Infrared Small UAV Target Detection Based on Residual Image Prediction via Global and Local Dilated Residual Networks

Houzhang Fang, *Member, IEEE*, Mingjiang Xia, Gang Zhou, Yi Chang, *Member, IEEE*, and Luxin Yan, *Member, IEEE*

***Abstract*— Thermal infrared imaging possesses the ability to monitor unmanned aerial vehicles (UAVs) in both day and night conditions. However, long-range detection of the infrared UAVs often suffers from small/dim targets, heavy clutter, and noise in the complex background. The conventional local prior-based and the nonlocal prior-based methods commonly have a high false alarm rate and low detection accuracy. In this letter, we propose a model that converts small UAV detection into a problem of predicting the residual image (i.e., background, clutter, and noise). Such novel reformulation allows us to directly learn a mapping from the input infrared image to the residual image. The constructed image-to-image network integrates the global and the local dilated residual convolution blocks into the U-Net, which can capture local and contextual structure information well and fuse the features at different scales both for image reconstruction. Additionally, subpixel convolution is utilized to upscale the image and avoid image distortion during upsampling. Finally, the small UAV target image is obtained by subtracting the residual image from the input infrared image. The comparative experiments demonstrate that the proposed method outperforms state-of-the-art ones in detecting real-world infrared images with heavy clutter and dim targets.**

***Index Terms*— Convolutional neural network (CNN), infrared small unmanned aerial vehicle (UAV) target, residual learning, target detection.**

I. Introduction

INFRARED unmanned aerial vehicles (UAVs) target detection has been attracting considerable attention due to their significant applications in the fields of civil and military security at low altitude. The performance of the UAV target detection relies heavily on the accuracy of small target detection at long distances. Current infrared small UAV detection has two major challenges. First, due to the long imaging distance from infrared sensors, the UAV target occupies only a small proportion of the whole observed image, and has no obvious texture and shape features. Second, the infrared UAV image often suffers from strong background clutter and noise, such as heavy clouds and buildings, thus a low signal-to-clutter ratio and a low signal-to-noise ratio. The above challenges make infrared small UAV detection a difficult task.

In this letter, we focus on the single-frame detection methods, which can be generally divided into three categories: local prior-based methods, nonlocal prior-based methods, and learning-based methods. In the local prior-based methods [1]–[3], the filtering methods [1] such as max-mean and max-median, are previously used to enhance the small target. However, some strong edges such as the edges of the cloud-sky background are also strengthened. Recently, the human visual system (HVS)-based methods [2], [3] utilize the local contrast to enhance the small target and suppress the background clutter. However, being highly sensitive to complex background clutter and noise, these methods can achieve good detection performance only for simple background, and are prone to fail for complex ones. The nonlocal prior-based methods show a competitive detection performance in recent years [4]–[8]. The small target detection is converted into an optimization problem of recovering a low-rank component and sparse component by employing the nonlocal self-similarity of the background and the sparsity of the small target. However, the performance of these methods will inevitably deteriorate when the background is filled with more complicated clutter and noise, which will destroy the nonlocal self-similarity of the background.

Recently, learning-based methods are also adopted to detect the infrared small target [9], [10]. Since the infrared small target has no evident texture and shape features, it is very difficult to perform feature learning directly for small targets with heavy background clutter using the convolutional neural network (CNN)-based object detection techniques, such as Faster R-CNN [9], thus a poor detection performance. Shi and Wang [10] regarded the small target as noise and proposed a denoising autoencoder network that uses the pre-trained Visual Geometry Group (VGG)-19 model as a basic network. Due to the loss of too many details during the pooling

Manuscript received April 13, 2021; revised May 16, 2021; accepted May 26, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41501371 and Grant 61971460, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JM4036, and in part by the Open Research Fund of the National Key Laboratory of Science and Technology on Multispectral Information Processing under Grant 6142113190103. *(Corresponding author: Houzhang Fang.)*

Houzhang Fang and Mingjiang Xia are with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: houzhangfang@xidian.edu.cn; xmj04027@gmail.com).

Gang Zhou is with the State Key Laboratory of Material Processing and Die and Mold Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: visualab@163.com).

Yi Chang is with the Artificial Intelligence Research Center, Pengcheng Laboratory, Shenzhen 518055, China (e-mail: owuchangyuo@gmail.com).

Luxin Yan is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: yanluxin@hust.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/LGRS.2021.3085495.

Digital Object Identifier 10.1109/LGRS.2021.3085495

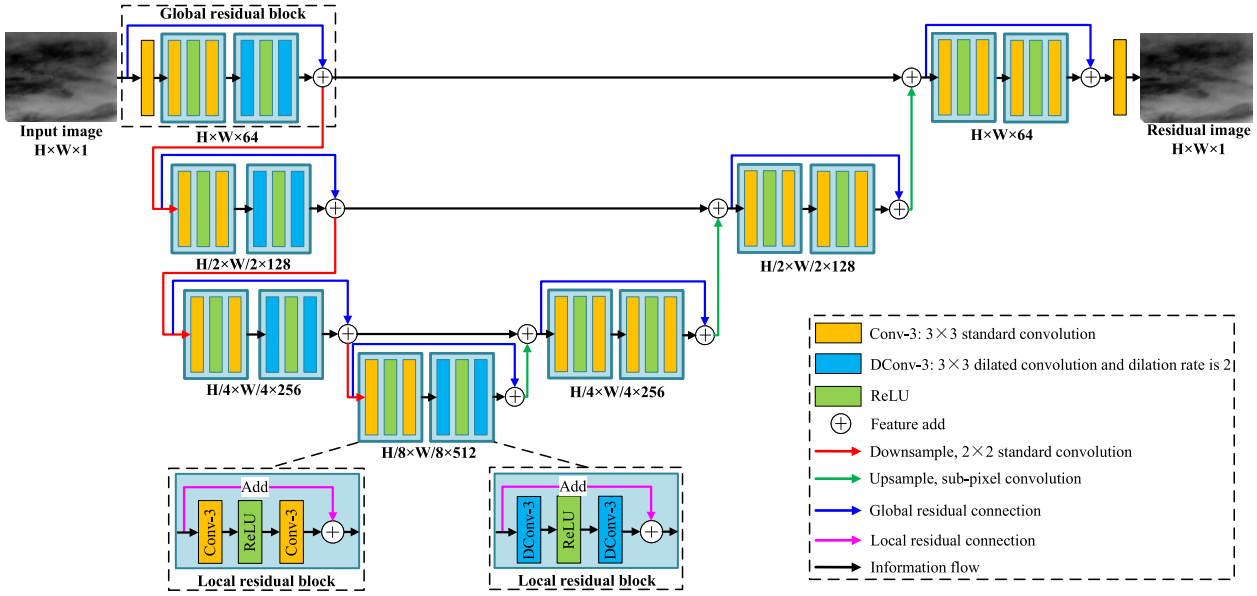

Fig. 1. Architecture of the proposed network. Each rectangle denotes a feature map extracted by the convolution kernels. The number below the rectangle represents the size of each feature map in a corresponding row. The operations are indicated by arrows; a description of each operation is explained in the bottom and the right bottom.

operation in the downsampling of the encoding process, it is hard to reconstruct the high-frequency details well in the decoding stage.

Our work is inspired by the residual image learning that is exploited for image restoration [11]. The DnCNN network in [11] is initially employed to predict image corruptions (e.g., noise) rather than clean images, which is generally an easier task. In this letter, we model the infrared small target detection as predicting the residual image by CNN. The background, clutter, and noise are regarded as the residual image between the input infrared image and the small UAV target image. Once the residual image is estimated, the small target image is obtained by subtracting the residual image from the input image. It is difficult to achieve satisfactory performance when directly using the DnCNN network [11] to predict the residual image due to its complex structures. To address the above issue, we propose a model that incorporates the global and the local dilated residual convolution blocks into the U-Net network. The global and the local dilated residual convolution blocks, as well as skip connections between the encoding and decoding stages, prevent the gradient from vanishing or exploding during the back-propagation, making the network training more stable. In particular, the global residual connection (GRC) can enhance the flow of shallow information to the deep layer to reduce the loss of feature map information. In addition, the introduction of the local residual block with the dilated convolution can enlarge the receptive field and capture more context information to reconstruct multiple structural components of the residual image without the need to increase the network depth. Furthermore, the efficient subpixel convolution (ESPC) [12] is employed to upscale the image and avoid image distortion during the decoding stage. In summary, the main contributions of this letter are listed as follows.

1) We propose a novel multiscale U-Net architecture that predicts the residual image between the input image and the target image for infrared small UAV target detection.

2) We propose a global residual block composed of two consecutive local residual blocks, which can not only capture the local and contextual features for image reconstruction, but also better fuse the features from the previous layer and the current layer at different scales via the GRC.

3) Our method significantly outperforms the state-of-the-art methods in terms of the qualitative and quantitative assessments of real infrared UAV images with heavy clutter, complex building, and dim targets.

II. Proposed Method

In this section, we present the proposed residual learning CNN model [dilated residual U-net (DRUNet)] for the small target detection in detail. Training a CNN model for our small target detection mainly involves two steps: 1) design of the network architecture and 2) model training from the training dataset. Shown in Fig. 1 is the basic network architecture.

A. Infrared Small Target Detection Based on the Residual Image Prediction

Generally, the infrared image model can be formulated as follows: $D = T + B + N$, where D, T, B, and N represent the original infrared image, the target image, the background image, and the random noise image, respectively. The goal of our residual learning network is to learn the nonlinear residual mapping $\mathcal{R}(D)$ from D to $D-T$, so the small target image can be achieved via $T = D - \mathcal{R}(D)$. Formally, the mean squared error between the residual image and the estimated ones from the original infrared image: $\ell(\Theta) = 1/2K \sum_{k=1}^{K} |(D - T) - \mathcal{R}(D)|_F^2$ is employed as the loss function to learn the network

parameters Θ, where $\{(D_k, B_k + N_k)\}_{k=1}^{K}$ denotes K infrared and residual image pairs.

The constructed network needs to predict the residual image that contains the structured regions (such as clutter, noise, and edges) and homogeneous (i.e., flat) regions, a very challenging task. In this study, we propose a model that adopts the U-Net-based multiscale network as the basic network, which is widely exploited as an image-to-image nonlinear mapping. The residual convolution block with normal convolution can capture the fine structures in the shallow layer and the large-scale edges in the deep layer; meanwhile, it can also maintain the relevance of local information. The dilated residual convolution block integrated into the U-Net can capture the contextual structure features in the corresponding layer. The constructed multiscale network can effectively represent the structural information at different scales, which is conducive to the reconstruction of the residual images with multiple complex components.

B. Network Architecture

The proposed architecture consists of a downsampling stage (left) and an upsampling (right) stage (Fig. 1), connected to each other via skip connections. The downsampling and upsampling stages can be regarded as the feature extraction and image reconstruction modules, respectively. The proposed framework has 36 convolutional layers in total. The U-Net contains four scale feature maps and each scale extracts the feature maps using a global residual block. Each global residual block has a normal local convolution residual block followed by a dilated convolution block composed of consecutive 3×3 dilated convolutions with a dilation rate of 2. Each normal or dilated convolution is followed by a rectified linear unit (ReLU) activation. In the downsampling stage, the input image of size $H \times W$ is delivered to two residual blocks. During the downsampling stage at each scale, the feature map is downsampled using a 2×2 convolution with stride 2 and is then fed to the next scale. The network in the $1/8$-scale is adopted to transfer the features from the downsampling stage to the upsampling stage. The upsampling stage differs from the downsampling stage in that it has an ESPC that upsamples the feature map. The feature fusion in the minimum scale ($1/8$-scale) is performed by adding the input and output features of the current scale element-wise. The upsampled feature maps first forward through two residual blocks, and then the output feature maps in the right half are fused with the ones from the left half of the downsampling stage in the same scale. The skip connections and the feature fusion structure help preserve the structured regions of the input infrared image and recover the spatial information lost during the downsampling.

C. Training and Testing of Network

In this section, we briefly describe the generation of the training data pairs, which consist of the input image with the UAV target and the reference image (i.e., residual image) without the UAV target. We collect twenty infrared sequences without/with small UAV targets under different complex backgrounds using two infrared imaging devices. For real infrared sequences without the UAV targets, we construct the training pairs by embedding the simulated small targets with different scales into ten real infrared background sequences. The number of small targets is 200 and each small target whose total spatial extent is less than 80 pixels. For real infrared sequences with the UAV targets, the key is how to generate the corresponding image that does not contain small UAV targets as a reference. Because the small UAV target occupies a few pixels and the local background pixels around the target are very similar, we use the adjacent background region pixels around the target to replace the target pixels. In our experiments, we find that the training dataset obtained by this method can achieve a high detection accuracy and a low false alarm rate, especially for infrared images with heavy background clutter and noise. The total number of the training pairs generated by the above two methods is 26 000. Six real infrared UAV sequences that have no intersection with the training set are used as the test set. The representative images of six test sequences and the details on how the training data is constructed can be found in supplementary materials.

III. Experimental Results and Analysis

A. Experiment Setup

In order to show the effectiveness of the proposed method, we compare our method with several state-of-the-art baseline methods: infrared patch image (IPI) [4], reweighted infrared patch tensor (RIPT) [5], partial sum of tensor nuclear norm (PSTNN) [7], and DnCNN [11]. We call the proposed DRUNet. For an objective evaluation, the detection probability P_d and false alarm rate F_a are adopted and defined as follows:

$$P_d = \frac{\#\ \text{true targets detected}}{\#\ \text{total true targets}} \tag{1}$$

$$F_a = \frac{\#\ \text{false pixels detected}}{\#\ \text{total pixels in images}}. \tag{2}$$

Other metrics, including the signal-to-clutter ratio gain (SCRG) [13], background suppression factor (BSF) [13], contrast gain (CG) [13], and intersection over union (IoU), are given in the supplementary material.

In this letter, the proposed method and DnCNN are tested using Python on a personal computer with a 2.10 GHz Intel Xeon E5-2620 V4 CPU and 128 GB memory. We use the Adam algorithm with mini-batch size 8 for optimization. The model is trained from scratch and the whole training process ends in 300 epochs. The initial learning rate is set to 10^{-5} and the learning rate will be multiplied by 0.1 at the 240th epoch. Other methods are tested using MATLAB on the same computer.

B. Comparisons With the Baseline Methods

We compare our method with the baseline methods in terms of background suppression and target detection. Table I lists the average SCRG, BSF, CG, and IoU values of five methods for six real infrared UAV sequences. Obviously, the proposed method consistently achieves the highest values for the four metrics. For each measure of SCRG, BSF, CG, and IoU, a higher score indicates better performance. Here, inf (i.e.,

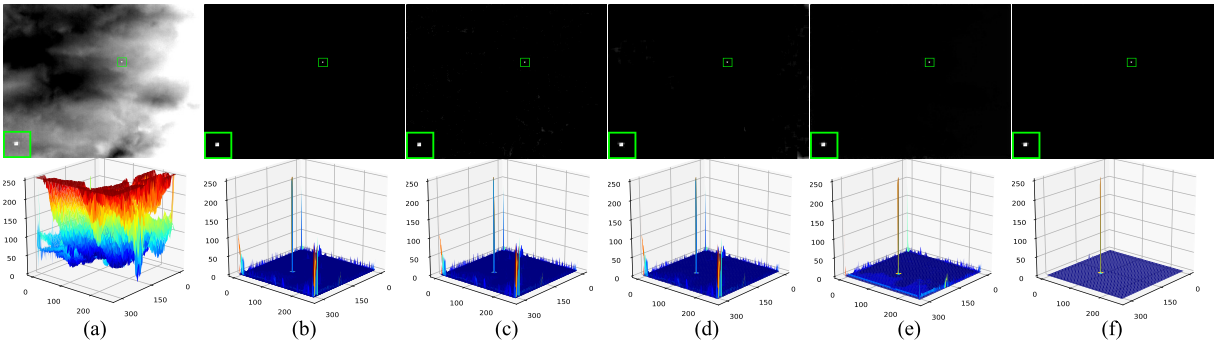

Fig. 2. Detected results of the five methods for Sequence 3 and the close-ups of the small UAV target for the detected result. (a) 66th frame of Sequence 3. (b)–(f) Detected results and the corresponding 3-D maps by the IPI, RIPT, PSTNN, DnCNN, and our method, respectively.

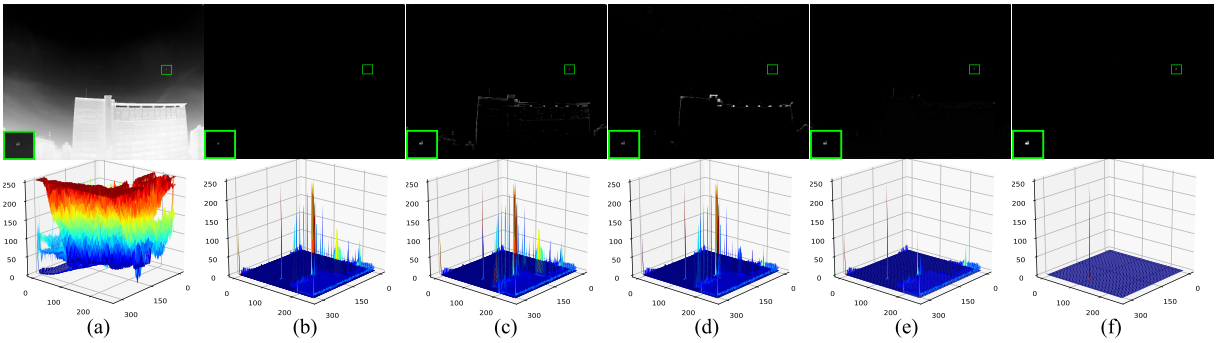

Fig. 3. Detected results of the five methods for Sequence 5 and the close-ups of the small UAV target for the detected result. (a) 147th frame of Sequence 5. (b)–(f) Detected results and the corresponding 3-D maps by the IPI, RIPT, PSTNN, DnCNN, and our method, respectively.

TABLE I

AVERAGE SCRG, BSF, CG, AND IOU VALUES OF FIVE METHODS FOR SIX TEST SEQUENCES

Methods	Indexes	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5	Seq. 6
IPI	SCRG	**inf**	**inf**	**inf**	**inf**	**inf**	**inf**
	BSF	37.14	95.68	128.88	123.36	**inf**	66.57
	CG	0.56	0.15	1.44	1.97	5.94	1.00
	IoU	0.07	0.15	0.48	0.39	0.14	0.27
RIPT	SCRG	**inf**	**inf**	21.37	25.2	**inf**	**inf**
	BSF	19.51	43.88	34.42	28.1	11.29	28.45
	CG	1.15	1.02	1.71	2.34	7.74	1.25
	IoU	0.47	0.39	0.59	0.61	0.39	0.33
PSTNN	SCRG	**inf**	**inf**	**inf**	**inf**	**inf**	**inf**
	BSF	19.75	31.98	22.14	24.92	12.14	18.43
	CG	1.1	1.3	1.84	2.57	4.78	1.37
	IoU	0.64	0.73	0.65	0.58	0.4	0.52
DnCNN	SCRG	**inf**	**inf**	**inf**	**inf**	**inf**	**inf**
	BSF	37.14	95.68	128.88	123.36	**inf**	66.57
	CG	0.56	0.15	1.44	1.97	5.94	1.00
	IoU	0.07	0.15	0.48	0.39	0.14	0.27
Ours	SCRG	**inf**	**inf**	**inf**	**inf**	**inf**	**inf**
	BSF	**167.49**	**inf**	**775.71**	**412.82**	**inf**	**351.74**
	CG	**1.68**	**1.70**	**1.87**	**2.63**	**11.18**	**1.69**
	IoU	**0.74**	**0.75**	**0.69**	**0.65**	**0.47**	**0.64**

infinity) means that the local background around the UAV is very clean. High SCRG, BSF, and CG values imply that the background clutter and noise in the small UAV images are better suppressed. High IoU indicates that the predicted box has higher localization accuracy.

Here, we present the detection results of the five methods for the 66th and the 147th frames of Sequences 3 and 5 as shown in Figs. 2 and 3, respectively. The close-ups of the small UAV target in each detection result are also given. It can be seen from Fig. 2 that, except for our method, the other four methods suffer from heavy clutter and noise residual. We can see from Fig. 3 that Sequence 5 has complex building background with strong edge structures. The IPI, RIPT, and PSTNN methods have much background residual, especially on the edges of the building, which indicates that these methods are not able to suppress the complex background well. The DnCNN method can suppress the complex background well but still has a little building residual. The complex background can almost be removed from the infrared image by our method, and finally, the detection of the small UAV targets will be very easy. That means the proposed method can learn the residual image more effectively. More experimental results are presented in the supplementary material.

Fig. 4 presents the receiver operation characteristic (ROC) curves of the five methods for six real UAV sequences on logarithmic scale. We can tell that the proposed method consistently obtains the highest detection probability in all cases regardless of the false alarm rates, indicating that our method has a better detection performance than the baseline methods. In particular, for all sequences, when the segmentation threshold reaches a certain level, the proposed method has no false alarm earlier, which again validates the effectiveness of the proposed method. The values of P_d and F_a for five methods on each test sequence can be found in the supplementary material.

C. Ablation Study

We perform an ablation study on six real sequences and report the impact of integrating the efficient subpixel

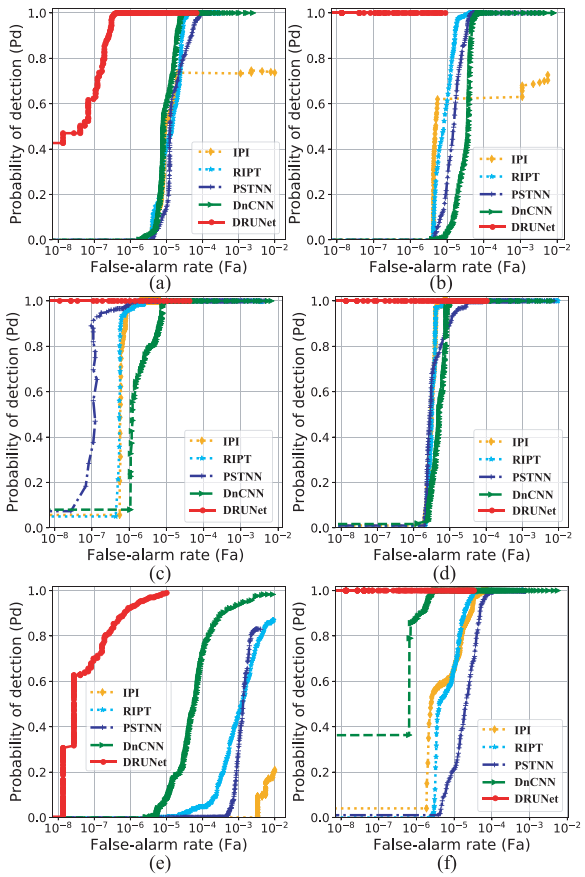

Fig. 4. ROC curves of the detection results for six real sequences in logarithmic scale. (a) Sequence 1. (b) Sequence 2. (c) Sequence 3. (d) Sequence 4. (e) Sequence 5. (f) Sequence 6.

TABLE II

IMPACT OF INTEGRATING THE ESPC, DILATED CONVOLUTION (DC), AND GRC IN THE U-NET ON OUR DATASET IN TERMS OF THE AVERAGE CG

ESPC	DC	GRC	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5	Seq. 6
-	-	-	1.53	1.68	1.83	2.6	10.85	1.61
✓	-	-	1.6	1.69	1.84	2.61	11.11	1.64
✓	✓	-	1.67	1.69	1.85	2.62	11.15	1.67
✓	✓	✓	**1.68**	**1.70**	**1.87**	**2.63**	**11.18**	**1.69**

convolution, dilated convolution, and GRC in the U-Net as shown in Table II. We can observe that integrating the ESPC can improve the performance of background suppression with the average CG value for each sequence. Furthermore, incorporating the dilated convolution once again increases the performance. Additionally, by introducing the GRC, which fuses the features from the previous layer and the output of two consecutive local residual blocks in the current layer, the overall performance is significantly improved.

IV. CONCLUSION

This letter reformulates infrared small UAV target detection into learning a nonlinear mapping from the space of infrared images to the residual image. The proposed method can learn the residual image very well with the constructed network, which proves very effective over small UAV target detection. Extensive real infrared data experiments show that the proposed method has a better detection performance under complex background clutter and noise than the baseline methods.

REFERENCES

[1] S. D. Deshpande, M. Er, V. Ronda, and P. Chan, "Max-mean and max-median filters for detection of small-targets," *Proc. SPIE*, vol. 3809, pp. 74–83, Oct. 1999.

[2] J. Han, Y. Ma, J. Huang, X. Mei, and J. Ma, "An infrared small target detecting algorithm based on human visual system," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 452–456, Mar. 2016.

[3] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 33–37, Jan. 2018.

[4] 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.

[5] 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.

[6] Y. Sun, J. Yang, Y. Long, Z. Shang, and W. An, "Infrared patch-tensor model with weighted tensor nuclear norm for small target detection in a single frame," *IEEE Access*, vol. 6, pp. 76140–76152, 2018.

[7] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, pp. 1–34, 2019.

[8] 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.

[9] 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.

[10] M. Shi and H. Wang, "Infrared dim and small target detection based on denoising autoencoder network," *Mobile Netw. Appl.*, pp. 1–15, 2019.

[11] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[12] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[13] H. Fang, M. Chen, X. Liu, and S. Yao, "Infrared small target detection with total variation and reweighted ℓ_1 regularization," *Math. Problems Eng.*, vol. 2020, Jan. 2020, Art. no. 1529704.