IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 14, 2020, accepted April 29, 2020, date of publication May 6, 2020, date of current version May 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992647

Multi-Scale Aerial Target Detection Based on Densely Connected Inception ResNet

MIAOHUI ZHANG[1,2], KANGNING PANG[1], CHENGCHENG GAO[1], AND MING XIN[3]
[1]Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475000, China
[2]Institute of Data and Knowledge Engineering, Henan University, Kaifeng 475000, China
[3]School of Computer Science and Engineering, Beihang University, Beijing 100083, China

Corresponding author: Ming Xin (xinming_bh@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802111, in part by the Foundation of Henan Education Department under Grant 19A520002, in part by the Fund of Henan Province Young Key Teacher under Grant 2017GGJS019, and in part by the Postdoctoral Science Fund of China under Grant 2015M582182.

ABSTRACT With the rapid development of unmanned aerial vehicles (UAVs), aerial targets detection has attracted extensive attention from researchers. The difficulty of aerial image detection lies in the small proportion of ground targets in aerial images and the wide variety of target sizes. After multiple down-sampling, the features of small targets are almost not available on the feature maps. To address these drawbacks, a densely connected Inception ResNet (RIDNet) is proposed. RIDNet is a lightweight multi-scale fusion detection network constructed with two residual inception units (RI): the RI-Dense model and the RI-Deconv model. The RI-Dense model consists of densely connected layers and shortcut connections. Each convolutional layer in RI-Dense has access to all the subsequent layers and passes on the information that needs to be preserved. The RI-Deconv fuses the global feature in a residual and hierarchical way, which continuously deconvolutes the output of RI-Dense and concatenates the result with the original output to get fusion layers. The fused layers absorb semantic information and detailed information from deep layers and shallow layers, respectively. Extensive experiments show the effectiveness of the proposed RIDNet. Ablation experiments also demonstrate that the RI-Dense model and RI-Deconv model can improve the mAP by 7.8% and 6.8%, respectively.

INDEX TERMS Object detection, convolutional network, unmanned aerial vehicle.

I. INTRODUCTION

For the past few years, the rapid development of unmanned aerial vehicle (UAV) has garnered an increasing interest for this technology in a wide range of applications. UAV-based remote sensing has become more modular, miniaturized and intelligent in recent years, and been widely used in various fields. In the field of security, particularly, the role of UAV is increasingly important [1]. If a manually evaluated image is used to search the targets, considerable false and missed detections could occur owing to the influence of subjective human factors. Therefore, high-precision and robust detection algorithms are urgently needed.

Detecting objects in aerial images is difficult and challenging due to the following reasons: (1) The long-distance between the target and UAV inevitably leads to a low resolution for images. (2) There exist huge variations in the appearance and color of targets with various orientations, which increases the inter-class similarity between desired targets and the complex backgrounds. The resolution of images taken by different devices also varies.

It is intractable to recognize targets in aerial images effectively. In order to solve these problems, a large number of detection methods have been proposed.

The majority of existing object detection methods applied to aerial images are implemented to distinguish objects from the background at a large scale, such as ships at sea [2], and planes at airports [3]. Traditional detection methods have many problems: The handy features are only for specific target detection, and they perform poorly in generalization and robustness. Besides, the region searching based traditional algorithms are time-consuming and slow [3].

More recently, deep learning-based algorithms have been dominating the top accuracy benchmarks for various visual

The associate editor coordinating the review of this manuscript and approving it for publication was Ruqiang Yan.

FIGURE 1. Samples of aerial targets. The scale of targets changes in a broad range.

detection tasks. The existing target detection models based on deep learning can be divided into two categories: those based on region recommendation [4]–[9] and those based on regression [10]–[13]. The region proposal based networks use the idea of region proposal and then classify them; while the regression-based networks use a single convolutional network to predict bounding boxes and class probabilities simultaneously from an input image.

To improve the detection results for aerial objects, an aerial target detection lightweight network is proposed, with fewer model parameters, a faster detection speed, and higher detection accuracy. The main contributions are as the following:

(1) The RI-Dense model is introduced to replace the original feature extraction network. This model integrates the ideas of InceptionNet and DenseNet, effectively connecting every feature layer in series along the feature channel to ensure information transmission. This proposed structure alleviates the problem of gradient disappearance, retains more features, and reduces the number of network parameters.

(2) A multi-scale feature fusion structure RI-Deconv is proposed, motivated by the idea of ResNet. RI-Deconv uses the deconvolution operation to perform multilayer feature fusion and constructs advanced feature maps with a high resolution and semantic information.

(3) A target stitching method is designed to combine the cropped targets. Large-sized pictures need to be cropped before being sent to the network. Consequently, some objects are inevitably cropped into different parts. Our method would stitch these divided parts together and reduce the missing rate.

(4) The proposed model is evaluated on the NWPU VHR-10 [14] and DOTA [15] datasets to test its performance. The experiment shows that the proposed network is effective in improving the detection accuracy.

The rest of this paper is organized as follows: section II introduces related work. Section III explains the details of the proposed method and materials. Section IV analyzes the experimental results. Section V makes a conclusion.

II. RELATED WORK

In recent years, many target detection methods based on deep learning have been developed, immensely promoting the advancement of target detection. These theories will be briefly introduced in the following part.

Over the past decade, some efforts have been devoted to addressing the problem of small object detection from aerial videos [16]–[19]. One widely applied strategy is to enlarge images to different scales directly, which achieves more detailed information about the small targets. For example, Chen *et al.* [20] presented an approach where the input is magnified to enhance the resolution of small objects. On the basis of this research principle, Cao *et al.* [21] fused feature modules to additional contextual information to deliver a better detection performance. By generating multiple feature maps with different resolutions, they are able to naturally handle objects of various sizes including small ones. Other approaches are based on the deep neural network in which multi-scale feature layers represent each small target characteristic [22], [23].

YOLO (You only look once) [10] is the first one-stage detector in object detection, a milestone in the history of one-stage detection model. YOLOV3 [13], an enhanced version of YOLO, achieves an outstanding detection accuracy. SSD [12] (Single Shot MultiBox Detector) is a one-stage detector proposed by W. Liu *et al.* The main contribution of this technology is the introduction of the multi-reference and multi-resolution detection technique which significantly improves the detection accuracy of one-stage detectors. Based on SSD and similar to FPN, DSSD [24] employs top-down pyramid CNN layers to improve the accuracy, but at the cost of computational efficiency. FSSD [25] inserts a fusion module at the bottom of the feature pyramid to enhance the accuracy of SSD. While keeping a fast speed, FSSD achieves marginal improvements upon SSD in accuracy. S3FD [26] is a highly accurate real-time face detector, based on the anchor model used initially for object detection. In order to overcome the limitations on small objects, S3FD introduces a scale compensation anchor matching strategy to improve recall rate, and a max-out background label to reduce the false positive detections.

After studies on these works, researchers put forward smaller target detection algorithms. Cheng *et al.* [14] train a RICNN model by optimizing a new objective function via imposing a regularization constraint. This explicitly enforces the feature representations of the training samples to be mapped closer to each other before and after rotating, hence achieving a rotation invariance. CISPNet [27] applies a context information scene perception (CISP) module to obtain the contextual information for targets of different scales and uses k-means clustering to set the aspect ratios and sizes of default boxes. Cheng *et al.* [28] propose a novel and effective method to learn a rotation-invariant and Fisher

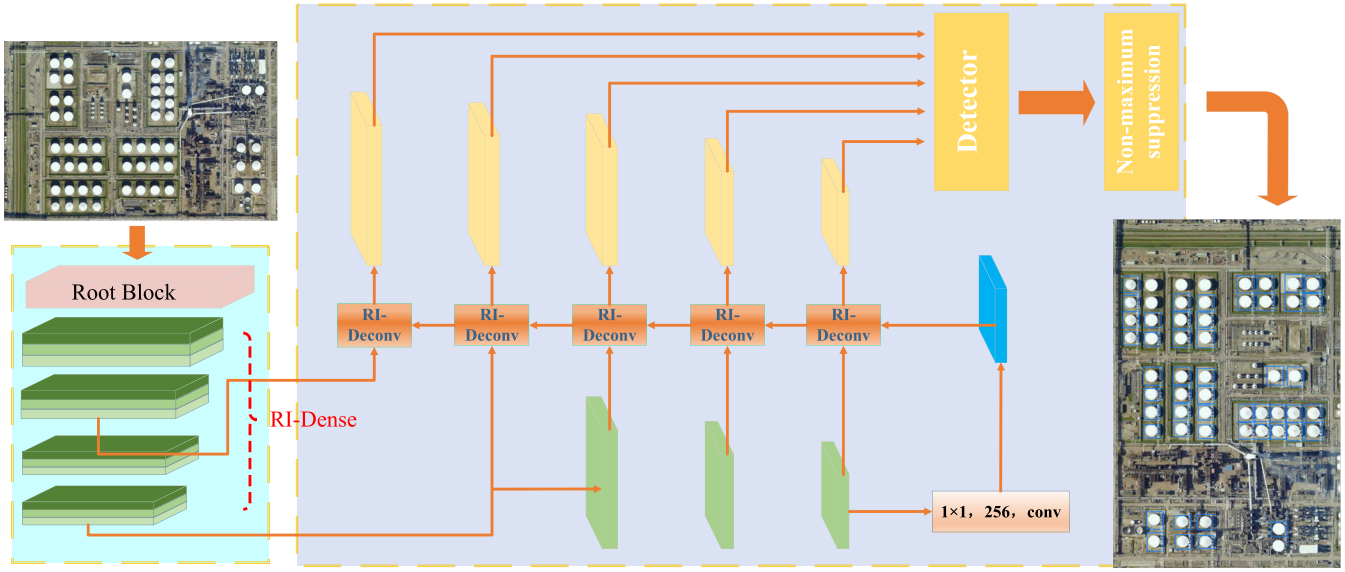

FIGURE 2. The framework of the proposed network.

discriminative CNN (RIFD-CNN) model by introducing and learning a rotation-invariant layer and a Fisher discriminative layer, respectively, on the basis of the existing high-capacity CNN architectures. REMSNet [29] combines a dense connectivity pattern and parallel multi-kernel convolution to build a lightweight and varied receptive field sizes model. In addition, they design a parallel multi-kernel deconvolution module and a spatial path to further aggregate different scales information. WFCNN [30] is a weight feature value convolutional neural network, consisting of one encoder and one classifier. The encoder uses the linear fusion method to hierarchically fuse semantic features. RADC-Net [31] proposes a residual attention based densely connected convolutional neural network, with a novel residual attention block designed to highlight local semantics relevant to the aerial scenes. Zhou *et al.* [32] suggest an effective framework for weakly supervised target detection in RSIs based on transferred deep features and negative bootstrapping for detection in remote sensing images. Li *et al.* [33] put forward a new FPN with multiangle anchors. A double-channel feature fusion network is proposed to learn local and contextual properties along two independent pathways. Zhang *et al.* [34], after analyzing the bottlenecks and development directions of deep learning in remote sensing target detection, provide a guidance for researches in this field. FaceBoxes [35] is a light-weight CNN for face detection which has a lightweight yet powerful network structure that consists of the Rapidly Digested Convolutional Layers (RDCL) and the Multiple Scale Convolutional Layers (MSCL). The MSCL, aiming to enrich the receptive fields and discretize anchors over different layers, is capable to handle faces of various scales. SVDNet [36] is designed based on a singular value decompensation algorithm, achieving a high detection robustness and desirable time performance. Diao *et al.* [37] combine the strength of the unsupervised feature learning of deep belief networks (DBNs) and visual saliency, which avoid an exhaustive search across the image and generate a small number of bounding boxes to locate the object quickly and precisely.

III. METHODS

SSD uses several feature layers to make predictions, which effectively improves the target detection accuracy. However, the feature layers in the SSD detect targets independently, resulting in reduced detection for small targets. A new one-stage detection model that inherits the idea of SSD is designed. The details will be introduced in this section.

A. THE RIDNet

The proposed network is illustrated in FIGURE 2. RIDNet consists of two parts: a feature extractor and an object detector. The RI-Dense structure, created as a feature extractor, consists of several RI-Dense modules, in which the input of each layer is the output of all previous modules. The dense connections between feature layers benefit to learning inner-class semantic features thoroughly. Therefore, the detection speed of the network is accelerated. Different from the SSD, our feature pyramid is made up of the fusion result of the RI-Deconv modules instead of the convolution. After each up-sampling process of the RI-Deconv structure is finished, the interference of the feature layer with less information can be reduced, and the feature recovery accuracy of the RIDNet can be enhanced, which improves the expressive power of the model.

B. FEATURE EXTRACTION NETWORK

The feature extractor consists of three modules: root module for preliminary feature extraction and feature size reduction,

TABLE 1. The structure of the feature extraction network.

Model	Structure	Output
Input		1024×1024×3
Root Block	Root module×1	128×128×32
RI-Dense Block 1	R-Dense module×6 Bottleneck layer×1	64×64×128
RI-Dense Block 2	R-Dense module×12 Bottleneck layer×1	32×32×256
RI-Dense Block 3	R-Dense module×12 Bottleneck layer×1	16×16×512
RI-Dense Block 4	R-Dense module×6 Bottleneck layer×1 without pooling	16×16×1024

TABLE 2. The structure of root model.

Layer	Rate	Kernel	Stride	Output
Dilay_1	1	3×3		512×512
Dilay_2	2	3×3	2	256×256
Dilay_3	3	3×3		128×128

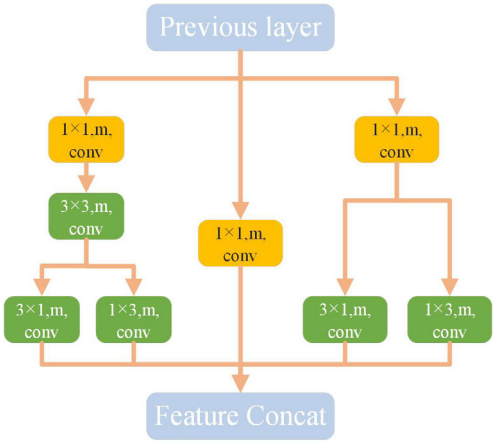

FIGURE 3. The framework of RI-Dense. 'm' means the number of channels.

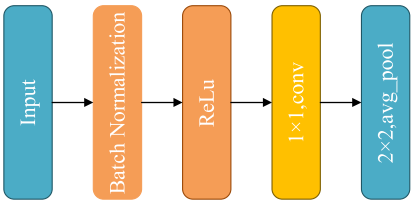

FIGURE 4. The structure of the Bottleneck Layer.

RI-Dense module for multi-feature layer connection, and bottleneck layer for dimension reduction. TABLE 1 describes the structure of the feature extraction network.

1) ROOT MODEL

In general, adjacent pixels own similar information, because of the large size of the input image. These pixels contain excessive redundant information. Dilated convolution [38] means adding holes that do not participate in the calculation of a standard convolution kernel. Through this procession, the receptive field becomes larger compared with the standard one, and the redundant information is effectively reduced. The root module contains three dilated convolution layers to eliminate redundant information and parameters. The details are shown in TABLE 1.

After processed by this three-layer dilated convolution, the receptive field becomes 13 × 13, while the standard convolution becomes 7 × 7. Compared with the standard convolution, the receptive field of dilated convolution increases by 2.4 times. Replacing the pooling operation by dilated convolution, the receptive field is increased without sacrificing the size of the feature map, and the redundant information in the image can be filtered out simultaneously.

2) RI-DENSE MODEL

This is a densely connected residual framework absorbing the idea of Inception ResNet to improve DenseNet. The framework is shown in FIGURE 3.

RI-Dense has three branches, all of which contain 1 × 1 convolution kernel to change dimensions. Depending on the shape of kernels, the feature scales extracted by different branches are various. 3 × 3 convolution kernels are used to extract the details of small targets. Furthermore, two 3 × 3 convolution kernels are used together to substitute a 5 × 5 convolution kernel to handle large targets. In order to further reduce the model parameters, the 3 × 3 convolution kernel in each branch is divided into a 1 × 3 and a 3 × 1 kernel. The dense connection between the RI-Dense models can ensure information transmission. Such an implementation benefits network training without consuming much computational resources.

3) BOTTLENECK LAYER

The bottleneck layer controls the dimension and scale of the RI-Dense module's output. Such a layer encourages the network to compress feature representations to the best fit in the available space, in order to get the best loss during training. FIGURE 4 shows the structure of the bottleneck layer. It consists of a 1 × 1 kernel and a 2 × 2 average pooling. They are added to reduce the channels of feature maps in the network, which otherwise tend to increase in each layer. This dimension alternation is achieved by using 1 × 1 kernels that have fewer output channels than input channels. A 1 × 1 convolution layer compresses the dimension, and then the scale of the feature map is compressed by a 2 × 2 average pooling layer. Bottleneck layers help by reducing the number of parameters in the network while allowing it to go deep and represent many feature maps.

C. RI-DECONV MODEL

One of the main contributions of SSD is that multiple feature layers are used for prediction. Nevertheless, this advantage

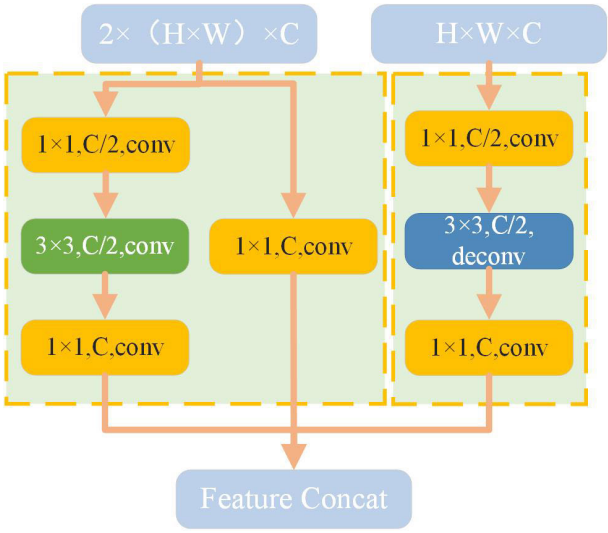

FIGURE 5. The framework of the RI-Deconv model.

is also limited due to the separate prediction of each feature map. Therefore, using multiple feature fusion can effectively increase the detection appearance of the network.

A feature fusion module named RI-Deconv is comprised and then constructs a feature pyramid with our proposed model. The RI-Deconv feature pyramid contains five fusion feature layers, and their scales are 2×2, 4×4, 8×8, 16×16, and 32×32 correspondingly. The structure is shown in FIGURE 5. In this phase, the feature map is restored step by step to the original image size by up-sampling. In order to acquire the deep-level semantic information and the shallow-level position information simultaneously, the RI-Deconv directly connects the corresponding size feature map from the deconvolution to the convolution in the up-sampling process.

As shown in FIGURE 5, the RI-Deconv module contains two inputs with different scales. The scale of the large feature map is twice that of the small feature layer. Both the two feature layers have the same number of channels. The large feature map has two branches: 1×1 shortcut and 3×3 convolution. These two paths compose a residual structure to address the problem of gradient disappearance. 3×3 convolution kernels provide a larger receptive field and increase the feature extraction ability of the network. The smaller feature map is expanded by four times after deconvolution. After that, the expanded feature map would fuse with the large feature map.

D. TARGET STITCHING

Because of the large size of pictures in the dataset, a series of 1024×1024 patches are cropped from the original images with a stride set to 512. Furthermore, the RIDNet takes the cropped blocks as the input for the model. Since the large-sized images are cropped into several patches, the objects are inevitably divided into several parts. A method for target stitching is designed to prevent targets from missing and repeating. Each fusion feature layer is detected and then maps the coordinates of results to the original image. Whether to fuse these prediction boxes is determined by their relative position.

Every pixel in SSD generates anchor boxes. The length of the default anchors are as follows:

$$S_k = S_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1) \tag{1}$$

where $k \in [1, m]$, m is the layer number, i.e., $m = 5$. $S_{\min} = 0.2$, $S_{\max} = 0.3$, $\min\_size = S_k$ and $\max\_size = S_{k+1}$ in layer K.

In order to enhance the model's identification ability, a series of aspect ratios are set for the anchor frame.

$$w_k = S_k\sqrt{R_n}, \quad h_k = \frac{S_k}{\sqrt{R_n}} \tag{2}$$

R_n means the ratio of anchors, and $R_n \in \{1, 2, 1/2, 3, 1/3\}$. A particular side length is added when $R_n = 1$:

$$w_k = h_k = \sqrt{S_k \times S_{k+1}} \tag{3}$$

The center of default anchor is $(\frac{a+0.5}{|f_k|}, \frac{b+0.5}{|f_k|})$, where f_k is the size of the k-th feature. $a, b \in \{0, 1, 2, \cdots, |f_k| - 1\}$. The mapping relationship between the anchor coordinate of the feature map and the original coordinate is as follows:

$$x_{\min} = \frac{c_x - \frac{w_b}{2}}{w_{\text{feature}}} w_{\text{img}} = \left(\frac{a+0.5}{|f_k|} - \frac{w_k}{2}\right) w_{\text{img}} \tag{4}$$

$$y_{\min} = \frac{c_y - \frac{h_b}{2}}{h_{\text{feature}}} h_{\text{img}} = \left(\frac{b+0.5}{|f_k|} - \frac{h_k}{2}\right) h_{\text{img}} \tag{5}$$

$$x_{\max} = \frac{c_x + \frac{w_b}{2}}{w_{\text{feature}}} w_{\text{img}} = \left(\frac{a+0.5}{|f_k|} + \frac{w_k}{2}\right) w_{\text{img}} \tag{6}$$

$$y_{\max} = \frac{c_y + \frac{h_b}{2}}{h_{\text{feature}}} h_{\text{img}} = \left(\frac{b+0.5}{|f_k|} + \frac{h_k}{2}\right) h_{\text{img}} \tag{7}$$

where (c_x, c_y) is the center of anchor coordinate on the feature map. w_b and h_b are the width and height of the anchor. w_{feature} and h_{feature} are the width and height of the feature map. w_{img} and h_{img} are the width and height of the original image. $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ are the coordinates of the upper left and lower right corner of the target on the original image.

An indicator I_{iou} is introduced to determine whether to fuse adjacent prediction boxes. When I_{iou} is higher than the threshold, two prediction boxes are fused. The threshold set in this paper is 0.4, and the class of fused prediction box is described as $Class_{fusion}$. FIGURE 6 illustrates the definitions of $I_{overlap}$ and I_{sum}.

$$I_{iou} = \frac{I_{overlap}}{I_{sum}} \tag{8}$$

$$I_{overlap} = \begin{cases} h_1 \cap h_2 \ (w_1 \cap w_2 = \varnothing) \\ w_1 \cap w_2 \ (h_1 \cap h_2 = \varnothing) \end{cases} \tag{9}$$

$$I_{sum} = \begin{cases} h_1 \cup h_2 \ (w_1 \cap w_2 = \varnothing) \\ w_1 \cup w_2 \ (h_1 \cap h_2 = \varnothing) \end{cases} \tag{10}$$

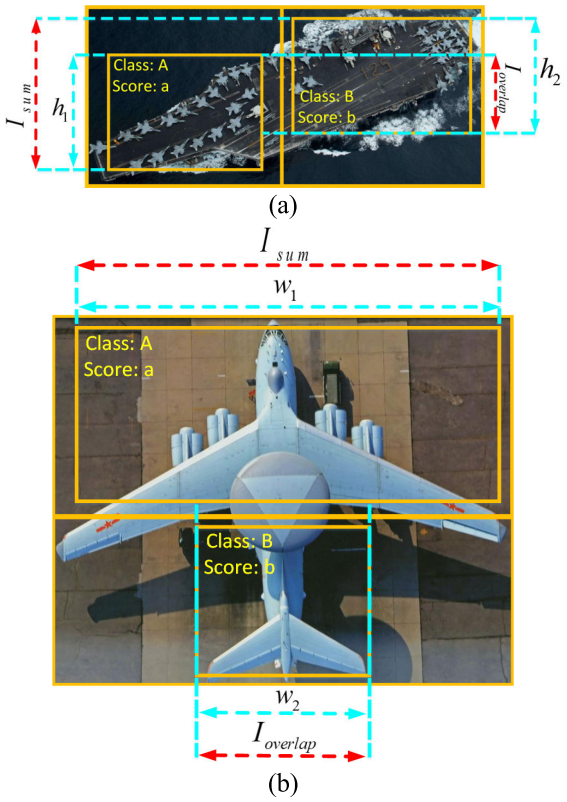

FIGURE 6. Two situations of I_{iou}. (a) and (b) show targets that have been divided into horizontal parts and vertical parts, respectively.

$$Class_{fusion} = \begin{cases} ClassA(Score_a > Score_b) \\ ClassB(Score_a < Score_b) \end{cases} \quad (11)$$

Firstly, the image pieces are stitched in the horizontal direction. After this operation, the image parts will become transverse strips, which will then be merged vertically. The divided targets will be stitched through this process.

IV. EXPERIMENTS AND ANALYSIS

A. MATERIALS

Experiments are implemented under Pytorch 1.0 framework by python language on a 64-bit computer with Ubuntu 18.04, CPU Intel i9-7900X CPU @ 3.3GHz, and NVIDIA Titan X 12G with CUDA9.2 and cuDNN7.5. The maximum training iteration is 120k. All parameters are randomly initialized with the xavier method. The model is fine-tuning by using SGD with 0.9 momentum, 0.0001 weight decay. The initial learning rate is set to 0.001, and it is decayed as cosine annealing for each batch. The batch size is set to 16.

The experiments are carried out on two public datasets: NWPU VHR-10 and DOTA. NWPU VHR-10 contains 800 aerial photos depicting 10 kinds of targets, among which 650 are targets and 150 are backgrounds. The samples of each category are shown in FIGURE 7. In the NWPU VHR-10 dataset, large-scale targets account for more than 15% of the image area, and small-scale targets account for less than 5%. With the target scales varying in an extensive range, it can test the network's ability to detect multi-scale

FIGURE 7. Sample display of NWPU VHR-10. This dataset contains 10 types of objects: Vehicle, Basketball Court, Baseball Diamond, Tennis Court, Ground Track Field, Airplane, Bridge, Storage Tank, Ship, and Harbor.

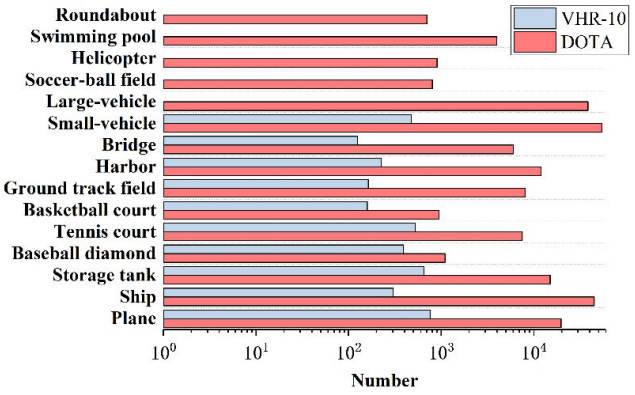

FIGURE 8. The numbers of samples in both datasets.

targets and small targets. 200 images are randomly selected for testing, and the rest 600 are used as training sets in the experiment. The operations of rotation (90 °, 180 °, 270 °), flipping (horizontal flipping, vertical flipping) and copying are carried out in order to expand the number of samples. FIGURE 8 shows the number of samples on both datasets.

The images in the DOTA dataset come from different platforms, with their size ranging from 600 × 600 to 4000 × 4000. The DOTA dataset can be classified into 15 categories, and the number of instances in it is more than 180,000, much larger than that in the NWPU VHR-10 dataset. DOTA includes targets of different scales with a high spatial resolution, which can better test the generalization performance and robustness of the model. FIGURE 9 shows the data samples.

B. COMPARISON WITH OTHER METHODS

The proposed method is compared with several popular models to verify the effectiveness of the RIDNet,

FIGURE 9. Sample display of DOTA. The objects in DOTA dataset occupy a small proportion in the images, and some objects are larger than others. For instance, the harbors are larger than the ships in the bottom left image.

TABLE 3. Comparisons of different models on NWPU VHR-10. The abbreviations in the TABLE is listed as: FRCN-FASTER-RCNN, ST-STORAGE TANK, BD-BASEBALL DIAMOND, TC-TENNIS COURT, BC-BASKETBALL COURT, GTF-GROUND TRACK FIELD.

Method	Airplane	Ship	ST	BD	TC	BC	GTF	Harbor	Bridge	Vehicle	mAP
FRCN[7]	72.51	67.47	52.86	**96.31**	72.77	78.45	84.39	68.79	67.62	63.80	72.50
YOLOV3[13]	**87.91**	81.73	**80.84**	90.96	91.32	93.56	**99.76**	81.17	**88.69**	79.76	**87.58**
SSD[12]	83.65	88.56	50.87	89.39	77.65	86.73	88.64	82.10	84.74	64.47	79.68
FSSD[25]	84.42	88.67	51.43	89.67	78.17	87.21	89.61	82.53	85.02	65.09	80.18
S3FD[26]	85.32	84.14	57.55	90.27	78.94	88.11	86.57	81.99	81.56	68.19	80.26
R-FCN[8]	82.01	**88.98**	74.21	89.05	81.49	84.88	82.84	83.67	83.61	77.65	82.84
Ours	83.76	82.83	78.21	91.09	**93.40**	**94.91**	93.68	**84.67**	86.67	**84.58**	87.38

including FRCN [7], YOLOV3 [13], SSD [12], FSSD [25], S3FD [26] and RFCN [8]. mAP (mean Average Precision) is used as a measure of performance. The results on the NWPU VHR-10 are shown in TABLE 3 and Figure 10. As displayed in TABLE 3, our method achieves a satisfactory performance in terms of mAP values on NWPU VHR-10 dataset. Figure 10 shows that the detection accuracy of different methods varies with the number of iterations. The mAP is recorded every 5000 iterations. Figure 12 displays the detection results of various methods.

The results of above-mentioned methods on DOTA database are shown in TABLE 4 and Figure 11. The mAP of the RIDNet is slightly lower than that of YOLOV3 on NWPU VHR-10. However, the mAP of the RIDNet is higher than that of YOLOV3 on DOTA. The reason for this phenomenon is that small objects occupy a large proportion on DOTA, which increases the difficulty of detection [39]. FIGURE 13 displays the detection results of RIDNet on DOTA. Furthermore, the time-consuming of the RIDNet is 47.4 ms. The proposed method covers a presentable computation cost while achieving a better detection accuracy with small model sizes.

C. ABLATION EXPERIMENT

Ablation experiments are conducted to verify the effectiveness of each technique proposed in this paper. The results are shown in TABLE 5 and TABLE 6.

TABLE 4. Comparisons of different models on DOTA. The abbreviations in the table is listed as: FRCN-FASTER-RCNN PL-PLANE, BD–Baseball diamond, BR-BRIDGE, GTF–Ground field track, SV–Small vehicle, LV–Large vehicle, SH-SHIP, TC–Tennis court, BC–Basketball court, ST–Storage tank, SBF–Soccer-ball field, RA–Roundabout, HA-HARBOR, SP–Swimming pool, and HC–Helicopter.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP	Time /ms	Model-Size /MB
FRCN	79.13	76.41	30.77	65.81	48.92	54.33	48.69	**82.30**	70.83	55.61	57.25	46.20	64.84	58.46	34.58	58.28	97.2	286.3
YOLO-V3	83.52	**77.55**	32.86	68.13	53.66	52.49	60.04	71.65	**75.05**	65.59	57.02	49.81	61.69	56.46	**41.85**	60.49	58.4	237
SSD	54.12	32.17	16.23	20.84	8.13	30.73	26.03	72.88	26.41	45.63	12.42	33.63	15.96	11.02	12.21	27.89	**20.6**	97.4
FSSD	55.34	32.37	16.65	22.01	8.15	31.18	26.72	73.12	26.84	47.06	13.94	34.51	16.76	11.71	12.94	28.62	25.76	123.6
S3FD	55.14	34.08	14.83	21.51	10.34	34.16	26.54	73.52	28.39	48.62	15.76	36.42	16.61	14.26	16.38	29.77	32.7	85.7
R-FCN	81.03	60.22	33.29	57.82	51.79	48.14	51.29	69.31	52.37	68.25	42.15	49.24	46.38	52.27	35.54	53.27	65.8	307.2
Ours	**86.34**	74.65	**34.09**	**73.36**	**57.32**	**60.63**	**73.29**	78.29	74.11	**84.47**	**68.20**	**55.83**	**71.16**	**61.92**	37.81	**66.10**	47.4	**39.4**

TABLE 5. mAP of different feature fusion.

Feature Layer					Metric		
2×2	4×4	8×8	16×16	32×32	mAP	$mAP^{R_{IoU}=0.50}$	$mAP^{R_{IoU}=0.75}$
√	√	√	√	√	87.38	95.42	89.51
	√	√	√	√	86.72	93.19	86.11
√	√	√	√		79.33	87.83	81.46
	√	√	√		77.57	84.62	79.73

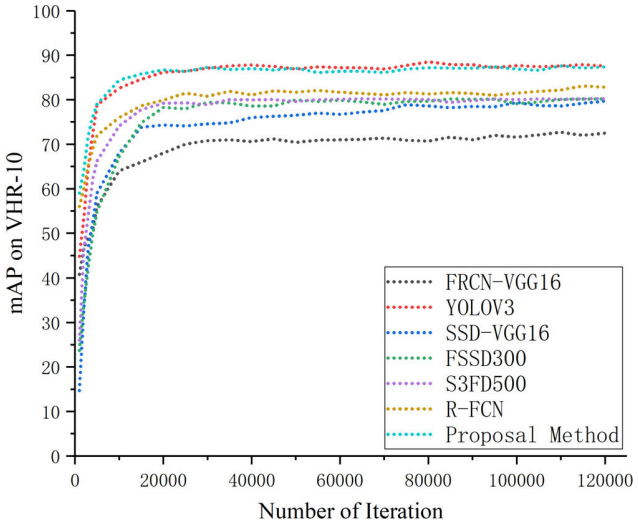

FIGURE 10. mAP-Iteration curves of seven models on NWPU VHR-10 dataset.

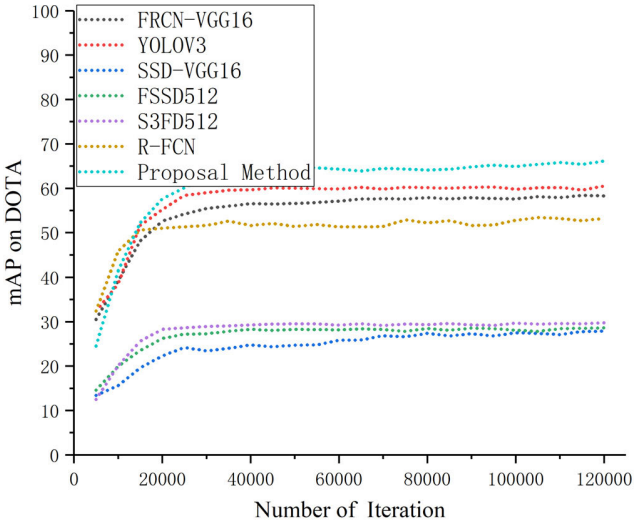

FIGURE 11. mAP-Iteration curves of seven models on DOTA dataset.

Experiment results show that the fusion of different feature layers has a corresponding effect on the detection ability. The detection result is the best when all five feature layers participate in the fusion. However, it is worth noting that the mAP is only slightly less than that of five fused feature maps when the other four feature maps are fused without the 2×2 feature map. It is because that after multiple convolutions and down-samplings, there leave almost no features on the 2×2 feature map. Thus, the 2×2 feature map has little effect on improving the detection accuracy.

The ablation experiment is conducted on the NWPU VHR-10 to test the effectiveness of proposed models. The experiment results are shown in TABLE 6. When DenseNet is used as the feature extraction network, RI-Deconv achieves the highest detection accuracy. When the RI-Dense structure is used as the feature extraction network, the detection accuracy will be improved. The RI-Deconv structure maintains the highest detection accuracy in each experiment, which shows that the proposed models are effective and efficient when detecting.

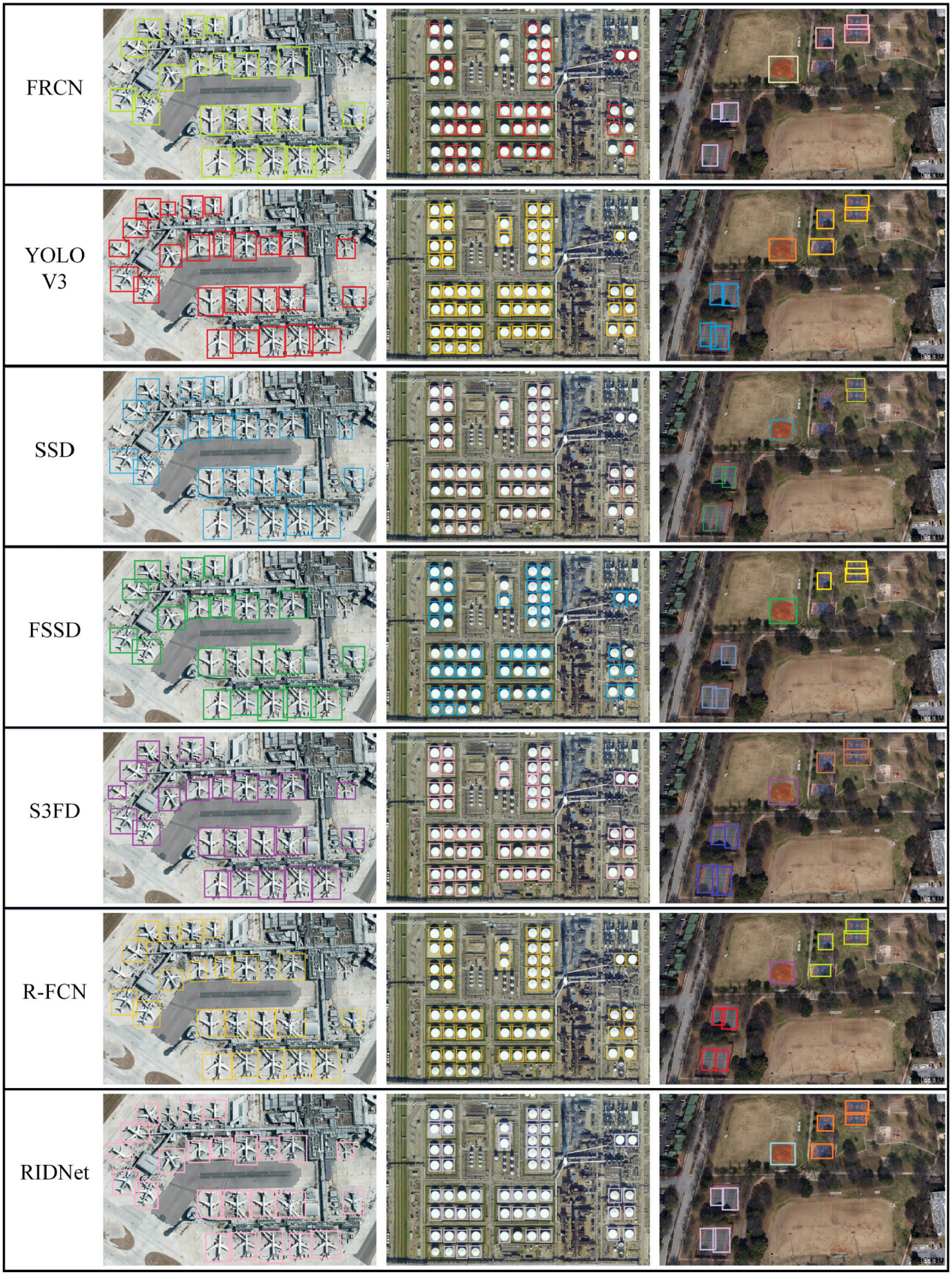

FIGURE 12. Samples of detection results for different models.

FIGURE 13. Detection results of the RIDNet on DOTA.

TABLE 6. Results of the ablation experiment. PFH is the pyramidal feature hierarchy of SSD; FPN is the feature pyramid network; RI-Deconv is the network proposed in this paper.

feature extraction network	PFH	FPN	RI-Deconv	mAP
DenseNet	√			74.6
		√		77.5
			√	**82.4**
RI-Dense	√			80.6
		√		82.5
			√	**87.4**

V. CONCLUSION

In this paper, the RI-Dense model and RI-Deconv model are proposed for small targets and multi-scale targets detection. Based on these models, a lightweight detection network RIDNet is designed, absorbing the ideas of deconvolution, DenseNet, InceptionNet, and ResNet. The RI-Dense model improves the efficiency of feature extraction, addresses the problem of gradient disappearance, and achieves high detection accuracy with fewer parameters. RI-Deconv module adds semantic information from deep layers and detailed information from shallow layers to the fusion layers, which can improve the performance of multi-scale detection. It fully utilizes the information extracted from multiple feature layers to improve detection accuracy. With a dense residual structure, our network is able to deal with objects of different sizes and improve the detection accuracy for small and weak objects. The network can handle large-sized images by cropping original input images. Moreover, the target stitching method guarantees that divided targets will be stitched back.

After experiments on two public datasets, it is found that the proposed algorithm, the RIDNet, has a better performance

in detection compared with other popular detection algorithms. Moreover, RIDNet is lightweight enough to be deployed on UAV. Ablation experiments also show that proposed models can effectively improve detection accuracy.

REFERENCES

[1] R. S. de Moraes and E. P. de Freitas, "Multi-UAV based crowd monitoring system," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 2, pp. 1332–1345, Apr. 2020.

[2] C. P. Schwegmann, W. Kleynhans, and B. P. Salmon, "Synthetic aperture radar ship detection using Haar-like features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 154–158, Feb. 2017.

[3] T.-B. Xu, G.-L. Cheng, J. Yang, and C.-L. Liu, "Fast aircraft detection using end-to-end fully convolutional network," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Oct. 2016, pp. 139–143.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[7] 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.

[8] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. Int. IEEE Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[10] 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. (CVPR)*, Jun. 2016, pp. 779–788.

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[12] W. Liu, "SSD: Single shot multibox detector," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 21–37.

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

[14] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[15] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[16] G. Wang, X. Wang, B. Fan, and C. Pan, "Feature extraction by rotation-invariant matrix representation for object detection in aerial image," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 6, pp. 851–855, Jun. 2017.

[17] T. Yang, J. Li, J. Yu, S. Wang, and Y. Zhang, "Diverse scene stitching from a large-scale aerial video dataset," *Remote Sens.*, vol. 7, no. 6, pp. 6932–6949, 2015.

[18] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1222–1230.

[19] T. Yang, X. Wang, B. Yao, J. Li, Y. Zhang, Z. He, and W. Duan, "Small moving vehicle detection in a satellite video of an urban area," *Sensors*, vol. 16, no. 9, p. 1528, 2016.

[20] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals using stereo imagery for accurate object class detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1259–1272, May 2018.

[21] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused SSD: Fast detection for small objects," *Proc. SPIE*, vol. 10615, Apr. 2018, Art. no. 106151E.

[22] X. Liu, T. Yang, and J. Li, "Real-time ground vehicle detection in aerial infrared imagery based on convolutional neural network," *Electronics*, vol. 7, no. 6, p. 78, 2018.

[23] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 354–370.

[24] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: http://arxiv.org/abs/1701.06659

[25] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*. [Online]. Available: http://arxiv.org/abs/1712.00960

[26] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3FD: Single shot scale-invariant face detector," in *Proc. Int. IEEE Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 192–201.

[27] W. Shi, J. Jiang, S. Bao, and D. Tan, "CISPNet: Automatic detection of remote sensing images from Google earth in complex scenes based on context information scene perception," *Appl. Sci.*, vol. 9, no. 22, p. 4836, 2019.

[28] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: Rotation-invariant and Fisher discriminative convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2884–2893.

[29] C. Liu, D. Zeng, H. Wu, Y. Wang, S. Jia, and L. Xin, "Urban land cover classification of high-resolution aerial imagery using a relation-enhanced multiscale convolutional network," *Remote Sens.*, vol. 12, no. 2, p. 311, 2020.

[30] C. Zhang, Y. Chen, X. Yang, S. Gao, F. Li, A. Kong, D. Zu, and L. Sun, "Improved remote sensing image classification based on multi-scale feature fusion," *Remote Sens.*, vol. 12, no. 2, p. 213, 2020.

[31] Q. Bi, K. Qin, H. Zhang, Z. Li, and K. Xu, "RADC-Net: A residual attention based convolution network for aerial scene classification," *Neurocomputing*, vol. 377, pp. 345–359, Feb. 2020.

[32] P. Zhou, G. Cheng, Z. Liu, S. Bu, and X. Hu, "Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping," *Multidimensional Syst. Signal Process.*, vol. 27, no. 4, pp. 925–944, Oct. 2016.

[33] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[34] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[35] S. Zhang, X. Wang, Z. Lei, and S. Z. Li, "Faceboxes: A CPU real-time and accurate unconstrained face detector," *Neurocomputing*, vol. 364, pp. 297–309, Oct. 2019.

[36] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with SVD networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5832–5845, Oct. 2016.

[37] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, "Efficient saliency-based object detection in remote sensing images using deep belief networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 137–141, Feb. 2016.

[38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[39] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. ACCV*. Cham, Switzerland: Springer, 2018, pp. 150–165.

MIAOHUI ZHANG received the B.S. degree in control theory and control engineering from Northeastern University, in 2002, the master's degree from the Graduate University of the Chinese Academy of Sciences, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China, in 2013. His current research interests include pedestrian detection and re-identification, abnormal behavior analysis, and video content understanding.

KANGNING PANG received the B.S. degree in automation from the Henan University of Science and Technology, China, in 2017. He is currently pursuing the M.S. degree with the School of Computer and Information Engineering, Henan University, China. His current interests include pattern recognition and computer vision.

CHENGCHENG GAO is currently pursuing the M.S. degree with the School of Computer and Information Engineering, Henan University. His current interests include pattern recognition and computer vision.

MING XIN received the B.S. degree in information management and information system from Southwest University, in 2002, and the M.S. degree in applied mathematics from Henan University, in 2008. She is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, China. She joined the School of Computer and Information Engineering, Henan University, in 2002, where she has been an Associate Professor, since 2013. Her current research interests include moving object detection and tracking and object recognition.