Infrared Small UAV Target Detection Based on Depthwise Separable Residual Dense Network and Multiscale Feature Fusion

Houzhang Fang, *Member, IEEE*, Lan Ding, Liming Wang, Yi Chang, *Member, IEEE*, Luxin Yan, *Member, IEEE*, and Jinhui Han

***Abstract*— Unmanned aerial vehicles (UAVs) have been widely applied in military and civilian fields, but they also pose great threats to restricted areas, such as densely populated areas and airports. Thermal infrared (IR) imaging technology is capable of monitoring UAVs at a long range in both day and night conditions. Therefore, the anti-UAV technology based on thermal IR imaging has attracted growing attention. However, the images acquired by IR sensors often suffer from small and dim targets, as well as heavy background clutter and noise. Conventional detection methods usually have a high false alarm rate and low detection accuracy. This article proposes a detection method that formulates the UAV detection as predicting the residual image (i.e., background, clutter, and noise) by learning the nonlinear mapping from the input image to the residual image. The UAV target image is obtained by subtracting the residual image from the input IR image. The constructed end-to-end U-shaped network exploits the depthwise separable residual dense blocks in the encoder stage to extract the abundant hierarchical features. Besides, the multiscale feature fusion and representation block is introduced to fully aggregate multiscale features from the encoder layers and intermediate connection layers at the same scale, as well as the decoder layers at different scales, to better reconstruct the residual image in the decoder stage. In addition, the global residual connection is adopted in the proposed network to provide long-distance information compensation and promote gradient backpropagation, which further improves the performance in reconstructing the image. The experimental results show that the proposed method achieves favorable detection performance in real-world IR images and outperforms other state-of-the-art methods in terms of quantitative and qualitative evaluation metrics.**

***Index Terms*— Feature fusion, infrared (IR) small target, residual image prediction, target detection, unmanned aerial vehicle.**

Manuscript received 21 May 2022; revised 20 July 2022; accepted 1 August 2022. Date of publication 11 August 2022; date of current version 25 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41501371 and Grant 61971460, in part by the Open Research Fund of the National Key Laboratory of Science and Technology on Multispectral Information Processing under Grant 6142113190103, and in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JM4036. The Associate Editor coordinating the review process was Dr. Hongrui Wang. *(Corresponding author: Liming Wang.)*

Houzhang Fang, Lan Ding, and Liming Wang are with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: houzhangfang@xidian.edu.cn; 1635126912@qq.com; wanglm@mail.xidian.edu.cn).

Yi Chang and Luxin Yan are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: owuchangyuo@gmail.com; yanluxin@hust.edu.cn).

Jinhui Han is with the College of Physics and Telecommunication Engineering, Zhoukou Normal University, Zhoukou 466001, China (e-mail: hanjinhui@zknu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIM.2022.3198490, provided by the authors.

Digital Object Identifier 10.1109/TIM.2022.3198490

I. Introduction

THE unmanned aerial vehicles (UAVs) have been widely used in military and civilian fields due to their high efficiency, low cost, and high mobility. However, the pervasive use of UAVs also poses significant safety and security challenges for critical areas, such as airports, densely populated areas, and military zones [1], [2], [3], [4], [5]. In recent years, the number of incidents involving unauthorized UAVs has been growing with the increasing use of UAVs. Therefore, the monitoring of the small UAVs at a long range is an urgent demand for providing sufficient time for early warning and interdiction of UAVs.

Due to low consumption, strong reliability, and long-distance monitoring ability, thermal infrared (IR) imaging technology has been one of the most important sensing modalities for UAV surveillance. In addition, the IR imaging technology possesses the ability to work all day, which can significantly improve the effectiveness of anti-UAV systems. However, the distance between IR sensors and UAVs is often long, and UAVs are usually relatively small in size. Therefore, UAV targets generally occupy only a tiny proportion of the entire IR image with no obvious texture structures and shape features. Moreover, in many real-life scenarios, the background is complex and frequently changes. Consequently, the small UAV targets are easily whelmed by surrounding background clutter and noise and, thus, a low contrast between the target and the background, as well as a low signal-to-clutter ratio [6], [7], [8], [9], [10]. The above issues have brought great challenges to UAV detection under complex backgrounds of low altitudes. Therefore, how to effectively improve the detection performance of the anti-UAV systems based on thermal IR imaging at long distances is still an open problem and is drawing considerable attention in recent years. Fig. 1 shows three representative IR UAV small target images (upper) and corresponding 3-D surfaces (bottom). It can be seen from these images that the UAV targets have very few pixels in entire images and no obvious features. Moreover, the backgrounds in Fig. 1(b) and (c) are complex and changeable. In Fig. 1(b), due to the drastic changes in the illumination, the intensity of the background

1557-9662 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

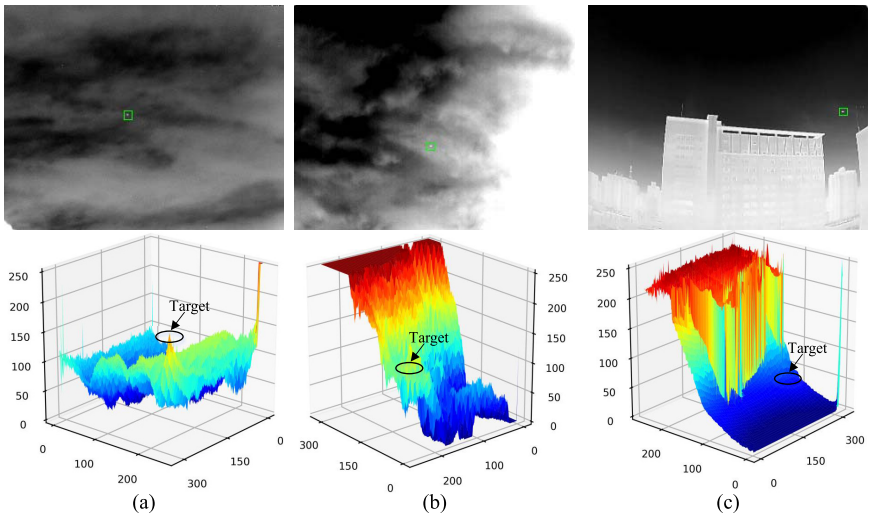


Fig. 1. Representative IR UAV small target images (upper) and corresponding 3-D surfaces (bottom). (a) Small and dim UAV target in the sky background. (b) Small UAV target in heavy cloudy and strong glare background. (c) Small UAV target in the background of the buildings.

region is significantly higher than that of the small target region, so it brings a great challenge to the detection of UAV targets.

Currently, available methods for general IR small targets are mainly classified into two categories: single-frame detection [9], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32] and multiframe sequence detection [33]. The multiframe sequence detection is time-consuming and normally necessary to meet strong prior assumptions. For instance, it is required for the background and the target to be similar in adjacent frames. However, it is hard to satisfy these assumptions due to the vibration of the sensor and the erratic movement of the target in real scenarios. By comparison, single-frame IR small target detection has the characteristics of simple and easy implementation. In addition, we can reliably identify small UAV targets on thermal images using a single-frame approach. Therefore, our research mainly focuses on single-frame IR UAV target detection.

Up until now, most existing methods for single-frame IR target detection can be divided into the following two classes: model-driven methods and data-driven methods. In the model-driven methods, such as max-mean and max-medium [34], which are used to enhance small targets and suppress background to perform small target detection, these methods are only suitable for targets that satisfy certain assumptions, such as high local contrast and strong salient. However, the above assumptions are hard to meet in real-world scenarios, leading to many false alarms. In recent years, methods based on nonlocal priors have also achieved good detection performance [11], [13], [14], [15], [16], [17], [18], [26], [35]. The main idea of these methods is to adopt the nonlocal self-similarity of the background and the sparsity of the small target in the IR image, thereby transforming the detection problem of small targets into an optimization problem of recovering the low-rank and the sparse components in the image. Nevertheless, these methods are very sensitive to complex backgrounds and random noise. As a result, they have poor detection performance for IR UAV images with complex backgrounds.

Generally, conventional model-driven methods used to detect the targets depend heavily on the handcrafted features and heuristic prior knowledge of the IR images, such as total variation regularization [17], [25], local contrast measure [18], [21], [31], and low-rank representation [11], [26]. The features adopted in these methods are only locally shallow or globally fixed. Since the deep learning methods have the ability to adaptively learn features from sample images, they can overcome the limitations caused by manual feature selection in traditional methods. Therefore, a variety of convolutional neural network (CNN)-based baseline detectors has been developed, such as Faster R-CNN [36], SSD [37], and YOLO series [38], [39], [40]. The above data-driven detectors based on CNNs have a strong generalization ability for high-resolution RGB images. However, they will still face a challenge in detecting IR small UAV targets as it is hard to fully learn the intrinsic features of small targets in complicated backgrounds with a low signal-to-clutter ratio, leading to many false alarms or missed detections.

In order to better detect IR small UAV targets in complex backgrounds, this article converts the detection of IR small UAV targets into a residual image prediction problem, in which the background, clutter, and noise are considered as the residual images between the input IR image and small UAV targets. Once the residual image is estimated, IR small UAV targets are obtained by subtracting the residual image from the input IR image. Our work is motivated by residual image learning initially proposed by Zhang *et al.* [41]. In [41], the denoising CNN (DnCNN) is utilized to predict the noise (residual) corruption with impressive performance for image denoising tasks. Nevertheless, the DnCNN faces great difficulties in recovering residual images with multiple complex structures involved in our research. In this article, we meet this challenge by developing a new

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

encoder–decoder framework that incorporates the depthwise separable residual dense blocks (DSRDBs) and the multiscale feature fusion and representation blocks (MFFRBs) into the U-Net++ [42] network. The proposed DSRDB in the encoder stage extracts abundant local level features by combining densely connected depthwise separable convolution blocks (DSCBs) with local feature fusion and local residual learning, which can enhance the representation abilities of the network. In addition, the MFFRB is integrated into the decoder stage to fully aggregate multiscale features from the encoder layers and intermediate connection layers at the same scale and the decoder layers at different scales, which can strengthen the residual image reconstruction capabilities of the model.

To sum up, the main contributions of this work are listed as follows.

1) We convert the IR small UAV target detection at a long distance into residual image prediction, making IR small UAV target detection more effective.
2) We propose a new encoder–decoder architecture. In the encoder stage, DSRDBs are employed to extract the abundant hierarchical structure features of the IR UAV images while reducing computational efforts. In addition, the network leverages MFFRBs to aggregate multiscale features in an effective fashion for reconstructing the residual image more accurately. In addition, global residual connections (GRCs) are incorporated into the network to provide information compensation, which further enhances the reconstruction performance.
3) The extensive experiments are conducted on eight real IR UAV sequences with different complicated backgrounds. The experimental results show that the proposed method outperforms the baseline methods in terms of accuracy, effectiveness, and robustness. Furthermore, the proposed method satisfies the real-time requirements in anti-UAV practical applications.

The remainder of this article is arranged as follows. Section II discusses related work on IR small object detection and the idea of residual image prediction. Section III introduces the proposed network architecture. The experimental results and analysis are presented in Section IV. In the end, we conclude the whole article in Section V.

II. Related Work

A. Infrared Small Target Detection

1) Methods Based on Background Suppression: The traditional background suppression methods, such as top-hat filtering [26], [43], maximum mean filtering, and maximum median filtering [34], suppress background clutter and noise by utilizing hand-designed spatial filters to accomplish the detection of the target. However, they only have good detection performance in simple and uniform backgrounds. It is no longer applicable when the background has strong clutter and noise because it is difficult to generate the background template image corresponding to the original IR image, resulting in the resulting image containing a lot of clutter and noise residual.

In recent years, deep learning-based background suppression methods have also received attention [44], [45], [46], [47]. Braham and Van Droogenbroeck [44] first proposed the application of CNNs for background subtraction. The CNN network receives two small patches of each pixel extracted from the input and background images, and predicts the probability of the pixel belonging to the foreground. Babaee *et al.* [45] also used a CNN-based subtraction method to estimate the background from video sequences, which includes background model generation, CNN for feature learning, and postprocessing. The CNN is adopted to extract the features from a given image–background pair and conducts background segmentation with a classifier from the extracted features. Lim and Yalim Keles [46] put forward a method based on multiscale and cascaded CNN, which uses the pretrained VGG-16 with the triple frame as the encoder. Finally, a decoder network is integrated to map the features to a pixel-level foreground probability map. Lin *et al.* [47] proposed a deep fully convolutional semantic network for background subtraction. The abovementioned background subtraction methods employed CNNs to predict the probability of a pixel belonging to the foreground and then perform threshold segmentation to obtain the target image. The performance of these methods depends heavily on the accuracy of feature extraction. The small UAV target in our problem is weak, while the background is complex and changeable. As a result, the above background subtraction methods will be prone to failure in reconstructing the background template. Our method can accurately reconstruct the background image due to the introduction of carefully designed DSRDB and MFFRB blocks, and in the final threshold segmentation, we can still obtain satisfactory detection results when the threshold is set to zero.

2) Methods Based on Target Enhancement: This type of method mainly exploits local information in the spatial domain and generates a saliency map by introducing a contrast metric to suppress the background and enhance the small target, and then, the target and the background are separated by threshold segmentation [19], [20], [21], [22], [23]. Chen and Xin [19] proposed an IR target detection method based on the visual contrast mechanism. The method first uses the weighted Gaussian difference method to calculate the target saliency map and then uses fuzzy control to optimize the parameters to obtain the saliency area. Finally, the real target is obtained by analyzing the contrast between the salient and surrounding areas. Chen and Xin [20] proposed an IR small target detection method based on a visual contrast mechanism. This method uses a top-hat filter and an adaptive threshold based on a constant false alarm rate applied to the original image to calculate the suspicious target region and then uses an effective local contrast metric to estimate the difference between the small target and the background area. Finally, a local contrast map is obtained from the original input image on the premise of combining local self-similarity and local contrast. This method not only suppresses background clutter and noise effectively but also improves the signal-to-clutter ratio of small targets. However, this type of method is always unsatisfactory in a complex background as it highlights the

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

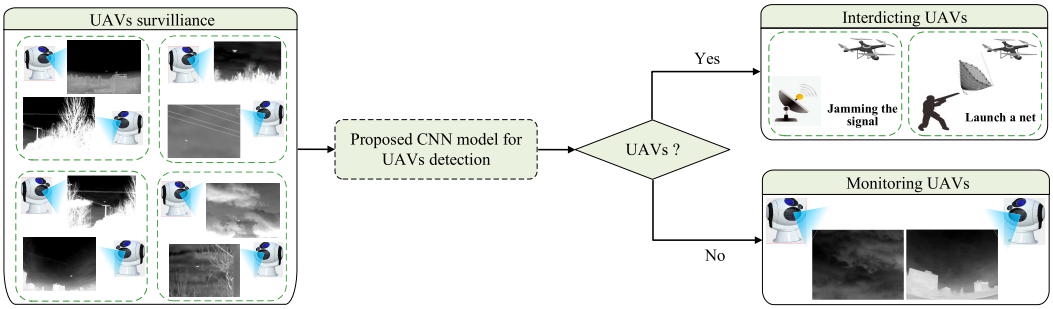


Fig. 2. Overview of the proposed system for IR UAV target detection using the proposed CNN model.

target and incorrectly enhances the high-intensity edge clutter of the background.

3) Methods Based on Target and Background Separation: Based on the nonlocal self-correlation property of the IR image and low-rank matrix decomposition theory, many models, such as the IR patch image (IPI) model [11], the reweighted IR patch tensor (RIPT) model [13], and the partial sum of tensor nuclear norm (PSTNN) model [15], have been developed by utilizing a low-rank matrix and a sparse matrix to approximately model the background and target components, respectively. Therefore, the IR small target detection is converted into an optimal solution problem of recovering low-rank and sparse components. Since the background in real IR UAV image is dynamically changing and is generally influenced by heavy clutter and noise, this means that it is difficult to accurately model the complex background with the low-rank matrix, and thus, false alarms are inevitably prone to appear. Besides, due to the high computational burden, these methods will face difficulties in meeting the real-time requirement in practical applications.

4) Object Detection Methods Based on Deep CNNs: In recent years, deep CNN-based object detection techniques have achieved great progress in vision tasks. The methods are generally divided into two classes: two-stage detectors, e.g., Faster R-CNN [36], and single-stage detectors, e.g., YOLO [40] and SSD [37]. The two-stage detectors employ a region proposal network (RPN) to produce regions of interest and then transfer the region proposals into object classification and the bounding-box regression network. The single-stage detectors estimate class probabilities and bounding box offsets from the whole feature maps. The detection performance of the above two types of methods largely depends on learning sufficient features from IR targets. Thus, it is very challenging for the CNN-based detectors to detect the small and dim UAV targets in a complicated background.

B. Residual Image Prediction

The idea of residual image learning based on DnCNN was first proposed by Zhang *et al.* [41] to solve the image denoising problem. The DnCNN network does not directly reconstruct the latent clean image but predicts the difference between the noisy observation and the clean image. The residual image is generally much easier to be predicted than the latent image. The DnCNN network implicitly removes the latent clean image from the input observed image with Gaussian noise and, finally, outputs a residual image. Shi and Wang [24] modeled IR small target detection as a denoising problem. They proposed an end-to-end denoising autoencoder network and regarded IR small targets as noise. The network [24] adopts VGG-16 as the basic network, which uses three average pooling operations for downsampling in the encoder layer, leading to too many features getting lost in the encoder stage, so it is difficult to reconstruct the high-frequency components well in the decoder stage. Overall, the current residual image learning methods still face challenges in accurately reconstructing multiple complex components of the UAV images.

III. Proposed Method

The real-time monitoring of UAV targets at long distances and in complex backgrounds is a challenging task. In this work, we propose a UAV target detection method based on residual image prediction, which can achieve favorable detection performance while reducing false alarms as much as possible. The framework of the constructed UAV detection system is overviewed in Fig. 2. The system first exploits a thermal imaging photoelectric turntable to monitor the field of view continuously, and then, the system autonomously transfers the collected online IR sequences from the imaging device to the proposed detection algorithm in real time to identify and locate the suspicious targets in IR images. If the system detects suspicious UAV targets, it will send an immediate alert to the relevant departments. Then, the relevant departments will take countermeasures to interdict it based on the location information output by the system. Otherwise, the system will continue to monitor.

Next, we will present the research work in the following two parts: 1) explicate IR small UAV target detection based on residual image prediction and 2) explain the proposed end-to-end U-shaped network in detail.

A. Infrared Small UAV Detection Based on Residual Image Prediction

In general, the IR image model can be expressed as $D = T + B + N$, where D, T, B, and N represent the observed IR image, the UAV target image, the background image, and the random noise image, respectively. The goal of residual image prediction is to learn a nonlinear mapping $\mathcal{M}(D)$ from the input image D to the residual image

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

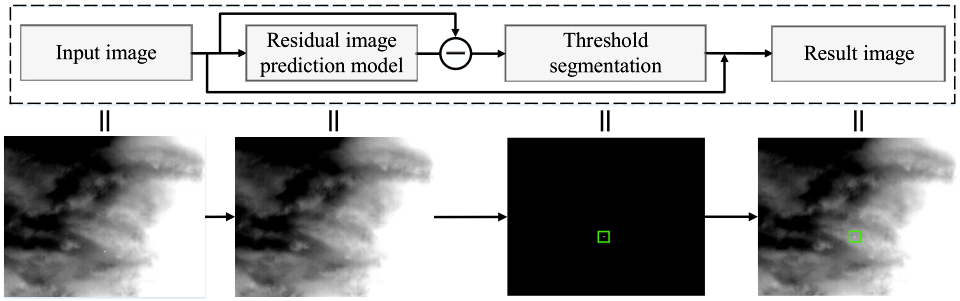


Fig. 3. Illustration of the detection process of IR UAV targets based on residual image prediction.

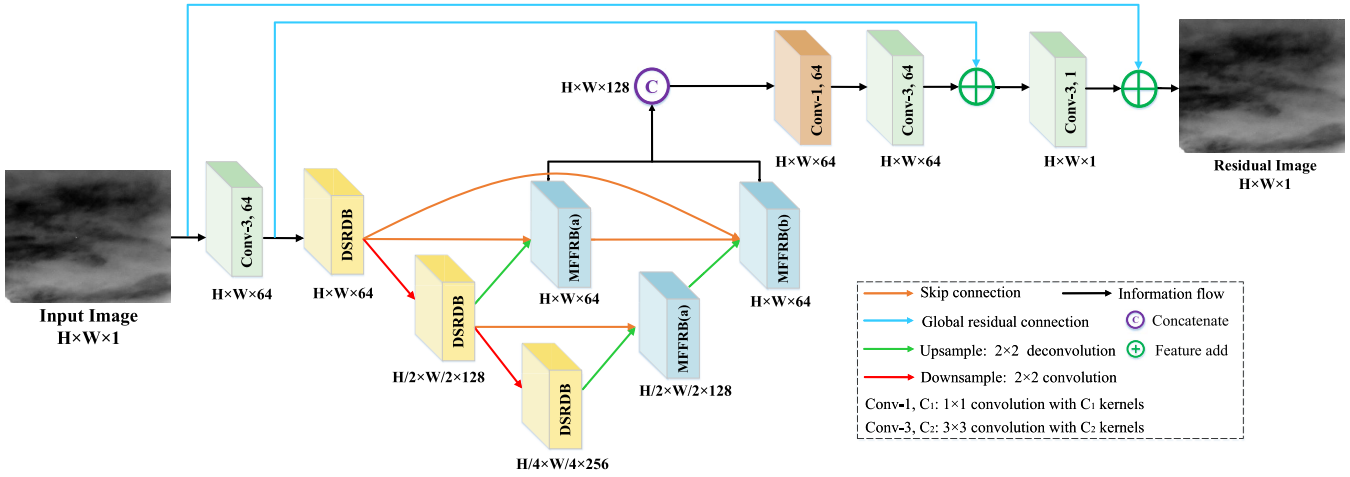


Fig. 4. Architecture of the proposed network. Each cube represents the operation to be performed on the current feature map. The number below the cube denotes the size of each feature map in a corresponding row. The operations are indicated by arrows; a description of each operation is explained at the right bottom. DSRDB denotes the depthwise separable residual dense block. MFFRB represents the multiscale feature fusion and representation block.

$D - T$. After the residual image is estimated, the UAV target is highlighted by subtracting the residual image from the input image, which is equivalent to suppressing the complex background containing clutter and noise in the IR image, as well as highlighting the UAV target. Finally, the target is extracted by the threshold segmentation, and the coordinate information is mapped back to the original image to obtain the final detection result. The above overall procedure is shown in Fig. 3. It is worth noting that, in order to highlight the effectiveness and robustness of our method, the threshold selected in our proposed method is 0. According to the formula $D = T + B + N$, when the residual image $B + N$ learned by the network is incomplete, the UAV target image obtained by subtracting the residual image $[D-(B + N)]$ from the original observation image will include a small amount of noise and clutter. When the segmentation threshold is 0, the gray value of this clutter and noise will also be set to 255 in the threshold segmentation. Therefore, in order to prove that our method can completely reconstruct the residual image, the obtained UAV target image does not contain any clutter and noise, so we set the segmentation threshold to 0.

B. Network Architecture

The proposed network architecture consists of two subnetworks: the encoder network (left part as shown in Fig. 4) and the decoder network (right part as shown in Fig. 4). The feature maps generated by the two subnetworks are connected through skip connections. The encoder stage is mainly responsible for the multiscale feature extraction of the image, and the decoder stage is in charge of reconstructing the residual image. The proposed network architecture contains a total of 63 convolutional layers. As shown in Fig. 4, the network mainly contains three scales of feature maps from input sizes 256×320 to 64×80. The feature maps of each scale are produced by DSRDBs. In addition, in the encoder stage, we use 2×2 convolution with the stride of 2 to perform downsampling and transfer the feature maps of the preceding stage to the next stage. When the size of the feature map is reduced to $H/4 \times W/4$, the encoder stage is transferred to the decoder stage. This work adopts a 2×2 deconvolution in the decoder stage with a scale of 2 to upsample the image. In order to make full use of the multiscale feature information generated in the encoder stage, the network fuses the feature map generated by the upsampling in the decoder stage with the feature map of the corresponding scale in the encoder stage through a skip connection.

Meanwhile, the network utilizes an intermediate feature aggregation layer based on a densely connected skip connection to narrow the semantic gap between the encoder and decoder layers. Finally, multiscale feature maps generated by different top-down pathways are fused in the network decoder

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

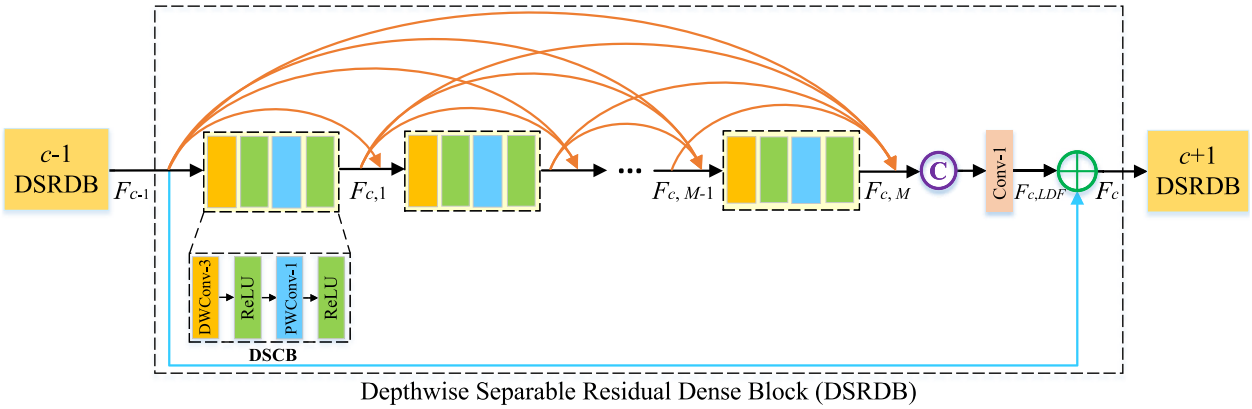


Fig. 5. Architecture of the proposed DSRDB. Each DSRDB consists of eight densely connected DSCBs. Conv-n: $n \times n$ standard convolution. DWConv-3: 3×3 depthwise convolution. PWConv-1: 1×1 pointwise convolution.

stage. The encoder–decoder networks integrate shallow fine features into deep semantic features, which is beneficial to residual image reconstruction. The long-distance information compensation of the network is provided by adopting the strategy of GRC. The skip connection and feature fusion structures can save as many structural features of the residual image as possible and help the network recover the spatial information lost during the encoder stage.

Next, we further clarify the proposed network from the encoder and decoder stages, respectively.

1) Bottom-Up Multiscale Feature Extraction Based on Residual Dense Connection: In the encoder subnetwork, the DSRDB structure, as shown in Fig. 5, consists of eight densely connected DSCBs, which is used to extract multilevel features of the image. Each DSCB is made up of two parts. The first part is to perform depthwise convolution by applying a single kernel size of 3×3 convolution kernel per input channel on the feature map. After the depthwise convolution, a pointwise convolution with a kernel size of 1×1 is used to fuse the feature map by learning linear combinations of the input channels.

If the depthwise separable convolution in the DSRDB is replaced with the ordinary convolution, the DSRDB will turn into the residual dense block (RDB). Table I gives the parameters and floating-point operations per second (FLOPs) for the RDB with the ordinary convolution and the proposed DSRDB with the deep separable convolution, respectively. We can find that the introduction of depthwise separable convolution can reduce the number of parameters and FLOPs for our model. Consequently, the proposed DSRDB can significantly alleviate the problems of high computational costs and insufficient memory. Each depthwise separable convolution in DSRDB is connected to other convolutions in the current block through dense connections. This dense connection allows the network to extract local feature information. In addition, due to the introduction of the dense connections in DSRDB, the feedforward information of other convolutional layers can also be preserved in the network so that the hierarchical features can be fully and effectively utilized. Furthermore, local residual connections in DSRDB promote the propagation of local gradient flows and further improve the learning and representation capabilities of the network.

TABLE I

PARAMETERS AND FLOPS OF RDB WITH THE ORDINARY CONVOLUTION AND DSRDB WITH THE DEEP SEPARABLE CONVOLUTION

Block	Parameters (K)	FLOPs (G)
RDB	1572.864	44.336
DSRDB	954.048	29.991

Our DSRDB contains local dense connections with DSCBs, local feature fusion, and local residual learning.

a) Local dense connections with DSCBs: To efficiently reduce the computational costs and save memory, DSCBs are integrated into the network. As shown in Fig. 5, the output of the ith convolutional layer can be written as

$$F_{c,1} = \psi(F_{c-1}) \tag{1}$$

$$F_{c,i} = \psi\left(\left[F_{c-1}, F_{c,1}, \ldots, F_{c,i-1}\right]\right), \quad 1 \le i \le M \tag{2}$$

where $F_{c,i}$ denotes the features of the ith layer in the cth DSRDB. Let F_{c-1} and F_c be the input and output feature maps of the cth DSRDB, respectively. $\psi(\cdot)$ represents the operation function used in the convolutional layer. $[F_{c-1}, F_{c,2}, \ldots, F_{c,i-1}]$ implies the concatenation of the feature maps from the convolutional layers $1, 2, \ldots, i-1$ in the cth DSRDB and the $c-1$th DSRDB. M represents the number of DSCBs in the cth DSRDB.

b) Local feature fusion: It concatenates features from the previous and current DSRDBs. In addition, a 1×1 convolutional layer is introduced to adaptively extract local dense features from concatenated features and reduce the channels of the fused feature maps. The local feature fusion is written as

$$F_{c,\mathrm{LDF}} = \phi\left(\left[F_{c-1}, F_{c,1}, \ldots, F_{c,M}\right]\right) \tag{3}$$

where $\phi(\cdot)$ represents the function of the 1×1 convolutional layer.

c) Local residual learning: To further enhance the representation learning capability of the network, local residual learning is incorporated by adding the input features to fused features. The resulting output of the cth DSRDB can be written as

$$F_c = F_{c-1} + F_{c,\mathrm{LDF}}. \tag{4}$$

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

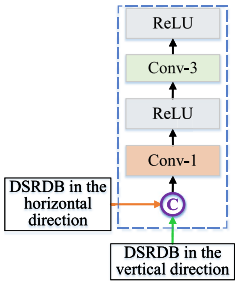


Fig. 6. MFFRB(a).

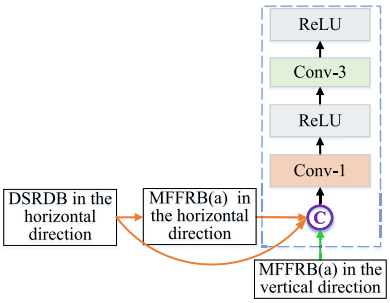


Fig. 7. MFFRB(b).

Local residual learning improves the information flow and, finally, produces better performance.

2) Top-Down Multiscale Feature Fusion and Representation for Residual Image Reconstruction: The encoder–decoder networks have been proven to be effective in reconstructing fine-grained details of the image on a complicated background, and the advantage can be attributed to the fact that the networks fully fuse low-level, fine-grained features from the encoder subnetwork and semantic, coarse-grained features from the decoder subnetwork by skip connections. Fusing the same-scale feature maps directly from the encoder and decoder stages often suffers from being semantically dissimilar. In order to reduce the semantic gap between the features from the encoder and decoder networks, an intermediate aggregation layer is added between the encoder and decoder subnetworks by a densely connected skip connection. In the decoder subnetworks, we utilize MFFRBs to fuse the multiscale features in horizontal and vertical directions from the preceding layers. The MFFRB in our network architecture can be divided into two parts MFFRB(a) and MFFRB(b), as shown in Figs. 6 and 7. After the feature fusion for the above each block, the 1×1 convolution is used to reduce the number of channels of the fused feature maps, and the 3×3 convolution is employed again to further aggregate the fused features.

In MFFRB(a), the upsampled feature maps from the vertical DSRDB are integrated with the same-scale feature maps in the horizontal DSRDB in a concatenation fashion. In MFFRB(b), the MFFRB block fuses features in three directions, i.e., the upsampled MFFRB(a) in the vertical direction, and the DSRDB and the intermediate aggregated MFFRB(a) in the horizontal direction.

The shallow features contain rich spatial location information, and the high-level features offer semantic information. Therefore, it is of vital importance to make full use of the features of different depths of the network for accurate residual image reconstruction. The MFFRB in the decoder subnetwork can help the network recover fine-grained details of the residual image. Furthermore, the feature information of different scales in each layer of the decoder stage is thoroughly utilized by using MFFRB, which further strengthens the feature representation and background reconstruction capabilities of the model.

3) Loss Function: Since the ℓ_1 loss is robust to outliers, we adopt ℓ_1 loss as the objective function, which is defined as

$$L_1(\widehat{y}, y) = \sum_{i=0}^{N} \left| y^{(i)} - \widehat{y}^{(i)} \right| \tag{5}$$

where N is the batch size, and $y^{(i)}$ and $\hat{y}^{(i)}$ are the real IR image without UAV targets and the residual image, respectively.

IV. Experimental Results and Analysis

In this section, we first introduce the experimental settings, including dataset preparation and evaluation metrics. Then, the analysis of hyperparameters of the network is illustrated. The qualitative and quantitative results in comparison with other popular detection methods are exhibited to demonstrate the effectiveness and robustness of the proposed method. Finally, we conduct extensive ablation studies to investigate the contributions of different components of the proposed network.

A. Experimental Settings

In our experiments, the proposed method and the DnCNN are conducted on a server with an NVIDIA RTX 2080Ti graphics card and a computer with a CPU 2.10-GHz Intel Xeon E5-2620 V4 CPU processor. Both devices have 128 GB of memory. Other methods are implemented in MATLAB (X64) R2017b software. We use the Adam algorithm for optimization. The minibatch size is 4 for training. The model is trained from scratch, and the training process ends up with 300 epochs. The initial learning rate is set to 10^{-5} and will be multiplied by 0.1 at the 240th epoch.

B. Data Preparation

In this section, we depict the generation of the training data pairs that are composed of the input image with the UAV target and the reference image (i.e., residual image) without the UAV target. We collect 18 IR sequences with/without small UAV targets in different backgrounds using two thermal IR imaging devices. We utilize ten real IR UAV image sequences to generate the training set, a total of 26 000 pairs of training data. Eight real IR UAV image sequences with small UAV targets that have no intersection with the training set are adopted as the testing set, a total of 2400 IR images.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

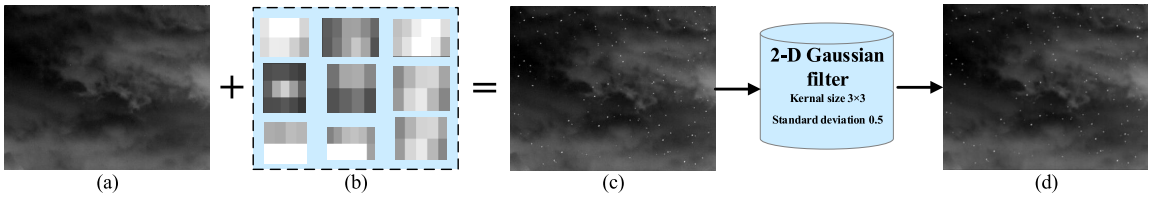


Fig. 8. Flowchart of generating input image for training. (a) Reference image without UAV targets. (b) Simulated small targets. (c) Intermediate image. (d) Input image required for network training.

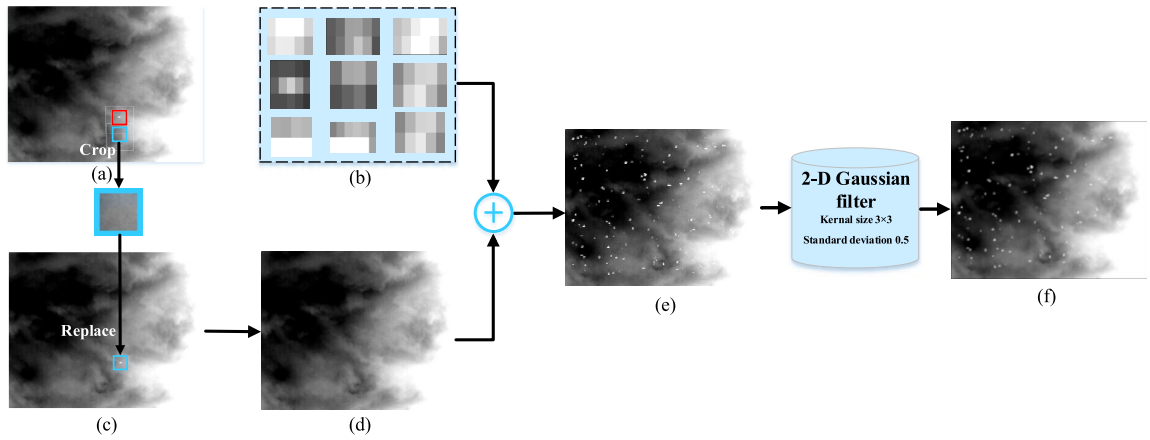


Fig. 9. Flowchart of generating reference image and input image for training. (a) Original image with UAV targets. (b) Simulated small targets. (c) Process of replacing the pixels of small UAV target using adjacent neighboring pixels around the UAV target. (d) Reference image without UAV targets. (e) Intermediate image. (f) Input image required for network training.

1) Construction of Training Data Pairs: Two approaches are employed to construct training data pairs according to whether the IR image contains the UAV targets in the real scenes and are listed as follows.

1) For the real IR image without UAV targets, also called the reference image. In the light of the characteristics of the UAV targets in the real scene, we first simulate 200 IR small targets with different gray levels and scales. According to the definition of IR small target by the International Society for Optical Engineering (SPIE), a small target is a target whose spatial pixel is less than 80 pixels of an IR image, The area of the UAV target that we simulated in our dataset ranges from 9 (about 3×3) to 30 (5×6) of a 256×320 image, which conforms to the definition of the small target in SPIE. Then, the intermediate image is constructed by incorporating simulated IR small targets randomly into the reference image. Finally, we use a 2-D Gaussian filter with a kernel size of 3×3 and a standard deviation of 0.5 to smooth the intermediate image to obtain the input image required for network training. The construction process is shown in Fig. 8.
2) For the real IR images with UAV targets, also called the original images. How to obtain an image without a UAV target as a reference image is a key to generating training data pairs. Because the IR radiation energy of the neighboring background around the target is lower than the target region, and there is a gray value transition between the target region and the surrounding background region. Thus, we first adopt the adjacent neighboring pixels around the UAV target to replace the pixels of the small UAV target to obtain a reference image. It is important to note that we must ensure that the small target is completely covered by the background. When the reference images are obtained, we embed 200 simulated targets into the reference images to construct the intermediate images. To acquire the input image, we also used the same Gaussian filter as mentioned above to smooth the intermediate image. The construction process is shown in Fig. 9.

2) Characteristics of Testing Data: Our method and the four baseline methods are tested on eight real IR UAV small target sequences with various backgrounds. Each test sequence has groups of 300 frames with the size of 256×320. The groups of 300 frames that we used in the experiment were randomly selected from a long sequence, so the 300 frames in each sequence are not continuous, and the randomly selected 300 frames are guaranteed to represent and characterize each test sequence.

The detailed description of eight real IR image sequences is shown in Table II. We also summarize the characteristics of eight sequences as follows. In sequence 1, the target is somewhat brighter than the background that is made up of scattered clouds, making it relatively simpler than other sequences. In sequences 2–4, the background has strong cloud and illumination interference, and the UAV target is brighter than the background. In sequences 5–7, the target is relatively dim, slightly brighter than the background, and the pixels

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

occupied by the target are about 2–4. Thus, the UAV targets are easily drowned in the complex background. Compared with other sequences, the background of sequence 8 is relatively complex with strong illumination variation and dense clouds. The complexity and change of the background greatly increase the difficulty of small UAV target detection.

C. Evaluation Metrics

In order to assess the target detection, background suppression, and UAV target enhancement capability of different methods, we utilize the following metrics to quantitatively evaluate the detection methods.

1) Precision (Prcn):

$$\text{Prcn} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

where the true positive (TP) is the number of correctly detected UAVs; the false positive (FP) is the number of falsely detected UAVs. It can be seen from the definition of precision that it reflects the proportion of a model that predicts correctly. Precision in our application represents the proportion of real UAVs to all identified UAVs.

2) Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

where the false negative (FN) is the number of missed detections, in which case the UAVs are not detected by the model. The definition of TP here is the same as the TP in the definition of Prcn. The recall is the proportion of targets that the model predicts correctly to all real targets. In our application, it denotes the ratio of the number of correctly identified UAVs to the number of all real UAVs in the test set.

3) F1-Score (F1): F1 is essentially a harmonic average of recall and precision, and we can use it to balance precision and recall. It can be defined as

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

4) Contrast Gain (CG): For target enhancement and background suppression, we introduce the indicator CG, which is defined as

$$\text{CG} = \frac{\text{CON}_{\text{out}}}{\text{CON}_{\text{in}}} \tag{9}$$

where CON_{out} and CON_{in} denote the contrast between the target and background of the processed output image and the original input image, respectively. The calculation of CON can be defined as

$$\text{CON} = |\mu_t - \mu_b|. \tag{10}$$

As shown in Fig. 10, we assume that the size of the UAV target is $a \times b$. Then, the background area of the UAV target is $(a + 2d) \times (b + 2d)$. In our experiment, we set $d = 20$ pixels. In the formula, μ_t and μ_b represent the average pixel value of the target region and that of the neighboring background area, respectively.

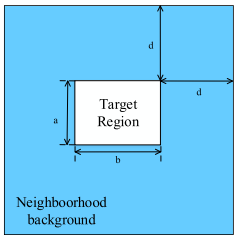


Fig. 10. External rectangle and neighboring background rectangle of a small target.

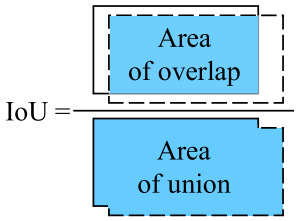


Fig. 11. Illustration of IoU calculation.

5) Probabilities of Detection (P_d) and False Alarm Rate (F_a): In the field of target detection, P_d and F_a are two commonly used evaluation metrics. P_d and F_a are described as

$$P_d = \frac{\text{Number of true targets detected}}{\text{Number of total true targets}} \tag{11}$$

$$F_a = \frac{\text{Number of detected false targets}}{\text{Number of image frames}}. \tag{12}$$

6) Intersection over Union (IoU): For a specific object, we can evaluate the localization accuracy by how much the prediction box and the ground-truth box overlap, and we usually use IoU to quantify the degree of overlap, which is defined as

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}}. \tag{13}$$

The schematic of (13) is shown in Fig. 11.

7) Receiver Operating Characteristic (ROC) Curve: The ROC curve is a method for dynamically evaluating the probability of detection and the false alarm rate of the model. It performs threshold segmentation on the detection results under different thresholds and then finds the positive sample detection frame and the negative sample detection frame.

D. Hyperparameters Analysis

The number of convolutional layers in DSRDB and the downsampling rate of the network are two important parameters and have a great impact on the detection accuracy and efficiency of the proposed method. In this part, we analyze the influence of two parameters on the performance of our method and discuss the selection of proper parameters.

1) Discussion on the Optimal Number of Convolution Layers in DSRDB: In the encoder stage of the network, the densely connected convolutional layer in the DSRDB is mainly responsible for extracting a large amount of local feature information. The number of convolutional layers in

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DETAILS OF EIGHT REAL TESTING SEQUENCES

Sequence	Target type	Range of target size	Target details	Background details
Seq. 1	DJI Phantom 4, Pro V2.0	3×3 to 4×4	A dim target, a long imaging distance, move irregularly between clouds, slightly brighter than the background	Less cloudy, sparse clouds defective pixels
Seq. 2	DJI Phantom 4, Pro V2.0	2×3 to 3×4	A long imaging distance, brighter than the background, move left and right among the clouds	Heavy cloudy, defective pixels
Seq. 3	DJI Mini2	2×2 to 3×4	Brighter than the background, move up and down in the clouds	Heavy cloudy and clutter backgrounds, strong illumination variations
Seq. 4	DJI Mini2	2×3 to 3×3	Brighter than the background, move up and down between clouds	Heavy cloudy, strong illumination variations
Seq. 5	DJI Mini2	1×2 to 2×2	Dim, slightly brighter than the background, move left and right above the buildings, a small size with a little change.	Heavy smog, building background
Seq. 6	DJI Phantom 4, Pro V2.0	1×2 to 2×2	Brighter than the background, a small size with a little change, move up and down between buildings	Heavy smog, building background, strong illumination variations
Seq. 7	DJI Mini2	1×2 to 2×2	Slightly brighter than the background, the gray value of target is similiar to the background in a few frames, move up and down between buildings	Heavy smog, building background
Seq. 8	DJI Phantom 4, Pro V2.0	2×2 to 3×3	Brighter than the background, a long imaging distance	Heavy cloudy, dense clouds, heavy smog, strong illumination variations

TABLE III
F1 AND CG VALUES OF THE PROPOSED MODEL UNDER DIFFERENT NUMBERS OF CONVOLUTION LAYERS IN THE DSRDB FOR EIGHT SEQUENCES

Number of convolution layers	Indexes	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5	Seq. 6	Seq. 7	Seq. 8
2	F1	0.6195	0.7914	0.9859	0.7806	0.4599	0.8766	0.4523	0.4974
	CG	1.60	1.65	1.81	2.60	11.12	1.69	17.68	2.28
4	F1	0.6214	0.7963	0.9885	0.7808	0.4834	0.8878	0.4704	0.5188
	CG	1.64	1.67	1.84	2.62	11.19	**1.69**	17.70	2.27
8	F1	**0.6272**	**0.8152**	**0.9922**	**0.7880**	**0.4890**	**0.9297**	**0.4822**	**0.5292**
	CG	1.68	**1.68**	**1.85**	**2.67**	**11.44**	**1.69**	**17.75**	**2.35**
16	F1	0.6017	0.8029	0.9773	0.7783	0.4608	0.8981	0.4693	0.5174
	CG	**1.69**	1.67	1.83	2.65	11.19	**1.69**	17.38	2.25

DSRDB has an influence on the receptive field of the layers at different depths, and then, this will influence the ability of the network to represent hierarchical features. Therefore, we need to explore the impact of the number of convolutional layers in DSRDB on the feature representation capability. Since the F1 comprehensively considers the precision and recall of a model, we utilize the F1 to measure the detection performance of the model. In addition, we also employ the CG to assess the model's capabilities to enhance the target and suppress the background. Table III shows the F1 and CG values of the proposed model under different numbers of convolutional layers for eight real IR sequences. Fig. 12 shows the evolution curve of the average CG and F1 versus the number of different convolutional layers for the proposed method. The minimum CG value of the proposed method for sequences 5 and 7 is greater than the maximum CG value of the other six sequences; for visual convenience, we draw

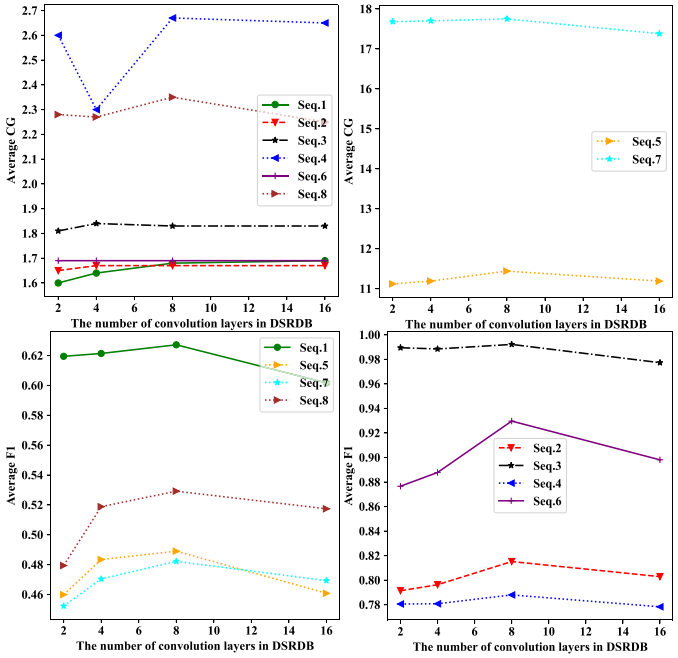


Fig. 12. Evolution curve of the average CG and F1 versus the number of different convolutional layers in DSRDB for the proposed method.

the curves of sequences 5 and 7 separately. The F1 values of sequences 1, 5, 7, and 8 are quite similar, as are the F1 scores of sequences 2–4 and 6; we separate the evolution curve of F1 versus the varied number of convolutions in DSRDB into two subgraphs to more intuitively reflect the trend of F1 with the number of convolutional layers in DSRDB.

It can be seen from Table III and Fig. 12 that, with the increase in the number of convolutional layers, the average

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

NETWORK PARAMETERS, FLOPS, AND TEST TIME OF THE PROPOSED MODEL FOR DIFFERENT NUMBERS OF CONVOLUTIONAL LAYERS IN THE DSRDB

Number of convolution layers	Parameters (K)	FLOP (s)	Time (s)
2	807.168	26.460	0.019
4	850.136	27.458	0.022
8	954.048	29.991	0.027
16	1248.280	37.202	0.147

TABLE V

F1 AND CG VALUES OF THE PROPOSED MODEL UNDER DIFFERENT DOWNSAMPLING RATES FOR EIGHT SEQUENCES

Down-sampling rate	Indexes	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5	Seq. 6	Seq. 7	Seq. 8
2	F1	0.5977	0.7669	0.9892	0.7862	0.4357	0.9048	0.4519	0.4993
	CG	1.65	1.67	1.82	2.67	11.00	**1.69**	0.96	0.51
4	F1	**0.6272**	**0.8152**	**0.9922**	**0.7880**	**0.4890**	**0.9297**	**0.4822**	**0.5292**
	CG	**1.68**	**1.68**	**1.85**	**2.67**	**11.44**	**1.69**	**17.75**	**2.35**
8	F1	0.6110	0.8040	0.9883	0.7807	0.4721	0.9091	0.4764	0.5032
	CG	1.64	1.66	1.80	2.59	11.23	**1.69**	17.69	2.32

TABLE VI

PARAMETERS, FLOPS, AND THE TEST TIME OF DIFFERENT DOWNSAMPLING RATES

Down-sampling rate	Parameters (K)	FLOP (s)	Time (s)
2	261.696	15.044	0.024
4	954.048	29.991	0.027
8	3675.136	51.720	0.034

F1 and CG values increase on the whole. The average CG and F1 values begin to stabilize or decrease when the number of convolutional layers is greater than 8. On the one hand, the change of F1 implies that, as the number of convolutional layers increases, the receptive field will gradually increase, and the detection performance of the proposed method can be enhanced. On the other hand, as the number of convolutional layers in each DSRDB increases to a certain level, the network can extract rich local structural features. Therefore, it can be observed from the changes of CG that, when the number of convolutional layers in DSRDB is 8, our model can reconstruct the residual image well, and the ability of background suppression and target enhancement reaches the optimum. However, too many local features in each DSRDB do not necessarily lead to improved performance and even decrease the performance in some cases because the global structure features are also important in completely reconstructing the residual image.

Table IV shows the network parameters, floating points of operations, and average computation time (in seconds) for a real IR sequence in our test set corresponding to the number of different convolutional layers. It can be seen from Tables III and IV that the proposed network achieves the best detection performance in terms of accuracy and efficiency when the number of the convolutional layers is 8.

2) Discussion on the Optimal Downsampling Rate of the Network: In the encoder stage, the network adjusts the size of the feature map through downsampling operations

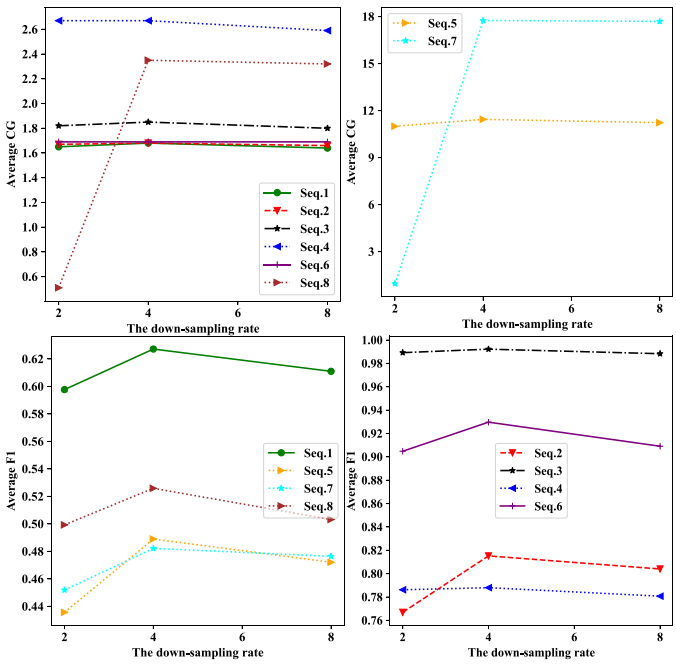


Fig. 13. Evolution curve of the average CG and F1 versus the downsampling rate for the proposed network.

to enlarge the receptive field. We aim at extracting and representing the important structure information. Too many downsampling operations will result in the loss of too many details. Therefore, the selection of the downsampling rate affects the detection performance of the network. Table V also presents the F1 and CG values of eight real IR sequences under different downsampling rates. Table VI shows the network parameters, FLOPs, and average computation time (in seconds) for a real IR sequence in our test set corresponding to the different downsampling rates. Fig. 13 shows the evolution curve of the average F1 and CG versus the downsampling rates for the proposed method. As the minimum CG value of the proposed method for sequences 5 and 7 is greater than the maximum CG value of the other six sequences, for visual convenience, we draw the curves of sequences 5 and 7 separately. The F1 values of sequences 1, 5, 7, and 8 are quite similar, as are the F1 scores of sequences 2–4 and 6. In order to more intuitively reflect the trend of F1 with the different downsampling rates, we divide the evolution curve of the F1 versus the different downsampling rates into two subgraphs.

It can be observed that a larger downsampling rate contributes to better performance when the downsampling rate is less than 4. However, the network will suffer some performance drop when further increasing the downsampling rate. The reason for this phenomenon is that too much downsampling will cause the loss of some details, which is not conducive to the detection of small-scale targets. It can be seen from Tables V and VI that F1 and CG values are optimal when the downsampling rate is 4, which implies that the proposed network achieves the best tradeoff between accuracy and efficiency at a downsampling rate of 4. Therefore, we use this value as the final downsampling rate of our model in all the other experiments.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
AVERAGE PRCN, RECALL, F1, CG, AND IOU VALUES OF SIX METHODS FOR EIGHT TEST SEQUENCES

Methods	Indexes	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5	Seq. 6	Seq. 7	Seq. 8
IPI	Prcn	0.0881	0.5465	0.8835	0.6634	0.0568	0.4120	0.0159	0.3186
	Recall	0.2322	0.0946	0.8622	0.6458	0.0656	0.0656	0.0286	0.1730
	F1	0.1278	0.5565	0.8847	0.6545	0.0610	0.5445	0.0204	0.3175
	CG	0.56	0.15	1.44	1.97	5.94	1.00	6.63	1.36
	IoU	0.07	0.15	0.48	0.39	0.14	0.27	0.40	0.24
RIPT	Prcn	0.0761	0.7332	0.7291	0.5354	0.0056	0.1901	0.0017	0.3250
	Recall	0.5165	0.6872	0.9083	0.7115	0.2991	0.7949	0.0996	0.4580
	F1	0.1327	0.7549	0.8174	0.6121	0.0110	0.3070	0.0034	0.3992
	CG	1.15	1.02	1.71	2.34	7.74	1.25	8.71	1.82
	IoU	0.47	0.39	0.59	0.61	0.39	0.33	0.22	0.41
PSTNN	Prcn	0.0763	0.4332	0.4551	0.3087	0.0042	0.0723	0.0002	0.0843
	Recall	0.4598	0.6362	0.9407	0.7215	0.1791	0.8404	0.0136	0.4993
	F1	0.1309	0.4675	0.6142	0.4329	0.0081	0.1331	0.0005	0.1453
	CG	1.10	1.30	1.84	2.57	4.78	1.37	2.89	**2.56**
	IoU	0.64	0.73	0.65	0.58	0.4	0.52	0.07	0.26
DnCNN	Prcn	0.1217	0.4926	0.5460	0.3006	0.0423	0.3725	0.0045	0.1295
	Recall	0.5801	0.4541	0.9188	0.7198	0.3253	0.9778	0.9778	0.4448
	F1	0.2012	0.2012	0.6860	0.4246	0.0749	0.5401	0.0086	0.2060
	CG	0.56	0.15	1.44	1.97	5.94	1.00	9.33	2.11
	IoU	0.07	0.15	0.48	0.39	0.14	0.27	0.47	0.52
CDAE	Prcn	0.0013	0.0021	0.1131	0.0036	0.0002	0.0008	0.0005	0.0670
	Recall	0.1251	0.2679	0.5177	0.1628	0.0177	0.1401	0.0428	0.4321
	F1	0.0026	0.0041	0.1857	0.0070	0.0001	0.0015	0.0009	0.1160
	CG	0.54	0.72	0.44	1.01	1.42	0.29	1.59	1.50
	IoU	0.12	0.11	0.14	0.11	0.21	0.09	0.17	0.20
Ours	Prcn	**0.6173**	**0.7399**	**0.9884**	**0.7866**	**0.4898**	**0.8696**	**0.4685**	**0.4951**
	Recall	**0.6374**	**0.9077**	**0.9960**	**0.7894**	**0.4882**	**0.9987**	**0.4968**	**0.5685**
	F1	**0.6272**	**0.8152**	**0.9922**	**0.7880**	**0.4890**	**0.9297**	**0.4822**	**0.5292**
	CG	**1.68**	**1.68**	**1.85**	**2.67**	**11.44**	**1.69**	**17.75**	2.35
	IoU	**0.78**	**0.75**	**0.67**	**0.64**	**0.44**	**0.64**	**0.59**	**0.71**

E. Comparisons With Other Baseline Methods

1) Baselines: To verify the effectiveness of the proposed approach, we compare our method with other baseline methods: IPI [11], RIPT [13], PSTNN [15], DnCNN [41], and CDAE [24].

We choose the IPI, RIPT, and PSTNN as the baseline methods for the following reasons: 1) they are generally regarded as original representative methods in the IR target detection field; 2) they also obtain the competitive performance in separating the IR small target from the complex sky background; and 3) the source codes of these methods are available to us, and thus, we can make a relatively fair comparison based on the original implementation. We choose the DnCNN and CDAE as the baseline methods because they are not only deep-learning-based but also residual learning-based methods and are highly relevant to our work.

2) Quantitative Evaluation: Table VII shows the average Prcn, Recall, F1, CG, and IoU values of the six detection methods on the eight real IR test sequences. It can be concluded from Table VII that the proposed method achieves the highest values in all test sequences in terms of Prcn, Recall, and F1, which means that our method has the best detection performance. It is shown from Table VII that the proposed method can highlight small targets and suppress background clutter effectively. In general, the higher the value of CG, the stronger the model's ability to enhance small targets and suppress background clutters. The higher the IoU value, the closer the predicted frame to the actual frame, and the more precise the small target location. We can observe from Table VII that our method obtains the highest IoU values, indicating that our method can accurately localize the small UAV targets.

Especially, the CG values of the proposed method for eight sequences are the highest, which suggests that the proposed network model can accurately predict the residual image, and then, IR small UAV target image can be extracted very well.

3) Qualitative Evaluation: The detection effectiveness of six methods is verified on eight real IR sequences. Figs. 14–17 show the detection results and 3-D gray maps of six methods on four representative sequences. The results of the other four sequences can be found in the Supplementary Material. The small targets of the result images are marked by green boxes, and the close-ups of the small targets are also given at the

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

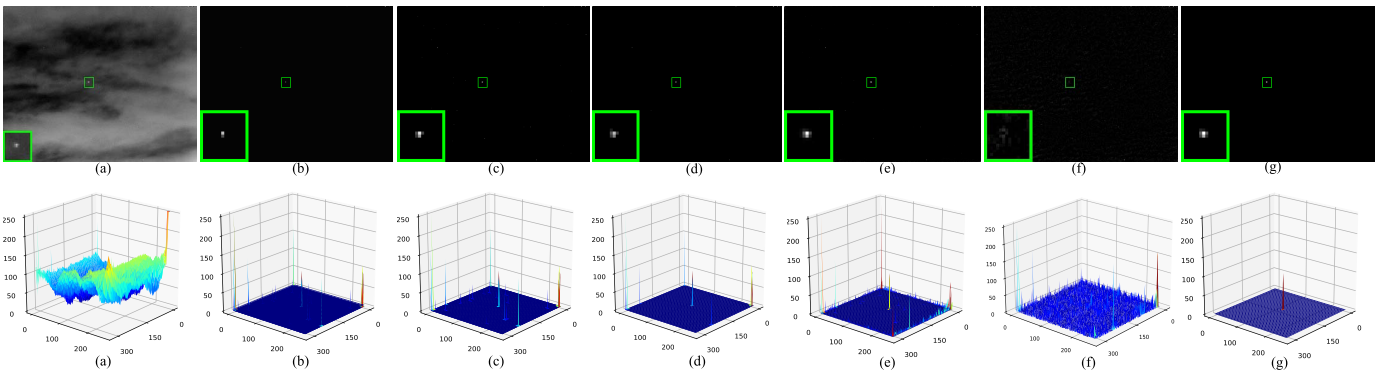


Fig. 14. Detected results of the six methods for Seq. 1 and the close-ups of the small UAV target for the detected result. (a) 66th frame of Seq. 1. (b)–(g) Detected results and the corresponding 3-D maps by the IPI, RIPT, PSTNN, DnCNN, CDAE, and our method, respectively.

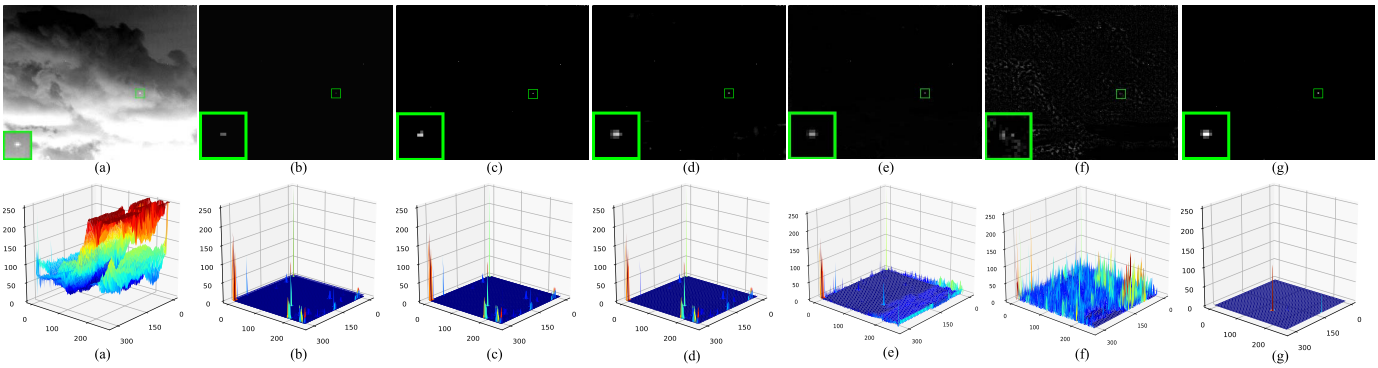


Fig. 15. Detected results of the six methods for Seq. 2 and the close-ups of the small UAV target for the detected result. (a) 54th frame of Seq. 2. (b)–(g) Detected results and the corresponding 3-D maps by the IPI, RIPT, PSTNN, DnCNN, CDAE, and our method, respectively.

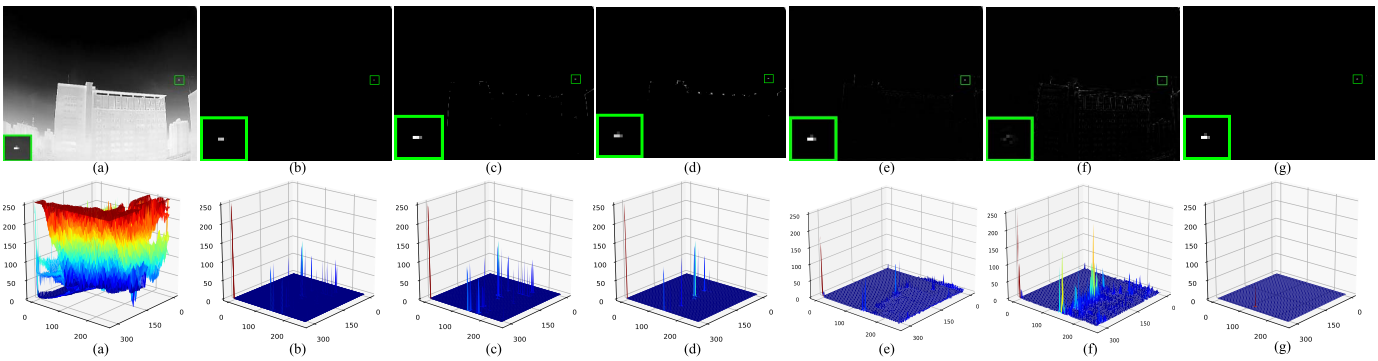


Fig. 16. Detected results of the six methods for Seq. 6 and the close-ups of the small UAV target for the detected result. (a) 271st frame of Seq. 6. (b)–(g) Detected results and the corresponding 3-D maps by the IPI, RIPT, PSTNN, DnCNN, CDAE, and our method, respectively.

lower left corners of the result images. It can be seen from the close-up images of small targets in Figs. 14–17(b) and (c) that the IPI and RIPT methods can enhance small targets to a certain extent, but the small target pixels obtained by these two methods are less than the target pixels in the real IR image. Take sequence 2 as an example. It can be seen from Fig. 15 that the scale of the UAV target in the original input image is small and the UAV target is in a strong cloud background, and the pixels in the UAV field are close to the edge pixels of the UAV target. These two traditional methods, IPI and RIPT, only use the nonlocal self-similarity of the background and the sparsity of the small target to complete the separation of the target and background. Therefore, it is easy for both methods to lose the UAV target edge when the target edge pixels are close to its neighboring pixels. However, our method utilizes rich multiscale structural information to accurately reconstruct residual images and, thus, preserves the original size of the UAV to the greatest extent compared with traditional methods. Visually, for sequence 2, it can be seen that the UAV in the original input image occupies approximately 12–15 pixels. IPI only recovers two pixels of the original UAV pixels, RIPT recovers 3 pixels, and our method recovers 10 pixels, which shows that our results outperform the other methods and are closest to the ground truth. Compared with IPI and RIPT, the

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

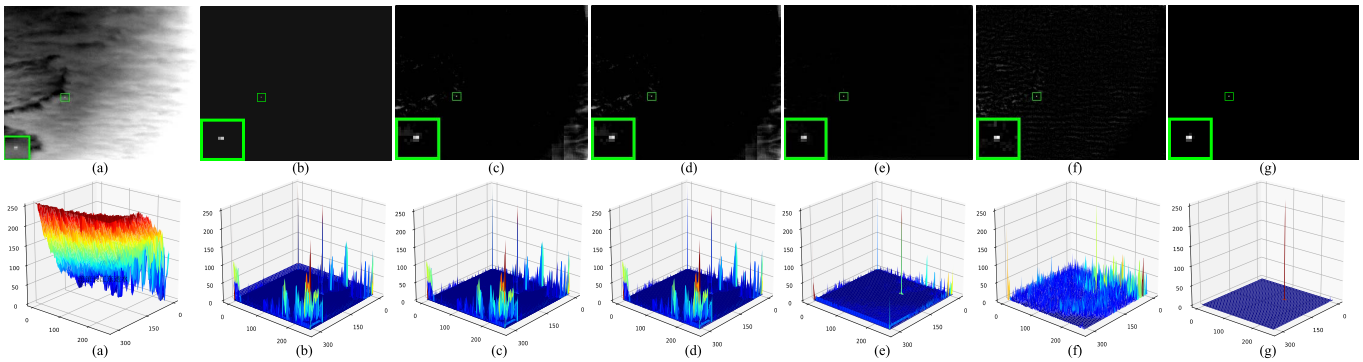


Fig. 17. Detected results of the six methods for Seq. 8 and the close-ups of the small UAV target for the detected result. (a) 84th frame of Seq. 8. (b)–(g) Detected results and the corresponding 3-D maps by the IPI, RIPT, PSTNN, DnCNN, CDAE, and our method, respectively.

PSTNN method has a stronger target enhancement ability, and the small targets are more complete than the targets obtained by the above two methods. Figs. 14–17(d) show that the UAV target in the detection results of PSTNN is relatively complete, about 2/3 of the original UAV target scale. However, for PSTNN, there is still room for improvement in the strong background clutters and noises.

Compared with the traditional methods, the DnCNN method based on deep learning can effectively enhance the target, but its poor background suppression also leads to some background residue. Especially for heavy cloud or glare interference complex building background, DnCNN performs poorly in background suppression. For example, sequence 6 is all complex building backgrounds with heavy cloudy and smog, and the UAV target is extremely dim and weak. From Fig. 16, it can be seen that there are still lots of background residues in the detection results of DnCNN. This is due to the limited ability of DnCNN to reconstruct residual images. In contrast, our method can reconstruct the residual images completely and accurately. From Figs. 14 to 17(f), it can be seen that the results by the CDAE method still have the background residual. It is worth noting that the CDAE method also does not highlight small and dark UAV targets either. This is because CDAE adopts a traditional encoder–decoder network consisting of basic operations of convolution and average pooling. This relatively simple network architecture cannot accurately and completely extract the feature information of the background and target and, thus, great difficulties in learning residual images with complex backgrounds. As seen in Fig. 16, our method has a consistent ability to suppress obvious background clutter and noise for all test images. The close-ups of the detection results show that the proposed method almost maintains the integrity of the target entirely. Besides, the proposed method has great capabilities to detect the extremely dim and weak UAV targets, which suggests that our method can estimate the residual image very well.

Figs. 14–17 also present the corresponding 3-D gray maps of the detection results of the six methods. For sequence 1, the background is relatively simple, and the target is brighter. From the 3-D distribution of images, traditional methods, such as IPI, RIPT, and PSTNN, can also accurately detect UAV targets but with a small amount of background residue, and DnCNN also has some background residue. Our method not only accurately detects the UAV target but also completely removes the background clutter and noise. For sequence 2, it can be seen from Fig. 15 that there still exists a large amount of background clutter and noise persists in the small UAV target image obtained by the three traditional methods of IPI, RIPT, and PSTNN. That is to say, although other baseline methods can also detect objects, their ability to suppress the background is weak. Sequence 6 has a complex building background with distinct edges, which greatly increases the difficulty of UAV target detection. As shown in Fig. 16, the results obtained by the IPI, RIPT, and PSTNN have many background residues, especially on the edges of buildings. For sequence 8, as shown in Fig. 17(b)–(d), due to the interference of heavy clouds, there remain a lot of background residues in the 3-D gray maps in the resulting image. It means that the conventional detection methods cannot well suppress the complex background. It can be mainly ascribed to the fact that the traditional detection methods struggle to model the images only utilizing locally shallow or globally fixed features, which have limited ability to cope with the complicated background, such as heavy cloud clutter and complex building background that are commonly seen in the anti-UAV applications. The DnCNN also has a weak feature representation ability for the complex background. Generally speaking, the existence of complex background clutter and noise greatly increases the difficulty of residual image learning. As can be seen from Figs. 14 to 17(f), we can observe that the CDAE method still suffers lots of clutters and noise residuals. Our method successfully overcomes the above challenges owning to the powerful image reconstruction capabilities of a novel U-shaped encoder–decoder network. The reason may be that our method uses DSRDB to extract abundant hierarchical features in the encoder stage, and thus, the features of the residual images can be extracted well. In addition, MFFRB fuses multiscale features from the input layer and the intermediate aggregation layer in the decoder stage. Thus, the residual image can be reconstructed completely. In summary, our method is an effective approach to enhance the small target while suppressing the background clutters and noise.

Fig. 18 shows the ROC curves on eight real IR sequences of six methods. Generally speaking, if the position of the ROC

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

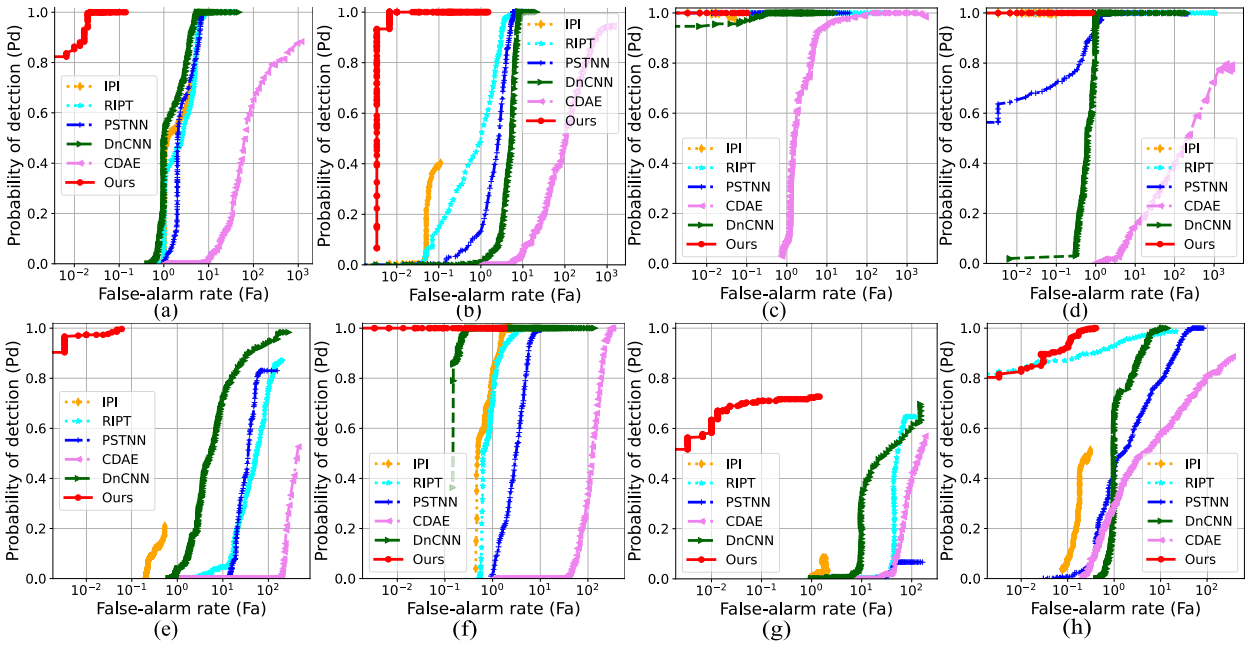


Fig. 18. ROC curves of the detection results for eight real sequences in logarithmic scale. (a) Seq. 1. (b) Seq. 2. (c) Seq. 3. (d) Seq. 4. (e) Seq. 5. (f) Seq. 6. (g) Seq. 7. (h) Seq. 8.

TABLE VIII
COMPARISON OF AVERAGE RUNNING TIME IN SECONDS

Method	Time	Platform
IPI	5.8855	CPU
RIPT	2.2619	CPU
PSTNN	0.4904	CPU
DnCNN	0.0086	GPU
CDAE	0.0171	GPU
Our method	0.0265	GPU

curve of a method is in the upper left corner of a graph, the method has better detection performance. It can be noted from Fig. 18 that, compared with other baseline methods, ours consistently achieves the highest detection rate. In particular, for sequences 2–4 and 6, the detection rate of our method almost reaches 1, showing that our method can detect small UAV targets. Moreover, among the eight real IR sequences, the corresponding curve of our method always rises the fastest, which shows that our method has a lower false alarm rate than other methods under the same detection rate. This indicates that our method can effectively suppress background clutter and noise.

The ROC curves of sequences 3 and 4 need special attention. For sequence 3, it can be found that the ROC curves of IPI, RIPT, and PSTNN are very close to ours. This is because, when the segmentation threshold is small, the P_d values of these three methods are relatively close. However, with the continuous increase in the threshold, more and more false alarms are excluded. According to the definition of F_a, the value of F_a decreases with the increase in the threshold. Therefore, if F_a of some methods becomes 0 earlier when the threshold is small, then this method is better.

F_a of the IPI method becomes 0 when the segmentation threshold is raised to 37, indicating that there is no false alarm beyond the threshold of 37. F_a of RIPT becomes 0 when the segmentation threshold is 98, suggesting that no false detection beyond the threshold of 98. F_a of PSTNN is only 0 when the segmentation threshold is 212, that is, there is no false detection beyond the threshold of 212. Compared with these three traditional methods, the F_a value of our method is equal to 0 at a threshold of just 33, that is to say, the proposed method has no false alarm earlier, demonstrating that our method has a stronger background suppression ability. In general, the proposed method consistently achieves the best detection ability and background clutters' suppression ability across almost all thresholds, thus demonstrating the superiority of our method.

For sequence 4, it can be found that the ROC curves of IPI and RIPT are very close to ours in sequence 4, and it is similar to the case of sequence 3. In the IPI method, F_a becomes 0 when the segmentation threshold is raised to 42, indicating that there is no false detection beyond the threshold of 42. F_a of RIPT is only 0 when the segmentation threshold is increased to 75, suggesting no false detection beyond the threshold of 75. Compared with these two traditional methods, F_a of our method is equal to 0 at a threshold of only 34, which proves that our method has a stronger background suppression ability.

In conclusion, our method can perform effectively and robustly for UAV targets with various complicated backgrounds.

4) Comparison of Average Running Time: Table VIII shows the average running time per frame (in seconds) for six different detection methods in the test set. It can be seen that the computational time for the proposed method to detect each frame is 0.0265 s. It is noted that the difference among the six methods in time may be attributed to the difference in platforms. Although the proposed method is lower in detection speed than DnCNN and CDAE, our method still achieves up

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
IMPACT OF INTEGRATING THE DSRDB, MFFRB, AND GRC IN THE U-NET++ ON OUR DATASET IN TERMS OF THE AVERAGE F1

DSRDB	MFFRB	GRC	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5	Seq. 6	Seq. 7	Seq. 8
-	-	-	0.4977	0.8009	0.9904	0.7634	0.4600	0.8732	0.4128	0.4634
✓	-	-	0.6165	0.8205	0.9924	0.7820	0.4686	0.9080	0.4561	0.4869
✓	✓	-	0.6243	0.8206	0.9933	0.7832	0.4804	0.9091	0.4798	0.5104
✓	✓	✓	**0.6271**	**0.8252**	**0.9941**	**0.7840**	**0.4890**	**0.9297**	**0.4822**	**0.5292**

TABLE X
IMPACT OF INTEGRATING THE DSRDB, MFFRB, AND GRC IN THE U-NET++ ON OUR DATASET IN TERMS OF THE AVERAGE CG

DSRDB	MFFRB	GRC	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Seq. 5	Seq. 6	Seq. 7	Seq. 8
-	-	-	1.59	1.63	1.80	2.61	11.03	1.66	10.64	2.15
✓	-	-	1.60	1.67	1.82	2.63	11.04	1.67	14.36	2.20
✓	✓	-	1.67	1.67	1.84	2.66	11.17	1.68	17.74	2.28
✓	✓	✓	**1.68**	**1.68**	**1.85**	**2.67**	**11.44**	**1.69**	**17.75**	**2.35**

to 38 frames/s (FPS) for 256 × 320 video frames, which meets real-time requirements for an anti-UAV video processing system. In summary, our method has a better balance between detection accuracy and speed than other methods.

F. Ablation Study

In order to verify the rationality of our model design, we perform a series of ablation experiments on eight real IR sequences. In our ablation study, we utilize U-Net++ as a baseline and explore the contribution to and effect on the model detection ability of the proposed components (i.e., DSRDB, MFFRB, and GRC). To verify the performance of these mechanisms, we conduct experiments with different settings: 1) testing the detection performance of eight real IR UAV sequences on the baseline method; 2) replacing Conv-3 + ReLU + Conv-3 + Local Residual Connection in the original U-Net++ encoding stage with our proposed DSRDB; 3) on the basis of using DSRDB, replacing Conv-3 in the original U-Net++ decoding stage + ReLU + Conv-3 + Local Residual Connection by our proposed MFFRB; and 4) using 2) and 3) simultaneously, and appending the GRC. In this ablation experiment, we evaluate the capabilities of the proposed components in terms of detection performance, target enhancement, and background suppression, respectively. We adopt F1 to evaluate the model's detection performance since it is the harmonic mean of accuracy and recall while using CG to measure the model's ability to enhance the target and suppress the background. Tables IX and X show the effects of the proposed method in terms of evaluation measures of F1 and CG.

It can be seen from Tables IX and X that the DSRDB introduced in the encoder stage significantly improves the detection performance and the ability of target enhancement and background suppression with the average F1 and CG values. With MFFRB integrated into the decoder stage, the capability of target enhancement and clutter suppression is improved greatly. For the eight real IR sequences, compared with the improvement of the CG after DSRDB is introduced, it can be observed from Table X that the CG value of this method is improved more significantly after the introduction of MFFRB, which indicates that, compared with the baseline method, the introduction of MFFRB can further improve the ability of target enhancement and background suppression. In addition, since the downsampling operation in the network encoder stage may cause the loss of the feature information of the input image, we incorporate a GRC into the network. The GRC provides long-distance information compensation of the network and facilitates the gradient backpropagation, thus further improving the detection ability of small targets.

G. Additional Experiments

1) Superiority of Our Method Compared With CNN-Based Baseline Detectors: Currently, CNN-based object detection methods have made great progress in the field of computer vision. In order to further verify the superiority of our method, we compare our method with the off-the-shelf baseline detectors: SSD and YOLOv3. For SSD, the initial learning rate is 10^{-4}, a total of 40 000 epochs are trained, and the learning rate is multiplied by 0.1 when iterating to 10 000, 15 000, and 19 000. For YOLOv3, the initial learning rate is 10^{-3}, the whole training epoch is 30 000, and the batch size is 8. The representative detection results for three methods are shown in Fig. 19. It can be observed from Fig. 19(b) and (c) that there are many false detections in the detection results of the SSD. YOLOv3 outputs multiple prediction boxes for one UAV target. The unsatisfactory detections obtained by SSD and YOLOv3 suggest that it is very challenging to accurately learn the features of the small UAV targets in a complicated background. By contrast, our method can well detect the small UAV target by adopting residual image prediction without false detections.

2) Robustness of the Proposed Residual Image Prediction Model in the Case of No UAV Targets: To further verify the background suppression ability of our method, we test six methods on a real IR image without UAV targets and only visually present the intermediate target images after the

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

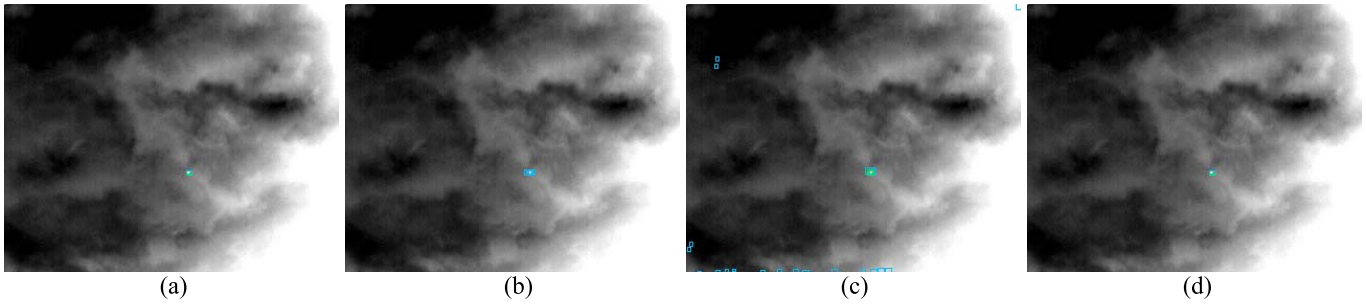


Fig. 19. Representative detection results of the YOLOv3, SSD, and our method. (a) 250th frame of Seq. 3. (b)–(d) Detected results by the YOLOv3, SSD, and our method, respectively. The green box denotes the ground truth, and the blue box denotes the detected result.

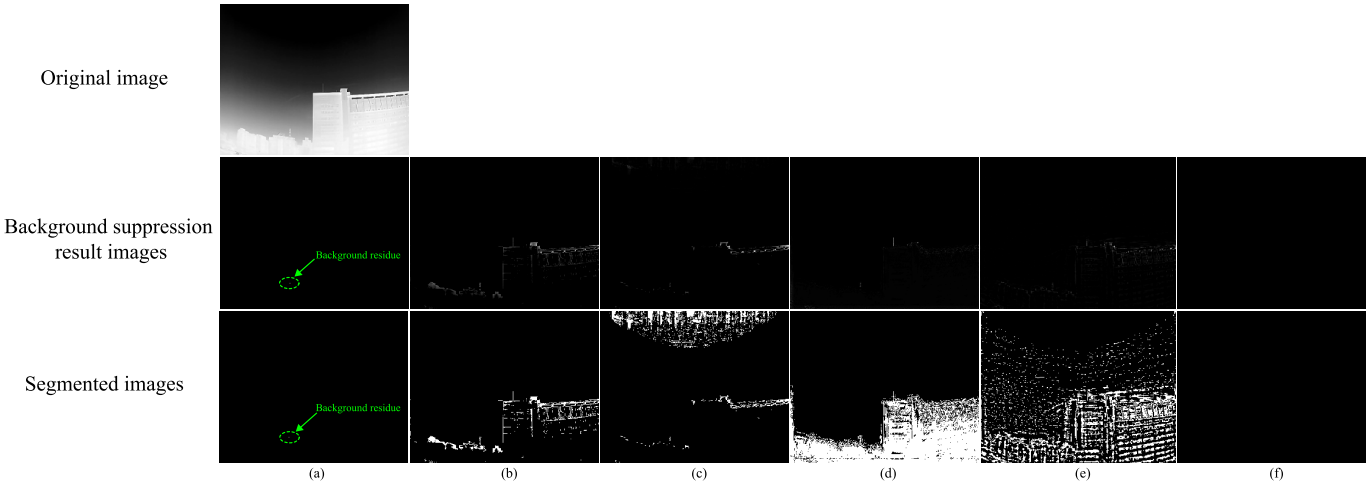


Fig. 20. Background suppression results of six methods on a real IR image without any targets. From left to right, (a)–(f) are the intermediate targets (middle) and segmented images (bottom) of IPI, RIPT, PSTNN, DnCNN, CDAE, and our method, respectively.

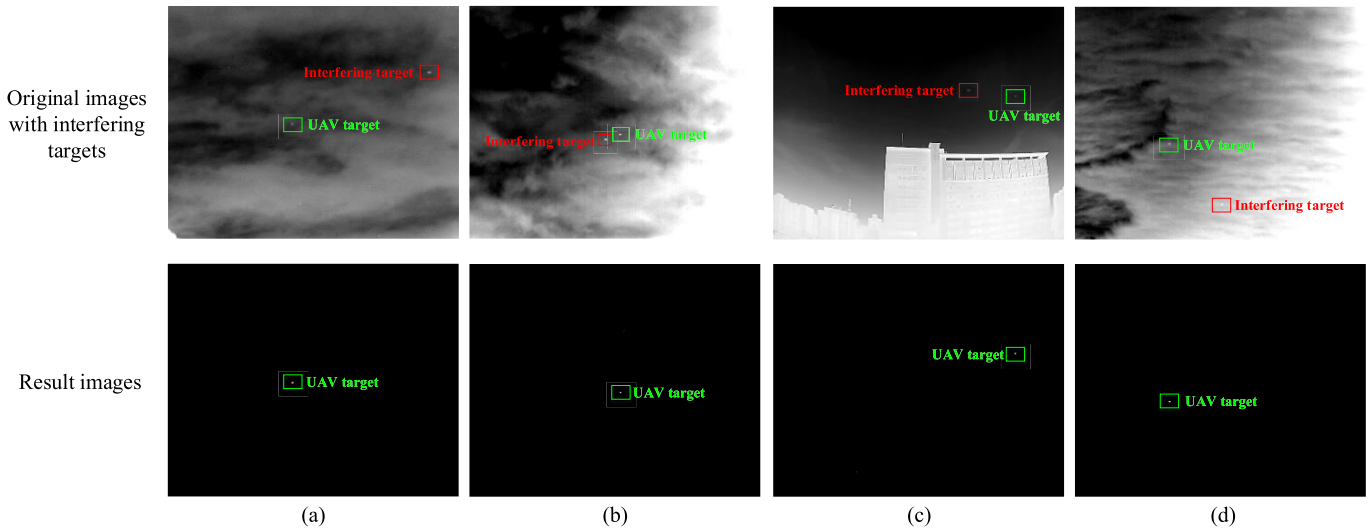


Fig. 21. Detected results of our method on images with interfering targets. From left to right, (a)–(d) are the input image with interfering targets and detection results of Seq. 1, Seq. 4, Seq. 5, and Seq. 8 respectively. Target regions are shown in small green boxes; interfering targets are displayed in small red boxes.

background suppression. For a more intuitive presentation of the background residues, we also show the results using threshold segmentation with a common threshold of zero. As it can be observed in Fig. 20(a)–(e), the original detected results and the segmented results by IPI, RIPT, PSTNN, DnCNN, and CDAE still contain background residues, which intuitively suggests that these methods cannot suppress the complex background well. By contrast, the results of our method are cleaner without background residues, which again validates that our method is able to reconstruct the residual image very well

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

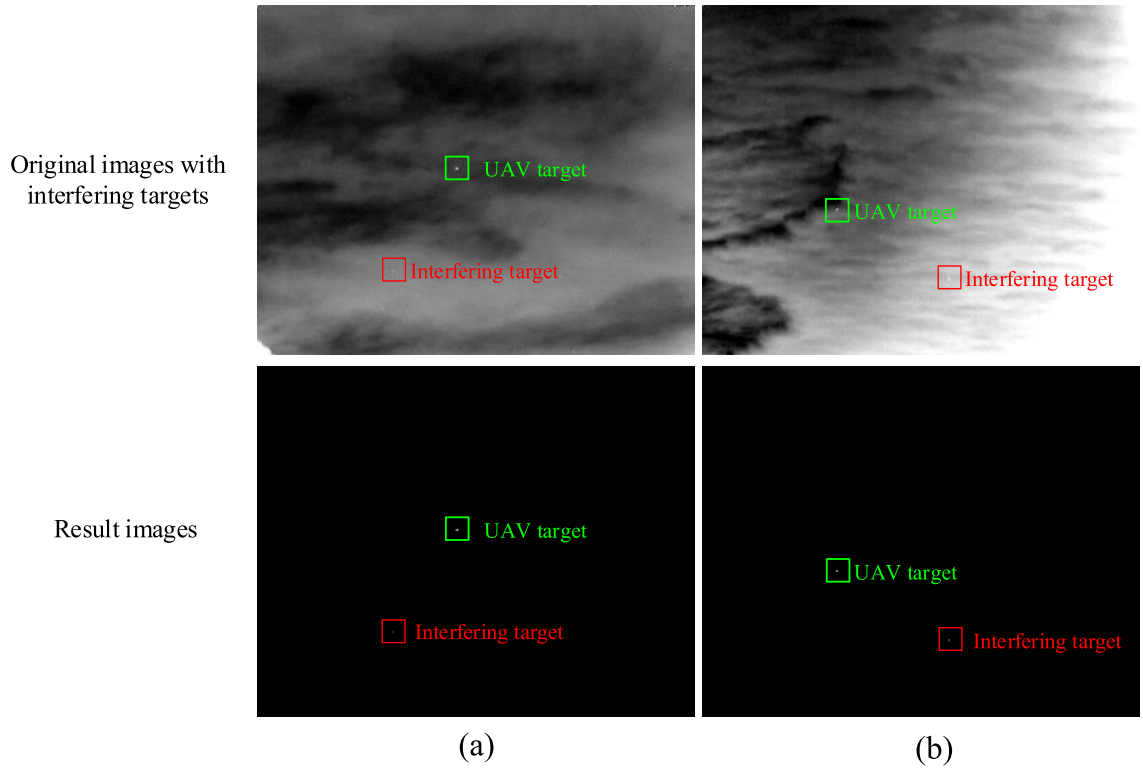


Fig. 22. Some detection failure cases with interfering targets by our method. From left to right, (a) and (b) are the input image with interfering targets and detection results of Seq. 1 and Seq. 8, respectively. Target regions are shown in small green boxes, and interfering targets are displayed in small red boxes.

and consistently yield better detection results. It means that our method is more robust in complex background conditions.

3) Influence of the Interfering Target on Our Method: In order to explore the influence of interfering targets on the detection ability of our method, we add interfering targets similar to small UAV targets (bird or ordinary aircraft) to sequence 1, sequence 4, sequence 5, and sequence 8 of testing data. We apply the proposed method to these four test sequences, and it can be seen from Fig. 21 that the proposed method has a robust anti-interference detection ability for small interfering targets.

However, our method also has false alarms in individual frames, as shown in Fig. 22. As our method strives for accurate background reconstruction ability, when the interfering objects of individual frames are extremely similar to the objects of interest, the proposed detection method has difficulty in distinguishing interfering targets from similar targets. However, our method can ensure that both the interfering target and the target of interest can be completely detected, and then, the interfering target can be further removed by using the spatial–temporal correlation information of the front and back frames in the traditional target tracking method.

V. Conclusion

This article proposes a new IR small UAV target detection idea by transforming the target detection into residual image prediction by learning the nonlinear mapping from the original image to the residual image. The constructed end-to-end U-shaped network can extract rich hierarchical structure features in the encoder by invoking the DSRDB. The residual image reconstruction can be well performed in the decoder stage by fully aggregating multiscale features from the encoder layers and intermediate decoder layers at the same scale and the decoder layers at different scales. Global residual learning is also employed to provide information compensation to further improve the reconstruction performance. Extensive experimental results on real IR UAV images demonstrate that our method has great potential to be a candidate in anti-UAV practical applications. In the future, we will deploy our scheme on an embedded platform for real applications. In addition, association information between adjacent frames can be also adopted to remove similar distractors.

Appendix

In this appendix, we give the dimension of feature maps after the concatenation operation in our network. The proposed network architecture consists of two subnetworks: the encoder network and the decoder network. In the encoding stage of the network, the feature maps are concatenated in DSRDB to extract multilevel local features. The DSRDB is shown in Fig. 5. The dimension of feature maps after the concatenation operation in DSRDB of different depths can be written as

$$C_{\mathrm{DSRDB}_i} = (N+1) \times 64 \times 2^i, \quad 0 \le i \le 2 \tag{14}$$

where C_{DSRDB_i} denotes the dimension of feature maps concatenated in DSRDB of different network depths, N represents the number of convolutional layers in each DSRDB, that is, the number of DSCBs, and i denotes the depth of the

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

network. 64 is the number of convolution kernels in the first convolutional layer of the whole network.

In the decoding stage of the network, MFFRB fuses the low-level, coarse-grained features of the encoding subnetwork with the high-level, fined-grained features of the decoding subnetwork through concatenation. The MFFRB in our network architecture can be divided into two parts MFFRB(a) and MFFRB(b), as shown in Figs. 6 and 7, respectively.

Therefore, the dimension of feature maps concatenated by the network in MFFRB(a) and MFFRB(b) of different depths can be defined as

$$C_{\mathrm{MFFRB}(a)_i} = 2 \times 64 \times 2^i, \quad 0 \le i \le 1 \tag{15}$$

$$C_{\mathrm{MFFRB}(b)_i} = 3 \times 64 \times 2^i, \quad i = 0 \tag{16}$$

where $C_{\mathrm{MFFRB}(a)_i}$ and $C_{\mathrm{MFFRB}(b)_i}$, respectively, denote the dimension of feature maps concatenated in MFFRB(a) and MFFRB(b) of different network depths, and i denotes the depth of the network.

Furthermore, multiscale feature maps are fused by concatenating the intermediate aggregated MFFRB(a) in the horizontal direction and MFFRB(b), as illustrated in Fig. 4. The dimension of feature maps after concatenation can be written as

$$C_0 = D \times 64, \quad D \ge 0 \tag{17}$$

where D represents the depth of the network. D is equal to 2 in our network architecture.

REFERENCES

[1] S. M. Giray, "Anatomy of unmanned aerial vehicle hijacking with signal spoofing," in *Proc. 6th Int. Conf. Recent Adv. Space Technol. (RAST)*, Jun. 2013, pp. 795–800.

[2] E. Vattapparamban, I. Güvenç, A. Yurekli, K. Akkaya, and S. Uluağaç, "Drones for smart cities: Issues in cybersecurity, privacy, and public safety," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Sep. 2016, pp. 216–221.

[3] X. Shi, C. Yang, W. Xie, C. Liang, Z. Shi, and J. Chen, "Anti-Drone system with multiple surveillance technologies: Architecture, implementation, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 68–74, Apr. 2018.

[4] Y. Zhao and Y. Su, "Cyclostationary phase analysis on micro-Doppler parameters for radar-based small UAVs detection," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 9, pp. 2048–2057, Sep. 2018.

[5] Y. Zhao and Y. Su, "Cyclostationary phase analysis on micro-Doppler parameters for radar-based small UAVs detection," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 9, pp. 2048–2057, Sep. 2018.

[6] W. Zhang, M. Cong, and L. Wang, "Algorithms for optical weak small targets detection and tracking: Review," in *Proc. Int. Conf. Neural Netw. Signal Process.*, vol. 1, Dec. 2003, pp. 643–647.

[7] R. Eysa and A. Hamdulla, "Issues on infrared dim small target detection and tracking," in *Proc. Int. Conf. Smart Grid Electr. Autom. (ICSGEA)*, Aug. 2019, pp. 452–456.

[8] K. Zhang, K. Yang, S. Li, and H.-B. Chen, "A difference-based local contrast method for infrared small target detection under complex background," *IEEE Access*, vol. 7, pp. 105503–105513, 2019.

[9] H. Fang, M. Xia, G. Zhou, Y. Chang, and L. Yan, "Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[10] H. Fang, X. Wang, Z. Liao, Y. Chang, and L. Yan, "A real-time anti-distractor infrared UAV tracker with channel feature refinement module," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1240–1248.

[11] C. Q. 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] X. Wang, Z. Peng, D. Kong, P. Zhang, and Y. He, "Infrared dim target detection based on total variation regularization and principal component pursuit," *Image Vis. Comput.*, vol. 63, pp. 1–9, Jul. 2017.

[18] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4204–4214, Jul. 2016.

[19] Y. Chen and Y. Xin, "Infrared small target detection based on visual contrast mechanism," *Laser Infr.*, vol. 47, no. 2, pp. 252–256, 2017.

[20] Y. Chen and Y. Xin, "An efficient infrared small target detection method based on visual contrast mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 7, pp. 962–966, Jul. 2016.

[21] 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.

[22] Y. Chen, B. Song, X. Du, and M. Guizani, "Infrared small target detection through multiple feature analysis based on visual saliency," *IEEE Access*, vol. 7, pp. 38996–39004, 2019.

[23] J. Liu, Z. He, Z. Chen, and L. Shao, "Tiny and dim infrared target detection based on weighted local contrast," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1780–1784, Nov. 2018.

[24] M. Shi and H. Wang, "Infrared dim and small target detection based on denoising autoencoder network," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1469–1483, 2019.

[25] H. Fang, M. Chen, X. Liu, and S. Yao, "Infrared small target detection with total variation and reweighted ell_1 regularization," *Math. Problems Eng.*, vol. 2020, pp. 1–19, Jan. 2020.

[26] H. Zhu, S. Liu, L. Deng, Y. Li, and F. Xiao, "Infrared small target detection via low-rank tensor completion with top-hat regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1004–1016, Oct. 2020.

[27] B. Zhao, C. Wang, Q. Fu, and Z. Han, "A novel pattern for infrared small target detection with generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4481–4492, May 2021.

[28] 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, Nov. 2021.

[29] P. Yang, L. Dong, and W. Xu, "Infrared small maritime target detection based on integrated target saliency measure," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2369–2386, 2021.

[30] M. Zhao, W. Li, L. Li, P. Ma, Z. Cai, and R. Tao, "Three-order tensor creation and tucker decomposition for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[31] J. Han, C. Liu, Y. Liu, Z. Luo, X. Zhang, and Q. Niu, "Infrared small target detection utilizing the enhanced closest-mean background estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 645–662, 2021.

[32] S. Sun, Y. Yin, X. Wang, and D. Xu, "Robust visual detection and tracking strategies for autonomous aerial refueling of UAVs," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 12, pp. 4640–4652, Dec. 2019.

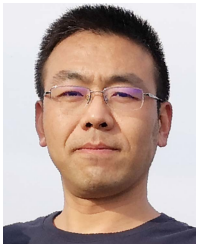
[33] C. Gao, L. Wang, Y. Xiao, Q. Zhao, and D. Meng, "Infrared small-dim target detection based on Markov random field guided noise modeling," *Pattern Recognit.*, vol. 76, pp. 463–475, Apr. 2018.

[34] S. D. Deshpande, M. H. 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.

[35] W. Xue, J. Qi, G. Shao, Z. Xiao, Y. Zhang, and P. Zhong, "Low-rank approximation and multiple sparse constraint modeling for infrared low-flying fixed-wing UAV detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4150–4166, 2021.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.

[36] 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.
[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 9905. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.
[38] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.
[39] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6517–6525.
[40] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[41] 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.
[42] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.
[43] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognit.*, vol. 43, no. 6, pp. 2145–2156, 2010.
[44] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2016, pp. 1–4.
[45] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, Apr. 2018.
[46] L. A. Lim and H. Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognit. Lett.*, vol. 112, pp. 256–262, Sep. 2018.
[47] C. Lin, B. Yan, and W. Tan, "Foreground detection in surveillance video with fully convolutional semantic network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4118–4122.



Houzhang Fang (Member, IEEE) received the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014.

He is currently an Associate Professor with the School of Computer Science and Technology, Xidian University, Xi'an, China. His research interests include target recognition and image restoration.



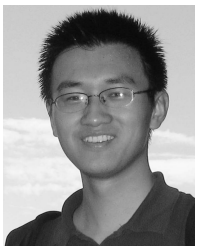
Lan Ding received the B.S. degree in software engineering from Ningxia University, Yinchuan, China, in 2021. She is currently pursuing the M.S. degree in electronic information (software engineering direction) with Xidian University, Xi'an, China.

Her research interests include computer vision and deep learning.



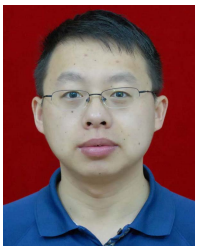
Liming Wang received the B.S. degree in software engineering and the M.S. and Ph.D. degrees in computer science from Xidian University, Xi'an, China, in 2004, 2007, and 2013, respectively.

He is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. His research interests include artificial intelligence in medicine and industry intelligence.



Yi Chang (Member, IEEE) received the B.S. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2011, and the M.S. degree in pattern recognition and intelligent systems and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014 and 2019, respectively.

From 2014 to 2015, he was a Research Assistant with Peking University, Beijing, China. From 2019 to 2021, he was a Post-Doctoral Researcher with Peng Cheng Laboratory, Shenzhen, China. He is currently an Assistant Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include image processing, computer vision, and machine learning.



Luxin Yan (Member, IEEE) received the B.S. degree in electronic communication engineering and the Ph.D. degree in pattern recognition and intelligence system from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2001 and 2007, respectively.

He is currently a Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include multispectral image processing, pattern recognition, and real-time embedded systems.



Jinhui Han received the Ph.D. degree in electrical circuits and systems from the Huazhong University of Science and Technology, Wuhan, China, in 2016.

He is currently a Lecturer with the College of Physics and Telecommunication Engineering, Zhoukou Normal University, Zhoukou, China. His current research interests include target detection and image processing.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 27,2022 at 02:00:55 UTC from IEEE Xplore. Restrictions apply.