This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3189225

Detecting Dim Small Target in Infrared Images via Sub-Pixel Sampling Cuneate Network

Xu He, Qiang Ling, Yuyuan Zhang, Zaiping Lin, and Shilin Zhou

***Abstract*—Infrared dim small target detection is regarded as a critical technology for the interpretation of space-based remote sensing images. In recent years, driven by deep learning technology and the surge of data, remarkable effects have been achieved for dim small target detection in infrared images. Nevertheless, the intrinsic feature scarcity and low signal-to-clutter ratio (SCR) characteristics pose tremendous challenges to deep learning-based detection methods. In this letter, we present a novel sub-pixel sampling cuneate network (SPSCNet) to detect dim small targets in infrared images. The overall model architecture is based on an end-to-end cuneate network with multiple groups of parallel high-to-low resolution subnetworks. Specifically, we design a multi-scale feature reweighted fusion (MSFRF) module to effectively fuse multi-scale feature maps which contain both low-level detail features and high-level semantics information. In addition, considering that the pooling operation may lose dim small targets with low SCR, we also exploit a sub-pixel sampling scheme to greatly retain the features of small targets. Moreover, to better test and verify the performance of the proposed method, we also develop an infrared dim small target (IDST) dataset to conduct more comparative experiments. Extensive experiments on the SIRST and IDST datasets illustrate that the proposed SPSCNet yields state-of-the-art performance in comparison with other detection algorithms.**

***Index Terms*—Infrared images, dim small target detection, multi-scale feature fusion, sub-pixel sampling, cuneate network.**

I. Introduction

DETECTING dim small targets in space-based infrared images is a significant computer vision application that is widely applied in various fields including early warning detection, military reconnaissance, and so on [1]. However, owing to the long-range imaging characteristics of space-based early warning satellite detection (SEWSD) systems, the infrared imaging generally presents a comparably low target-to-image pixel ratio (TIPR) and signal-to-clutter ratio (SCR), resulting in scarce texture and shape features of the infrared dim small target. In addition, complex and fast-changing background clutter can also severely degrade the performance of the detection model. Therefore, detecting dim small target in infrared images is still a momentous and challenging task.

Conventional model-based methods model infrared dim small target as an isolated point where the nearby regions are the background of slow transition. For example, background estimation-based models usually make strong prior assumptions about the properties of the background, such as the Top-hat filter [2], average absolute gray difference (AAGD) [3], and absolute directional mean difference (ADMD) [4], which estimates background areas through predetermined background prior assumptions. However, these methods lack generalization for various scenarios and engender many false alarms. Human visual contrast-based models consider that the dim small target has local saliency and split the target with conspicuousness by devising various local contrast filters. For example, local contrast measure (LCM) [5] and multi-scale patch-based contrast measure (MPCM) [6] capture preliminary suspicious targets with strong local contrast by sliding an internal fixed-size window over the whole image to filter out veritable targets. Besides, Zhang et al. [7] designed a local intensity and gradient (LIG) operator to restrain clutter and boost the target saliency. Nonlocal correlation-based models, such as the infrared patch-image (IPI) [8] model and the reweighted infrared patch-tensor (RIPT) [9] model, convert the problem of infrared dim small detection to the low-rank and sparse decomposition (LRSD) task using the nonlocal autocorrelation peculiarity of infrared images. To exploit the inter-frame similarity of the target and the low-rank constraint of the background, two improved LRSD models using total variation (TV) and the partial sum of tensor nuclear norm (PSTNN) were proposed [10], [11]. Although these methods get promising effects in some images, they still easily lose efficacy in real changeable scenarios due to the strong dependence on prior knowledge about infrared images. Furthermore, these methods only utilize handcrafted shallow features, lacking pivotal semantic information used to discriminate real targets from the complex background.

Recently, deep learning-based infrared dim small target detection methods gradually leap to the eyes and attract increasing attention from scholars. To be specific, these approaches cast aside cumbersome expert analysis and boost semantic discriminability of the target with the strong feature learning ability of deep convolutional neural network (DCNN). For example, Wang et al. [12] applied a two-path conditional generative adversarial network (cGAN) to balance and optimize the miss detection (MD) and false alarm (FA) subtasks. Hou et al. [13] combined handcrafted features and DCNN features to establish a novel detection framework. Dai et al. [14] proposed implementing segmentation network framework U-Net [15] to the infrared dim small target detection task and devised an asymmetric contextual module (ACM) to displace the skip connection structure in U-Net. Furthermore, Dai et al. [16] further developed the ACM structure and designed a novel attentional local contrast network (ALCNet), which combined

This work was supported by the National Natural Science Foundation of China under Grant 61605242, Grant 61602499, and Grant 61471371. (*Corresponding author: Qiang Ling.*)

X. He, Q. Ling, Y. Zhang, Z. Lin, and S. Zhou are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail:hexu_710@nudt.edu.cn; lq910131@163.com; zhangyuyuan22@nudt.edu.cn; linzaiping@nudt.edu.cn; slzhou@nudt.edu.cn).

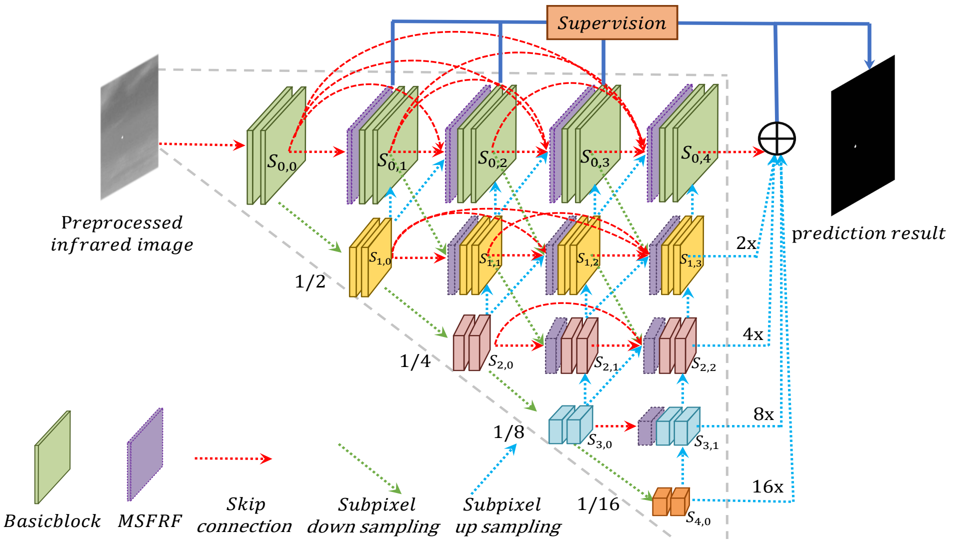


Fig. 1. Overview of our SPSCNet. *MSFRF* represents the multi-scale feature reweighted fusion module. ⊕ represents a combined operation of channel concatenation and 1×1 convolution. *Basicblock* is a frequently-used network structure in ResNet [17]. *Supervision* indicates deep supervision technology [18].

the segmentation network-based model and conventional LCM mechanism. Nevertheless, limited by the paradigm of U-Net, the feature fusion scheme only occurs on the feature maps of the current scale and adjacent high-level semantics layer, lacking adjacent low-level features that contain more detailed information on dim small targets. To this end, we exploit a multi-scale feature reweighted fusion (MSFRF) module via a spatial feature reweighted scheme to fully aggregate features of the current scale and the neighbouring high-level and low-level features, which replenish the semantics and detailed information of the target with effect. Moreover, convolutional pooling operation for attenuating the feature map size, which U-Net relies on, may lose the dim small target due to the intrinsic feature abandonment mechanism. Therefore, we design a novel parameterless subpixel sampling scheme to maximize the retention of features. The overall detection framework is designed as a cuneate network named sub-pixel sampling cuneate network (SPSCNet). In addition, we also develop an infrared dim small target (IDST) dataset to verify the effectiveness of our method.

II. Sub-Pixel Sampling Cuneate Network

A. Overall Cuneate Detection Framework

As illustrated in Fig. 1, the overall structure of SPSCNet is based on a cuneate network. The preprocessed infrared image will be first fed into the cuneate network to gradually extract the multi-scale features. The cuneate network is composed of five groups of parallel large-to-small scale subnetworks, i.e., $\{S_{c,u}, c \in [0 \sim 4], u \in [0 \sim (4-c)]\}$, in which the resolution of feature maps in each subnetwork $\{S_{c,u}\}$ is $\frac{1}{2^c}$ of the input image. Each subnetwork contains an MSFRF module and two Basicblocks except for the first subnetwork $\{S_{c,0}\}$ without the MSFRF module in each group subnetwork. The MSFRF module is designed to effectively enhance multi-scale feature fusion. The skip connection structure in DenseNet [19] is also applied in each group subnetwork to strengthen the transmission of features. To reduce the parameter quantity of model, we replace the 3×3 convolution of Basicblock in ResNet with a depthwise separable convolution [20]. Meanwhile, to better retain the characteristics of an infrared dim small target, we exploit a sub-pixel sampling scheme to substitute for the conventional sampling strategies, e.g., pooing and bilinear interpolation. Finally, at the end of each group of subnetworks, the shallow-layer features maps with abundant spatial detail features and deep-layer feature maps with rich semantic features are upscaled to the same size and fused by a combined operation of concatenation and convolution for more robust feature maps. Furthermore, we apply deep supervision [18] technology to better train our network and adopt the Soft-IoU loss [21] as the training loss function. The Soft-IoU loss in the k-th $(k = 1,2,3,4)$ deep supervision prediction branch of the proposed cuneate network is defined as follows:

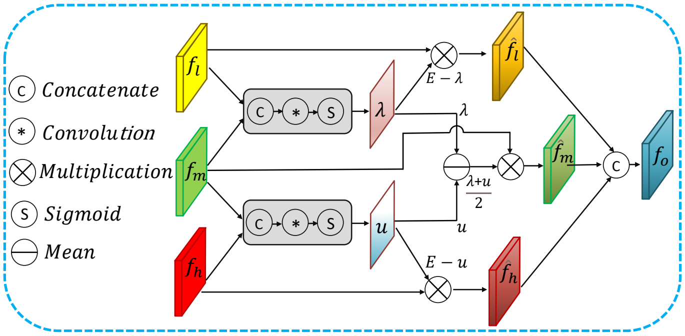


Fig. 2. Illustration of multi-scale feature reweighted fusion module.

$$l^k_{\text{SoftIoU}}(p,y) = \frac{\sum_{i,j}(Sig(p_{i,j})\cdot y_{i,j}) + \delta}{\sum_{i,j}(Sig(p_{i,j}) + y_{i,j} - Sig(p_{i,j})\cdot y_{i,j}) + \delta} \quad (1)$$

where $p_{i,j}$ and $y_{i,j}$ denote the values at point (i,j) on the k-th prediction feature map and the real mask label, respectively. *Sig* denotes the sigmoid activation function. The smoothing factor δ is set to 1. The final network training optimization loss is the mean value of all prediction branch losses.

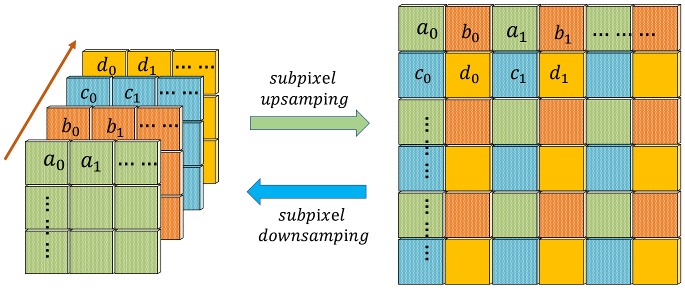


Fig. 3. The diagram of the sub-pixel upsampling and downsampling scheme when the sampling step r is equal to 2.

TABLE I
DESCRIPTION OF SIRST AND OUR IDST DATASETS.

Dataset	Frames	Target Number	TMP	TIPR/%	Target/image	Dataset Split
SIRST [14]	427	533	33.03	0.0605	1∼7	3:1:1
IDST	514	729	19.35	0.0089	1∼3	3:1:1

The abbreviations are defined as: TMP: Target mean pixels, TIPR: Target-to-image pixel ratio.

B. Multi-Scale Feature Reweighted Fusion

As shown in Fig. 1, we stack multiple high-to-low resolution subnetworks together to build the cuneate network. Different from the encoder and decoder structure in U-Net, which replenishes the semantic features by fusing the upscaled adjacent deep-layer feature maps, we design an MSFRF module to further make use of the adjacent shallow-layer and deep-layer feature maps in our cuneate network more effectively. The structure of MSFRF is illustrated in Fig. 2. Assume $f_m \in \mathbb{R}^{H\times W\times C_m}$ represents the output feature maps of the former subnetwork by skip connection. Meanwhile, $f_h \in \mathbb{R}^{H\times W\times C_h}$ and $f_l \in \mathbb{R}^{H\times W\times C_l}$ denote the adjacent upscaled deep-layer feature maps and downscaled shallow-layer feature maps, respectively. W, H, C_m, C_h, C_l represent the width, height, and channels of f_l, f_m, f_h. First, the shallow-layer fusion weight map $\lambda \in \mathbb{R}^{H\times W\times 1}$ and deep-layer fusion weight map $\mu \in \mathbb{R}^{H\times W\times 1}$ can be calculated as follows:

$$\begin{aligned} \lambda &= Sig(Conv(Concat(f_m, f_l))), \\ \mu &= Sig(Conv(Concat(f_m, f_h))). \end{aligned} \tag{2}$$

where the $Conv$ and $Concat$ denote 1×1 convolution and channel concatenation, respectively. Then, the feature maps f_l, f_m, and f_h will be reweighted by above weight maps λ and μ as follows:

$$\begin{aligned} \hat{f}_l &= (E-\lambda)\otimes f_l, \\ \hat{f}_h &= (E-\mu)\otimes f_h, \\ \hat{f}_m &= (\lambda \ominus \mu)\otimes f_m. \end{aligned} \tag{3}$$

where $\otimes$ and $\ominus$ denote the broadcast multiplication and pixel-by-pixel mean operation, respectively. $E \in \mathbb{R}^{H\times W\times 1}$ represents the feature map with all pixel values of 1. Finally, the output fused feature maps f_o can be obtained by concatenating the above reweighted feature maps $\hat{f}_l$, $\hat{f}_h$, and $\hat{f}_m$. By this spatial feature reweighted mechanism, our cuneate network can adaptively aggregate the multi-layer feature maps with rich detail and semantic features for better detection performance.

C. Sub-Pixel Sampling Scheme

The conventional downsampling pattern in DCNN always adopts maxpooling or averagepooling that results in the lose of dim small targets. Meanwhile, the main upsampling scheme such as deconvolution and uppooling often employ the strategy of making up 0 or interpolation, which damages the intrinsic features of small targets to a great extent. Inspired by the pixel accuracy and subpixel theory of the imaging mechanism, there exist smaller sub-pixels between two pixels in an image. Based on this idea, we devise a novel sub-pixel sampling scheme to better maintain the features of the infrared target. The specific principle of sub-pixel sampling is shown in Fig. 3. By combining the single pixel on the feature maps of multiple channels into one unit on a new feature map, the pixel on each original feature map is equivalent to the sub-pixel on a new feature map. From this perspective, the feature maps information on all channels will be fully utilized. Fig. 3 illustrates an up and down sub-pixel sampling scheme when sampling step r equals 2. We can further increase the sampling step r to form larger step size sampling such as 4×, 8×, and 16× sub-pixel upsampling. Furthermore, as illustrated in Fig 3, the upsampling and downsampling can be regarded as a pair of inverse transformations. Meanwhile, it can be seen that this kind of sub-pixel sampling scheme is parameterless with high efficiency, which further cuts down the parameters of network.

III. EXPERIMENTS

A. Implementation Details and Evaluation Metrics

In this letter, we implement comparative experiments on the SIRST [14] and IDST datasets. The description of these two datasets is shown in Table I. From Table I, we can conclude that the targets in IDST dataset have lower TMP and TIPR values which ratchets up the difficulty of detection. We divide the datasets into training set, validation set, and test set in the light of the ratio of 3:1:1. We resize all input infrared images to 256 × 256 pixels. All values in the input image will be normalized to the range of [-1,1] to expedite the network convergence. The initialization method and the optimization approach of our network adopt Xavier and Adagrad, respectively. The batch size and learning rate in the process of network training are 16 and 0.05. The software platform of our model is based on PyTorch 1.0.0 and Python 3.7.1. Our models were implemented on a computer with an Intel Xeon E5-2678 v3 CPU and an Nvidia GeForce 1080Ti GPU with 11 GB memory. To test the performance of different detection methods, the probability of detection (P_d), false alarm rate (F_a), and intersection-over-union (IoU) are adopted in our experiments. These evaluation metrics are defined as follows:

$$P_d = \frac{\text{number of correctly detected targets}}{\text{number of total targets}} \tag{4}$$

$$F_a = \frac{\text{number of falsely predicted pixels}}{\text{number of all image pixels}} \tag{5}$$

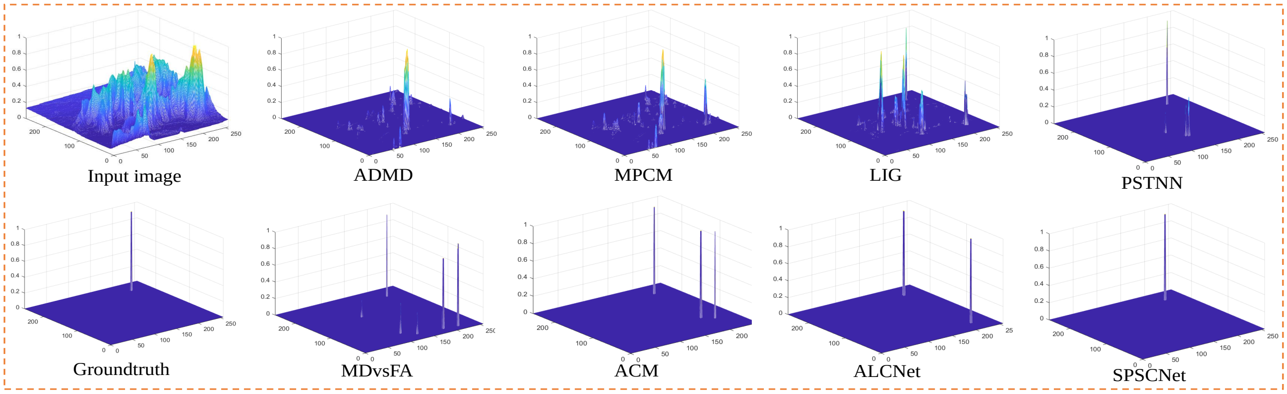


Fig. 4. Input image, groundtruth, and corresponding detection results of eight algorithms on the SIRST dataset.

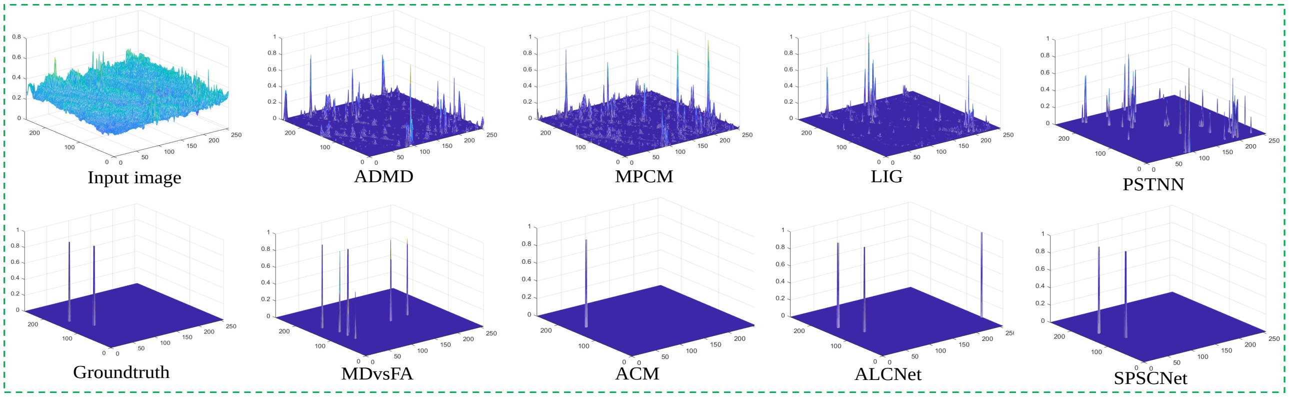
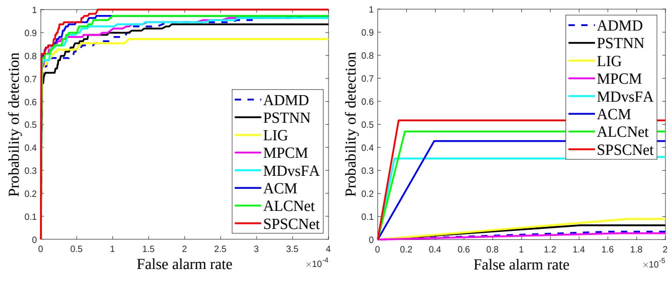


Fig. 5. Input image, groundtruth, and corresponding detection results of eight algorithms on the IDST dataset.



(a) (b)

Fig. 6. ROC curves of eight detection methods on the (a) SIRST dataset and (b) IDST dataset.

$$IoU = \frac{\mathrm{AI}}{\mathrm{AU}} \tag{6}$$

where AI and AU denote the area of intersection and area of union between the result of detection and ground truth. In addition, to distinguish performance differences between different algorithms, receiver operation characteristic (ROC) curve, which records the variation curve of ordinate P_d under varying abscissa F_a is adopted in our comparative experiments.

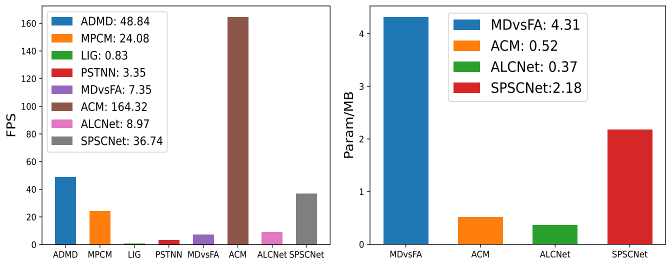
B. Experimental Results and Analysis

To manifest the superiority of our approach, we compare our SPSCNet with some state-of-the-art methods, including conventional model-based methods (ADMD [4], MPCM [6], LIG [7], and PSTNN [11]) and learning-based methods (MDvsFA [12], ACM [14], and ALCNet [16]) on the SIRST and IDST datasets. The proposed SPSCNet and seven comparison methods are tested on the test set of the SIRST and IDST datasets, and the results of P_d, F_a, and IoU are presented in Table II. We can observe that the metrics of P_d and IoU of our SPSCNet are all higher than the comparative four models-based methods and three learning-based methods, which demonstrates the precision advantage of our method. Meanwhile, the lower F_a index also proves that our method has stronger false alarm suppression ability and learns more discriminative dim small target features from complex backgrounds. In addition, as shown in Table II, our method with the sub-pixel sampling scheme obtains higher IoU and lower F_a than that with conventional sampling, which demonstrates the effectiveness of sub-pixel sampling. Furthermore, we visualized the final detection results of our methods and seven comparison methods. The 3D visualization maps of these methods on the SIRST and IDST datasets are shown in Fig. 4 and Fig. 5. We can observe that our method can accurately detect the number and the precise position of multiple infrared dim small targets. In contrast, other comparison methods have more or less missed detection and false alarm targets. Fig. 6 records the ROC curves of our SPSCNet and the other seven methods on the SIRST and IDST datasets. In theory, the closer the ROC curve is to the upper left corner, the better the detection performance of the model. It can be seen from Fig. 6 that the proposed method in this

TABLE II
P_d, F_a, AND IoU OF EIGHT METHODS ON THE SIRST AND IDST DATASETS.

Method	SIRST			IDST		
	Pd	Fa(e-05)	IoU	Pd	Fa(e-05)	IoU
ADMD	0.8440	21.930	0.1941	0.0414	97.932	0.0040
MPCM	0.8532	20.724	0.2607	0.0483	100.00	0.0039
LIG	0.8716	16.448	0.4076	0.1655	340.00	0.0992
PSTNN	0.8899	15.880	0.4782	0.0759	94.721	0.0156
MDvsFA	0.9358	12.420	0.6197	0.3793	5.2171	0.1613
ACM	0.9633	11.267	0.6772	0.4345	5.1299	0.2279
ALCNet	0.9725	10.947	0.7225	0.4966	4.9119	0.2932
SPSCNet[1]	0.9851	10.372	0.7413	0.5104	4.9056	0.3078
SPSCNet	**0.9908**	**9.1020**	**0.7533**	**0.5241**	**4.8828**	**0.3111**

SPSCNet[1] denotes our network with conventional sampling scheme (max-pooling and uppooling).



(a) (b)

Fig. 7. (a) Frame per second (FPS) of eight methods. (b) Parameters of the four learning-based methods.

letter has a better performance on two infrared dim small target datasets, especially on the IDST dataset with smaller targets, which further indicates the superiority of our SPSCNet.

Except for the above performance indicators of P_d, F_a, and IoU, we also record the frame per second (FPS) of these eight models in Fig. 7(a) and the parameter quantity of four learning-based models in Fig. 7(b). For fair comparison, we test all eight models on the test set of the IDST dataset. It can be drawn that the proposed SPSCNet gets 36.74 FPS on the IDST dataset. The FPS of our method is only lower than ACM [14] (164.32 FPS) and ADMD [4] (48.84 FPS). However, the parameters of our method reach 2.18 MB which is only less than MDvsFA [12] (4.31 MB) among the four learning-based methods. These test results demonstrate that our model still has some shortcomings in efficiency. We have increased the scale of the network and greatly improved the detection accuracy. However, the parameters and inference time of the model are also increased accordingly. Therefore, using certain model compression techniques such as pruning and distillation to reduce the bloated and useless parameters while maintaining high detection accuracy is the next significant research field in future work. parameters while maintaining high detection accuracy is the next significant research field in future work.

IV. Conclusion

In this letter, we propose a sub-pixel sampling cuneate network for infrared dim small target detection. The overall model framework is based on a cuneate network with multiple parallel high-to-low resolution subnetworks. To address the concern of dim small targets being lost in conventional sampling scheme, we design a parameterless sub-pixel sampling scheme to make full use of the features of multiple channels. Moreover, we explore a novel multi-scale feature reweighted fusion module to reconcile the features of different scales. The experimental results on the SIRST and IDST datasets prove the effectiveness of our method.

References

[1] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, early access, Feb. 16, 2022, doi: 10.1109/MGRS.2022.3145502.

[2] J.-F. Rivest and R. Fortin, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, pp. 1886–1893, 1996.

[3] S. Aghaziyarati, S. Moradi, and H. Talebi, "Small infrared target detection using absolute average difference weighted by cumulative directional derivatives," *Infr. Phys. Technol.*, vol. 101, pp. 78–87, 2019.

[4] S. Moradi, P. Moallem, and M. F. Sabahi, "Fast and robust small infrared target detection using absolute directional mean difference algorithm," *Signal Process.*, vol. 177, p. 107727, 2020.

[5] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, 2013.

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

[7] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Infrared small-target detection using multiscale gray difference weighted image entropy," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 1, pp. 60–72, 2016.

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

[9] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, 2017.

[10] M. Wan, G. Gu, Y. Xu, W. Qian, K. Ren, and Q. Chen, "Total variation-based interframe infrared patch-image model for small target detection," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2021.

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

[12] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 8509–8518.

[13] Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, and W. Zhang, "Ristdnet: Robust infrared small target detection network," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[14] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE/CVF Wint. Conf. Appl. Comput. Vision*, 2021, pp. 950–959.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*, 2015, pp. 234–241.

[16] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[18] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," 2015. [Online]. Available: https://arxiv.org/abs/1505.02496

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[21] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. 12th Int. Symp. Vis. Comput. (ISVC)*, 2016, pp. 234–244.