DCL-Net: Augmenting the Capability of Classification and Localization for Remote Sensing Object Detection

Enhai Liu, Yu Zheng, Bin Pan, Xia Xu and Zhenwei Shi

Abstract

Deep learning based remote sensing object detectors are usually composed of two branches: classification and localization. Recently proposed object detectors often follow the pipeline that classification and localization branches share the same feature maps, which leads to strong coupling relationship between them. However, when tackling remote sensing images, this strong coupling relationship may impair the performance of the detectors, because the top-view perspective of remote sensing images may result in conflicts between classification and location branches. To address this issue, we propose a decoupled classification localization network (DCL-Net) by considering the different characteristics between the two branches. Two modules are developed to suppress the strong coupling: receptive field aggregation module (RFAM) and bottom-up path aggregation module (PAM). For the classification branch, RFAM can learn the relationship between objects and context information by simulating the human receptive field, and improves the robustness of the classification branch to rotational distortions. For the localization branch, PAM can enhance the entire feature hierarchy by transferring the rich detailed information of low-level features, which helps the detector to achieve precise bounding box regression. Compared with existing methods, the major contribution of DCL-Net is that the independence of the classification and localization branches can be significantly enhanced, which may be beneficial to the detection accuracy for the objects in remote sensing images. Experiments on public data sets validate the effectiveness of our detector.

Index Terms

Convolutional neural networks (CNNs), remote sensing object detection, decoupled network.

This work was supported by the National Key R&D Program of China under the Grant 2019YFC1510900, the National Natural Science Foundation of China under the Grant 62001251 and 62001252, the China Postdoctoral Science Foundation under the Grant 2020M670631, the Natural Science Foundation of Hebei under the Grant F2019202062 and F2020202008, the Science and Technology Program of Tianjin under the Grant 18YFCZZC00060 and 18ZXZNGX00100. (*Corresponding author: Bin Pan.)*

Enhai Liu and Yu Zheng are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China (e-mail: liuenhai@scse.hebut.edu.cn; zhengyu.hebut@outlook.com).

Bin Pan (Corresponding author) is with the School of Statistics and Data Science, Nankai University, Tianjin 300071, China, and also with the Key Laboratory of Pure Mathematics and Combinatorics, Ministry of Education, China (e-mail:panbin@nankai.edu.cn).

Xia Xu is with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: xuxia@nankai.edu.cn).

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

I. INTRODUCTION

BENEFITING from the advancement of earth observation technology, nowadays a large number of remote sensing images are available, which has contributed to many applications such as object detection, semantic segmentation, etc [1]–[5]. Remote sensing object detection refers to locating the objects of interest from aerial or satellite images, and further determining which classes they belong to. As one of the basic tasks in remote sensing image interpretation, object detection is of great significance for urban planning, precision agriculture and other civilian and military applications [6]–[8].

The past decades have witnessed tremendous efforts of researchers for remote sensing object detection. Earlier studies were mainly based on handcrafted features that lack abstract semantic information [9], [10], leading to limited detection performance. In recent years, thanks to convolutional neural networks (CNNs) can automatically learn high-level features from the raw images, it has been applied to remote sensing object detection and made remarkable progress [11]. For example, the literature [12]–[15] based on multi-reference and multi-resolution detection are effective ways to address the inconsistencies of remote sensing object scales. The literature [16]–[20] have shown that appropriate learning of context can help reduce false alarms of classification task, especially when object appearance characteristics are ambiguous because of small object size, cluttered background or occlusion and so on. Moreover, since most objects in remote sensing images have indefinite direction and mutual clustering, some researchers have proposed methods based on oriented bounding boxes, which overcame the shortage of traditional horizontal bounding boxes [21]–[25]. In addition, many works have proven that attention mechanism can effectively make the network select more critical information for the current task from numerous information and suppress other useless information [26]–[30].

However, the aforementioned CNN-based detectors may suffer the problem of strong coupling between classification and localization branches when tackling remote sensing images. Coupling, also known as coupling degree, is a measure of the degree of correlation between modules. Generally, the classification and localization branches of remote sensing detectors share the same feature maps inspired by general object detection methods, which leads to strong coupling between two branches. However, due to the particularity of the perspective of remote sensing images compared to natural scene ones, the classification and localization branches may conflict during a detection process. 1) The sensitivity of classification and localization branches to feature translation and rotation is different. The classification branch is more prone to rotation and translation invariance, while the localization branch is quite sensitive to rotation and translation [31]–[33]. This conflict may be less important in natural images with small changes in orientation and limited aspect ratios [34]. However, because remote sensing objects are shot at high altitudes, arbitrary direction and large aspect ratio of objects are common problems. Therefore, the classification and localization branches of remote sensing object detectors are somewhat incompatible. 2) classification requires high-level features that contain strong semantic information, while localization relies on low-level features with rich detail information [35], [36]. Based on the above reasons, it is necessary to design an object detection network considering the different characteristics of two branches and alleviating the coupling degree between them.

In this paper, we propose a method termed decoupled classification localization network (DCL-Net), which

aims to eliminate the direct relationship between classification and location branches. In DCL-Net, we design two intermediate layers that act on the classification and localization branches of Region Proposal Network (RPN) to generate two disentangled features based on the original shared feature. For the intermediate layers, we adopt the following design principles that take the characteristics in classification and localization branches into account, respectively.

In the classification branch, we propose a multi-branch convolution structure named receptive field aggregation module (RFAM) to obtain more robust features. RFAM is inspired by the following observations: 1) The direction of objects in remote sensing images is arbitrary, but the classification branch requires rotation-invariant features. 2) Objects are easily confused in the cluttered background where context learning can provide key information to distinguish them. 3) According to the human visual perception system, the scale of population Receptive Field (pRF) is positively related to the eccentricity of their retinotopic maps [37], which indicates that the area closer to the center has stronger influence on object recognition. Based on the above observations, RFAM is developed. Firstly, RFAM learns the relationship between objects and context information by simulating the human receptive field. On the one hand, it exploits the context information to eliminate the uncertainty of the object. On the other hand, it enhances the feature of the central region of the object to prevent confusion with contextual information. In addition, it can enhance the robustness of classification branch to rotational distortions.

In the localization branch, we propose a structure named bottom-up path augmentation module (PAM), which can transfer shallow information to achieve better localization. PAM is inspired by the idea that object localization relies on low-level features with rich detail information such as local texture and edges [35], [38]. For remote sensing detection task with complex background, PAM enhances the entire feature hierarchy by aggregating the shallow features to deep features gradually in bottom-up strategies, resulting in more robust and accurate localization features.

In summary, the main contributions of our work are as follows:

- Aiming at the conflict between feature requirements of classification and localization in object detection for remote sensing images, we propose an innovative decoupling classification and location network to guide the independent learning of the two branches.
- We design a multi-branch convolution structure to enhance the distinguishability of classification features and the robustness to rotation.
- We enhance the entire feature hierarchy by transferring the rich detailed information of shallow features, which improves the localization performance.

Code will available at http://levir.buaa.edu.cn/Code.htm.

II. The Proposed Method

In this section, we first outline our DCL-Net. Subsequently, the details of RFAM and PAM are elaborated in Section II-B, II-C, respectively.

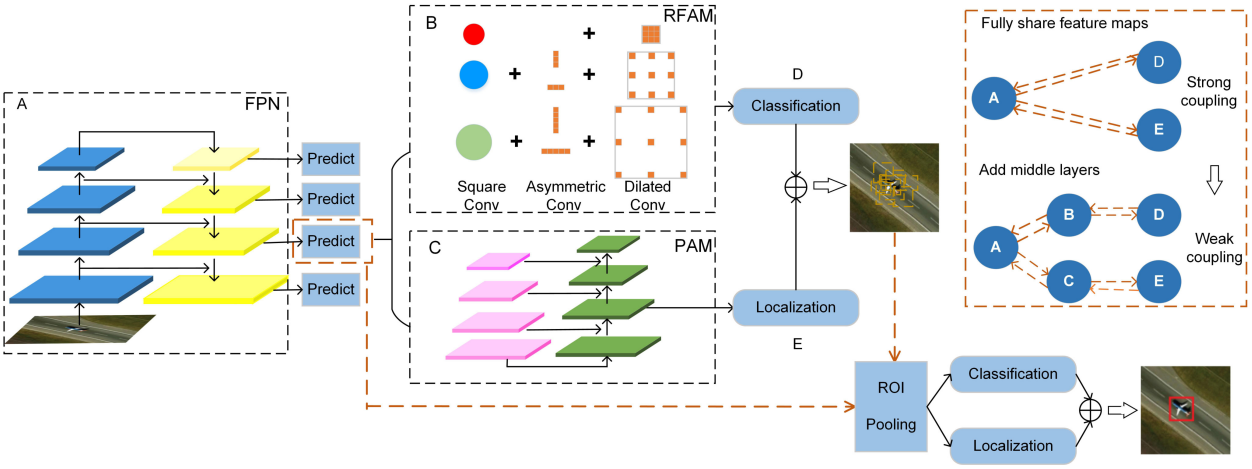


Fig. 1. The overall architecture of our proposed DCL-Net. As shown on the right side of the figure, modules D and E have a strong coupling relationship due to they share module A. This strong coupling relationship is alleviated by adding intermediate layers B and C, which are respectively applicable to D and E.

A. *Overview of DCL-Net*

Before outlining our method, we first elaborate on RPN [39] to intuitively analyze the strong coupling problem of classification and localization branches. RPN is used to generate a series of regional proposals, each with an objective score. The network slides a 3×3 window on the output feature maps of basic network. Then the features of each sliding window are mapped to a low-dimensional, and these features are fed into two sibling 1×1 convolution layers, where the classification layer outputs the reference box as the confidence of the foreground, and the other is the localization layer outputting the reference box with a coordinate offset relative to the ground truth.

From the RPN structure described above, it is clear that the classification and localization branches fully share feature maps, leading to strong coupling between two branches. However, this is not the best in remote sensing object detection due to the different feature requirements of two branches, resulting in poor performance of the RPN. Therefore, we propose DCL-Net to alleviate the strong coupling between classification and localization branches by generating two disentangled features for these two tasks. DCL-Net guarantees respective feature quality of two branches, thus providing high-quality regional proposals for the next stage. The overall structure of DCL-Net is depicted in Fig. 1. Firstly, we use Feature Pyramid Networks (FPN) [36] to extract multi-scale features $P_2 - P_5$ of raw input images. Then, RFAM and PAM act on the features of each specific scale $P_l (l \in [2, 5])$ to generate two type of features suitable for classification branch and localization branch respectively, which decoupled the RPN and generated high-quality regional proposals. Finally, the region proposals are further classified and coordinate regression. Detailed descriptions of each part are given below.

B. *Classification branch based on RFAM*

For the classification branch, our proposed RFAM is a multi-branch convolution block base on [40]. RFAM internal structure mainly includes three parts: the multi-branch convolution layer, the asymmetric convolutional

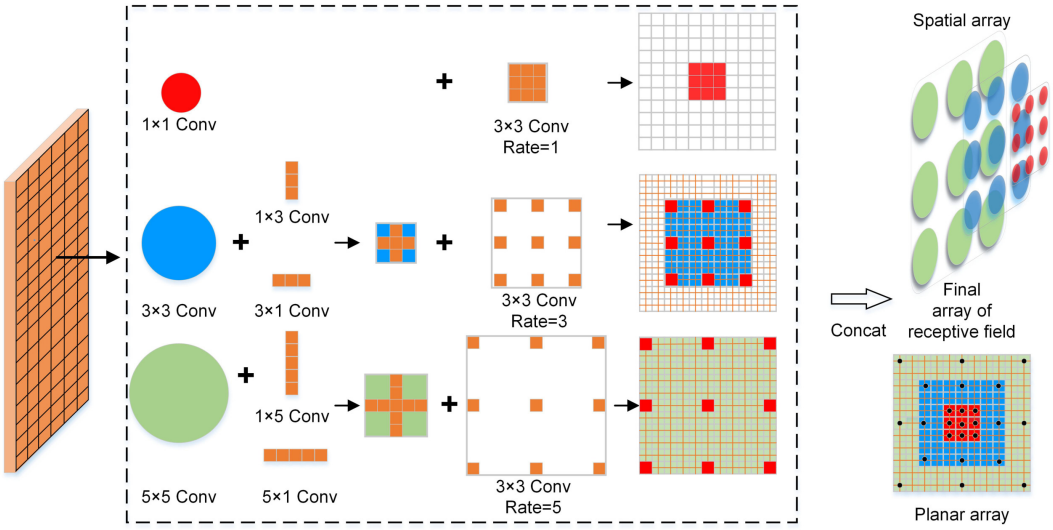


Fig. 2. RFAM is constructed by combining multi-branch convolution layer, asymmetric convolution layer and dilated convolution layer. Convolution kernels of different sizes are used to simulate pRFs. Asymmetric convolution layers to strengthen the central skeleton parts of square convolution kernels. The dilated convolution is used to control the relationship between the scale and eccentricity of pRFs. The final RF is generated by cascading all branches, which emphasizes the significance of the area nearer the center.

layer, and the dilated convolution layer. In the first part, the convolution kernels of different sizes enhance the robustness of classification features to scale change by simulating the scale of pRFs. In the second part, the asymmetric convolutional layer is responsible for improving the robustness of classification branch to rotational distortions. In the last part, the dilated convolution can increase the size of ERF (Effective receptive field) and imitate the relationship between the scale and eccentricity of pRF with the former parts. These three parts emphasize the importance of the area near the center (as illustrated in Fig. 2) by simulating human receptive fields, thereby enhancing the feature discriminability of classification features. The overall structure is illustrated in Fig. 3. Next, we will elaborate on the three parts of the RFAM and its functions.

1) Multi-branch convolution layer: The object scale of the remote sensing image varies greatly, the object with more global information distribution prefers a larger convolution kernel, and the object with more local information distribution prefers a smaller convolution kernel. Convolution kernels of different sizes can generate multi-scale receptive fields, increasing network adaptability to scale.

We mainly adopted 1×1, 3×3 and 5×5 sizes of convolution kernels to generate the multi-scale receptive field by simulating the pRFs. We learn to capture the visual pattern of large objects through convolutions of 3×3 and 5×5, so that the output features have sufficiently large receptive fields. In addition, to capture small objects, the output feature maps should have smaller receptive field. The convolution kernel of 1×1 retains the receptive field of the previous layer, only increases the non-linearity of the input patterns, slowing down the receptive field growth of some output features to accurately capture small objects. It is worth noting that we set padding='SAME' for different convolution kernels, so the size of the output feature map has nothing to do with the width and height of the convolution kernel, only the stride. Meanwhile, we also use the same stride to ensure the consistency of the

feature map size, which facilitates subsequent integration.

2) Asymmetric convolution layer: Natural scene data sets are taken from a horizontal perspective, and the object direction has a certain vertical directivity relative to the ground. However, since the remote sensing images are captured from overhead view, the objects will have arbitrary orientations. The literature [41] points out that when the asymmetric convolution kernel operates on the flipped image, it can get the same output as the original image at the axisymmetric position. So we integrate asymmetric convolution to enhance the robustness of the structure to rotational distortions. On the other hand, the asymmetric convolution kernel can provide rectangular receptive fields, which is beneficial to the recognition of long rectangular objects such as ships and vehicles.

We extend the 3×3, 5×5 square convolution kernels in the multi-branch convolution layer by one-dimensional asymmetric convolutions to enrich the feature space. Specifically, we add two parallel branches with the kernel of $1\times d$, $d\times 1$ to the convolution kernel $d\times d$ of the square, and then fuse the output of the three paths, noting that each branch here has applied BatchNormal [42]. Here is the formula expression.

For a convolutional layer, it takes a feature map with C-channel as input and it has D filters. We use K to represent convolution kernel, I to represent the input. F represents the output feature map with D channels. The output feature map channel of the j-th filter in this layer is

$$F_{:,:,j} = \sum_{s=1}^{C} I_{:,:,s} * K^{(j)}_{:,:,s} \tag{1}$$

where $I_{:,:,s}$ is the s-th channel of I, $K^{(j)}_{:,:,s}$ is the s-th channel of $K^{(j)}$, $*$ denotes the convolution operation,.

Like the activation layer, convolutional layer, fully connected layer, and pooling layer, BatchNormal is also a layer of the network. It addresses the internal covariate shift problem by normalizing the input of layer, thus accelerating convergence, avoiding overfitting, and reducing the network's insensitivity to initialization weights. Adding the Batch Normalization layer, the output channel becomes

$$F_{:,:,j} = (\sum_{s=1}^{C} I_{:,:,s} * K^{(j)}_{:,:,s} - \mu_j)\frac{\gamma_j}{\sigma_j} + \beta_j \tag{2}$$

where μ_j is the mean channel value, and σ_j is variance of batch normalization. γ_j and β_j are the learnable scale and shift.

For the j-th convolution kernel, $K^{'(j)}$ represents the fused convolution kernel, $\overline{K}^{(j)}$ and $\hat{K}^{(j)}$ represent the $1\times d$ and $d\times 1$ convolution kernel. b_j represents the bias. we have

$$K^{'(j)} = \frac{\gamma_j}{\sigma_j}K^{(j)} \oplus \frac{\overline{\gamma}_j}{\overline{\sigma}_j}\overline{K}^{(j)} \oplus \frac{\hat{\gamma}_j}{\hat{\sigma}_j}\hat{K}^{(j)} \tag{3}$$

$$b_j = -\frac{\mu_j\gamma_j}{\sigma_j} - \frac{\overline{\mu}_j\overline{\gamma}_j}{\overline{\sigma}_j} - \frac{\hat{\mu}_j\hat{\gamma}_j}{\hat{\sigma}_j} + \beta_j + \overline{\beta}_j + \hat{\beta}_j. \tag{4}$$

So for any filter j, we have

$$F_{:,:,j} + \overline{F}_{:,:,j} + \hat{F}_{:,:,j} = \sum_{s=1}^{C} I_{:,:,s} * K^{'(j)}_{:,:,s} + b_j \tag{5}$$

where $F_{:,:,j}$ denotes $d\times d$ branch, $\overline{F}_{:,:,j}$ denotes $1\times d$ branch, $\hat{F}_{:,:,j}$ denotes $d\times 1$ branch.

3) Dilated convolution layer: Remote sensing image objects are easily confused in cluttered background and need to use the information beyond the object area to help detect, which is often called context information. Considering that CNN's effective field theory [43] points out that the contribution of pixels in the theoretical receptive field is Gaussian distribution. As the Gaussian distribution decays rapidly from the center, the ERF is only part of the theoretical receptive field. In short, the area of the input image that can be seen by convolutional neural network features is not as large as in theory. Therefore, increasing the size of the ERF can help to obtain more global features and more contextual information, which can eliminate the uncertainty or ambiguity of the object and reduce false alarms.

Dilated convolution was originally proposed to solve the problem that downsampling will reduce image resolution and left out information in semantic segmentation tasks [44], [45]. It expands the convolution filter and uses sparse parameters. We leverage dilated convolution to increase ERF significantly to aggregate more contextual information for the classification branch. In addition, it imitated the relationship between the scale and eccentricity of pRFs with the former parts, which is helpful to focus on the area near the center. Therefore, the classification branch can better learn the relationship between the objects and the context information, and further enhance the discernibility and robustness of the features.

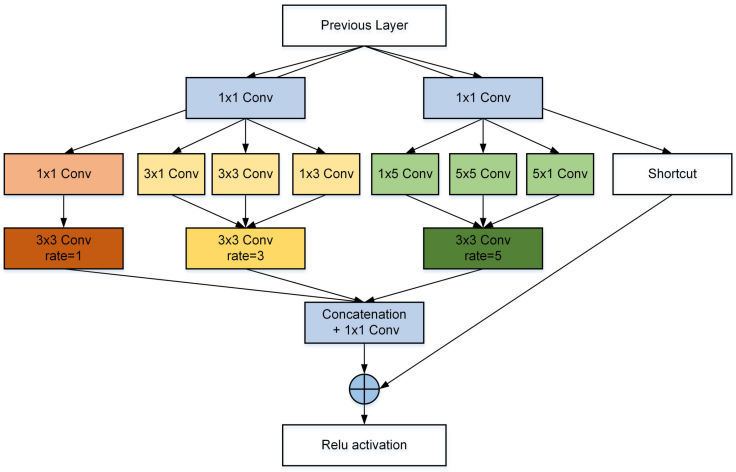


Fig. 3. The architectures of RFAM. Concatenation and 1×1 conv are used to integrate feature maps generated by different branches and $\oplus$ means element-wise addition operation.

C. Localization branch based on PAM

For localization branch, we are specifically inspired by the insightful point [38] that the lower neurons are mainly activated by low-level features such as local texture and edges, while the higher neurons are activated by the whole object or large area of the object. In other words, the shallow features of the network are rich in geometry and location information, which is beneficial to accurate localization of neural network instances. Meanwhile, the features obtained by sharing the basic network may not be enough to locate some objects due to the clutter of

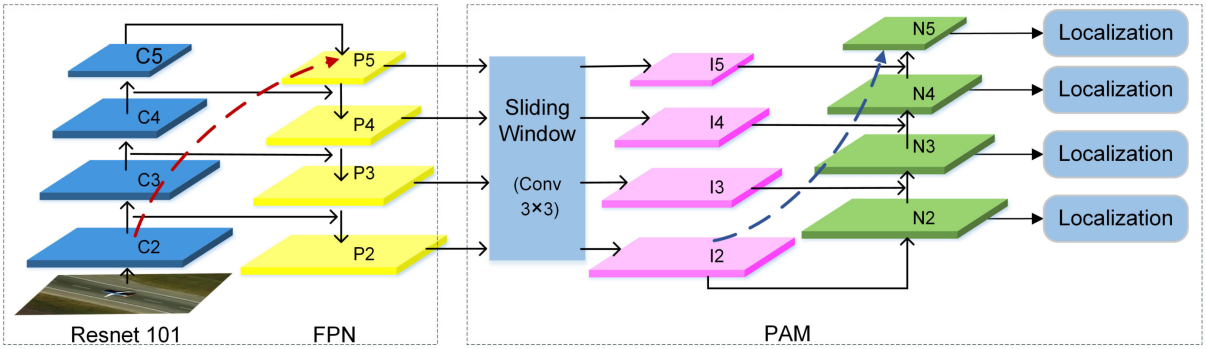


Fig. 4. The architecture of PAM. The dashed red line represents the bottom-up process of FPN, shallow features pass to the top layer through dozens or even more than one hundred network layers. The dashed blue line represents the bottom-up process of PAM, shallow features are passed to the top layer through a network layer with less than 10 layers, which shortens the path of shallow information transmission.

remote sensing image background. Therefore, it is necessary to enhance the shallow information propagation of RPN localization branches, thus improving the accuracy of the object localization.

For the basic network, we use FPN to generate multi-scale features containing strong semantic information. Moreover, feature maps with different resolutions perceive objects of different scales, which is contribute to optimizing the detection of small objects. FPN consists of three parts: the bottom-up path, top-down path, and lateral connection. The first part is the forward propagation of the backbone network. The second part refers to the process of turning deep feature maps with stronger semantic information into higher resolution through upsampling. Finally combining the upsampling result with the bottom-up feature map through lateral connection.

The localization branch of RPN uses the multi-scale feature maps generated by FPN to refine anchor boxes. In the bottom-up process of FPN, shallow features propagate the topmost layer through dozens or even more than one hundred network layers (dashed red line in Fig. 4). It is obvious that after so many layer transfer, the loss of shallow feature information will be more serious. Hence, we extend the localization branch of RPN by adding an additional bottom-up pathway (PAM) based on [35] to shorten the path of shallow information propagation (dashed blue line in Fig. 4). In short, our method further propagates the low-level details to enhance the localization ability of the localization feature hierarchy. The overall structure is shown in Fig. 4. The detail of the lateral connection is depicted in Fig. 5. Next, we will elaborate on the bottom-up path augmentation structure.

We adopt ResNet as the backbone and use $\{C_2, C_3, C_4, C_5\}$ to represent the last feature maps of each stage of ResNet. Note that the stride of $\{C_2, C_3, C_4, C_5\}$ relative to the input image is $\{4, 8, 16, 32\}$. After the top-down processing of FPN, the further expression corresponding to the above feature maps is represented by $\{P_2, P_3, P_4, P_5\}$. Low-level feature $P_l (l \in [2, 4])$ is obtained from high-level features P_{l+1} and corresponding features $C_l (l \in [2, 5])$ through a lateral connection. So the lower layer features are represented as:

$$P_l = f_3^l[f_1^l(C_l \oplus f_u(P_{l+1})], l \in [2, 4] \tag{6}$$

where $\oplus$ denotes element-wise addition. $f_u(\cdot)$ denotes nearest neighbor upsampling, to ensure the same resolution as the corresponding combination feature map C_l. $f_1^l(\cdot)$ denotes 1×1 convolution of l layer, to reduce channel

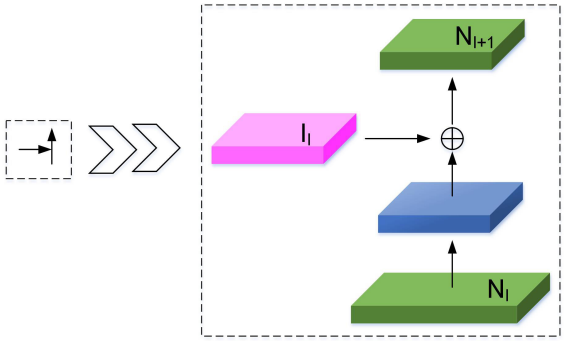


Fig. 5. A block illustrates the lateral connection and the bottom-up path. $\oplus$ means element-wise addition operation.

dimensions. $f_3^l(\cdot)$ denotes the convolution of 3×3 of l layer to eliminate the aliasing effect cauesd by up-sampling. Note that when l = 5, $P_l = f_1^l(C_l)$. $\{I_2, I_3, I_4, I_5\}$ denote the feature maps after 3×3 convolution in the RPN, which can be represented as

$$I_l = f_3^l(P_l), l \in [2, 5]. \tag{7}$$

We use $\{N_2, N_3, N_4, N_5\}$ to denote the feature maps of each layer after bottom-up path augmentation. The low most layer of the bottom-up pathway can be represented as

$$N_l = I_l, l = 5. \tag{8}$$

The other higher layers are added with the lower layers using addition. Hence these layers can be represented as

$$N_l = f_3^l[I_l \oplus f_d^l(N_{l-1})], l \in [2, 5] \tag{9}$$

where $\oplus$ denotes element-wise addition, $f_d^l(\cdot)$ is downsampling function corresponding to l layer which adopts 3×3 convolution with stride = 2.

D. Loss Function

When training our network, objective function follows the multi-task loss of Faster R-CNN. As follows, the form two items represent the classification and localization loss of RPN in the first stage, and the last two items represent the classification and positioning loss in the second stage.

$$\begin{aligned} L = &\frac{\lambda_1}{N_{cls1}} \sum_i L_{cls1}(p_i^{gc_1}, p_i^{pc_1}) + \frac{\lambda_2}{N_{reg1}} \sum_i L_{reg1}(t_i^{gt_1}, t_i^{pt_1}) \\ &+ \frac{\lambda_3}{N_{cls2}} \sum_i L_{cls2}(p_i^{gc_2}, p_i^{pc_2}) + \frac{\lambda_4}{N_{reg2}} \sum_i L_{reg2}(t_i^{gt_2}, t_i^{pt_2}) \end{aligned} \tag{10}$$

where i is the index of an anchor in a mini-batch. $p_i^{gc_1}$ represents two kinds of labels for the anchor. If $p_i^{gc_1}$ = 1 indicates positive label, otherwise $p_i^{gc_1}$= 0 indicates positive label. $p_i^{pc_1}$ is the probability that the anchor box is predicted as the object. $p_i^{gc_2}$ denotes the ground-truth label of the object. $p_i^{pc_2}$ is the probability of the category to which the proposal belongs. $t_i^{gt_1}$ and $t_i^{gt_2}$ denote the offset of the anchor box and the proposal relative to the

ground-truth box, respectively. $t_i^{pt_*}$ represents the offset of the predicted at different stages. L_{cls*} and L_{reg*} are Softmax cross-entropy and smooth L1 loss respectively. λ_* is a hyperparameter to control the balance of various losses.

Algorithm 1 DCL-Net Algorithm.

Input: Image X and its predefined anchor boxes B.

Output: The score of the proposals C, the coordinates of the proposals S.

Feed X and B into the network to extract multi-scale feature maps P=[P_2, P_3, P_4, P_5] based on FPN.
for $i = 2$ to 5 **do**
  Obtain I_i by sliding 3×3 window on P_i.
  Obtain multi-scale receptive field RF_1 ,RF_2 ,RF_2 using convolution kernels of different sizes based on I_i .
  Obtain RF_1', RF_2' ,RF_3' by extending RF_1, RF_2 ,RF_3 based on dilated convolution and Eq. (2,3,4,5) .
  Final receptive field RF is obtain by contacting RF_1', RF_2', and RF_3'.
  R_i is a fusion of identity map of I_i and RF.
  C = $Classification(R_i)$.
  if $i = 2$ **then**
    N_i = I_i.
  else
    Generate a feature map N_i by propagating the information of N_{i-1} based on Eq. (9) .
  end if
  S = $Regression(N_i, B)$.
end for

III. EXPERIMENTS

In this section, we first elaborate on experimental data sets, evaluation metric and implementation details. Then we conducted ablation studies to analyze the different components of our DCL-Net. Finally, we perform comparisons with several other recent object detection methods to further investigate the effectiveness of DCL-Net.

A. Data Sets Description

1) NWPU VHR-10: NWPU VHR-10 is a public data set containing 10 categories (e.g., airplanes, vehicles, ground track fields) for remote sensing object detection. It consists of 800 satellite images, 715 of which are from Google Earth with 0.5-2.0 m spatial resolution, and 85 images are from Vaihingen with 0.08 m spatial resolution. The data set is randomly split training set, verification set and testing set according to the proportion of 6 : 2 : 2.

2) RSOD: RSOD is a public data set for remote sensing object detection, all images are from Google Earth. The data set includes 4993 aircraft in 446 images, 191 playgrounds in 189 images, 1586 oil tanks in 165 images, and 180 overpasses in 176 images. The resolution range of this data set is also wide, from 0.3 to 3 m. This data set is randomly split training, validation and test set according to the proportion of 6 : 2 : 2.

B. Evaluation Metrics

We adopt the extensive used object detection evaluation indicator mean average precision (mAP) as the standard for quantitative evaluation model performance. Set TP, FP, TN and FN as the number of true positives, false positives, true negatives, and false negatives. Recall reflects the proportion of positive samples rightly predicted by the detector to the total positive samples. Precision reflects the proportion of TPs in the positive samples predicted by the detector.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{12}$$

The predicted box is considered as TP if the Intersection Over Union (IOU) between it and the ground truth bounding box exceeds 0.5, otherwise, it is FP. In multi-class object detection, each class can plot a curve according to recall and precision. AP is the area under the curve, and the mAP is the mean of all classes of AP values.

C. Implementation Details

We adopt ResNet-101 pre-trained on the ImageNet data set as the backbone network. We choose C_2, C_3, C_4, C_5 layers to construct the feature pyramid. We set the anchor boxes sizes of each layer as 32^2, 64^2, 128^2, 256^2, 512^2 and the anchor box ratios to $1:2$, $1:1$, $2:1$ for NWPU VHR-10 and RSOD, respectively. During training, we employ MomentumOptimizer as an optimizer. For NWPU VHR-10 data set, the model is trained 30k iterations totally, and the initial learning rate is 1e-3. For RSOD data set, we trained 40k iterations totally, the learning rate is initialized as same as NWPU VHR-10 data set. In the first stage, if IoU$>$0.7, the anchor box is defined as a positive sample and IoU$<$0.3 is assigned a negative sample. Both thresholds are set to 0.5 in the second stage. Our experiments were implemented based on the deep learning framework TensorFlow.

D. Ablation Study

To better understand DCL-Net, we analyze the contributions of each component of the proposed method, including the RFAM and PAM. We first evaluate the baseline on the NWPU data set, then gradually integrate these technologies. Experimental results are summarized in Table I and Fig. 7.

As shown in Table I, the following models are trained in ablation study: baseline setup: FPN as our ablation study baseline; baseline + RFAM: the classification branch of the baseline network with RFAM included only; baseline + PAM: the localization branch of the baseline network with PAM included only; baseline + RFAM + PAM: the complete implementation of the proposed DCL-Net, in which the classification and location branches do not fully share features.

1) Effect of RFAM: As shown in Table I, the baseline + RFAM achieves a 1.05% mAP improvement compared with baseline. For most objects, such as ships, vehicles, harbors, tennis courts, storage tanks, baseball diamonds, bridges, the detection results have improved. Fig. 6 visualizes the classification features corresponding to RFAM (the fourth column) and the shared features corresponding to baseline (the second column). We can see that RFAM

TABLE I

ABLATION STUDY OF DCL-NET ON THE NWPU VHR-10 DATA SET.

method	Ship	Vehicle	Airplane	Bridge	Harbor	Ground track field	Tennis court	Storage tank	Basketball court	Baseball diamond	mAP (%)
Baseline	90.38	89.41	**99.99**	81.90	92.79	99.46	93.14	94.25	89.07	97.25	92.76
Baseline + RFAM	**91.63**	90.69	99.97	91.32	93.52	98.59	93.33	94.49	87.54	97.01	93.81
Baseline + PAM	90.46	90.26	99.84	84.29	**94.19**	**99.75**	94.39	94.31	**90.80**	**98.02**	93.63
Baseline + RFAM + PAM (DCL-Net)	91.05	**91.51**	99.94	**92.02**	94.17	99.24	**95.29**	**94.75**	90.12	97.45	**94.55**

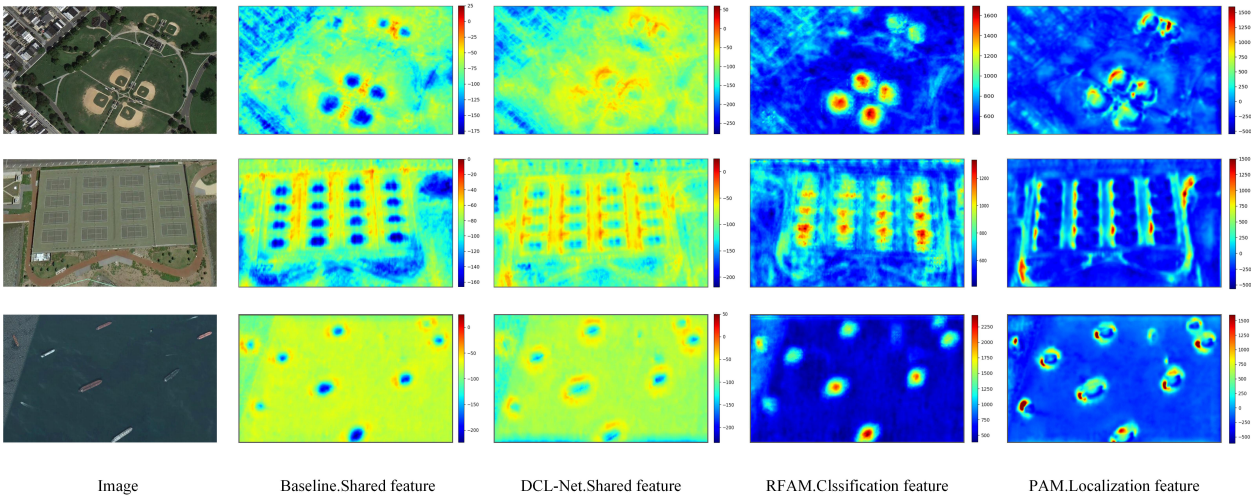


Fig. 6. Visualization of feature maps. To see the objects clearly, we select the low-level (P2) feature maps for visualization. The first column is raw images. The second column is the P2 level feature maps corresponding to the baseline. The third column is the P2 level feature maps corresponding to our DCL-Net. The fourth column and the fifth column correspond to the P2 level feature maps processed by RFAM and PAM respectively.

makes feature activation more hierarchical, where a larger weight is assigned to the smaller convolution kernel near the center, thereby highlighting the importance of the objects center area. This simulates the relationship between the size and eccentricity of RFs in the human receptive field mechanism, thus enhancing the distinguishability of objects features. Moreover, because the remote sensing image is taken from high altitude, the direction of the objects are always arbitrary, but their central area is usually rigid. Therefore, the high activation of the center also ensures the rotation insensitivity of the object features to a certain extent. All of these are helpful to improve the ability of RPN to recognize foreground objects. In addition, most bridges are part of the road, and the appearance of the two kinds of objects is almost the same. Baseline + RFAM is 9.42% higher than baseline for bridges. This is due to the dilated convolution contained in our RFAM expands the effective receptive field and provides more context information, which helps to eliminate the uncertainty of objects. At the same time, the prominence of the object center area also prevents the confusion between the object and the context information, thereby improving the ability of RPN to recognize foreground bridges and background roads.

2) Effect of PAM: As shown in Table I, the baseline + PAM increases mAP from 92.76% to 93.63%, an increase of 0.87%. Among them, the improvement of large objects is obvious, such as the baseball fields and basketball courts.

TABLE II
DETECTION RESULTS ON NWPU VHR-10 DATA SET UNDER DIFFERENT IOU THRESHOLD.

Method	$AP_{50:95}$	AP_{60}	AP_{70}	AP_{80}
Baseline	53.5	88.1	71.4	36.6
DCL-Net	55.2	89.9	73.6	37.3

This owes to the fact that large objects often contain rich semantic information and lack of shallow information such as edges and local textures that are conducive to precise locating. In the localization branch, the information path between the low-level and the high-level is established through adding a bottom-up pathway, which reduces the difficulty of positioning information flow, thereby enhancing the entire feature hierarchy and further improving the frame locating capability. The visualization results in Fig. 6 also prove this point. We can see that the features learned by PAM (the fifth column) are more about the shape and edge features of objects.

3) Effect of RFAM + PAM: The proposed DCL-Net consists of two modules: RFAM and PAM. Experimental results show that both RFAM and PAM can improve the detection performance, but the combined effect is better. Specifically, DCL-Net improves the baseline mAP by 1.79% from 92.76% to 94.55%. The mAP can be improved by about 0.74% and 0.92% compared with RFAM and PAM respectively. Better than all other models in Table I. This is because in remote sensing object detection, the detector usually places classification and localization branches on the same backbone, but the requirements of two branches are different. After the addition of the RFAM and PAM modules, the classification and localization branches do not fully share features, which alleviates the conflict between them. Moreover, RFAM and PAM structures can guide the classification and Localization branches to multi-task optimization, which can promote the two branches learn features that are suitable for themselves, thereby improving the performance of the detector. In addition, since the evaluation metrics of VOC adopts a fixed IOU threshold (0.5), in order to further explore the localization ability of our method, we also use COCO metrics on NWPU dataset, that is, AP is calculated for different IOU thresholds (0.5-0.95, step size is 0.05), and then average. As listed in table II, our AP is higher than the baseline, which indicates that our method can achieve better localization. Especially when the IOU threshold is 0.7, compared with the baseline, our method has the larger gains than the IOU threshold of 0.5, an increase of 2.2%, which further shows that our method can generate more accurate detection boxes. From the visualization results in in Fig. 6, we can see that for the shared feature map of the baseline (the second column), the activation of the object edge is significantly higher than that of the object interior. For the same feature map, the two features constrain each other. That is, when the localization feature (edge feature) is activated high, the activation of internal feature is relatively low. When we reduce the degree of coupling, we can see that the interior and edge of the object are significantly improved compared with the baseline for the shared features of our method (the third column). This shows that our method can better balance classification and localization features. The fourth and fifth columns are the decoupled classification and localization features. The separation of the response areas of the two tasks can be clearly seen, which is more conducive to their respective roles, thereby improving the overall detection performance.

TABLE III
PERFORMANCE EVALUATION OF EIGHT DIFFERENT METHODS ON THE NWPU VHR-10 DATA SET. THE HIGHEST VALUES IN EACH ROW ARE REPRESENTED BY BOLD NUMBERS.

method	Faster R-CNN	SSD300	RetinaNet	R-FCN	Cascade R-CNN	SCRDet	FMSSD	DCL-Net(Our)
Ship	94.39	83.74	90.27	**95.83**	95.65	89.4	89.9	91.05
Vehicle	82.02	38.43	86.89	83.97	85.33	90.1	88.2	**91.51**
Airplane	99.35	90.61	95.78	99.94	99.36	**100.0**	99.7	99.94
Bridge	74.42	**98.18**	97.06	75.89	84.62	74.5	80.1	92.02
Harbor	87.94	88.21	83.86	89.82	87.84	**99.4**	75.6	94.17
Ground track field	97.37	**99.99**	99.76	98.72	99.72	99.2	99.6	99.24
Tennis court	92.03	87.61	90.80	93.55	93.05	83.2	86.0	**95.29**
Storage tank	66.65	77.37	86.30	66.91	66.41	**97. 2**	90.3	94.75
Basketball court	85.25	69.28	90.63	90.65	94.29	87.5	**96.8**	90.12
Baseball diamond	93.37	97.44	93.06	97.67	96.70	97.0	**98.2**	97.45
mAP (%)	87.28	83.09	89.62	89.30	90.30	91.8	90.4	**94.55**

TABLE IV
COMPUTATIONAL COST ON ON NWPU VHR-10 DATA SET.

Method	FPS	FLOPs	Params
Baseline	2.08	259.8M	120.8M
RFAM	1.13	340.3M	153.0M
PAM	1.38	330.5M	149.1M
DCL-Net	0.89	411.1M	181.3M

4) Computational Cost : We conducted ablation studies on the inference speed of our method on NWPU dataset. Table IV summarizes some relevant computational complexity metrics of our proposed method. It is worth noting that all results are tested on a single Nvidia GeForce GTX 1080Ti GPU. Compared with the baseline, we inevitably increase time cost. It can be clearly seen that the inference speed of the baseline FPN is 2.08 FPS, while our method is only 0.89 FPS. In addition, FLOPs (Floating-point Operations Per Second) and params amount increased by 151.3M and 60.5M respectively compared with baseline. Part of the extra time is mainly due to the use of large convolution kernel to obtain a large receptive field. In future work, we will also consider ensuring that fewer parameters are introduced while obtaining a large receptive field to further reduce computational complexity.

E. Peer Methods Comparison

1) Experiments on NWPU VHR-10: To further prove the effectiveness of the proposed DCL-Net, we compare it with the following deep learning detectors: Faster R-CNN [39], SSD [46], RetinaNet [47], FPN, R-FCN [48], Cascade R-CNN [49], SCRDet [21] and FMSSD [50]. It is worth noting that all detectors adopt ResNet-101 pre-trained on the ImageNet data set, except for SSD adopting VGG-16 [51]. Table III shows the results of quantitative

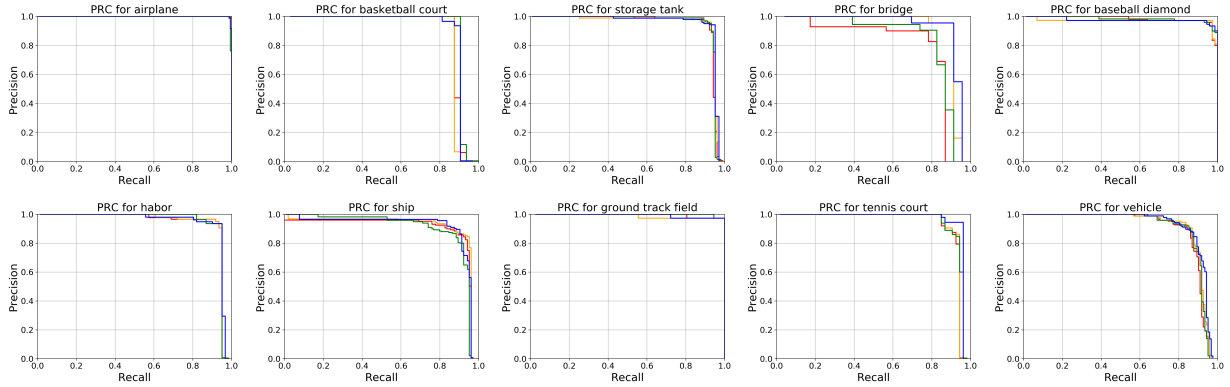


Fig. 7. PRCs of the baseline (red), baseline + RFAM (orange), baseline + PAM (green) and DCL-Net (blue) for NWPU VHR-10.

comparisons. By analyzing the results of different algorithms, it is known that the VGG-16-based SSD network detection accuracy is relatively low, especially the mAP of the vehicle is 38.43%. Compared with the one-stage SSD network, the accuracy of R-FCN has been significantly improved, and its mAP is 89.30%. On the one hand, R-FCN uses a deeper ResNet-101 network to mine more discriminating depth features. On the other hand, R-FCN introduces position-sensitive score map to alleviate the translation invariance contradiction between classification and localization. R-FCN significantly improves the detection performance of ships, vehicles and airplanes, which also shows the effectiveness of balancing the classification and localization tasks in remote sensing detection. Our DCL-Net achieves an improvement of 5.25% compared with R-FCN. In particular, our method for bridge improves AP by 16.13% compared to R-FCN, which is mainly due to the introduction of context information solves the problem of the blurred appearance of remote sensing objects. In addition, compared with other detectors, our detector also achieves the best detection results, with an improvement of 7.27% compared to Faster R-CNN, 11.46% compared to SSD, 4.93% compared to RetinaNet and 4.25% compared to Cascade R-CNN. Besides the object detector mentioned above, we also compared with the recently proposed SCRDet and FMSSD algorithms designed for remote sensing images, which have improved by 4.15% and 2.75% respectively. Moreover, our method shows excellent results for objects with large changes in direction, such as ship, vehicle and airplane, which attributes to our RFAM improving the robustness of rotation. To better show the detection performance, we have selected some examples, and the visualization results are shown in Fig. 8(a), from which we can see that DCL-Net has better detection results.

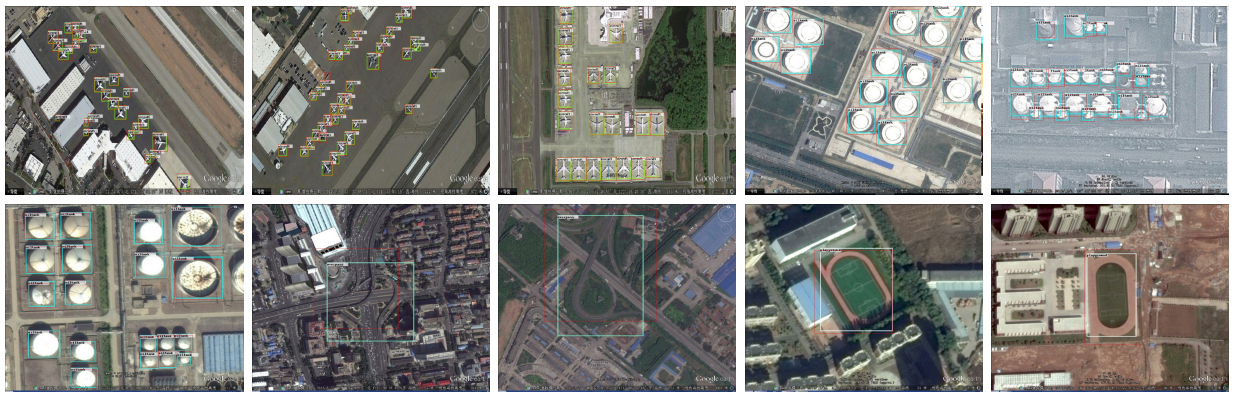
2) Experiments on RSOD: We compare with some detectors on yhe RSOD data set to verify the generality of our DCL-Net. Except for SSD adopting VGG-16, all other methods still use ResNet-101 pre-trained on the ImageNet data set as the backbone. Table V describes the quantitative comparison results. Compared with the baseline FPN, there are improvements in the four categories of the RSOD data set, achieving the higher mAP of 94.61%. Among them, the oil tank and overpass have been significantly improved, increasing by 3.25% and 2.88% respectively. For the oil tank, its shallow features such as texture information are less than other categories. Compared with the baseline, our method not only transmits the shallow information as much as possible but also introduces context information by expanding the ERF to help identify oil tanks and overpasses. In addition, our proposed DCL-Net has

TABLE V

PERFORMANCE EVALUATION OF SENVEN DIFFERENT METHODS ON THE RSOD DATA SET. THE HIGHEST VALUES IN EACH ROW ARE REPRESENTED BY BOLD NUMBERS.

Method	Faster R-CNN	SSD300	RetinaNet	FPN	Cascade R-CNN	NAS-FPN	DCL-Net(Our)
Aircraft	83.54	71.89	80.57	92.94	84.03	89.88	**94.05**
Playground	97.81	98.58	96.97	99.87	99.08	97.88	**99.99**
Oiltank	98.11	90.72	96.69	89.99	**98.80**	92.50	93.24
Overpass	88.62	90.21	90.25	88.27	**92.06**	89.37	91.15
mAP (%)	92.02	87.85	91.19	92.77	93.49	92.41	**94.61**

(a) Prediction of DCL-Net on NWPU VHR-10



(b) Prediction of DCL-Net on RSOD

Fig. 8. Some detection results on NWPU VHR-10 and RSOD test data set. The red box is the ground truth.

the best detection results compared with other methods, with an improvement of 2.59% compared to Faster R-CNN, 6.76% compared to SSD, 3.42% compared to RetinaNet, 1.12% compared to Cascade R-CNN, 2.2% compared to NAS-FPN [52]. This demonstrates the superiority of our method compared with the existing methods. In particular, the playground and aircraft have also achieved the highest results. On the one hand, it benefits from the multi-scale feature prediction of baseline FPN, on the other hand, our method enhances the ability of deep feature location by transmitting shallow information, which facilitate large object detection such as playground. Some visualization

results are demonstrated in Fig. 8(b).

IV. Conclusion

In this paper, we propose a new object detector in remote sensing called DCL-Net, which could alleviate the strong coupling relationship of classification and localization branches. We elaborate on the conflicts between classification and localization branches in remote sensing object detection, and proved that the strong coupling between the two branches is not optimal. To address this problem, we design a decoupling structure to augment the capability of classification and localization of network, where RFAM and PAM are developed and included. For the classification branch, RFAM can properly learn the relationship between the object and the context information, and improve its robustness to rotational distortions, which effectively solve the problems of blurred appearance characteristic and arbitrary direction of remote sensing objects. For the localization branch, PAM enhances the entire feature hierarchy with the rich detailed information of shallow features, which is helpful for localization branch to locate accurately in the remote sensing images with complex background. Experiments indicate that compared with recently proposed deep learning object detection methods, DCL-Net has advanced performance.

In the future, we will further study the relationship between classification branch and localization branch and their respective characteristics, so that the adaptive feature representation can be better in multi-task learning.

References

[1] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1100–1111, 2017.

[2] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 911–923, 2018.

[3] P. Ren, M. Di, H. Song, C. Luo, and C. Grecos, "Dual smoothing for marine oil spill segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 82–86, 2015.

[4] J. Peng, Y. Zhou, W. Sun, Q. Du, and L. Xia, "Self-paced nonnegative matrix factorization for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[5] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[6] Z. Yang, Y. Liu, L. Liu, X. Tang, J. Xie, and X. Gao, "Detecting small objects in urban settings using slimnet model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8445–8457, 2019.

[7] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware convolution neural network for ship detection in surveillance video," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[8] M. ElMikaty and T. Stathaki, "Detection of cars in high-resolution aerial images of complex urban environments," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5913–5924, 2017.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[11] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[12] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proceedings of the Springer Asian Conference on Computer Vision*, 2018, pp. 150–165.

[13] W. Guo, W. Yang, H. Zhang, and G. Hua, "Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network," *Remote Sensing*, vol. 10, no. 1, p. 131, 2018.

[14] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7147–7161, 2018.

[15] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 3–22, 2018.

[16] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2337–2348, 2017.

[17] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *arXiv preprint arXiv:1903.00857*, 2019.

[18] Y. Gong, Z. Xiao, X. Tan, H. Sui, C. Xu, H. Duan, and D. Li, "Context-Aware convolutional neural network for object detection in vhr remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 34–44, 2019.

[19] J. Zhang, C. Xie, X. Xu, Z. Shi, and B. Pan, "A contextual bidirectional enhancement method for remote sensing image object detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4518–4531, 2020.

[20] C. Tao, L. Mi, Y. Li, J. Qi, Y. Xiao, and J. Zhang, "Scene context-driven vehicle detection in high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7339–7351, 2019.

[21] 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 International Conference on Computer Vision*, 2019, pp. 8232–8241.

[22] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based cnn for ship detection," in *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2017, pp. 900–904.

[23] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for detecting oriented objects in aerial images," *arXiv preprint arXiv:1812.00155*, 2018.

[24] J. Koo, J. Seo, S. Jeon, J. Choe, and T. Jeon, "RBox-CNN: rotated bounding box based cnn for ship detection in remote sensing image," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018, pp. 420–423.

[25] W. Liu, L. Ma, and H. Chen, "Arbitrary-oriented ship detection framework in optical remote-sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 6, pp. 937–941, 2018.

[26] X. Ying, Q. Wang, X. Li, M. Yu, H. Jiang, J. Gao, Z. Liu, and R. Yu, "Multi-attention object detection model in remote sensing images based on multi-scale," *IEEE Access*, vol. 7, pp. 94 508–94 519, 2019.

[27] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in vhr remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 310–314, 2018.

[28] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, 2019.

[29] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, "Feature-attentioned object detection in remote sensing imagery," in *Proceedings of the IEEE International Conference on Image Processing*, 2019, pp. 3886–3890.

[30] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8983–8997, 2019.

[31] R. Zhu, S. Zhang, X. Wang, L. Wen, H. Shi, L. Bo, and T. Mei, "ScratchDet: Training single-shot object detectors from scratch," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2268–2277.

[32] Z. Zhang, X. Chen, J. Lie, and K. Zhou, "Rotated feature network for multi-orientation object detection," *arXiv preprint arXiv:1903.09839*, 2019.

[33] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting RCNN: On awakening the classification power of faster rcnn," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 453–468.

[34] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5909–5918.

[35] 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.

[36] 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.

[37] B. A. Wandell and J. Winawer, "Computational neuroimaging and population receptive fields," *Trends in Cognitive Sciences*, vol. 19, no. 6, pp. 349–357, 2015.

[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the Springer Uropean Conference on Computer Vision*, 2014, pp. 818–833.

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[40] S. Liu, D. Huang *et al.*, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 385–400.

[41] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1911–1920.

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[43] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 4898–4906.

[44] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[45] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo, and X. Lan, "DSSNet: A simple dilated semantic segmentation network for hyperspectral imagery classification," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[46] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the Springer European Conference on Computer Vision*, 2016, pp. 21–37.

[47] 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.

[48] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.

[49] Z. Cai and N. Vasconcelos, "Cascade R-cnn: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[50] P. Wang, X. Sun, W. Diao, and K. Fu, "Fmssd: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3377–3390, 2019.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[52] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7036–7045.