Infrared Small Target Detection with Scale and Location Sensitivity

Qiankun Liu[1] Rui Liu[1,2] Bolun Zheng[3] Hongkui Wang[3] Ying Fu[1,2*]
[1]Beijing Institute of Technology [2]Yangtze Delta Region Academy of Beijing Institute of Technology
[3]Hangzhou Dianzi University
{liuqk3, liurui20}@bit.edu.cn, {blzheng, wanghk}@hdu.edu.cn, fuying@bit.edu.cn

Abstract

Recently, infrared small target detection (IRSTD) has been dominated by deep-learning-based methods. However, these methods mainly focus on the design of complex model structures to extract discriminative features, leaving the loss functions for IRSTD under-explored. For example, the widely used Intersection over Union (IoU) and Dice losses lack sensitivity to the scales and locations of targets, limiting the detection performance of detectors. In this paper, we focus on boosting detection performance with a more effective loss but a simpler model structure. Specifically, we first propose a novel Scale and Location Sensitive (SLS) loss to handle the limitations of existing losses: 1) for scale sensitivity, we compute a weight for the IoU loss based on target scales to help the detector distinguish targets with different scales: 2) for location sensitivity, we introduce a penalty term based on the center points of targets to help the detector localize targets more precisely. Then, we design a simple Multi-Scale Head to the plain U-Net (MSHNet). By applying SLS loss to each scale of the predictions, our MSHNet outperforms existing state-of-the-art methods by a large margin. In addition, the detection performance of existing detectors can be further improved when trained with our SLS loss, demonstrating the effectiveness and generalization of our SLS loss. The code is available at https://github.com/ying-fu/MSHNet.

1. Introduction

Infrared small target detection (IRSTD) is an important computer vision task, which has a wide range of applications, such as maritime surveillance [20, 34], traffic management [26, 27, 30] and so on. However, due to the long-distance camera capture, and the noise and clutter interference, infrared targets are often small and dim, making it difficult to detect infrared small targets effectively.

To enhance the detection performance of infrared small targets, numerous methods have been proposed. Early traditional methods can be divided into filtering-based methods [5, 17], local-contrast-based methods [9, 10], and low-rank-based methods [2, 28, 29, 31]. However, these traditional methods rely on manually designed features, and thus can not generalize well to environment changes. Recently, with the development of Deep Learning (DL), IRSTD has been dominated by DL-based methods [1, 15, 21, 24]. Different from traditional methods, DL-based methods can automatically learn useful features through a gradient descent algorithm with the constraints of loss functions, making them more robust to various scenarios.

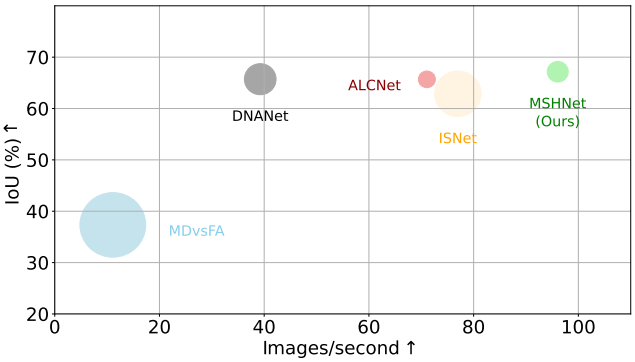

Figure 1. Visualization of the detection performance (IoU), inference time consumption (Images/second) as well as the number of floating point of operations (area of circles) of some deep-learning-based methods. It can be seen that our MSHNet achieves a better balance between these three metrics than other methods. Results are evaluated on IRSTD-1k [30].

However, existing DL-based methods primarily focus on designing complex model structures for feature extraction, leaving the loss functions for IRSTD under-explored. For instance, Li *et.al.* [14] customize a dense nested interactive module to achieve multi-layer feature fusion, and Wu *et.al.* [25] nest the U-Net structure to achieve feature aggregation. Although discriminative features can be extracted by the complex model structures, the detection performance is still limited by the under-explored loss functions. For ex-

*Corresponding author

ample, the widely used Intersection over Union (IoU) loss and Dice loss [18] lack sensitivity to the scales and locations of targets. As shown in Fig. 2, targets with different scales (top row) and locations (bottom row) may share the same IoU loss or Dice loss. This insensitivity to scales and locations makes it challenging for detectors to distinguish targets of different scales and locations, which ultimately limits the detection performance.

In this paper, we focus on boosting detection performance with a more effective loss function but a simpler model structure. Specifically, we first propose a novel Scale and Location Sensitive (SLS) loss to handle the limitations of existing losses. The merits of the proposed SLS loss include: (1) Scale sensitivity. We compute a weight for the IoU loss based on the predicted and ground-truth scales of targets. The larger the gap between predicted and ground-truth scales is, the more attention will be paid by the detector. (2) Location Sensitivity. We design a location penalty based on the predicted and ground-truth center points of targets. Compared with traditional L1 and L2 distances, the designed location penalty produces the same value for fewer different location errors, making the detector locate targets more precisely. Then, we introduce a simple Multi-Scale Head to the plain U-Net (MSHNet), which produces multi-scale predictions for each input. Through leveraging SLS loss at different scales, our MSHNet outperforms existing state-of-the-art (SOTA) methods by a large margin. With the absence of complex structures, our detector achieves a better balance between detection performance, floating point of operations (FLOPs) and inference time consumption, as shown in Fig. 1. Moreover, we further train different existing detectors with our SLS loss and achieve better detection performance, demonstrating the effectiveness and generalization of our SLS loss.

In summary, our main contributions are:

- We propose a novel scale and location sensitive loss for infrared small target detection, which helps detectors distinguish objects with different scales and locations.
- We propose a simple but effective detector by introducing a multi-scale head to the plain U-Net, which achieves SOTA performance without bells and whistles.
- We apply our loss to existing detectors and show that the detection performance can be further boosted, demonstrating the effectiveness and generalization of our loss.

2. Related Work

In this section, we first make a brief introduction to existing methods for IRSTD. Then we provide a review of related works from the perspectives of loss functions and model structures for IRSTD.

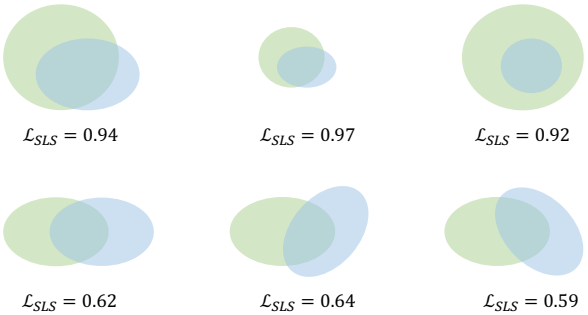

Figure 2. Top row: our SLS loss for the targets of different scales, where IoU loss (=0.4) and Dice loss (=0.57) have the same values for different cases. Bottom row: our SLS loss for the targets of different locations, where IoU loss (=0.3) and Dice loss (=0.43) have the same values for different cases.

2.1. Infrared Small Target Detection

Existing IRSTD methods can be roughly divided into traditional methods and Deep-Learning-based (DL-based) methods. Among them, traditional methods depend on the hand-crafted priors and can be further divided into filtering-based methods [5, 17], local-contrast-based methods [9, 10, 12], and low-rank-based methods [2, 6, 8, 19, 28, 29, 32]. To get useful features with such hand-crafted priors, lots of hyper-parameters need to be manually fine-tuned, making them less robust to the interference of noises and clutters. Differently, DL-based methods [3, 7, 14, 15, 21, 25, 30, 33] can automatically learn useful features with the help of loss functions and gradient descent algorithms. Though impressive detection performance has been achieved, existing DL-based methods mainly focus on the design of model structures for the pursuit of more effective features, leaving the loss functions for IRSTD under-explored.

Different from existing DL-based methods, we focus on boosting detection performance with a more effective loss function but a simpler model structure. A better balance between detection performance, FLOPs and inference time consumption can be achieved.

2.2. Loss Functions for IRSTD

As one of the components in DL-based methods, the loss function plays a critical role in the learning process of deep models by quantifying the disparity between predictions and ground-truths. The commonly adopted IoU loss and Dice loss [18] suffer from insensitivity to the scales and locations of targets, rendering the detectors for distinguishing the targets of different scales and locations accurately.

In order to achieve better detection performance, researchers have developed several loss functions. For example, the loss for adversarial training [21], edge loss for the detecting of target edges [30] and the likelihood loss between target and background maps [11]. However, these

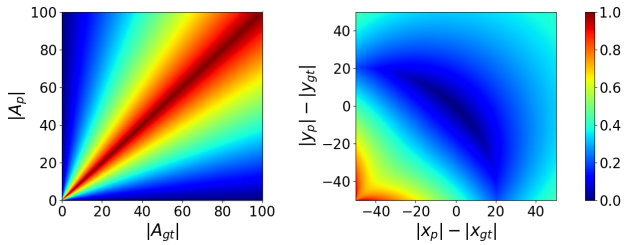

Figure 3. Left: the value of weight w in scale sensitive loss with respect to the number of predicted pixels and ground-truth pixels (*i.e.*, predicted and ground-truth scales). Right: the normalized location sensitive loss with respect to the location error between the predicted center point and ground-truth center point. The range of [0, 100] pixels is shown for illustration.

losses are tailored for specific network architectures, limiting their broader utility. Different from these dedicated losses, the generalized IoU (GIoU) [16] and complete IoU (CIoU) [35] losses have been adopted for box-level IRSTD [13, 23]. However, these IoU variant losses still lack the sensitivity of scales and locations.

In contrast to these losses, we formulate a general loss function that is better suitable for IRSTD. It can distinguish targets of different scales and locations, enabling different detectors to achieve better detection performance.

2.3. Model Structures for IRSTD

The deep model is another key component in DL-based methods. Liu *et al.* [15] pioneer the application of deep learning to IRSTD. The adopted five-layer Multi-Layer Perceptron (MLP) network demonstrates the superiority of DL-based methods in IRSTD. Recently, lots of works have focused on the design of model structures to get more effective features. Li *et al.* [14] customize a dense nested interactive module to achieve multi-layer feature fusion. Zhang *et al.* [30] utilize Taylor finite difference and orientation attention strategy to extract edge information of targets. However, such complex model structures not only bring more computational cost but still suffer from moderate detection performance due to the lack of effective loss functions.

Differently we introduce a simple multi-scale head to the plain U-Net, rather than designing complex structures. By applying our SLS loss to different scales, SOTA performance is achieved with less time consumption.

3. Scale and Location Sensitive Loss

The scale and location sensitive (SLS) loss, denoted as $\mathcal{L}_{SLS}$, is designed to handle the insensitivity of scales and locations in existing losses. It consists of a scale sensitive loss and a location sensitive loss. Formally,

$$\mathcal{L}_{SLS} = \mathcal{L}_S + \mathcal{L}_L, \tag{1}$$

where $\mathcal{L}_S$ and $\mathcal{L}_L$ present the scale sensitive loss and location sensitive loss, respectively. In the following, we introduce the SLS loss in detail starting from the scale sensitive loss, which is based on the commonly used IoU loss.

3.1. Scale Sensitive Loss

Let A_p and A_{gt} be the set of predicted pixels and ground-truth pixels of targets, the IoU loss between them can be formulated as:

$$\mathcal{L}_{IoU} = 1 - \frac{|A_p \cap A_{gt}|}{|A_p \cup A_{gt}|}. \tag{2}$$

Though having been widely used in IRSTD, the IoU loss is insensitive to the scales and locations of targets, as shown in Fig. 2. The scale sensitive loss is implemented by providing a weight to the IoU loss:

$$\begin{aligned} \mathcal{L}_S &= 1 - w\frac{|A_p \cap A_{gt}|}{|A_p \cup A_{gt}|}, \\ \text{s.t.} \quad w &= \frac{\min(|A_p|, |A_{gt}|) + \mathrm{Var}(|A_p|, |A_{gt}|)}{\max(|A_p|, |A_{gt}|) + \mathrm{Var}(|A_p|, |A_{gt}|)}, \end{aligned} \tag{3}$$

where $\mathrm{Var}(\cdot, \cdot)$ is the function that gets the variance of provided scalars.

On the left of Fig. 3, we visualize the values of w with respect to the number of pixels in A_p and A_{gt}. It can be observed that the larger the gap between $|A_p|$ and $|A_{gt}|$ is, the smaller the w is, which results in a larger scale sensitive loss (on the assumption that the IoU between A_p and A_{gt} is fixed). The intuition behind the design of w is that the detector should pay more attention to the target with a larger loss if the predicted and ground-truth scales (*i.e.*, the numbers of pixels in A_p and A_{gt}) are quite different.

3.2. Location Sensitive Loss

The location sensitive loss is calculated based on the predicted and ground-truth center points of targets. Given the sets of predicted pixels A_p and ground-truth pixels A_{gt}, the corresponding center points for A_p and A_{gt} are obtained by averaging the coordinates of all pixels, which are denoted as $\mathbf{c}_p = (x_p, y_p)$ and $\mathbf{c}_{gt} = (x_{gt}, y_{gt})$, respectively. Then, we convert the coordinates of these two center points into the polar coordinate system. Take $\mathbf{c}_p$ for example, the corresponding distance d_p and angle θ_p in the polar coordinate system are:

$$\begin{aligned} d_p &= \sqrt{x_p^2 + y_p^2}, \\ \theta_p &= \arctan(\frac{y_p}{x_p}). \end{aligned} \tag{4}$$

The location sensitive loss can be obtained by:

$$\mathcal{L}_L = (1 - \frac{\min(d_p, d_{gt})}{\max(d_p, d_{gt})}) + \frac{4}{\pi^2}(\theta_p - \theta_{gt})^2, \tag{5}$$

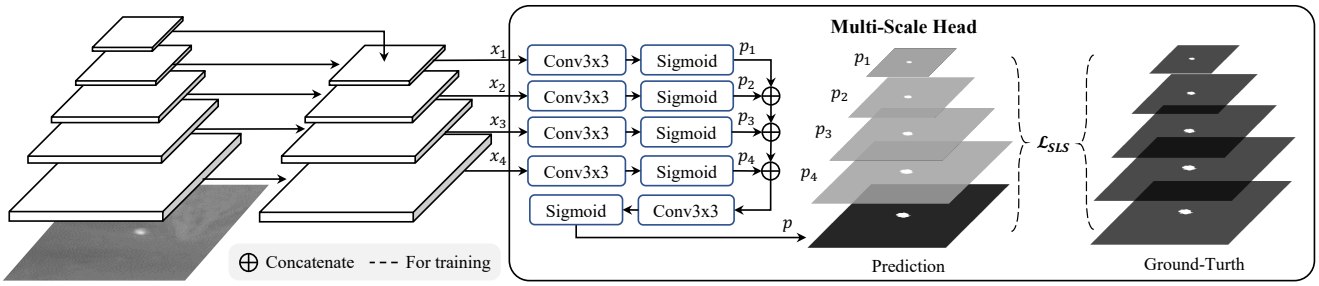

Figure 4. Overview of the proposed MSHNet. Our MSHNet is implemented based on a plain U-Net without bells and whistles. Only a simple multi-scale head is introduced. For each scale, the feature map is fed into a dedicated head, producing a prediction with the same spatial shape as the feature map. Different scales of predictions are upsampled (if needed) and concatenated together to get the final prediction. In the training stage, our SLS loss is applied to each of these predictions since it is scale sensitive.

where d_{gt} and θ_{gt} are the distance and angle of $\mathbf{c}_{gt}$ in the polar coordinate system, respectively.

On the right of Fig. 3, we show how the location loss changes with respect to different location errors between $\mathbf{c}_p$ and $\mathbf{c}_{gt}$. As we can see, though some different location errors share the same loss value (these location errors still can be distinguished with their gradients), the location loss distinguishes most of the different location errors effectively, making the detector sensitive to different types of location errors and locate the targets more accurately.

4. MSHNet Dectector

In this section, we introduce our MSHNet detector, which is implemented by introducing a simple but effective multi-scale head to the plain U-Net. The overview of MSHNet is shown in Fig. 4. We take the commonly used U-Net as the backbone network. The feature maps that have different scales in the decoder are fed into different prediction heads to get different scales of predictions. All the predictions from different feature maps are finally concatenated (upsampling before concatenation is adopted if needed) to get the final prediction. In the training stage, our SLS loss is applied to each of the predictions. In the following, the multi-scale head is first described in detail. Then we introduce the utilization of SLS loss in MSHNet.

4.1. Multi-Scale Head

Let $\mathbf{x}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ be the feature map at the i-th scale in the decoder of U-Net, where $H_i \times W_i$ is the spatial size and C_i is the number of channels. Following the common settings within existing works, there are 4 scales in U-Net, which means that $i \in \{1, 2, 3, 4\}$. Supposing the input spatial size is $H \times W$, then $H_i = \frac{H}{2^{4-i}}$ and $W_i = \frac{W}{2^{4-i}}$.

The i-th prediction $\mathbf{p}_i \in \mathbb{R}^{H_i \times W_i \times 1}$ is obtained by the corresponding prediction head, which is implemented by a convolution layer and a sigmoid activation function:

$$\mathbf{p}_i = \text{Sigmoid}(\text{Conv}(\mathbf{x}_i)). \tag{6}$$

Note that different prediction heads have their own dedicated parameters. The final prediction $\mathbf{p} \in \mathbb{R}^{H \times W \times 1}$ is obtained based on all the 4 predictions:

$$\mathbf{p} = \text{Sigmoid}(\text{Conv}([\Uparrow(\mathbf{p}_1, 8), \Uparrow(\mathbf{p}_2, 4), \Uparrow(\mathbf{p}_3, 2), \mathbf{p}_4])), \tag{7}$$

where $\Uparrow(\cdot, \cdot)$ is the operation that spatially upsamples the first argument with the second argument as the factor, and $[\cdot, ..., \cdot]$ concatenates all the provided arguments along the channel dimensionality.

4.2. Training MSHNet with SLS Loss

Since our SLS loss is scale sensitive and there are several scales in the predictions of MSHNet, we apply SLS loss to all of the predictions. The inspiration is that our SLS loss produces different loss values for different scales even if they share the same spatial layout (refer to the first two cases in the top row, Fig. 2). We hypothesize that by applying our SLS loss to different scales, the targets that are with different scales can attract different attention from the detector, resulting in an overall better detection performance.

Let $\mathbf{p}_{gt} \in \{0, 1\}^{H \times W \times 1}$ be the ground-truth label. The final loss for MSHNet is:

$$\mathcal{L} = \frac{1}{5}\left(\sum_{i=1}^{4} \mathcal{L}_{SLS}(\mathbf{p}_i, \Downarrow(\mathbf{p}_{gt}, 2^{4-i})) + \mathcal{L}_{SLS}(\mathbf{p}, \mathbf{p}_{gt})\right), \tag{8}$$

where $\Downarrow(\cdot, \cdot)$ is the operation (*i.e.*, max-pooling) that spatially downsamples the first argument with the second argument as the factor, and $\mathcal{L}_{SLS}(\cdot, \cdot)$ is our SLS loss.

5. Experiments

In this section, the adopted datasets and metrics are firstly introduced, followed by the implementation details. Then, we compare the proposed method MSHNet with existing IRSTD detectors. Finally, discussions are provided to show the effectiveness of our Scale and Location Sensitive loss and MHSNet detector.

Method	Description	IRSTD-1k			NUDT-SIRST		
		IoU↑	P_d ↑	F_a ↓	IoU↑	P_d ↑	F_a ↓
Top-Hat [17]	Filtering	10.06	75.11	1432	20.72	78.41	166.7
Max-Median [5]		6.998	65.21	59.73	4.197	58.41	36.89
WSLCM [10]	Local Contrast	3.452	72.44	6619	2.283	56.82	1309
TLLCM [9]		3.311	77.39	6738	2.176	62.01	1608
IPI [8]	Low Rank	27.92	81.37	16.18	17.76	74.49	41.23
NRAM [29]		15.25	70.68	16.93	6.927	56.40	19.27
RIPT [2]		14.11	77.55	28.31	29.44	91.85	344.3
PSTNN [28]		24.57	71.99	35.26	14.85	66.13	44.17
MSLSTIPT [19]		11.43	79.03	1524	8.342	47.40	888.1
MDvsFA [21]	Deep Learning	37.34	83.71	88.52	35.86	85.22	95.37
ALCNet [4]		65.68	89.25	27.71	72.89	96.19	30.40
ISNet [30]		62.88	92.59	27.92	67.86	92.59	34.65
DNANet [14]		65.71	91.84	17.61	79.98	96.93	12.78
MSHNet (Ours)		**67.16**	**93.88**	**15.03**	**80.55**	**97.99**	**11.77**

Table 1. Quantitative results of different methods. Results for the metrics of IoU(%), P_d(%) and $F_a(10^{-6})$ are presented. The best values are highlighted with **bold**. It can be seen that our MSHNet achieves the best results on different metrics and datasets.

5.1. Datasets and Metrics

Datasets. The experiments are conducted on two datasets *i.e.*, IRSD-1k [30] and NUDT-SIRST [14]. There are 1,001 and 1,327 infrared images in IRSTD-1k and NUDT-SIRST, respectively. Following existing works [14, 30], the images in IRSTD-1k are divided into the training and testing splits with a ratio of 4:1, while the images in NUDT-SIRST are equally divided into the training and testing splits.

Evaluation Metrics. We use IoU and false alarm rate (F_a) as the pixel-level evaluation metric and assess target-level performance using the probability of detection (P_d). Different metrics reveal the performance of the detector from different aspects. F_a and P_d focus on recall and false alarms, while IoU takes both into consideration.

Formally, The false alarm rate is:

$$F_a = \frac{P_{false}}{P_{all}}, \tag{9}$$

where P_{false} is the number of false positive pixels and P_{all} is the number of all pixels in the image.

The probability of detection is:

$$P_d = \frac{N_{pred}}{N_{all}}, \tag{10}$$

where N_{pred} is the number of correctly predicted targets and N_{all} is the number of all targets.

5.2. Implementation Details

The proposed method is implemented with PyTorch framework. Following existing works, the input size of the detector is set to 256×256. We train different models using AdaGrad on 2 RTX3090 GPUs. The batch size is set to 4 and the learning rate is set to 0.05.

5.3. Comparison to Existing Methods

We first compare the proposed MSHNet with existing methods. Different types of methods are evaluated, including traditional methods and deep-learning-based methods. The evaluated traditional methods are the filtering-based Top-Hat [17] and Max-Median [5], local-contrast-based WSLCM [10] and TLLCM [9], low-rank-based IPI [8], NRAM [29], RIPT [2], PSTNN [28] and MSLSTIPT [19]. The deep-learning-based methods are MDvsFA [21], ALCNet [4], ISNet [30], and DNANet [14]. For a fair comparison, all the evaluated deep-learning-based models are retrained with their official codes to their convergence on NUDT-SIRST [14] and IRSTD-1k [30] datasets.

Quantitative Results. The quantitative results of different methods are presented in Tab. 1. Overall, our MSHNet performs the best on all metrics and all datasets.

As expected, traditional methods perform poorly on these challenging datasets due to the limitations of manually designed priors for feature extraction. In contrast, DL-based methods can automatically learn useful features and yield better results than traditional methods. Nevertheless, existing DL-based methods lack the consideration of different scales and locations of targets in the training stage. The predictions of existing DL-based methods suffer from the incomplete shape and missed detection of valid pixels, *e.g.*, lower IoU and P_d. Compared with DNANet, which is one of the existing SOTA DL-based methods, our MSHNet achieves 1.45% and 0.57% higher IoU, 2.04% and 1.06% higher P_d on IRSTD-1k and NUAA-SIRST datasets, re-

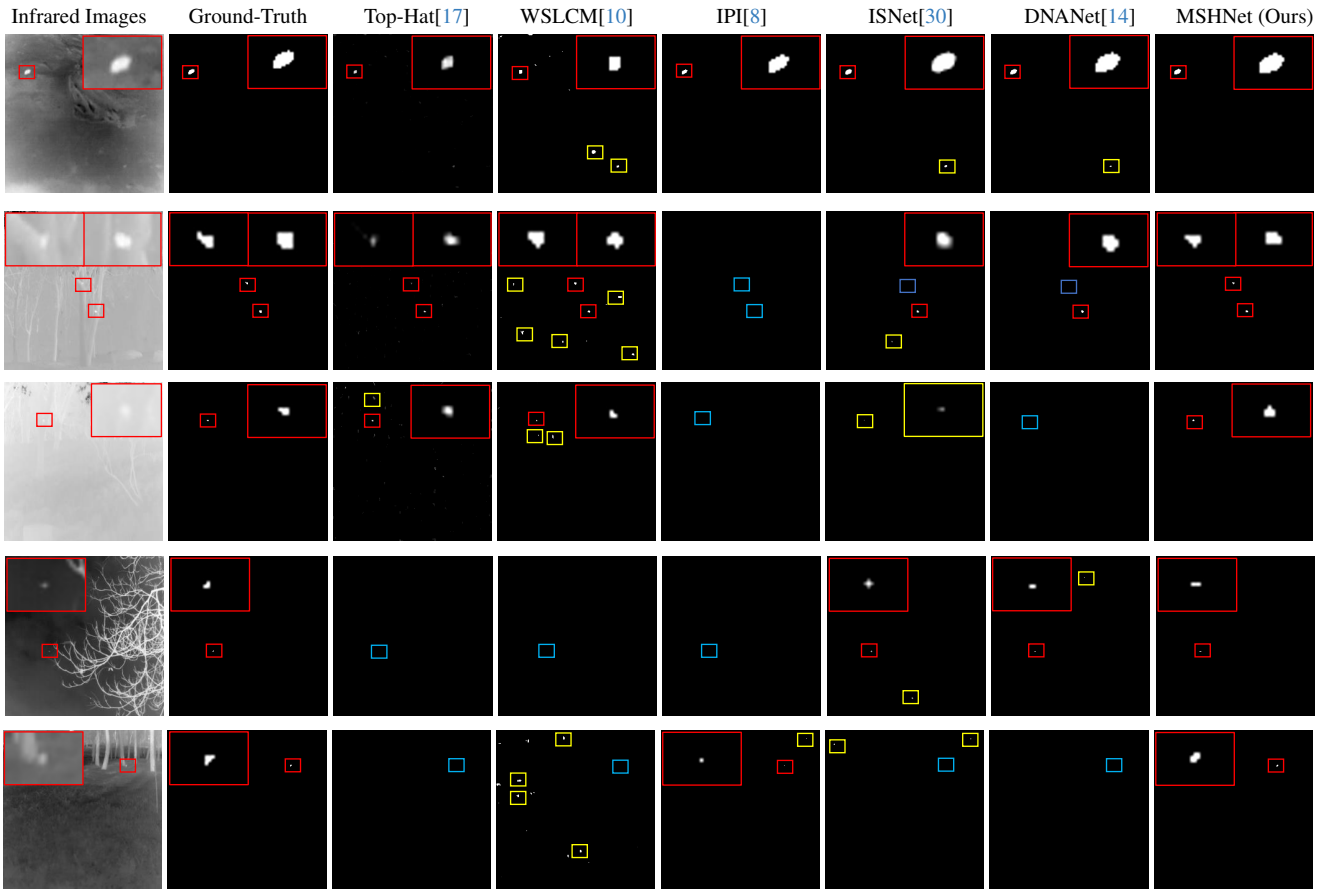

Figure 5. Visual comparison of detection results on several infrared images. Correctly detected targets, missed targets, and false alarms are framed by red, blue, and yellow boxes, respectively. A close-up view of the target is shown in image corners.

spectively. In addition, the false alarm (F_a) on both datasets are all reduced by a large margin, demonstrating the superiority of our method.

As mentioned before, there are no complex model structures in MSHNet, which achieves a better balance between detection performance, floating point of operations (FLOPs) and inference time consumption. The results of DL-based methods are presented in Fig. 1. It can be observed that MSHNet has the best detection performance and the least inference time consumption. It also has the least number of FLOPs except for ALCNet. We own the less inference time consumption of MSHNet to the smaller number of FLOPs and the simple architectures (complex structures in deep models bring more inference time consumption [22]). The better detection performance demonstrates the effectiveness of the proposed multi-scale head and the scale and location sensitive loss.

Qualitative Results. For different kinds of methods, we visualize the detection results of the representative methods in Fig. 5 and Fig. 6. It can be observed that our method has better capability for the detection of challenging small targets. Traditional methods are susceptible to noise interference, resulting in a large number of false alarms. Moreover, due to the small scale and low contrast of the target, traditional methods and other DL-based methods struggle to effectively extract feature information of small targets from challenging scenarios, resulting in a considerable number of missed targets. In contrast, our methods can accurately distinguish small targets from these low-contrast infrared images. Take the case in the 5-th row in Fig. 5 for example, except for MSHNet and WLSLCM, other methods fail to detect the target. Moreover, from the peaks and the area under the peaks in Fig. 6, it can be observed that our method predicts targets with higher confidence and obtains a closer number of peaks with ground-truth. We attribute the superiority of MSHNet to the scale and location sensitivity brought by our SLS loss and multi-scale head.

5.4. Discussions

We provid some discussions to show the effectiveness of our method. Experiments are conducted on IRSTD-1k dataset.

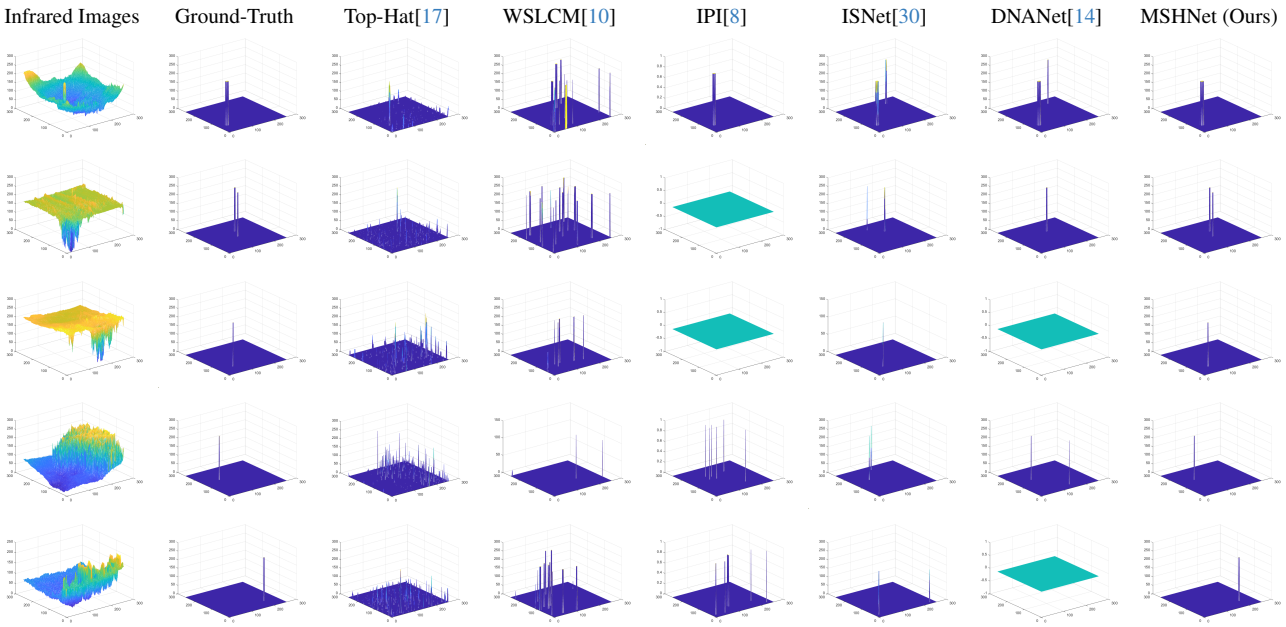

Figure 6. 3D visualization results corresponding to those in Fig. 5. Different colors represent distinct grayscale values, ranging from blue to yellow, with values gradually increasing. Please zoom in for better viewing.

Detection Results for Different Scales of Targets. Since our SLS loss and MSHNet both are scale sensitive, we show how the detection performance changes with respect to different scales of targets. To do this, we split the targets into three different scale levels according to the number of pixels: (0,10], (10, 40] and (40, ∞]. Results are shown in Tab. 2. Due to the limited space, only several DL-based methods are evaluated (specifically, ISNet, DNANet and our MSHNet). Compared with DNANet, MSHNet performs better for all scale levels of targets. Compared with ISNet, MSHNet achieves comparable performance with ISNet for (0,10] and (10,40] scales but much better performance for (40,∞] scale. The conclusion is that MSHNet can pay attention to different scale levels of targets, resulting in an overall better performance.

Scale	Method	IoU↑	P_d ↑	F_a ↓
(0,10]	DNANet [3]	47.26	91.27	22.53
	ISNet [14]	**50.94**	94.53	30.91
	MSHNet (Ours)	49.51	**95.24**	**20.68**
(10,40]	DNANet [3]	63.14	91.85	13.25
	ISNet [14]	**65.41**	**91.91**	16.60
	MSHNet (Ours)	64.79	91.85	**8.03**
(40,∞]	DNANet [3]	78.46	93.94	**6.78**
	ISNet [14]	66.45	87.87	58.80
	MSHNet (Ours)	**79.20**	**96.97**	11.02

Table 2. Comparisons of different methods in terms of different scales of targets. Results for the metrics of IoU(%), P_d(%) and $F_a(10^{-6})$ are presented. Our MSHNet achieves overall the best performance.

Impact of Scale and Location Sensitive Loss. We compare our SLS loss with the commonly used IoU loss and Dice loss. The results of several DL-based methods trained with different losses are presented in Tab. 3. Overall, the detection performance of different detectors is all improved by our SLS loss in terms of IoU metric, demonstrating the effectiveness and generalization of SLS loss. However, it is hard to tell which loss performs the best on the metric of P_d. Interestingly, adopting our SLS loss results in a poorer performance on F_a metric. The reason is that more false alarm pixels may be treated as positive targets since our SLS loss tries to distinguish the targets of all different scales.

Next, we show the effectiveness of scale sensitive loss ($\mathcal{L}_S$) and location sensitive loss ($\mathcal{L}_L$) separately. To do this, we train MSHNet with different losses. The results are presented in Tab. 4. Note that the difference between $\mathcal{L}_S$ and $\mathcal{L}_{IoU}$ lies in the newly introduced weight w (refer to Sec. 3.1). By comparing the results in the first two rows, we can find that the introduction of w (refer to Eq. (3)) is positive to the metrics of IoU and F_a, indicating that the detector produces more accurate shapes for targets as well as less false alarms. While comparing the results in the last two rows, it can be observed that more false alarms will be produced when the location sensitive loss is introduced. However, both IoU and P_d are greatly improved.

Finally, we make a study on different types of location sensitive loss. As mentioned in Sec. 3.2, our location sen-

Loss	DNANet [3]			ISNet [14]			MSHNet (Ours)		
	IoU↑	P_d ↑	F_a ↓	IoU↑	P_d ↑	F_a ↓	IoU↑	P_d ↑	F_a ↓
$\mathcal{L}_{IoU}$	65.71	91.84	17.61	62.88	**92.59**	27.92	64.83	91.16	**5.28**
$\mathcal{L}_{Dice}$	65.58	**92.17**	**12.75**	62.94	90.91	19.26	65.36	92.18	14.12
$\mathcal{L}_{SLS}$ (Ours)	**67.09**	92.15	35.15	**64.42**	92.26	**15.14**	**67.16**	**93.88**	15.03

Table 3. Comparisons of different losses and detectors. Results for the metrics of IoU(%), P_d(%) and $F_a(10^{-6})$ are presented. The main IoU metrics of different detectors are consistently boosted by our scale and location sensitive loss.

Loss	IoU↑	P_d ↑	F_a ↓
$\mathcal{L}_S$ w/o. w	64.83	91.16	5.28
$\mathcal{L}_S$	65.82	89.46	**4.06**
$\mathcal{L}_S + \mathcal{L}_L$	**67.16**	**93.88**	15.03

Table 4. Ablation study of the scale sensitive loss ($\mathcal{L}_S$) and location sensitive loss ($\mathcal{L}_L$). Results are obtained by MSHNet. The metrics of IoU(%), P_d(%) and F_a (10^{-6}) are adopted. Note that $\mathcal{L}_S$ without w is equivalent to $\mathcal{L}_{IoU}$.

Loss	IoU↑	P_d ↑	F_a ↓
$\mathcal{L}_S + \mathcal{L}_{L_2}$	64.89	92.86	7.29
$\mathcal{L}_S + \mathcal{L}_{L_1}$	64.75	93.54	**6.18**
$\mathcal{L}_S + \mathcal{L}_L$	**67.16**	**93.88**	15.03

Table 5. Ablation study of different types of location sensitive loss. Results are obtained by MSHNet. The metrics of IoU(%), P_d(%) and F_a (10^{-6}) are adopted.

sitive loss can distinguish different location errors more effectively . To show the effectiveness of such distinguishability, we use the L_2 and L_1 distances between $\mathbf{c}_p$ and $\mathbf{c}_{gt}$ as the location sensitive loss (denoted as $\mathcal{L}_{L_2}$ and $\mathcal{L}_{L_1}$, respectively). Results are shown in Tab. 5. As we can see, the distinguishability of different location errors is effective in getting better IoU and P_d metrics. However, more false alarms are produced (higher F_a). Such phenomenon is also observed from Tab. 4. The reason may be that the distinguishability of different location errors makes the detector more sensitive to small targets, resulting in the potential of treating some noises to targets.

Impact of Multi-Scale Heads. Now we ablate the number of scales in our multi-scale head. Results are shown in Tab. 6. As we can see, different predictions achieve different detection performance. Overall, the more scales adopted, the better the detection performance is achieved. For example, the IoU is improved from 63.10% to 67.16% when the number of scales is increased from 1 to 4. We use 4 scales by default.

6. Conclusion

In this paper, we focus on boosting the performance of infrared small target detection with an effective loss function

Num. of Scales	Supervised Predictions	IoU↑	P_d ↑	F_a ↓
1	p	63.10	86.73	19.21
1	p_4, p	64.69	93.87	36.74
2	p_3, p_4, p	64.35	**94.89**	35.38
3	p_2, p_3, p_4, p	65.90	93.54	24.37
4	p_1, p_2, p_3, p_4, p	**67.16**	93.88	**15.03**

Table 6. Ablation study of the number of scales in MSHNet. While reducing the number of adopted scales, the remaining smallest scale is removed. Results for the metrics of IoU(%), P_d(%) and F_a (10^{-6}) are presented.

but a simpler model structure. To do this, a Scale and Location Sensitive (SLS) loss is first proposed. The merits of SLS loss include: (1) it pays more attention to the targets that have large gaps between the predicted and ground-truth scales; (2) it distinguishes different types of location errors between the predicted and ground-truth center points of targets. Then a simple Multi-Scale Head is introduced to the plain U-Net (MSHNet), which produces multi-scale predictions for each input. Through applying SLS loss to different scales of predictions, different scales of targets can attract different attention from the detector, resulting in an overall better detection performance. Experimental results show that MSHNet achieves SOTA detection performance with the better balance of inference time and the number of floating point of operations. While applying our SLS loss to other existing detectors, the overall detection performance can be boosted. However, more false alarms may be introduced. Through the analysis in Sec. 5.4, we find that such a phenomenon is brought by the location sensitive loss in SLS loss which potentially treats some noises as targets. In future works, we try to handle this by designing more suitable location sensitive loss.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62331006,62171038, 62088101, and 62371175), the R&D Program of Beijing Municipal Education Commission (KZ202211417048), the Key R&D Program of Zhejiang (2023C01044), the Fundamental Research Funds for the Provincial Universities of Zhejiang (GK239909299001-013), and the Fundamental Research Funds for the Central Universities.

References

[1] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 1

[2] 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, 5

[3] 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, 7, 8

[4] 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. 5

[5] 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*, pages 74 – 83. International Society for Optics and Photonics, SPIE, 1999. 1, 2, 5

[6] Ying Fu, Yongrong Zheng, Lin Zhang, Yinqiang Zheng, and Hua Huang. Simultaneous hyperspectral image super-resolution and geometric alignment with a hybrid camera system. *Neurocomputing*, 384:282–294, 2020. 2

[7] Ying Fu, Jian Chen, Tao Zhang, and Yonggang Lin. Residual scale attention network for arbitrary scale image super-resolution. *Neurocomputing*, 427:201–211, 2021. 2

[8] 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. 2, 5, 6, 7

[9] Jinhui Han, Saed Moradi, Iman Faramarzi, Chengyin Liu, Honghui Zhang, and Qian Zhao. A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Geoscience and Remote Sensing Letters*, 17(10):1822–1826, 2019. 1, 2, 5

[10] 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, 5, 6, 7

[11] Qingyu Hou, Zhipeng Wang, Fanjiao Tan, Ye Zhao, Haoliang Zheng, and Wei Zhang. Ristdnet: Robust infrared small target detection network. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 2

[12] Hua Huang, Guangtao Nie, Yinqiang Zheng, and Ying Fu. Image restoration from patch-based compressed sensing measurement. *Neurocomputing*, 340:145–157, 2019. 2

[13] Moran Ju, Jiangning Luo, Guangqi Liu, and Haibo Luo. Istdet: An efficient end-to-end neural network for infrared small target detection. *Infrared Physics & Technology*, 114: 103659, 2021. 3

[14] 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*, 32:1745–1758, 2022. 1, 2, 3, 5, 6, 7, 8

[15] Ming Liu, Hao-yuan Du, Yue-jin Zhao, Li-quan Dong, Mei Hui, and SX Wang. Image small target detection based on deep learning with snr controlled sample generation. *Current Trends in Computer Science and Mechanical Automation*, 1: 211–220, 2017. 1, 2, 3

[16] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 3

[17] Jean-Francois Rivest and Roger Fortin. Detection of dim targets in digital infrared imagery by morphological image processing. *Optical Engineering*, 35(7):1886 – 1893, 1996. 1, 2, 5, 6, 7

[18] 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*, pages 240–248. Springer, 2017. 2

[19] 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. 2, 5

[20] 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

[21] 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, 5

[22] Jiahao Wang, Songyang Zhang, Yong Liu, Taiqiang Wu, Yujiu Yang, Xihui Liu, Kai Chen, Ping Luo, and Dahua Lin. Riformer: Keep your vision backbone effective but removing token mixer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14443–14452, 2023. 6

[23] 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. 3

[24] Kaixuan Wei, Angelica I Avilés-Rivero, Jingwei Liang, Ying Fu, Hua Huang, and Carola-Bibiane Schönlieb. Tfpnp: Tuning-free plug-and-play proximal algorithms with applications to inverse imaging problems. *J. Mach. Learn. Res.*, 23(16):1–48, 2022. 1

[25] 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. 1, 2

[26] Xinyi Ying, Li Liu, Yingqian Wang, Ruojing Li, Nuo Chen, Zaiping Lin, Weidong Sheng, and Shilin Zhou. Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15528–15538, 2023. 1

[27] Ke Zhang, Shuyan Ni, Dashuang Yan, and Aidi Zhang. Review of dim small target detection algorithms in single-frame infrared images. In *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference*, pages 2115–2120, 2021. 1

[28] 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, 5

[29] Landan Zhang, Lingbing Peng, Tianfang Zhang, Siying Cao, and Zhenming Peng. Infrared small target detection via non-convex rank approximation minimization joint $l_{2,1}$ norm. *Remote Sensing*, 10(11):1821, 2018. 1, 2, 5

[30] 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. 1, 2, 3, 5, 6, 7

[31] Tao Zhang, Ying Fu, and Cheng Li. Hyperspectral image denoising with realistic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2248–2257, 2021. 1

[32] Tao Zhang, Ying Fu, and Cheng Li. Deep spatial adaptive network for real image demosaicing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3326–3334, 2022. 2

[33] Tao Zhang, Ying Fu, and Jun Zhang. Guided hyperspectral image denoising with realistic data. *International Journal of Computer Vision*, 130(11):2885–2901, 2022. 2

[34] Mingjing Zhao, Wei Li, Lu Li, Jin Hu, Pengge Ma, and Ran Tao. Single-frame infrared small-target detection: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):87–119, 2022. 1

[35] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12993–13000, 2020. 3