D2-Net: Dilated Contextual Transformer and Depth-wise Separable Deconvolution for Remote Sensing Imagery Detection

Huaping Zhou[1]
Anhui University
of Science and Technology

Qi Zhao[2]*
Anhui University
of Science and Technology

Kelei Sun[3]
Anhui University
of Science and Technology

***Abstract*—Remote sensing-based object detection faces challenges in arbitrary orientations, complex backgrounds, dense distributions, and large aspect ratios. Considering these issues, this paper introduces a novel method called D2-Net, which incorporates a transformer structure into a convolutional neural network. First, a new feature extraction module called dilated contextual transformer block is designed to minimize the loss of object information due to complex backgrounds and dense targets. In addition, an efficient approach using depth-wise separable deconvolution as an up-sampling method is developed to recover lost feature information effectively. Finally, the circular smooth label is incorporated to compute the angular loss to complete the rotated detection of remote sensing images. Experimental evaluations are conducted on the DOTA and HRSC2016 datasets. On the DOTA dataset, the proposed method achieves 79.2% and 78.00% accuracy in horizontal and rotated object detection, respectively; it achieves 94.00% accuracy in the rotated detection of the HRSC2016 dataset. The proposed model shows a significant performance improvement over other comparative models on the dataset, which verifies the effectiveness of our proposed approach.**

Keywords*—YOLOv7; dilated contextual transformer; depth-wise separable deconvolution; circular smooth label; remote sensing

I. Introduction

Due to advances in computer processing power, object detection has developed rapidly over the past decade. This task typically accomplishes by utilizing single-stage detectors, typified by the YOLOs models [1], [2], [3], [4], [5], and dual-stage models exemplified by the RCNN series [6], [7], [8], [9].

Despite significant advances in generic target detection, the mission in remote sensing images (RSIs) faces numerous challenges due to characteristics such as substantial variations in scale, crowded and small targets, arbitrary orientations, and large aspect ratios[10]. Therefore, detection using oriented bounding boxes (OBBs), which can handle object rotation, has become critical in remote sensing applications. Existing rotated object detection models are often constructed with pure convolutional neural networks (CNNs) or CNN-transformer hybrid structures. And the former has a lot of representative work. Pixels-IoU Loss improves performance for complex backgrounds and large aspect ratios[11] but increases training time. Rotational region convolutional neural network (R2CNN) introduces joint prediction of axis-aligned bounding boxes and inclined minimum area boxes to complete text recognition in any direction[12]. A joint image cascade (ICN) and feature pyramid network (FPN) can capture semantic features at multiple scales [13]. Adaptive period embedding (APE) proposed by Zhu et al. represented oriented targets in a novel way and length-independent IoU (LIIoU) suitable for long targets [14]. Kim B et al. developed TricubeNet, which locates oriented targets according to visual cues such as heat maps rather than oriented box offset regression [15].

Although CNNs have achieved impressive performance, they are limited by the difficulty of obtaining long-range dependencies, resulting in deficient performance in remote sensing detection. In contrast, the unique structure of the transformer allows it to compensate well for the shortcomings of CNN. Many hybrid CNN-transformer networks have achieved satisfactory results [16], [17], [18], [19], [20]. The RoI transformer technique utilizes spatial transformations on Regions of Interest (RoIs) and learns the spatial transformation parameters by using OBB annotations as supervision. This approach results in fewer mismatches during detection [16]. To address the boundary loss and spatial receptive field issues in RSIs, Dai et al. developed a rotating object detection transformer-based model (RODFormer) [17]. Another improved detector, CLT-Det, leverages correlation learning and a transformer to tackle the problem of large-scale variation and dense targets [18]. TransConvNet uses a self-attention block and CNN to aggregate broad and specific details, offsetting the CNN's lack of rotational invariance [19]. Li et al. propose an adaptive points learning method that effectively obtains geometric information for instances of arbitrary orientations [20].

The above information suggests that incorporating a transformer module into CNN can help overcome the model's difficulty in global feature modeling. And recent researches show that simple hybrid networks can acquire the same effect as many excellent complex models [21]. Therefore, this paper presents the dilated contextual transformer block (DCoT) combined with efficient layer aggregation networks (ELAN) in YOLOv7[5] to improve the model's feature extraction capability. DCoT extraction provides more feature information with a larger receptive field, allowing shallow location information to combine effectively with deep semantic information, which improves detection ability in complex backgrounds and dense objects. Second, a depth-wise separable deconvolution (DS-DeConv) module is proposed to enable the model to generate more diverse feature information during upsampling, thereby improving its ability to detect small and dense objects. Finally, the Circular Smooth Label (CSL)[22] is integrated into the baseline YOLOv7[5] to complete the rotation detection process without being affected by boundary discontinuities. Extensive experiments were conducted on DOTA v1.0[10] and HRSC2016[23] datasets to validate the efficacy of the proposed method. The experimental results demonstrate that the proposed model enhances the detection capacity of RSIs. Moreover, it achieves real-time detection with a slight reduction in the number of parameters, striking a balance between accuracy and speed.

The main contributions of this paper can be summarized as follows. First, we use DCoT to improve the model's ability to obtain contextual information, which enhances the model's ability to detect complex backgrounds and dense targets in RSIs. Second, we use DS-DeConv for upsampling, which effectively preserves detailed feature information, enhancing the model's detecting ability of small objects. Finally, CSL is integrated into YOLOv7 to complete the rotated detection of multi-directional objects in RSIs. The proposed model outperforms other comparative models in detection.

The upcoming sections are structured as follows. Section 2 the related work, including pure CNN and CNN-transformer hybrid detection models. Section 3 provides a detailed description of the proposed methods integrated into the D2-Net. Section 4 is the experimental details and analyses of the experimental results. Finally, the conclusion is presented in Section 5.

II. Related Work

A. Pure CNN Detection Models

Depending on whether models generate region proposals, detection models consist of two types: single-stage detection methods and two-stage detection algorithms.

The single-stage detection methods directly predict object class and location without region proposal, resulting in faster inference and lower computational complexity than two-stage models. Redmon J et al. proposed the first generation of YOLO [1], which starts with real-time object detection time. This model views target recognition as a regression task and detects the presence of an object by determining whether the object's center point falls within a particular grid cell, which is obtained by dividing the image into multiple grid cells. Inevitably, it cannot solve problems of dense, small, and large aspect ratio targets and other issues that inspire other researchers to make further progress. To improve the accuracy of small target detection, SSD [24] feeds multiple features extracted from different layers of the feature extraction model to the object prediction module. It also simplifies the training process for targets with different shapes by assigning different scales and aspect ratios to the prior bounding boxes associated with each grid cell. The method used convolutional layers instead of fully connected layers and produced the same results as contemporaneous two-stage detection models. More recently, an enhanced SSD [25] introduces interactive multiscale attention to acquiring more effective feature representation capability. Retinanet [26] incorporates focal loss and effectively addresses the class imbalance problem, resulting in high speed and accuracy performance.

Two-stage detectors also gained significant attention due to their remarkable accuracy and robustness. RCNN[6] treats the detection task as a classification problem. In the first stage, it extracts region proposals from each image, then predicts targets' categories after computing features in CNN. FPN [27] regards layers with consistent feature map sizes as a stage and achieves the top-down integration of multi-scale feature maps through successive stages. It distributes features based on object scale, merging deep-level semantic information with shallow-level fine-grained information to perform more accurately. Mask R-CNN [9] innovates RoI alignment to mitigate date missed owing to feature quantization during the RoI pooling process.

B. Transformer Detection Models

Since transformers were introduced to computer vision, many distinctive models emerged. Vision transformers divide the image into multiple patches, provide them with positional embedding, and then feed the feature information into the head for detection.[28] This allows the model to be independent of image size. DINO improves DETR-like models in terms of performance and efficiency by using a comparative denoising training method, a hybrid query selection method for anchor initialization, and a look-forward double scheme for box prediction.[29] Biformer proposes a novel dynamic sparse attention via bi-level routing for more flexible computational allocation and content awareness, enabling dynamic query-aware sparsity.[30]

C. CNN-Transformer Hybrid Detection Models

The emergence of the Transformer structure compensates for the shortcomings of the pure CNN structure in obtaining long-range dependencies and contextual information, leading to numerous Transformer-related models. However, the pure transformer models have high memory consumption and complexity. So more models fuse the transformer module with CNN by insertion or replacement to achieve a balance. RoI transformer [16] conducts spatial transformations on RoIs, learning transformation parameters supervised by OBB annotations, which solves dense RSI targets and RoI-target mismatches. RODFormer [17] addresses boundary loss and spatial receptive field lack in RSI detection via a structured transformer model. CLT-Det [18] presents a correlation learning detector for solving the problem of large-scale variation and dense targets. TransConvNet [19] merges a self-attention block and CNN, aggregating the detailed and specific information to compensate for the CNN's deficiency in rotational invariance. Li et al. proposed a robust adaptive points learning methodology to extract the geometric information of instances of arbitrary orientations [20].

To summarize, the combination of transformer and CNN can effectively overcome the limitation of CNN structures in capturing features at varying scales and improve the accuracy and robustness of object detection.

III. Methods

In the following parts of this section, we begin with a short introduction to the overall architecture of the proposed D2-Net, taking YOLOv7 [5] as the baseline model. Next, we present

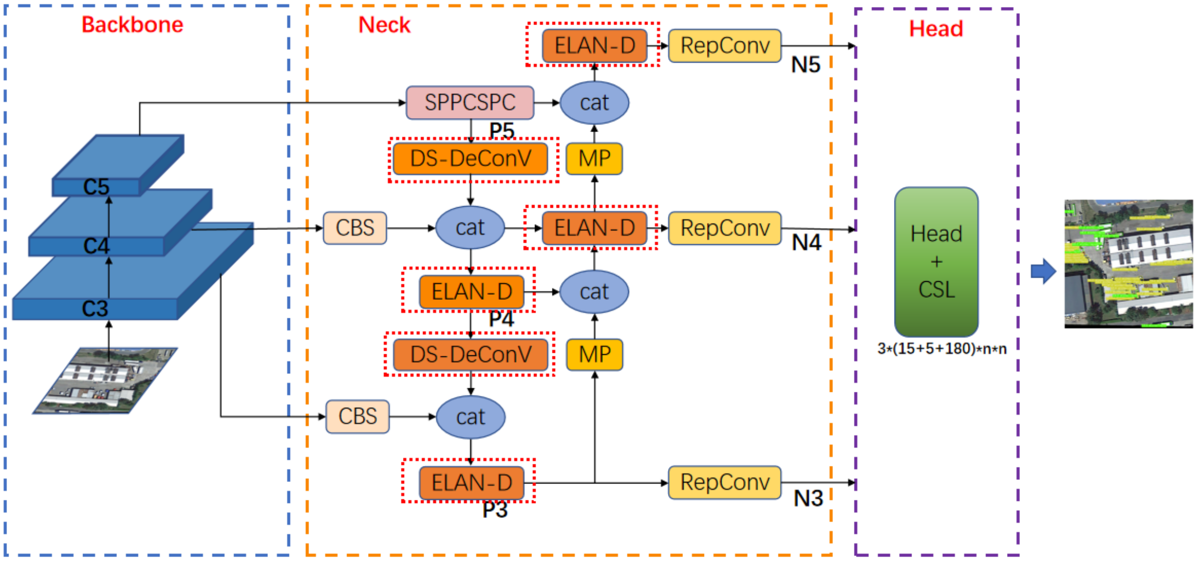


Fig. 1. The overall structure of our network. The SPPCPSC,MP, RepConv are modules of the original YOLOv7. And the detailed composition of each block in Fig. 1 is illustrated in Fig. 2 and Fig. 3.

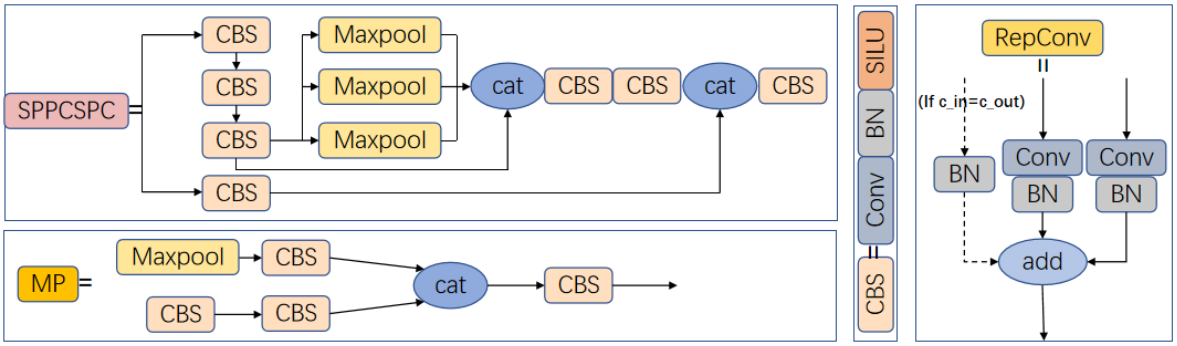


Fig. 2. Detailed block consistency of the neck network. The ELAN-D module is depicted in Fig. 4.

a detailed description of the DCoT block and the depth-wise separable deconvolution. Finally, we briefly discuss the CSL [22], which is integrated into our model to accomplish the task of rotation detection. Fig. 1 and Fig. 2 show the overall structure and detailed block consistency, respectively.

A. The D2-Net Structure

As depicted in Fig. 1, the backbone network extracts feature maps c_i, which are then sent to the neck network, where $i = 3, 4, 5$ represents the level of features, and C_i has a stride of 2^i and is $1/2^i$ pixel density of the input image size $W \times H$. The neck network consists of two modules. The initial component is the FPN [25] architecture, which propagates semantic features from higher to lower resolutions. The second module utilizes the PAFPN [31] module. To compensate for the loss of fine-grained information caused by resolution reduction, an ascending feature merging is employed to transfer location details to feature maps at deeper layers. Furthermore, depth-wise separable deconvolution makes the most suitable up-sample method by itself, and the improved ELAN module is adopted to improve the reception capability of contextual information of the network. Different scales feature maps containing detailed semantic and rich localization information are output to the RepConv block. Finally, the head network with CSL predicts object categories and position information regarding the angular problem as classification.

In our method, after being processed by the improved neck network, the output feature representations with various resolutions achieve a balance between semantic information in deep and shallow spatial details, leading to improve detection performance.

B. Contextual Transformer Block with Dilated Convolution

Drawing inspiration from the self-attention mechanism in Transformer models, numerous scholars have investigated the effectiveness of hybrid networks mixed by CNNs and transformers in computer vision task scenarios [16], [17], [18], [19]. And as existing researches prove, through the simple fusion of CNNs and transformers, object detection models pay more attention to more useful features so that the performances of those models are improved. Therefore, the hybrid network, including the transformer module, has a good prospect in the RSIs detection task.

The traditional self-attention modules utilize input feature information obtained from various spatial positions to process input data. Nevertheless, these modules acquire knowledge of all possible query-key connections by training on individual query-key pairs. This process occurs independently, without considering the contextual information between their interactions. The CoT [32] architecture can integrate abundant contextual information and Contribute significantly to the visual representation of 2D images. Nevertheless, the standard convolution operation will lose much localization information in feature processing. Therefore, we replace it with dilated convolution to form DCoT, which effectively makes the network increase the receptive field while obtaining more information. Then we displace the last three CBS modules of ELAN with DCoT to form ELAN-D (see Fig. 4), which reduces the calculation amount and FLOPs. By combining the strengths of the Transformer and CNN, the DCoT module can capture both global and detailed local information from input features. This approach improves the network model's ability to represent input information features, leveraging the advantages of each component. It showed the architecture of the DCoT block in Fig. 3.

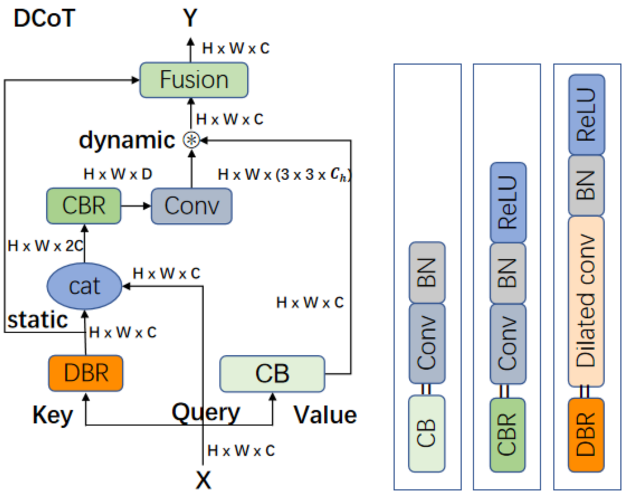


Fig. 3. The detailed structure of the DCoT block and its module. H, W, and C denote the height, width, and number of channels of the input data X, ⊛ denotes local matrix multiplication.

For input feature X, it is processed through three pathways, namely Q(queries), K(keys), and V(values), to generate more feature information. The keys undergo dilated convolution to capture local information and increase the receptive field. Then, K is concatenated with X to supplement local information and passed through a CBR module and a standard convolution to generate Q. Finally, Q is multiplied with V and fused with K to obtain the final output Y. The Q, K, and V can be written as:

$$Q = [K, X]W_{CBR}W_C \tag{1}$$

$$K = XW_{DBR} \tag{2}$$

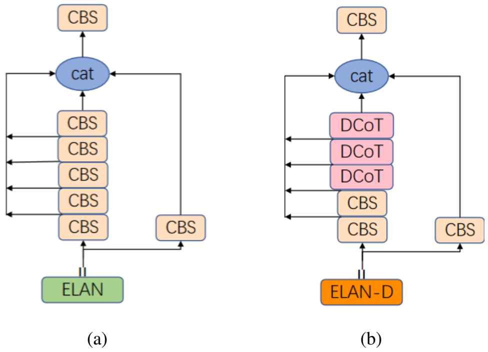


Fig. 4. The architectures of the ELAN block and ELAN-D block. (a) shows the detailed ELAN structure, and (b) shows the detailed ELAN-D block.

$$V = XW_{CB} \tag{3}$$

where X is the input feature, $W_{\square}$ are different convolutional blocks.

C. Depth-wise Separable Deconvolution for Up-sampling

During object detection with deep learning, the resolution of the feature map tends to decrease as the network deepens, leading to a loss of information. Thus, up-sampling is essential for an algorithm. In the YOLO algorithms, nearest neighbor interpolation is employed for up-sampling. However, focusing solely on the nearest pixels has also resulted in image quality and details loss, especially for tiny targets. Deconvolution is also a commonly used up-sampling method. Compared with neighbor interpolation, it performs better than in preserving feature information. However, it produces more parameters as well. Deep separable convolution [33] disassembles traditional convolution into depth convolution and point convolution, which can make the model more efficient and parameter reduction.

In this paper, we propose the DS-DeConv block for up-sampling. With this method, more diverse pixel values can be produced when recovering the feature map's resolution, which makes the Acquired feature map preserve more details and features of the original feature map.

We also introduce group convolution and change the filter size of deconvolution to decrease the parameter quantity caused by deconvolution. Our DS-DeConv method improves network model accuracy in up-sampling with a slight increase in parameters. Fig. 5 illustrates the principal diagram of DS-DeConv, while the number of deconvolution groups is adjusted based on the channel quantities in the network.

D. Rotationally Detection

Currently, bounding boxes in object detection consist of HBBs, rotated bounding boxes, and custom bounding boxes. The characteristics of remote sensing detection include the random and diverse directions of the objects to be detected. And to achieve more accurate detection of these rotating objects, the rotating bounding box is used for it.

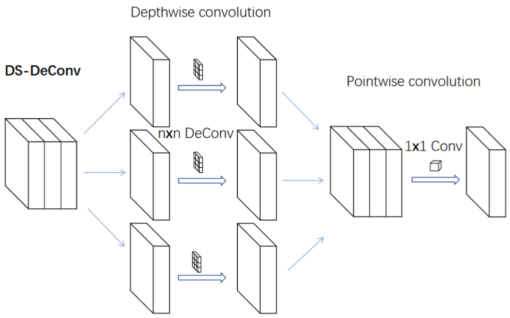


Fig. 5. The structure of depth-wise separable deconvolution.

The rotated detection method based on parametric regression mainly consists of the five parameters and the eight-parameter method. However, in rotation detection, the target parameters for learning are periodic, which causes the learned parameters to be located at the boundary periodicity, resulting in discontinuity issues and an abrupt rise of loss. Therefore, we use CSL [22] to solve the boundary discontinuity problem, as depicted in Fig. 6.

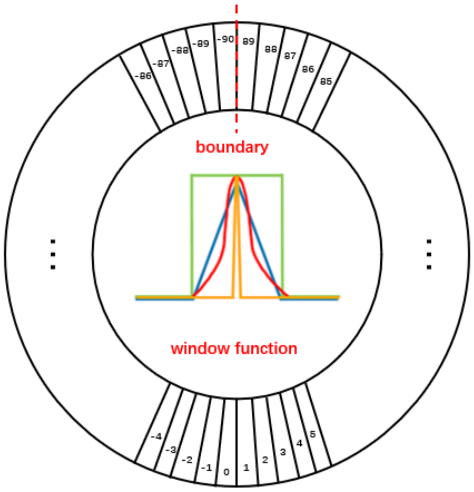


Fig. 6. The schematic diagram of the CSL.

The CSL is expressed as follows:

$$\mathrm{CSL}_{(x)} = \begin{cases} g(x), & \theta - r < x < \theta + r \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $g(x)$, r, and θ represents the window function, radius, and the current bounding box angle, respectively. By converting angle prediction from a regression task to a classification task, the boundary discontinuity issue can be effectively resolved with minimal loss of accuracy.

IV. Experiments and Results Analysis

A. Datasets

1) DOTA Dataset: The DOTA dataset [10] contains 2806 high-resolution aerial images collected from various sensors and platforms and encompasses 15 categories. It is split into three subsets for training, validation, and testing, including 1411 images, 458 images, and 937 images, respectively, containing 188282 instances in total. The image size varies from 800 × 800 to 4000 × 4000 pixels.

2) HRSC2016 Dataset: The HRSC2016 dataset[23] includes 1061 remote sensing images from six distinct ports. The dataset is divided into three parts, 436 images for training (a total of 1207 labeled examples), 444 images for testing (a total of 1228 labeled examples), and 181 images for validation (a total of 541 labeled examples). The images have varying resolution, ranging from 300 × 300 to 1500 × 900 pixels.

B. Implementation Details and Evaluation Index

Considering the adverse influence of high and inconsistent resolution images, we reprocess the original data of these two datasets. For the DOTA dataset, we cropped the images to 1024×1024 resolution with 200 pixels overlapping area. Then 15749 images were extracted for training and 5297 images for evaluation, and the final test results are obtained through the official evaluation server. The network is trained with the SGD optimizer in the training process. The lr (learning rate) is 0.001, and momentum and weight decay are 0.937 and 0.0005. We train 300 epochs with batch size 16 on two GeForce RTX 3090 GPUs. For the HRSC2016 dataset, we resized all the images to (768, 768). The network is trained with the SGD optimizer for training. The learning rate is 0.01, and momentum and weight decay are 0.937 and 0.0005. We train 200 epochs with batch size 8 on GeForce RTX 3060 GPU.

We adopt the Average Precision (AP) and the mean AP (mAP @0.5) metric in the comparative experiments to evaluate the multi-class detection accuracy. They can be calculated as follows:

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{5}$$

$$\mathrm{AP} = \int_0^1 P d_r \tag{6}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{7}$$

TP is the correctly classified target number, while FP is the background number recognized as target. The accuracy rate P can be defined as the proportion of correctly detected targets among all detection results. The mAP is the average of AP values of all classes. In the ablation experiments, FLOPs and speed are also used to estimate the differences in algorithm capability. Speed is also used to estimate the differences in algorithm capability.

C. Ablation Experiments

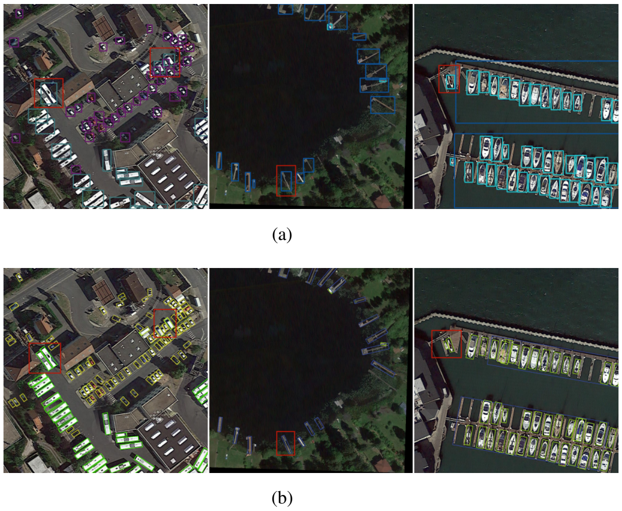
In this section, we choose YOLOv7 as the baseline model to conduct ablation experiments on the DOTA dataset to verify the effectiveness of the introduced DCoT block, DS-DeConv, and CSL. It should be noted that this paper aims to address the problem of rotated RSI detection, so unnecessary ablation experiments on horizontal detection are not shown. The batch

TABLE I. The Result of the Ablation Experiment

\	YOLOv7	CSL	DS-DeConv	DCoT	FLOPs(G)	Speed(ms)	mAP/HBB(%)	mAP/OBB(%)
①	✓				**103.4**	**90.9**	73.70	\
②	✓	✓			106.5	43.7	75.60(+1.90)	74.71
③	✓	✓	✓		106.5	45.2	77.10(+1.50)	75.12(+0.41)
④	✓	✓		✓	106.4	39.4	76.4(-0.7)	75.76(+0.64)
⑤	✓	✓	✓	✓	106.4	39.4	**79.20(+2.10)**	**77.96(+2.84)**

TABLE II. The detailed result of the ablation experiment. PL: plane, BD: baseball diamond, BR: bridge, GFT: ground field track, SV: small vehicle, LV: large vehicle, SH: ship, TC: tennis court, BC: basketball court, ST: storage tank, SBF: soccer-ball field, RA: roundabout, HA: harbor, SP: swimming pool, HC: helicopter.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
①	93.80	73.40	48.00	**72.90**	71.50	88.80	89.50	94.90	72.10	**76.80**	**67.50**	57.60	85.80	62.40	50.50	73.70
②	**98.40**	81.70	**50.90**	59.00	87.50	92.10	97.50	96.80	85.70	70.30	48.10	57.80	**87.60**	64.00	57.10	75.60
③	**98.40**	81.60	48.90	58.60	87.00	91.10	97.20	97.20	82.70	78.70	50.20	**59.10**	86.50	63.80	75.90	77.10
④	97.30	78.40	47.90	65.00	84.60	91.20	97.00	96.60	85.40	75.70	57.80	**59.50**	84.70	58.60	66.90	76.40
⑤	98.30	**84.50**	47.90	61.70	**87.70**	**92.90**	**97.50**	**98.10**	**88.00**	76.40	57.60	58.40	86.30	**66.40**	**86.40**	**79.20**


(a)

(b)

Fig. 7. Some contrastive detection results. (a) is the result of the baseline; (b) is the result of the D2-Net. And the differences are highlighted in red.

size for training was 16, and the performance metrics were evaluated every 10 epochs during the training process. A total of 300 iterations were completed to train both the baseline and improved models. FLOPs, speed, and mAP are used as evaluation indicators in the experiments. Table I shows the results of our improvements and Table II shows detailed AP values of each category conducted on the DOTA dataset. And the bold font is the best result.

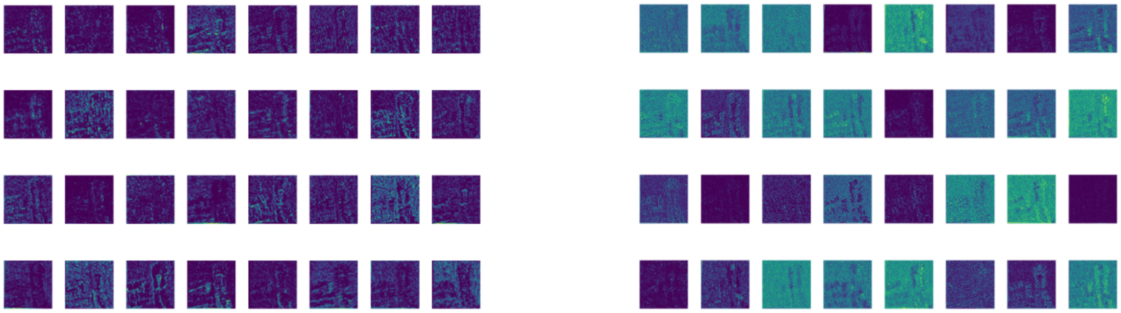
As seen from Table I speed and mAP of the OBB task are commonly lower than those in the HBB task, which is attributed to the angle issue when serving the rotated detection task. Attentively, to ensure the effectiveness of the baseline, its experiments were all performed at $640 * 640$ resolution, while other experiments were conducted at $1024 * 1024$ resolution. And the baseline speed is 40.98 at $1024 * 1024$ resolution. Despite the speed and mAP having decreased, the effect has been improved in the actual detection(see Fig. 7). In the horizontal task, compared with the original YOLOv7, ②③④⑤ showed improvement of 1.9%, 3.4%, 2.7% and 5.5%. Relative to the YOLOv7 with CSL added, ③④⑤ achieved 0.41%, 1.05% and 3.25% improvement. According to Table II, it can be found that the proposed method has greatly improved in the categories of small vehicles, harbors, and ships, obtaining 16.2%, 35.9%, and 8% improvement, respectively, compared with the baseline model.

In Fig. 7, three images are chosen for comparing the detection results from the dataset. The results of the two rows are the baseline model, and the D2-Net model proposed in this paper, respectively. There are plenty of small and dense objects in the leftmost images of Fig. 7(a) and Fig. 7(b). It can be seen from the red highlights that the baseline model loses some targets, while the proposed model detects them very effectively. The background of the middle image is similar to the object, and the baseline's results are affected, while the proposed model works well. The right image contains many targets with large aspect ratios, and the D2-Net is more accurate than the baseline when boxing targets and no targets are lost.

In Fig. 8(a) and Fig. 8(b), to prove the feature extraction capability of the DcoT modules, we made the first 32 feature maps visualization in the same stage of both baseline and the D2-Net. It can be observed that the proposed model can effectively eliminate irrelevant information from the background and has good extraction capability for detecting targets. It is the DCoT modules that enable the network to fully utilize feature information and concentrate on detecting targets with distinguishable features. Fig. 9(a) and Fig. 9(b) show the first upsampling heatmaps of the baseline and D2-Net. The latter preserves more useful feature information around objects and eliminates unnecessary noise. It proves the DS-DeConv's effectiveness when detecting small targets.

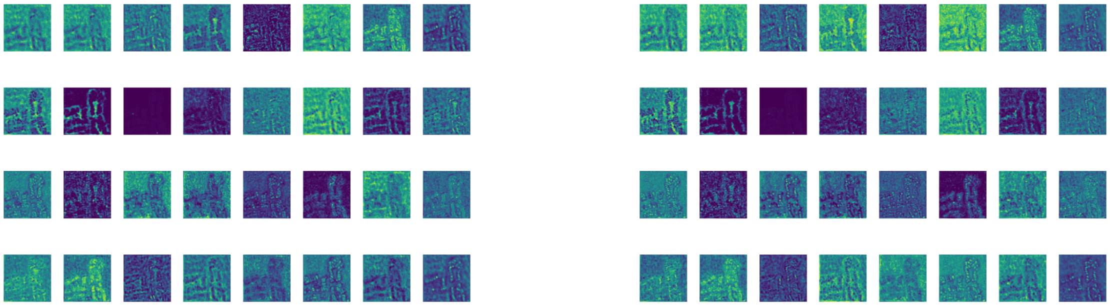
D. Comparison with other OBB Methods

In this section, we choose the YOLOv7 as the baseline. We compare our model performance with other state-of-the-art methods for the DOTA-v1.0 and HRSC2016 datasets. In compared models, RoI Trans [16], RODFormer [17], and CLT-Det [18] adopted a hybrid network using CNNs and transformer blocks, and the others applied pure CNNs structure.



(a) (b)

Fig. 8. The DCoT visualizations of the first 32 features: (a) represents the baseline result, and (b) is the result of the D2-Net.



(a) (b)

Fig. 9. The first upsampling visualizations of the first 32 features: (a) represents the baseline result, and (b) is the result of the D2-Net.

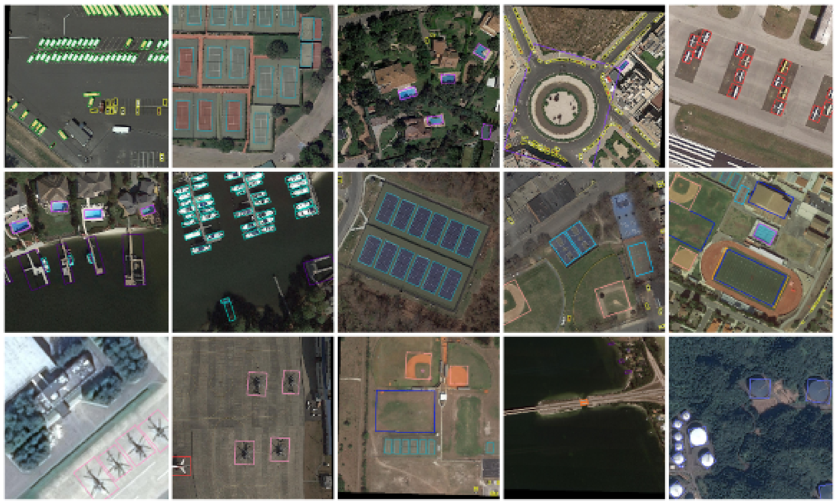


Fig. 10. Visualization of the detection results of our method on the DOTA data set.

TABLE III. OBB Task Performance Comparisons on the DOTA-v1.0 Test Set (AP (%) for Each Category and Overall mAP @0.5 (%). In the Column, the Bold Denotes the Best Detection Results

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
OBB																
R2CNN[12]	80.94	65.67	35.34	67.44	59.52	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RADet[34]	79.45	76.99	48.05	65.83	65.46	74.40	68.86	89.70	78.14	74.97	49.92	64.63	66.14	71.58	62.16	69.09
Axis Learning[35]	79.53	77.15	38.59	61.15	67.53	70.49	76.30	89.66	79.07	83.53	47.27	61.01	56.28	66.06	36.05	65.98
RoI Trans[16]	88.64	78.52	43.44	75.92	75.92	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
DRN[36]	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
RODFormer[17]	89.76	79.64	**56.61**	71.57	78.60	85.29	**89.93**	90.53	87.73	83.05	60.19	60.34	66.03	69.75	64.95	75.60
CLT-Det[18]	89.31	**85.69**	53.97	**77.11**	79.66	79.01	88.55	**90.89**	85.36	86.56	63.92	68.47	75.65	70.65	66.91	77.45
CSL[22]	90.25	85.53	54.64	75.31	70.44	73.51	77.62	90.84	**86.15**	86.69	69.60	68.04	73.83	71.10	68.93	76.17
Ours	**90.54**	82.99	43.20	55.39	**81.71**	**87.40**	89.82	90.80	83.58	**89.71**	**76.62**	**72.27**	75.00	67.57	**82.87**	**77.96**
HBB																
GraphFPN[37]	89.32	68.88	50.41	60.42	70.91	79.45	86.18	90.80	83.11	80.35	53.01	60.98	75.95	64.36	58.71	71.52
SCRDet[38]	90.18	81.88	55.30	**73.29**	72.09	77.65	78.06	90.91	82.44	86.39	64.53	63.45	75.77	78.21	60.11	75.35
Mask OBB[39]	89.69	**87.07**	**58.51**	72.04	78.21	71.47	85.20	89.55	84.71	**86.76**	54.38	**70.21**	78.98	77.46	70.40	76.98
YOLOv7[5]	93.80	73.40	48.00	72.90	71.50	88.80	89.50	94.90	72.10	76.80	**67.50**	57.60	85.80	62.40	50.50	73.70
CGL[40]	89.53	82.85	56.53	**76.52**	79.29	83.39	88.19	90.90	86.67	85.07	63.40	68.23	77.82	**78.77**	50.23	77.16
Ours	**98.30**	84.50	47.90	61.70	**87.70**	**92.90**	**97.50**	**98.10**	**88.00**	76.40	57.60	58.40	**86.30**	66.40	**86.40**	**79.20**

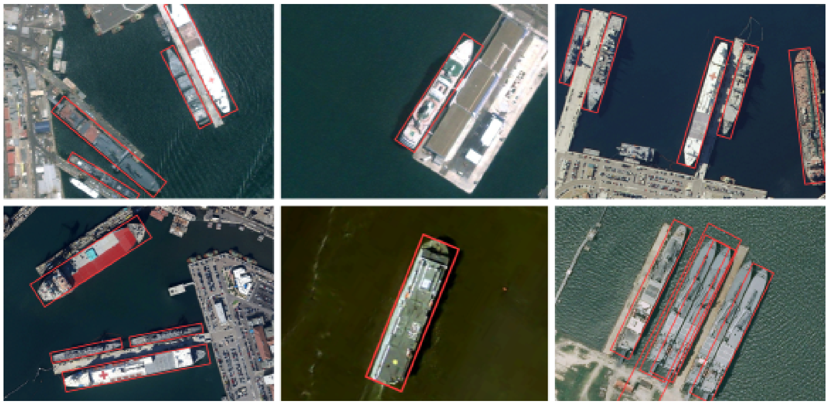


Fig. 11. The visualization of the detection results of our method on the HRSC2016 dataset.

1) Results on DOTA-v1.0: As reported in Table III, The comparative experiments on the DOTA dataset consist of the OBB and HBB tasks. In the OBB task, we achieved the mAP of 77.96%, which gains 1.79% higher than the CSL with CNNs structure, and 0.51% higher than CLT-Det with a hybrid framework. Moreover, the prediction performance on densely distributed small objects, like storage tanks and small vehicles, has improved enormously, reaching 89.71% and 81.71%, which are 3.02% and 2.05% higher than the second best, respectively. Besides, soccer ball fields, large vehicles, and helicopters also perform well, reaching 76.62%, 87.4%, and 82.87%, respectively. In the HBB task, the proposed model is 5.5% (from 73.70 to 79.20%) higher than the baseline. The top-3 mAP is plane, tennis court, and ship, achieving 98.3%, 98.1%, and 97.5%, respectively. In general, the above statement demonstrates the effectiveness of our model, and Fig. 10 visualizes some detection results of our method on the DOTA dataset.

2) Results on HRSC2016: The HRSC2016 dataset consists of plenty of oriented ships. As shown in Table IV, many classical detection algorithms have attained excellent performance in this dataset, such as R2CNN [12], RoI Trans [16], CLT-Det [18], CSL [22], Axis Learning [35], and Oriented R-CNN [41]. Our model uses the A block for enhancing feature extraction and depth-wise separable deconvolution for upsampling. It achieves an mAP value of 94.00% with the 768×768 resolution, outdoing several of the mentioned methods. And the visualization of some detection results is depicted in Fig. 11.

TABLE IV. Performance Comparisons on the HRSC2016 OBB Task. The Best Result is Highlighted in Bold

Methods	mAP	Resolution
R2CNN[12]	73.07	800×800
Axis Learning[35]	78.15	800×800
RoI Trans[16]	86.20	512×800
SLA[42]	89.51	768×768
CLT-Det[18]	89.72	512×800
CSL[35]	89.62	800×800
Oriented R-CNN[41]	90.50	1333×800
Attention-Points[43]	90.59	1333×800
Ours	94.00	768×768

V. Conclusions

In this paper, we proposed an effective one-stage model called D2-Net for rotated remote sensing image detection based on the YOLOv7 model. we innovate the DCoT block combining dilated convolution with contextual transformer block for feature extraction and enhancing the ability to detect

Objects with tiny sizes and dense distribution of RSIs, which can fully utilize the global and local information of objects and enlarge the receptive field. Then, We designed the DS-DeConv for up-sampling, which mitigates the effects of complex backgrounds and low resolution. It improves the resolution and quality of the up-sampled feature maps, enabling the detector to capture the details and shapes of the targets more effectively. Additionally, the CSL is employed for determining the angle loss and accomplishing the prediction of rotated objects in RSIs. In the end, we conducted experiments on the DOTA and HRSC2016 datasets to prove the effectiveness of D2-Net. Although detection capability surpasses other commonly employed algorithms, the speed and FLOPs has decreased. Thus, we will further enhance the feature representation and improve the model's detection speed with a more lightweight model.

ACKNOWLEDGMENT

The authors thank Anhui University of Science and Technology for its support. Secondly, they thank ITVR-AUST laboratory for its support.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[2] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263-7271.

[3] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[4] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[5] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2023, pp. 7464-7475.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[7] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961-2969.

[10] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3974-3983.

[11] Z. Chen, K. Chen, W. Lin, J. See, H. Yu, Y. Ke, and C. Yang, "Piou loss: Towards accurate oriented object detection in complex environments," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, 2020, pp. 195-211.

[12] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," arXiv preprint arXiv:1706.09579, 2017.

[13] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III, 2019, pp. 150-165.

[14] Y. Zhu, J. Du, and X. Wu, "Adaptive period embedding for representing oriented objects in aerial images," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7247-7257, 2020.

[15] B. Kim, J. Lee, S. Lee, D. Kim, and J. Kim, "TricubeNet: 2D kernel-based object representation for weakly-occluded oriented object detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 167-176.

[16] J. Ding, N. Xue, Y. Long, G. S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2849-2858.

[17] Y. Dai, J. Yu, D. Zhang, T. Hu, and X. Zheng, "RODFormer: High-Precision Design for Rotating Object Detection with Transformers," Sensors, vol. 22, no. 7, p. 2633, 2022.

[18] Y. Zhou, S. Chen, J. Zhao, R. Yao, Y. Xue, and A. El Saddik, "CLT-Det: Correlation learning based on transformer for detecting dense objects in remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-15, 2022.

[19] X. Liu, S. Ma, L. He, C. Wang, and Z. Chen, "Hybrid network model: Transconvnet for oriented object detection in remote sensing images," Remote Sensing, vol. 14, no. 9, p. 2090, 2022.

[20] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1829-1838.

[21] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng, and Z. Zha, "A battle of network structures: An empirical study of cnn, transformer, and mlp," arXiv preprint arXiv:2108.13002, 2021.

[22] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, 2020, pp. 677-694.

[23] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," 2017.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016, pp. 21-37.

[25] S. Zhou and J. Qiu, "Enhanced SSD with interactive multi-scale attention features for object detection," Multimedia Tools and Applications, vol. 80, pp. 11539-11556, 2021.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980-2988.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117-2125.

[28] Dosovitskiy, A. et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[29] Zhang, H. et al., "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," arXiv preprint arXiv:2203.03605, 2022.

[30] Zhu, L. et al., "BiFormer: Vision Transformer with Bi-Level Routing Attention," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10323-10333, 2023.

[31] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759-8768.

[32] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[34] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images," Remote Sensing, vol. 12, no. 3, p. 389, 2020.

[35] Z. Xiao, L. Qian, W. Shao, X. Tan, and K. Wang, ”Axis learning for orientated objects detection in aerial images,” Remote Sensing, vol. 12, no. 6, p. 908, 2020.

[36] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, C. Ma, and C. Xu, ”Dynamic refinement network for oriented and densely packed object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11207-11216.

[37] G. Zhao, W. Ge, and Y. Yu, ”GraphFPN: Graph feature pyramid network for object detection,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2763-2772.

[38] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, ”Scrdet: Towards more robust detection for small, cluttered and rotated objects,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8232-8241.

[39] J. Wang, J. Ding, H. Guo, W. Cheng, T. Pan, and W. Yang, ”Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images,” Remote Sensing, vol. 11, no. 24, p. 2930, 2019.

[40] X. Chen, C. Wang, Z. Li, M. Liu, Q. Li, H. Qi, D. Ma, Z. Li, and Y. Wang, ”Coupled Global–Local object detection for large VHR aerial images,” Knowledge-Based Systems, vol. 260, p. 110097, 2023, Elsevier.

[41] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, ”Oriented R-CNN for object detection,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3520-3529.

[42] Q. Ming, L. Miao, Z. Zhou, J. Song, and X. Yang, ”Sparse label assignment for oriented object detection in aerial images,” Remote Sensing, vol. 13, no. 14, p. 2664, 2021.

[43] C. T. C. Doloriel and R. D. Cajote, ”Improving the Detection of Small Oriented Objects in Aerial Images,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 176-185.