IENet: Interacting Embranchment One Stage Anchor Free Detector for Orientation Aerial Object Detection

Youtian Lin
CCST, HRBEU
Harbin, China
linyoutian.loyot@gmail.com

Pengming Feng
SGIIT, CAST
Beijing, China
P.feng.cn@outlook.com

Jian Guan
CCST, HRBEU
Harbin, China
j.guan@hrbeu.edu.cn

Abstract

Object detection in aerial images is a challenging task due to its lack of visiable features and variant orientation of objects. Currently, amount of R-CNN framework based detectors have made significant progress in predicting targets by horizontal bounding boxes (HBB) and oriented bounding boxes (OBB). However, there is still open space for one-stage anchor free solutions. This paper proposes a one-stage anchor free detector for orientional object in aerial images, which is built upon a per-pixel prediction fashion detector. We make it possible by developing a branch interacting module with a self-attention mechanism to fuse features from classification and box regression branchs. Moreover a geometric transformation is employed in angle prediction to make it more manageable for the prediction network. We also introduce an IOU loss for OBB detection, which is more efficient than regular polygon IOU. The propsed method is evaluated on DOTA and HRSC2016 datasets, and the outcomes show the higher OBB detection performance from our propsed IENet when compared with the state-of-the-art detectors.

1. Introduction

Recently, with the advance development of deep convolutional neural networks, object detection has achieved tremendous success in natural images. Detecting objects in aerial images is also achieved significant progress using mainstream object detection methods (e.g., Faster-RCNN [33], YOLO [32], SSD [24]). However, in aerial images, objects are captured at a downward perspective, and objects are always arbitrary oriented, so it is difficult to apply standard detection methods to oriented objects in remote sensing and aerial images. This task comes with following significant challenges:

- In aerial iamges, most objects are having similar shape and fewer appearance features than natural images(e.g., House, Vehicle). These object could lead to misdetection because of the shape is more evident than the appearance for the model in this circumstances.
- The highly complex background and variant appearances of targets increase the difficulty of target detection, especially for small and densely distributed targets.
- The birdviews perspective increases the complexity of the variant orientation of objects, thus the model is obstinate to obtain the parameters to represent the diversity angle.

To address these challengs, most of the two-stage oriented object detector already reported great performances. These methods have benefited a lot from the R-CNN mechanism [13, 20, 6]. However, most of the object in detection datasets are labeled by horizontal bounding boxes, which could lead to region overlap between objects, so datasets like DOTA [37] come with high-grade label oriented bounding box, could solve the overlap issue. [43] handle the orientation regression by adding different angle anchors to the regression head in the region proposal step and Roi regression step, this allows the existing R-CNN methods to produce oriented bounding box by recognizing the object orientation angle. Nevertheless, the features extraction layers (e.g., RoI Pooling [7], Deformable Convolution [8], Spatial Transformer [17]) in most of the R-CNN framework has a limitation when it comes to predict the oriented bounding box, which will lead to extracts the overlap feature between the objects. Hence [10] proposes RoI Transformer to extract the rotation region feature for orientation objects sufficiently.

In common knowledge, despite the R-CNN's two-stage mechanism increase the accuracy of the detection, the mechanism also increases the computing complexity in the training step and decreases the inference speed in the test step. This obstacle exit orientated object detection in aerial images when using the R-CNN framework, these networks

are required careful design and refined select in component and hyperparameter. To eliminate the above problem, and it is essential to design a unified one-stage detector for orientated object detection in aerial images.

In order to reduce the training times and inference times, a one-stage detector fully utilizes the advantage of the fully convolutional networks(FCNs) [29]. By the power of the feature pyramid networks (FPN) [22] and pre-define anchor boxes, a one-stage detector achieved state-of-the-art performance [45, 23, 44]. Nevertheless, the anchor boxes strategy also takes many computing times. Therefore, [34] come up with per-pixel prediction fashion, inspire by semantic segmentation, this method outperforms most of the one-stage anchor base detector with anchor free solution. The anchor base method predicts the HBB, which could transform to OBB by just adding the angle anchor. Therefore, [21] develop a one-stage detector for oriented scene text detection, which directly uses the HBB as an anchor to regress the OBB. This method performs state-of-the-art results on text detection. However, even though the method is for oriented object detection, but the scene text is very different from aerial image object, which is harder to predict because of the sizeable dense cluster object have more misalignment in HBB. Futher, there is no anchor free solution in the one-stage oriented object detection in aerial images, and this is because the one-stage detector does not have an excellent feature extractor like RoI Pooling, which is an essential part of R-CNN detector. Therefore, the lack of the feature extractor is severe for the model to recognize the object orientation.

The one-stage and two-stage detection method do not have an effective solution for OBB. However, it is essential to design a model, which would balance the performance and speed in predicting OBB.

In this paper, we develop a one-stage anchor free network for orientated object detection in aerial images. To our knowledge, this work will be the first one-stage anchor free solution for orientated object detection in aerial images. However, most detection models treat OBB as an auxiliary task on box regression, but we form the angle prediction as an independent task. We built our network based on the FCOS network and add our orientation task to the network using an independent branch to regress the orientation parameter. Furthermore, we develop a new way to extract more features for oriented prediction, a branch fusion black based on attention mechanism named Interacting Embranchment (IE) black, to the best use of both features in classification and regression branch to provide more practical features for angle prediction. IE black forces the task to interact with features from all branches in the network. The interacting behavior helps the network to select more consistent and relevant features, through this process, the training is more stable, and the oriented prediction can be elevated. We introduce a simple OBB version of IOU loss to acquire computation efficiency and also allow the network to produce high accuracy OBB. To further prove the efficiency of our method, we modify the state-of-the-art one-stage detector to adapt to predict OBB. Nevertheless, we apply a geometric transformation for OBB representation, which split the angle prediction task into two separate tasks. We show in the same backbone setting our network outperform the adapt baseline network, we also compare with the state-of-the-art two-stage detector, to reveal our model obtain high performance while maintaining great computing and memory efficiency.

In summary, our contributions are shown as follows:

- We propose a one-stage anchor free detector for oriented object detection in aerial images by treating the orientation prediction as an independent task. Moreover, we apply a geometric transformation and an OBB version of IOU loss to regress the OBB better.
- We use self-attention mechanisms to develop IE module to force the orientation prediction task to interact the features in classification and regression branches to further improve the accuracy of orientation detection.
- We show the outcome from our method achieve when compare with state-of-the-art the one-stage detectors on public datasets for orientation detection in aerial images.

2. Related Work

2.1. Two-stage Detector

The first two-stage method on object detection is introduced by by [14]. Two-stage detector solves the object detection by looking at the image two times, and the first look is to generate a region proposal set, which detects the possible region of the object. The second look is to extract the feature from the backbone feature maps for each region proposal and send the feature to a classifier to identify the object category. Later, [12] design an RoI pooling layer to extract the feature in a fully convolutional way. In this way, RoI pooling accelerates the processing speed. [26] improve this two-stage framework in some detail.

However, this progress can not merely apply to the oriented object detection, because of these methods is base on horizontal bounding boxes. [30, 28] design a rotated anchor to generate Rotated Region Proposal(R-RoI), and use a Rotated RoI Warping to extract feature from an R-RoI. However, the R-RoI based method involves generating a lot of rotated proposals. According to [1, 43], rotated proposal anchors are challenging to be embedded in the neural network, which would cost extra time for rotated proposal generation. Therefore, due to the computation cast for ro-

tated anchor, [10] propose a method to avoid the rotated anchor computation by transforming the RoI to RRoI using a light fully connected layer. Moreover, they also add an IOU loss to matching two OBBs, which can effectively avoid the problem of misalignment. These two-stage detectors obtain the high performentce while sacrifice the computation cost. So we use a anchor free one-stage detector to directly predict the object without the complex computation on anchor matching and RoI feature extraction.

2.2. One-stage Detector

The one-stage approach obtains high performance and running speed, this is because of one-stage detectors are usually more computationally efficient than two-stage detectors, such as [24]. However, the anchor base detectors which detect the object by predicting the offsets with the dense anchor boxes would create a massive imbalance between positive and negative anchor boxes during training. [23] propose Focal Loss to address this imbalance issue. Nevertheless, this still requires much computation on anchor, [34] introduce an anchor free one-stage detector built upon the previous work [23] which is using a per-pixel prediction manner. This emancipate the model from high dense computation on achor matching.

Most of the one-stage oriented object detectors that achieve high performance are the text scene detectors [2, 39], and [38] use mask to form the OBB. These method could be directly employ on aerial image datasets which the objects are labeled by OBB. However, text scene detection is far different from aerial object detection which have dissimilar challenge mention in Section 1. The IENet also use one-stage to directly regress all the parameter represent the object, but with the help of out geometric transformation which split the angle prediction to two geometric parameter prediction. This make the model predict parameter distributed in a lower dimension. Futher, to solve the OBB detection in anchor free solution, IENet is constructed on [34].

2.3. Self Attention Mechanisms

Self-attention mechanisms [35, 3] is originally proposed to solve the machine translation which capture global dependencies. Recently, the self-attention is apply in computer vision task and [5, 31] prove self-attention capture more interrelated feature for the task. Futher, [18, 40, 41] present non-local operation for capturing long-range dependencies, and achieve state-of-the-art classification accuracy. And [16] also experiments on object detection and instance segmentation, which also produce high mAP. In this work, we conduct our IE module with self attention, and according to above work, the module have the ability to compute the relation between the feature maps, and filter out the fine feature for OBB.

3. IENet

Most of the object detection method [45, 11, 19] using the downstream image size to fit the feature map size, and the final prediction is constructed by resizing the output prediction. Despite this is a more natural way to solve the detection task, but also come with some drawbacks, which is the large resize error in final prediction. So, most of the method also predicts an offset to reduce the resize error. We found this resize error is more affected in aerial images. Hence, our method based on [34], in which a per-pixel prediction fashion solves the object detection task. The regression points which select in output feature maps are corresponding to a pixel point in an image coordinate. Therefore, in this manner, we can avoid the resize error, which the final predictions already represent the points in images. The general description about our one-stage detection model is illustrate in Figure 2.

In this section, we show our proposed model in detail. We first introduce the representation of the oriented object detection bounding box in Section 3.1. Then, we describe our network architecture in Section 3.2. Futher, we interpret the IE black and self-attention mechanism in Section 3.3. In Section 3.4, we give a construction of the loss function, which is used to train the model. At last, we show also give a detail about the model inference process.

3.1. The Representation of Oriented Bounding Boxes

In our method, each oriented object are represent as $[x_{min}, y_{min}, x_{max}, y_{max}, o]$. In the representation, the $[x_{min}, y_{min}, x_{max}, y_{max}]$ describe the object horizontal bounding boxes, and the parameter $[o]$ represent the orient angle of the object bounding boxes. However, the network has trouble to predict the object in this representation. Therefore, in order to let the network to predict the object accurately, we use a geometric transformation to reconstruct the representation of the OBB object.

As shown in Figures 1 (b), we first reconstruct the HBB follow by [34], which is using a regression point to calculate the offset between the regression point and HBB boundary. Hence, $[l, t, r, b]$ represent as left, top, right, bottom respectively. Than, in Figures 1 (a) we convert the orient angle as $[w, h]$. Therefore, the anlge was split to two different prediction task. In this way the original OBB represent as $[l, t, r, b, w, h]$, which is easier for the network to predict. In this paper, HBB is the extensive box of OBB, and notice we use this box for box regression.

In next section, we show a network architecture which aim to solve the oriented object detection by predict the OBB representation describe in this section.

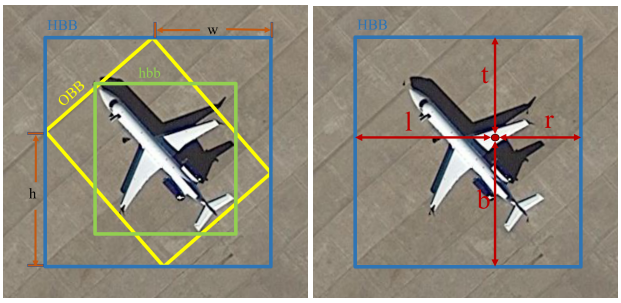

Figure 1. (a) Geometric transformation for OBB to its surrounding HBB, where h and w are the transformation parameters, and hbb is the compact box for object. (b) shows the IENet works by predicting a 4D vector (l, t, r, b) to encode the location of an HBB at each foreground pixel.

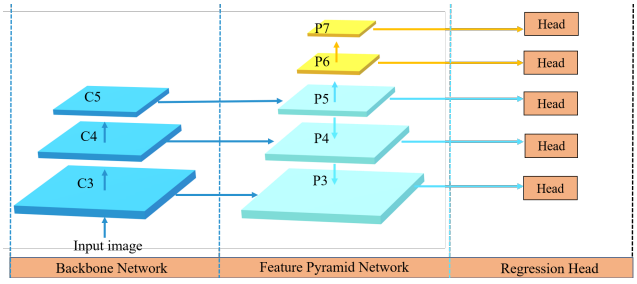

Figure 2. The one-stage detection model, where C3, C4, and C5 denote the feature maps of the backbone network and P3, P4, P5, P6 and P7 are the feature levels used for the final prediction. The input image use a backbone network to extract feature to FPN, and a shared prediction head is use to do the classification, box regression and orientation regression task.

3.2. Network Architecture

Most of the aerial images datasets is lack of the precision and amount. So as describe in [34], a convolution neural backbone network is apply to the network architecture, the backbone network is pretrained on ImageNet [9], and fine tune in our target datasets, which is refer to DOTA and HRSC2016. In this way, the network is able to extract more fine feature from the aerial images.

$$\begin{aligned} x &= \lfloor \tfrac{s}{2} \rfloor + xs, \\ y &= \lfloor \tfrac{s}{2} \rfloor + ys \end{aligned} \tag{1}$$

where $[x, y]$ is the location on the image, and $[xs, ys]$ is the location on the feature. s denotes the number of the stride in feature maps.

The box regression branch predicts the object HBB offset, and this will output 4D vector represent as $[l, t, r, b]$ for each location in the feature map and also corresponding to an image location. The offsets are calculated by:

$$\begin{aligned} l &= x - x_{min}, \quad t = y - y_{min} \\ r &= x_{max} - x, \quad b = y_{max} - y \end{aligned} \tag{2}$$

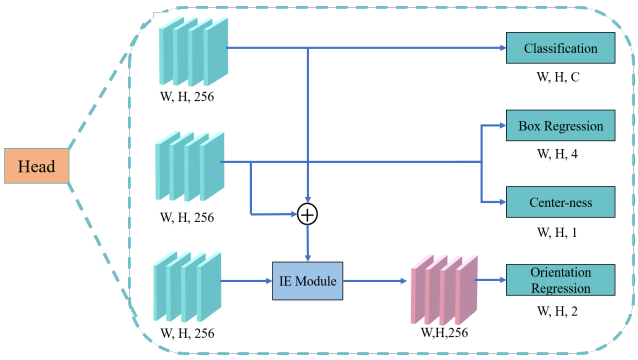

Figure 3. The prediction head contain of IENet. The prediction head contain three separate branchs, each branch for different tasks, which is Classification, HBB regression and orientation regression, respectively. W, H indicate the domestream output sizes from backbone network, C is the number of category, and $\bigoplus$ donate the element-wise addition.

In most case [27, 25, 21], require a new task in the detection model usually by adding a new convolution layer directly on a box regression or classification branch. The box regression is a task which predicts the box boundary offset, and the classification is to recognize the object category. However, both tasks do not have much relation with the angle prediction task, hence only add a new layer to the classification or box regression branch directly, must not work in perspective. Therefore, we add a new branch to regress the 2D vector further represent as $[w, h]$, which also is the parameters that represent the object orientation. Moreover, we also use two convolutional layers to predict $[w, h]$ parameters, respectively. We call this branch the orientation branch, which is the third regression branch on the prediction head. The design of our prediction head are shown in Figure3.

In Figure 3, we use a IE module to extrct feature from the other branchs, and combine them to the orientation feature, to generate the final feature for orientation regression. All branchs are first use four convolutional layers with 256 feature maps output. An addition convolution layer are used to do the prediction.

3.3. IE Black via Self-Attention

To provide more features and elevate the oriented prediction accuracy, similar with [42, 36], we build an interacting embranchment black using a self-attention module to obtain the features that come from both classification and box regression branch, and these features could be rearranging by self-attention mechanisms. The self-attention could establish a relationship between those feature maps and also decide which feature is better for oriented regression. The features will combine with an attention map then add to the orientation branch, as shown in Figures 2. In this way, ori-

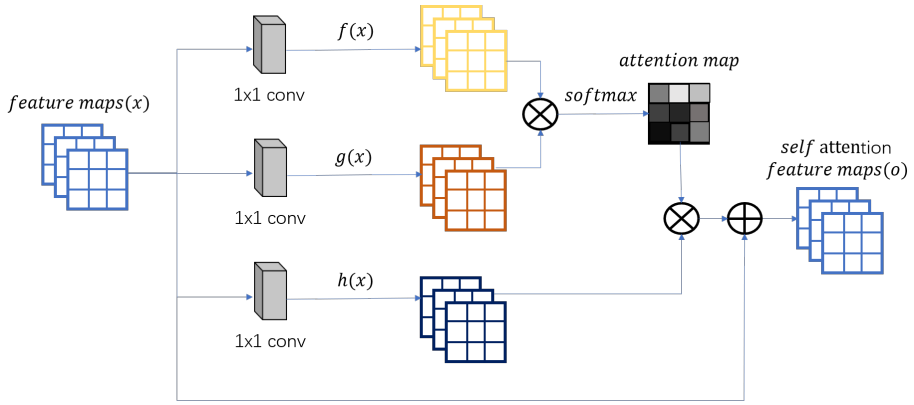

Figure 4. The IE module is constract by self attention mechanism. The $\otimes$ indicate the matrix multiplication, and the $\oplus$ donate the element-wise addition.

entation branch maintains self features for angle prediction task, and also obtain more useful and associative features cames from other branches. As shown in Figures 4, we form the self-attention module by just three 1x1 convolutional and a softmax layer. Then the feature is projected to three feature space by $f(x), g(x), h(x)$ using three different convolutional layers, respectively. $f(x)$ and $g(x)$ together form an attention map via softmax function. Moreover, the attention map indicates the relativity amount of the input features and gives a retroaction on $h(x)$, which presents the original feature maps. The attention maps α is formed by:

$$\alpha = softmax(f\left(x\right)^{T} g\left(x\right)) \tag{3}$$

where $f\left(x\right)^{T} g\left(x\right)$ output a $N \times N$ feature maps s, and N is the number of input feature maps x. A $softmax$ function is apply to each row in s. Therefore,

$$\alpha_{j,i} = \frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{N} \exp\left(s_{ij}\right)} \tag{4}$$

Then the output of the attention layer is $o = (o_1, o_2, \ldots, o_j, \ldots, o_N)$. N donotes the numbers of the inputs and outputs feature maps.

$$o_j = \sum_{i=1}^{N} \alpha_{j,i} h\left(x_i\right) \tag{5}$$

Follow [36], the output of the attention layer is multiplied a scale parameter γ and add back the input feature map, so the self-attention module output is given by:

$$y_i = \gamma o_i + x_i \tag{6}$$

3.4. Loss Function

In order to train the network, we give the loss function, which is calculated over all locations on feature maps:

$$L = \frac{1}{N_{pos}} L_{cls} + \frac{\lambda}{N_{pos}} L_{reg} + \frac{\omega}{N_{pos}} L_{ori} \tag{7}$$

where N_{pos} indicate the number of positive target in groundturth. L_{cls} donote the classification loss calculate by focal loss function [23]. In the loss L_{reg} is calculates by:

$$\begin{aligned} L_{reg} =& BCE(P_{centerness}, G_{centerness}) \\ &+ Smooth_{L1}(P_{ltrb}, G_{ltrb}) \\ &+ (1 - \text{IOU}((P_{ltrb}, G_{ltrb})) \end{aligned} \tag{8}$$

In the regression loss, the centerness loss is constructed follow [34], and a standard binary cross entropy loss is employed to centerness. The purpose of the centerness loss is to encourage the network to choose a regression point that closes to the target center point. Furthermore, the centerness also affects the confidence of the predicted object. P_{ltrb}, G_{ltrb} indicate the prediction and ground truth HBB. The prediction of the HBB comes from the regression branch, and we further use IOU to calculate the IOU loss between HBB. L_{ori} is constructed by:

$$\begin{aligned} L_{ori} =& Smooth_{L1}(P_{wh}, G_{wh}) \\ &+ (1 - \text{IOU}(P_{wh}, G_{wh}, P_{ltrb}, G_{ltrb})) \end{aligned} \tag{9}$$

P_{wh}, G_{wh} denotes the prediction from the orientation branch and the ground truth parameters. Same as Equation 8, we also use Smooth-L1 loss and IOU loss. Moreover, we compute the IOU loss using OBB. Thus the parameters $[l, t, r, b]$ and $[w, h]$ combine to transform HBB to OBB. To compute the OBB IOU is too computational in the training process, we form a different version of the IOU for OBB, which is the inner box calculate by:

$$\begin{aligned} l_n &= |l - w|, \quad t_n = |t - h|, \\ r_n &= |r - w|, \quad b_n = |b - h| \end{aligned} \tag{10}$$

where $[l_n, t_n, r_n, b_n]$ donate the inner box offset. Therefore, we have the groundtruth and predict inner box offsets. Further a simple version of IOU on OBB can be calculated using the offsets.

3.5. Inference

In this section, we explain the inference in our model. Given an input image, the backbone network generated Ns feature maps, and we use the last three layers output as the input of the FPN. The FPN fuse three feature maps and generate the final feature maps for prediction head. The prediction head contains three branches, and each branch is designed to fulfill different tasks, hence classification branch for classifying task, box regression branch for bounding box prediction task, and orientation branch for predict the orientation parameters task. The prediction head is shared amount the three feature maps. Each branch produces the prediction map same size as the feature maps generated by backbone network, therefore the location on each prediction map $P_{x,y}$ can be projected to a location on the image by Equation 1, for each location we select those which classification confidence is higher than 0.5 as a definite prediction. However, we use center-ness prediction to multiply the category prediction, so we set the threshold to 0.05. In the end, the model predicts a 4D vectors $[l, t, r, b]$ and a 2D vectors $[w, h]$, and then these parameters could be transformed to OBB refer to Section3.1.

4. Experiments

We evaluated our proposed IENet on the challenge datasets DOTA and HRSC2016. Both datasets include a mass of object which express in arbitrary oriented. The datasets description are shown as follow:

- **DOTA** The dataset contain 2806 high resolution images with 15 categories. The DOTA images contain 188, 282 instances, and the instances in this data set vary greatly in scale, orientation, and aspect ratio.
- **HRSC2016** The dataset contain 1061 images with 29 categories. The images size in HRSC2016 is various. The images size ranges from 300×300 to 500×500.

The average precision (AP) over categories is employed to the above datasets as the measurement to characterize the performance of detectors. Some selected results from proposed IENet are shown in Fig.4.1 for interest.

4.1. Trainin Details

Datasets Setting In this work, follow [10] all images from datasets are cropped to 1024×1024 pixels for memory efficiency, and for data augmentation, we resize the images at scales(1.0, 0.5), we also apply random flip and random rotade from (0, 90, 180, 270) to avoid an imbalance between the categories in datasets. And these setting are used both in training and texting.

Network Setting ResNet-101 [15] is used as the backbone networks in all the results. Batch size is set to be 16 and the learning rate is initialised with 0.01, we use stochastic gradient descent (SGD) [4] for 100K iterations and weight decay and momentum are set as 0.0001 and 0.9. The learning rate is reduced by a factor of 10 at the end of the last 20K training step (80k-100K).

4.2. Comparisons with State-of-the-art methods

A one-stage orientation detector baseline method built on FCOS, which is inspired by one of the state-of-the-art one stage detectors [34]. We modified the regression head at the end of the network to enable FCOS to directly regress orientation parameters h and w, which are predicted by adding a convolutional layer in the regression branch, hence it can be used for OBB detection. We also compare our method with two published top performance two-stage orientation detectors, RoI Transformer [10] and Faster R-CNN OBB detector [37]. The comparisons using DOTA and HRSC2016 datasets are shown in Table.1 and Table.2, respectively.

The result show that, our method outperforms the baseline method according to mAP measurement by 9.75% on DOTA and 6.45% on HRSC2016 datasets, respectively. When compared with two-stage detectors, our proposed IENet beats the FR-O method by 3.01%. Although it is hard to exceed the performance of the RoI Transformer detector [10], IENet still works better on 5 of 15 categories of DOTA dataset with fewer network parameters and less computational complexity.

To evaluate the efficiency of IENet, we train the networks on eight GTX 1080Ti (12GB) GPUs, The trade-off between accuracy and speed is shown in Table3.

The comparison in Table.3 shows that, when compared with the anchor-free one-stage detector, IENet can achieve great improvement on accuracy while maintain low complexity and small model. When compared with the two stage detectors, although IENet is not always dominant in accuracy, it has advantages of efficiency and lightweight model.

4.3. IENet Ablation Studies

We experiments the contribution of our method, which is geometric transformation and the IE module, and we study the effect of the DCN [8] on our model. FCOS is motify to predict the OBB, which is directly regress the unroted bounding box of OBB and a angle parameter.

Geometric transformation. We train FCOS with DCN and directly regress the OBB which predict five parameters. Our model divided the angle into two parameters $[w, h]$, which we have six parameters to predict. In Table 4 2nd and 3rd entries, even withour the DCN our geometric transformation still outperform the FCOS by 6.99%.

Interacting Embranchment Module. In Table 4 3rd and 4th entries, we show our with the aid of DCN, our IE

Table 1. Numerical results (AP) comparisons with state-of-the-art methods on DOTA. The short names for categories are defined as: BD-Baseball diamond, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, TC-Tennis court, BC-Basketball court, SC-Storage tank, SBF-Soccer-ball field, RA-Roundabout, SP-Swimming pool, and HC-Helicopter. RoI Trans means method with RoI transfer in [**?**], FR-O means orientation Faster-RCNN model and IE is our IENet model.

methods	mAP	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harber	SP	HC
RoI Trans	**66.95**	**88.02**	76.99	36.70	**72.54**	**70.15**	61.79	**75.77**	**90.14**	**73.81**	85.04	**56.57**	**62.63**	53.30	59.54	41.91
FR-O	54.13	79.42	**77.13**	17.70	64.05	35.30	38.02	37.16	89 .41	69.64	59.28	50.30	52.91	47.89	47.40	**46.30**
Baseline	47.39	79.32	58.02	23.39	32.36	33.31	34.86	35.59	64.55	42.33	78.16	43.41	55.60	39.70	53.62	36.70
IE(ours)	57.14	80.20	64.54	**39.82**	32.07	49.71	**65.01**	52.58	81.45	44.66	**78.51**	46.54	56.73	**64.40**	**64.24**	36.75

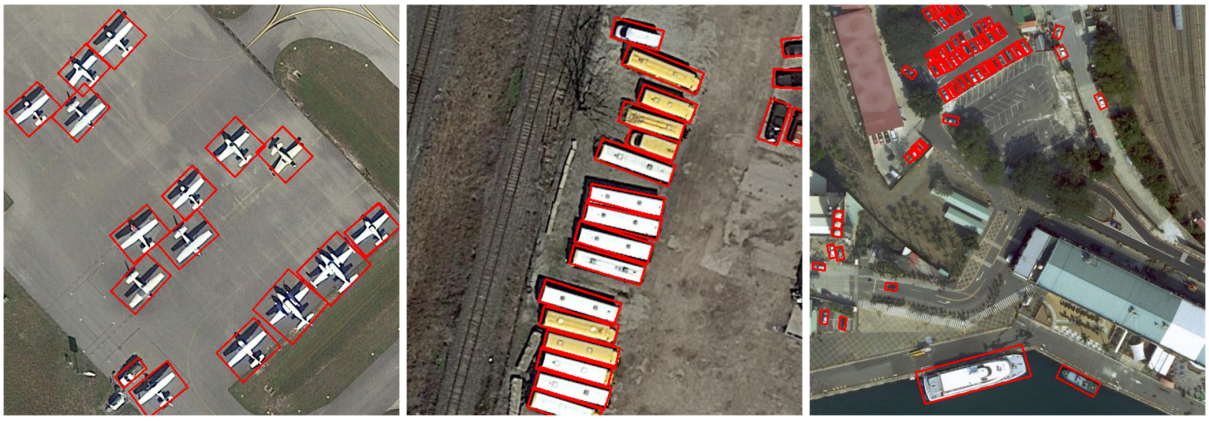

Figure 5. Selected results for DOTA datasets from proposed IENet, targets are presented by red orientation rectangle, where (a) is results for planes, (b) is for large vehicles and (c) is for densely parked small vehicles together with ships.

Table 2. mAP comparisons with state-of-the-art methods on HRSC2016, where RoI Trans means method with RoI transfer in [10], FR-O means orientation Faster R-CNN method and IENet-baseline is baseline method.

methods	RoI Trans	Baseline(ours)	IENet(ours)
mAP	86.20	68.56	75.01

Table 3. Speed-accuracy trade-off comparison for IENet, where RoI Trans means method with RoI transfer in [10], FR-O means orientation Faster-RCNN model and the IE is our IENet model. Tr-time denotes time for training, inf-time denotes inference time and Params denotes the total parameter numbers.

methods	mAP	tr-time(s)	inf-time(s)	params(MB)
RoI Tran	66.95	0.236	0.084	273
FR-O	54.13	0.221	0.102	270
Baseline	47.39	0.109	0.056	208
IE(ours)	57.14	0.111	0.059	212

module could help the model increase the sores by 2.6%. However, in Table 4 4th and 5th entries, the DCN has only a slightly improvement on mAP, which is increase by 0.16%. We believe the results indicate that our IE module already extract the befitting feature for the final prediction. Therefore with or without the DCN, our model still obtain a nearly equal result with our geometric transformation and IE module.

Our method use a brand new way to predict the OBB,

Table 4. Ablative experiments of the IE module on the DOTA. We use ResNet-101 for all the experiments in the table. The experiments study the improvement of our propose geometric transformation and IE module. Geo trans donate the geometric transformation.

	DCN	Geo trans	IE Module	mAP
FCOS	✓			47.39
Ours		✓		54.38
		✓	✓	56.98
	✓	✓	✓	57.14

which is geometric transformation on OBB split the anlge to two split parameters. Therefore, without optimizing the hyper-parameters, our performence a achieve better results compare with state-of-the-art RoI Transformer. We believe with proper optimizing, our IENet can obtain high numerical of mAP as the state-of-the-art.

5. Conclusion

In this paper, a one-stage orientation detector, IENet, was presented which was an anchor free solution for predicting OBB. IENet was presented for oriented target detection in aerial images. A one-stage anchor-free keypoint based architecture was employed and a novel rotation prediction method was proposed following a geometric transformation. Moreover, an IE module based on a self atten-

tion mechanism was used as feature interacting module to combine features for orientation prediction. Comparison results showed the improvement of accuracy from our proposed IENet and because of the interacting behavior; IENet was proved to be more computationally efficient when compared with the state-of-the-art orientation detectors, and the efficiency of our geometric transformation and IE module were proved to obtain high performence; In future work, we seek to another feature interacting method other than self attention mechanism to extract majestic feature for OBB, and to achieve state-of-the-art result when compared with detectors.

References

[1] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, pages 150–165. Springer, 2018.

[2] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[5] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005.

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for detecting oriented objects in aerial images. *arXiv preprint arXiv:1812.00155*, 2018.

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019.

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv preprint arXiv:1907.12273*, 2019.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[18] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[19] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[20] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4):985–996, 2017.

[21] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018.

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[25] Yuliang Liu, Lianwen Jin, Zecheng Xie, Canjie Luo, Shuaitao Zhang, and Lele Xie. Tightness-aware evaluation protocol for scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9612–9620, 2019.

[26] Yudong Liu, Yongtao Wang, Siwei Wang, TingTing Liang, Qijie Zhao, Zhi Tang, and Haibin Ling. Cbnet: A novel

composite backbone network architecture for object detection, 2019.

[27] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. *arXiv preprint arXiv:1906.02371*, 2019.

[28] Zikun Liu, Jingao Hu, Lubin Weng, and Yiping Yang. Rotated region based cnn for ship detection. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 900–904. IEEE, 2017.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[30] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018.

[31] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[34] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[37] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[38] Enze Xie, Yuhang Zang, Shuai Shao, Gang Yu, Cong Yao, and Guangyao Li. Scene text detection with supervised pyramid context network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9038–9045, 2019.

[39] Linjie Xing, Zhi Tian, Weilin Huang, and Matthew R. Scott. Convolutional character networks, 2019.

[40] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[41] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[42] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[43] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1745–1749, 2018.

[44] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9259–9266, 2019.

[45] Xingyi Zhou, Dequan Wang, and Philipp Krhenbhl. Objects as points, 2019.