International Journal of Cognitive Informatics and Natural Intelligence
Volume 17 • Issue 1

Fusing DCN and BBAV for Remote Sensing Image Object Detection

Honghuan Chen, College of Internet of Things Technology, Hangzhou Polytechnic, China*
Keming Wang, Hangzhou Polytechnic, China

ABSTRACT

At the oriented object detection in aerial remote sensing images, the perceptual field boundaries of ordinary convolutional kernels are often not parallel to the boundaries of the objects to be detected, affecting the model precision. Therefore, an object detection model (DCN-BBAV) that fuses deformable convolution networks (DCNs) and box boundary-aware vectors (BBAVs) is proposed. Firstly, a BBAV is used as the baseline, replacing the normal convolution kernels in the backbone network with deformable convolution kernels. Then, the spatial attention module (SAM) and channel attention mechanism (CAM) are used to enhance the feature extraction ability for a DCN. Finally, the dot product of the included angles of four adjacent vectors are added to the loss function of the rotation frame parameter, improving the regression precision of the boundary vector. The DCN-BBAV model demonstrates notable performance with a 77.30% mean average precision (mAP) on the DOTA dataset. Additionally, it outperforms other advanced rotating frame object detection methods, achieving impressive results of 90.52% mAP on VOC07 and 96.67% mAP on VOC12 for HRSC2016.

KEYWORDS

Aerial Remote Sensing Images, Box Boundary-Aware Vectors, Channel Attention Mechanism, Deformable Convolution Networks, Object Detection, Spatial Attention Mechanism

INTRODUCTION

Aerial object detection is a key computer vision task (Ding et al., 2021; Wang et al., 2020; Hu et al., 2022) that has been getting increasing attention in recent years and plays a significant role in remote image understanding. Unlike general object detection, aerial object localization presents particularly tricky questions, including nonaxis-aligned objects in arbitrary directions (Ding et al., 2019; Han et al., 2021; Pan et al., 2020) and dense distributions in complex contexts (Guo et al., 2021; Yang et al., 2018, 2021). For example, aircraft object detection techniques are mainly interfered with by external factors, for instance, noise, weather, light intensity, shadows, and background (Xiaolin et al., 2021) in remote sensing images.

Mainstream methods usually treat aerial object detection as a question of rotating object localization (Han et al., 2020; Yang et al., 2020a). Among them, the angle-based direct orientation

DOI: 10.4018/IJCINI.335496 *Corresponding Author

regression method is dominant in this research area, and it comes from general detectors (Lin et al., 2016; Li et al., 2022a; Lu et al., 2022; Zhang et al., 2022) with additional orientation parameters. While promising performance has been achieved, direct orientation prediction still suffers from a number of problems, including loss discontinuities and regression inconsistencies (Wang et al., 2022; Yang et al., 2020a, 2021; Yang & Yan 2022). The reasons are the bounded periodicity of the angular directions and the orientation definition of the rotating bounding box. Detectors based on orientation regression may not be able to accurately predict the orientation despite their attractive localization results.

To effectively address the aforementioned issues, the representation of airborne objects is revisited in order to prevent orientation estimation that is overly sensitive. Point sets are exceptionally capable of capturing important semantic features in conventional general-purpose detectors, such as RepPoints, as a fine-grained object representation (Yao et al., 2022). However, its basic transformation function can only generate upright-horizontal bounding boxes, which are unable to precisely calculate the orientation of airborne objects with precision. Additionally, RepPoints ignores a measure of the learned point's quality and merely regresses significant points. Poor performance for complex scenes and nonaxis-aligned objects with dense distribution may result from this in aerial images.

According to deformable convolutional networks (DCNs) (Zhou et al., 2022) and box boundary-aware vectors (BBAVs) (Yi et al., 2020), a new object detector oriented to aerial imagery, called DCN-BBAV, is proposed. It introduces adaptive point representations for different orientations, shapes, and attitudes. Compared to traditional directional regression methods, the suggested method captures the underlying geometry of arbitrarily oriented aerial instances in addition to finely localized aerial detection. Specifically, to fit the aerial objects, initial adaptive points are generated from the centroid and then refined. In addition, accurate feature extraction is performed on tilted objects using a deformable convolutional kernel, a scheme that measures the quality of the oriented repoints in terms of classification, localization, and feature correlation of the points. The scheme enables the detector to assign representative oriented direction vectors by capturing nonaxis-aligned information from nearby objects or background noise. In addition, a spatial attention mechanism (SAM) and channel attention mechanism (CAM) are presented to improve feature extraction, enabling points of vulnerability to discover their case owners in the complicated context of the aerial scene. The framework of the method obtains more accurate orientation and more precise detection performance compared to directional regression-based approaches.

In summary, the main contributions are:

1) Using a BBAV as a baseline model, which uses predicted centroids and boundary-aware vectors for directional object detection, effectively addressing the challenge of multiscale and arbitrary direction object detection. Owing to the nonhorizontal nature of the boundary feature distribution of the directional object, the use of ordinary convolutional kernels is easily affected by the background features. Therefore, the ordinary convolutional kernels in the feature extraction network are replaced by DCN kernels, which will enhance the discrimination of the variable target shapes.
2) To improve the feature extraction effect of deformable convolution, add the spatial attention module to a feature extraction network and the channel attention mechanism after the feature pyramid network.
3) Add the dot product of the angles of four neighboring vectors in the loss function of the rotating frame parameter, so that the angle of the angle tends to be a right angle, which improves the accuracy of the boundary vector regression.

RELATED WORKS

Remote Sensing Image Object Detection

Recent methods for aerial object detection mostly use orientation regression over traditional object detectors. By foreseeing the rotation angle of the bounding box, SCRDet (Yang et al., 2018), CADNet (Zhang et al., 2019), DRN (Pan et al., 2020), R3Det (Yang et al., 2020b), ReDet (Han et al., 2021), and directional RCNN (Xie et al., 2021) all perform well. By regressing the quadrilateral, Gliding Vertex (Xu et al., 2019) and RSDet (Wang et al., 2022) enhanced the detection outcomes. Angle regression was changed to angle classification by Yang et al. (2022) to overcome the boundary discontinuity in angle-based orientation estimation. To get more reliable results for directional object detection, Yang et al. (2021) parameterized the rotating bounding box as a 2D Gaussian distribution. These techniques are mostly employed to enhance orientation estimation by using rotated angular representations.

Most conventional object detection methods (Lu et al., 2022; Zhang et al., 2022; Yao et al., 2022; Tian et al., 2019) concentrate on vertically or axis-aligned objects, which may have trouble detecting densely distributed nonaxis-aligned objects in complex backgrounds. To solve this problem, Wang et al. (2019) deployed the inception lateral connection network (ILCN) to augment the feature pyramid network (FPN) with the semantic attention network (SAN) in order to offer semantic features that can effectively distinguish objects of interest from congested backgrounds. Under the guidance of oriented bounding boxes, Ding et al. (2019) suggested spatial transformations on axis-aligned ROIs and learned nonaxis-aligned representations. In order to train the network, SCRDet++ (Yang et al., 2020c) enhanced nonaxis-aligned features and increased object responses. Han et al. (2020) created a module for character alignment to reduce the mismatch across axis-aligned convolutional features and random object-oriented objects. A DRN (Pan et al., 2020) suggested a character selection module to consolidate nonaxis-aligned information from various kernel sizes, shapes, and orientations, and perform additional regression by using a dynamic filter generator. Guo et al. (2021) used a convex packet representation to acquire irregular shapes and configurations with the intention of avoiding feature aliasing via learnable feature adaptation. However, many models typically exhibit high false positives when facing targets with diverse types and arbitrary directions. In order to better solve the problem of multiscale and arbitrary direction object detection, Yi et al. (2020) and Yu et al. (2022) proposed different detection methods based on BBAVs. Among them, Yi et al. (2020) extended the target detector based on horizontal key points to directional object detection tasks, combined Cartesian coordinate design with directional boundary classification, effectively solving the problem of objects learning for any direction, Yu et al. (2022) proposed the ASFF-BBAV by introducing multiscale adaptive spatial feature fusion (ASFF) on the basis of the BBAV, which fuses multiscale convolutional neural network Res2Net with adaptive spatial features, effectively enhancing the adaptability of the detection model to objects of different sizes. These two studies indicate that using BBAVs can effectively solve the problem of boundary feature extraction for multiscale and arbitrary directional targets. However, their feature extraction ability for targets with variable shapes still needs to be improved.

In addition, some efforts have been made to enhance the interpretability of models and improve their performance by incorporating adversarial generative networks. Ferdous et al. (2019) proposed a two-stage detector that utilizes GAN for image super-resolution and SSD for object detection. Similarly, Rabbi et al. (2020) integrated ESRGAN and edge-enhanced GAN to develop an end-to-end small object detection network, employing Faster RCNN and SSD for object detection. However, the primary focus of these methods is to enhance image resolution, and their effectiveness in tasks involving directional object detection is limited.

Problems and Solution Strategies

Summarizing the above analysis, many anchorless-based directional object detection models are prone to be affected by the background or other objects due to the structural characteristics of the

ordinary convolutional kernel, causing problems such as inaccurate semantic information when feature extraction is performed for the boundaries of arbitrarily oriented objects. Aiming to solve this problem, using a BBAV as a baseline model, it is proposed to use a deformable convolution kernel instead of an ordinary convolution kernel to address the problem. It can not only effectively deal with multiscale and arbitrary direction targets but also effectively solve the problem of target shape variability. In addition, the mechanism of attention is used to enhance the feature extraction capability of a variability convolution kernel, further increasing the precision of the proposed DCN-BBAV model.

THE PROPOSED DCN-BBAV-BASED OBJECT DETECTION METHOD

The proposed algorithm takes the BBAV model as the baseline and changes the 3 × 3 convolutional kernel in the backbone's feature extraction network ResNet101 to a DCN. The features extracted by the ordinary convolutional kernel are the feature regions parallel to the image boundaries, and thus the feature regions contain a lot of semantic information that is not related to the object, which is not conducive to the enhancement of the model's detection accuracy, whereas a DCN can effectively solve this problem. Figure 1 shows the DCN-BBAV model structure:

The input image in Figure 1 is changed to 608 × 608 prior to transmission over the network. A U-shaped network supports the structure's design. During the upsampling procedure, skip connections are utilized to merge feature maps. Four mappings constitute the output of the architecture: a box parameter map B, a heatmap P, an orientation map α, and an offset map O. The heatmap and offset map are utilized to determine the centroid's location. The input image of the model is assumed to be Input $\in R3 \times H \times W$, where 3 indicates the channels' number, H stands for the height, and W stands for the width. Four branches are transformed from the output feature map $X \in RC \times (H/S) \times (W/S)$ ($C = 256$ in this paper), they are orientation map ($\alpha \in R1 \times (H/S) \times (W/S)$), heatmap ($P \in RK \times (H/S) \times (W/S)$), box parameter ($B \in R10 \times (H/S) \times (W/S)$), and offset ($O \in R2 \times (H/S) \times (W/S)$), where $S = 4$ represents the scale, and K represents the number of data set categories. Three 3 × 3 ordinary convolutional kernels and two convolutional layers with 256 channels are used to implement the transformation.

Deformable Convolutional Networks

In feature extraction of an image using a convolutional kernel, for each output y (P0), nine positions are sampled from the input x. The set of sampling points R = {(–1,–1),(–1,0),(–1,1),(0,–1),(0,0),(0,1),(1,–1),(1,0),(1,1)} is shown in Figure 2:

The initial version of deformable convolution optimizes the locations of feature extraction, and the second version of deformable convolution multiplies each feature point by a coefficient ∈

Figure 1. The DCN-BBAV model structure

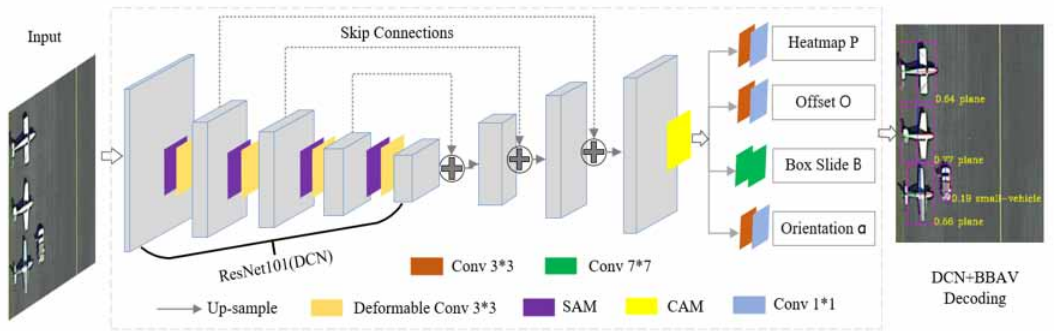


Figure 2. The set of sampling points

[0,1], which is also learned through training, so that the output channel in deformable convolution is increased from 2N to 3N. The coefficients obtained through learning are generally small, so it is necessary to use bilinear interpolation to round the sampling locations.

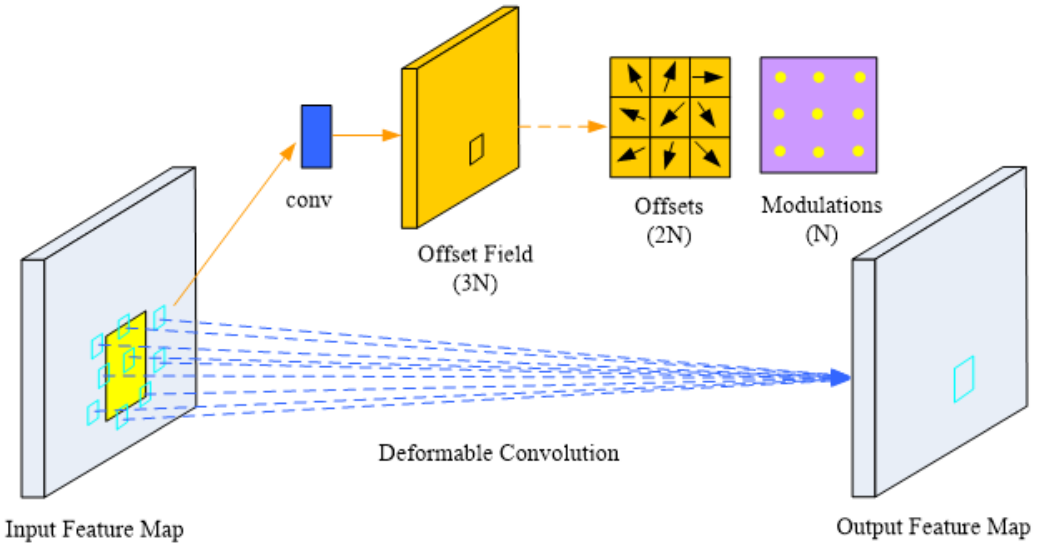
$$y(P_0)=\sum_{L_n\in R}\omega(P_0)\cdot x(P_0+L_n+O_n)\cdot M_n\ , \tag{1}$$

where O_n is the offset and M_n is the weight. The DCN calculation process is shown in Figure 3:

Thermograms

Utilizing the heatmap, it is possible to locate the center of an object. The heatmap has *k* channels that correspond to different classes of objects. The mapping on each channel passes through a sigmoid function, which represents the value of each pixel point of the heatmap as the confidence level of the object. Considering that the center point of a directional detection box is (x, y), the probability density

Figure 3. The DCN calculation process



values for all coordinates of the entire heatmap are obtained using the coordinates of this point as the mean and adaptively generating the variance based on the size of the box, using a two-dimensional Gaussian distribution. Training the heatmap results in only the center point being positive. Negative values exist for all other points, including those in the Gaussian bump. Due to the imbalance problem, learning positive centroids directly would be difficult. To solve this problem, based on the work in the literature (Albattah et al., 2022), the penalty for points within the Gaussian bump is reduced, and focal loss is used to train the heatmap:

$$L_h = -\frac{1}{N}\sum_i \begin{cases} (1-p_i)^a \log(p_i) & if \quad \hat{p}_i = 1 \\ \left(1-\hat{p}_i\right)^\beta p_i^\alpha \log(1-p_i) & otherwise \end{cases}, \tag{2}$$

where p and $\hat{p}$ indicate the predicted and true values, correspondingly, i represents the pixel coordinate, N represents the total number of objects, and α and β adjust the weights of positive and negative samples. In the proposed DCN-BBAV, α and β take the values of 2 and 4, respectively.

Center Offset

The model inference sets the maximum point in the heatmap as the center of the detection object c. The initial value of this point is an integer, and after downsampling, it becomes a floating-point number. The error can be eliminated by offsetting o, as shown in Equation (3):

$$o = \left(\frac{\overline{c}_x}{s} - \left\lfloor \frac{\overline{c}_x}{s} \right\rfloor, \frac{\overline{c}_y}{s} - \left\lfloor \frac{\overline{c}_y}{s} \right\rfloor\right). \tag{3}$$

Optimizing the offset with a smoothed L1 loss (Qian et al., 2023):

$$L_O = \frac{1}{N}\sum_{k=1}^{N} Smooth_{L_1}(O_k - \hat{O}_k), \tag{4}$$

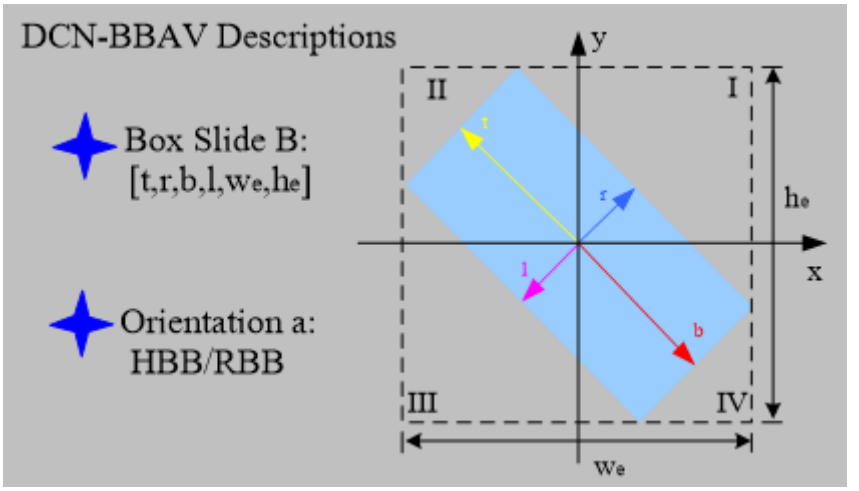
where the objects' total number representation is N, the ground truth offset representation is $\hat{O}$, and k represents the objects.

Rotation Box Parameters

The baseline method for capturing the oriented bounding box is named Center-WH-θ. This method has several drawbacks. First, when the angle is very small, it has an effect on the model loss, resulting in a relatively large intersection that connects the predicted box and the real box. Furthermore, the w and h of each object's OBB are calculated in a separate rotated coordinate system with an angle θ relative to the *y*-axis. Consequently, it is difficult for the network to acquire the box parameters of all objects simultaneously. The OBB is described using the box boundary-sensing vectors in this paper. The BSAV includes the upper t, right r, lower b, and left l vectors from the center point of the object. The four quadrants of the Cartesian coordinate system are occupied by these four categories of vectors. All arbitrarily oriented objects share a common coordinate system, thereby enhancing generalization by facilitating the transfer of mutual information. The rotating frame parameter is $b = \left[t, r, b, l, w_e, h_e\right]$, where w_e and h_e denote the width and height of the external horizontal rectangular frame of the OBB, respectively. The rotating box is depicted in Figure 4:

The rotating frame loss function is:

Figure 4. The rotating box



$$L_b = \frac{1}{N}\sum_{k=1}^{N} Smooth_{L_1}(b_k - \hat{b}_k), \tag{5}$$

where b and $\hat{b}$ represent the predicted and ground truth box parameters, correspondingly. In this paper, the angle loss function between the boundary vectors is added:

$$L_b = \frac{1}{N}\sum_{k=1}^{N}(t_k \cdot l_k + t_k \cdot r_k + b_k \cdot l_k + b_k \cdot r_k), \tag{6}$$

where $t \cdot l$, $t \cdot r$, $b \cdot l$, $b \cdot r$ are dot products of vectors, such that the angle between neighboring vectors is maintained at 90° during vector regression.

Classification of Rotating Bounding Boxes

A BBAV categorizes rotating boxes into two types, HBB and RBB, with RBB comprising all rotating bounding boxes, excluding horizontal boxes. When the network encounters the case of horizontal boxes, the orientation category and external dimensions can assist the network in capturing accurate OBBs. Extra external dimension parameters further strengthen the description of OBBs, as shown in Figure 4.

$$\hat{a} = \begin{cases} 1(RBB) & IOU(OBB, HBB) < 0.95 \\ 0(HBB) & otherwise \end{cases}. \tag{7}$$

The loss function is as follows:

$$L_a = -\frac{1}{N}\sum_{i}^{N}(\hat{a}_i \log(a_i) + (1-\hat{a}_i)\log(1-a_i)). \tag{8}$$

Among this, a represents the predicted orientation class, and $\hat{a}$ represents the true orientation classes.

EXPERIMENTS

During model training and inference, all input images were adjusted to a resolution of 608 × 608, and the proposed model was implemented using the PyTorch framework. Data enhancement, such as inversion, translation, and addition of noise are used for preprocessing during model training. Adm is utilized by the optimization algorithm with an initial rate of learning of 1.25 × 10–4. On the DOTA data set, nearly 100 epochs were trained. Eighty epochs are included in the HRSC2016 data set. For the HRSC2016 data set, the efficiency of the proposed DCN-BBAV was measured on a single NVIDIA TITAN X GPU. AdamW (Yu & Ji, 2022) is used as the optimizer, the weight attenuation was set to 0.0001, and the initial learning rate was set to 0.000025.

Model Evaluation Criteria

As the evaluation index, the mean average precision mean (mAP) is used as the criterion for evaluating the model in this paper. In the object detection task, three kinds of results are generally used: true positive samples (TP), false positive samples (FP), and false negative samples (FN). On the basis of these three results, the algorithm's performance is assessed using the accuracy rate P, the recall rate R, and the average precision AP as the criterion, which are calculated using the following formula:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$AP = \int_0^1 P(R)dR \tag{11}$$

$$mAP = \sum_{i=1}^{n} \frac{AP_i}{n}. \tag{12}$$

The mAP is the averaged value of the average accuracy of all categories. Firstly, P and R are calculated for each prediction frame, the R value obtained each time is used as a threshold, when R is greater than the threshold, the corresponding maximum accuracy is calculated, then the average of the accuracy is calculated to get the AP, and finally, the AP of all the different categories of objects is averaged to get the mAP.

Data Sets

The remote sensing data sets used for the experiment are DOTA (Xia et al., 2017) and HRSC2016 (Liu et al., 2017). Among them, DOTA possesses 2,806 aerial images with resolution sizes starting at 800 × 800 to 4,000 × 4,000, containing a total of 15 categories and 188,282 instances. The labeling method is a quadrilateral determined by four points, which can be applied to the horizontal frame detection task and the rotated frame detection task. The set used for training is 1,411 sheets, the set used for validation is 458 sheets, and the set used for tests is 937 sheets. The 15 categories included are: planes, bridges, harbors, athletic fields, small vehicles, large vehicles, roundabouts, swimming pools, ships, soccer fields, tennis courts, basketball courts, oil storage tanks, baseball fields, and helicopters, with the acronyms PL, BR, HA, GTF, SV, LV, RA, SP, SH, SBF, TC, BC, ST, BD, and HC, respectively. HRSC2016 is labeled in a format that uses a directional rotating frame approach. It contains 1,061 images with 2,976 instances, of which the set used for training contains 436 images and 1,207 instances, the set used for validation contains 181 images and 541 instances, and the set

used for test contains 444 images and 1,228 instances. The resolution of the data set ranges from 300 × 300 to 1,500 × 900, and all the objects in HRSC2016 are treated as a category of "ship" for training and testing in this paper.

EXPERIMENTAL RESULTS

Figure 5 depicts detection outcomes on the data set DOTA. Table 1 displays the performance comparison results between this proposed DCN-BBAV and other models, including one-stage models RSDet (Wang et al., 2022), R3Det (Yang et al., 2020b), S2ANet (Han et al., 2020), two-stage models MaskOBB (Yang et al., 2020c), CenterMap (Wang et al., 2020), ReDet (Han et al., 2021) and anchor-free-based models DRN (Pan et al., 2020), CFA (Guo et al., 2021), BBAV (Yi et al., 2020), ASFF-BBAV (Yu et al., 2022), and PPSS (Song et al., 2023), which shows that the mAP of the DCN-BBAV reaches 77.30%. The PPSS model utilizes the classification information, regression information, and distribution characteristics of point sets to represent objects. However, in practical situations, this selection strategy may lead to the omission of important object information in certain cases or the selection of samples with higher noise levels. On the other hand, the DCN-BBAV focuses primarily on oriented object detection. It enhances the feature extraction capability by introducing the DCN and combines it with the representation of BBAV, directing the model toward oriented object detection.

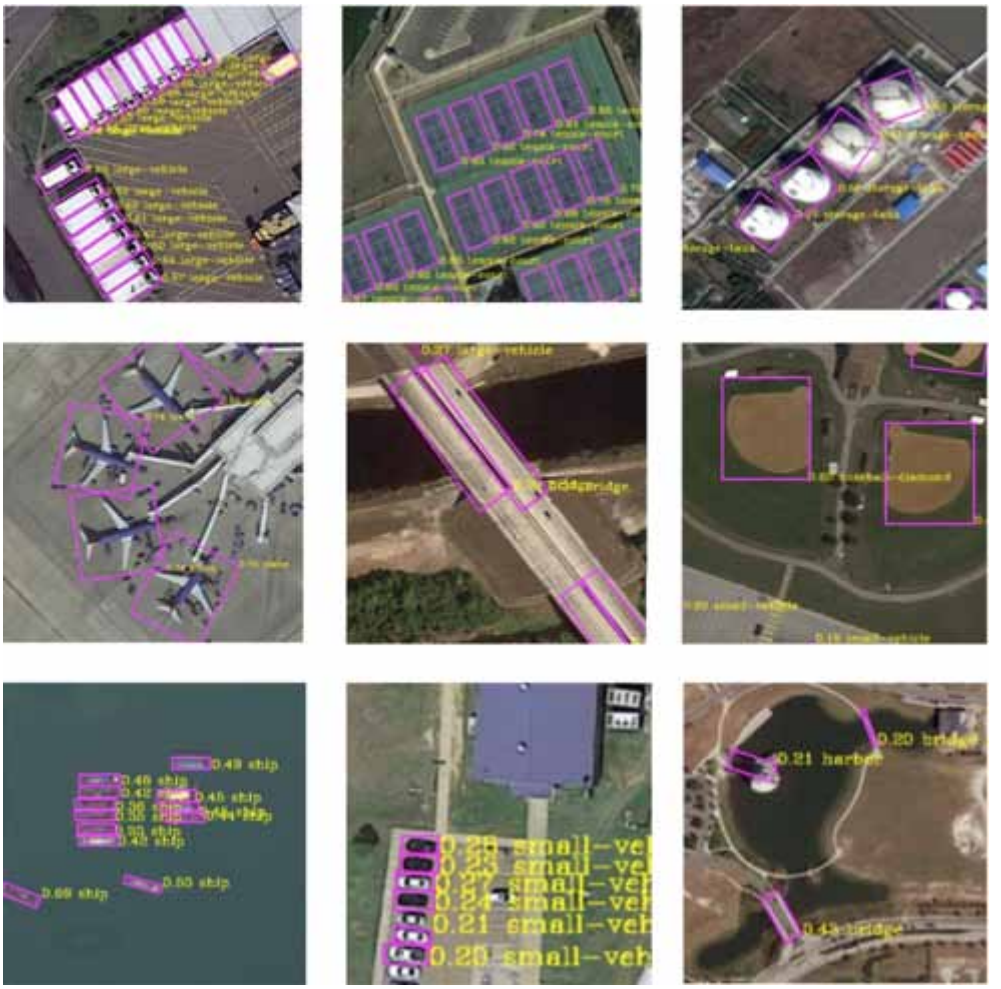
Table 2 depicts the performance comparison results of the DCN-BBAV with CenterMap (Wang et al., 2020), ROI-Transformer (Ding et al., 2019), DRN (Pan et al., 2020), R3Det (Yang et al. 2020b), FPN-CSL (Yang & Yan, 2022), S2ANet (Han et al., 2020), Oriented R-CNN (Xie et al., 2021), BBAV (Yi et al., 2020), ASFF-BBAV (Yu et al., 2022), PPSS (Song et al., 2023) and CF-ORNet (Wang et al., 2023). The mAP50 (VOC2007) of the proposed DCN-BBAV reaches 90.52%, while the mAP50 (VOC2012) reaches 96.67%, which is superior to several other comparative object detection models, and it is consistent with the comparison results on the DOTA data set.

Compared to the CF-ORNet, the DCN-BBAV exhibits significant performance advantages. This is attributed to the fact that the CF-ORNet, through the knowledge distillation process, fails to

Table 1. Results of different algorithms on data set DOTA (%)

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC
One-stage									
RSDet	R-152-FPN	90.10	82.00	53.80	68.50	70.20	78.70	73.60	91.20
R3Det	R-152-FPN	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81
S2ANet	R-50-FPN	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83
Two-stage									
MaskOBB	R-50-FPN	89.61	85.09	51.85	72.90	75.28	73.23	85.57	90.37
CenterMap	R-50-FPN	88.88	81.24	53.15	60.65	78.62	66.55	78.10	88.83
ReDet	ReR-50-ReFPN	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.52
Anchor-free									
DRN	H-104	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57
CFA	R-101-FPN	89.26	81.72	51.81	67.17	79.99	78.25	84.46	90.77
BBAV	R-101-FPN	88.63	84.06	52.13	69.56	78.26	80.40	88.06	90.87
ASFF-BBAV	Res2Net50-FPN	89.84	84.90	51.55	74.42	78.54	83.28	87.36	90.52
PPSS	R-50-FPN	88.99	82.28	54.02	73.32	81.01	81.88	88.13	90.85
DCN-BBAV (ours)	R-101-FPN	90.17	87.14	52.18	78.18	82.33	81.41	84.31	91.12

Figure 5. Results of DOTA data set visualization



adequately preserve and transmit valuable information from the first stage, thereby influencing the final fine-grained recognition performance.

In contrast, the DCN-BBAV adopts a nonparametric form, specifically in the loss function, to ensure the orientation of the bounding box. This not only reduces the parameter count but also enhances the effectiveness of detection. Figure 6 depicts detection results on the data set HRSC2016. Observations reveal that the visualization performance of the Bounding Box is outstanding, accurately locating targets in various orientations. This is attributed to the introduction of the DCN, which adaptively adjusts the receptive field, accommodating diverse shapes and orientations of targets, thereby more effectively capturing local features and enhancing detection accuracy. Additionally, by employing rotation box parameters to capture boundary information, rotation boxes, compared to traditional Bounding Boxes, exhibit greater flexibility in adapting to the orientation of targets, further improving localization accuracy.

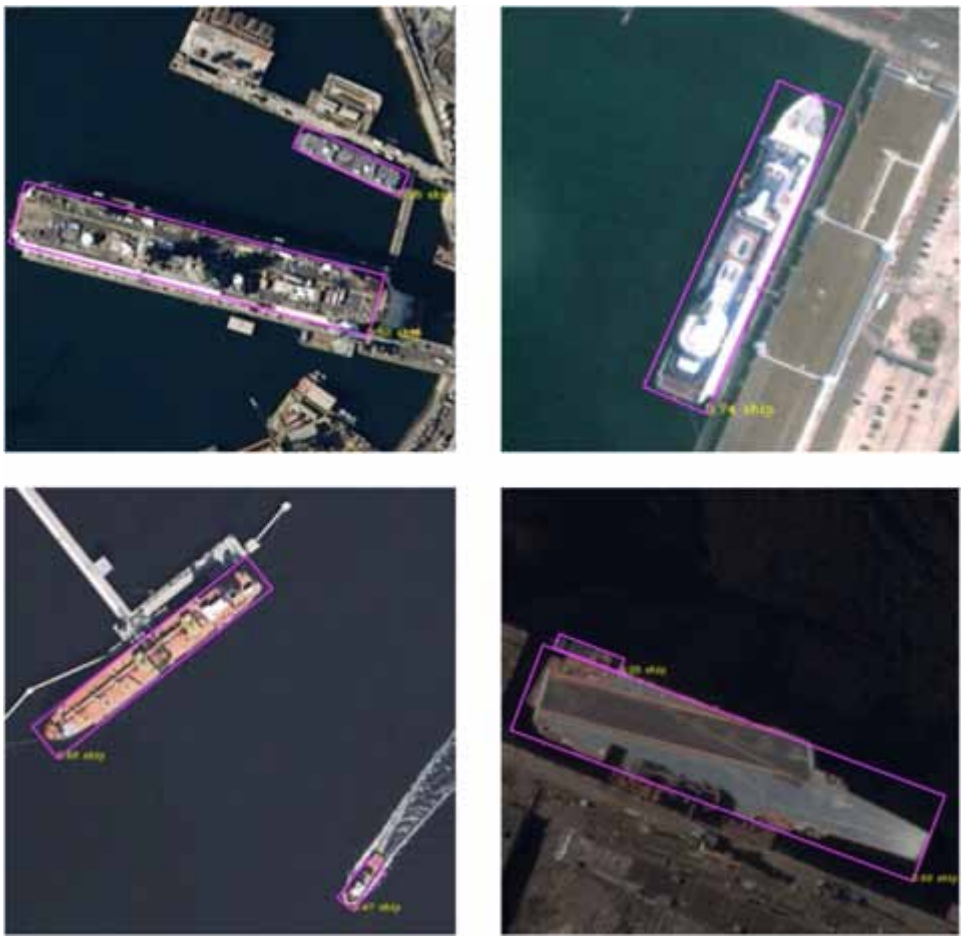
Ablation Experiments

The results of ablation experiments conducted on the DOTA data set and HRSC2016 data set are presented in Table 3, respectively. In Table 3, several combinations yielded better results in which

Table 2. Results of different algorithms on the data set HRSC2016 (%)

Method	Backbone	mAP50 (07)	mAP50 (12)
R3Det	R-101-FPN	89.26	96.01
FPN-CSL	R-101-FPN	89.62	96.10
ROI-Transformer	R-101-FPN	86.20	—
DRN	H-104	—	92.70
CenterMap	R-50-FPN	—	92.80
S2ANet	R-101-FPN	90.17	95.01
Oriented R-CNN	R-50-FPN	90.40	96.50
BBAV	R-101-FPN	88.60	94.85
ASFF-BBAV	Res2Net-50-FPN	90.30	—
PPSS	R-50-FPN	89.53	—
CF-ORNet	R-50-FPN	84.26	—
DCN-BBAV (ours)	R-101-FPN	90.52	96.67

Figure 6. Results of HRSC2016 data set visualization



the performance of the BBAV was marginally improved with the introduction of the SAM and CAM. Similarly, with the introduction of the DCN, the performance of the BBAV is also slightly improved. The performance of the BBAV can be greatly improved by introducing the DCN, SAM, and CAM at the same time. Among them, the best results of the proposed DCN-BBAV model can be obtained when the last 26 3×3 convolutional kernels of ResNet101 are replaced by deformable convolutional kernels, and four spatial attention modules and one channel attention module were introduced to the proposed DCN-BBAV model, validating the efficacy of each module. Aiming to be more intuitive, the PR curve is plotted with mAP50 (VOC2007) on the HRSC2016 data set as an example, as illustrated in Figure 7.

By introducing the DCN, SAM, and CAM, we observe significant improvements in both the recall rate and mAP. The incorporation of the DCN aids in capturing specific shape and directional information of the targets, thereby enhancing the model's ability to recognize objects. Additionally, the SAM and CAM individually enhance the model's focus on spatial and channel information. The SAM is dedicated to reinforcing the model's attention to different locations in the image, allowing for a more accurate capture of target boundary information. On the other hand, the CAM focuses on the importance of different channels in the image, enabling better capture of abstract features. The synergistic effects of these comprehensive optimization strategies result in superior performance of the model in handling object detection tasks.

CONCLUSION

The drawbacks of using an ordinary convolutional kernel to extract features for directional target detection were analyzed, and it was concluded that an ordinary convolutional kernel has low accuracy in boundary feature extraction for targets in different directions. A target detection algorithm (DCN-BBAV) that combines the DCN and BBAV was proposed with this problem as its starting point. Comprehensive experiments showed that the DCN-BBAV helps to improve the detection accuracy. The DCN-BBAV obtained the highest accuracy without any additional features and accomplished a significant improvement over the baseline on the DOTA as well as HRSC2016 data sets.

However, the proposed DCN-BBAV model has some limitations:

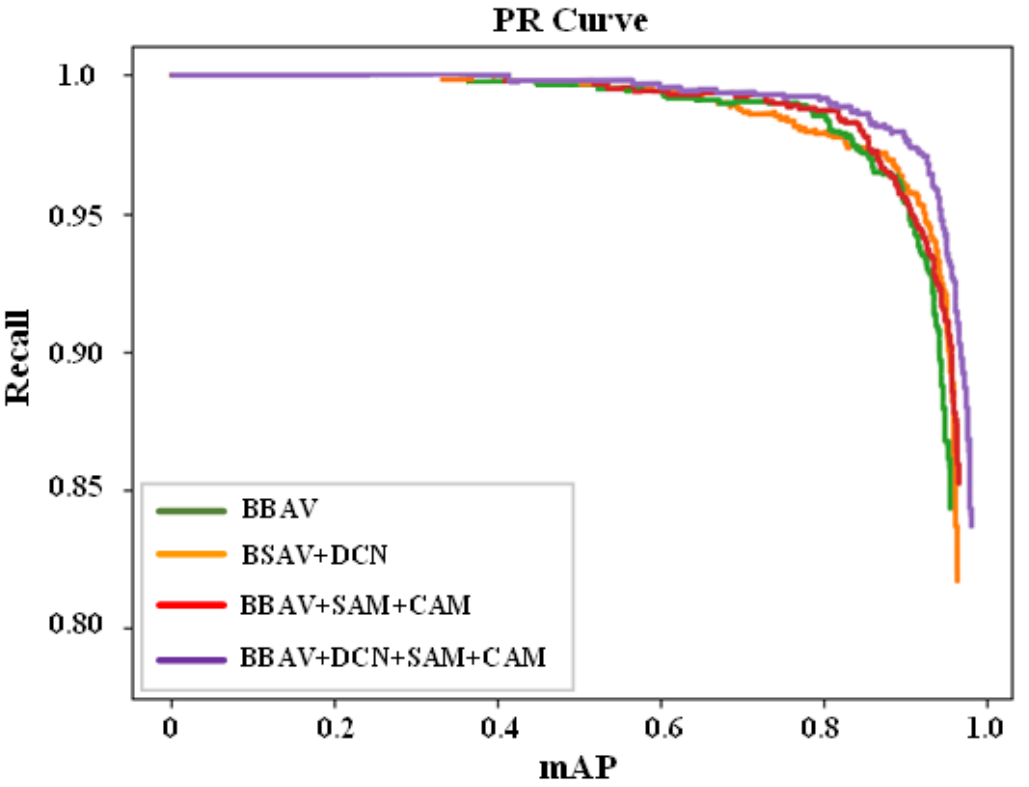
In terms of detection accuracy, both the backbone network and the attention module of the DCN-BBAV model have areas for improvement. Therefore, its detection accuracy will be further improved in the future by introducing other feature enhancement techniques (Shi et al., 2022) and parameter tuning (Bai et al., 2022).

The feature extraction process of the DCN-BBAV model is more complicated, while real application scenarios require high real-time performance of the model. Therefore, the model will be

Table 3. mAP values of ablation experiments on DOTA and HRSC2016 data sets (%)

Method	Backbone	DOTA	HRSC2016	
		mAP	mAP50 (VOC07)	mAP50 (VOC12)
BBAV	ResNet101	75.36	88.60	94.85
BBAV+DCN	ResNet101(DCN*26)	76.33	89.96	96.25
BBAV+SAM+CAM	ResNet101	76.09	89.87	96.08
BBAV+DCN+SAM+CAM	ResNet101(DCN*3)	76.97	90.32	96.41
BBAV+DCN+SAM+CAM	ResNet101(DCN*26)	77.30	90.52	96.67
BBAV+DCN+SAM+CAM	ResNet101(DCN*30)	77.18	90.38	96.48
BBAV+DCN+SAM+CAM	ResNet101(DCN*33)	76.38	90.27	96.30

Figure 7. PR curves for mAP50(VOC07) on the HRSC2016 data set



optimized in the future, that is, it will be improved to a single-stage (Li et al., 2022b) multidirectional object detector to improve the training and detection efficacy of the model without compromising the detection precision.

In summary, future research will primarily focus on two aspects. Firstly, we will explore the incorporation of additional feature enhancement techniques to further enhance the model's performance in object detection tasks. Secondly, there are plans to optimize the model's feature extraction process, potentially by transforming it into a single-stage multidirectional object detector, aiming to improve the model's training and detection effectiveness without compromising the detection accuracy.

REFERENCES

Albattah, W., Masood, M. F., Javed, A., Nawaz, M., & Albahli, S. (2022). Custom CornerNet: A drone-based improved deep learning technique for large-scale multiclass pest localization and classification. *Complex & Intelligent Systems*, *9*(2), 1299–1316. doi:10.1007/s40747-022-00847-x

Bai, J., Ren, J., Yang, Y., Xiao, Z., Yu, W., Havyarimana, V., & Jiao, L. (2022). Object detection in large-scale remote-sensing images based on time-frequency analysis and feature optimization. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 5405316. doi:10.1109/TGRS.2021.3119344

Ding, J., Xue, N., Long, Y., Xia, G., & Lu, Q. (2019, June 15–20). *Learning RoI transformer for oriented object detection in aerial images* [Conference session]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. doi:10.1109/CVPR.2019.00296

Ding, J., Xue, N., Xia, G., Bai, X., Yang, W., Yang, M. Y., Belongie, S. J., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2021). Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7778–7796. doi:10.1109/TPAMI.2021.3117983 PMID:34613910

Ferdous, S. N., Mostofa, M., & Nasrabadi, N. M. (2019, May). *Super resolution-assisted deep aerial vehicle detection* [Conference session]. Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications SPIE, Baltimore, MD, USA. 10.1117/12.2519045

Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., & Ye, Q. (2021, June 20–25). *Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection* [Conference session]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. doi:10.1109/CVPR46437.2021.00868

Han, J., Ding, J., Li, J., & Xia, G. (2020). Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11. doi:10.1109/TGRS.2021.3062048

Han, J., Ding, J., Xue, N., & Xia, G. (2021, June 20–25). *ReDet: A rotation-equivariant detector for aerial object detection* [Conference session]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. doi:10.1109/CVPR46437.2021.00281

Hu, J., Xie, L., Gu, X., Xu, W., Chang, M., & Xu, B. (2022, October 31–November 2). *Information-interaction feature pyramid networks for object detection* [Conference session]. 2022 IEEE 34th International Conference on Tools With Artificial Intelligence (ICTAI), Macao, China. doi:10.1109/ICTAI56018.2022.00197

Li, B., Yao, Y., Tan, J., Zhang, G., Yu, F., Lu, J., & Luo, Y. (2022a, June 18–24). *Equalized focal loss for dense long-tailed object detection* [Conference session]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.00686

Li, Y., Kong, C., Dai, L., & Chen, X. (2022b). Single-stage detector with dual feature alignment for remote sensing object detection. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. doi:10.1109/LGRS.2021.3130379

Lin, T., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2016, July 21–26). *Feature pyramid networks for object detection* [Conference session]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. doi:10.1109/CVPR.2017.106

Liu, Z., Yuan, L., Weng, L., & Yang, Y. (2017). A high resolution optical satellite image dataset for ship recognition and some new baselines. *International Conference on Pattern Recognition Applications and Methods,* 324–331. https://www.semanticscholar.org/paper/A-High-Resolution-Optical-Satellite-Image-Dataset-Liu-Yuan/e6a32b4df848fd74b43486c5232ebd362eb90416

Lu, J., Li, D., Wang, M., Mi, B., Wang, P., Dai, Z., & Zheng, F. (2022, October 7–9). *Object detection system based on Faster R-CNN* [Conference session]. 2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Hamburg, Germany. doi:10.1109/AIAM57466.2022.00027

Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., & Xu, C. (2020, June 13–19). *Dynamic refinement network for oriented and densely packed object detection* [Conference session]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA. doi:10.1109/CVPR42600.2020.01122

Qian, X., Zhang, N., & Wang, W. (2023). Smooth GIoU loss for oriented object detection in remote sensing images. *Remote Sensing (Basel)*, *15*(5), 1259. doi:10.3390/rs15051259

Rabbi, J., Ray, N., Schubert, M., Chowdhury, S., & Chao, D. (2020). Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. *Remote Sensing (Basel)*, *12*(9), 1432. doi:10.3390/rs12091432

Shi, T., Gong, J., Hu, J., Zhi, X., Zhang, W., Zhang, Y., Zhang, P., & Bao, G. (2022). Feature-enhanced CenterNet for small object detection in remote sensing images. *Remote Sensing (Basel)*, *14*(21), 5488. doi:10.3390/rs14215488

Song, J., Miao, L., Zhou, Z., Ming, Q., & Dong, Y. (2023). Optimized point set representation for oriented object detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, *20*, 6010505. doi:10.1109/LGRS.2023.3314517

Tian, Z., Shen, C., Chen, H., & He, T. (2019, October 27–November 2). *FCOS: Fully convolutional one-stage object detection* [Conference session]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea. doi:10.1109/ICCV.2019.00972

Wang, J., Ding, J., Guo, H., Cheng, W., Pan, T., & Yang, W. (2019). Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing (Basel)*, *11*(24), 2930. doi:10.3390/rs11242930

Wang, J., Li, F., & Bi, H. (2022). Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13. doi:10.1109/TGRS.2022.3175520

Wang, J., Yang, W., Li, H., Zhang, H., & Xia, G. (2020). Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(5), 4307–4323. doi:10.1109/TGRS.2020.3010051

Wang, L., Zhang, J., Tian, J., Li, J., Zhuo, L., & Tian, Q. (2023). Efficient fine-grained object recognition in high-resolution remote sensing images from knowledge distillation to filter grafting. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–16. doi:10.1109/TGRS.2023.3335484

Xia, G., Bai, X., Ding, J., Zhu, Z., Belongie, S. J., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2017, June 18–23). *DOTA: A large-scale dataset for object detection in aerial images* [Conference session]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. doi:10.1109/CVPR.2018.00418

Xiaolin, F., Fan, H., Ming, Y., Tongxin, Z., Ran, B., Zenghui, Z., & Zhiyuan, G. (2021). Small object detection in remote sensing images based on super-resolution. *Pattern Recognition Letters*, *153*, 107–112. doi:10.1016/j.patrec.2021.11.027

Xie, X., Cheng, G., Wang, J., Yao, X., & Han, J. (2021, October 10–17). *Oriented R-CNN for object detection* [Conference session]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada. doi:10.1109/ICCV48922.2021.00350

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G., & Bai, X. (2019). Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(4), 1452–1459. doi:10.1109/TPAMI.2020.2974745 PMID:32086194

Yang, F., Fan, H., Chu, P., Blasch, E., & Ling, H. (2019, October 27–November 2). *Clustered object detection in aerial images* [Conference session]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea. doi:10.1109/ICCV.2019.00840

Yang, X., Hou, L., Zhou, Y., Wang, W., & Yan, J. (2020a, June 20–25). *Dense label encoding for boundary discontinuity free rotation detection* [Conference session]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. doi:10.1109/CVPR46437.2021.01556

Yang, X., Liu, Q., Yan, J., & Li, A. (2020b). R3det: Refined single-stage detector with feature refinement for rotating object. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(4), 3163–3171. doi:10.1609/aaai.v35i4.16426

Yang, X., & Yan, J. (2022). On the arbitrary-oriented object detection: Classification based approaches revisited. *International Journal of Computer Vision*, *130*(5), 1340–1365. doi:10.1007/s11263-022-01593-w

Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., & Tian, Q. (2021). Rethinking rotated object detection with Gaussian Wasserstein distance loss. *Proceedings of the 38th International Conference on Machine Learning, 139*, 11830–11841. https://icml.cc/virtual/2021/spotlight/9046

Yang, X., Yan, J., Yang, X., Tang, J., Liao, W., & He, T. (2020c). SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 2384–2399. doi:10.1109/TPAMI.2022.3166956 PMID:35412976

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., & Fu, K. (2018, October 27–November 2). *SCRDet: Towards more robust detection for small, cluttered and rotated objects* [Conference session]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea. doi:10.1109/ICCV.2019.00832

Yao, X., Shen, H., Feng, X., Cheng, G., & Han, J. (2022). R^2IPoints: Pursuing rotation-insensitive point representation for aerial object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–12. doi:10.1109/TGRS.2022.3230411

Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., & Metaxas, D. N. (2020, January 3–8). *Oriented object detection in aerial images with box boundary-aware vectors* [Conference session]. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA. doi:10.1109/WACV48630.2021.00220

Yu, D., & Ji, S. (2022). A new spatial-oriented object detection framework for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 4407416. doi:10.1109/TGRS.2021.3127232

Yu, D., Xu, Q., Guo, H., Xu, J., Lu, J., Lin, Y., & Liu, X. (2022). Anchor-free arbitrary-oriented object detector using box boundary-aware vectors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 2535–2545. doi:10.1109/JSTARS.2022.3158905

Zhang, G., Lu, S., & Zhang, W. (2019). CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(12), 10015–10024. doi:10.1109/TGRS.2019.2930982

Zhang, Y., Hu, Q., Xu, G., Ma, Y., Wan, J., & Guo, Y. (2022). *Not all points are equal: Learning highly efficient point-based detectors for 3D LiDAR point clouds* [Conference session]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.01838

Zhou, L., Sun, P., Li, D., & Piao, J. (2022). A novel object detection method in city aerial image based on deformable convolutional networks. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 31455–31465. doi:10.1109/ACCESS.2022.3156953

Honghuan Chen was born in Shaoxing, Zhejiang, China, in 1989. He received his M.S. degree in control theory and science from the College of Automation at Hangzhou Dianzi University in 2015. Since 2019, he has been a lecturer at the College of Internet of Things Technology, Hangzhou Polytechnic. He is currently pursuing his Ph.D. at Hangzhou Dianzi University. His research interests include semantic analysis, natural language processing, and attention mechanism.

Keming Wang graduated in 2008 with a bachelor's degree in Communication Engineering from Dalian Maritime University. Since 2021, he has been an Assistant Professor in College of Internet of Things Technology at Hangzhou Polytechnic. His primary research areas are remote sensing and marine engineering.