RISTDnet: Robust Infrared Small Target Detection Network

Qingyu Hou, Zhipeng Wang, Fanjiao Tan, Ye Zhao, Haoliang Zheng, and Wei Zhang

***Abstract*— The infrared (IR) small target detection algorithm with a high detection rate, low false alarm rate, and high real-time performance has significant application value in the field of IR remote sensing. IR small targets in complex backgrounds have low contrast and low signal-to-noise ratio (SNR). Therefore, small target detection is more difficult. Traditional IR small target detection is generally implemented by local contrast methods (LCM), nonlocal autocorrelation methods (NAM), and adaptive segmentation. In this letter, a robust infrared small target detection network (RISTDnet) is proposed based on deep learning. In RISTDnet, a feature extraction framework combining handcrafted feature methods and convolutional neural networks is constructed, a mapping network between feature maps and the likelihood of small targets in the image is established, and a threshold is applied on the likelihood map to segment real targets. Experimental results show that the RISTDnet can detect small targets with different sizes and low SNRs in complex backgrounds and have better effectiveness and robustness against existing algorithms.**

***Index Terms*— Convolutional neural network (CNN), infrared (IR) small target, multiscale detection, robust infrared small target detection network (RISTDnet).**

I. Introduction

INFRARED small target detection plays an important role in the field of infrared (IR) image processing. Detecting low signal-to-noise ratio (SNR) small targets under complex backgrounds is often a challenging task because targets in the image are usually small and weak, without specific shape, texture, and structural information.

In recent decades, numerous IR small target detection algorithms have been proposed, which includes two steps: single-frame detection and multiframe association. In this letter, we only focus on single-frame detection. Traditional single-frame detection methods can be further divided into two categories: local contrast methods (LCMs) and nonlocal autocorrelation methods (NAMs).

A. Local Contrast Methods

It is very important to propose a reasonable definition of local contrast for LCM, which can suppress complex background and enhance small target as much as possible. Chen *et al.* [1] applied LCM to measure the difference between each pixel position and its neighborhood. Han *et al.* [2] proposed an improved LCM (ILCM), which further considers the mean estimation of the central subblock and improves the efficiency of the algorithm by increasing the step size of the sliding window. Wei *et al.* [3] proposed a multiscale patch-based contrast measure (MPCM). It uses the gray ratio between each pixel location and its adjacent areas as an enhancement factor, which can effectively enhance the true target. Class LCM also includes relative local contrast measure (RLCM) [4] and multiscale modified LCM (MLCM) [5]. Han *et al.* [6] adopted ratio-difference joint local contrast measure (RDLCM) to enhance true small target and suppress background.

B. Nonlocal Autocorrelation Methods

In the past decade, NAM uses the nonlocal autocorrelation of IR backgrounds and the sparsity of target to separate the target from the background and have been developed rapidly. The infrared patch image (IPI) model separates small target and background by sparse representation and low-rank matrix recovery [7]. The detection problem is solved by optimizing the objective function.

The two obvious disadvantages of IPI are target over shrinkage and noise residual, which are mainly caused by using nuclear norm as the low-rank regularization term. In order to solve these problems, more low-rank matrix recovery techniques are introduced into the IPI model to achieve better performance. Dai *et al.* [8] proposed the nonnegative infrared patch image (NIPPS) model and the weighted infrared patch tensor (RIPT) model [9]. In addition, different forms of tighter rank substitution are used to solve the background residual problem of the traditional IPI model. Zhang *et al.* [10] used nonconvex rank approximation minimization joint $l_{2,1}$ norm (NRAM) instead, and Zhou *et al.* [11] applied an effective integration of the Schatten 1/2 quasi-norm regularization and reweighted sparse enhancement to improve performance.

However, these methods are usually based on handcrafted features, and the performance depends on the effectiveness of features. In recent years, some target detection methods based on deep learning are proposed, such as faster R-convolutional neural network (CNN) [12], YoloV2 [13], and YoloV3 [14], which have high performance on large-scale targets with obvious texture. Deep learning can learn features from a large

Manuscript received October 16, 2020; revised December 6, 2020; accepted December 27, 2020. *(Corresponding author: Fanjiao Tan.)*

The authors are with the School of Aeronautics, Harbin Institute of Technology, Harbin 150001, China (e-mail: houqingyu@126.com; hit_wangzhipeng@163.com; tanfj333@163.com; 1059882633@qq.com; 842073444@qq.com; wzhang@hit.edu.cn).

Color versions of one or more figures in this letter are available at https://doi.org/10.1109/LGRS.2021.3050828.

Digital Object Identifier 10.1109/LGRS.2021.3050828

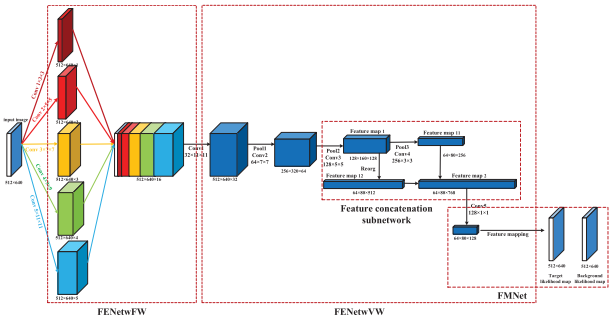


Fig. 1. Network structure of RISTDnet.

number of training data automatically, and it is more helpful to describe the rich and unique information of data than handcrafted features. However, the texture features of the IR small target are not obvious extremely. Therefore, the existing methods using deep learning [12]–[14] are not suitable for the detection of the IR small target. Aiming at the deficiencies of deep learning technology, an improved IR small target detection method based on deep learning is proposed.

The following contributions are made in this letter.

1) Robust infrared small target detection network (RISTDnet) is proposed based on deep learning.
2) In RISTDnet, a feature extraction framework combining handcrafted feature methods and CNNs is constructed, and a feature mapping network (FMNet) between feature map and the likelihood map of small targets or background in the image is established.
3) Experimental results on several real images and simulation images demonstrate that our method is better than the currently existing methods.

II. Proposed Method

First, the likelihood map is calculated by RISTDnet, where the pixel value represents the probability that the pixel belongs to a small target or background. Then, a threshold is applied to the likelihood map to extract real targets.

A. RISTDnet

Taking the 512 × 640 IR image as an example, the network structure is described as follows. The input image is normalized after subtracting the minimum gray value and dividing it by the gray value range.

The specific network structure of RISTDnet is shown in Fig. 1. The input image is a single-channel IR image, and the output data is the same size as the input image. The value of each pixel represents the probability value of the pixel as the target or background. RISTDnet consists of a feature extraction network based on convolution kernel with fixed weight (FENetwFW), a Feature extraction network based on convolution kernel with variable weight (FENetwVW), and a Feature mapping network (FMNet).

In Fig. 1, the FENetwFW uses five sizes of convolution kernels: 3 × 3, 5 × 5, 7 × 7, 9 × 9, and 11 × 11. The number of convolution kernels is 1, 2, 3, 4, and 5, respectively. The structure is shown in Fig. 2. The result of each convolution

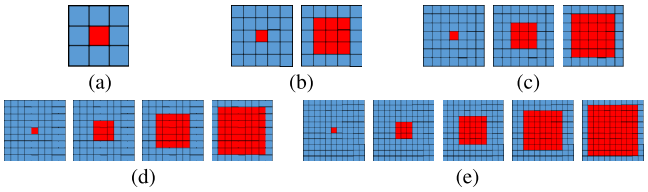


Fig. 2. Five sets of convolution kernels in FENetwFW. (a) 3 × 3 structure. (b) 5 × 5 structure. (c) 7 × 7 structure. (d) 9 × 9 structure. (e) 11 × 11 structure.

TABLE I
Structure of FENetwVW

Type	Number of convolution kernels	Convolution kernel size / step size	Output (input)
Input layer	–	–	512×640
Conv1	32	11×11/1	512×640
Pool1	–	2×2/2	256×320
Conv2	64	7×7/1	256×320
Pool2	–	2×2/2	128×160
Conv3	128	5×5/1	128×160
Pool3	–	2×2/2	64×80
Conv4	256	3×3/1	64×80
Conv5	128	1×1/1	64×80

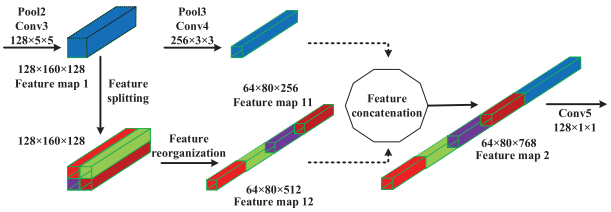


Fig. 3. Process of feature concatenation in FCsubnet.

kernel is the average value of the pixels in the red grid minus the average value of the pixels in the blue grids. The purpose of these sets of convolution kernels is to extract the contrast information of the small target.

The FENetwFW contains 15 convolution kernels with the abovementioned fixed weight, the number of fixed weight convolution kernels with a high response is 5, 4, 3, 2, and 1 for the target with sizes 1 × 1, 3 × 3, 5 × 5, 7 × 7, and 9 × 9, respectively, and the number of corresponding feature maps is 5, 4, 3, 2, and 1. Finally, these 15 feature maps and the original image constitute the feature map of 16 channels. It can be seen that the strategy can achieve an adequate number of feature channels for the smaller targets and then ensure the detectability for the smaller target.

The input image and feature maps formed by five sets of convolution kernels are concatenated as the input of the FENetwVW, and the structure of FENetwVW is shown in Table I.

The FENetwVW contains a Feature concatenation subnetwork (FCsubnet), which is a network structure derived from passthrough. After splitting and rearranging high-resolution feature maps, FCsubnet connects the results with low-resolution feature maps to form a multiscale feature map, which is more conducive to detecting small targets of different sizes. The detailed structure of FCsubnet is shown in Fig. 3.

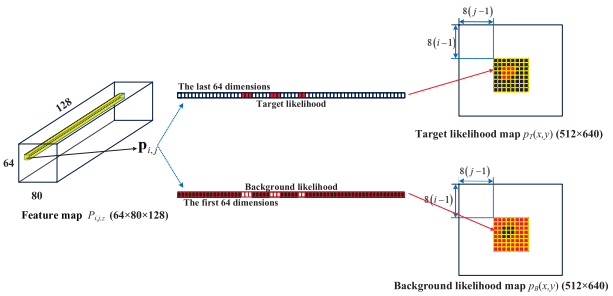


Fig. 4. Feature mapping process in FMNet.

The FMNet is used to map the feature map to the target/background likelihood map. The calculation process is shown in Fig. 4. Let $p_{i,j,z}$ denote an element in a feature map with a size of $64 \times 80 \times 128$, where, $i = 1, 2, \ldots, 64$; $j = 1, 2, \ldots, 80$; and $z = 1, 2, \ldots, 128$. The relationship between the target likelihood map $p_T(x, y)$ and the background likelihood map $p_B(x, y)$ $(x = 1, 2, \ldots, 512, y = 1, 2, \ldots, 640)$ and $p_{i,j,z}$ can be expressed as

$$\begin{aligned} \text{Patch}_T &= p_T[8(i-1)+(1:8), 8(j-1)+(1:8)] \\ &= \text{Sigmoid}\begin{bmatrix} p_{i,j,1} & p_{i,j,2} & \cdots & p_{i,j,8} \\ p_{i,j,9} & p_{i,j,10} & \cdots & p_{i,j,16} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i,j,57} & p_{i,j,58} & \cdots & p_{i,j,64} \end{bmatrix} \end{aligned} \tag{1}$$

$$\begin{aligned} \text{Patch}_B &= p_B[8(i-1)+(1:8), 8(j-1)+(1:8)] \\ &= \text{Sigmoid}\begin{bmatrix} p_{i,j,65} & p_{i,j,66} & \cdots & p_{i,j,72} \\ p_{i,j,73} & p_{i,j,74} & \cdots & p_{i,j,80} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i,j,123} & p_{i,j,124} & \cdots & p_{i,j,128} \end{bmatrix} \end{aligned} \tag{2}$$

where Patch_T represents a matrix in the range of $8(i-1)+1$ to $8(i-1)+8$ rows, $8(j-1)+1$ to $8(j-1)+8$ columns in $p_T(x, y)$. Patch_B represents a matrix in the range of $8(i-1)+1$ to $8(i-1)+8$ rows, $8(j-1)+1$ to $8(j-1)+8$ columns in $p_B(x, y)$. Sigmoid means that the matrix elements are calculated by the sigmoid function.

It can be seen that FENetwFW extracts handcrafted multi-scale features in RISTDnet, and FENetwVW performs deep features extraction on the basis of handcrafted features. Therefore, RISTDnet is a fusion extraction framework of handcrafted features and deep features for small targets' detection.

B. Network Training

In this letter, the background image is obtained with data augmentation strategies, such as geometric transformation and radiation transformation, and then, the target is added to the background image randomly corroding to the position, SNR, and size. Targets' position obeys uniform distribution. The range of target SNR is 0.5∼5, and the range of size is $1 \times 1 \sim 13 \times 13$. Several transformed background images are shown in Fig. 5. The definition of target SNR is given by

$$\text{SNR} = \frac{|\mu_t - \mu_b|}{\sigma_b} \tag{3}$$

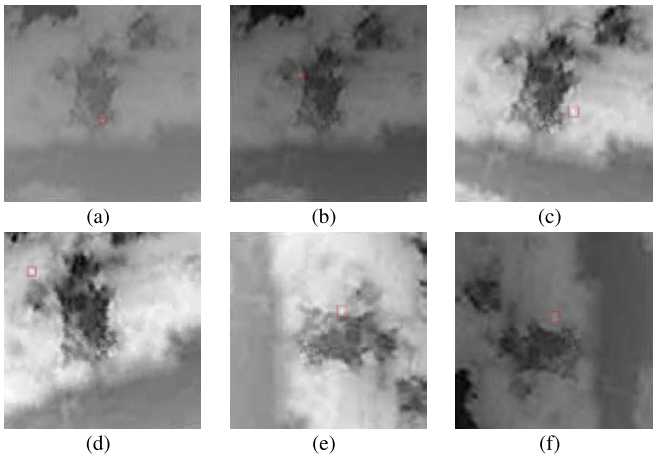


Fig. 5. Several training images are generated by background data augmentation and target random addition. (a) Original images. (b) Contrast weakened images. (c) Contrast enhanced images. (d) Partially enlarged images. (e)Rotated clockwise images. (f) Rotated counterclockwise images.

where μ_t is the maximum of the target intensity, μ_b is the average background intensity corresponding to the target position, and σ_b is the mean square deviation of flat area in the image.

The input of network training is normalized IR image. The loss function is calculated by the target/background likelihood map and the label of the training data set. The network training realizes the optimization of the loss function based on the gradient descent algorithm and the solution of weights of all convolution kernels in FENetwVW. The probability loss function is as follows:

$$\begin{aligned} \text{loss} = \frac{1}{K}\sum_{k=1}^{K}\sum_{x=1}^{640}\sum_{y=1}^{512}\Big[&\left\|p_B^k(x,y)\text{-}\hat{p}_B^k(x,y)\right\|^2 \\ &+\left\|p_T^k(x,y)-\hat{p}_T^k(x,y)\right\|^2\Big] \end{aligned} \tag{4}$$

where k is the frame number, $k = 1, 2, \ldots, K$; $p_T^k(x, y)$ and $p_B^k(x, y)$ represent the target likelihood map and background likelihood map based on the network predictability. The target likelihood map $\hat{p}_T^k(x, y)$ and the background likelihood map $\hat{p}_B^k(x, y)$ are generated by the label, expressed as

$$\hat{p}_T^k(x,y) = \begin{cases} 1, & (x,y) \in \text{target pixels} \\ 0, & (x,y) \in \text{background pixels} \end{cases} \tag{5}$$

$$\hat{p}_B^k(x,y) = \begin{cases} 1, & (x,y) \in \text{background pixels} \\ 0, & (x,y) \in \text{target pixels.} \end{cases} \tag{6}$$

C. Threshold Operation

The normalized IR image is processed by FENetwFW, FENetwVW, and FMNet, and the target likelihood map is obtained. Threshold segmentation and connected domain extraction are adopted in the likelihood map to get the position and scale information of the target.

The probability threshold Th is selected to segment the target likelihood map. The values of small targets' pixels in the likelihood map are greater than Th, while the values of background pixels are less than Th.

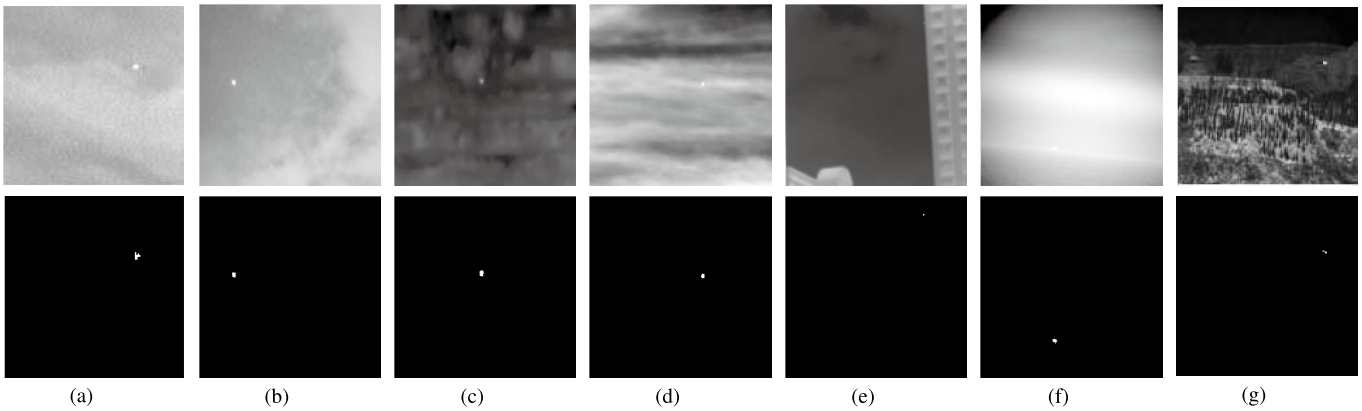


Fig. 6. IR images and corresponding RISTDnet detection results. The size of the target in different scenes is as follows: (a) 4×4, (b) 8×6, (c) 8×8, (d) 7×5, (e) 3×3, (f) 9×5, and (g) 5×5.

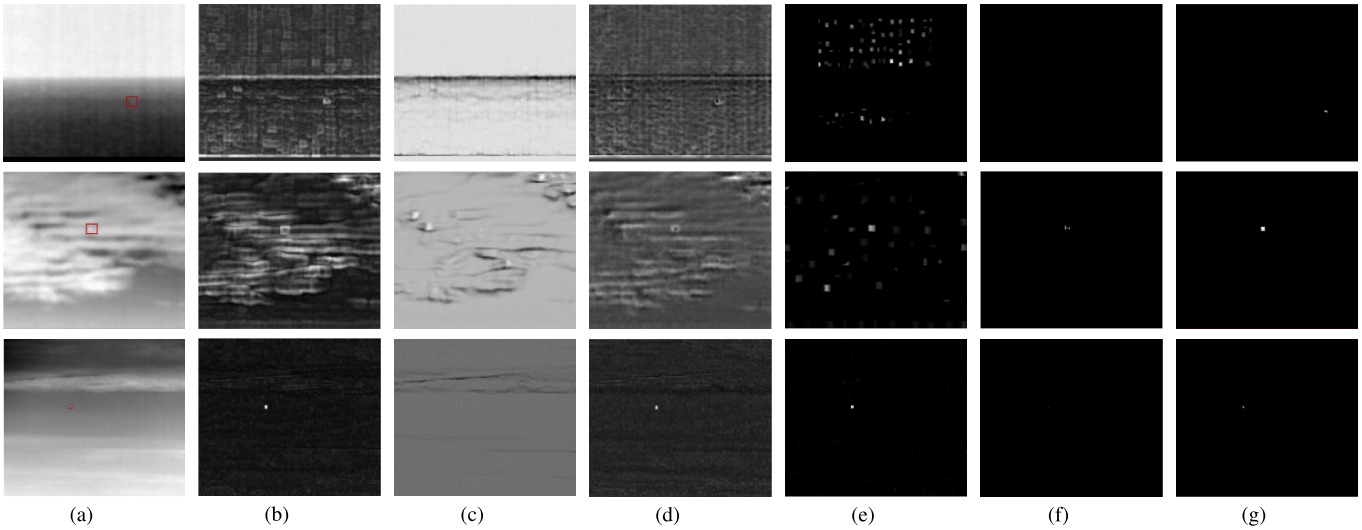


Fig. 7. Detection results. (a) Raw IR images. (b) MLCM results. (c) MPCM result. (d) ILCM result. (e) RLCM result. (f) RIPT results. (g) RISTDnet results.

III. Experiments and Analysis

In order to verify the performance of the RISTDnet detection algorithm proposed in this letter, the test program for RISTDnet is written based on TensorFlow 1.12.0 and Python 3.6.4.

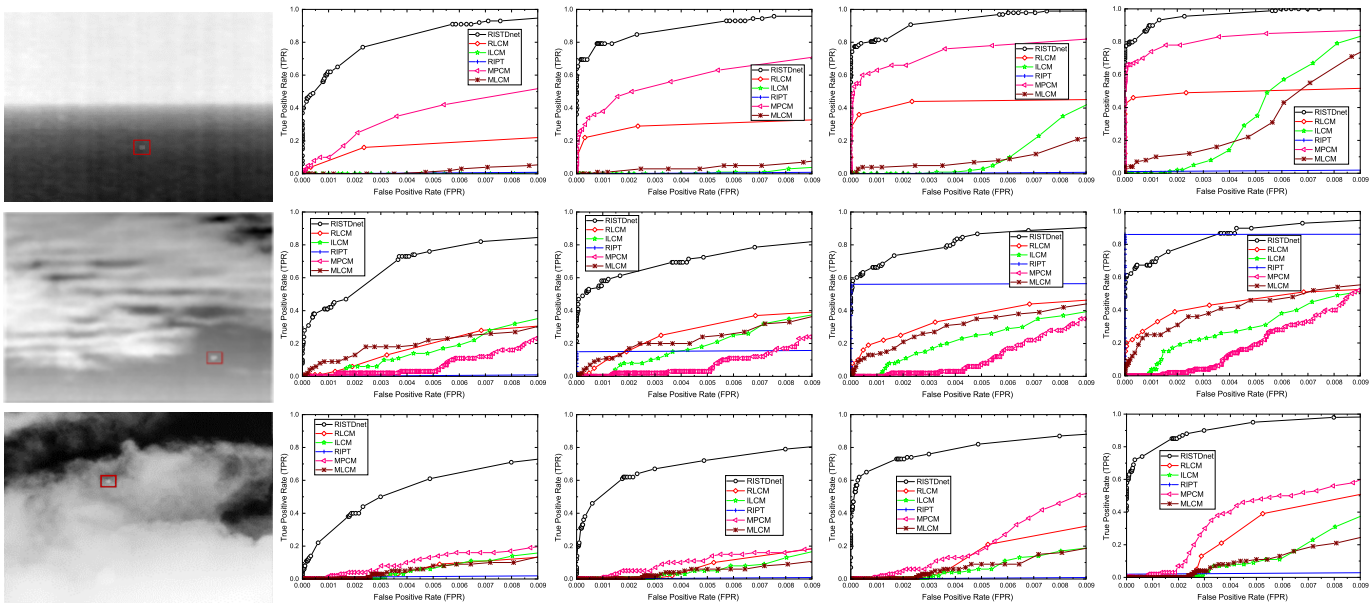
The training data set contains 4122 images, which includes four types of background: cloud background, cloudless background, building-sky background, and ground-sky background. The test data set contains 4000 images, which includes four types of background: forest-sky background, cloud background 1, cloud background 2, and sea-sky background (as shown in Figs. 6–8). The test image set not included in the training data set is shown in the first line of Fig. 6. These images not only contain complex background clutter but also contain targets of different sizes and intensities. Use the algorithm proposed in this article to detect these images, and the target detection results are shown in the second line of Fig. 6. It can be seen that the algorithm in this letter has better detection performance for small targets in complex background images.

In order to prove the advantages of RISTDnet, we compare its performance with MLCM [5], MPCM [3], ILCM [2], RLCM [4], and RIPT [9] (the codes of these algorithms are obtained from the websites of the authors, and the default parameter settings are used). Three test images are shown in Fig. 7(a). The results of target enhancement processing corresponding to different algorithms are given in Fig. 7(b)–(g), Fig. 7(g) shows the target likelihood map. It can be seen that the target likelihood map corresponding to RISTDnet has obvious advantages, and the contrast between the target and the background is significantly higher than other algorithms.

In addition, in order to describe the performance comparison between the proposed algorithm and other algorithms more systematically, the receiver operating characteristic (ROC) curve is used as the performance quantitative index, where the false positive rate (FPR) and the true positive rate (TPR) are defined as

$$\begin{aligned} \mathrm{FPR} &= \frac{\text{number of detected false targets}}{\text{total number of pixels in the whole image}} \\ \mathrm{TPR} &= \frac{\text{number of detected true targets}}{\text{total number of real targets}}. \end{aligned} \tag{7}$$

This letter mainly compares the detection performance of low SNR targets. The first line in Fig. 8(b)–(e) represents the ROC curve when the SNR of the target is 1, 1.5, 2, and 2.5, respectively, in the corresponding scenarios; the second line



(a) (b) (c) (d) (e)

Fig. 8. ROC curves of different methods on datasets 1–3. Row 1: sea-sky background. Row 2: cloud background1. Row 3: cloud background2. Different columns represent ROC curves of different methods for various SNR targets.

and the third line in Fig. 8(b)–(e) represent the ROC curve when the SNR of the target is 0.5, 1, 1.5, and 2, respectively, in the corresponding scenarios.

The ROC curves obtained by different algorithms for these images are shown in Fig. 8. It can be seen from Fig. 8 that the algorithm in this letter has a better performance on low SNR, indicating the advantages of RISTDnet in low SNR target detection.

The IR small target detection algorithm should have a constant false alarm rate (CFAR) detection ability for the complex background. Considering that the false alarm rate is restricted by the high detection threshold, we test the relationship between Th and FPR in several simulation scenarios. By setting $Th = 0.99$, the FPR can be limited to less than 3×10^{-5}. Meanwhile, RISTDnet realized by convolution neural network can reach 101 fps in NVIDIA RTX2080Ti.

IV. Conclusion

In this letter, a RISTDnet for IR small target detection based on deep learning methods is proposed. RISTDnet transforms the input image into a target/background likelihood map, and a threshold is applied to extract real targets. Experimental results show that the proposed RISTDnet can deal with small targets of different sizes and low SNR under complex backgrounds, and the proposed RISTDnet also has better effectiveness and real-time performance against existing algorithms.

References

[1] C. L. 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, Jan. 2014, doi: 10.1109/TGRS.2013.2242477.

[2] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, "A robust infrared small target detection algorithm based on human visual system," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168–2172, Dec. 2014, doi: 10.1109/LGRS.2014.2323236.

[3] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, Oct. 2016, doi: 10.1016/j.patcog.2016.04.002.

[4] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, "Infrared small target detection utilizing the multiscale relative local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 612–616, Apr. 2018, doi: 10.1109/LGRS.2018.2790909.

[5] S. Yao, Y. Chang, and X. Qin, "A coarse-to-fine method for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 256–260, Feb. 2019, doi: 10.1109/LGRS.2018.2872166.

[6] J. Han, S. Liu, G. Qin, Q. Zhao, H. Zhang, and N. Li, "A local contrast method combined with adaptive background estimation for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1442–1446, Sep. 2019, doi: 10.1109/LGRS.2019.2898893.

[7] 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, Dec. 2013, doi: 10.1109/TIP.2013.2281420.

[8] Y. Dai, Y. Wu, Y. Song, and J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infr. Phys. Technol.*, vol. 81, pp. 182–194, Mar. 2017, doi: 10.1016/j.infrared.2017.01.009.

[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, Aug. 2017, doi: 10.1109/JSTARS.2017.2700023.

[10] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, "Infrared small target detection via non-convex rank approximation minimization joint l2,1 norm," *Remote Sens.*, vol. 10, no. 11, p. 1821, Nov. 2018, doi: 10.3390/rs10111821.

[11] F. Zhou, Y. Wu, Y. Dai, and P. Wang, "Detection of small target using schatten 1/2 quasi-norm regularization with reweighted sparse enhancement in complex infrared scenes," *Remote Sens.*, vol. 11, no. 17, p. 2058, Sep. 2019, doi: 10.3390/rs11172058.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767