remote sensing

MDPI

Article

Mask OBB: A Semantic Attention-Based Mask Oriented Bounding Box Representation for Multi-Category Object Detection in Aerial Images

Jinwang Wang [1,†], Jian Ding [2,†], Haowen Guo [1], Wensheng Cheng [1], Ting Pan [1] and Wen Yang [1,2,*]

[1] School of Electronic Information, Wuhan University, Wuhan 430072, China; jwwangchn@whu.edu.cn (J.W.); ghw@whu.edu.cn (H.G.); cwsinwhu@whu.edu.cn (W.C.); ting.pan@whu.edu.cn (T.P.)
[2] State Key Lab. LIESMARS, Wuhan University, Wuhan 430072, China; jian.ding@whu.edu.cn
* Correspondence: yangwen@whu.edu.cn; Tel.: +86-27-68754367
† These authors contributed equally to this work.

Received: 4 November 2019; Accepted: 3 December 2019; Published: 6 December 2019

check for updates

Abstract: Object detection in aerial images is a fundamental yet challenging task in remote sensing field. As most objects in aerial images are in arbitrary orientations, oriented bounding boxes (OBBs) have a great superiority compared with traditional horizontal bounding boxes (HBBs). However, the regression-based OBB detection methods always suffer from ambiguity in the definition of learning targets, which will decrease the detection accuracy. In this paper, we provide a comprehensive analysis of OBB representations and cast the OBB regression as a pixel-level classification problem, which can largely eliminate the ambiguity. The predicted masks are subsequently used to generate OBBs. To handle huge scale changes of objects in aerial images, an Inception Lateral Connection Network (ILCN) is utilized to enhance the Feature Pyramid Network (FPN). Furthermore, a Semantic Attention Network (SAN) is adopted to provide the semantic feature, which can help distinguish the object of interest from the cluttered background effectively. Empirical studies show that the entire method is simple yet efficient. Experimental results on two widely used datasets, i.e., DOTA and HRSC2016, demonstrate that the proposed method outperforms state-of-the-art methods.

Keywords: oriented object detection; aerial images; convolutional neural networks

1. Introduction

It is a fundamental problem in Earth Vision to achieve accurate and robust object detection in aerial images, which is very challenging due to four issues:

- arbitrary orientations: unlike natural images in which objects are generally oriented upward, objects in aerial images often appear with arbitrary orientations since aerial images are typically taken with a bird's-eye view [1,2].
- densely packed objects: it is hard to separate small crowded objects like vehicles in parking lots [3].
- huge scale variations: scale changes of objects in aerial images captured with various platforms and sensors are usually huge [2,4].
- cluttered background: the background in aerial images is cluttered and normally contains a large number of uninteresting objects [5].

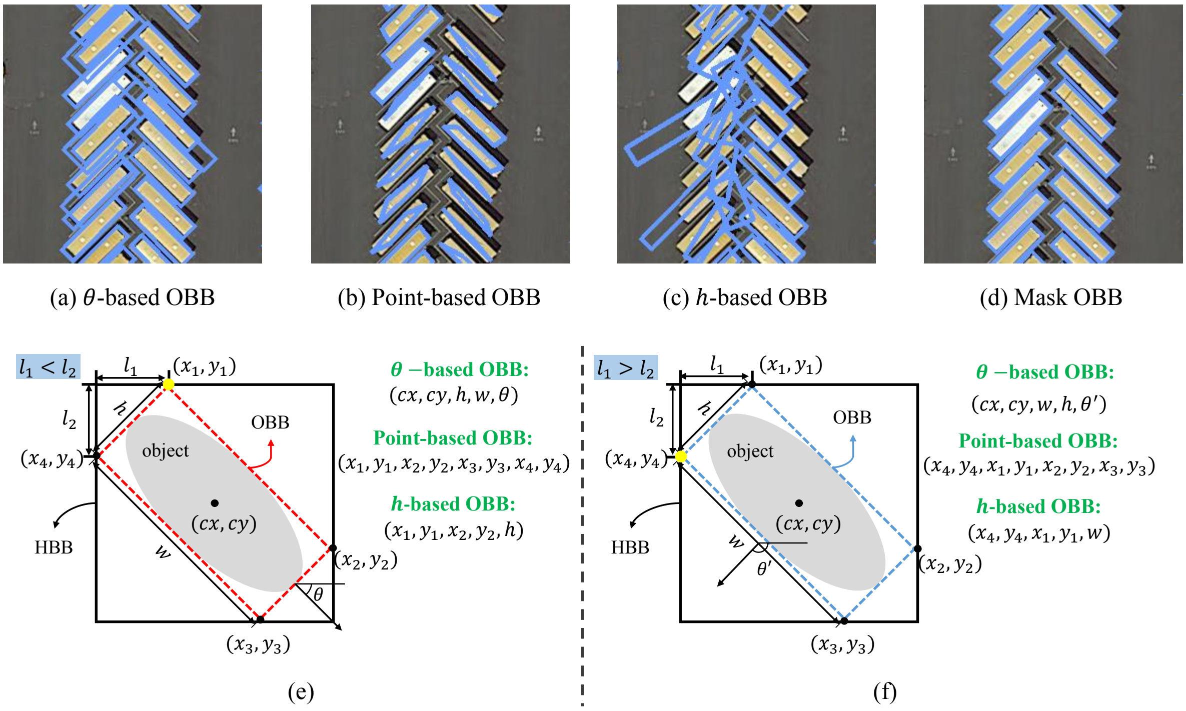
To tackle these issues, we need a robust object detection method for aerial images which is resilient to the aforementioned appearance variations.

With the development of deep learning technology, modern generic object detection methods based on a horizontal bounding box (HBB) have achieved great success in natural scenes. They can be organized into two main categories: two-stage and single-stage detectors. Two-stage detectors are firstly introduced by a Region-based Convolutional Neural Network (R-CNN) [6]. R-CNN generates object proposals by Selective Search [7], then classifies and refines the proposal regions by a Convolutional Neural Network (CNN). To eliminate the duplicated computation in the R-CNN, Fast R-CNN [8] extracts the feature of the whole image once, then generates region features through Region of Interest (RoI) Pooling. Faster R-CNN [9] introduces a Region Proposal Network (RPN) to generate the region proposals efficiently. Some researchers further extend the work of Faster R-CNN for better performance, like Region-based Fully Convolutional Network (R-FCN) [10], Deformable R-FCN [11], Light Head R-CNN [12], Scale Normalization for Image Pyramids (SNIP) [13], SNIP with Efficient Resampling (SNIPER) [14], etc. Unlike two-stage detectors, single-stage detectors directly estimate class probabilities and bounding box offsets with a single CNN like You Only Look Once (YOLO) [15–17], Single Shot Multibox Detector (SSD) [18] and RetinaNet [19]. Compared with two-stage detectors, one-stage detectors are much simpler and more efficient, because there is no need to produce region proposals.

Similarly, object detection methods based on HBB are widely used for object detection in aerial images. Han et al. [20] propose R-P-Faster R-CNN for detecting small objects in aerial images. Xu et al. [21] use Deformable Convolutional Network (DCN) [11] to address geometric modeling in aerial image object detection and propose a Ratio Constrained Non Maximum Suppression (arcNMS) to reduce the increase of false region proposals. Guo et al. [22] propose a multi-scale CNN and multi-scale object proposal network for geospatial object detection in high resolution satellite images. Li et al. [23] propose a hierarchical selective filtering network (HSF-Net) to detect ships with various scales efficiently. Pang et al. [24] design a Tiny-Net backbone and a global attention block to detect tiny objects in large-scale aerial images. Dong et al. [25] propose the Sig-NMS to replace traditional NMS for improving the detection accuracy of small objects. These methods greatly promote the development of object detection in the remote sensing field.

However, the HBBs are not suitable for describing objects in aerial images since the objects in aerial images are often of arbitrary orientations. To deal with this challenge, instead of using HBB, some datasets [2,26–29] use oriented bounding boxes (OBBs) to annotate objects in aerial images. OBBs can not only compactly enclose oriented objects, but also retain the orientation information which is very useful for further processing. Many works [1,2,30–32] handle this problem as a regression task and directly regress oriented bounding boxes. We call them regression-based methods. For instance, DRBox [33] redesigns the SSD [18] to regress oriented bounding boxes by multi-angle prior oriented bounding boxes. Xia et al. [2] propose the FR-O which regresses the offsets of OBBs relative to HBBs. ICN [30] joints image cascade and feature pyramid network to extract features for regressing the offsets of OBBs relative to HBBs. Ma et al. [34] design a Rotation Region Proposal Network (RRPN) to generate prior proposals with the object orientation information, and then regress the offsets of OBBs relative to oriented proposals. R-DFPN [35] adopts RRPN and puts forward the Dense Feature Pyramid Network to solve the narrow width problems of ships. Ding et al. [1] design a rotated RoI learner to transform horizontal RoIs to oriented RoIs by a supervised method. All these regression-based methods can be summarized as the problem of regression for the offsets of OBBs relative to HBBs or OBBs, and they rely on the accurate representation of OBB.

Nevertheless, regression-based methods encounter the problem of ambiguity in the regression target. For example, OBB represented as $\{(x_i, y_i)|i = 1, 2, 3, 4\}$ (point-based OBB) has 4 different representation ways if we just change the order of vertexes. In order to get the uniqueness of OBB representation, some tricks, such as defining the first vertex by a certain rule [2], are used to leverage the ambiguity problem. The ambiguity is still unsolved, because two similar region features may have obvious different OBB representations. For instance, Figure 1e,f show the ambiguity problem of regression-based OBB intuitively. When the angle of OBB is near $\pi/4$ or $3\pi/4$, the ambiguity of adjacent points is the most serious. Specifically, by the definition in [2], OBB in Figure 1e can be represented as $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ (point-based OBB), but in Figure 1f which is very similar with Figure 1e, the OBB needs to be represented as $(x_2, y_2, x_3, y_3, x_4, y_4, x_1, y_1)$ (point-based OBB). Although OBBs (θ-based OBBs, point-based OBBs, and h-based OBBs) in Figure 1e,f are completely different, but they have similar feature map. Due to the ambiguity, the training is hard to converge, and the mAP of the HBB task is often much higher than OBB task even with the same backbone network. In this paper, we give an experimental analysis of different regression-based methods. Figure 1a–c are the visualization results of different regression-based OBB representations, and we can see that the detection results are terrible when the angles of objects are near $\pi/4$ or $3\pi/4$.


(a) θ-based OBB (b) Point-based OBB (c) h-based OBB (d) Mask OBB

(e) (f)

Figure 1. (**a**–**c**) Failure modes of regression-based OBB representations. Specifically, (**a**) is the result from θ-based OBB representation (cx, cy, h, w, θ), (**b**) is the result from point-based OBB representation $\{(x_i, y_i)|i = 1, 2, 3, 4\}$, and (**c**) is the result from h-based OBB representation (x_1, y_1, x_2, y_2, h). (**d**) Result from Mask OBB representation. (**e**,**f**) Borderline states of regression-based OBB representations. The full line, dashed line and gray region represent horizontal bounding box, oriented bounding box and oriented object. The feature map of the left instance should be very similar to the right one, but by the definition in [2] to choose the first vertex (yellow vertex of OBB in (**e**,**f**)), the coordinates of θ-based OBB, point-based OBB and h-based OBB representations differ greatly. The ambiguity makes the learning unstable and explains the results in (**a**–**c**). The representation of Mask OBB can avoid the problem of ambiguity and obtain better detection results.

In order to eliminate the ambiguity, we represent an object region as a binary segmentation map and treat the problem of detecting an oriented object as pixel-level classification for each proposal in this paper. Then, the oriented bounding boxes are generated from the predicted masks by post-processing, and we call this kind of OBB representation a mask-oriented bounding box representation (Mask OBB). By using Mask OBB, the convergence is faster and the gap of mAP between HBB task and OBB task is greatly reduced while compared with regression-based methods. As shown in Figure 1d, detection results of Mask OBB are better than regression-based OBB representation methods. Despite its relevance, segmentation-based oriented object detection methods in remote sensing have been poorly exploited when compared with regression-based methods. There are just some segmentation-based methods in the field of oriented text detection. For instance, Ref [36] presents

Fused Text Segmentation Networks to detect and segment the text instance simultaneously. Ref [37] finishes the detection and recognition task on mask branch by predicting word segmentation maps. These segmentation-based methods are restricted to single-category (text) object detection while there are many different categories to discern for aerial images, such as the dataset DOTA [2]. Our proposed segmentation-based method Mask OBB can handle multi-category-oriented object detection in aerial images. It is based on an instance segmentation framework Mask R-CNN [38] which is proposed by adding a mask branch on Faster R-CNN to obtain pixel-level segmentation predictions. To the best of our knowledge, this work is the first multi-category segmentation-based oriented object detection method in the remote sensing field.

Besides the arbitrary orientation problem, huge scale changes of objects in aerial images is also a challenging problem. Some works [30,32,35] use a Feature Pyramid Network (FPN) [39] to handle the scale problem by fusing low-level and high-level features. In this paper, we design an Inception Lateral Connection Network (ILCN) to further enhance the FPN for solving the scale change problem. Unlike the original FPN, we use the inception structure [40–43] instead of one 1 × 1 convolutional layer as the lateral connection. In the ILCN, besides the original 1 × 1 convolutional layer, three additional layers with different receptive fields are added. We call this enhanced FPN an Inception Lateral Connection Feature Pyramid Network (ILC-FPN). Experimental results show ILC-FPN can handle large-scale variations in aerial images efficiently.

In addition, in aerial images, the background is cluttered and normally contains a large number of uninteresting objects. For distinguishing interesting objects from cluttered background, attention mechanism which is proven to be promising in many vision applications, such as image classification [44–46] and general object detection [47,48] is used in some aerial image object detection works [49–51]. Specifically, inspired by [45,49,52] proposes a Feature Attention FPN (FA-FPN) which contains channel-wise attention and pixel-wise attention to effectively capture the foreground information and restrain the background in aerial images. CAD-Net [50] designs a spatial-and-scale-aware attention module to guide the network to focus on more informative regions and features as well as more appropriate feature scales in aerial images. Chen et al. [51] proposes a multi-scale spatial and channel-wise attention (MSCA) mechanism to make the network pay more attention to object in aerial images as human vision. Unlike the aforementioned attention modules which are unsupervised, we use the semantic segmentation map converted from oriented bounding boxes as the target of semantic segmentation network and design a Semantic Attention Network (SAN) to learn semantic features for predicting HBBs and OBBs efficiently.

Overall, our complete model can achieve 75.33% and 76.98% mAP on OBB task and HBB task of DOTA dataset, respectively. At the same time, it achieves 96.70% mAP on OBB task of HRSC2016 dataset. The main contributions of this paper can be summarized as follows:

- We address the influence of ambiguity of regression-based OBB representation methods for oriented bounding box detection, and propose a mask-oriented bounding box representation (Mask OBB). As far as we know, we are the first to treat the multi-category oriented object detection in aerial images as a problem of pixel-level classification. Extensive experiments demonstrate its state-of-the-art performance on both DOTA and HRSC2016 datasets.
- We propose an Inception Lateral Connection Feature Pyramid Network (ILC-FPN), which can provide better features to handle huge scale changes of objects in aerial images.
- We design a Semantic Attention Network (SAN) to distinguish interesting objects from cluttered background by providing semantic features when predicting HBBs and OBBs.

The rest of the paper is organized as follows: Section 2 presents the proposed method, including Mask OBB, ILC-FPN, SAN, and overall network architecture. Then we give the experimental results in Section 3. Finally, we discuss the model settings in Section 4 and draw the conclusions in Section 5.

2. Methodology

2.1. Regression-Based OBB Representations

In order to detect OBBs, the definition of OBBs is essential. There are several formats to represent the OBBs. Ref. [1,34] use $\{(cx, cy, h, w, \theta)\}$ (θ-based OBB) to represent OBB, where (cx, cy) is the center coordinate of OBB, h, w, θ are the height, width and angle of OBB, respectively. Ref. [2,30,31] use $\{(x_i, y_i)|i = 1, 2, 3, 4\}$ (point-based OBB) to represent OBB, where (x_i, y_i) is the i-th vertex of OBB. Ref. [53] uses $\{(x_1, y_1, x_2, y_2, h)\}$ (h-based OBB) to represent OBB, where (x_1, y_1) and (x_2, y_2) are first vertex and second vertex of OBB, and h is the height of OBB. We call these formats regression-based OBB representations. Figure 1e,f demonstrate these formats.

Although these representations ensure the uniqueness in the OBB's definition with some rules, there still allow extreme conditions. In these conditions, a tiny change of OBB angle would result in a large change of OBB representation. We denote angle values in these conditions as discontinuity points. For oriented object detectors, similar features extracted by the detector with close positions are supposed to generate similar position representations. However, OBB representations of these similar features would differ greatly near discontinuity points. This would force the detector to learn totally different position representations for similar features. It would impede the detector training process and deteriorate detector's performance obviously.

Specifically, for point-based OBB representation, to ensure the uniqueness of OBB definition, Xia et al. [2] choose the vertex closest to the "top left" vertex of the corresponding HBB as the first vertex. Then the other vertexes are fixed in clockwise order, so we get the unique representation of OBB. Nevertheless, this mode still allows discontinuity point, as illustrated in Figure 1e,f. When the l_1 on the horizontal bounding box is shorter than the l_2, the OBB is represented with $R_1 = (x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ (point-based OBB), as Figure 1e shows. Otherwise, the OBB is represented with $R_2 = (x_4, y_4, x_1, y_1, x_2, y_2, x_3, y_3)$ (point-based OBB), as Figure 1f shows. When the length of l_1 increases with θ, till θ approaches and surpasses $\pi/4$, the OBB representation would jump from R_1 to R_2, and vice versa. Hence $\pi/4$ is a discontinuity point in this mode.

For h-based OBB and θ-based OBB, $\pi/4$ is still a discontinuity point. As shown in Figure 1 (e) and (f), with θ oscillating near $\pi/4$, the h-based OBB representation would switch between (x_1, y_1, x_2, y_2, h) and (x_4, y_4, x_1, y_1, w). The θ-based OBB representation would switch back and forth between (cx, cy, h, w, θ) and (cx, cy, w, h, θ') similarly.

2.2. Mask OBB Representation

For handling the ambiguity problem, we represent oriented object as binary segmentation map which ensures the uniqueness naturally, and the problem of detecting OBB can be treated as pixel-level classification for each proposal. Then the OBBs are generated from the predicted masks by post-processing, and we call this kind of OBB representation as mask-oriented bounding box representation (Mask OBB). Under this representation, there is no discontinuity point and ambiguity problem.

Furthermore, aerial image datasets like DOTA [2] and HRSC2016 [26] give the regression-based oriented bounding boxes as ground truth. Specifically, DOTA and HRSC2016 use point-based OBBs $\{(x_i, y_i)|i = 1, 2, 3, 4\}$ [2] and θ-based OBBs $\{(cx, cy, h, w, \theta)\}$ [26] as the ground truth, respectively. However, for pixel-level classification, pixel-level annotations are essential. In order to handle this problem, pixel-level annotations are converted from original OBB ground truths. Specifically, pixels inside OBBs are labeled as positive and pixels outside are labeled as negative, and then, we obtain the pixel-level annotations which will be treated as the pixel-level classification ground truth. Figure 2 illustrates the point-based OBBs and converted Mask OBBs on DOTA images. The highlight points are original ground truth, and the highlight regions inside point-based OBBs are new ground truth for pixel-level classification, which is well known as instance segmentation problem. Unlike point-based OBB, h-based OBB and θ-based OBB, Mask OBB is unique in the definition no matter how point-based

OBB changes. Using Mask OBB, the problem of ground truth ambiguity can be solved in nature, and there is no discontinuity points allowed in this mode.

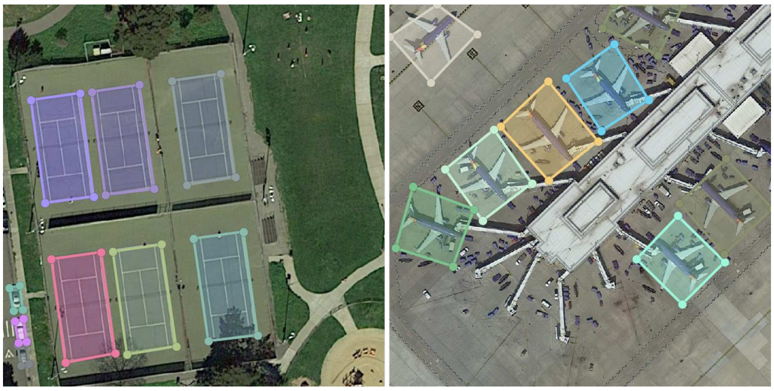


Figure 2. Samples for illustrating mask-oriented bounding box representation (Mask OBB). The highlight points are original ground truth (point-based OBB), and the highlight regions inside point-based OBBs are ground truth for pixel-level classification.

In the inference stage, the predicted Mask OBB needs to be converted to point-based OBB and θ-based OBB for performance evaluation on DOTA and HRSC2016 dataset, respectively. We calculate the minimum area oriented bounding box of predicted segmentation map by Topological Structural Analysis Algorithm [54]. Minimum area oriented bounding box has the same representation as θ-based OBB, which can be directly used by HRSC2016 dataset for calculating mAP. For DOTA, the four vertexes of minimum area oriented bounding box can be used for evaluating performance.

2.3. Overall Pipeline

The overall architecture of our method is illustrated in Figure 3. Our network is a two-stage method based on Mask R-CNN [55], which is known as an instance segmentation framework. In the first stage, a number of region proposals are generated by a Region Proposal Network (RPN) [9]. In the second stage, after the RoI Align [55] operation for each proposal, aligned features extracted from ILC-FPN features and SAN's semantic features are fed into the HBB branch and OBB branch to generate the HBBs and instance masks. Finally, the OBBs are obtained by OBB branch based on predicted instance masks.

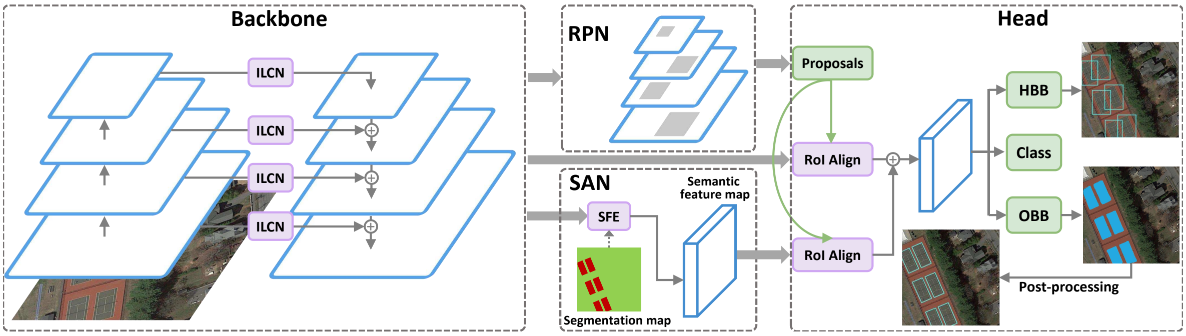


Figure 3. Overview of the pipeline for our method. ILCN is the Inception Lateral Connection Network. RPN is the Region Proposal Network. SAN is the Semantic Attention Network, and SFE is the Semantic Feature Extraction module in SAN. Horizontal bounding boxes and oriented bounding boxes are generated by the HBB and OBB branches, respectively.

In this work, we apply the ILC-FPN which will be detailed in Section 2.4 with ResNet [56] as backbone. Each level of the pyramid can be used for detecting objects at a different scale. We denote the output as $\{C_2, C_3, C_4, C_5\}$ for conv2, conv3, conv4, and conv5 of ResNet, and call the final feature map set of ILC-FPN as $\{P_2, P_3, P_4, P_5, P_6\}$. Note that $\{P_2, P_3, P_4, P_5, P_6\}$ have strides of $\{4, 8, 16, 32, 64\}$ pixels with respect to the input image.

Region Proposal Network (RPN) [9] is used to generate region proposals for the second stage on the outputs of ILC-FPN. Following [39], we assign anchors of a single scale to each level. Specifically, we set five anchors with areas of $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ pixels on five levels $\{P_2, P_3, P_4, P_5, P_6\}$, respectively. Different anchor aspect ratios $\{1{:}2,\ 1{:}1,\ 2{:}1\}$ are also adopted at each level. Thus, in total, there are 15 anchors over the pyramid. Note that no special design for objects in aerial images is adopted in RPN.

RoI Align [55] is adapted to extract the region features of the proposals produced by RPN from the outputs of ILC-FPN and SAN. Compared with RoI Pooling [9], RoI Align preserves more accurate location information, which is quite beneficial for the segmentation task in the OBB branch. Through RoI Align, all proposals are resized to 7 × 7 for HBB branch and 14 × 14 for OBB branch. HBB branch aims to regress HBB and classify objects. OBB branch predicts a 28 × 28 mask from each proposal by four convolutional layers and one deconvolutional layer. In training stage, OBB branch just generates masks to calculate OBB branch loss, and in the inference stage, an OBB branch generates OBB set $\{(cx, cy, h, w, \theta)\}$ based on predicted masks by post-processing.

2.4. Inception Lateral Connection Feature Pyramid Network

Objects in aerial images are very complicated, and relative scales vary greatly between different categories. For example, the size of ground track field is about 1500 times the size of small vehicle in DOTA. Even for the same category object, such as ship, the sizes are range from about 10 × 10 to 400 × 400 pixels in the different images. The huge scale gap leads to poor performance of normal object detection methods. In order to handle this, we need to use a strong feature extraction network to enhance the backbone.

In convolutional neural network, low-level features lack semantic information, but have the accurate location information. On the contrary, high-level features have rich semantic information but relatively rough location information. Making full use of low-level features and high-level features can handle the scale problem to a certain extent. FPN [39] is an effective method to fuse low-level and high-level features via the top-down pathway and lateral connection. However, original FPN [39] simply utilizes one 1 × 1 convolutional neural network as lateral connection to fuse features C_i and P_{i+1}. This fusion strategy can not effectively handle the features of very large objects.

In order to solve this problem, we design an Inception Lateral Connection Network (ILCN), which uses inception structure to enhance feature propagation. Figure 4 shows the architecture of ILCN. Based on ILCN, we design an enhanced FPN, and we call it Inception Lateral Connection Feature Pyramid Network (ILC-FPN). Unlike the original FPN which uses one 1 × 1 convolutional layer as the lateral connection, we use inception structure [42] as the lateral connection. Besides the original 1 × 1 convolutional layer, three additional layers are added in the lateral connection. As shown in Figure 4, these extra layers include the 5 × 5 convolutional layer, the 3 × 3 convolutional layer, and the max pooling layer. The output of the new lateral connection is a concatenation of these layers' outputs. Thanks to the different convolution kernel sizes in the ILC-FPN, the same level feature in FPN can better handle large scale variations. Through the experimental results, we can observe that the use of ILC-FPN can significantly improve the detection performance due to a better feature fusion strategy.

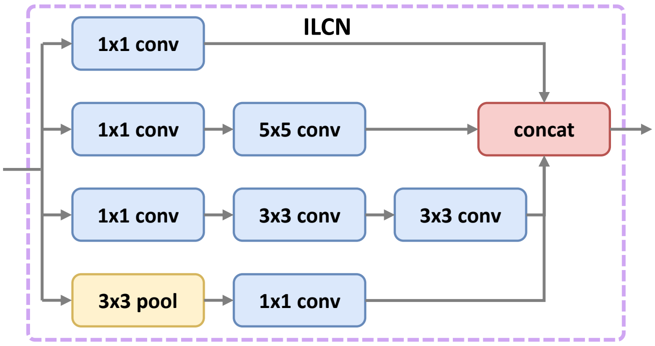


Figure 4. Illustration of the Inception Lateral Connection Network (ILCN). ILCN is the modified inception blocks as the lateral connection in ILC-FPN to better fuse features for enhancing the original FPN.

2.5. Semantic Attention Network

To further help model extract the interesting objects from the cluttered background in aerial images. We design a Semantic Attention Network (SAN) to extract semantic feature of the whole image. Note that, RPN, the Semantic Attention Network, OBB branch and HBB branch are jointly trained end-to-end.

In the Semantic Attention Network, the feature extraction module is the key which is called a Semantic Feature Extraction (SFE) module. To generate the semantic feature from the outputs of ILC-FPN, we design a simple architecture to incorporate higher-level feature maps with context information and lower-level feature maps with location information for better feature representation. Figure 5 illustrates the details of SFE. For the output of ILC-FPN, low-level feature maps are upsampled and high-level feature maps are downsampled to the same spatial scale. The result is a set of feature maps at the same scale, which are then element-wise summed. Four 3 × 3 convolutions are then used to obtain the output of SFE, and one 1 × 1 convolution is used on the SFE's output to generate class label map and semantic feature map.

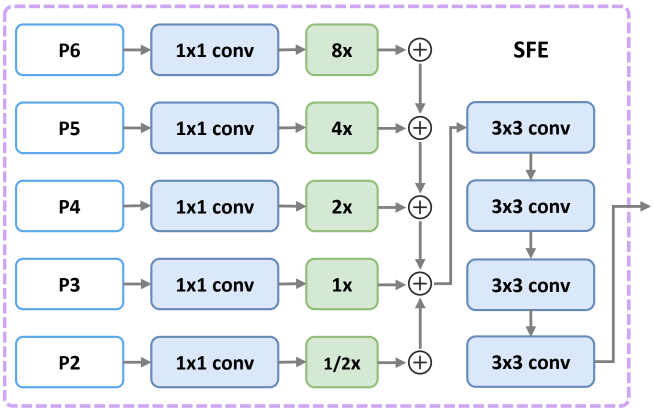


Figure 5. Illustration of the Semantic Feature Extraction (SFE) module. SFE is the feature extraction module in Semantic Attention Network (SAN). SFE simply upsamples and downsamples the outputs of ILC-FPN to provide features for predicting HBBs and OBBs.

The class label map is used to calculate the semantic segmentation loss, and the semantic feature map is used to be fused with an HBB branch feature map and OBB branch feature map. Specifically, given a proposal from RPN, we use RoI Align to extract a feature patch (e.g., 7 × 7 for HBB branch and 14 × 14 for OBB branch) from the corresponding level of ILC-FPN outputs as the region feature. At the same time, we also apply RoI Align on the semantic feature map to obtain other feature patch of the same shape as the region feature, and then combine the features from both branches by element-wise sum, and the result will be treated as the new region feature for HBB branch and OBB branch. Experimental results demonstrate that SAN can improve the detection performance on OBB task and HBB task.

In addition, aerial image datasets have no semantic segmentation ground truth. For calculating the semantic segmentation loss, we generate the semantic segmentation ground truth by the OBB ground truth. Specifically, pixels inside OBB are labeled as certain class of OBB and pixels outside are labeled as background in the whole image. Figure 6 demonstrates the OBB ground truth and corresponding semantic segmentation ground truth.

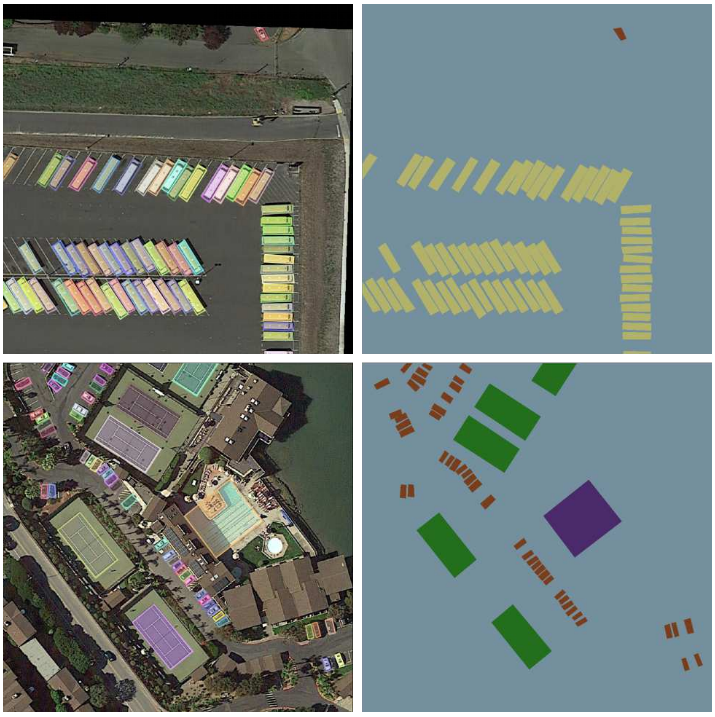


Figure 6. OBB ground truth (**left**) and the corresponding semantic segmentation ground truth (**right**). Different colors in semantic segmentation ground truth mean different categories.

2.6. Multi-Task Learning

Multi-task learning has been proved to benefit the performance of multiple tasks [55]. It enables the network to learn HBBs, OBBs and semantic segmentation at the same time from the same input image. The overall loss function takes the form of a multi-task learning:

$$L = L_{RPN} + \alpha_1 L_{CLS} + \alpha_2 L_{HBB} + \alpha_3 L_{OBB} + \alpha_4 L_{SEG}, \tag{1}$$

where L_{RPN} is region proposal network loss, L_{CLS} is classification loss, L_{HBB} is the horizontal bounding box loss, L_{OBB} is the oriented bounding box loss, and L_{SEG} is the semantic segmentation loss. Specifically, L_{RPN}, L_{CLS} and L_{HBB} are the same as Faster R-CNN [9], L_{OBB} are the same as instance segmentation branch of Mask R-CNN [55] in the training stage, and L_{SEG} is computed as per-pixel cross entropy loss between the predicted and the ground truth labels. $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are the weights of these sub-losses.

The ground truth of the OBB branch is described in Section 2.2. We calculate OBB's horizontal bounding box as the ground truth of RPN and HBB branch. The ground truth of semantic segmentation branch is described in Section 2.5.

3. Experiments

In this section, we describe the implementation of the proposed method in detail and demonstrate the performance of our proposed method on DOTA [2] dataset and HRSC2016 [26] dataset with state-of-the-art methods.

3.1. Datasets and Evaluation Metrics

3.1.1. DOTA Dataset

DOTA [2] is a dataset for multi-category object detection in aerial images. It contains 2806 images from different cameras and platforms. The image sizes vary from about 800 × 800 to 4000 × 4000 pixels. There are 15 object categories: baseball diamond (BD), ground track field (GTF), small vehicle (SV), large vehicle (LV), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), swimming pool (SP), helicopter (HC), bridge (BR), harbor (HA), ship (SP), plane (PL). Each object in this dataset is annotated with an arbitrary quadrilateral which is the same as point-based OBB. The training, validation and test sets include 1/2, 1/6 and 1/3 of the dataset, respectively. DOTA aims for two tasks: Horizontal Bounding Box Task (HBB task) and Oriented Bounding Boxes Task (OBB task), and provides an evaluation server. It is one of the largest and the most challenging aerial image object detection datasets.

3.1.2. HRSC2016 Dataset

HRSC2016 [26] is a dataset for ship detection in aerial images. It contains 1061 images collected from Google Earth and has more than 20 categories of ships. The image size ranges from 300 × 300 to 1500 × 900 pixels. It consists of 436 training images, 181 validation images and 444 test images. Objects in HRSC2016 are annotated with θ-based OBBs.

3.1.3. Evaluation Metrics

For DOTA, we submit our results on test set to official evaluation server to obtain the mean Average Precision (mAP). For HRSC2016 dataset, we report the standard VOC-style AP metrics with Intersection Over Union (IoU) threshold of 0.5.

3.2. Implementation Details

Our model is implemented with PyTorch [57]. We use SGD with a weight decay of 0.0001 and momentum of 0.9 on 4 NVIDIA Titan Xp GPUs with a total of 8 images per mini-batch (2 images per GPU). We train 12 epochs in total with an initial learning rate of 0.01, and decrease it by a factor of 0.1 at epoch 9 and 11. The batch size of RPN and Fast R-CNN is set to 256 and 512 per image with a sample ratio 1 : 3 of positive to negatives. In the multi-task loss function, we set $\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1, \alpha_4 = 0.2$. We use ResNet-50 [56] with FPN [39] as the backbone for all experiments, if not specified otherwise. All models are trained on the training and validation sets, then evaluated on the test set.

Multi-scale training and testing (MSTT) and data augmentation (Data Aug.) technologies are applied in final results of DOTA and HRSC2016 datasets when compared with state-of-the-art detectors in Section 3.4.

For DOTA dataset, we use three scales $\{(1024, 1024), (896, 896), (768, 768)\}$ to apply MSTT in both training and inference stage. For data augmentation, we resize the original images at two scales (1.0 and 0.5) before dividing the images into patches. After resizing, we divide the resized images into 1024 × 1024 patches with an overlap of 200 in both training and inference stage. In addition, each image is randomly flipped with a probability of 0.5 and randomly rotate an angle from an angle set $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ in training stage.

For HRSC2016 dataset, in the training stage, long sides of the input images are resized to 1024 pixels, and short sides are randomly resized to a range of $[800, 1024]$ pixels, and in the inference stage, we use three scales $\{(1280 \times 1024), (1024 \times 800), (800 \times 600)\}$ to do MSTT. For data augmentation, we firstly do 5 times data augmentation by randomly rotating the input images with an angle range of $[-90°, 90°]$ before training. Then, in the training stage, each image is randomly flipped with a probability of 0.5 and we randomly rotate an angle from an angle set $\{0°, 90°, 180°, 270°\}$.

For more details, in the training stage, the proposal number of RPN is set to 2000, the HBB branch runs on all these proposals and the OBB branch just runs on positive proposals which have IoU overlap with a ground-truth bounding box of at least 0.5. In the inference stage, the proposal number of RPN is set to 2000, we run the HBB branch on these proposals, following Non Maximum Suppression (NMS) [58], the OBB branch is then applied to 500 horizontal bounding boxes with the highest scores. Finally, the oriented bounding boxes are generated from the predicted mask in OBB branch by post-processing.

3.3. Comparison of Different OBB Representations

OBBs can be represented in a variety of ways, as shown in Figure 1. θ-based OBB, point-based OBB and h-based OBB are the most common representation methods. In this section, we firstly study the different "first vertex" definition methods which will affect the performance of point-based OBB and h-based OBB in Table 1, and then, we study the effect of different OBB representations in Table 2. For a fair comparison, we re-implement above three bounding box representations on the same basic network structure as Mask OBB. Note that in our implementations, we use the box-encoding functions $\phi_\theta(g; p), \phi_{\text{point}}(g; p), \phi_h(g; p)$ to encode the ground truth boxes of θ-based OBB, point-based OBB and h-based OBB with respect to their matching proposals generated by RPN which can be obtained by following equations:

$$\begin{aligned}
\phi_\theta(g; p) &= [10 \cdot \frac{g_x - p_x}{p_w}, 10 \cdot \frac{g_y - p_y}{p_h}, 5 \cdot \log(\frac{g_w}{p_w}), 5 \cdot \log(\frac{g_h}{p_h}), 10 \cdot (g_\theta - p_\theta)] \\
\phi_{\text{point}}(g; p) &= [10 \cdot \frac{g_{x_1} - p_{x_1}}{p_w}, 10 \cdot \frac{g_{y_1} - p_{y_1}}{p_h}, 10 \cdot \frac{g_{x_2} - p_{x_2}}{p_w}, 10 \cdot \frac{g_{y_2} - p_{y_2}}{p_h}, \\
&\quad 10 \cdot \frac{g_{x_3} - p_{x_3}}{p_w}, 10 \cdot \frac{g_{y_3} - p_{y_3}}{p_h}, 10 \cdot \frac{g_{x_4} - p_{x_4}}{p_w}, 10 \cdot \frac{g_{y_4} - p_{y_4}}{p_h}] \\
\phi_h(g; p) &= [10 \cdot \frac{g_{x_1} - p_{x_1}}{p_w}, 10 \cdot \frac{g_{y_1} - p_{y_1}}{p_h}, 10 \cdot \frac{g_{x_2} - p_{x_2}}{p_w}, 10 \cdot \frac{g_{y_2} - p_{y_2}}{p_h}, \\
&\quad 10 \cdot \frac{g_h - p_h}{p_h}]
\end{aligned}$$

where g and p denote ground truth box and proposal, respectively.

Table 1. Comparison with different first vertex definition methods on the mAP of point-based OBB and h-based OBB representations. "Best point" method significantly outperforms "extreme point" method on the OBB task of DOTA dataset. The best result in the same OBB representation is highlighted in bold.

Dataset	First Vertex	OBB Representation	Backbone	OBB (%)	HBB (%)	Gap (%)
DOTA	extreme point	point-based OBB	ResNet-50-FPN	64.40	68.72	4.32
		h-based OBB	ResNet-50-FPN	62.95	70.73	7.78
	best point	point-based OBB	ResNet-50-FPN	**69.35**	70.65	**1.30**
		h-based OBB	ResNet-50-FPN	**67.36**	70.46	**3.10**

For the first vertex definition, we compare two different methods. One is the same as [2], which chooses the vertex closest to the "top left" vertex of the corresponding HBB, and we call this method as "best point". The other one is defined by ourself, which chooses the "extreme top" vertex of OBB as the first vertex, then other vertexes are fixed in clockwise order, and we call this method as "extreme point". As shown in Table 1, "best point" method significantly outperforms "extreme point" method on the OBB task of DOTA dataset. We can learn that different "first vertex" definition methods will significantly affect mAPs of OBB task. Thus if we want to obtain great performance on the OBB task by using point-based OBB and h-based OBB representations, we should design a special "first vertex" definition method which can represent OBB uniquely.

For different OBB representations, there is a higher gap between the HBB and OBB performance for both θ-based OBB, point-based OBB and h-based OBB representation than Mask OBB. Theoretically, changing from prediction of HBB to OBB should not affect the classification precision, but as shown in Table 2, the methods which use regression-based OBB representations have higher HBB task performance than OBB task performance. We argue that the reduction is due to the low quality localization, which is caused by the discontinuity point as discussed in Section 2.1. There should not be such a large gap between the performance of HBB and OBB task if the representation of OBB is defined well. The result of Mask OBB verified that. In addition, mAPs on HBB and OBB tasks of Mask OBB are nearly all higher than the other three OBB representations in our implementations.

Table 2. Comparison with different methods on the gap of mAP between HBB and OBB. The best result in the gap is highlighted in bold.

Implementations	OBB Representation	Backbone	OBB (%)	HBB (%)	Gap (%)
Ours	θ-based OBB	ResNet-50-FPN	69.06	70.22	1.16
	point-based OBB	ResNet-50-FPN	69.35	70.65	1.30
	h-based OBB	ResNet-50-FPN	67.36	70.46	3.10
	Mask OBB	ResNet-50-FPN	69.97	70.14	**0.17**
FR-O [2]	point-based OBB	ResNet-50-C4	54.13	60.46	6.33
ICN [30]	point-based OBB	ResNet-50-FPN	68.16	72.45	4.29
SCRDet [59]	θ-based OBB	ResNet-101-FPN	72.61	75.35	2.74
Li et al. [49]	θ-based OBB	ResNet-101-FPN	73.28	75.38	2.10

For other implementations, FR-O [2] uses point-based OBB and gets 60.46% HBB mAP and 54.13% OBB mAP, and the gap is 6.33%. ICN [30] also uses point-based OBB and gets 72.45% HBB mAP and 68.16% OBB mAP, and the gap is 4.29%. SCRDet [59] uses θ-based OBB and gets 72.61% OBB map and 75.35% HBB map, and the gap is 2.70%. Li et al. [49] also uses θ-based OBB and gets 73.28% OBB map and 75.38% HBB map, and the gap is 2.10%. Note that the performances of ICN, SCRDet and Li et al. are obtained by using other modules and data augmentation technology. The gaps between HBB task and OBB task of these methods (6.33%, 4.29%, 2.70%, 2.10%) are all higher than Mask OBB (0.17%). Therefore, We can draw the conclusion that Mask OBB is a better representation on the oriented object detection problem. Figure 7 shows some visualization results in our implementations by using different OBB representation methods on OBB task of DOTA dataset. We can observe that detection results are very bad when the angles of objects relative to HBBs are near $\pi/4$ or $3\pi/4$ in θ-based OBB, point-based OBB and h-based OBB, but Mask OBB can compactly enclose oriented objects.

3.4. Comparison with State-of-the-Art Detectors

We compared the performance of our method with the state-of-the-art methods on the OBB task and the HBB task of two datasets DOTA and HRSC2016.

3.4.1. Results on DOTA Dataset

We compare our method with the state-of-the-art methods on OBB and HBB tasks of DOTA dataset in Tables 3 and 4. Besides the official baseline given by DOTA, we also compare proposed model with RRPN [34], R^{2}CNN [53], R-DFPN [35], ICN [30], RoI Transformer [1], SCRDet [59] and

Li et al. [49] which have been introduced in Section 1. Note that some methods only report mAP of the OBB detection task. The results in Tables 3 and 4 are obtained by using Soft NMS [60], MSTT and Data Aug. By using ResNet-50, our method achieves 74.86% and 75.98% mAP on OBB task of DOTA, respectively, and outperforms all methods which even use ResNet-101. In addition, by using ResNeXt [61], our method achieves 75.33% and 76.98% mAP on the OBB task and HBB task of DOTA, respectively. We note that our method attains a little gap (1.65%) between its OBB task mAP and HBB task mAP. Our method outperforms all methods evaluated on this dataset. Figures 8 and 9 show some visualization results on the DOTA dataset.

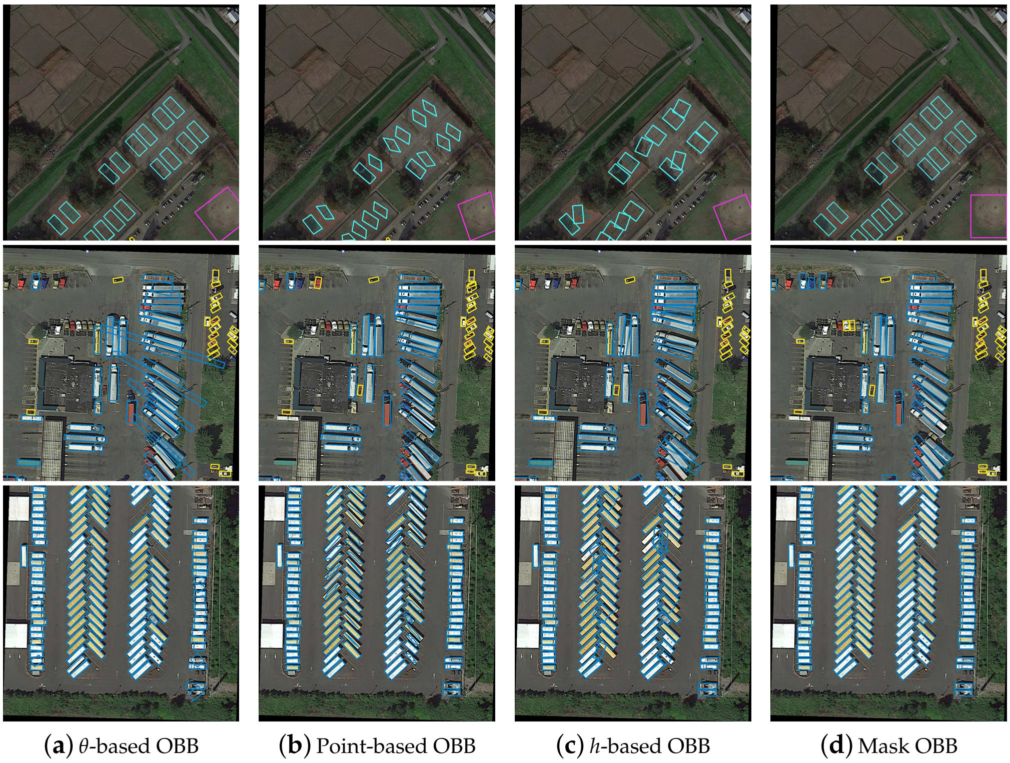

(**a**) θ-based OBB (**b**) Point-based OBB (**c**) h-based OBB (**d**) Mask OBB

Figure 7. Visualization of detection results by using different OBB representation methods on OBB task of DOTA dataset. Compared with other OBB representation methods, Mask OBB can compactly enclose oriented objects.

Table 3. Quantitative comparison of the baselines and our method on the OBB task in the test set of DOTA (%). The best result in each category is highlighted in bold.

Method	**Backbone**	**FPN**	**PL**	**BD**	**BR**	**GTF**	**SV**	**LV**	**SH**	**TC**	**BC**	**ST**	**SBF**	**RA**	**HA**	**SP**	**HC**	**mAP**
SSD [18]	VGG-16	-	39.83	9.09	0.64	13.18	0.26	0.39	1.11	16.24	27.57	9.23	27.16	9.09	3.03	1.05	1.01	10.59
YOLOv2 [16]	Darknet-19	-	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.35	34.65	16.02	37.62	47.23	25.5	7.45	21.39
R-FCN [10]	ResNet-101	-	37.80	38.21	3.64	37.26	6.74	2.60	5.59	22.85	46.93	66.04	33.37	47.15	10.60	25.19	17.96	26.79
FR-H [9]	ResNet-50	-	47.16	61.00	9.80	51.74	14.87	12.80	6.88	56.26	59.97	57.32	47.83	48.70	8.23	37.25	23.05	32.29
FR-O [2]	ResNet-50	-	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.40	52.52	46.69	44.80	46.30	52.93
R-DFPN [35]	ResNet-101	✓	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
R^{2}CNN [53]	ResNet-101	-	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN [34]	ResNet-101	-	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
ICN [30]	ResNet-101	✓	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
RoI Trans. [1]	ResNet-101	✓	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet [59]	ResNet-101	✓	89.98	80.65	52.09	68.36	68.36	60.32	72.41	**90.85**	**87.94**	**86.86**	**65.02**	66.68	66.25	68.24	65.21	72.61
Li et al. [49]	ResNet-101	✓	**90.21**	79.58	45.49	**76.41**	73.18	68.27	79.56	90.83	83.40	84.68	53.40	65.42	**74.17**	69.69	64.86	73.28
Ours	ResNet-50	✓	89.61	85.09	51.85	72.90	75.28	73.23	85.57	90.37	82.08	85.05	55.73	68.39	71.61	**69.87**	**66.33**	74.86
Ours	ResNeXt-101	✓	89.56	**85.95**	**54.21**	72.90	**76.52**	**74.16**	**85.63**	89.85	83.81	86.48	54.89	**69.64**	73.94	69.06	63.32	**75.33**

Table 4. Quantitative comparison of the baselines and our method on the HBB task in the test set of DOTA (%). The best result in each category is highlighted in bold.

Method	Backbone	FPN	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
SSD [18]	VGG-16	-	57.85	32.79	16.14	18.67	0.05	36.93	24.74	81.16	25.10	47.47	11.22	31.53	14.12	9.09	0.00	29.86
YOLOv2 [16]	Darknet-19	-	76.90	33.87	22.73	34.88	38.73	32.02	52.37	61.65	48.54	33.91	29.27	36.83	36.44	38.26	11.61	39.20
R-FCN [10]	ResNet-101	-	81.01	58.96	31.64	58.97	49.77	45.04	49.29	68.99	52.07	67.42	41.83	51.44	45.15	53.30	33.89	52.58
FR-H [9]	ResNet-50	-	80.32	77.55	32.86	68.13	53.66	52.49	50.04	90.41	75.05	59.59	57.00	49.81	61.69	56.46	41.85	60.46
FPN [39]	ResNet-50	✓	88.70	75.10	52.60	59.20	69.40	**78.80**	84.50	90.60	81.30	82.60	52.50	62.10	76.60	66.30	60.10	72.00
ICN [30]	ResNet-101	✓	90.00	77.70	53.40	73.30	73.50	65.00	78.20	90.80	79.10	84.80	57.20	62.10	73.50	70.20	58.10	72.50
IoU-Adaptive [62]	ResNet-101	✓	88.62	80.22	53.18	66.97	76.30	72.59	84.07	90.66	80.95	76.24	57.12	66.65	74.08	66.36	56.85	72.72
SCRDet [59]	ResNet-101	✓	**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
Li et al. [49]	ResNet-101	✓	90.15	78.60	51.92	**75.23**	73.60	71.27	81.41	90.85	83.94	84.77	58.91	65.65	76.92	**79.36**	68.17	75.38
Ours	ResNet-50	✓	89.60	85.82	56.50	71.18	77.62	70.45	85.04	90.18	80.10	85.30	56.60	69.43	75.45	76.71	69.70	75.98
Ours	ResNeXt-101	✓	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**

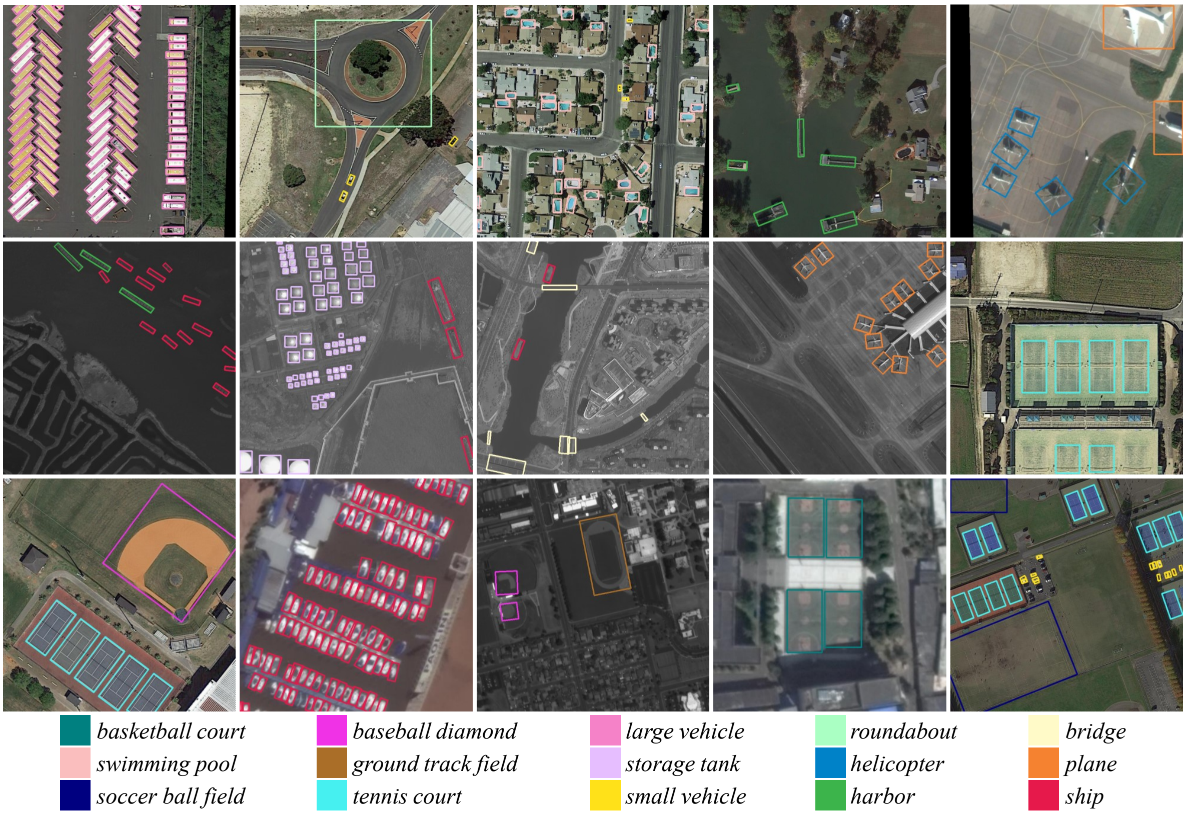


Figure 8. Visualization of detection results using our method on OBB task of DOTA.

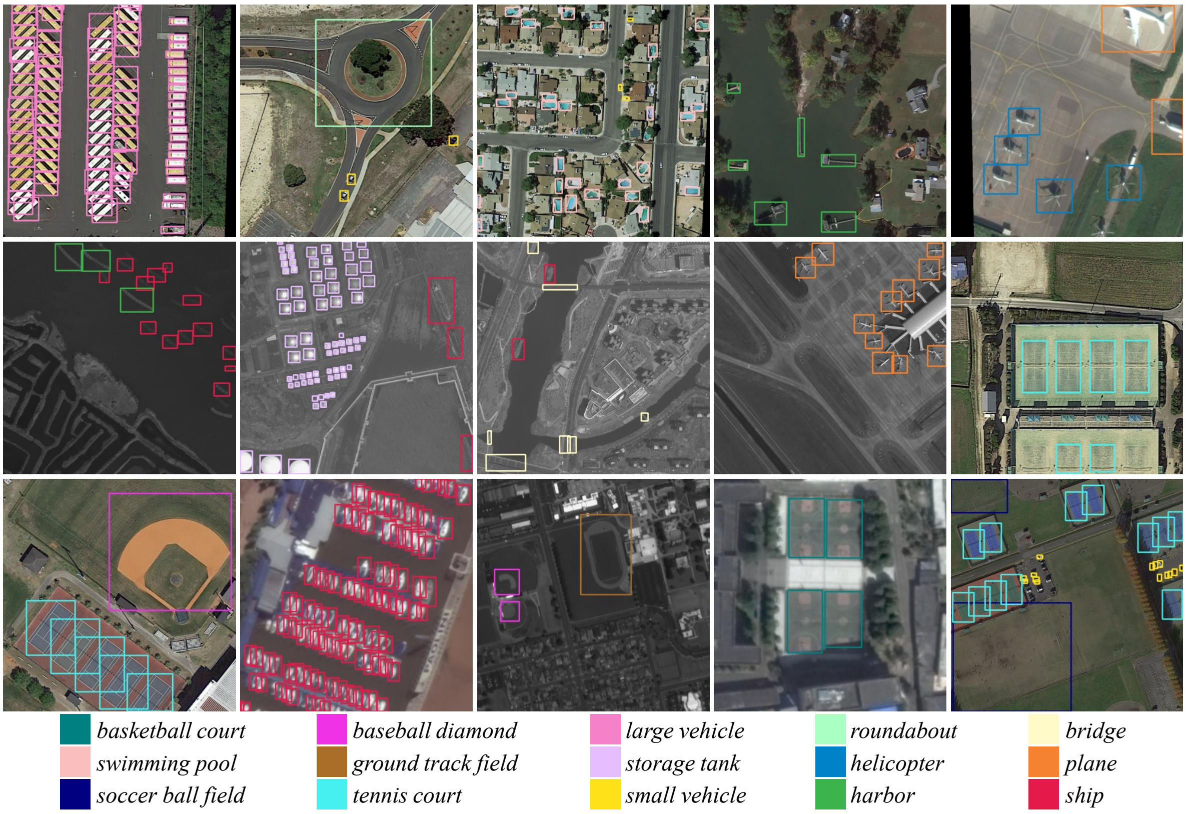


Figure 9. Visualization of detection results using our method on HBB task of DOTA.

3.4.2. Results on HRSC2016 Dataset

Table 5 shows the comparison results with state-of-the-art methods of HRSC2016 OBB detection task. Our full model achieves 96.70% mAP of OBB detection task. It outperforms all other methods evaluated on this dataset with a promotion around **4.8** points in mAP. Some visualization results on HRSC2016 dataset are displayed in Figure 10.

Table 5. Comparison with the state-of-the-art methods on HRSC2016 OBB task. The best result is highlighted in bold.

Method	mAP (%)
BL2 [63]	69.6
RC1 [63]	75.7
RC2 [63]	75.7
R^{2}PN [64]	79.6
RRD [31]	84.3
RoI Trans. [1]	86.2
RBOX-CNN [65]	91.9
Ours	**96.7**

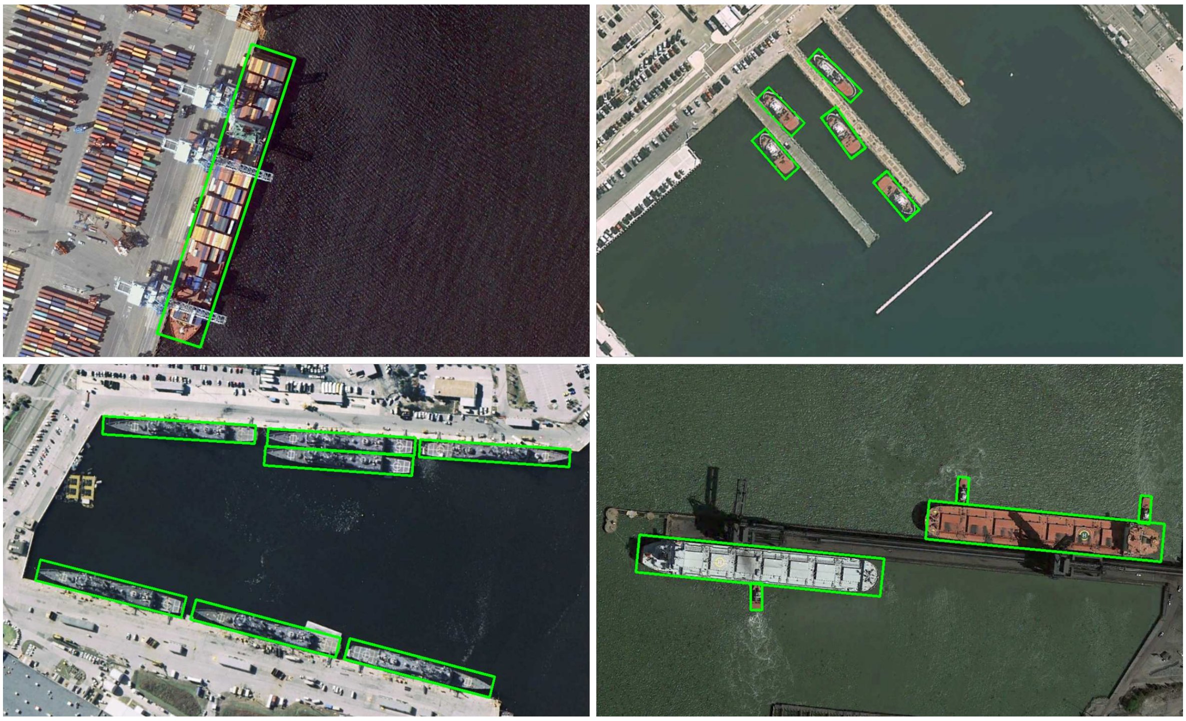


Figure 10. Visualization of detection results using our method on HRSC2016.

4. Discussion

4.1. Ablation Study

To verify the effectiveness of our approach, we do a series of comparative experiments on DOTA test set. Table 6 summarizes the results of our models with different settings on DOTA dataset. The detailed comparison is given in the following.

Baseline setup. Mask R-CNN which is extended for oriented object detection task without other components is used as the baseline of the ablation experiments. To ensure the fairness and accuracy of the experiment, all experimental data and parameter settings are strictly consistent. We use ResNet-50-FPN as the backbone and mAP as the indicator of model performance. The results of mAP on DOTA reported here are obtained by submitting results to the official DOTA evaluation server. In our implementation, it gets 69.97% and 70.14% mAPs for OBB task and HBB task.

Effect of ILC-FPN. As discussed in Section 2.4, the inception structure can help FPN to better handle large scale variations in aerial images. Through the experimental results in Table 6, we can observe that the use of ILC-FPN can significantly improve the detection performance of 1.12% on OBB task and 1.47% on HBB task because ILCN in ILC-FPN can help FPN to extract more discriminative features of objects in aerial images.

Table 6. Ablation study of each component in our proposed method on DOTA. ILC-FPN is the Inception Lateral Connection Feature Pyramid Network, SAN is the Semantic Attention Network.

Dataset	Algorithm	OBB (%)	HBB (%)
DOTA	baseline	69.97	70.14
	baseline + ILC-FPN	71.09 (↑ 1.12)	71.61 (↑ 1.47)
	baseline + SAN	70.69 (↑ 0.72)	72.02 (↑ 0.88)
	baseline + ILC-FPN + SAN	71.43 (↑ 1.46)	72.41 (↑ 2.27)

Effect of SAN. Table 6 shows our Semantic Attention Network can improve mAP of 0.72% on OBB task and 0.88% on HBB task compare baseline. Compared baseline+ILC-FPN, it also improves

0.34% on OBB task and 0.80% on HBB task. It shows the importance of semantic features on the whole image in aerial image detection.

4.2. Failure Cases

Figure 11 shows some failure cases. In the DOTA, errors are most likely to occur on long and narrow objects, whose parts are possible to be detected as objects. For instance, as shown in the second and the third images of the row 1, the local part of harbor and bridge are detected in some situations. Some failure examples are caused by objects with large extent like large-vehicle and basketball court. However, this is not always the case, as the ship can be seen in the second image of the row 2, and some other failure cases are caused by the huge objects like ground track field and soccer ball field shown in the third image of row 2.

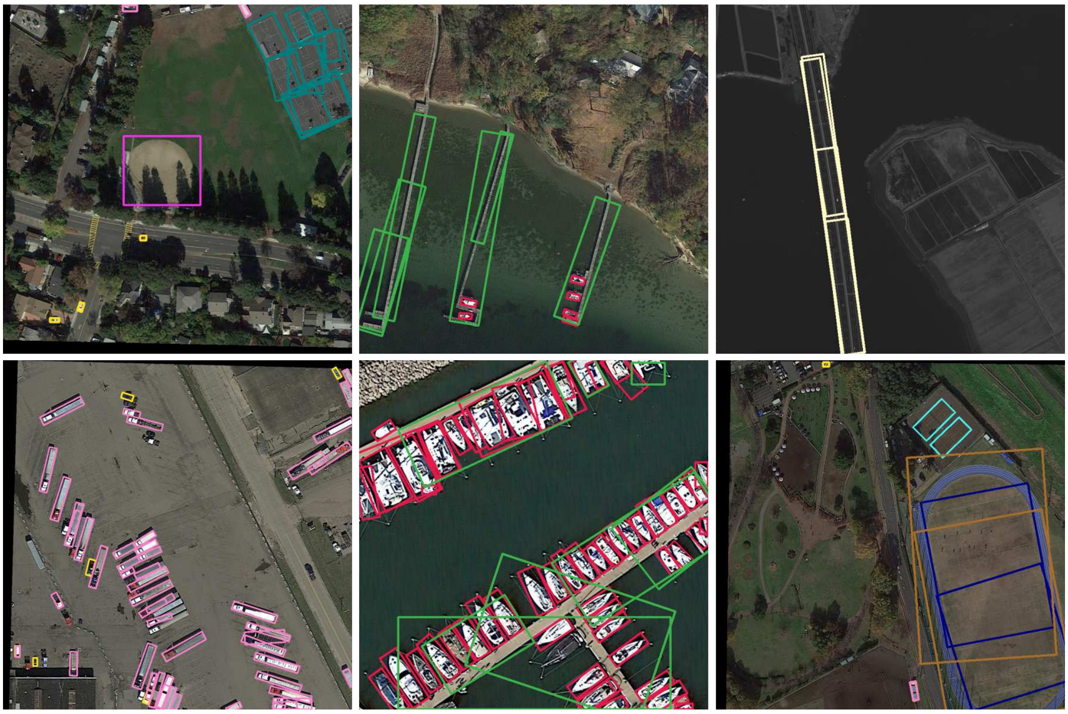


Figure 11. Some typical failure predictions of our method.

5. Conclusions

In this paper, we analyzed the influence of different OBB representations for oriented object detection in aerial images, which exposes shortcomings of the typical regression-based OBB representation methods like θ-based, point-based and h-based OBB representation methods. Based on the analysis, the Mask OBB representation is proposed to tackle the ambiguity in regression-based OBB. In addition, we proposed the Inception Lateral Connection Feature Pyramid Network (ILC-FPN) which usees Inception Lateral Connection (ILCN) to enhance the feature extraction ability of FPN for handling the scale variation problem in aerial images. Furthermore, we proposed the Semantic Attention Network (SAN) to extract semantic features to further enhance features of generating HBBs and OBBs. Experimental results on the DOTA and HRSC2016 datasets demonstrated the importance of representations in multi-category arbitrary-oriented object detection. Notably, our method achieves 75.33% and 76.98% on OBB task and HBB task of DOTA dataset, respectively. At the same time, it achieves 96.70% on OBB task of HRSC2016 dataset.

Author Contributions: Funding acquisition, W.Y.; Investigation, J.W. and J.D.; Methodology, J.W. and J.D.; Supervision, W.Y.; Writing—original draft, J.W.; Writing—review & editing, J.D., H.G., W.C., T.P. and W.Y.

Funding: The research was partially supported by the National Natural Science Foundation of China (NSFC) under Grant 61771351 and the project for innovative research groups of the Natural Science Foundation of Hubei Province (No. 2018CFA006).

Acknowledgments: Sincere thanks are given to Gui-Song Xia for his helpful discussions and suggestions. The authors would also like to express their sincere thanks to the anonymous reviewers for their valuable comments and contributions. In addition, the numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Detecting Oriented Objects in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, 16–19 June 2019.
2. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
3. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
4. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
5. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2337–2348. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [CrossRef]
7. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
8. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–15 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2015; pp. 91–99.
10. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016; pp. 379–387.
11. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 764–773.
12. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Light-head r-cnn: In defense of two-stage object detector. *arXiv* **2017**, arXiv:1711.07264.
13. Singh, B.; Davis, L.S. An analysis of scale invariance in object detection snip. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3578–3587.
14. Singh, B.; Najibi, M.; Davis, L.S. SNIPER: Efficient multi-scale training. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 3–8 December 2018; pp. 9310–9320.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
16. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
17. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.

18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 21–37.
19. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
20. Han, X.; Zhong, Y.; Zhang, L. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
21. Xu, Z.; Xu, X.; Wang, L.; Yang, R.; Pu, F. Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery. *Remote Sens.* **2017**, *9*, 1312. [CrossRef]
22. Guo, W.; Yang, W.; Zhang, H.; Hua, G. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. *Remote Sens.* **2018**, *10*, 131. [CrossRef]
23. Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X.X. HSF-Net: Multiscale Deep Feature Embedding for Ship Detection in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7147–7161. [CrossRef]
24. Pang, J.; Li, C.; Shi, J.; Xu, Z.; Feng, H. $\mathcal{R}^2$-CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5512–5524. [CrossRef]
25. Dong, R.; Xu, D.; Zhao, J.; Jiao, L.; An, J. Sig-NMS-Based Faster R-CNN Combining Transfer Learning for Small Target Detection in VHR Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8534–8545. [CrossRef]
26. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
27. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
28. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
29. Liu, K.; Mattyus, G. Fast multiclass vehicle detection on aerial images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1938–1942.
30. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. *arXiv* **2018**, arXiv:1807.02700.
31. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.S.; Bai, X. Rotation-Sensitive Regression for Oriented Scene Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2018; pp. 5909–5918.
32. Li, Q.; Mou, L.; Xu, Q.; Zhang, Y.; Zhu, X.X. R^3-Net: A Deep Network for Multi-oriented Vehicle Detection in Aerial Images and Videos. *arXiv* **2019**, arXiv:1808.05560.
33. Liu, L.; Pan, Z.; Lei, B. Learning a rotation invariant detector with rotatable bounding box. *arXiv* **2017**, arXiv:1711.09405.
34. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
35. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
36. Dai, Y.; Huang, Z.; Gao, Y.; Xu, Y.; Chen, K.; Guo, J.; Qiu, W. Fused Text Segmentation Networks for Multi-oriented Scene Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Beijing, China, 20–24 August 2018; pp. 3604–3609.
37. Lyu, P.; Liao, M.; Yao, C.; Wu, W.; Bai, X. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2018; pp. 71–88.
38. Wang, J.; Ding, J.; Cheng, W.; Yang, W.; Xia, G. Mask-OBB: A Mask Oriented Bounding Box Representation for Object Detection in Aerial Images. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
39. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

40. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
41. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
42. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 26 June–1 July 2016; pp. 2818–2826.
43. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the National Conference on Artificial Intelligence, San Diego, CA, USA, 28–29 June 2017; pp. 4278–4284.
44. Chaib, S.; Liu, H.; Gu, Y.; Yao, H. Deep Feature Fusion for VHR Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4775–4784. [CrossRef]
45. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
46. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
47. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R.B. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2874–2883.
48. Wang, X.; Girshick, R.B.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
49. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; Yang, J. Feature-Attentioned Object Detection in Remote Sensing Imagery. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 3886–3890.
50. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015-10024. [CrossRef]
51. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-Scale Spatial and Channel-wise Attention for Improving Object Detection in Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2019**. [CrossRef]
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
53. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational region CNN for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
54. Suzuki, S.; Be, K.A. Topological structural analysis of digitized binary images by border following. *Comput. Vis. Graph. Image Process.* **1985**, *30*, 32–46. [CrossRef]
55. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
57. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the Advances in Neural Information Processing Systems Workshop, Long Beach, CA, USA, 4–9 December 2017.
58. Girshick, R.; Iandola, F.; Darrell, T.; Malik, J. Deformable part models are convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 437–446.
59. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.

60. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–Improving Object Detection With One Line of Code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
61. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
62. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H. IoU-Adaptive Deformable R-CNN: Make Full Use of IoU for Multi-Class Object Detection in Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]
63. Liu, Z.; Hu, J.; Weng, L.; Yang, Y. Rotated region based CNN for ship detection. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 900–904.
64. Zhang, Z.; Guo, W.; Zhu, S.; Yu, W. Toward Arbitrary-Oriented Ship Detection With Rotated Region Proposal and Discrimination Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1745–1749. [CrossRef]
65. Koo, J.; Seo, J.; Jeon, S.; Choe, J.; Jeon, T. RBox-CNN: Rotated bounding box based CNN for ship detection in remote sensing image. In Proceedings of the International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 6–9 November 2018; pp. 420–423.