IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 28, 2020, accepted May 10, 2020, date of publication May 14, 2020, date of current version May 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994379

Robust Object Detection in Aerial Imagery Based on Multi-Scale Detector and Soft Densely Connected

MIAOHUI ZHANG[1,2], BO ZHANG[1], MENGYA LIU[3], AND MING XIN[4]

[1]Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475001, China
[2]Institute of Data and Knowledge Engineering, Henan University, Kaifeng 475001, China
[3]School of Physics and Electronics, Henan University, Kaifeng 475001, China
[4]School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Corresponding author: Ming Xin (xinming_henu@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 Object detection in aerial images is vital for autonomous guidance, navigation and control, and situational awareness. However, there are still many challenges facing researchers in this filed, including the target scales, the perspectives in taking pictures, and the highly complex background. The present paper introduces a robust object detector which is optimized for handling with multi-scale objects and the overhead capturing perspective object instances in aerial images. Firstly, in the feature extraction stage, an effective multi-scale detector (MSD) is designed to search for objects with different scales in feature maps. After that, when detecting a small target from a cluttered background, both the shallow and deep layer features are densely connected by the deconvolution after tackling the issues of low dimensionality in deep layers and inadequate representation of small objects. In the experiments part, we analyze the impacts of the above mentioned components on the model and make a comparison between the method at issue and other state-of-the-art approaches on two publicly-available datasets captured by satellites and high-altitude UAVs. The results show that the proposed method, which is applicable to a wider range of aerial images, is more effective and robust.

INDEX TERMS Object detection, aerial images, multi-scale detection, small object relative scale (ORS).

I. INTRODUCTION

As airborne cameras and remote sensing systems keep developing, it is more and more common for high-resolution aerial images that are captured by unmanned airborne vehicles (UAVs) and satellites to provide data for researchers. As a result, object detection in aerial images becomes an essential technique in the attempt to automatically obtain instance-level information [1], [2]. Those machine vision systems and algorithms based on object detection, widely used in many critical applications such as military reconnaissance and intelligent transportation, are often applied to collecting information about the areas surrounding an object. But due to the inevitable problems of small object sizes, large camera motions and occlusions, their application in aerial images remains a challenging task [3], [4]. What's more, the object scales vary greatly in aerial images due to different shooting heights and viewpoints of the airborne camera, making the object detection more demanding.

Recently, after deep-learning-based methods are introduced, researches on object detection have seen great progress [5]–[7]. These technologies and algorithms, however, generally suffer from poor localization accuracy or missed detections when applied into aerial images, in that their network architectures are designed and developed for general datasets which are different from aerial images, especially in terms of the object sizes and image fractions occupied by an object. Besides, aerial images usually include not only objects of fixed shapes and scales, such as ships, airplanes and vehicles, but also that of varied shapes and scales, such as bridges and harbors. Furthermore, aerial images captured from a top-down view store much less information than general images captured from the horizontal point of view. One of the major reasons of that is the

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

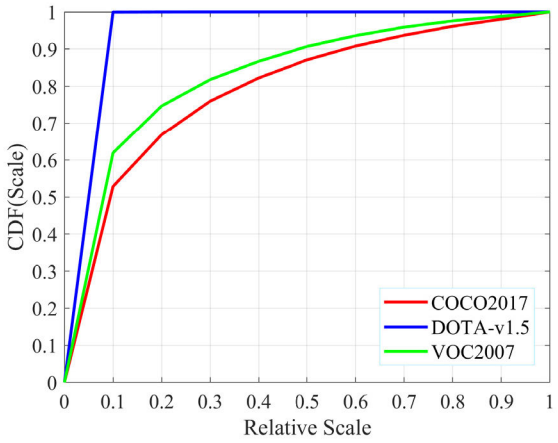
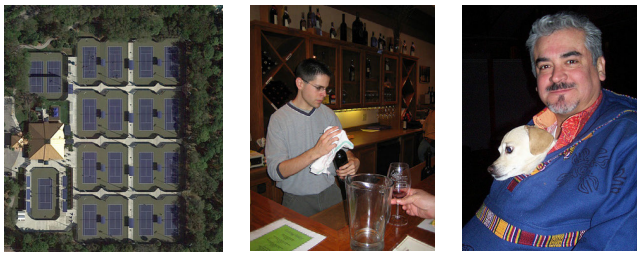


FIGURE 1. The cumulative distribution of the instance target area proportion of the dataset.



(a) DOTA-v1.5 (b) COCO (c) VOC

FIGURE 2. Examples from DOTA-v1.5 and general datasets.

remote distance which, even with an imaging equipment of a high-resolution, inevitably limits the details shown in aerial images. The resolution with the ground sampling distance (GSD) of images captured by satellites usually is lower than 1m. Generally speaking, it is from objects with large distances rather than sensors that small object cases in aerial images are derived. Therefore, object detection is more challenging for aerial images than natural images [8]–[11].

Though very efficient in detecting objects, deep learning methods are laborious when processing aerial images. COCO and VOC are general object detection datasets captured in natural scenes, widely used to evaluate the performance of object detection models. Most images in these datasets are shot in the horizontal direction and a close range. The instance target occupies a large proportion in the image, as shown in Fig. 2(b) and Fig. 2(c). Different from the COCO and VOC datasets, DOTA-v1.5 is an aerial image dataset captured by satellite, as shown in Fig. 2(a). The images in this dataset are basically shot from an overlook view, and the instances are relatively small. Therefore, Aerial images are facing serious challenges regarding to the targeted objects, such as the small object relative scale (ORS), the low-resolution of objects and the variations of object scale. The Cumulative Distribution Functions (CDF) of DOTA-v1.5 and another two publicly available datasets are displayed in Fig. 1. The curve of DOTA-v1.5 is close to the top-left corner, and the share of objects taking up less than 1% of the whole image area is over 90%.

This paper puts forward a robust object detection method to process aerial images. The overall architecture of the proposed model is shown in Fig. 3. Firstly, inspired by attention mechanisms in the human visual system, we designed a multi-scale detector (MSD) with different convolution kernel sizes and residual structures. Secondly, we maximized the mutual information between multi-scale objects and features by combining the MSD with top-down convolution mechanisms which are specifically designed for feature extraction. The suggested model is made up of four major parts. In Fig. 3, the gray blocks in the bottom left corner represent the backbone network, such as VGG-16 [12], Resnet-50 [13], and Darknet-53. These backbone networks contain the traditional convolution and pooling operation, which are widely used for the preliminary feature extraction in the typical CNN model. The second part is a bottom-up pathway utilizing the layer in the convolutional blocks of the backbone network with MSD. These blocks get smaller from the bottom-up because of pooling layers. The blocks with different colors represent different features of various scales, with the blue blocks representing MSD. The third part is a top-down pathway. In this part, feature blocks of the same size in part 2 are lateral connected and the blocks with the same size are labelled with the same color. And the last part is the predictor head which contains the fully connected layers and loss functions used to predict regression boxes and class scores.

For convenience, a glossary is provided in Table 1 defining important terms used in this paper.

TABLE 1. Glossary.

Initials	Definition
MSD	Multi-scale detector
SNR	Signal-noise ratio
UAVs	Unmanned airborne vehicles
GSD	Ground sampling distance
CDF	Cumulative distribution functions
AR	Average recall
PRC	Precision-recall curve
mAP	Mean average precision
UA	User accuracy
PA	Producer's accuracy
LSK+VLAD	Vector of locally aggregated descriptors based on local steering kernel
FCN (feature fusion)	Fully Convolutional Neural Networks with feature fusion

II. RELATED WORK

In general, object detection in aerial images is conducted with a sliding window method which involves some hand-crafted features and a classifier or a cascade of classifiers [14], [15]. However, there exist shortcomings with this approach. First of all, those hand-crafted or shallow-learning based features might exert adverse impacts on the representation ability and effectiveness of aerial detection. Secondly, the sliding

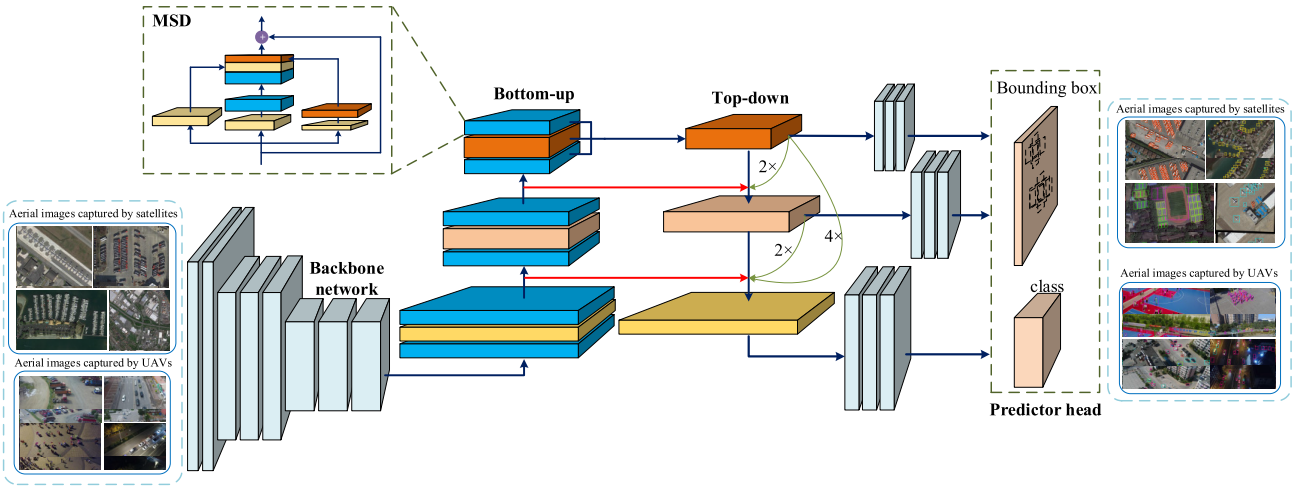


FIGURE 3. The overall architecture of proposed model.

window approach may impose a computation burden on detecting. Those features are so low-level that they are not distinct enough, therefore earlier methods often meet challenges imposed by interclass variations. In recent years, deep learning methods resorting to features obtained by convolutional neural networks (CNN) have greatly increased the accuracy of detection on benchmark datasets compared with traditional ones utilizing hand-crafted features for detecting objects in aerial images [16]–[25]. In addition, the existing deep-learning based methods, as their name suggests, are able to self-learn features from given data in deep layers, an ability enabling them to perform better.

CNN-based object-detection methods, for example Region-based Convolutional Neural Network (R-CNN), Fast R-CNN and Faster R-CNN [26], often go through two successive stages: bounding box proposal and classification. However, they have two major shortcomings when applied to aerial images. Firstly, most of these methods consist of two processes: proposal generation and object classification, making the detection process too complex to operate in a high speed. Secondly, in the detecting process, they utilize axis-aligned bounding boxes which include both objects and backgrounds, making it more difficult to detect in an accurate manner [27]. Ref. [28] makes an elaboration on the potential and limitations of Faster R-CNN when it is used for tracking medium-sized objects in cases of pedestrian detection. Those methods of one stage, such as You Only Look Once (YOLO) [29], Single Shot Multi-Box Detector (SSD) [30] and Retinanet [31], carry out the bounding box and classification processes simultaneously [32]. These two kinds of methods, however, have some differences in their performance regarding to the computing speed and detection accuracy, factors that can also be influenced by the type of CNN backbone employed, such as Googlenet [33], VGGNet [12], Resnet [13], Darknet-53 [29] or Densenet [34].

Objects captured by aerial images, with low signal-noise ratio (SNR), provide only limited visual information for researchers, making the targeted objects difficult to be distinguished from the cluttered backgrounds. In addition, the aerial images collected by satellites or UAVs are considerably different. In ref. [18], [20], [21], [27], researchers show the optimized structure and detection performance of Faster R-CNN, Oriented SSD, feature pyramid networks (FPN) and FCN with feature fusion for aerial images captured by satellites. These network architectures are explored and optimized for satellite datasets which increased computational costs. In the image datasets captured by UAVs [1], [4], [19], [22], objects are typically centered and occupy a fixed fraction in the pictures. These network structures are generally different from above methods in that they are able to reach an area of interest flexibly and take pictures with various levels of details. Liu and Ding propose a modified image representation model, a vector of locally aggregated descriptors based on local steering kernel (LSK+VLAD) for detecting vehicles in aerial images collected by satellites and UAVs [35]. However, this model still does not live up to expectations when targeted vehicles are partially occluded or are similar to other objects in the backgrounds in shape.

The remaining of the paper is organized as follows: Section 3 describes the main algorithms used, including MSD bottleneck structure, Multi-scale feature fusion, and loss function. Section 4 gives the experiment results. The last section makes a conclusion and recommendations for future research work.

III. MATERIALS AND METHODS

A. MULTI-SCALE DETECTOR

Traditional convolutional neural network architectures for object detection are explored and optimized for particular datasets [36], [37], resulting in poor localization accuracy or missed detections in aerial images. In general, Max Pooling and Spatial Pyramid Pooling are used to solve the multi-scale problem in bottom-up feature extraction. Many objects with small scales are missed out during the Max Pooling process. Since the convolution kernel size is relatively simple in the traditional backbone, such as VGG-16, Resnet-50, and

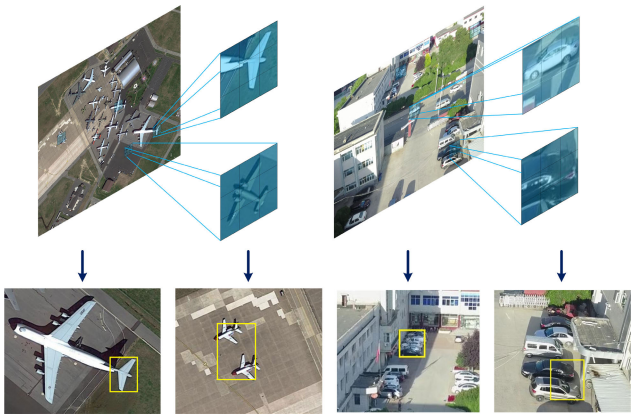
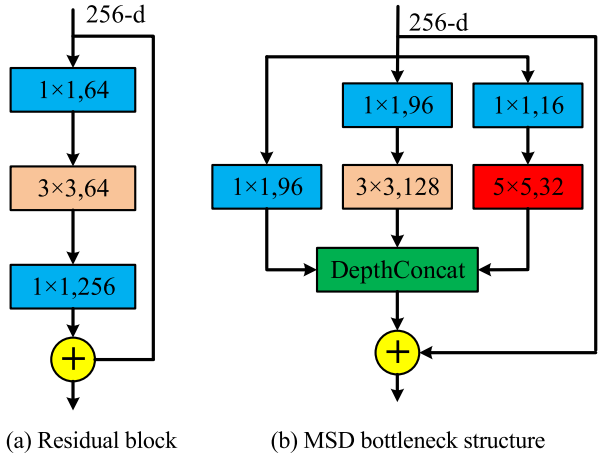


FIGURE 4. False detections caused by single convolution kernel size.



(a) Residual block (b) MSD bottleneck structure

FIGURE 5. Comparison between the Residual block and the proposed MSD structure.

Darknet-53, instance features in feature maps are hard to extract completely. As a result, the problems of poor localization and false detections are easily to come up, as illustrated in Fig. 4.

The residual block displayed in Fig. 5 (a) is a typical bottleneck structure which is usually employed in CNNs backbone [13]. The residual block usually uses a 3×3 convolution layer to extract features and two 1×1 convolution layers with shortcut connections. Shortcut connections are those skipping one or more layers. This design can effectively solve the problem of gradient disappearance in the deep convolutional neural network. Instead of using a 3×3 convolution layer to extract features as practiced in the residual block, we look for other architectures which have a stronger multi-scale feature extraction ability while also keeping a good computational load. To be specific, we add a group of 5×5 filters and a group of 1×1 filters in the block and connect them in a hierarchical residual-like manner, as shown in Fig.5 (b). Here, 1×1 filters are used mainly as dimension reduction modules to remove computational bottlenecks that would otherwise limit the size of our networks. This allows for not just increasing the depth, but also the width of our networks without significant performance penalty. MSD is an architecture combining all of the layers while connecting all the output filter banks together into a single output vector which constitutes the input of the next stage. Since our proposed bottleneck structure employs different scales of convolution filters to extract features, the multi-scale problem in aerial images can be effectively solved.

The main idea behind the MSD structure is to seek an optimal local sparse structure in a convolution neural network, so as to effectively extract and cover instances that are of multiple scales and densely packed. We propose a layer-by-layer model in which the correlation components of the layer input shall be analyzed firstly and then split into units of a high correlation. It is assumed that every such unit matches certain region of the input image, and these units will be grouped into 1×1 filter, 3×3 filter, and 5×5 filter banks. It means that in lower layers, the units that are correlated with each other will congregate in local regions and be covered by a layer of 1×1 convolutions. However, in deep layers, the number of more spatially spread out clusters covered by convolutions over larger patches will be smaller, and the number of patches over larger regions will also be decreased. The 5×5 filter in MSD can be combined with the 3×3 filter to extract features from some difficult samples, such as small and overlapped instances. This convolution combination covers multi-scale objects effectively, and reduces potential false detection and poor localization in Figure 4. As shown in Fig. 3, MSD is dispersed in different sizes of convolution groups in the bottom-up pathway, rather than simply repeated as a complex backbone in space. Furthermore, this decision is more based on simple computation rather than necessity. It also denotes that the proposed model is a combination of all layers with the output filter banks gathered within a single output vector which forms the next input. Fig. 6 shows multi-scale objects in aerial images.

B. MULTI-SCALE FEATURE FUSION

Feature pyramids composed of multi-scale features are the recognition systems' basic element to detect objects of multiple scales. Lin *et al.* [42] put forward a feature pyramid network, an in-network feature hierarchy that draws feature maps of different spatial resolutions but leads to huge semantic gaps under the influence of various depths. However, these high-resolution maps, showing low-level features, can have an adverse impact on their representational capacity when performing the detection task. To solve this problem, we make use of a model that is characterized by low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. This architecture has a similar result with that of FPN, but the process of up-sampling is done in different ways with a soft-dense connection that is rich in semantics at all levels, and that is built quickly from a single input image scale. Dense connection is a feed-forward connection fashion which connects one layer to all other layers [34]. Traditional dense connections map each layer to the inputs of subsequent layers. In order to transfer the multi-scale features in the bottom-up pathway, we used a kind of soft-dense connection

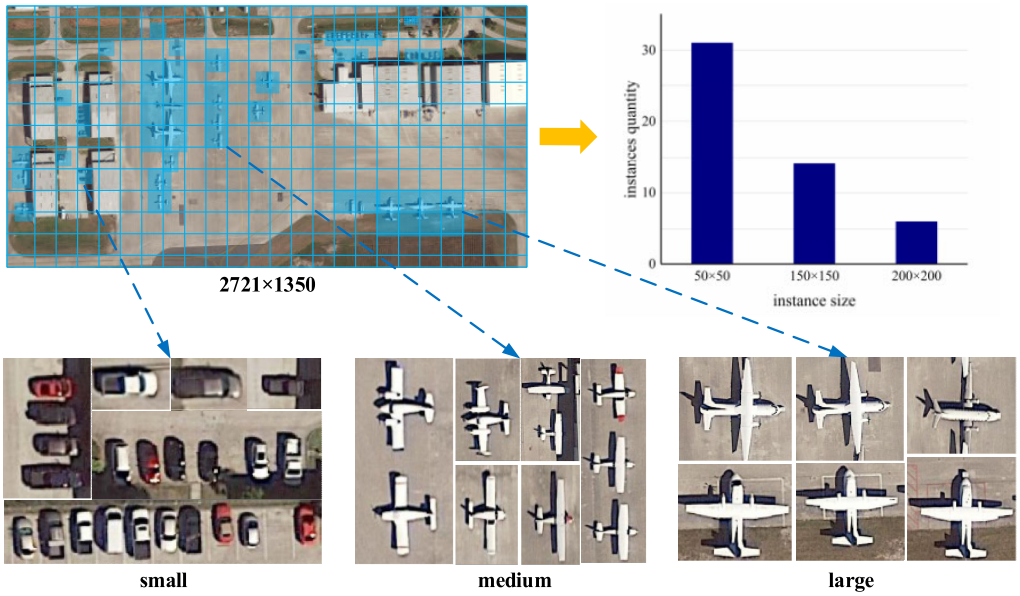


FIGURE 6. Multi-scale objects in aerial image.

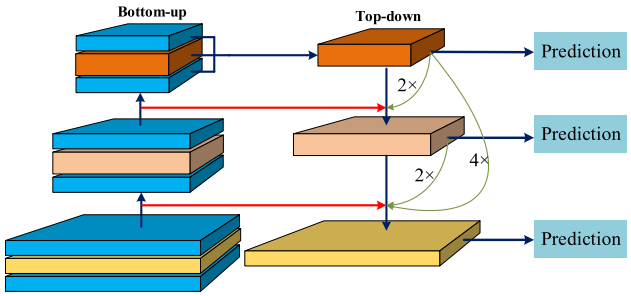


FIGURE 7. Feature fusion by MSD with soft-dense connection.

strategy whose calculation cost is relatively small. The principle advantage is that the multi-scale feature representation it produces is semantically strong at all levels, including the high-resolution levels.

Fig. 7 shows how the feature fusion works. Since this convolution parameter setting first proposed in darknet-53 can effectively keep the size of feature maps unchanged, we adopt the same convolution parameter setting in the proposed backbone network in this paper. The parameter setting of convolution in Darknet-53 can keep the feature size unchanged after convolution. It computes a feature hierarchy which contains multiple scales of feature maps with a scaling step of 2. There are often a large number of layers that create same-sized output maps in the same network stage. The output of the last MSD layer in each stage is chosen as a stage output of feature maps which is also called the reference set of feature maps.

The top-down pathway come into higher resolution and semantically stronger features by up-sampling from higher pyramid levels. The traditional up sampling method usually leads to the discontinuity between pixels. Bilinear interpolation can keep the continuity between adjacent pixels and avoid the gradient decrease caused by resolution decrease. The calculation of bilinear interpolation is defined as follows:

$$P = \frac{P_1 - P_2}{y_1 - y_0} \times v + P_1 \tag{1}$$

where P is the pixel value of the interpolation point. P_1 and P_2 are coordinates of the interpolation points in x direction, and they are defined as follows:

$$P_1 = \frac{f\,(x_1, y_0) - f\,(x_0, y_0)}{x_1 - x_0} \times u + f\,(x_0, y_0) \tag{2}$$

$$P_2 = \frac{f\,(x_1, y_1) - f\,(x_0, y_1)}{x_1 - x_0} \times u + f\,(x_0, y_1) \tag{3}$$

where the coordinate of P is $(i + u, j + v)$, with i, j belonging to the integer part and u, v to the decimal part. $f\,(x_0, y_0)$ and $f\,(x_1, y_1)$ are pixel values of P in the upper left and lower right, respectively. Then, a top-down model with lateral connections is constructed to build high-level semantic feature maps at all scales. These features are enhanced by that from the bottom-up pathway via lateral connections, each of which merges the feature maps with the same spatial size from the bottom-up pathway and that from the top-down pathway together.

C. LOSS FUNCTION

The joint loss function is commonly used in one-stage object detection neural networks. It is the simple addition of differences, including coordinate errors, confidence errors, and classification errors. The loss function is represented as follows:

$$Loss = \sum_{i=0}^{S^2} coordErr + confErr + clsErr \tag{4}$$

where S^2 is the number of grids. Coordinate errors consist of central coordinate errors and width-height coordinate errors. The coordinate errors are defined as follows:

$$coordErr = coordErr_{central} + coordErr_{WH} \tag{5}$$

where the central coordinate errors and width-height coordinate errors are defined as follows:

$$coordErr_{central} = \sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\left(x_i-\hat{x}_i\right)^2+\left(y_i-\hat{y}_i\right)^2\right] \quad (6)$$

$$coordErr_{WH} = \sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj} \times\left[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right] \quad (7)$$

where B represents the number of anchor boxes calculated by K-means in each cell; (x, y) is the center coordinates of each cell;w, h are the width and height of the prediction box; I_{ij}^{obj} tells whether there is a pedestrian object in the j_{th} prediction frame of the i_{th} cell: if a true target exists, the value is 1, otherwise the value is 0. The confidence errors can be expressed by:

$$confErr = -\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left[\begin{matrix} I_{ij}^{obj}+ \\ \lambda_{noobj} I_{ij}^{noobj}\end{matrix}\right] \times\left[\begin{matrix}\hat{C}_i^j\log\left(C_i^j\right)+ \\ \left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\end{matrix}\right] \quad (8)$$

where λ_{noobj} is the weight of the classification loss function, and it equals to 0.5. The classification error can be expressed by:

$$clsErr = \sum_{i=0}^{S^2} I_{ij}^{obj}\sum_{c\in classes}\left[\begin{matrix}\hat{P}_i^j\log\left(P_i^j\right)+ \\ \left(1-\hat{P}_i^j\right)\log\left(1-P_i^j\right)\end{matrix}\right] \quad (9)$$

Among which C_i^j is used for category determination, $C_i^j = 1$ is equivalent to classification correctness, otherwise, it is equal to 0, and P_i^j is the confidence degree. The confidence error and classification error are calculated by cross-entropy function, and $\left(\hat{x}, \hat{y}, \hat{w}, \hat{h}, \hat{C}, \hat{P}\right)$ is the corresponding predicted value as demonstrated in equations (5), (7), (8) and (9).

IV. EXPERIMENTS

A. DATASETS AND EVALUATION METRICS

The publicly available DOTA-v1.5 [38] and VisDrone2019 [39] datasets captured by satellites and high-altitude UAVs are utilized in the following experiments. The DOTA-v1.5 dataset is captured by satellites. The size of each image in the dataset is about $4,000 \times 4,000$ pixels and there are many kinds of objects with various scales, orientations, and shapes in sample images. These DOTA images, annotated by aerial image interpretation experts, are classified into 16 categories by the type of objects. This completely annotated DOTA dataset consists of 188,282 instances, with all of which labeled by an oriented bounding box rather than an axis-aligned one, a common practice to annotate objects in natural scenes. The VisDrone2019 benchmark dataset consists of 10, 209 static images that are formed by 2.6 million bounding boxes of targets and that are captured by many drone-mounted cameras with various locations, environments, objects, and density. In this dataset, 6,471 images are used for training, 548 for validation and 3,190 for testing, with the size of each image being about $2,000 \times 1.500$ pixels. Though captured at different locations, these images from the above three subsets share similar environments and attributes.

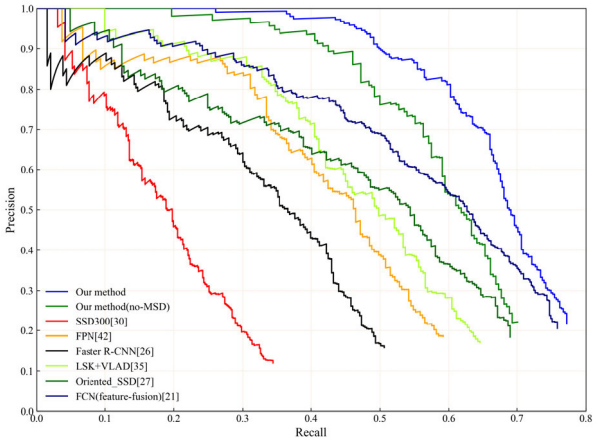


FIGURE 8. PRC of different methods.

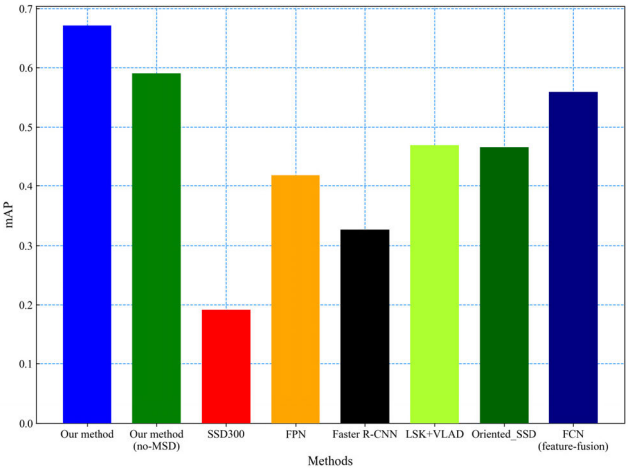


FIGURE 9. mAP value of different methods.

We made a quantitative evaluation of the performance of our method with four commonly used methods: average recall (AR), precision-recall curve (PRC), mean average precision (mAP), and F1-Score.

mAP is the average of the average precision (AP) of each class. AP metric is determined by the area under PRC [40]. The average recall (AR) metric is used to evaluate the performance of each approach [41]. The higher the value of mAP and AR, the better the performance. In addition, F1-Score can be defined as follows:

$$F_1 = \frac{2\times Recall\times Precision}{Recall + Precision} \quad (10)$$

(a) large vehicles (b) planes (c) harbors

(d) planes (e) baseball diamonds (f) basketball courts and soccer fields

(g) basketball courts and soccer fields (h) vehicles in the parking lot (i) planes

(j)planes and vehicles (k) vehicles on the road (l) vehicles and tennis courts

FIGURE 10. The detection results of DOTA-v1.5 dataset.

where recall and precision are calculated by:

$$Recall = \frac{TruePositive}{TruePositive + False\ Negative} \tag{11}$$

$$Precision = \frac{TruePositive}{TruePositive + False\ Positive} \tag{12}$$

F1-Score is a comprehensive evaluation index and larger values of F1-Score indicate better performance of the model. In remote sensing image classification tasks, Recall and Precision are often referred to as Producer's accuracy (PA) and User accuracy (UA). In order to compare with the traditional object detection model, we used recall and precision to replace UA and PA for calculation.

In our experiments, the detections with an IOU value greater than 0.5 are defined as true, otherwise, false. We conducted all of the experiments on a desktop computer which is equipped with an Intel Xeon E5-2620 v3 CPU (6 Core, 2.4 GHz), 32 GB memory, an Nvidia GeForce GTX Titan X 12GB GPU and Ubuntu 16.04 OS.

B. DOTA-V1.5 RESULTS

The proposed method is compared with three popular proposal-generation algorithms and three optimized methods

TABLE 2. Statistical information of datasets.

Dataset	DOTA-v1.5	VisDrone2019
resolution	4,000×4,000	2,000×1,500
categories	16	10
BBox quantity	188,282	542,000
# training set	1,411	6,471
# validation set	458	548
# testing set	937	3,190

TABLE 3. Comparison between different methods on DOTA-v1.5.

Method	AR	mAP	F1
SSD300	27.2	19.2	27.7
Faster R-CNN	40.8	32.7	38.5
FPN	47.1	41.9	44.2
Oriented-SSD	55.7	46.6	47.2
LSK+VLAD	52.6	47.0	45.5
FCN (feature fusion)	61.3	56.0	52.2
Ours	64.3	67.0	56.8
Ours (no-MSD)	57.6	59.0	53.8

TABLE 4. Comparison between different methods on Visdrone2019.

Method	AR	mAP	F1
SSD300	40.0	21.0	21.4
Faster R-CNN	38.9	29.8	24.3
FPN	51.9	28.8	33.2
Oriented-SSD	47.0	33.7	35.3
LSK+VLAD	66.0	60.2	43.6
FCN (feature fusion)	63.8	57.7	43.8
Ours	75.6	64.0	49.8

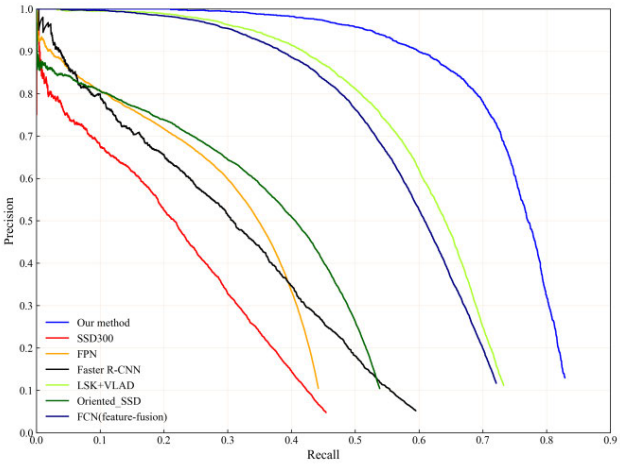


FIGURE 11. PRC of different methods under VisDrone2019 dataset.

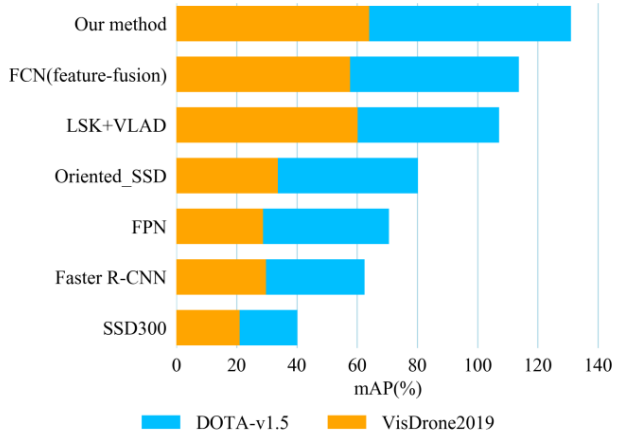


FIGURE 12. mAP values in different methods under two datasets.

in our experimental environment to illustrate how it performs better. The performance of our proposed methods are evaluated with the recall metric which equals to the ratio of the number of bounding boxes above a certain IOU overlap threshold to the entire number of ground-truth bounding boxes. Considering the comprehensive performance, the IOU threshold of the PRC is set to 0.5 in following experiments.

The plots of PRC and mAP are demonstrated in Fig. 8 and Fig. 9 respectively. The experiment is repeatedly conducted on the DOTA-v1.5 dataset, and the results represent the value of recall and precision. As for the three popular algorithms, it is evident that FPN performs better than SSD300 and Faster R-CNN. This is mainly because the feature Pyramid [42] employed by FPN is more suitable to detect small targets. Unlike general detection, the recalls obtained from SSD300 were unexpectedly low, because strategies of the original SSD, used for detection and processed without fully connected (FC) layers, probably became ineffective for detecting small objects. At the same time, Faster R-CNN, similar to common two-stage detections, utilized the last feature map of the backbone model as an input to the RPN and applied the axis-aligned bounding boxes to the detection process. The fact that axis-aligned boxes generally capture not only targeted objects but also backgrounds has increased the difficulty of an accurate detection.

Meanwhile, our method was compared with the other three optimized algorithms. As can be seen from Fig. 8 and Fig. 9, contrary to the typical SSD, the mAP of Oriented-SSD is greatly improved. Oriented-SSD offsets are predicted for each default box to better match the object shape, and its accuracy is equivalent to that of LSK+VLAD. Because of its strong feature extraction and fusion ability, FCN(feature fusion) has the highest detection accuracy among the three optimized methods. In addition, our method is also performed in another variant without the MSD bottleneck structure. As shown in Fig. 8 and Fig. 9, the test accuracy of our method is higher than that of any other algorithms, because the proposed framework utilizes many kinds of scales convolution filters to extract features with MSD and fusion of feature mapping from different stages for small target detection.

(a) Traffic in the path (b) Daytime traffic (c) Road overlook

(d) Pedestrians and vehicles (e) The crowd is on the basketball court (f) The village fair

(g) Traffic with motion blur (h) Vehicles parked on the road (i) Traffic on the road at night

(j) Walking crowd (k) The crowd in the square (l) Traffic intersection under high exposure

FIGURE 13. The detection results of VisDrone2019 dataset.

Table 3 makes comparisons between AR, mAP, and F1 for several algorithms, showing that the suggested method outperforms others, including classic detection methods as well as the state-of-the-art ones. Compared with the highest accuracy of other methods, the proposed method can improve the performance of mAP for DOTA-v1.5 by 19.79%, and the accuracy of the proposed method with MSD structure is 13.56% higher than that without MSD structure. Furthermore, F1-Score is also improved correspondingly. Our method has the highest AR value of 64.34%. A high value for AR indicates that our method can effectively reduce false detections. Results of the detection executed on the DOTA-v1.5 dataset are shown in Fig. 10. The dense instances such as the blue box are displayed in Fig. 10(j, k, and l). Most of objects occupying less than 1% of the total image area are the main reason for the low overall accuracy.

C. VISDRONE2019 RESULTS

The test-set of VisDrone2019 provides only aerial images captured by high-altitude UAVs which are considerably different from that by satellites, particularly in object sizes and the image fraction occupied by an object. In order to test

the robustness of the proposed method which has already been adopted on the DOTA dataset, we employ it on the VisDrone2019 dataset within the same environment introduced above. Fig. 11 and Table 4 display the performances of PRCs, AR, mAP and F1. Contrary to others, our method has achieved satisfactory results and developed the generalization ability. As displayed in Fig. 8 and Fig. 11, the recall value of Oriented-SSD decreases significantly compared with the two PRC curves, which is due to the fact that VisDrone2019 includes more complex image scenes, such as those in the early morning and the late night, those under a high exposure, those affected by the motion blur, and those from different angles. The experiment shows that the multi-scale feature of the target in the regression box or candidate box is very important. LSK+VLDA and FPN obtain a high recall value by extracting pyramid features in the CNN stage and carrying out some feature fusion. Compared with LSK+VLDA, our method improves the mAP by 6.24%. In addition, the proposed method obtains the highest F1-Score and can improve its performance for VisDrone2019 by 13.92%. Based on the feature pyramid, uses MSD to extract multi-scale features at different stages and employs a soft-dense connection to fuse the extracted multi-scale features. Our approach achieves the best performance while keeping a similar computational load.

To prove the robustness of our method, Fig. 12 shows the stacked histogram of mAP values got from different methods executed on the two datasets. The orange part represents the mAP value calculated on the VisDrone2019 dataset and the blue part on DOTA-v1.5. The robustness of different methods on different datasets are compared in this figure. It can be noticed that some methods, such as Oriented- SSD and Faster R-CNN, perform well on one of the two datasets, but poor on the other, because they are explored and optimized only for certain datasets. But our method shows a strong robustness and a great generalization ability on both datasets. Detection results on the VisDrone2019 dataset are shown in Fig. 13.

V. CONCLUSION AND FUTURE RESEARCH WORK

This paper firstly analyzes the disadvantages of existing object detection methods in aerial images. In order to overcome these shortages caused by multi-scale targets especially small instances, a robust object detector based on the deep neural network is proposed. The main contributions are as follows:

(1) In the feature extraction stage, a multi-scale detector (MSD) embedded in the deep neural network framework at the bottom-up pathway is explored to search for instances with multiple scales.

(2) A soft dense connection is proposed in the top-down pathway of the network. The multi-scale feature information extracted from MSD is transferred to all subsequent layers through a small number of links

To demonstrate the effectiveness of our proposed method, AR, PRC, Map and F1-Score are used to evaluate and compare the performances of the above-mentioned methods, and extensive experiments are also conducted on two publicly-available datasets: DOTA-v1.5 and VisDrone2019. The experimental results demonstrate that, the proposed method can improve the performance of mAP for DOTA-v1.5 and VisDrone2019 by 19.79% and 6.24% respectively. In addition, by evaluating the other seven typical and state-of-the-art methods with F1-Score metrics, the proposed method can improve the performance for DOTA-v1.5 and VisDrone2019 by 8.81% and 13.92% respectively.

Despite the superior performance, our method has some limitations. One of the limitations is hard example detection. In addition, MSD is a bit time-consuming in the inference stage even we reduce the dimension for feature input. For the future work, we will focus on the further optimization of MSD and built a finer architecture for feature maps extraction.

REFERENCES

[1] T. Tang, Z. Deng, S. Zhou, L. Lei, and H. Zou, "Fast vehicle detection in UAV images," in *Proc. Int. Workshop Remote Sens. with Intell. Process. (RSIP)*, May 2017, pp. 1–5.
[2] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery : A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016.
[3] J. Li, Y. Dai, C. Li, J. Shu, D. Li, T. Yang, and Z. Lu, "Visual detail augmented mapping for small aerial target detection," *Remote Sens.*, vol. 11, no. 1, p. 14, 2019.
[4] J. Zhou, C.-M. Vong, Q. Liu, and Z. Wang, "Scale adaptive image cropping for UAV object detection," *Neurocomputing*, vol. 366, pp. 305–313, Nov. 2019.
[5] A. Alshehri, Y. Bazi, N. Ammour, H. Almubarak, and N. Alajlan, "Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery," *IEEE Access*, vol. 7, pp. 119873–119880, 2019.
[6] Y. Zhang, Y. Zhang, Z. Shi, J. Zhang, and M. Wei, "Design and training of deep CNN-based fast detector in infrared SUAV surveillance system," *IEEE Access*, vol. 7, pp. 137365–137377, 2019.
[7] X. Ding, J. Zhang, and J. Yang, "A robust two-stage algorithm for local community detection," *Knowl.-Based Syst.*, vol. 152, pp. 188–199, Jul. 2018.
[8] W. Li, H. Li, Q. Wu, X. Chen, and K. N. Ngan, "Simultaneously detecting and counting dense vehicles from drone images," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9651–9662, Dec. 2019.
[9] A. Eltantawy and M. S. Shehata, "An accelerated sequential PCP-based method for ground-moving objects detection from aerial videos," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5991–6006, Dec. 2019.
[10] 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.
[11] W. Song, B. Zhong, and X. Sun, "Building corner detection in aerial images with fully convolutional networks," *Sensors*, vol. 19, no. 8, p. 1915, 2019.
[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2014.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[14] M. Shleymovich, M. Medvedev, and S. Lyasheva, "Image analysis in unmanned aerial vehicle on-board system for objects detection and recognition with the help of energy characteristics based on wavelet transform," *Proc. SPIE*, vol. 10342, Apr. 2017, Art. no. 1034210.
[15] A. ElTantawy and M. S. Shehata, "KRMARO: Aerial detection of small-size ground moving objects using kinematic regularization and matrix rank optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1672–1686, Jun. 2019.
[16] H. Zhou, L. Wei, D. Creighton, and S. Nahavandi, "Orientation aware vehicle detection in aerial images," *Electron. Lett.*, vol. 53, no. 21, pp. 1406–1408, Oct. 2017.

[17] J. Zhong, T. Lei, and G. Yao, "Robust vehicle detection in aerial images based on cascaded convolutional neural networks," *Sensors*, vol. 17, no. 12, p. 2720, 2017.
[18] L. W. Sommer, T. Schuchert, and J. Beyerer, "Deep learning based multi-category object detection in aerial images," *Proc. SPIE*, vol. 10202, May 2017, Art. no. 1020209.
[19] R. Opromolla, G. Fasano, and D. Accardo, "A vision-based approach to UAV detection and tracking in cooperative applications," *Sensors*, vol. 18, no. 10, p. 3391, 2018.
[20] H. Tayara and K. Chong, "Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network," *Sensors*, vol. 18, no. 10, p. 3341, 2018.
[21] Y. Xu, M. Zhu, P. Xin, S. Li, M. Qi, and S. Ma, "Rapid airplane detection in remote sensing images based on multilayer feature fusion in fully convolutional neural networks," *Sensors*, vol. 18, no. 7, p. 2335, 2018.
[22] T. Liu, H. Y. Fu, Q. Wen, D. K. Zhang, and L. F. Li, "Extended faster R-CNN for long distance human detection: Finding pedestrians in UAV images," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2018, pp. 1–2.
[23] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3735–3739.
[24] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2849–2858.
[25] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.
[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[27] T. Tang, S. Zhou, Z. Deng, L. Lei, and H. Zou, "Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks," *Remote Sens.*, vol. 9, no. 11, p. 1170, 2017.
[28] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–457.
[29] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Apr. 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767
[30] W. Liu, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[32] S.-J. Hong, Y. Han, S.-Y. Kim, A.-Y. Lee, and G. Kim, "Application of deep-learning methods to bird detection using unmanned aerial vehicle imagery," *Sensors*, vol. 19, no. 7, p. 1651, 2019.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[35] Liu, Ding, Zhu, Xiu, Li, and Li, "Vehicle detection in aerial images using a fast oriented region search and the vector of locally aggregated descriptors," *Sensors*, vol. 19, no. 15, p. 3294, 2019.
[36] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 30, 2019, doi: 10.1109/TPAMI.2019.2938758.
[37] H. Zhou, L. Wei, C. P. Lim, D. Creighton, and S. Nahavandi, "Robust vehicle detection in aerial images using Bag-of-Words and orientation aware scanning," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7074–7085, Dec. 2018.
[38] 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.
[39] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," Apr. 2018, *arXiv:1804.07437*. [Online]. Available: https://arxiv.org/abs/1804.07437
[40] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, p. 336, 2017.
[41] J. Hosang, R. Benenson, P. Dollar, and B. Schiele, "What makes for effective detection proposals?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, Apr. 2016.
[42] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

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.

BO ZHANG received the M.Sc. degree in control science and engineering from Henan University, in 2019, where he is currently pursuing the degree with the School of Computer and Information Engineering, Henan University. His supervisor is Prof. Miaohui Zhang. His research interests include pattern recognition and machine vision.

MENGYA LIU received the M.Sc. degree in optical engineering from Henan University, in 2019, where she is currently pursuing the degree with the School of Physics and Electronics. Her research interests include system optimization and mathematical calculation.

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.