Anchor Retouching via Model Interaction for Robust Object Detection in Aerial Images

Dong Liang, *Member, IEEE*, Qixiang Geng, Zongqi Wei, Dmitry A. Vorontsov, Ekaterina L. Kim, Mingqiang Wei, *Senior Member, IEEE*, and Huiyu Zhou

***Abstract*— Object detection has made tremendous strides in computer vision. Small object detection with appearance degradation is a prominent challenge, especially for aerial observations. To collect sufficient positive/negative samples for heuristic training, most object detectors preset region anchors in order to calculate intersection-over-union (IoU) against the ground-truth data. In this case, small objects are frequently abandoned or mislabeled. In this article, we present an effective dynamic enhancement anchor network (DEA-Net) to construct a novel training sample generator. Different from the other state-of-the-art (SOTA) techniques, the proposed network leverages a sample discriminator to realize interactive sample screening between an anchor-based unit and an anchor-free unit to generate eligible samples. Besides, multi-task joint training with a conservative anchor-based inference scheme enhances the performance of the proposed model while reducing computational complexity. The proposed scheme supports both oriented and horizontal object detection tasks. Extensive experiments on two challenging aerial benchmarks (i.e., Dataset of Object deTection in Aerial images (DOTA) and HRSC2016) indicate that our method achieves SOTA performance in accuracy with moderate inference speed and computational overhead for training. On DOTA, our DEA-Net which integrated with the baseline of RoI-transformer surpasses the advanced method by 0.40% mean-average-precision (mAP) for oriented object detection with a weaker backbone network (ResNet-101 vs. ResNet-152) and 3.08% mAP for horizontal object detection with the same backbone. Besides, our DEA-Net which integrated with the baseline of ReDet achieves the SOTA performance by 80.37%. On HRSC2016, it surpasses the previous best model by 1.1% using only three horizontal anchors. The source code and the training set are made publicly available at https://github.com/QxGeng/DEA-Net.**

Manuscript received August 1, 2021; revised November 2, 2021; accepted December 10, 2021. Date of publication December 16, 2021; date of current version March 29, 2022. This work was supported in part by the AI+ Project of the Nanjing University of Aeronautics and Astronautics under Grant XZA20003; in part by the Natural Science Foundation of China under Grant 62172218, Grant 61772268, and Grant 62172212; and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20190065. *(Corresponding author: Dong Liang.)*

Dong Liang, Qixiang Geng, Zongqi Wei, and Mingqiang Wei are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, and also with the MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China (e-mail: liangdong@nuaa.edu.cn; gengqx@nuaa.edu.cn; weizongqi@nuaa.edu.cn; mingqiangw@nuaa.edu.cn).

Dmitry A. Vorontsov and Ekaterina L. Kim are with the Faculty of Physics, National Research Lobachevsky State University of Nizhny Novgorod, 603950 Nizhny Novgorod, Russia (e-mail: vorontsovda@mail.ru; kim@phys.unn.ru).

Huiyu Zhou is with the School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: hz143@leicester.ac.uk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TGRS.2021.3136350, provided by the authors.

Digital Object Identifier 10.1109/TGRS.2021.3136350

***Index Terms*— Aerial observation, dynamic enhancement anchor (DEA), object detection.**

I. Introduction

OBJECT detection is one of the fundamental and challenging problems in computer vision. Tremendous successes have been achieved on object detection with the development of deep convolution neural networks (DCNNs) in recent years. Different from the objects in natural scenes which are often captured from horizontal perspectives, aerial images are typically taken from a bird's eye view at a high altitude, suggesting that objects in aerial images usually are of a small size and diverse orientations with complex background [1]. A large number of detectors [2]–[5] have been designed for aerial observations, most of which are based on a two-stage detector (e.g., fast R-CNN [6] and faster R-CNN [7]) or a one-stage detector (e.g., RetinaNet [8] and you only look once (YOLO) [9]).

Region anchors are designed as the regression references and the classification candidates to predict the proposals in two-stage detectors or final bounding boxes in one-stage detectors. Most of the anchor-based detectors utilize a uniform anchoring scheme, and then positive and negative samples are selected through intersection-over-union (IoU) with ground truth. For example, the anchor boxes with IoU $>$ 0.5 are treated as positive samples and IoU $<$ 0.3 as negative samples [7]. In practice, such a strategy may cause two main problems.

1) *Anchor quantization errors and noisy training samples:* We take faster-RCNN [7] as an example. If the base anchor size is set to 32 and the IoU threshold is set to 0.5, objects with area $<32^2 \times 0.5$ (512 pixels) will be excluded from the positive proposals. As shown in Fig. 1(a), the anchor box (red) and the ground truth (green) have a large quantization discrepancy. This discrepancy will lead to much confusion for both box localization and classification. On the other hand, empirical evidence shows that objects with an area $<$512 pixels occupy approximately 30% of an aerial image of Dataset of Object deTection in Aerial images (DOTA) [1], where inaccurate anchor boxes or misclassification of the samples lead to unstable convergence of the model.
2) *Mismatch between the pyramid levels and the samples:* This is based on the consensus that upper feature maps have more semantic information suitable for detect-

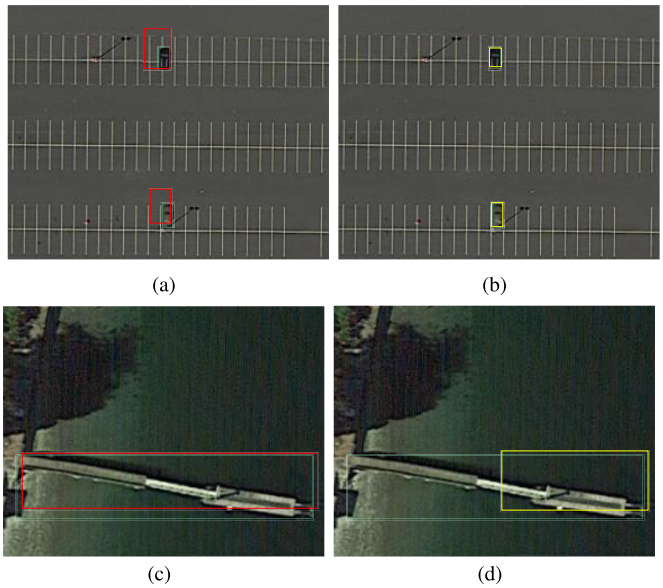

Fig. 1. Comparison of anchor proposals. (a) and (c) Anchor-based method (e.g., faster-RCNN [7]) and regression bounding boxes. (b) and (d) Anchor-free method (e.g., FCOS [10]). Anchor-free regression bounding boxes (yellow boxes) have better consistency in small object detection than the anchor proposals (red boxes), but for large objects with large aspect ratios, the anchor proposals have better consistency. Green boxes are the ground truth.

ing big instances, whereas lower feature maps have more fine-grained details suitable for detecting small instances. Integrated within a feature pyramid, large anchor proposals are typically associated with upper feature maps, and small anchor proposals are associated with lower feature maps. Inaccurate bounding boxes with large background areas would cause a mismatch between the feature pyramid levels and the training samples, largely affecting the model training. In other words, the selected feature level to train each sample may not be correct.

To deal with these issues, image feature pyramids with more levels can be used to better detect small objects. Another common solution is to enlarge the quantity of the anchors with diverse sizes and aspect ratios. These two solutions have evident drawbacks—both of them lead to significant computational overhead, especially when processing large-scale aerial images or training the network with a heavy backbone.

As shown in Fig. 1(b), the regression bounding boxes in an anchor-free detector (e.g., fully convolutional one-Stage object detection (FCOS) [10]) can be potentially leveraged as positive region proposals because they are free from anchor quantization errors. On the other hand, compared with the anchor-based detectors, the anchor-free detectors usually fail to generate an accurate bounding box when the objects are of a large size and an extreme aspect ratio [10]–[12], just like the example shown in Fig. 1(d). Most anchor-based detectors (including the baseline faster R-CNN) regress from the anchor box with four offsets between the anchor box and the object box, while FCOS regresses from one point with four distances to the bound of the object. It means that for a positive sample, the regression's starting status of faster R-CNN is a box, while FCOS is a point. The box itself contains prior of the shape,

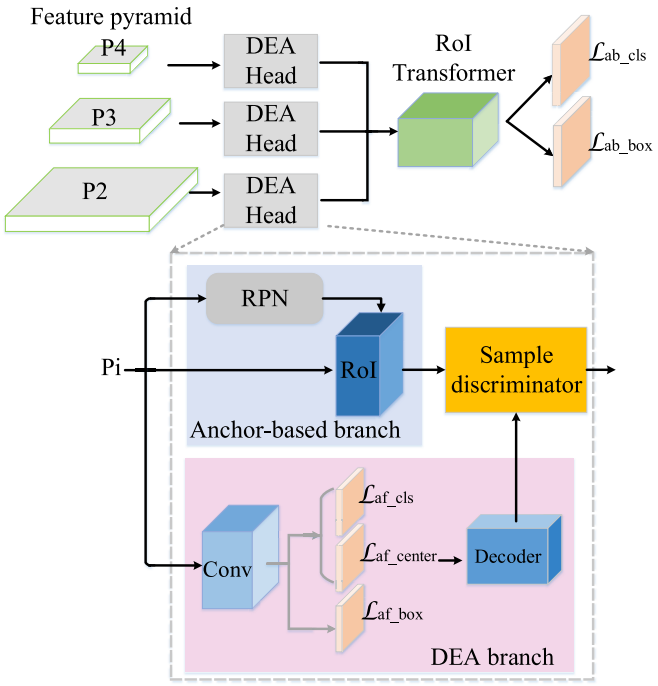

Fig. 2. Architecture of our proposed DEA-Net for oriented object detection. The DEA head serves each level of the feature pyramid to generate higher-quality training samples, including an anchor-based model, an anchor-free model, and a sample discriminator. RoI transformer [2] processes the horizontal and the oriented proposals.

and the regression is only the two small offset of X and Y directions. In contrast, FCOS needs to independently return the offsets of $+/-X$ and Y, four directions, from a start point, without any shape prior. The regression error in any direction would greatly affect the shape of the box, especially for large objects with a large aspect ratio. This observation has been reported in [13].

For the anchor-based approaches, the anchor quantization errors can be ignored for large objects. The anchor boxes are designed to discretize all possible instance boxes into a finite number of boxes with predefined locations, scales, and aspect ratios. We need more anchors with a smaller size and denser layouts or more angles in arbitrary-oriented detection to cover small objects, which may lead to extensive computation cost and imbalanced problems of positive and negative samples. Achieving spatial alignment with small ground-truth objects is challenging and prone to the miss of the corresponding positive anchors based on this strategy.

Inspired by the above observations, in this article, we propose an effective dynamic enhancement anchor network (DEA-Net) to enhance the learning of small objects efficiently. The overall architecture is shown in Fig. 2. The DEA head serves each level of the feature pyramid consisting of an anchor-based module, an anchor-free module, and a sample discriminator. The sample discriminator merges the complementary anchor-based and the anchor-free proposals and generates representative and informative samples with accurate locations and sizes, while avoiding positive/negative confusion. Besides, multi-task joint training with a conservative anchor-based inference scheme enhances the performance of the model while avoiding complexity augmentation.

We conduct extensive experiments on both oriented and horizontal object detection tasks. Experiments on the aerial image benchmarks DOTA [1] and HRSC2016 [14] show that our proposed DEA-Net makes substantial improvement, compared to the baseline methods, and achieves state-of-the-art (SOTA) performance in accuracy [i.e., 80.37% mean-average-precision (mAP) (+0.14%) and 90.56% mAP (+1.10%)] for oriented object detection tasks. Besides, experiments on DOTA [1] for horizontal object detection achieves SOTA performance in accuracy with 78.43% mAP (+3.08%). By combining the anchor-based and anchor-free branch efficiently, our method maintains a fair inference speed and training computational overhead.

To the best of our knowledge, this is the first time to simultaneously consider the impact of both the object's scale and aspect ratio and then distinguish and process them separately for training. In summary, our main contributions consist of: 1) an effective sample generator based on DEA head to enhance the performance of detecting small objects by combining the advantages of anchor-based and anchor-free models and 2) a novel and robust DEA-Net, which can achieve the start-of-the-art oriented and horizontal object detection performance in aerial images.

The remainder of this article is organized as follows. In Section II, we discuss the related work. In Section III, we describe the proposed method in detail. The experimental results are presented and discussed in Section IV, and the conclusions and future work are given in Section V.

II. Related Work

A. Anchor-Based and Anchor-Free Models

The current mainstream detectors can be divided into two categories: 1) anchor-based methods [6]–[9] and 2) anchor-free methods [10]–[12]. In anchor-based methods, the network is trained to regress the offsets between the anchors and ground-truth bounding boxes. However, these methods take advantages of the task-oriented settings of anchors, leading to complex parameter tuning. Moreover, since the scales and aspect ratios of anchors are fixed, it has the difficulty to handle the objects with large shape variations, especially for small objects. Anchor-free methods directly regress the bounding box without using preset anchors. They usually have a streamlined network structure due to discarding of the dense anchors. However, they also meet difficulties in learning large variations of the bounding boxes without prior knowledge. DEtection TRansformer (DETR) [15] utilizes self-attention to build a novel detection architecture, whose detection precision can compete against those of the two-stage object detectors, but it has the weakness in detecting small objects with high computational overheads in the published literature. Due to the dilemma of the above methods, an emerging line of work attempts to design a detector by combining anchor-based and anchor-free methods. GA-RPN [16] constructs a region proposal network in an anchor-free manner to predict the proposals for faster R-CNN. Feature selective anchor-free module (FSAF) [17] attaches an anchor-free module at each feature pyramid level to select appropriate features of each object for RetinaNet. SFace [18] attaches an anchor-free model to an anchor-based detector and combines the outputs of two models to improve the performance of the detector. Different from the other state of the arts such as [16]–[18], we focus on the collaboration of anchor-based and anchor-free methods from the perspective of sample discrimination. It makes interactive sample screening invulnerable to the diversity of the scale distributions.

B. Improved Anchor-Based Detection Models

The recent improvement of the anchor selection strategies mainly focuses on two aspects as follows.

1) Weight the predicted anchor boxes to distinguish the potential importance and quality differences. MetaAnchor [19] and soft anchor-point object detection (SAPD) [20] belong to this type. MetaAnchor directly weights the generated anchors, while SAPD leverages both soft-weighted anchor points and soft-selected pyramid levels.
2) Propose a refining anchor box assignment strategy. For example, dynamic anchor feature selection (DAFS) [21] uses an anchor refinement module (ARM) to adjust the locations and sizes of anchors and filter out negative anchors and then select new pixels in a feature map for each refined anchor.

Ming *et al.* [4] define the dynamic anchor with a matching degree to evaluate both spatial and feature alignment for anchor assignment. Nevertheless, all the above methods ignore the influence of the object's scale and aspect ratio on anchor assignment. In remote-sensing scenarios, for example, we observe that small objects are frequently abandoned or mislabeled due to the predefined anchor sampling interval and the IoU rule, which could potentially destroy the original sample distribution in the feature space. On the other hand, anchor-free-based detectors often fail to generate an accurate bounding box for large objects with a large aspect ratio. Different from [4], [19]–[21], the significant differences of our method are: 1) consider the impact of both the object's scale and aspect ratio simultaneously and 2) distinguish between different scales and aspect ratios and then use appropriate strategies to process them separately. The anchor-free module is utilized to generate more positive samples of small objects which are ignored in the anchor-based module. For some objects of a large size and extreme aspect ratios, we preserve the anchors in the anchor-based model which have higher IoUs as the positive samples. Compared with the existing methods (weighting the anchor [19], [20] or refinement assignment strategy [4], [21]), our idea has also been experimentally proved to be effective. In particular, it is more suitable for object detection in remote-sensing images, because in remote-sensing images, small-sized (such as vehicles) and large-sized objects with extreme aspect ratios (such as ports, bridges) are very common. The work adaptive training sample selection (ATSS) [22] is a comparative analysis of the Retinanet and FCOS, which proposes a general training strategy to serve them separately, so this method naturally does not increase any overhead. In contrast, our method involves model design,

a training scheme (relies on a sample discriminator for interactive sample screening and with multi-task joint training), and an inference scheme (a conservative anchor-based scheme to freezing the anchor-free branch to suppress computational complexity). Such a comprehensive scheme improves the performance in remote-sensing scenarios.

C. Object Detection in Aerial Images

Object detection in aerial images often faces a large number of small objects with arbitrary orientations in complex environments. Detecting objects with oriented bounding boxes (OBBs) is a non-trivial extension of horizontal object detection, which are mostly built on anchor-based detectors [2], [23]–[25]. R^3Det [26] adopts cascade regression to refine the predicted boxes. DCL [5] utilizes a densely coded label encoding mechanism for angle classification. SCRDet [3] improves the performance of small objects by reducing the anchor strides to preset smaller and more anchors, which incurs extensive computational costs. Dynamic anchor learning (DAL) [4] utilizes a comprehensive scheme for spatial alignment, feature alignment ability, and regression uncertainty for label assignment. RoI transformer [2] applies spatial transformations on RoIs and learn the transformation parameters under the supervision of OBB annotations, which is with lightweight and can be easily embedded into detectors for oriented object detection. In our work, our method is based on the RoI transformer to deal with OBB, and we constrain random discarding and positive/negative confusion of small objects and produce qualified training samples with accurate locations and scopes without introducing complicated modules. Our experiments also confirm that it is unnecessary to preset a large number of specially designed anchors with large computational overheads.

III. Proposed Method

In this section, we introduce the technical details of our proposed DEA-Net and instantiate our DEA module by showing how to apply the scheme to the object detectors with a feature pyramid for object detection in aerial images. Specifically, we first introduce the details of our proposed DEA module and then introduce the sample discriminator which facilitates the learning of small objects. Then, we show the details of our overall network architecture. Finally, we show how to join training with inference of our DEA-Net.

A. Dynamic Enhancement Anchor

In the literature, anchor-based methods find it difficult to fully learn small objects by selecting positive and negative samples for training through the examination of IoU overlap. As shown in Fig. 3, in the anchor-based module, if the IoU between the preset anchors and the ground-truth boxes of small objects is lower than the threshold of the positive samples, the samples are treated as discarded samples (i.e., -1 as sample label) or negative samples (i.e., 0 as sample label). The training of anchor-based detectors for small objects is not sufficient because of the lack of positive samples, severely affecting the detection performance.

As aforementioned, the regression bounding boxes in the anchor-free detector usually have higher IoU for the small objects than the anchor-based detector. However, for some objects of a large size and extreme aspect ratios, anchor-free detectors may have poor performance. Therefore, instead of utilizing the anchor-free method to replace the anchor-based method to generate samples for training, we combine their advantages to deal with positive sample selection of different scales.

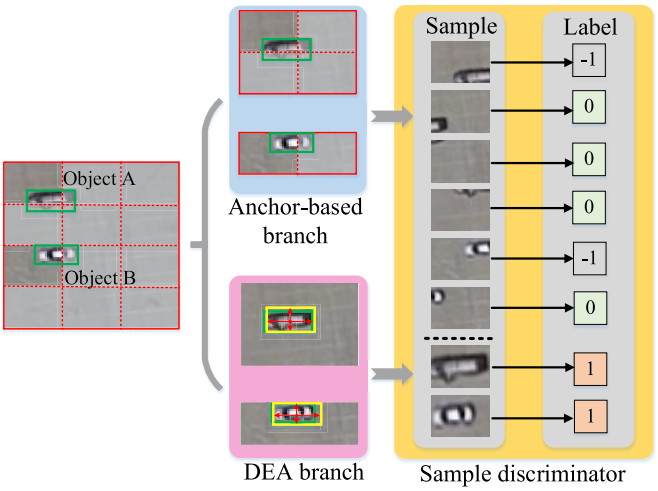

Fig. 3. DEA module to provide better positive samples of small objects. Specifically, in the original anchor-based module, the preset anchors are assigned as negative samples (0) or discarded samples (-1). However, our DEA module can generate high-quality bounding boxes of higher IoU, assigned as positive samples (1) for better training the network.

We construct an anchor-free sample generation module that shares the feature pyramid with the anchor-based module and integrate it with an anchor-based detector. Via an interactive sample screening procedure in the sample discriminator, the anchor-free module is utilized to generate more positive samples of small objects which are ignored in the anchor-based module. For some objects of a large size and extreme aspect ratios, we preserve the anchors in the anchor-based module which have higher IoUs as the positive samples.

B. Interactive Sample Screening

Current studies [10], [18] have reported that the designed anchor boxes are the key to successful anchor-based detectors, and the detection performance is sensitive to the size, aspect ratio, and the number of the anchor boxes. Therefore, the anchors in these anchor-based detectors must be carefully tuned for each specific task on different datasets. For example, to deal with the challenge of small object detection, one needs to design smaller anchors beforehand and densely locate them on the input image. This handling leads to extensive computational costs and the imbalanced problem. Therefore, our proposed dynamic enhancement method aims to use the preset number of the preset anchors to improve the detection performance of small objects with less computational cost. In our solution, the DEA head serves each level of the feature pyramid consisting of an anchor-based branch, an anchor-free branch (DEA branch), and a sample discriminator to realize interactive sample screening. More detailed, the sample

Algorithm 1 Sample Discriminator

Input:

$\mathcal{G}$ is the set of ground-truth boxes
$\mathcal{P}$ is the set of feature pyramid levels
$\mathcal{A}$ is the set of anchor boxes of the RPN outputs
$\mathcal{V}$ is the set of predicted vector of anchor-free branch
$\mathcal{T}_{\mathcal{P}}$ is the threshold of positive samples
$\mathcal{T}_{\mathcal{N}}$ is the threshold of negative samples

Output:

$\mathcal{S}_{\mathcal{E}}$ is the set of enhancement samples
$\mathcal{S}_{\mathcal{P}}$ is the set of positive samples
$\mathcal{S}_{\mathcal{N}}$ is the set of negative samples

for each ground-truth box $g \in \mathcal{G}$ **do**
  for each feature pyramid level $p_i \in \mathcal{P}$ **do**
    decoder predicted vector $\mathcal{V}$ to bounding-boxes $\mathcal{B}$: $\mathcal{B} = Decoder(\mathcal{V})$;
  end for
  calculate the Intersection-over-Union (IoU) between g and $b_j \in \mathcal{B}$: $\mathcal{IB}_g = IoU(b_j, g)$;
  calculate the Intersection-over-Union (IoU) between g and $a_i \in \mathcal{A}$: $\mathcal{IA}_g = IoU(a_i, g)$;
  if $\mathcal{IB}_g^j \geq \mathcal{T}_{\mathcal{P}}$ and $\mathcal{IB}_g^j \geq \mathcal{IA}_g^i$ **then**
    $\mathcal{S}_{\mathcal{E}} = \mathcal{S}_{\mathcal{E}} \bigcup \mathcal{B}_g^j$
  else if $\mathcal{IA}_g^i \geq \mathcal{T}_{\mathcal{P}}$ and $\mathcal{IA}_g^i \geq \mathcal{IB}_g^j$ **then**
    $\mathcal{S}_{\mathcal{P}} = \mathcal{S}_{\mathcal{P}} \bigcup \mathcal{A}_g^i$
  else if $\mathcal{IA}_g^i \leq \mathcal{T}_{\mathcal{N}}$ **then**
    $\mathcal{S}_{\mathcal{N}} = \mathcal{S}_{\mathcal{N}} \bigcup \mathcal{A}_g^i$
  end if
end for
$\mathcal{S}_{\mathcal{P}} = \mathcal{S}_{\mathcal{P}} \bigcup \mathcal{S}_{\mathcal{E}}$
return $\mathcal{S}_{\mathcal{E}}$, $\mathcal{S}_{\mathcal{P}}$, $\mathcal{S}_{\mathcal{N}}$;

discriminator merges the complementary anchor-based and the anchor-free proposals and generates representative and informative samples with accurate locations and sizes, while avoiding positive/negative confusion.

Algorithm 1 describes how the proposed sample discriminator works with an input image. For each ground-truth box $g = [x, y, w, h, c]$ on the input image, where (x, y) is the left-top corner of the box, (w, h) are the box width and height, respectively, and c is the class label, our anchor-free branch will generate prediction vectors

$$\mathcal{V} = [v_t^{m,n}, v_l^{m,n}, v_b^{m,n}, v_r^{m,n}, c^{m,n}] \tag{1}$$

where $v_t^{m,n}$, $v_l^{m,n}$, $v_b^{m,n}$, and $v_r^{m,n}$ are the distances between the current pixel location (m, n) and the top, left, bottom, and right boundaries of the box, respectively, and $c^{m,n}$ is the prediction class label. We first decode the prediction vectors $\mathcal{V}$ to form the bounding boxes $\mathcal{B}$

$$\mathcal{B} = [x^{m,n}, y^{m,n}, w^{m,n}, h^{m,n}, c^{m,n}] \tag{2}$$

where $x^{m,n} = m - v_l^{m,n}$, $y^{m,n} = n - v_t^{m,n}$, $w = v_l^{m,n} + v_r^{m,n}$, and $h = v_t^{m,n} + v_b^{m,n}$.

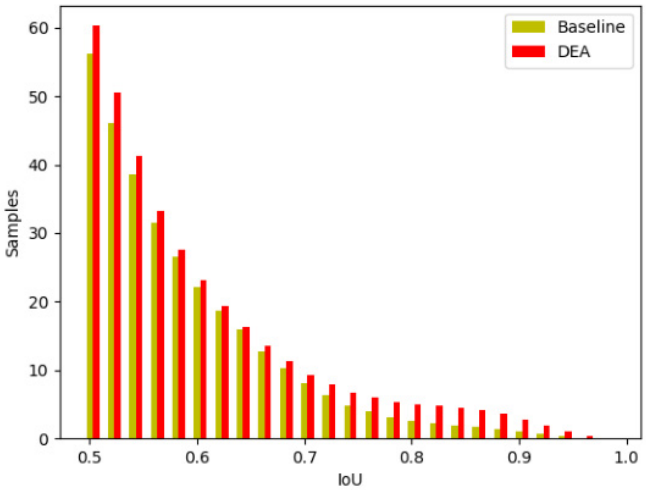

Fig. 4. IoU distributions of the baseline (faster R-CNN [7]) and our DEA-net. The x-axis represents the IoU between the samples and the ground-truth boxes. The y-axis represents the average number of samples on each image.

Then, we calculate the IoU between the regressed bounding boxes $\mathcal{B}$ and the ground-truth g, labeled as $\mathcal{IB}_g$, and the IoU between the anchors of the RPN output in the anchor-based module $\mathcal{A}$ and the ground-truth g, labeled as $\mathcal{IA}_g$. Afterward, we select samples as follows:

$$\left(\mathcal{IB}_g^j \geq \mathcal{IA}_g^i\right) \cap \left(\mathcal{IB}_g^j \geq \mathcal{T}_{\mathcal{P}}\right) \tag{3}$$

where $\mathcal{T}_{\mathcal{P}}$ is the threshold of the positive samples (i.e., 0.5 in this article). It denotes that the bounding box in the anchor-free model have better consistency than the anchor proposal, and we assign box $\mathcal{B}_g^j$ to the enhanced samples $\mathcal{S}_{\mathcal{E}}$ with

$$\left(\mathcal{IA}_g^i \geq \mathcal{T}_{\mathcal{P}}\right) \cap \left(\mathcal{IA}_g^i \geq \mathcal{IB}_g^j\right). \tag{4}$$

We assign anchor $\mathcal{A}_g^i$ as positive samples $\mathcal{S}_{\mathcal{P}}$. If we have

$$\mathcal{IA}_g^i \leq \mathcal{T}_{\mathcal{N}} \tag{5}$$

where $\mathcal{T}_{\mathcal{N}}$ is the threshold of the negative samples (i.e., 0.3 in this article), we assign anchors $\mathcal{A}_g^i$ as negative samples $\mathcal{S}_{\mathcal{N}}$. Finally, we add the enhancement samples $\mathcal{S}_{\mathcal{E}}$ to the positive samples $\mathcal{S}_{\mathcal{P}}$.

The threshold of positive and negative samples affects the accuracy of the detection. Because if the threshold is higher, the numbers of positive samples will decrease. If the threshold is lower, the quality of the samples will decrease. We keep this hyperparameter that is widely used in the baseline just like faster R-CNN, and we find that doing so would potentially cause the problem of insufficient positive samples. The introduction of DEA alleviates the above-mentioned risks, and it directly and independently generates more qualified positive samples. As shown in Fig. 4, we study the IoU distributions of the samples generated by faster R-CNN [7] and our DEA-net. Statistics shows that our DEA network provides more positive samples with higher IoUs compared with the original anchor-based detector.

In our method, the positive samples dynamically generated by the DEA module are all horizontal bounding box (HBB), which are the same as the anchor preset in the anchor-based branch. Then, we leverage a sample discriminator to

realize interactive sample screening between the anchor-based branch and DEA branch to generate eligible samples. Finally, we utilize RoI-transformer [2] to obtain the feature of rotated objects.

C. Network Architecture

We build a DEA-Net for both oriented and horizontal object detection tasks in aerial images. Fig. 2 illustrates the overall architecture of our DEA module integrated with faster RCNN [7] and RoI-transformer [2] for oriented object detection. We deploy ResNet [27] as the backbone, which has been pre-trained on the ImageNet [28]. Then, we construct a multi-scale feature pyramid [29] in the top-down pathway from the backbone network with levels from $\mathcal{P}_2$ to $\mathcal{P}_6$ and $\mathcal{P}_i$ has $1/2^i$ resolution of the input image. Then, we construct a DEA head to $\mathcal{P}_i$, which contains the anchor-based module and our proposed DEA module. We construct the anchor-based module following the technique reported in [7], including the RPN head network to generate horizontal region proposals.

For the DEA module, we construct an anchor-free module following the approach shown in [10]. We add four convolutional layers after the feature maps $\mathcal{P}_i$ created by the standard feature pyramid network (FPN) for classification, centralization, and regression. We decode the prediction vectors of the anchor-free module to form bounding boxes and then we select better positive samples from the two modules through the sample discriminator. Finally, for the task of oriented object detection, we build the rotated head inspired by RoI-transformer [2] which transforms the horizontal proposals to the rotated ones for arbitrary-oriented detection and a standard faster R-CNN [7] is used for horizontal object detection. The anchor-free and anchor-based modules work jointly in a multi-task style and share the features at each pyramid level.

A recent work for improving one-stage detectors is to introduce an individual prediction branch to estimate the quality of localization, where the predicted quality facilitates the classification to improve detection performance [30]. The authors compared IoU-branch and centerness-branch and believed that IoU-branch performs consistently better than centerness as a measurement of localization quality. The convincing reason is that centerness scores are much smaller than IoU scores, which causes the final scores of bounding boxes are potentially small and then removed by non-maximum suppression (NMS). In our method, we utilize DEA-branch (an anchor-free branch with centerness loss) to assist the training process of the anchor-based detector to generate eligible training sample according to the ground-truth sample rather than generating the final output score. This avoids the divergence of the two branches in the inference stage. We introduce centerness loss as a part of loss for training the DEA branch, and we select samples based on the IoU between regressed bounding boxes and the ground truth, not based on the centerness score.

D. Training and Inference

1) Multi-Task Joint Training: Integrated with faster RCNN [7], our DEA module is trained jointly with the anchor-based module in a multi-task style, as shown in Fig. 2. We define $\mathcal{L}_{\text{ab}}$ as the total loss of the anchor-based module, and $\mathcal{L}_{\text{af}}$ as the total loss of the anchor-free module. We combine the losses from the anchor-based and anchor-free modules as the loss of the entire network. Then, the total optimization loss for the whole network is

$$\mathcal{L} = \mathcal{L}_{\text{ab}} + \mathcal{L}_{\text{af}}. \tag{6}$$

For the multi-task loss in the anchor-based detection module, following [7], we optimize the target of the detection by regressing anchor boxes. The loss function for each anchor can be formulated as

$$\mathcal{L}_{\text{ab}}(\{p_i\}, \{t_i\}) = \mathcal{L}_{\text{ab\_cls}}\left(p_i, p_i^*\right) + p_i^* \mathcal{L}_{\text{ab\_reg}}\left(t_i, t_i^*\right) \tag{7}$$

where the classification loss $\mathcal{L}_{\text{ab\_cls}}$ is the cross entropy loss, p_i is the predicted probability of anchor i being an object, and p_i^* represents its ground-truth label ($p_i^* = 1$ for positive samples and $p_i^* = 0$ for negative samples). The regression loss $\mathcal{L}_{\text{ab\_reg}}$ is smooth L1 loss [6], t_i is the vector of the predicted box, and t_i^* represents the ground-truth box.

For the anchor-free module, following [10], the loss function for each location can be formulated as

$$\begin{aligned}\mathcal{L}_{\text{af}}(\{p_{m,n}\}, \{t_{m,n}\}) = {} & \mathcal{L}_{\text{af\_cls}}\left(p_{m,n}, p_{m,n}^*\right) \\ & + \mathbb{1}_{\{p_{m,n}^* > 0\}} \mathcal{L}_{\text{af\_reg}}\left(t_{m,n}, t_{m,n}^*\right) \\ & + \mathbb{1}_{\{p_{m,n}^* > 0\}} \mathcal{L}_{\text{af\_center}}\left(t_{m,n}, t_{m,n}^*\right)\end{aligned} \tag{8}$$

where classification loss $\mathcal{L}_{\text{af\_cls}}$ is focal loss [8], $p_{m,n}$ is the prediction of class labels, and $p_{m,n}^*$ represents the ground-truth label. The regression loss $\mathcal{L}_{\text{af\_reg}}$ is IoU loss [31]. $\mathbb{1}_{\{p_{m,n}^* > 0\}}$ is the indicator function, being 1 if $p_{m,n}^* > 0$ and 0 otherwise. $t_{m,n}$ is a vector of the predicted box and $t_{m,n}^*$ represents the ground truth. The centerness loss $\mathcal{L}_{\text{af\_center}}$ is the cross entropy loss.

2) Inference With Anchor-Based Module: Our DEA-Net utilizes the anchor-free and the anchor-based modules to jointly train the network to strengthen its feature representation ability and provide high-quality samples to the training task. During the inference stage, we feed the images to the anchor-based module whilst freezing the anchor-free module. This is mainly due to the fact that the anchor-free module has relatively poor consistency in locating bounding boxes, especially for the objects of a large aspect ratio. Freezing the anchor-free module could also avoid complicated fusion computation to control the computational overhead for the inference. We use the confidence score 0.05 and set the threshold of NMS to be 0.1 to generate the final detection results. We demonstrate the effectiveness of the proposed scheme in the following ablation experiments.

3) Discussion: In essence, our method is still anchor-based two-stage detection pipeline. One critical problem we solve is the "anchor over-quantization" problem, which would cause small targets to be ignored in the anchor laying process, thereby breaking the original statistical distribution of training samples and finally affecting the performance of the trained model. Instead of increasing the anchor laying density (which would greatly increase training overhead), we design a sample discriminator in the training stage. Unlike the solutions that combine anchor-based and anchor-free methods to

detect + fusion, the proposed sample discriminator (as shown in Fig. 2 and Algorithm 1) comprehensively evaluates the consistency of anchor boxes (produced in the anchor-based branch) and the inferred box produced by the DEA branch (an anchor-free branch) with ground truth. As demonstrated in Fig. 3, small targets that are split into negative samples by anchor boxes are completely retained in the DEA module, free from the quantization errors of anchors. In the sample discriminator, these target regions located by the DEA module are further regarded as positive samples to compensate for quantization errors in the anchor-based branch, as shown in Fig. 4.

IV. Experimental Work

A. Settings

1) Datasets: DOTA [1] is one of the large datasets for object detection in aerial images with both OBB and HBB annotations. It contains 2806 aerial images with 188 282 annotated instances from different sensors and platforms. The image size ranges from around 800 × 800 to 4000 × 4000 pixels and contains objects exhibiting in a wide variety of scales, orientations, and shapes. DOTA contains 15 object categories, including plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). In our experiments, following [1], [3], three-sixths of the original images are randomly selected to form the training set, one-sixth as the validation set, and two-sixths as the testing set. *HRSC2016* [14] is a challenging dataset for ship detection in aerial images with large aspect ratios and arbitrary orientations. These images were collected from Google Earth, which contain 1061 images and more than 20 categories of ships in various appearances. The image size ranges from 300 × 300 to 1500 × 900. In our work, following [14], the training, validation, and test sets include 436, 181, and 444 images, respectively. For HRSC2016, only oriented object detection can be carried out.

2) Image Size: For DOTA and HRSC2016, we generate a series of 1024 × 1024 patches from the original images with a stride of 824 for training, validation, and testing.

3) Baseline Setup: We use the standard two-stage detector faster R-CNN [7] as the baseline. It utilizes ResNet-101 as backbone. FPN [29] is adopted to construct a feature pyramid. Predefined horizontal anchors are set on each feature level, i.e., $P2$–$P6$. Here, we do not use any rotation anchor. For oriented object detection, we add the rotated head developed in RoI-transformer [2] which transforms the horizontal proposals to the rotated ones. For a fair comparison, all the experimental data and parameter settings are strictly consistent as those reported in [1], [2], and [14].

To verify the universality of our approach, we also embed our approach to ReDet [32] which incorporates rotation-equivariant networks into the detector to extract rotation-equivariant features. It uses ReResNet-50 [32] as backbone, and FPN [29] is adopted to construct a feature pyramid. And then it also adds the rotated head developed in RoI-transformer [2] for arbitrary-oriented detection.

4) Hyper-Parameters: For the hyper-parameters, following [2], [4], in DOTA and HRSC2016, only three horizontal anchors are set with the aspect ratios of $\{1/2, 1, 2\}$, the base anchor scale is set as $\{8^2\}$, and the anchor strides of each level of the feature pyramid are set to be $\{4, 8, 16, 32, 64\}$.

For the positive and negative sample selection, following [2], [7], we set the threshold of the positive samples as $\mathcal{T}_{\mathcal{P}} = 0.5$ and the threshold of the negative samples as $\mathcal{T}_{\mathcal{N}} = 0.3$.

We set $\gamma = 2$ and $\alpha = 0.25$ for the focal loss in $\mathcal{L}_{\text{af\_cls}}$.

5) Implementation Details: In order to verify the effectiveness of our method, we perform ablation studies on the DOTA dataset, and avoid utilizing any bells-and-whistles training strategy and data augmentation in the ablation study.

For the peer comparison on DOTA and HRSC2016, like [2]–[4], we only conduct rotation augmentation using 4 angles $(0, 90, 180, 270)$ to simply avoid the imbalance between different categories.

Stochastic gradient descent is used as the optimizer. The initial learning rate is set to 0.005 and divided by ten at each decay step. Weight decay and momentum are set to 0.0001 and 0.9, respectively. Following [1], [14], the total iterations for DOTA and HRSC2016 are 80 and 20 k, respectively. We train the models on RTX 2080Ti with a batch size of 1.

6) Evaluation and Metrics: Following [1], the standard mAP is used as the primary evaluation metric for accuracy. Moreover, to verify the model efficiency, the model parameters (#Params) and giga floating-point operations per second (GFLOPs)/frames/s are also taken into consideration. The results of DOTA reported in our work are obtained by submitting our predictions to the official DOTA evaluation server.[1]

B. Ablation Study

Our ablation study is carried out on DOTA [1] for oriented object detection with ResNet-101 [27], which aims to: 1) verify the effectiveness of our method on different backbone networks; 2) verify the effectiveness of our proposed units integrated with the baseline; and 3) verify the effectiveness of the inference schemes.

1) Effectiveness and Efficiency on Different Backbones: In Table I, we show the experimental results of different backbone networks with our proposed units on the test set of DOTA. We use mAP to examine our proposed module with ResNet-50, ResNet-101, and ResNet-152, respectively. Note that for aerial image, the object detection using OBB is much more important but more difficult than using HBB, that is why in Table I we perform ablation study on OBB task rather than HBB. We observe that adding our proposed module to the backbone increases mAP by 0.52%, 0.90%, and 0.43%.

In Table II, we report the model #Params and GFLOPs/frames/s for the evaluation of model efficiency. It is clear that using our proposed DEA module increases a little computational cost. For example, average increases on these

[1] https://captain-whu.github.io/DOTA/

TABLE I

EFFECTIVENESS OF OUR PROPOSED METHOD WITH DIFFERENT BACKBONE NETWORKS ON THE TEST SET OF DOTA [1] FOR ORIENTED OBJECT DETECTION. "+DEA" INDICATES THE IMPLEMENTATION OF OUR PROPOSED MODULE ON THE BACKBONE NETWORKS

R-50	R-101	R-152	+DEA	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (%)
✓				88.25	77.05	51.94	64.98	77.98	76.83	87.02	90.78	82.75	82.84	55.56	62.70	73.92	67.74	58.59	73.26
✓			✓	87.97	79.14	51.13	65.43	79.87	78.24	87.66	90.59	83.28	85.65	53.72	63.39	73.83	69.50	57.32	**73.78**$_{+0.52}$
	✓			88.53	77.70	51.59	68.80	74.02	76.85	86.98	90.24	84.89	77.68	53.91	63.56	75.88	69.48	55.50	73.06
	✓		✓	88.32	79.18	52.03	69.50	78.21	77.98	87.76	90.21	85.12	83.53	54.35	62.08	73.52	70.62	56.94	**73.96**$_{+0.90}$
		✓		88.56	77.71	54.03	72.76	74.15	77.48	87.17	90.17	76.39	83.95	45.68	64.50	76.22	69.53	53.41	72.78
		✓	✓	88.43	79.21	51.28	69.46	78.17	79.19	87.21	89.89	78.20	85.98	45.94	63.56	74.77	70.97	55.83	**73.21**$_{+0.43}$

three backbones are around 4.47 M model #Params with around 13.91 GFLOPs, and with around 2 frames/s reduction. Considering the model performance and the amount of the calculation, in the following experiments, we select ResNet-101 as our backbone network.

2) Effectiveness of the Proposed Units: In Table I, we show the performance of our DEA module integrated with three backbones for 15 categories of DOTA. We witness that our module can bring improvements for the bounding box mAP by 0.90%. We can observe that our module has large improvements on small objects. Specifically, for small vehicles (SVs), our method can increase AP by 4.19%, and for storage tank (ST), AP can be increased by 5.85%.

3) Efficiency of Our Proposed Units: We also compare the efficiency and effectiveness of our proposed method against those of the method of presetting more small anchors, as shown in Table III. The hyper-parameters settings of the comparison experiment are as follows.

1) The base anchor scale of the baseline and our method is $\{8^2\}$, and we set the base anchor scale of the method of presetting more small anchors as $\{2^2, 4^2, 8^2\}$.
2) The aspect ratios of our method and the method of presetting more small anchors are both $\{1/2, 1, 2\}$.
3) The anchor strides of each feature map of the two methods are both {4, 8, 16, 32, 64}.

It is clear that the method of presetting more anchors increases more computational cost than the proposed method: the GFLOPs increase of 21.67 (+Anchor) versus 13.91 (+DEA), with the frames/s reduction of 3.2 (+Anchor) versus 2.1 (+DEA). This is because in the method of presetting more anchors, the number of anchors has tripled and these anchors would participate in the calculation of the HBBs and the rotated bounding boxes. In terms of the performance of the two methods, the method of presetting more anchors would indeed improve the performance of some small objects (e.g., small vehicles). Our method has more improvements on small objects, because our DEA module can dynamically generate positive samples which match objects better.

4) Effectiveness of the Inference Schemes: We also compare the effectiveness of different inference schemes on DOTA for horizontal object detection after having trained the anchor-based and anchor-free baselines with our proposed method, as shown in Table IV. Compared with the two baselines, our DEA-Net of freezing the anchor-free module can increase mAP by 0.96% and 6.40%. When we fuse the outputs of the anchor-based and anchor-free modules,

TABLE II

EFFICIENCY OF OUR PROPOSED METHOD WITH DIFFERENT BACKBONE NETWORKS ON THE TEST SET OF DOTA [1] FOR ORIENTED OBJECT DETECTION. "+DEA" INDICATES THE IMPLEMENTATION OF OUR PROPOSED MODULE ON THE BACKBONE NETWORKS

R-50	R-101	R-152	+DEA	GFLOPs / FPS	#Params (M)
✓				211.30 / 14.8	55.13 6
✓			✓	225.21 / 12.5	59.90
	✓			289.19 / 12.7	74.12
	✓		✓	303.10 / 10.6	78.89
		✓		367.17 / 11.1	89.77
		✓	✓	381.08 / 9.2	94.54

mAP can have a minor improvement (0.96% vs. 1.07% and 6.40% vs. 6.51%), compared with the former. Meanwhile, fusing the two modules for inference, the inference speed becomes clearly slower (10.8 vs. 6.2 frames/s). That is why after having trained the DEA-net, we freeze the anchor-free branch and only utilize the anchor-based module for inference.

5) Visualizations: We show some of the visual comparisons for oriented object detection between the baseline and the proposed method in Fig. 5. The proposed method achieves notably better precision for small object detection, such as small vehicles, storage tanks, ships, and airplanes.

C. Comparisons With State-of-the-Arts

1) Results on DOTA: We compare the proposed approach with some SOTA methods on the test set of DOTA, as shown in Table V. When our approach integrated with RoI-transformer [2], our DEA-Net achieves 77.77% mAP for oriented object detection and 78.43% mAP for horizontal object detection and outperforms many advanced methods. Of these 15 categories, DEA-Net ranks at the top for four categories for oriented object detection and ten categories for horizontal object detection. Moreover, DEA-Net surpasses the advanced method by 0.40 mAP for oriented object detection with a weaker backbone network (ResNet-101 vs. ResNet-152) and 3.08 mAP for horizontal object detection with the same backbone. Visualization results on the test set of DOTA are shown in Fig. 6. DEA-Net can accurately predict the categories and have satisfactory performance on small objects, such as small vehicles, storage tanks, and ships.

We also compare the proposed approach with some newest methods on the test set of DOTA, as shown in Table VI. To verify the universality of our approach, we embed our

TABLE III

COMPARISON OF EFFICIENCY BETWEEN OUR PROPOSED METHOD AND THE METHOD OF PRESET MORE SMALL ANCHORS ON DOTA [1] FOR ORIENTED OBJECT DETECTION. "BASELINE" INDICATES THE FASTER RCNN WITH THE BACKBONE OF RESNET-101. "+ANCHOR" INDICATES THE IMPLEMENTATION OF MORE SMALL ANCHORS ON THE BASELINE NETWORKS. "+DEA" INDICATES THE IMPLEMENTATION OF OUR PROPOSED MODULE ON THE BASELINE NETWORKS

Baseline	+Anchor	+DEA	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	GFLOPs / FPS	#Params (M)	mAP (%)
✓			88.53	77.70	51.59	68.80	74.02	76.85	86.98	90.24	84.89	77.68	53.91	63.56	75.88	69.48	55.50	289.19 / 12.7	74.12	73.06
✓	✓		88.41	80.14	53.79	70.70	77.82	76.98	86.98	90.75	83.90	81.13	51.95	61.41	74.89	69.15	59.27	310.86 / 9.5	74.13	73.82
✓		✓	88.32	79.18	52.03	69.50	78.21	77.98	87.76	90.21	85.12	83.53	54.35	62.08	73.52	70.62	56.94	303.10 / 10.6	78.89	**73.96**

TABLE IV

EFFECTIVENESS OF DIFFERENT INFERENCE SCHEMES WITH OUR DEA-NET ON DOTA [1] FOR HORIZONTAL OBJECT DETECTION IN AERIAL IMAGES. RESNET-101 [27] IS THE BACKBONE

Inference schemes	FPS	mAP (%)
Anchor-based baseline Faster RCNN [7]	13.0	73.89
Anchor-free baseline FCOS [10]	11.0	68.45
Ours (DEA-Net) freezing anchor-free branch	10.8	74.85$_{+0.96/+6.40}$
Ours (DEA-Net) anchor-free + anchor-based fusion	6.2	74.96$_{+1.07/+6.51}$

TABLE V

COMPARISONS WITH OTHER SOTA METHODS ON THE TEST SET OF DOTA [1] FOR BOTH ORIENTED AND HORIZONTAL OBJECT DETECTION IN AERIAL IMAGES. "OURS" MEANS THE IMPLEMENTATION OF THE DEA MODULE ON THE BASELINE MODEL. "R-" IN THE BACKBONE COLUMN DENOTES RESNET [27], AND "H-" DENOTES THE HOURGLASS NETWORK [42]

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (%)
Oriented object detection																	
FR-O [1] (CVPR 2018)	R-101	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 [33] (ISO4 2018)	R-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^2CNN [34] (preprint 2017)	R-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 [23] (TMM 2018)	R-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 [35] (ACCV 2018)	R-101	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.02	53.64	62.90	67.02	64.17	50.23	68.16
RoI Trans [2] (CVPR 2019)	R-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
CAD-Net [36] (TGRS 2019)	R-101	87.80	82.40	49.40	73.50	71.10	63.50	76.70	**90.90**	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
DRN [37] (CVPR 2020)	H-104	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70
O^2-DNet [38] (ISPRS 2020)	H-104	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
SCRDet [3] (ICCV 2019)	R-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
R^3Det [26] (preprint 2019)	R-152	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74
CSL [39] (ECCV 2020)	R-152	**90.25**	**85.53**	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	**69.60**	**68.04**	73.83	71.10	68.93	76.17
DAL [4] (AAAI 2021)	R-50	89.69	83.11	55.03	71.00	78.30	81.90	**88.46**	90.89	84.97	**87.46**	64.41	65.65	**76.86**	72.09	64.35	76.95
R^3Det-DCL [5] (CVPR 2021)	R-152	89.26	83.60	53.54	72.76	79.04	82.56	87.31	90.67	86.59	86.98	67.49	66.88	73.29	70.56	**69.99**	77.37
Ours (RoI Trans + DEA)	R-101	89.18	83.46	**55.14**	71.69	**79.59**	**83.08**	88.10	90.88	87.09	86.73	63.99	65.14	75.81	**78.01**	68.69	**77.77**$_{+0.40}$
Horizontal object detection																	
SSD [40] (ECCV 2016)	R-101	44.74	11.21	6.22	6.91	2.00	10.24	11.34	15.59	12.56	17.94	14.73	4.55	4.55	0.53	1.01	10.94
YOLOv2 [9] (CVPR 2016)	R-101	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 [41] (NIPS 2016)	R-101	79.33	44.26	36.58	53.53	39.38	34.15	47.29	45.66	47.74	65.84	37.92	44.23	47.23	50.64	34.90	47.24
FR-H [1] (CVPR 2018)	R-101	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 [29] (CVPR 2017)	R-101	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 [35] (ACCV 2018)	R-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
SCRDet [3] (ICCV 2019)	R-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
Ours (RoI Trans + DEA)	R-101	89.18	**83.34**	**58.94**	71.69	**80.23**	**83.97**	**88.26**	90.88	**87.09**	**87.44**	64.24	**65.04**	76.40	**80.88**	**68.81**	**78.43**$_{+3.08}$

TABLE VI

COMPARISONS WITH SOME CURRENT SOTA METHODS ON THE TEST SET OF DOTA [1] FOR ORIENTED OBJECT DETECTION IN AERIAL IMAGES. "OURS" MEANS THE IMPLEMENTATION OF THE DEA MODULE ON THE BASELINE MODEL. "R-" IN THE BACKBONE COLUMN DENOTES RESNET [27], AND "RER-" DENOTES ROTATION-EQUIVARIANT RESNET [32]

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (%)
S^2A-Net [43] (TGRS 2021)	R-50	88.89	83.60	57.74	81.95	79.94	83.19	**89.11**	90.78	84.87	87.81	70.30	**68.25**	78.30	77.01	69.58	79.42
ReDet [32] (CVPR 2021)	ReR-50	88.81	82.48	**60.83**	80.82	78.34	86.06	88.31	**90.87**	88.77	87.03	68.65	66.90	**79.26**	**79.71**	74.67	80.10
GWD [44] (ICML 2021)	R-152	89.66	**84.99**	59.26	**82.19**	78.97	84.83	87.70	90.21	86.54	86.85	**73.47**	67.77	76.92	79.22	74.92	80.23
Ours (ReDet + DEA)	ReR-50	**89.92**	83.84	59.65	79.88	**80.11**	**87.96**	88.17	90.31	**88.93**	**88.46**	68.93	65.94	78.04	79.69	**75.78**	**80.37**$_{+0.14}$

approach to one of these current detectors ReDet [32], which is a SOTA rotation detector that explicitly encodes rotation equivariance and rotation invariance. We integrate our DEA module with ReDet and conduct data augmentation following the way in [32] (i.e., multi-scale data and random rotation), our method achieves 80.37% mAP for oriented object detection,

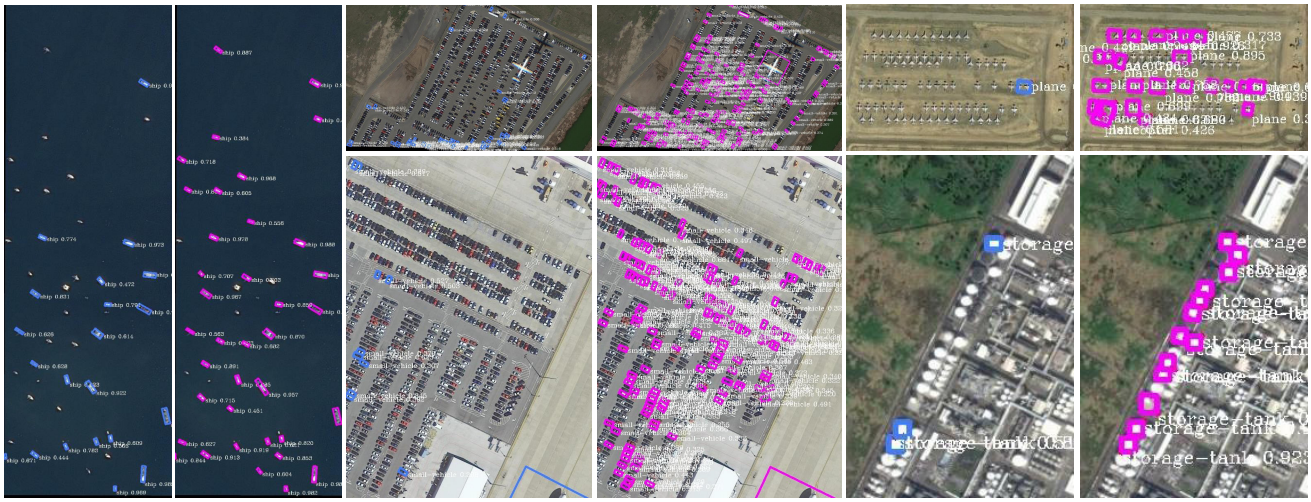

Fig. 5. Comparison against the baseline (faster RCNN [7] + RoI-transformer [2]) on DOTA [1] for oriented object detection with ResNet-101 [27]. Blue boxes indicate the results of the baseline and pink boxes are the results of our proposed DEA-Net.

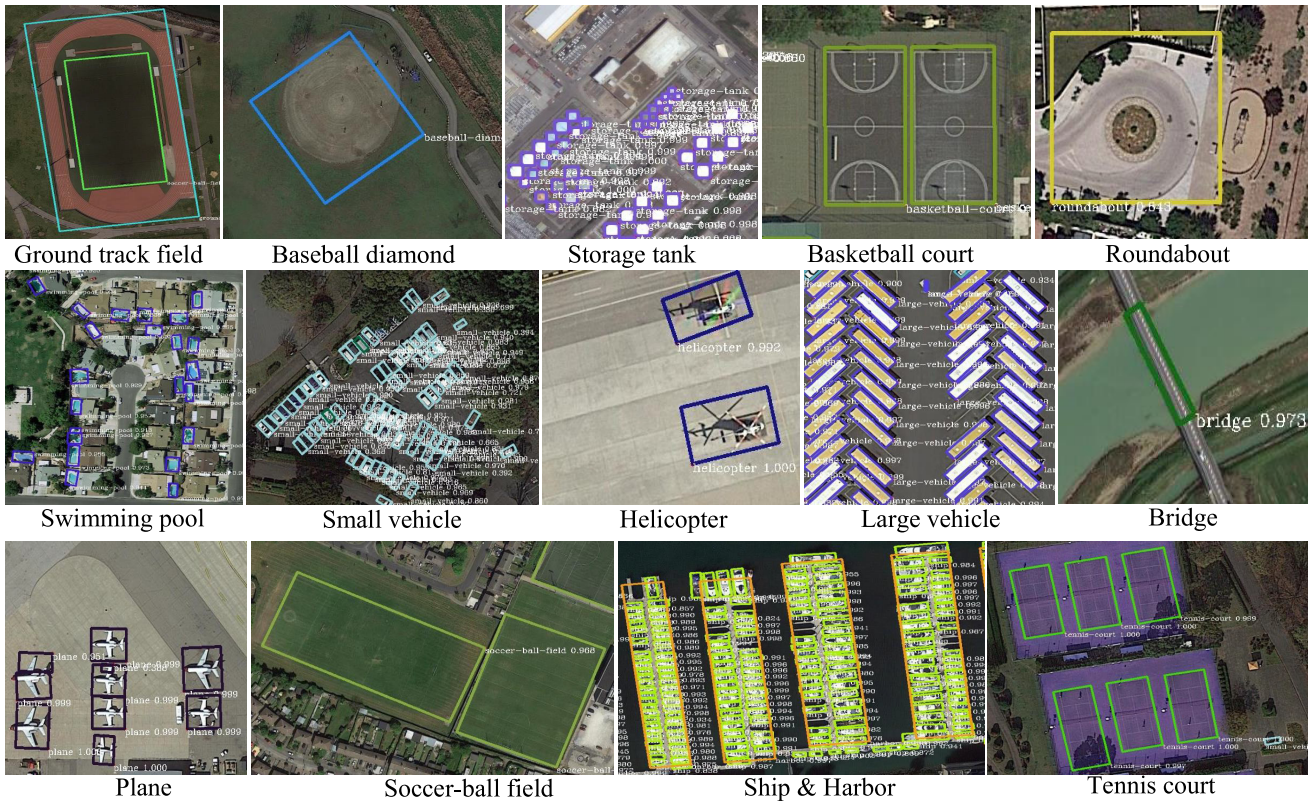

Fig. 6. Visualization results for oriented object detection on the test set of DOTA [1].

and of these 15 categories, it ranks at the top for five categories.

2) Results on HRSC2016: The comparisons with the other SOTA methods on the test set of HRSC2016 [14] are shown in Table VII. We can observe that our method achieves the SOTA performance in mAP by 90.56%, which surpasses the previous best model by 1.1%. Particularly, in our experiments, our DEA-Net uses only three horizontal anchors with the aspect ratios of {1/2, 1, 2}, but outperforms the other frameworks with a large number of anchors. Our proposed method also achieves better precision for objects with a large aspect ratio. The experiments show that it is critical to effectively utilize the predefined anchors and select high-quality samples where our DEA module can regress the bounding boxes at the locations of the objects without presetting a large number of rotated anchors.

3) Discussion: The object detection using OBB is much more difficult than using HBB. Tables V and VII also support this phenomenon. In Table V, for OBB task, the proposed methods outperform SOTA by 0.4% but for HBB it

TABLE VII

COMPARISONS WITH OTHER SOTA METHODS ON THE TEST SET OF HRSC2016 [14] FOR ORIENTED OBJECT DETECTION IN AERIAL IMAGES. "R-" IN THE BACKBONE COLUMN DENOTES THE RESNET [27], AND "V-" DENOTES THE VGG NETWORK [48]. MAP IS OBTAINED ON THE VOC 2007 EVALUATION METRIC

Methods	Backbone	mAP (%)
R^2CNN [34] (preprint 2017)	R-101	73.07
RCI&RC2 [14] (ICPRAM 2017)	V-16	75.70
RRPN [23] (TMM 2018)	R-101	79.08
R^2PN [45] (GRSL 2018)	V-16	79.60
RRD [46] (CVPR 2018)	V-16	84.30
RoI Trans [2] (CVPR 2019)	R-101	86.20
Gliding Vertex [47] (TPAMI 2020)	R-101	88.20
R-RetinaNet [8] (ICCV 2017)	R-101	89.18
R^3Det [26] (preprint 2019)	R-101	89.26
RetinaNet-DAL [4] (AAAI 2021)	R-101	89.77
R^3Det-DCL [5] (CVPR 2021)	R-101	89.46
Ours (RoI Trans + DEA)	R-101	**90.56**$_{+1.10}$

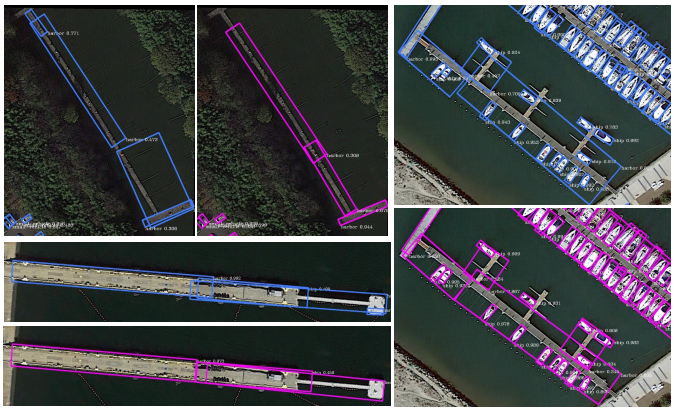

Fig. 7. Some failed visualized comparisons of baseline and our method. The figures with blue boxes are the results of the baseline (faster R-CNN + RoI-transformer) and pink boxes are the results of the proposed DEA-Net.

outperforms SOTA by 3.08%. In Tables V and VII, we can also find that, for OBB task, recent SOTA method can have only <1% gain compared with the previous state of the art methods, for example, R3Det-DCL [5] (International Conference on Computer Vision and Pattern Recogintion (CVPR) 2021) 77.37% versus DAL [4] (the Association for the Advance of Artificial Intelligence (AAAI) 2021) 76.95% on DOTA.

We also list some failed visualization comparisons of baseline and our method, as shown in Fig. 7. The results show that our method, similar to the baseline, cannot generate accurate bounding boxes when detecting objects with extreme aspect ratios (e.g., harbor). This may be because the DEA module fail to generate accurate positive samples when the objects are of a large size and an extreme aspect ratio, while the anchor-based branch (e.g., the baseline) also cannot regress accurate bounding boxes when the aspect ratio of preset anchors are not match the objects of extreme aspect ratio. We show more visual comparisons of the results in the appendix. Specifically, we show some of the visual comparisons for oriented object detection between the baseline and the proposed method on HRSC2016 in Fig. 8. We also show some visualized comparisons for horizontal object detection on DOTA in Fig. 9. The visualization results show that our proposed method can achieve better performance both in the case of ships with OBBs and objects with HBBs.

APPENDIX A
SOME VISUALIZED COMPARISON OF OBJECT DETECTION WITH OBB ON HRSC2016

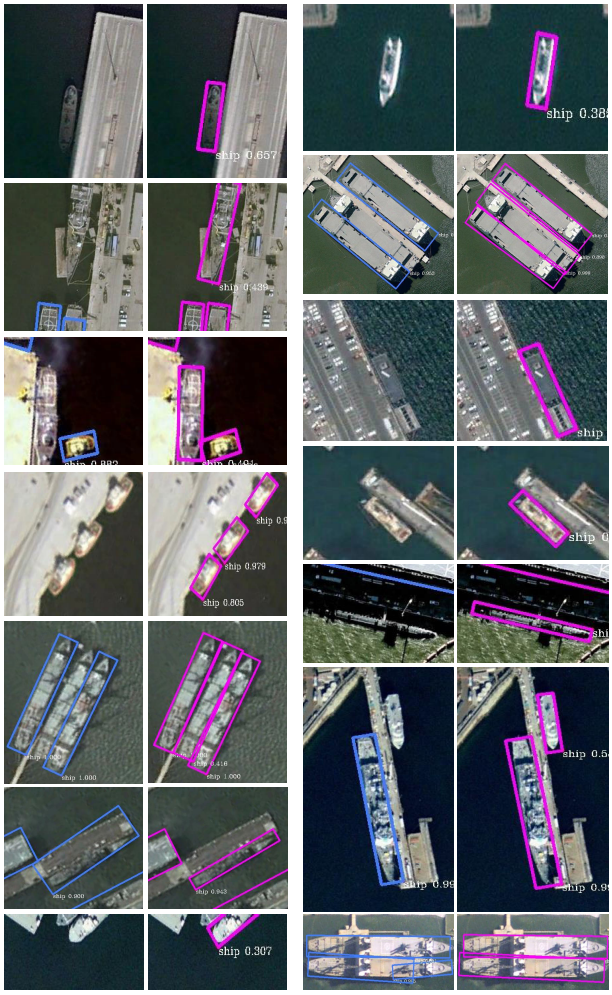

Fig. 8. Some visualized comparison of object detection with OBB on HRSC2016. The figures with blue boxes are the results of the baseline (RoI-transformer + faster R-CNN) and pink boxes are the results of our proposed DEA-Net.

V. CONCLUSION AND FUTURE WORK

In this work, a simple yet effective DEA module was proposed to facilitate the learning of small objects. We implemented our DEA module on the standard object detection backbone network with an FPN (i.e., DEA-Net) and conducted extensive experiments on both oriented and horizontal object detection in aerial images. Experimental results on the challenging DOTA and HRSC2016 indicated that our proposed DEA-Net could achieve SOTA performance in accuracy with moderate computational overhead.

In the future, we will extend the proposed DEA-Net to a broader range of natural scenes. Besides, exploring how to use

APPENDIX B
SOME VISUALIZED COMPARISONS OF OBJECT DETECTION WITH HBB ON DOTA

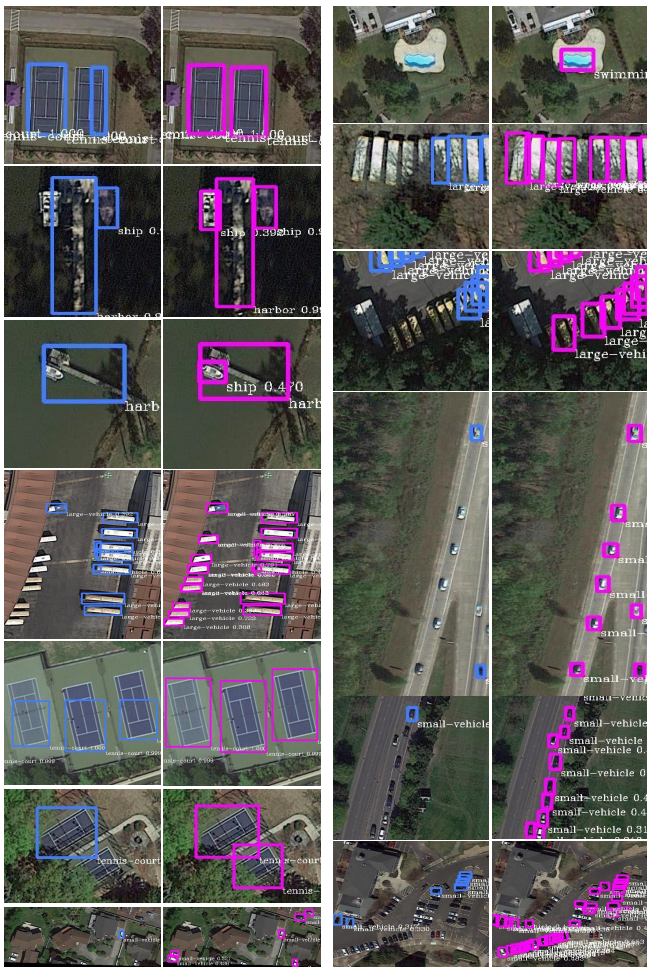

Fig. 9. Some visualized comparisons of object detection with HBB on DOTA. The figures with blue boxes are the results of the baseline (faster R-CNN) and pink boxes are the results of our proposed DEA-Net.

DEA-Net for semantic and panoramic segmentation is also a promising direction.

ACKNOWLEDGMENT

The authors would like to thank Prof. Sheng-Jun Huang from the Nanjing University of Aeronautics and Astronautics and Prof. Gui-Song Xia from the Wuhan University for their help discussion and comments.

REFERENCES

[1] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[2] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2849–2858.

[3] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8232–8241.

[4] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, 2021, pp. 2355–2363.

[5] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15819–15829.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[10] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[11] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[12] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[13] K. Duan, L. Xie, H. Qi, S. Bai, and Q. Tian, "Corner proposal network for anchor-free, two-stage object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 399–416.

[14] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[16] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2965–2974.

[17] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849.

[18] J. Wang, Y. Yuan, B. Li, G. Yu, and S. Jian, "SFace: An efficient network for face detection in large scale variations," 2018, *arXiv:1804.06559*.

[19] T. Yang, X. Zhang, Z. Li, W. Zhang, and J. Sun, "MetaAnchor: Learning to detect objects with customized anchors," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 318–328.

[20] C. Zhu, F. Chen, Z. Shen, and M. Savvides, "Soft anchor-point object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 91–107.

[21] S. Li, L. Yang, J. Huang, X.-S. Hua, and L. Zhang, "Dynamic anchor feature selection for single-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6609–6618.

[22] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[23] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[24] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[25] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[26] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," 2019, *arXiv:1908.05612*.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[30] X. Li *et al.*, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 21002–21012.

[31] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520.
[32] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2786–2795.
[33] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, 2018.
[34] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.
[35] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 150–165.
[36] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.
[37] X. Pan *et al.*, "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11207–11216.
[38] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.
[39] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 677–694.
[40] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.
[41] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2016, p. 379—387.
[42] T. Xu and W. Takano, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2021, pp. 483–499.
[43] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
[44] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 11830–11841.
[45] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1745–1749, Nov. 2018.
[46] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5909–5918.
[47] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

Dong Liang (Member, IEEE) received the B.S. degree in telecommunication engineering and the M.S. degree in circuits and systems from Lanzhou University, Lanzhou, China, in 2008 and 2011, respectively, and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University, Hokkaido, Japan, in 2015.

He is currently an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He has published several journal articles, including IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition*, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His research interests include model communication in pattern recognition and image processing.

Dr. Liang was awarded the Excellence Research Award from Hokkaido University in 2013.

Qixiang Geng received the B.S. degree in computer science and technology from Nanjing Audit University, Nanjing, China, in 2019. He is currently pursuing the master's degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing.

He is currently a Research Intern with Senstime, Beijing, China. His research interest includes computer vision, especially in object detection.

Zongqi Wei received the B.S. degree in computer science and technology from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2019. He is currently pursuing the master's degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing.

He is currently a Research Intern with TikTok, Beijing, China. His research interest includes computer vision, especially in object detection and video object segmentation.

Dmitry A. Vorontsov received the B.S., M.S., and Ph.D. degrees from the Faculty of Physics, National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia, in 2003, 2005, and 2008, respectively.

He was an Engineer, a Junior Researcher, a Senior Researcher, and an Associate Professor with the National Research Lobachevsky State University of Nizhny Novgorod, from 2004 to 2020. He was a Visiting Researcher with Hokkaido University, Hokkaido, Japan, from 2013 to 2014. He was awarded the Japan Society for the Promotion of Science (JSPS) Grant in 2013. His research interests include machine vision and nonlinear characterization of optical facilities.

Ekaterina L. Kim received the Ph.D. degree from the Faculty of Physics, National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia, in 2003.

She is currently an Associate Professor with the National Research Lobachevsky State University of Nizhny Novgorod. She was awarded a Civilian Research and Development Foundation (CRDF) Grant from 2004 to 2005. Her research interests focus on machine vision and nonlinear characterization of optical facilities.

Mingqiang Wei (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong (CUHK), Hong Kong, in 2014.

He was an Assistant Professor at the Hefei University of Technology, Hefei, China, and a Post-Doctoral Fellow at CUHK. He is a Professor at the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His research interests focus on 3-D vision, computer graphics, and deep learning.

Dr. Wei was a recipient of the CUHK Young Scholar Thesis Award in 2014. He is currently an Associate Editor of *The Visual Computer* journal, *Journal of Electronic Imaging*, and *Journal of Image and Graphics*.

Huiyu Zhou received the B.E. degree in radio technology from the Huazhong University of Science and Technology, Wuhan, China, in 1990, the M.Sc. degree in biomedical engineering from the University of Dundee, Dundee, U.K., in 2002, and the Ph.D. degree in computer vision from Heriot-Watt University, Edinburgh, U.K., in 2006.

He is currently a Full Professor with the School of Computing and Mathematical Sciences, University of Leicester, Leicester, U.K. He has published over 350 peer-reviewed articles in his field.