally efficient, they typically suffer from low robustness in diverse environments.

Recently, deep learning-based approaches, especially convolutional neural networks (CNNs) and Transformer architectures, have shown impressive performance in infrared small target detection. Feature fusion strategies have been widely adopted in theIRSTD task [19–22]. Dai et al. [23] proposed an asymmetric contextual fusion strategy that introduces a bottom-up modulation pathway based on pointwise channel attention, enabling effective exchange between high-level semantics and fine-grained low-level details. Inspired by nested architectures [24], DNA-Net [25] introduced a densely nested interaction module to enable progressive fusion between multi-level features. SCTransNet [26] integrates spatial-channel cross transformer blocks (SCTBs) into a U-shaped architecture, enabling both local detail extraction and global context modeling. This hybrid design enhances semantic discrimination between small infrared targets and complex backgrounds.

These methods often rely on complex network designs and large-scale high-quality training data, yet the scarcity and costly annotation of infrared remote sensing data limit further performance gains and hinder the development of IR target detection algorithms. Leveraging large-scale pre-trained foundation models offers a promising way to alleviate the annotation burden. Due to the significant style differences between natural and infrared images, SAM [27] cannot be directly applied to infrared small target detection tasks. To alleviate the annotation burden in infrared small target detection tasks and improve the generalization ability and detection performance of infrared small target detection models, we propose the E2SAM. The network follows an encoder-decoder architecture based on the pre-trained SAM model, with two branches: one for target mask prediction and the other for extracting target edge information. To optimize the pre-trained encoder for infrared target detection, we design a Fast Fourier Convolution Adapter (FFC Adapter) to extract high-frequency components from infrared images, addressing the challenges of infrared image characteristics. Additionally, the edge detection branch extracts image edges, which are combined with high-frequency components to guide the foundational model in effectively extracting infrared features.

Our contributions can be summarized as follows:

- We propose E2SAM, a novel infrared adaptation of SAM that introduces a dual-branch architecture.
- We propose the edge-enhanced branch extracts explicit edge features to enrich target representations and improve detection performance.
- We design a Fast Fourier Convolution (FFC) adapter to extract high-frequency information from infrared images, effectively mitigating the domain gap between natural and infrared imagery and enhancing the base model’s feature extraction capability.
- Extensive experiments on benchmark infrared datasets demonstrate that our method outperforms existing state-of-the-art approaches in terms of both precision and robustness, even when trained on a small sample size.

2 Related Work

2.1 Single-Frame Infrared Small Target Detection

As an important branch of target detection, infrared remote sensing target detection has produced a series of important research results after years of development. Early single-frame infrared small target detection methods are mainly model-driven methods, which can be divided into three major categories: filtering-based methods [5, 6], human visual system-based methods [7–12], and low rank based methods [13–18]. These methods primarily rely on manually crafted features.

In recent years, with the rapid growth of infrared data and significant improvements in computing power, deep learning methods have been widely applied in infrared remote sensing target detection. Wang et al. [28] propose a conditional Generative Adversarial Network MDvsFA, comprising of two generators and one discriminator, with each focusing on reducing either missed detection rate and false alarm rate. Dai et al. [23] proposed an infrared small target detection network ACM-Net based on the asymmetric contextual modulation (ACM) feature fusion. Liu et al. [25] proposed the Densely Nested Attention Network (DNANet), which integrates and fully exploits the contextual information of small targets. UIU-Net [29] uses residual U-blocks to preserve feature resolution despite network depth increases. ISTDU-Net [30] introduces feature map groups during down-sampling to prioritize small target features, and a fully connected layer in the jump connection suppresses background structures. Liu et al. [31] proposed a segmentation approach that combines CNNs with transformers. By leveraging the self-attention mechanism of transformers, the method captures interactive features over a broader spatial range, enabling improved spatial context awareness. SCTransNet [26] introduces a spatial channel cross-Transformer block (SCTB) to improve feature discrimination by capturing both local and global dependencies. While infrared small targets typically lack rich texture information, they contain abundant physical cues such as edge shapes and frequency characteristics. Zhang et al. [32] proposed a novel Infrared Shape Network (ISNet). It focuses on accurately detecting the shapes of small targets by incorporating TFD-inspired edge blocks to enhance edge information, and employing Two-Direction Attention Aggregation (TOAA) to suppress noise and capture detailed shape features.

While deep learning methods have advanced, their performance heavily relies on large-scale and high-quality annotated datasets. However such datasets remains scarce and costly to obtain. To mitigate this dependence, we leverage foundational segmentation models pre-trained on natural images forIRSTD.

2.2 Segment Anything Model

SAM [27] is a general object segmentation model proposed by the Meta AI team, aiming to achieve the goal of “segmenting anything.” It demonstrates strong generalization and versatility through training in large-scale, diverse datasets, becoming a representative foundational model in computer vision. The core architecture of SAM includes an image encoder, a prompt encoder, and a mask decoder. The image encoder uses Vision Transformer (ViT) [33] to extract global features, the prompt encoder processes various

forms of user prompts (such as points, boxes, or masks), and the mask decoder generates precise segmentation results based on these prompts. Its efficient architectural design ensures real-time responsiveness, enabling SAM to dynamically adjust input prompts and generate segmentation masks in real time.

Subsequently, several studies have fine-tuned the structure of SAM to enhance its performance on more challenging tasks [4, 17, 23, 35]. For instance, MedSAM [34] creates a large and diverse data set to fine-tune SAM for medical image segmentation. While acquiring extensive professional data is often challenging, parameter-efficient fine-tuning has gained increasing attention. Parameter-efficient fine-tuning [35, 36] typically involves freezing the parameters of foundational models while introducing learned adapters to customize them for specialized tasks. For instance, to adapt SAM to medical images, Med-SA [35] placed MLP-based bottleneck adapters along the self-attention and residual connection paths of the ViT block in SAM. Given that SAM is a prompt-based interactive segmentation model, AutoSAM [37] proposed learning an image-based prompt encoder to adapt the frozen SAM for medical image segmentation, rather than incorporating adapters into its core feature block.

2.3 Edge-Aware Enhancement for Small Target Detection

Infrared (IR) small targets typically occupy only a few pixels in the image, resulting in limited semantic and structural information. This scarcity of intrinsic features, coupled with their low contrast against complex backgrounds. Since convolutional neural networks (CNNs) primarily rely on texture cues for recognition, integrating edge-aware enhancement can effectively highlight structural boundaries, thus compensating for the weak appearance features of small targets. To this end, edge information is an important feature for object detection [38], and therefore, this information is needed to be preserved in the enhanced images to effectively amplify target representations and improve detection performance.

Recently, increasing attention has been paid to leveraging the edge information of small targets, including infrared shape network (ISNet) [32], gradient-guided learning network (GGL-Net) [39] and shape-biased representation network (SRNet) [40]. Yang et al. [41] proposes an Edge-enhanced Labeling Network, called SAM2-ELNet, which incorporates a labeling module and an edge attention mechanism. In addition, various studies have explored the use of edge information to enhance performance in medical image analysis tasks. For example, Zhou et al. [42] introduce a learnable Dense Residual Gradient Operator (DRGO) to enhance edge-aware feature representations. To preserve edge details, Zeng et al. [43] incorporate an Edge Enhancement Module (EEM) that effectively retains and integrates edge information to refine segmentation results.

3 Proposed Method

3.1 Network Overview

In this section, we provide a detailed discussion of the overall framework of the E2SAM, as shown in Fig. 1. The framework consists of two branches: one for target mask prediction and the other for extracting target edge information. Both branches adopt an

encoder-decoder structure. Given that SAM [27] is supervised on natural images, its style differs significantly from infrared remote sensing images. To optimize the performance of the pre-trained encoder, we design a Fast Fourier Convolution Adapter (FFC Adapter) to extract high-frequency components of infrared images. Additionally, the encoder of the edge detection branch is used to extract image edges. Finally, the mask decoder generates the target mask.

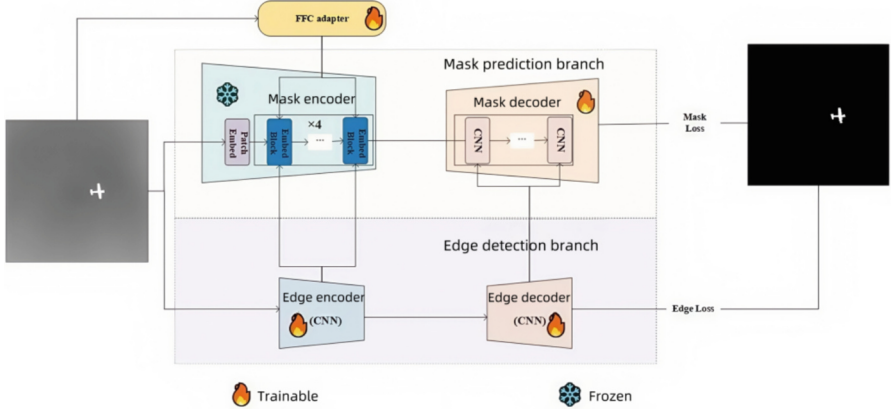


Fig. 1. Overall architecture of the proposed model.

3.2 Branch a: Edge Feature Extraction

The edge detection branch utilizes a lightweight encoder-decoder structure. The encoder hierarchically extracts edge features through convolutional layers and down-sampling operations, formulated as:

$$F_i^{ee} = CNN(Down(F_{i-1}^{ce})), i = 1, \dots, 4, \quad (1)$$

$$F_0^{ee} = CNN(I), \quad (2)$$

where F_i^{ee} denotes the edge features extracted by the encoder at the i -th level, $CNN(\cdot)$ represents convolutional feature extraction, and $Down(\cdot)$ performs spatial downsampling.

The decoder reconstructs the feature maps via upsampling and convolutional layers, progressively refining edge details:

$$F_{i+1}^{ed} = CNN(UP(F_i^{ed})), i = 1, \dots, 4 \quad (3)$$

$$F_1^{ed} = F_4^{ee} \quad (4)$$

Here, F_i^{ed} represents the edge features at the i -th decoder level, and $UP(\cdot)$ denotes upsampling.

3.3 Fast Fourier Convolution Adapter (FFC Adapter)

To bridge the domain gap between natural and infrared images, the FFC Adapter is introduced to explicitly capture high-frequency components. As shown in Fig. 2, it consists of three modules: The Patch Embed (PE) module first divides the input image $I \in \mathbb{R}^{H \times W \times 1}$ into patches and projects them into embeddings. Following this, the FFC Block performs Fourier domain analysis to extract high-frequency information from these embeddings. Finally, a Linear Layer transforms the resulting features to match the target dimension for subsequent processing in the network.

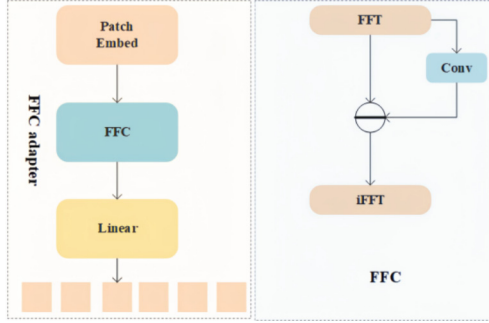


Fig. 2. Structural of the Fast Fourier Convolution (FFC) adapter.

The process is formalized as:

$$F_{he} = \text{Linear}(\text{FFC}(\text{PE}(I))), \quad (5)$$

where F_{he} denotes the high-frequency embeddings, $\text{Linear}(\cdot)$, $\text{FFC}(\cdot)$, and $\text{PE}(\cdot)$ correspond to three distinct modules: the Patch Embed block for dividing input into patches, the FFC block for high-frequency information extraction, and the Linear block for dimension transformation, respectively.

3.4 Branch B: Mask Prediction

The mask prediction branch consists of a mask encoder and a mask decoder. The encoder employs ViT-Tiny as the feature extraction backbone, initialized with EfficientSAM weights pre-trained on a large-scale data set. Specifically, given an input infrared image $I \in \mathbb{R}^{H \times W \times 1}$, the image is first divided into multiple patches, which are then fed into the Embed Block to obtain corresponding feature embeddings $F_i^{me} \in \mathbb{R}^{L \times C}$, where the subscript i and superscript me denote the i -th Embed Block and mask encoder respectively, and L represents the token length of the feature embedding. The encoder integrates high-frequency and edge features to adaptively extract infrared-specific features. This process is formulated as:

$$F_i^{me} = \text{EB}(F_{i-1}^{me} + F_{he} + F_{i-1}^{ee}), i = 1, \dots, 4, \quad (6)$$

$$F_0^{me} = \text{PE}(I). \quad (7)$$

where $EB(\cdot)$ and $PE(\cdot)$ represent the feature extraction Embed Block and patch-dividing Patch Embed block respectively.

Compared to the structurally complex mask encoder, the mask decoder shares a similar architecture with the edge decoder, consisting of lightweight CNN blocks primarily designed for feature aggregation and spatial resolution recovery.

Similar to the edge decoder, the mask decoder differs mainly in that it integrates the edge features F_i^{ed} extracted by the edge decoder during the resolution recovery process. The decoding process of the mask decoder can be described as follows:

$$F_{i+1}^{md} = CNN\left(UP\left(F_i^{md} + F_i^{ed}\right)\right), i = 1, \dots, 4, \quad (8)$$

$$F_1^{md} = F_4^{me}. \quad (9)$$

The final outputs are a predicted mask and an edge map, supervised by mask and edge losses, respectively.

3.5 Loss Function

Dice Loss. To evaluate the consistency between predicted masks and the ground truth, we employ the Dice loss, which originates from the Dice similarity coefficient. It quantifies the proportion of the overlap relative to the total region of both sets, formulated as:

$$L_{Dice} = 1 - \frac{2 \cdot |P \cap T| + \epsilon}{|P| + |T| + \epsilon}, \quad (10)$$

where P denotes the predicted positive region, T represents the ground truth, $|P \cap T|$ indicates their intersection, and ϵ is a smoothing factor to prevent division by zero.

Edge Loss. For boundary supervision, we adopt a joint loss function that integrates Dice loss with the binary cross-entropy (BCE) criterion:

$$L_{edge} = L_{Dice}^{edge} + \lambda L_{BCE}^{edge}, \quad (11)$$

where λ is a hyper parameter that balances the two loss terms, empirically set to 10.

The final training objective combines both the edge loss L_{edge} and the Dice loss for mask prediction L_{Dice}^{mask} :

$$L_{total} = L_{edge} + L_{Dice}^{mask}. \quad (12)$$

4 Experiments

To validate the effectiveness of the proposed method, we conduct comprehensive experiments from multiple perspectives, including few-shot performance and generalization capability. Specifically, the method is evaluated on public datasets and compared with existing deep learning-based approaches. Furthermore, ablation studies are carried out to assess the contribution of each module.

4.1 Datasets and Evaluation Metrics

Datasets. We conduct experiments on two publicIRSTD datasets, including NUDT-SIRST [25] andIRSTD-1k [32]. NUDT-SIRST contains 1327 infrared images, with 663 training images and 664 test images.IRSTD-1k has 1001 infrared images, which is divided into 800 training images and 201 test images.

Evaluation Metrics. We use the following three metrics to evaluate the comparison methods together with ours.

Mean Intersection Over Union (mIoU).

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}[i]}{\text{T}[i] + \text{P}[i] - \text{TP}[i]} \quad (13)$$

where N represents the total number of samples. $\text{TP}[i]$, $\text{T}[i]$, and $\text{P}[i]$ denote the total number of correctly predicted pixels, ground truth target pixels, and predicted target pixels, respectively, pixels belong to real targets, and pixels predicted as the target, respectively. The value of mIoU reflects the similarity between the detected target and the real target to a certain extent.

Probability of Detection (Pd).

$$\text{Pd} = \frac{N_{\text{pred}}}{N_{\text{all}}} \quad (14)$$

where N_{pred} is the number of targets correctly predicted by the model and N_{all} is the total number of real targets. We determine whether it is a positive sample by comparing the deviation of the predicted and real target centroids. Specifically, if the deviation of the target centroid is less than the deviation threshold, we consider these targets as correctly predicted targets; otherwise, we regard these pixels as erroneously predicted pixels.

False-Alarm Rate (Fa). Fa is the ratio of the total number of target pixels incorrectly predicted by the model N_{false} to the total number of pixels in the image N_{all} .

$$\text{Fa} = \frac{N_{\text{false}}}{N_{\text{all}}} \quad (15)$$

4.2 Implementation Details

To validate the effectiveness of our method, we conducted experiments with the following configuration: The experimental environment utilized two NVIDIA GTX 3090 GPUs for training, with CUDA version 11.3, Ubuntu 20.4.6 operating system, and PyTorch 1.12.0 deep learning framework running on Python 3.8. For images with varying resolutions in the dataset, we uniformly resized them to 1024×1024 pixels.

We then employed a combination of random flipping and rotation for data augmentation. During training, we adopted the Adam (adaptive moment estimation) optimizer

with an initial learning rate of 5×10^{-4} . The total training epochs were set to 500, with learning rate decay to one-tenth of the original value at the 200th and 300th epochs. Additionally, validation was performed every 20 epochs to update the weight.

parameters and save the optimal model. Considering both parameter efficiency and performance, we employ the image encoder from EfficientSAM [44] as the mask encoder in our proposed framework. During training, we freeze the mask encoder parameters and only fine-tune the FFC adapter, mask decoder, and edge detector. For inference, only the mask prediction branch generates the final output.

4.3 Experimental Results

Few-shot Comparison Experiment. To evaluate the effectiveness of the proposed method under few-shot settings, we conducted extensive experiments on two public datasets: NUDT-SIRST and IRSTD-1K. Specifically, we randomly sampled 30%, 50%, and 80% of the original training sets from both datasets to construct training subsets.

Quantitative Results. The comparison results with other methods on both datasets are presented in Tables 1 and 2. From left to right in Tables 1 and 2, it can be observed that the detection performance of the model improves consistently as the size of the training data increases (from 30% to 50% and then to 80%). This demonstrates that the performance of data-driven methods heavily depends on the scale of the training data. The underlying reason is that data-driven models aim to fit the common patterns within the data, and a larger, more diverse dataset allows for better generalization and higher detection accuracy.

Table 1. Few-shot comparison experiments on NUDT-SIRST dataset with mIoU (%), pd (%), and fa (10^{-6}) metrics.

Method	30%			50%			80%		
	mIoU	Pd	Fa	mIoU	Pd	Fa	mIoU	Pd	Fa
ACMNet	60.72	85.79	8.13	64.23	95.03	24.36	65.26	95.87	35.07
UIUNet	69.46	81.27	19.44	83.04	96.93	6.92	85.07	95.34	12.18
DNANet	87.09	97.46	22.13	88.18	97.57	7.54	90.65	98.20	2.53
ISTDU	84.37	96.93	19.58	86.26	97.57	26.35	88.83	97.78	34.20
IRSAM	61.33	62.71	23.28	64.03	64.50	5.88	64.53	64.83	4.57
Ours	88.11	97.06	18.32	89.19	97.81	15.32	91.95	98.13	9.65

Table 2. Few-shot comparison experiments on IRSTD-1K dataset with mIoU (%), pd (%), and fa (10^{-6}) metrics.

Method	30%			50%			80%		
	mIoU	Pd	Fa	mIoU	Pd	Fa	mIoU	Pd	Fa
ACMNet	58.14	85.19	25.93	60.48	88.55	17.63	63.18	87.88	31.22
UIUNet	57.89	80.81	13.11	60.17	82.49	7.648	63.30	84.85	19.26
DNANet	60.94	87.53	16.80	61.01	87.54	11.16	65.64	87.21	15.01
ISTDU	60.33	86.54	18.28	61.72	90.91	12.49	65.36	89.23	6.756
IRSAM	62.14	74.06	35.91	63.06	81.19	17.09	63.02	75.55	41.72
Ours	61.21	85.64	14.10	63.29	89.99	9.26	66.07	90.15	10.32

From a dataset-specific perspective, the proposed method consistently achieves the highest mIoU scores under all few-shot training settings, reaching 88.11%, 89.19%, and 91.95%, respectively. This can be attributed to the strong few-shot generalization capability of the SAM image encoder. Specifically, under the 30% training data setting, our method achieves an mIoU of 88.11%, outperforming DNANet’s 87.90%, indicating superior ability in capturing fine structural details. In addition, our method achieves a lower false alarm rate (Fa) compared to DNANet, suggesting fewer false detections.

Similarly, on the IRSTD-1K dataset, our method continues to show superior performance. With 30% of the training data, our approach achieves the best mIoU and second-best Fa, while Pd reaches a competitive 85.64%. With 50% of the training data, the proposed method surpasses the runner-up ISTDU-Net (61.72% mIoU), and achieves the second-best Pd and Fa, at 89.99% and 9.26×10^{-6} , respectively. When the full 80% of the training data is used, our method achieves the best performance with an mIoU of 66.07% and a Pd of 90.15%, as well as the second-best Fa.

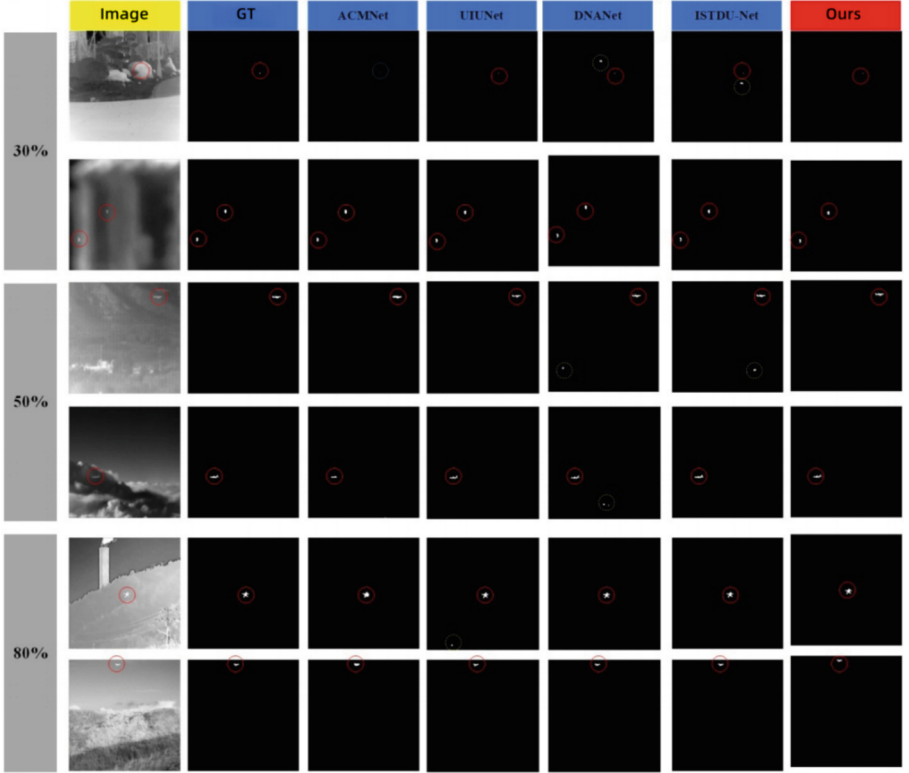


Fig. 3. Visual comparison of detection results under different training sample sizes (30%, 50%, and 80%)

Qualitative Results. To intuitively demonstrate the detection performance of different methods, we conducted qualitative visualization experiments under various representative scenarios, as shown in Fig. 3. It can be observed that the proposed method consistently exhibits superior performance across datasets of different sizes. Specifically, our method detects more targets with more accurate shapes and significantly fewer false alarms. In challenging scenarios where targets are extremely small, difficult to distinguish from the background, or exhibit low contrast, our method can still accurately localize targets with nearly no false alarms, as illustrated in the first row of the Fig. 3. Notably, from rows 2 to 6 in the Fig. 3, the proposed method consistently outperforms other methods in both target localization accuracy and shape fidelity. The detected masks are closer to the ground truth annotations, which can be attributed to the integration of the mask prediction branch and the edge detection branch.

Generalization Comparison Experiment. To evaluate the generalization capability of the proposed method across different scenarios, we conducted cross-dataset experiments using public datasets and compared the results with those of other state-of-the-art methods, as shown in Table 3.

The experimental setup is as follows: we selected training and testing sets from different datasets to simulate domain shift conditions. Specifically, two publicly available datasets, NUDT-SIRST and IRSTD-1K, were used to construct two experimental settings:

- Setting A: The model is trained on the training set of NUDT-SIRST and evaluated on the test set of IRSTD-1K.
- Setting B: The model is trained on the training set of IRSTD-1K and evaluated on the test set of NUDT-SIRST.

Table 3. Cross-dataset generalization comparison experiments.

Method	Setting A			Setting B		
	mIoU(%)	Pd(%)	Fa(10^{-6})	mIoU(%)	Pd(%)	Fa(10^{-6})
ACMNet	21.56	85.52	359.28	35.00	66.77	53.06
UIUNet	28.78	83.17	105.26	41.16	66.46	23.16
DNANet	27.49	74.75	81.44	43.27	69.42	34.95
ISTDU	34.20	84.51	78.02	39.18	67.30	53.50
Ours	31.16	88.45	44.37	44.30	70.69	19.95

Quantitative Results. From the overall analysis of the results presented in Table 3, it is evident that the proposed method, which leverages a pre-trained foundation model, exhibits significantly better generalization performance compared to other data-driven approaches. In both experimental settings, our method achieves outstanding detection performance. Specifically, under Setting A, the proposed method achieves the best overall performance across most evaluation metrics. Under Setting B, the proposed method achieves the best results across all three metrics: mIoU, Pd, and Fa, with values of 44.30%, 70.69%, and 19.95×10^{-6} , respectively. These cross-dataset results demonstrate the strong generalization capability of our approach across different infrared imaging scenarios.

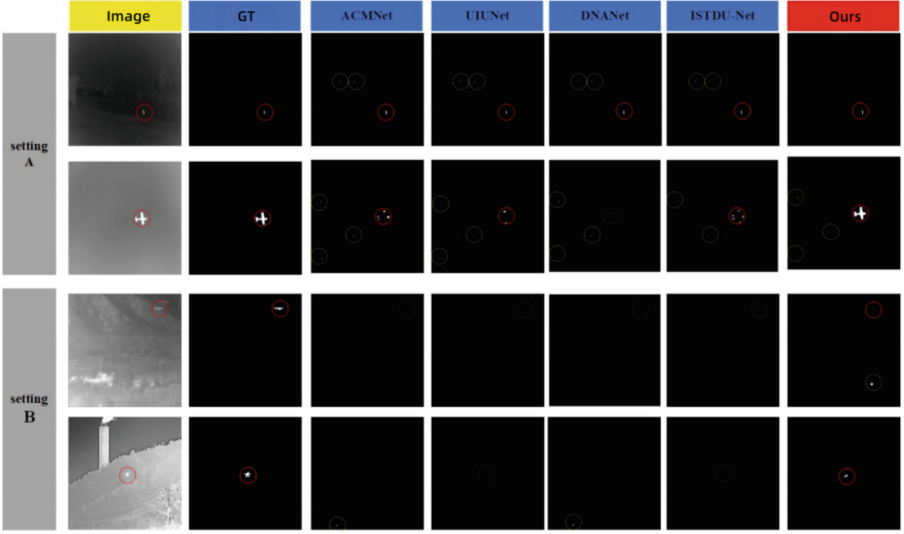


Fig. 4. Visual comparison of detection results under different generalization scenarios.

Qualitative Results. To visually compare the generalization ability of different methods, we present qualitative detection results under several typical generalization scenarios, as shown in Fig. 4. The results demonstrate that the proposed method exhibits superior generalization performance across these scenarios.

From the analysis of the detection results in the first and second rows, it can be observed that for targets with high distinguishability, all methods are able to detect the target, but our method significantly reduces the false alarm rate. For previously unseen background scenes, other methods fail to correctly recognize the target, while our method successfully locates the target, as shown in the last two rows of the Fig. 4.

4.4 Ablation Study

To validate the effectiveness of the components of the proposed method, ablation experiments were conducted on the RSTD-1K data set, with the results presented in Table 4. The experimental details are as follows:

- *w/o FFC Adapter:* In this variant, we removed the Fast Fourier Convolution (FFC) adapter from the network.
- *w/o Edge Branch:* In this variant, we removed the edge detection branch responsible for extracting edge information.
- *w/o Edge Encoder:* In this variant, we disabled the skip connection between the mask encoder and the edge encoder.
- *w/o Edge Decoder:* In this variant, we disabled the skip connection between the mask decoder and the edge decoder.

Table 4. Impact of fundamental components.

Method	mIoU(%)	Pd(%)	Fa(10^{-6})
Ours	67.26	90.55	6.71
w/o FFC adapter	63.65	86.89	9.34
w/o edge branch	61.74	84.85	28.24
w/o edge encoder	62.30	88.55	24.18
w/o edge decoder	62.37	85.86	14.82

From the overall analysis of the data in the Table 4, it can be seen that the proposed network achieves optimal scores in all three metrics. When the FFC adapter is removed from the network, the detection performance declines to varying degrees, as shown by a 3.6% decrease in mIoU, a nearly 4% decrease in Pd, and a slight increase in false alarm rate (Fa). This may be because the FFC adapter learns the frequency domain representation of targets in infrared images during training, extracting high-frequency features from the images to guide the base model in extracting infrared target features, thus reducing performance deviation caused by style domain differences.

From the data in the last three rows of the Table 4, it can be concluded that the proposed dual-branch network design enhances the model’s detection performance. When the edge detection branch is removed, the model performance drops significantly, with mIoU decreasing from 67.27% to 61.74%, Pd dropping from 90.55% to 84.85%, and the false alarm rate (Fa) increasing to four times the original value. This is because the lightweight edge detection branch effectively extracts the physical edge information of targets in infrared images during supervised training, and its fusion with the encoder and decoder of the mask branch through skip connections helps the mask branch accurately detect the targets, thus improving detection performance.

Additionally, to explore the impact of integrating edge information at different stages of the mask encoder and decoder on the model’s detection performance, we removed the skip connections in either the encoder or decoder of the proposed method. The results, shown in rows 4 and 5 of the Table 4, indicate a decline in performance compared to the proposed method (data in row 1), with the worst performance observed when both connections are removed (row 3 data). The data and analysis from the Table 4 validate the effectiveness of the components in the proposed method.

5 Discussion

While the proposed E2SAM demonstrates strong performance across few-shot and cross-domain scenarios, its real time applicability and scalability remain important considerations for practical deployment. Moreover, performance degradation in precise shape reconstruction is still observed under large domain shifts. Future work may consider developing domain adaptation strategies to reduce style discrepancy between natural and infrared images. Additionally, pre-training SAM-like models directly on large-scale infrared datasets could enhance feature extraction for this specific modality. Exploring semi-supervised or zero-shot learning approaches and integrating prompt-based

adaptation mechanisms could also improve model flexibility in low-resource settings. Extending E2SAM to handle temporal dynamics in video sequences or more complex multi-target scenes remains another promising direction to strengthen its practical utility.

6 Conclusion

In this article, we present a novel E2SAM for infrared small target detection tasks that fully leverages the powerful feature extraction capability of the base model. At the same time, the algorithm is specifically designed with an improved architecture suitable for infrared remote sensing image target detection, addressing the fact that the base model is trained with supervision on natural images. To reduce the interference caused by the differences in image characteristics, we designed a dual-branch network architecture and a Fast Fourier Convolution (FFC) adapter. The edge detection branch extracts target edge features through supervised learning, while the FFC adapter learns the high-frequency features of the image and fuses these features with edge features, guiding the base model to more effectively extract target features in infrared images. Experiments on public datasets show that the proposed method can maintain excellent detection performance even when trained on a small sample size. Additionally, the method demonstrates outstanding scene generalization ability, validating its effectiveness in practical applications. The experimental results further demonstrate that the proposed design can effectively alleviate the impact of the characteristic differences between infrared and natural images on detection performance, providing an efficient and robust solution for infrared image target detection based on pre-trained models.

References

1. Zhang, M., et al.: Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1730–1738 (2022)
2. Pang, D., Ma, P., Feng, Y., Shan, T., Tao, R., Jin, Q.: Tensor spectral k-support norm minimization for detecting infrared dim and small target against urban backgrounds. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–13 (2023)
3. Becker, L.: Influence of ir sensor technology on the military and civil defense. In: Quantum Sensing and Nanophotonic Devices III, vol. 6127, pp. 180–194. SPIE (2006)
4. Yan, P., Hou, R., Duan, X., Yue, C., Wang, X., Cao, X.: Stdmanet: Spatio-temporal differential multiscale attention network for small moving infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–16 (2023)
5. Rivest, J.-F., Fortin, R.: Detection of dim targets in digital infrared imagery by morphological image processing. *Opt. Eng.* **35**(7), 1886–1893 (1996)
6. Deshpande, S.D., Er, M.H., Venkateswarlu, R., Chan, P.: Max-mean and max-median filters for detection of small targets. In: Signal and Data Processing of Small Targets 1999, vol. 3809, pp. 74–83. SPIE (1999)
7. Chen, C.P., Li, H., Wei, Y., Xia, T., Tang, Y.Y.: A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **52**(1), 574–581 (2013)
8. Han, J., Ma, Y., Zhou, B., Fan, F., Liang, K., Fang, Y.: A robust infrared small target detection algorithm based on human visual system. *IEEE Geosci. Remote Sens. Lett.* **11**(12), 2168–2172 (2014)

9. Han, J., Moradi, S., Faramarzi, I., Liu, C., Zhang, H., Zhao, Q.: A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Geosci. Remote Sens. Lett.* **17**(10), 1822–1826 (2019)
10. Han, J., et al.: Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **18**(9), 1670–1674 (2020)
11. Kim, S., Lee, J.: Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track. *Pattern Recognit.* **45**(1), 393–406 (2012)
12. Wang, X., Lv, G., Xu, L.: Infrared dim target detection based on visual attention. *Infrared Phys. Technol.* **55**(6), 513–521 (2012)
13. Gao, C., Meng, D., Yang, Y., Wang, Y., Zhou, X., Hauptmann, A.G.: Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Process.* **22**(12), 4996–5009 (2013)
14. Dai, Y., Wu, Y.: Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **10**(8), 3752–3767 (2017)
15. Zhang, L., Peng, L., Zhang, T., Cao, S., Peng, Z.: Infrared small target detection via non-convex rank approximation minimization joint $\ell_{2,1}$ norm. *Remote Sens.* **10**(11), 1821 (2018)
16. Zhang, L., Peng, Z.: Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **11**(4), 382 (2019)
17. Zhu, H., Liu, S., Deng, L., Li, Y., Xiao, F.: Infrared small target detection via low-rank tensor completion with top-hat regularization. *IEEE Trans. Geosci. Remote Sens.* **58**(2), 1004–1016 (2019)
18. Dai, Y., Wu, Y., Song, Y., Guo, J.: Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. *Infrared Phys. Technol.* **81**, 182–194 (2017)
19. Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **59**(11), 9813–9824 (2021)
20. Zhang, T., Li, L., Cao, S., Pu, T., Peng, Z.: Attention-guided pyramid context networks for detecting infrared small target under complex background. *IEEE Trans. Aerosp. Electron. Syst.* **59**(4), 4250–4261 (2023)
21. Tong, X., et al.: Msaffnet: A multiscale label-supervised attention feature fusion network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–16 (2023)
22. Zhang, M., Yue, K., Zhang, J., Li, Y., Gao, X.: Exploring feature compensation and cross-level correlation for infrared small target detection. In: *Proc. 30th ACM Int. Conf. Multimedia*, pp. 1857–1865 (2022)
23. Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Asymmetric contextual modulation for infrared small target detection. In: *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 950–959 (2021)
24. 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 (DLMIA 2018, ML-CDS 2018, MICCAI 2018 Workshops)*, pp. 3–11. Springer (2018)
25. Li, B., et al.: Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **32**, 1745–1758 (2022)
26. Yuan, S., Qin, H., Yan, X., Akhtar, N., Mian, A.: Sctransnet: Spatial-channel cross transformer network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* (2024)
27. Kirillov, A., et al.: Segment anything. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 4015–4026 (2023)
28. Wang, H., Zhou, L., Wang, L.: Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 8509–8518 (2019)

29. Wu, X., Hong, D., Chanussot, J.: Uiu-net: U-net in u-net for infrared small object detection. *IEEE Trans. Image Process.* **32**, 364–376 (2022)
30. 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.* **19**, 1–5 (2022)
31. Liu, F., Gao, C., Chen, F., Meng, D., Zuo, W., Gao, X.: Infrared small and dim target detection with transformer under complex backgrounds. *IEEE Trans. Image Process.* **32**, 5921–5932 (2023)
32. Zhang, M., et al.: Isnet: Shape matters for infrared small target detection. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 877–886 (2022)
33. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
34. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nat. Commun.* **15**(1), 654 (2024)
35. Wu, J., et al.: Medical sam adapter: Adapting segment anything model for medical image segmentation. *Med. Image Anal.* **102**, 103547 (2025)
36. Huang, C., et al.: Adapting visual-language models for generalizable anomaly detection in medical images. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 11375–11385 (2024)
37. Shaharabany, T., Dahan, A., Giryas, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370* (2023)
38. Zhan, C., Duan, X., Xu, S., Song, Z., Luo, M.: An improved moving object detection algorithm based on frame difference and edge detection. In: *Proc. 4th Int. Conf. Image Graphics (ICIG 2007)*, pp. 519–523. *IEEE* (2007)
39. Zhao, J., Yu, C., Shi, Z., Liu, Y., Zhang, Y.: Gradient-guided learning network for infrared small target detection. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5 (2023)
40. Lin, F., Ge, S., Bao, K., Yan, C., Zeng, D.: Learning shape-biased representations for infrared small target detection. *IEEE Trans. Multimedia* **26**, 4681–4692 (2023)
41. Yang, J., et al.: Sam2-elnnet: Label enhancement and automatic annotation for remote sensing segmentation. *arXiv preprint arXiv:2503.12404* (2025)
42. Zhou, M., Zhang, Y., Xu, X., Wang, J., Khalvati, F.: Edge-enhanced dilated residual attention network for multimodal medical image fusion. In: *2024 IEEE Int. Conf. Bioinformatics Biomedicine (BIBM)*, pp. 4108–4111. *IEEE* (2024)
43. Zeng, Y.: Mgf-net: A multi-grained feature integration network for enhanced medical image segmentation. *arXiv preprint arXiv:2502.13808* (2025)
44. Xiong, Y., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 16111–16121 (2024)