

SAIST: Segment Any Infrared Small Target Model Guided by Contrastive Language-Image Pretraining

Mingjin Zhang¹, Xiaolong Li^{1*}, Fei Gao¹, Jie Guo^{1*}, Xinbo Gao^{1,2}, Jing Zhang³

¹Xidian University, China ²Chongqing University of Posts and Telecommunications, China

³School of Computer Science & School of Artificial Intelligence, Wuhan University, China

{mjinzhang, fgao}@xidian.edu.cn, 23011210739@stu.xidian.edu.cn

jguo, xbgao@mail.xidian.edu.cn, jingzhang.cv@gmail.com

Abstract

Infrared Small Target Detection (IRSTD) aims to identify low signal-to-noise ratio small targets in infrared images with complex backgrounds, which is crucial for various applications. However, existing IRSTD methods typically rely solely on image modalities for processing, which fail to fully capture contextual information, leading to limited detection accuracy and adaptability in complex environments. Inspired by vision-language models, this paper proposes a novel framework, SAIST, which integrates textual information with image modalities to enhance IRSTD performance. The framework consists of two main components: Scene Recognition Contrastive Language-Image Pretraining (SR-CLIP) and CLIP-guided Segment Anything Model (CG-SAM). SR-CLIP generates a set of visual descriptions through object-object similarity and object-scene relevance, embedding them into learnable prompts to refine the textual description set. This reduces the domain gap between vision and language, generating precise textual and visual prompts. CG-SAM utilizes the prompts generated by SR-CLIP to accurately guide the Mask Decoder in learning prior knowledge of background features, while incorporating infrared imaging equations to improve small target recognition in complex backgrounds and significantly reduce the false alarm rate. Additionally, this paper introduces the first multimodal IRSTD dataset, MIRSTD, which contains abundant image-text pairs. Experimental results demonstrate that the proposed SAIST method outperforms existing state-of-the-art approaches.

1. Introduction

Infrared small target detection (IRSTD) is crucial for applications like traffic management and maritime rescue

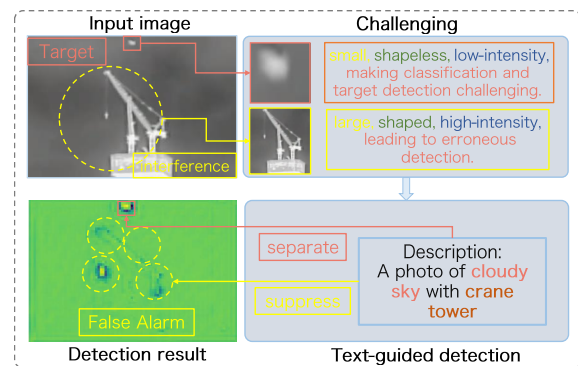


Figure 1. Illustration of the advantages of text-guided detection over image-only detection in identifying small infrared targets in complex backgrounds. The textual description offers valuable context, helping to reduce interference and distinguish the target from the background effectively.

[5, 18, 29]. Unlike visible light, infrared sensors detect thermal radiation, enabling target acquisition in adverse conditions such as fog, rain, snow, and at night [44]. However, due to the long-range imaging mechanism, infrared images often lack distinctive features like color, shape, and texture, while the background typically shows strong structural features with higher intensity and sharper contours [9, 43]. These challenges—complex backgrounds, clutter, and noise—hinder accurate detection, complicating IRSTD.

IRSTD methods can be divided into traditional algorithms and deep learning-based ones. Early methods focused on filtering and image enhancement [1, 6, 12], which performed well in simpler scenarios but struggled with complex environments due to reliance on prior knowledge and handcrafted features. With the rise of deep learning and large IRSTD datasets, newer approaches using Generative Adversarial Networks (GANs) [30], U-Net [7, 21, 38], and transformers [31, 35, 37] have been introduced.

Despite advances, existing methods are limited by their reliance on *single-modality inputs*. Recently, the CLIP-SAM model, which combines Contrastive Language-

*Corresponding authors

Image Pretraining (CLIP) and the Segment Anything Model (SAM), has shown promise in enhancing cross-modal feature fusion for various tasks [20]. CLIP [24] captures semantic relationships between images and text, offering rich supervision, while SAM [16] provides precise segmentation. However, applying CLIP-SAM to IRSTD remains challenging due to the *differences between infrared and natural images*. Specifically: **1)** Infrared targets lack typical visual features, limiting the generalization of the CLIP model; **2)** Small, weak infrared targets and noisy backgrounds complicate target detection, making SAM’s application difficult. To address these issues, we propose using CLIP to extract semantically rich prior knowledge from text, guiding SAM to better capture target details amidst complex backgrounds.

This paper introduces a novel IRSTD approach named SAIST by combining text and image modalities within a multimodal learning framework. First, we design the Scene Recognition CLIP (SR-CLIP) to capture both global scene and local object features. SR-CLIP creates similarity matrices for objects and scenes, facilitating key information extraction. It also integrates a cross-modal interaction mechanism, incorporating scene-specific visual cues into learnable prompts that refine text features and mitigate the vision-language domain gap. Building on this, we diversify CG-SAM to utilize the text prompts from SR-CLIP to guide SAM in learning background features, enhancing accurate identification and extraction across diverse scenes. By combining global semantic and background features, and leveraging the radiative transfer properties of targets and backgrounds, CG-SAM achieves effective target-background separation. Additionally, we present MIRSTD, the first multimodal IRSTD dataset, which combines image-text pairs from NUAA-SIRST, NUDT-SIRST, and IRSTD-1K datasets. Text descriptions are generated by GPT-4V and carefully reviewed manually. Experiments validate the superiority of SAIST over state-of-the-art (SOTA) methods.

In summary, our main contributions are as follows:

- We introduce SAIST, the first approach to integrate multimodal vision and language for IRSTD. On challenging benchmarks, it outperforms SOTA methods, excelling in both objective metrics and subjective evaluations.
- We design two key components, SR-CLIP and CG-SAM, to effectively fuse multimodal information and separate targets and backgrounds in complex scenes.
- We create a novel multimodal IRSTD dataset combining infrared images and textual data, offering rich resources for training and evaluating multimodal learning methods.

2. Related work

2.1. Infrared small target detection

Traditional IRSTD methods use image processing or handcrafted features, such as low-rank (*e.g.*, RIPT [6],

PSTNN [36]), human visual system-based (*e.g.*, TLLCM [4], WSLCM [12]), and filter-based approaches (*e.g.*, top-hat [1]). Low-rank methods perform well in low SNR but struggle with false alarms in low-contrast areas. Human visual methods are effective with large objects but not in complex scenes, and filter-based techniques suppress uniform noise but fail in complex backgrounds. **Deep learning IRSTD methods** include CNN-based models (*e.g.*, MD-vsFA [30], ISNet [39], Dim2Clear [40], CSRNet [21], DMFNet [11], GCI-Net [41]) that focus on local features, with some improving SNR, detection, or small target extraction, and hybrid models (*e.g.*, IAANet [31] and RK-former [37]) that combine CNNs and Vision Transformers (ViTs) to capture both local and global context. While IRSTD has made significant strides, current methods rely mainly on image modalities, which fail to capture full contextual information, limiting detection accuracy in complex backgrounds. Integrating textual information can offer substantial benefits, providing richer scene understanding and aiding target extraction in challenging environments.

2.2. CLIP and SAM Collaboration

In recent years, foundational models like CLIP and SAM have garnered significant attention for their exceptional performance across various tasks. CLIP excels at aligning multimodal features, while SAM specializes in segmentation through diverse prompts. Recent research has explored synergies between these models, such as CRIS [32] integrates both models to boost performance. ClipSAM [20] uses CLIP’s semantic understanding for anomaly localization and coarse segmentation, later refined by SAM. Open-Vocabulary SAM enhances SAM’s recognition by transferring knowledge from SAM to CLIP via distillation and adapters. However, applying pre-trained CLIP-SAM models, originally trained on natural images, to IRSTD faces challenges due to domain differences. In this work, we leverage CLIP’s semantic understanding from scene descriptions to guide SAM in extracting target information from complex infrared backgrounds.

2.3. Datasets for IRSTD

Prior studies have developed large-scale datasets [7, 19, 39], which have advanced IRSTD research. However, these datasets only include image data with target masks, limiting the detection of small targets in complex backgrounds. Recent multimodal approaches, such as text-guided detection [17, 22, 25], have shown performance improvements. Inspired by this, we create the MIRSTD dataset, which combines both image and text modalities. The text modality in MIRSTD provides descriptions of scenes in infrared images from datasets like NUAA-SIRST, IRSTD-1k, and NUDT-SIRST, helping to mitigate the challenge of weak target signals against strong background noise.

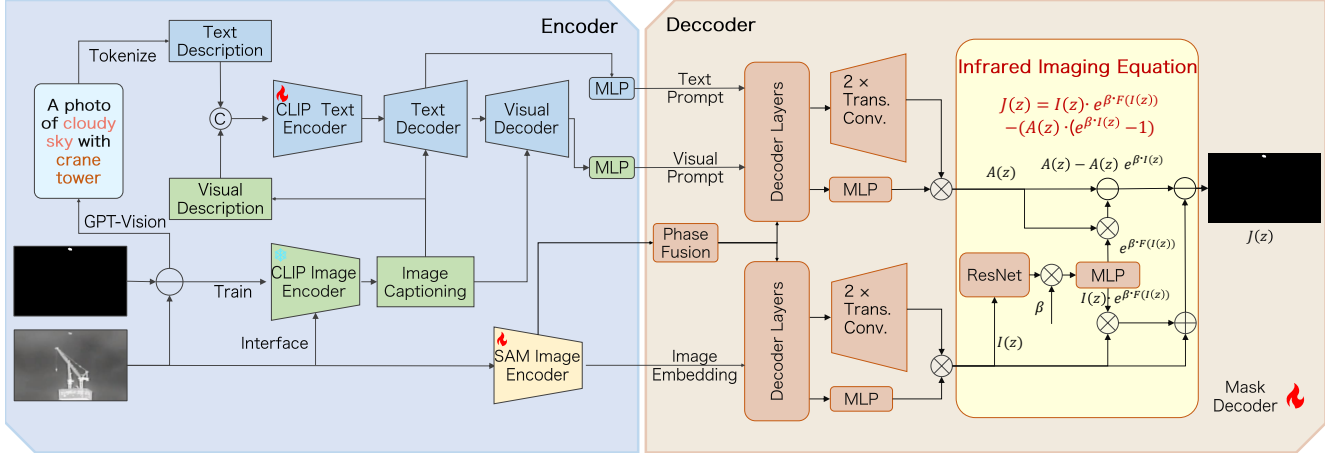


Figure 2. Overview of the proposed encoder-decoder architecture. The CLIP text encoder converts textual descriptions into feature representations, while the frozen image encoder processes infrared images without target. SR-CLIP leverages cross-modal interactions to refine textual features, bridging the visual-textual domain gap. The SAM image encoder extracts image embeddings, fine-tuned with LoRA to enhance task adaptability. The generated textual prompts and image embeddings are fed into the SAM mask decoder, along with an infrared imaging model, to effectively suppress background noise and detect small targets.

3. Methodology

3.1. Overall Architecture

The overall architecture of SAIST is illustrated in Fig. 2. In this framework, the text encoder of SR-CLIP converts textual descriptions into feature representations, while the image encoder is frozen. Next, SR-CLIP employs a cross-modal interaction mechanism to integrate the visual and textual description sets, refining the textual features and reducing the domain gap between vision and language, thus generating precise textual and visual prompts. Subsequently, the pre-trained and frozen SAM image encoder extracts image embeddings, which are fine-tuned using Low-Rank Adaptation (LoRA) [14] to enhance the model’s adaptability to the specific task. Finally, the generated textual prompts and image embeddings are fed into the SAM mask decoder, in conjunction with a specially designed infrared imaging model, to effectively suppress background noise and accurately detect small targets.

3.2. SR-CLIP

In infrared imaging, due to the inherent characteristics of remote sensing, the target’s radiative intensity is typically weak, while the background’s intensity is higher and more prominent. In this case, the information related to the target is difficult to describe clearly and accurately. In contrast, the background’s radiative intensity is more stable and easier to identify, allowing the overall structure and features of the scene to be effectively reflected through intuitive textual descriptions. There, we design SR-CLIP to fully leverage the complementary relationship between background textual descriptions and visual features, generating precise

prompts that guide SAM to effectively suppress background interference in infrared images.

3.2.1. Visual / Text description set

For the scene recognition task in infrared images, scenes are composed of multiple semantic concepts, including distinct objects and backgrounds. These features necessitate the consideration of various semantic information when learning discriminative patterns. To capture comprehensive scene information, we introduce a novel form of textual prompting, referred to as scene prompts. This prompt creates a textual description set:

$$\{a \text{ photo of } [scene] \text{ with } [object], [object], \text{ and } [object]\},$$

which provides both scene-level and object-level textual descriptions. Simultaneously, to bridge the image-text domain gap, we extract local object features from Clip’s image encoder, acquire global visual features with the help of a learnable projection layer, and generate a set of visual feature descriptions using the image captioning [13, 34].

$$\{[scene], [object], [object] [object]\},$$

which provides both scene-level and object-level visual descriptions.

3.2.2. Image Captioning

In scene recognition, objects often share considerable similarity, and there is a strong correlation between objects and the scene, which reflects the unique characteristics of that particular scene. By integrating both interactions, the model can capture more granular semantic information, thereby improving its discriminative power in specific contexts. To

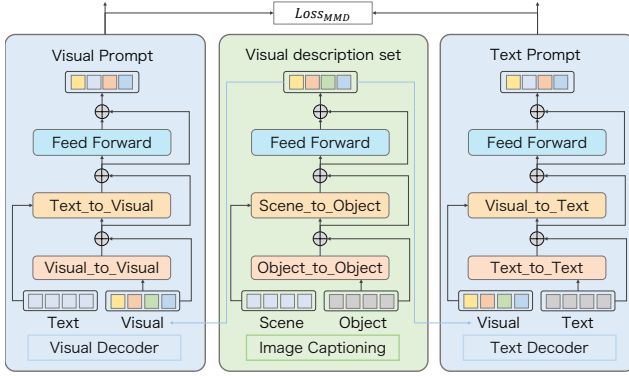


Figure 3. Overview of the Cross-modal Interaction Mechanism. Image captioning generates visual feature descriptions that are processed by a text encoder. The text decoder then maps visual information into the text, enhancing text feature expressiveness. Finally, the visual decoder recovers visual information from the encoded text to improve image-text alignment.

this end, we construct both an object-object similarity matrix and a scene-object correlation matrix:

Object-Object Similarity Matrix:

$$\begin{bmatrix} o_{11} & \cdots & o_{1n} \\ \vdots & \ddots & \vdots \\ o_{n1} & \cdots & o_{nn} \end{bmatrix} = \frac{\exp\left(\frac{o_i^\top o_j}{\|o_i\| \|o_j\|}\right)}{\sum_{k=1}^n \exp\left(\frac{o_i^\top o_k}{\|o_i\| \|o_k\|}\right)}, \quad (1)$$

where $o_{i,j}$ represents the similarity between the feature vector of object i and object j .

Scene-Object Correlation Matrix:

$$[s_1 \quad \cdots \quad s_n] = \frac{\exp\left(\frac{s^\top o_i}{\|s\| \|o_i\|}\right)}{\sum_{k=1}^n \exp\left(\frac{s^\top o_k}{\|s\| \|o_k\|}\right)}, \quad (2)$$

where s_i represents the correlation between the feature vector of the scene and the feature vector of object j .

Therefore, based on the similarity and correlation matrices, we design the Object-to-Object and Scene-to-Object modules to effectively capture the interrelationships between objects and facilitate the integration of scene and object details. Meanwhile, by incorporating image captioning techniques, we generate precise visual description.

3.2.3. Cross-Modal Interaction Mechanism

CoOp [42] can be viewed as a textual domain prompt, where the learnable context is simply a trainable vector without incorporating any visual information. Inspired by CoOp, we propose a cross-modal interaction mechanism that introduces the visual description set into the text description set to refine the textual features. This approach mitigates the domain gap between the visual and textual modalities, thereby enhancing the performance of the CLIP model in downstream tasks. Initially, as shown in Fig. 2

we integrate the visual and textual description sets to form a comprehensive input description set, which is then processed by the CLIP text encoder. Meanwhile, to further enhance alignment, we expand the output dimension of the final learnable linear projection layer in CLIP, enabling it to generate more expressive and informative textual features.

$$\text{Description} = \text{Concat}(T_{\text{scene}}, T_{\text{objects}}, V_{\text{scene}}, V_{\text{objects}}), \quad (3)$$

$$\text{Text} = \text{CLIP Text Encoder}(\text{Description}). \quad (4)$$

To effectively transfer fine-grained semantic features from the visual description set to the textual description set, we designed a text decoder. As illustrated in Fig. 3, the text decoder first employs the Text-to-Text module to refine and extract textual information. Then, based on the correlation between image features and textual descriptions, it leverages the Visual-to-Text module to dynamically map visual information onto textual descriptions. This process facilitates the recovery of semantic details relevant to the image content, thereby enhancing the expressiveness and informativeness of the textual descriptions.

$$\mathcal{T}_{\text{Prompt}} = \text{Text Decoder}(\text{Visual}, \text{Text}), \quad (5)$$

Meanwhile, in order to recover the visual information from the set of text descriptions encoded by CLIP text encoder, we designed a visual decoder. As illustrated in Fig. 3, the decoder effectively extracts the visual features of the background through the Visual-to-Visual module. It then leverages the Text-to-Visual module to perform cross-modal conversion, capturing the semantic relationships between the image and textual descriptions. This process enhances the mutual understanding and alignment between visual and textual modalities, improving the coherence and expressiveness of the descriptions.

$$\mathcal{V}_{\text{Prompt}} = \text{Visual Decoder}(\text{Visual}, \text{Text}). \quad (6)$$

Loss Function: In cross-modal tasks, textual and visual description sets come from different modalities, and their feature spaces may differ significantly. To enable effective use of both modalities, it's crucial to align their distributions in a shared semantic space. The Maximum Mean Discrepancy (MMD)[2, 27] measures the squared distance between the means of two distributions X_S and X_T in the feature space:

$$D_{\mathcal{H}}(X_S, X_T) = \sup_{\|\phi\|_{\mathcal{H}} \leq 1} \|\mathbb{E}_S[\phi(X_S)] - \mathbb{E}_T[\phi(X_T)]\|_{\mathcal{H}}^2, \quad (7)$$

where ϕ is the feature map, and $\mathbb{E}_S[\phi(X_S)]$ and $\mathbb{E}_T[\phi(X_T)]$ are the expected values of the mapped samples. To align

textual and visual prompts, we propose an MMD loss:

$$\begin{aligned} \mathcal{L} = \mathcal{L}(D_{\mathcal{H}}(X_S, X_T)) &= \frac{1}{|X_n^s|^2} \sum_{i=1}^{|X_n^s|} \sum_{j=1}^{|X_n^s|} k(x_i^{(s)}, x_j^{(s)}) \\ &- \frac{2}{|X_n^s| |X_n^t|} \sum_{i=1}^{|X_n^s|} \sum_{j=1}^{|X_n^t|} k(x_i^{(s)}, x_j^{(t)}) \\ &+ \frac{1}{|X_n^t|^2} \sum_{i=1}^{|X_n^t|} \sum_{j=1}^{|X_n^t|} k(x_i^{(t)}, x_j^{(t)}). \end{aligned} \quad (8)$$

Textual prompt and Visual prompt feature. The samples $x_i^{(s)}$ and $x_j^{(s)}$ represent the feature descriptions in the source domain X_S , and these features are used to represent the Text prompt. Similarly, $x_i^{(t)}$ and $x_j^{(t)}$ represent the feature descriptions in the target domain X_T for the Visual prompt. These features are utilized in the MMD calculation to minimize the distributional gap between textual and visual representations, optimizing the alignment of both modalities for improved performance in cross-modal matching and generation tasks.

Kernel function for similarity calculation. $k(x_i, x_j)$ calculates the similarity between samples x_i and x_j . The kernel function is used to compute similarity, which helps in measuring the relationships between samples in the source and target domains.

Number of samples in the source and target domain. $|X_n^s|$, $|X_n^t|$ denotes the number of samples in the source domain and target domain. It is used to normalize the number of samples, ensuring that the influence of each sample is balanced in the calculation.

Minimizing this loss reduces the distributional gap between the textual and visual prompt, aligning the information for better performance in image-text matching.

3.3. CG-SAM

Infrared small targets often exhibit limited visual features, while the background is typically filled with significant clutter and noise. In this context, the SAM encoder faces an inherent trade-off between providing detailed target information and suppressing background noise. Specifically, the early layers of the ViT architecture retain more local details, which is crucial for recognizing small targets with limited visual features. However, these early layers also preserve a substantial amount of background noise, increasing the complexity of target detection. Therefore, directly using the Mask Decoder to recover clear target masks from complex backgrounds presents a significant challenge.

3.3.1. Infrared Imaging Equation

To improve SAM's performance in IRSTD, we propose a novel approach based on the fundamental principles of in-

frared imaging. The Beer-Lambert law [23, 28], a fundamental principle in optics and spectroscopy, is widely applied in the study of various light transmission problems. Ideally, the transmission of infrared radiation from a target to a sensor can be described by the Beer-Lambert law. Thus, by drawing on the Beer-Lambert law and integrating it with the core principles of infrared imaging, we propose a novel approach to model the propagation of infrared radiation and its effects on imaging, especially in scenarios where background interference leads to degradation of the observed target. In the basic model of infrared imaging, the fundamental equation is defined as:

$$I(z) = J(z) \cdot t(z) + A(z) \cdot (1 - t(z)), \quad (9)$$

where:

- $I(z)$: The pixel intensity of the infrared observation image at position z , which includes background radiation and interference.
- $J(z)$: The true infrared radiation intensity of the target at position z , representing the target's brightness in the infrared spectrum.
- $t(z)$: The transmission factor, representing the energy attenuation of infrared radiation from the target to the observer, with values in the range $[0, 1]$.
- $A(z)$: The background radiation refers to the infrared intensity from the background at position z .

The definition of the transmission factor is given by:

$$t(z) = e^{-\beta \cdot d(z)}, \quad (10)$$

where:

- β : The infrared radiation coefficient, determining the absorption level of infrared radiation by different components. This value typically varies depending on the infrared wavelength (e.g., near-infrared, shortwave infrared, or mid-to-far infrared).
- $d(z)$: The scene depth, representing the distance between the target and the imaging device. In infrared imaging, a greater distance results in lower transmission, as infrared radiation undergoes greater loss in the atmosphere and other media.

Given a 2D image, we need a function to calculate its corresponding depth. This process can be expressed as:

$$d(z) = F(I(z)). \quad (11)$$

Considering the scene's varying distances, classification techniques can be employed. Cao et al.[3] treated the depth estimation problem as a pixel-level classification task.

$$F(I(z)) = \text{ResNet}(I(z)). \quad (12)$$

Using the above equations, we can calculate the infrared target's radiation intensity:

$$J(z) = \frac{I(z)}{t(z)} - \frac{A(z)}{t(z)} + A(z). \quad (13)$$

Substituting the transmission factor yields:

$$J(z) = I(z) \cdot e^{\beta \cdot F(I(z))} - A(z) \cdot (e^{\beta \cdot F(I(z))} - 1). \quad (14)$$

Through the formulations outlined above, we can more effectively separate the target from the background, thereby improving the performance of infrared small target detection. This approach constructs an interpretable mathematical model based on the fundamental principles of infrared imaging, providing a robust and explainable solution for infrared small target detection.

3.3.2. CG-SAM

Based on the infrared imaging equation (Eq. 14), we achieve precise infrared small target detection by inversely separating infrared degraded images from background images. As shown in Fig. 2, the `text_prompt` and `visual_prompt` are input into the SAM mask decoder, where they are fused with the early and late global semantic feature maps extracted by the ViT encoder. Building upon this, `text_prompt` and `visual_prompt` effectively extract and identify background features $A(z)$ from different scenes through the scene prompting mechanism. Similarly, we feed the learnable output token from SAM and the image embeddings from the SAM encoder into the mask decoder to further refine the target mask representation. This process gradually recovers target features, but due to background interference, it ultimately generates the infrared degraded image $I(z)$ with background noise.

Finally, based on the degraded image $I(z)$ and the background features $A(z)$, we apply the infrared imaging equation to achieve precise detection of infrared small targets in complex backgrounds while effectively suppressing background interference. Through this integrated process, the CG-SAM can accurately identify small targets and significantly improve the accuracy of target detection in infrared images with strong background noise.

3.4. Datasets

The goal of data generation is to create descriptive text for infrared images to support small target detection in complex backgrounds. To achieve this, we enhanced the IRSTD-1k, NUAA-SIRST, and NUDT-SIRST datasets by offering textual descriptions, resulting in the multimodal image-text dataset collections MIRSTD. Each infrared image, with the target mask removed, is encoded as a Base64 string and used as a text prompt for a visual-language model. ChatGPT generates concise descriptions that summarize key scene information, which are checked manually. The expert’s prompt follows the format: a photo of a scene with [object], [object], and [object] offering a brief description of the scene and its objects. The final output is a 25-word description that captures the image’s essential background information.

4. Experiments

4.1. Experimental Settings

Dataset: We evaluate our model on the MIRSTD dataset, which consists of 427 and 1,000 real infrared images with one or more small targets, and 1,327 synthetic infrared images with small targets. All images are resized to 1024×1024 . For each dataset, we split the images into three disjoint subsets: 50% for training, 30% for validation, and 20% for testing.

Implementation Details: Our SAIST model is trained using the Adam optimizer with an initial learning rate of 0.0001, and a Cosine Decay Learning Rate Scheduler over 400 epochs. We use a batch size of 4 and train on a single Nvidia GeForce A800 GPU with 80 GB of memory. For comparison, we evaluate against state-of-the-art CNN-based methods for IRSTD, including ISNet [39], GCINet [41], UIUNet [33], DNANet [19], Dim2Clear [40], ALCNet [8], and ACMNet [7], as well as hybrid methods such as SCTransNet [35], IAAANet [31], SAM [16] and SAM-HQ [15]. Traditional methods Top-Hat [1] and IPI [10] are also included. We adopt Intersection over Union (IoU) for pixel-level evaluation, and Probability of Detection (Pd) and False Alarm Rate (Fa) for object-level evaluation.

4.2. Comparative Experiments

Quantitative results: As shown in Tab. 1, traditional methods struggle with complex scenarios, especially CNNs, which fail to distinguish targets from backgrounds, resulting in higher false alarm and missed detection rates. The proposed SAIST outperforms existing SOTA methods across all metrics, particularly in reducing false alarms. SR-CLIP improves textual prompt generation through visual-textual alignment, while CG-SAM leverages both SR-CLIP’s guidance and infrared imaging equations to enhance small target recognition in complex backgrounds.

Visual Results: Fig. 4 shows detection results for SR-CLIP and CG-SAM, outperforming other IRSTD methods. SR-CLIP benefits from precise visual and textual guidance, while CG-SAM effectively integrates infrared imaging equations. Both methods excel in identifying small targets in complex backgrounds, reducing F_a and background noise, thus enhancing detection accuracy and robustness.

ROC: As shown in Fig. 5, our SAIST significantly outperforms the other methods, where the Area Under the ROC Curve (AUC) of our SAIST is larger than that of the traditional and deep learning based methods.

4.3. Ablation Study

To evaluate the components of SAIST, we conducted ablation studies on the NUAA-SIRST dataset, using CLIP + SAM as the baseline model. We systematically introduced the CG-SAM and SR-CLIP modules. As shown in the

Method	NUDT-SIRST			IRSTD-1k			NUAA-SIRST			Type
	IoU \uparrow	P _d \uparrow	F _a \downarrow	IoU \uparrow	P _d \uparrow	F _a \downarrow	IoU \uparrow	P _d \uparrow	F _a \downarrow	
IPI [10]	30.93	81.98	17.99	27.92	81.37	16.18	1.09	87.05	30467	Traditional
Top-Hat [1]	22.40	89.90	174.1	10.06	75.11	1432	1.508	79.74	16456	Traditional
MSLSTIPT [26]	8.34	47.40	881	11.43	79.03	1524	1.080	0.052	8183	Traditional
ACM [7]	68.90	97.05	11.29	62.41	91.44	35.58	70.77	93.08	9.33	CNN
ALCNet [8]	81.40	96.51	9.26	62.05	90.58	21.78	74.31	73.12	20.21	CNN
FC3-Net [38]	78.56	93.86	23.922	65.07	91.54	15.55	72.44	98.14	10.85	CNN
ISNet [39]	81.77	96.3	44.47	68.77	95.56	15.39	80.02	99.02	4.61	CNN
DNA-Net [19]	88.19	98.62	9.00	69.38	93.3	11.66	79.26	98.48	5.35	CNN
UIU-Net [33]	92.19	97.77	15.44	69.96	93.98	22.07	78.25	97.45	4.29	CNN
Dim2Clear[40]	81.37	96.23	9.17	66.34	93.75	20.93	77.29	99.10	6.72	CNN
GCI-Net [41]	92.43	98.25	8.96	67.75	93.89	12.84	78.81	99.34	4.11	CNN
IAANet [31]	90.22	97.26	8.32	66.25	93.15	14.20	74.22	93.53	22.70	Hybrid
SCTransNet [35]	94.09	96.95	4.29	68.03	93.27	10.74	77.50	96.95	13.92	Hybrid
SAM [16]	84.20	97.38	13.28	68.12	93.64	8.38	75.21	97.07	9.78	SAM
SAM-HQ [15]	88.02	98.31	14.48	68.85	93.54	9.56	75.27	97.22	6.87	SAM
SAIST (Ours)	95.23	99.28	1.31	72.14	96.18	4.76	80.82	99.56	0.87	CLIP + SAM

Table 1. Comparison with existing IRSTD approaches on the NUDT-SIRST, IRSTD-1k and NUAA-SIRST datasets. The evaluation metrics are IoU (10^{-2}), P_d (10^{-2}) and F_a (10^{-6}), the best results are highlighted.

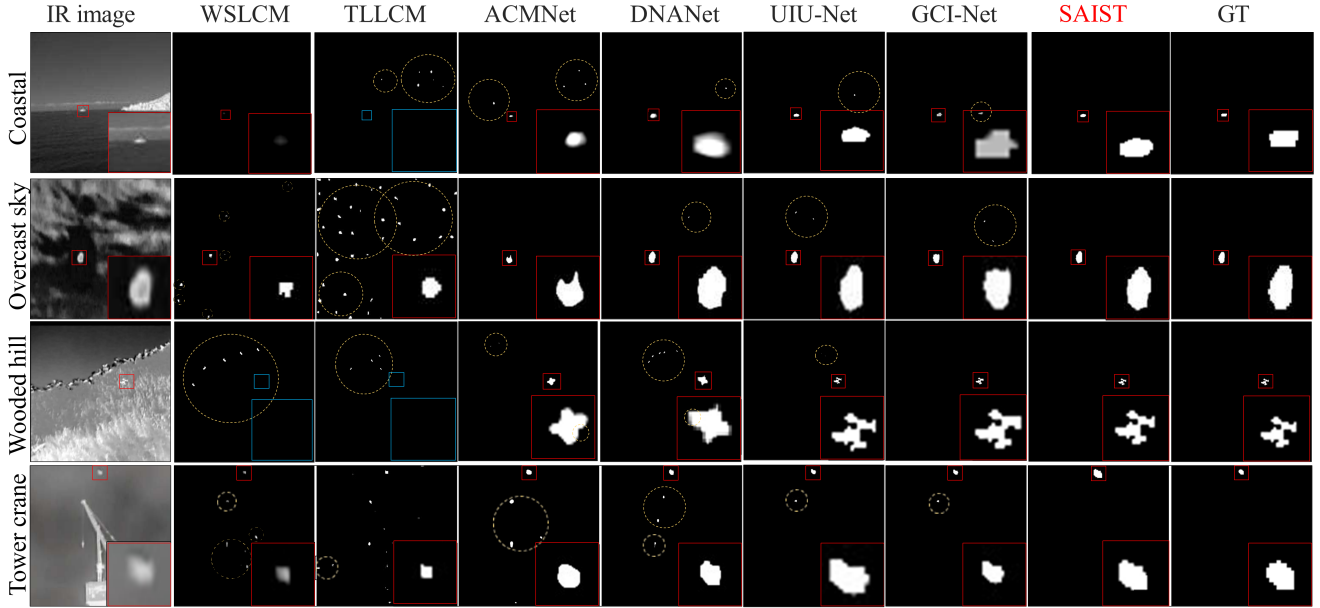


Figure 4. Visual results of different IRSTD methods. The boxes in red, yellow, and blue represent correct, missed, and false detections, respectively. Close-up views are shown in the corners.

Tab. 1, removing CG-SAM significantly reduces the IoU and P_d scores while increasing the F_a score. Additionally, as illustrated in the Fig. 7, compared to other methods, CG-SAM can precisely identify small targets in complex backgrounds. These results validate the effectiveness of CG-SAM in enhancing small target representation and suppressing background interference. Similarly, removing SR-CLIP increases the F_a score. Meanwhile, Fig. 6 demonstrates that SR-CLIP effectively suppresses background interference and noise. These results suggest that SR-CLIP gen-

erates accurate visual-textual prompts for CG-SAM. Ultimately, as shown in Table. 3, although SAIST does not have an advantage in inference time and the number of trainable parameters, it achieves the best performance across all evaluation metrics, demonstrating the excellence of its design.

4.3.1. Impact of the SR-CLIP

In this ablation study, we provide identical text prompts to CLIP, CoOp, and SR-CLIP, utilizing their respective vision and text encoders for image understanding and caption gen-

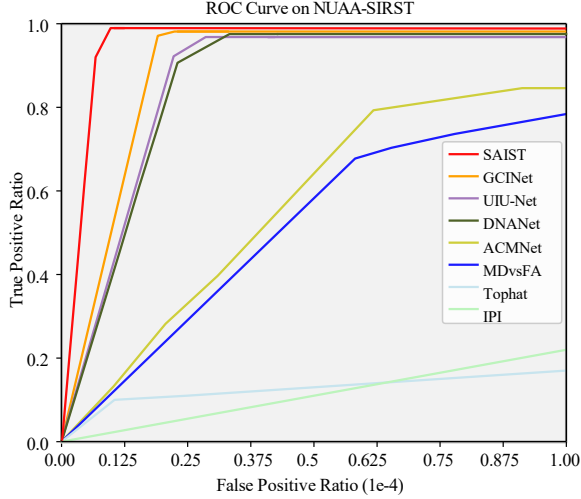


Figure 5. ROC curves on NUAASIRST database.

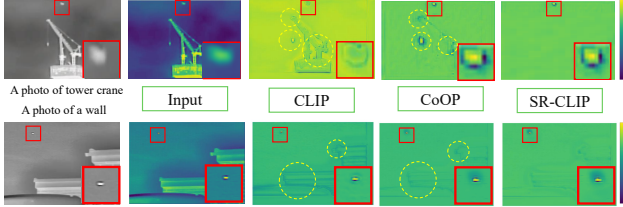


Figure 6. Visual comparison of CLIP, CoOp and SR-CLIP.

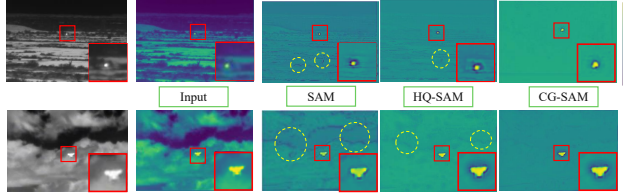


Figure 7. Visual comparison of SAM, HQ-SAM and CG-SAM.

Table 2. Ablation study of SR-CLIP and CG-SAM in $IoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$ on NUAASIRST.

Method	IoU	Pd	Fa
SAIST	80.82	99.56	0.87
w/o SR-CLIP	80.02	98.77	6.61
w/o CG-SAM	78.87	97.42	8.53
w/o CG-SAM & SR-CLIP	76.82	96.50	12.37

eration, to systematically evaluate the performance of SR-CLIP. As shown in Fig. 6 and the Tab. 4, we demonstrate that SR-CLIP effectively suppresses background interference and noise, enabling the generation of precise visual-textual prompts for CG-SAM. These results highlight the efficacy of SR-CLIP in improving the alignment between the visual and textual modalities.

4.3.2. Impact of the CG-SAM

In this ablation study, we utilize the same SAM encoder while employing different Mask Decoders: SAM, HQ-

Table 3. Comparison of different models in $IoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$ on MIRSTD.

Model	Trainable Params (M)	Inference (s)	Avg IoU	Avg Pd	Avg Fa
SAM + CLIP	383.28	0.079	78.49	96.67	8.45
SAIST	389.57	0.081	82.73	98.34	2.31

Table 4. Ablation study of CLIP, CoOp, and SR-CLIP in $IoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$ on NUAASIRST.

Method	IoU	Pd	Fa
CLIP + CG-SAM	80.02	98.77	6.61
CoOp + CG-SAM	80.32	98.56	4.87
SR-CLIP + CG-SAM	80.82	99.56	0.87

Table 5. Ablation study of SAM, HQ-SAM, and CG-SAM in $IoU(\%)$, $P_d(\%)$, $F_a(10^{-6})$ on NUAASIRST.

Method	IoU	Pd	Fa
SAM + SR-CLIP	78.87	97.42	5.23
HQ-SAM + SR-CLIP	79.45	98.89	4.56
CG-SAM + SR-CLIP	80.82	99.56	0.87

SAM, and CG-SAM. Identical fine-grained visual and textual prompts, generated by SR-CLIP, are provided as inputs. The performance of CG-SAM is systematically evaluated. As demonstrated in Fig. 7 and the Tab. 5, our results confirm that CG-SAM is capable of accurately identifying small targets in complex backgrounds. These findings validate that, with the aid of infrared imaging equations, CG-SAM enhances the representation of small targets while effectively suppressing background interference.

5. Conclusion

This paper presents SAIST, a novel framework that integrates visual and textual modalities for enhanced IRSTD. Leveraging SR-CLIP to capture cross-modal relationships, SAIST generates precise prompts to overcome the limitations of image-only methods. Guided by these prompts, CG-SAM employs interpretable infrared imaging equations to extract targets while suppressing noise. We also introduce MIRSTD, the first multimodal IRSTD dataset with rich image-text pairs. Experiments demonstrate the effectiveness of multimodal integration in IRSTD.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272363, Grant 92470108, Grant 62371362, and Grant U21A20514; in part by the Joint Laboratory for Innovation in Satellite-Borne Computers and Electronics Technology Open Fund 2023 under Grant 2024KFKT001-1; in part by the Proof of Concept Foundation of Xidian University Hangzhou of Technology under Grant No. GNYZ2023YL0301; and part by the Shaanxi Provincial Key R&D Program Project under Grant 2024SF-YBXM-330.

References

- [1] 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, 7
- [2] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 4
- [3] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2018. 5
- [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. 2
- [5] G Cuccurullo, L Giordano, D Albanese, Luciano Cinquanta, and M Di Matteo. Infrared thermography assisted control for apples microwave drying. *Journal of food engineering*, 112(4):319–325, 2012. 1
- [6] 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
- [7] 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. 1, 2, 6, 7
- [8] 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*, 59(11):9813–9824, 2021. 6, 7
- [9] 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
- [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. 6, 7
- [11] Tan Guo, Baojiang Zhou, Fulin Luo, Lei Zhang, and Xinbo Gao. Dmfnet: Dual-encoder multistage feature fusion network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 2
- [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
- [13] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6), 2019. 3
- [14] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. 3
- [15] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7
- [16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 6, 7
- [17] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 2
- [18] Wing-Cheung Law, Zhourui Xu, Ken-Tye Yong, Xin Liu, Mark T Swihart, Mukund Seshadri, and Paras N Prasad. Manganese-doped near-infrared emitting nanocrystals for in vivo biomedical imaging. *Optics express*, 24(16):17553–17561, 2016. 1
- [19] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 2022. 2, 6, 7
- [20] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *arXiv preprint arXiv:2401.12665*, 2024. 2
- [21] Fanzhao Lin, Kexin Bao, Yong Li, Dan Zeng, and Shiming Ge. Learning contrast-enhanced shape-biased representations for infrared small target detection. *IEEE Transactions on Image Processing*, 33:3047–3058, 2024. 1, 2
- [22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2
- [23] Alexandre Mallet, Roumiana Tsenkova, Jelena Muncan, Cyrille Charnier, Éric Latrille, Ryad Bendoula, Jean-Philippe Steyer, and Jean-Michel Roger. Relating near-infrared light path-length modifications to the water content of scattering media in near-infrared spectroscopy: toward a new bouguer–beer–lambert law. *Analytical Chemistry*, 93(17):6817–6823, 2021. 5
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [25] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng

- Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2
- [26] 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, 2021. 7
- [27] Danica J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016. 4
- [28] Donald F Swinehart. The beer-lambert law. *Journal of chemical education*, 39(7):333, 1962. 5
- [29] Michael Teutsch and Wolfgang Krüger. Classification of small boats in infrared images for maritime surveillance. In *2010 International WaterSide Security Conference*, pages 1–7. IEEE, 2010. 1
- [30] 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. 1, 2
- [31] Kewei Wang, Shuaiyuan Du, Chengxin Liu, and Zhiguo Cao. Interior attention-aware network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 1, 2, 6, 7
- [32] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2
- [33] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Uiu-net: U-net in u-net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2022. 6, 7
- [34] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3
- [35] Shuai Yuan, Hanlin Qin, Xiang Yan, Naveed Akhtar, and Ajmal Mian. Sctransnet: Spatial-channel cross transformer network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. 1, 6, 7
- [36] Landan Zhang and Zhenming Peng. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sensing*, 11(4):382, 2019. 2
- [37] Mingjin Zhang, Haichen Bai, Jing Zhang, Rui Zhang, Chaoyue Wang, Jie Guo, and Xinbo Gao. Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1730–1738, 2022. 1, 2
- [38] Mingjin Zhang, Ke Yue, Jing Zhang, Yunsong Li, and Xinbo Gao. Exploring feature compensation and cross-level correlation for infrared small target detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1857–1865, 2022. 1, 7
- [39] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. Isnet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 877–886, 2022. 2, 6, 7
- [40] Mingjin Zhang, Rui Zhang, Jing Zhang, Jie Guo, Yunsong Li, and Xinbo Gao. Dim2clear network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 2, 6, 7
- [41] Mingjin Zhang, Ke Yue, Boyang Li, Jie Guo, Yunsong Li, and Xinbo Gao. Single-frame infrared small target detection via gaussian curvature inspired network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. 2, 6, 7
- [42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 4
- [43] Hu Zhu, Haopeng Ni, Shiming Liu, Guoxia Xu, and Lizhen Deng. Tnlrs: Target-aware non-local low-rank modeling with saliency filtering regularization for infrared small target detection. *IEEE Transactions on Image Processing*, 29:9546–9558, 2020. 1
- [44] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 1