

Dynamic Anchor Learning for Arbitrary-Oriented Object Detection

Qi Ming, Zhiqiang Zhou*, Lingjuan Miao, Hongwei Zhang, Linhao Li

School of Automation, Beijing Institute of Technology, China
 chaser.ming@gmail.com, {zhzhzhou, miaolingjuan}@bit.edu.cn,
 zhanghw.hongwei@gmail.com, lilinhao@bit.edu.cn

Abstract

Arbitrary-oriented objects widely appear in natural scenes, aerial photographs, remote sensing images, etc., and thus arbitrary-oriented object detection has received considerable attention. Many current rotation detectors use plenty of anchors with different orientations to achieve spatial alignment with ground truth boxes. Intersection-over-Union (IoU) is then applied to sample the positive and negative candidates for training. However, we observe that the selected positive anchors cannot always ensure accurate detections after regression, while some negative samples can achieve accurate localization. It indicates that the quality assessment of anchors through IoU is not appropriate, and this further leads to inconsistency between classification confidence and localization accuracy. In this paper, we propose a dynamic anchor learning (DAL) method, which utilizes the newly defined matching degree to comprehensively evaluate the localization potential of the anchors and carries out a more efficient label assignment process. In this way, the detector can dynamically select high-quality anchors to achieve accurate object detection, and the divergence between classification and regression will be alleviated. With the newly introduced DAL, we can achieve superior detection performance for arbitrary-oriented objects with only a few horizontal preset anchors. Experimental results on three remote sensing datasets HRSC2016, DOTA, UCAS-AOD as well as a scene text dataset ICDAR 2015 show that our method achieves substantial improvement compared with the baseline model. Besides, our approach is also universal for object detection using horizontal bound box. The code and models are available at <https://github.com/ming71/DAL>.

Introduction

Object detection is one of the most fundamental and challenging problem in computer vision. In recent years, with the development of deep convolutional neural networks (CNN), tremendous successes have been achieved on object detection (Ren et al. 2015; Dai et al. 2016; Redmon et al. 2016; Liu et al. 2016). Most detection frameworks utilize preset horizontal anchors to achieve spatial alignment with ground-truth(GT) box. Positive and negative samples are then selected through a specific strategy during training phase, which is called label assignment.

*Corresponding author

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



Figure 1: Predefined anchor (red) and its regression box (green). (a) shows that anchors with a high input IoU cannot guarantee perfect detection. (b) reveals that the anchor that is poorly spatially aligned with the GT box still has the potential to localize object accurately.

Since objects in the real scene tend to appear in diverse orientations, the issue of oriented object detection has gradually received considerable attention. There are many approaches have achieved oriented object detection by introducing the extra orientation prediction and preset rotated anchors (Ma et al. 2018; Liao, Shi, and Bai 2018). These detectors often follow the same label assignment strategy as general object detection frameworks. For simplicity, we call the IoU between GT box and anchors as input IoU, and the IoU between GT box and regression box as output IoU. The selected positives tend to obtain higher output IoU compared with negatives, because their better spatial alignment with GT provides sufficient semantic knowledge, which is conducive to accurate classification and regression.

However, we observe that the localization performance of the assigned samples is not consistent with the assumption mentioned above. As shown in Figure 1, the division of positive and negative anchors seems not always related to the detection performance. Furthermore, we have counted the distribution of the anchor localization performance for all candidates to explore whether this phenomenon is universal. As illustrated in Figure 2(a), a considerable percentage (26%) of positive anchors are poorly aligned with GT after regression, revealing that the positive anchors cannot ensure accurate localization. Besides, more than half of the candidates that achieve high-quality predictions are regressed from negatives (see Figure 2(b)) which implies that a consid-

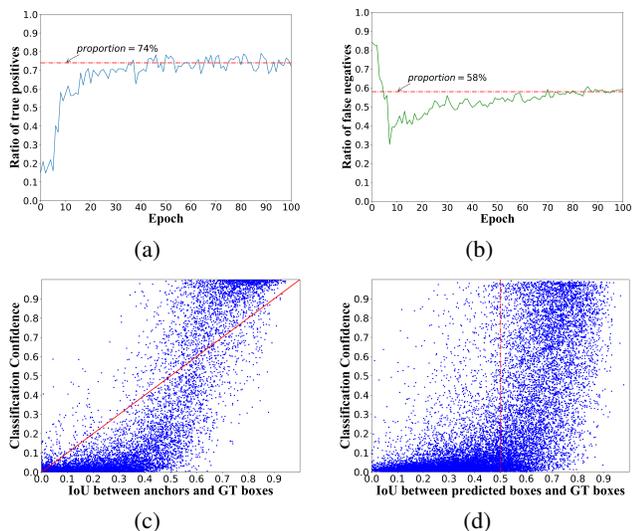


Figure 2: Analysis of the classification and regression capabilities of anchors that use input IoU for label assignment. (a) Only 74% of the positive sample anchors can localize GT well after regression (with output IoU higher than 0.5), which illustrates that many false positive samples are introduced. (b) Only 42% of the high-quality detections (output IoU is higher than 0.5) come from matched anchors, which implies that quite a lot of negative anchors (58% in this example) have the potential to achieve accurate localization. (c) Current label assignment leads to a positive correlation between the classification confidence and the input IoU. (d) High-performance detection results exhibit a weak correlation between the localization ability and classification confidence, which is not conducive to selecting accurate detection results through classification score during inference.

erable number of negatives with high localization potential have not been effectively used. In summary, we conclude that the localization performance does not entirely depend on the spatial alignment between the anchors and GT.

Besides, inconsistent localization performance before and after anchor regression further lead to inconsistency between classification and localization, which has been discussed in previous work (Jiang et al. 2018; Kong et al. 2019; He et al. 2019). As shown in Figure 2(c), anchor matching strategy based on input IoU induces a positive correlation between the classification confidence and input IoU. However, as discussed above, the input IoU is not entirely equivalent to the localization performance. Therefore, we cannot distinguish the localization performance of the detection results based on the classification score. The results in Figure 2(d) also confirm this viewpoint: a large number of regression boxes with high output IoU are misjudged as background.

To solve the problems, we propose a Dynamic Anchor Learning (DAL) method for better label assignment and further improve the detection performance. Firstly, a simple yet effective standard named *matching degree* is designed to assess the localization potential of anchors, which compre-

hensively considers the prior information of spatial alignment, the localization ability and the regression uncertainty. After that, we adopt matching degree for training sample selection, which helps to eliminate false-positive samples and dynamically mine potential high-quality candidates, as well as suppress the disturbance caused by regression uncertainty. Next, we propose a matching-sensitive loss function to further alleviate the inconsistency between classification and regression, making the classifier more discriminative for proposals with high localization performance, and ultimately achieving high-quality detection.

Extensive experiments on public datasets, including remote sensing datasets HRSC2016, DOTA, UCAS-AOD, and scene text dataset ICDAR 2015, show that our method can achieve stable and substantial improvements for arbitrary-oriented object detections. Integrated with our approach, even the vanilla one-stage detector can be competitive with state-of-the-art methods on several datasets. In addition, experiments on ICDAR 2013 and NWPV VHR-10 prove that our approach is also universal for object detection using horizontal box. The proposed DAL approach is general and can be easily integrated into existing object detection pipeline without increasing the computational cost of inference.

Our contributions are summarized as follows:

- We observe that the label assignment based on IoU between anchor and GT box leads to suboptimal localization ability assessment, and further brings inconsistent classification and regression performance.
- The matching degree is introduced to measure the localization potential of anchors. A novel label assignment method based on this metric is proposed to achieve high-quality detection.
- The matching-sensitive loss is proposed to alleviate the problem of the weak correlation between classification and regression, and improves the discrimination ability of high-quality proposals.

Related Work

Arbitrary-Oriented Object Detection

The current mainstream detectors can be divided into two categories: two-stage detector (Ren et al. 2015; Dai et al. 2016) and one-stage detector (Redmon et al. 2016; Liu et al. 2016). Existing rotation detectors are mostly built on detectors using horizontal bounding box representation. To localize rotated objects, the preset rotated anchor and additional angle prediction are adopted in the literature (Liu, Ma, and Chen 2018; Ma et al. 2018; Liao, Shi, and Bai 2018; Liu et al. 2017a). Nevertheless, due to the variation of the orientation, these detectors are obliged to preset plenty of rotated anchors to make them spatially aligned with GT box. There are also some methods that detect oriented objects only using horizontal anchors. For example, RoI Transformer (Ding et al. 2019) uses horizontal anchors but learns the rotated RoI through spatial transformation, reducing the number of predefined anchors. R³Det (Yang et al. 2019a) adopts cascade regression and refined box re-encoding module to achieve state-of-the-art performance with horizontal

anchors. Although the above approaches have achieved good performance, they cannot make a correct judgment on the quality of anchors, and thus cause improper label assignment which brings adverse impact during the training process.

Label Assignment

Most anchor-based detectors densely preset anchors at each position of feature maps. The massive preset anchors lead to serious imbalance problem, especially for the arbitrary-oriented objects with additional angle setting. The most common solution is to control the ratio of candidates through a specific sampling strategy (Shrivastava, Gupta, and Girshick 2016; Pang et al. 2019). Besides, Focal Loss (Lin et al. 2017b) lowers the weight of easy examples to avoid its overwhelming contribution to loss. The work (Li, Liu, and Wang 2019) further considers extremely hard samples as outliers, and use gradient harmonizing mechanism to conquer the imbalance problems. We demonstrate that the existence of outliers is universal, and our method can prevent such noise samples from being assigned incorrectly.

Some work have observed problems caused by using input IoU as a standard for label assignment. Dynamic R-CNN (Zhang et al. 2020a) and ATSS (Zhang et al. 2020b) automatically adjust the IoU threshold to select high-quality positive samples, but they fail to consider whether the IoU itself is credible. The work (Li et al. 2020) points out that binary labels assigned to anchors are noisy, and construct cleanliness score for each anchor to supervise training process. However, it only considers the noise of positive samples, ignoring the potentially powerful localization capabilities of massive negatives. HAMBox (Liu et al. 2020) reveals that unmatched anchors can also achieve accurate prediction, and attempts to utilize these samples. Nevertheless, its compensated anchors mined according to output IoU is not reliable; moreover, it does not consider the degradation of matched positives. FreeAnchor (Zhang et al. 2019) formulates object-anchor matching as a maximum likelihood estimation procedure to select the most representative anchors, but its formulation is relatively complicated.

Proposed Method

Rotation Detector Built on RetinaNet

The real-time inference is essential for arbitrary-oriented object detection in many scenarios. Hence we use the one-stage detector RetinaNet (Lin et al. 2017b) as the baseline model. It utilizes ResNet-50 as backbone, in which the architecture similar to FPN (Lin et al. 2017a) is adopted to construct a multi-scale feature pyramid. Predefined horizontal anchors are set on the features of each level P_3, P_4, P_5, P_6, P_7 . Note that rotation anchor is not used here, because it is inefficient and unnecessary, and we will further prove this point in the next sections. Since the extra angle parameter is introduced, the oriented box is represented in the format of (x, y, w, h, θ) . For bounding box regression, we have:

$$\begin{aligned} t_x &= (x - x_a) / w_a, & t_y &= (y - y_a) / h_a \\ t_w &= \log(w / w_a), & t_h &= \log(h / h_a) \\ t_\theta &= \tan(\theta - \theta_a) \end{aligned} \quad (1)$$

where x, y, w, h, θ denote center coordinates, width, height and angle, respectively. x and x_a are for the predicted box and anchor, respectively (likewise for y, w, h, θ). Given the ground-truth box offsets $\mathbf{t}^* = (t_x^*, t_y^*, t_w^*, t_h^*, t_\theta^*)$, the multi-task loss is defined as follows:

$$L = L_{cls}(p, p^*) + L_{reg}(\mathbf{t}, \mathbf{t}^*) \quad (2)$$

in which the value p and the vector \mathbf{t} denote predicted classification score and predicted box offsets, respectively. Variable p^* represents the class label for anchors ($p^* = 1$ for positive samples and $p^* = 0$ for negative samples).

Dynamic Anchor Selection

Some researches (Zhang et al. 2019; Song, Liu, and Wang 2020) have reported that the discriminative features required to localize objects are not evenly distributed on GT, especially for objects with a wide variety of orientations and aspect ratios. Therefore, the label assignment strategy based on spatial alignment, i.e., input IoU, leads to the incapability to capture the critical feature demanded for object detection.

An intuitive approach is to use the feedback of the regression results, that is, the output IoU to represent feature alignment ability and dynamically guide the training process. Several attempts (Jiang et al. 2018; Li et al. 2020) have been made in this respect. In particular, we tentatively select the training samples based on output IoU and use it as soft-label for classification. However, we found that the model is hard to converge because of the following two issues:

- Anchors with high input IoU but low output IoU are not always negative samples, which may be caused by not sufficient training.
- The unmatched low-quality anchors that accidentally achieve accurate localization performance tend to be misjudged as positive samples.

The above analysis shows that regression uncertainty interferes with the credibility of the output IoU for feature alignment. Regression uncertainty has been widely discussed in many previous work (Feng, Rosenbaum, and Dietmayer 2018; Choi et al. 2019; Kendall and Gal 2017; Choi et al. 2018), which represents the instability and irrelevance in the regression process. We discovered in the experiment that it also misleads the label assignment. Specifically, high-quality samples cannot be effectively utilized, and the selected false-positive samples would cause the unstable training. Unfortunately, neither the input IoU nor the output IoU used for label assignment can avoid the interference caused by the regression uncertainty.

Based on the above observations, we introduce the concept of *matching degree* (MD), which utilizes the prior information of spatial matching, feature alignment ability and regression uncertainty of the anchor to measure the localization capacity, which is defined as follows:

$$md = \alpha \cdot sa + (1 - \alpha) \cdot fa - u^\gamma \quad (3)$$

where sa denotes a priori of spatial alignment, whose value is equivalent to input IoU. fa represents the feature alignment capability calculated by IoU between GT box and regression box. α and γ are hyperparameters used to weight

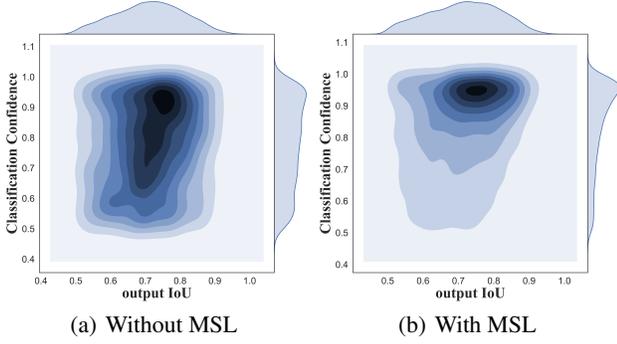


Figure 3: The correlation between the output IoU and classification score with and without MSL.

the influence of different items. u is a penalty term, which denotes the regression uncertainty during training. It is obtained via the IoU variation before and after regression:

$$u = |sa - fa| \quad (4)$$

The suppression of interference during regression is vital for high-quality anchor sampling and stable training. Variation of IoU before and after regression represents the probability of incorrect anchor assessment. Note that our construction of the penalty term for regression uncertainty is very simple, and since detection performance is not sensitive to the form of u , the naive, intuitive yet effective form is employed.

With the newly defined matching degree, we conduct dynamic anchor selection for superior label assignment. In the training phase, we first calculate matching degree between the GT box and anchors, and then the anchors with matching degree higher than a certain threshold (set to 0.6 in our experiments) are selected as positives, while the rest are negatives. After that, for GT that do not match any anchor, the anchor with the highest matching degree will be compensated as positive candidate. To achieve more stable training, we gradually adjust the impact of the input IoU during training. The specific adjustment schedule is as follows:

$$\alpha(t) = \begin{cases} 1, & t < 0.1 \\ 5(\alpha_0 - 1) \cdot t + 1.5 - 0.5 \cdot \alpha_0, & 0.1 \leq t < 0.3 \\ \alpha_0, & t \geq 0.3 \end{cases} \quad (5)$$

where $t = \frac{iters}{Max_Iteration}$, $Max_Iteration$ is the total number of iterations, and α_0 is the final weighting factor that appears in Eq. (3).

Matching-Sensitive Loss

To further enhance the correlation between classification and regression to achieve high-quality arbitrary-oriented detection, we integrate the matching degree into the training process and propose the matching-sensitive loss function (MSL). The classification loss is defined as:

$$L_{cls} = \frac{1}{N} \sum_{i \in \psi} FL(p_i, p_i^*) + \frac{1}{N_p} \sum_{j \in \psi_p} w_j \cdot FL(p_j, p_j^*) \quad (6)$$

where ψ and ψ_p respectively represent all anchors and the positive samples selected by the matching degree threshold. N and N_p denote the total number of all anchors and positive anchors, respectively. $FL(\cdot)$ is focal loss defined in RetinaNet (Lin et al. 2017b). w_j indicates the matching compensation factor, which is used to distinguish positive samples of different localization potential. For each ground-truth g , we first calculate its matching degree with all anchors as md . Then positive candidates can be selected according to a certain threshold, matching degree of positives is represented as md_{pos} , where $md_{pos} \subseteq md$. Supposing that the maximal matching degree for g is md_{max} , the compensation value is denoted as Δmd , we have:

$$\Delta md = 1 - md_{max}. \quad (7)$$

After that, Δmd is added to the matching degree of all positives to form the matching compensation factor:

$$w = md_{pos} + \Delta md. \quad (8)$$

With the well-designed matching compensation factor, the detector treats positive samples of different localization capability distinctively. In particular, more attention will be paid to candidates with high localization potential for classifier. Therefore, high-quality predictions can be taken through classification score, which helps to alleviate the inconsistency between classification and regression.

Since matching degree measures the localization ability of anchors, and thus it can be further used to promote high-quality localization. We formulate the matching-sensitive regression loss as follows:

$$L_{reg} = \frac{1}{N_p} \sum_{j \in \psi_p} w_j \cdot L_{smooth_{L_1}}(t_j, t_j^*) \quad (9)$$

where $L_{smooth_{L_1}}$ denotes the smooth- L_1 loss for regression. Matching compensation factor w is embedded into regression loss to avoid the loss contribution of high-quality positives being submerged in the dominant loss of samples with poor spatial alignment with GT boxes. It can be seen from Figure 3(a) that the correlation between the classification score and the localization ability of the regression box is not strong enough, which causes the prediction results selected by the classification confidence to be sometimes unreliable. After training with a matching-sensitive loss, as shown in Figure 3(b), a higher classification score accurately characterizes the better localization performance represented by the output IoU, which verifies the effectiveness of the proposed method.

Experiments

Datasets

We conduct experiments on the remote sensing datasets HRSC2016, DOTA, UCAS-AOD and a scene text dataset ICDAR 2015. The ground-truth boxes in the images are all oriented bounding boxes. The HRSC2016 (Liu et al. 2017b) is a challenging remote sensing ship detection dataset, which contains 1061 pictures. The entire dataset is divided into training set, validation set and test set, including 436, 181

	Different Variants			
with Input IoU	✓	✓	✓	✓
with Output IoU		✓	✓	✓
Uncertainty Suppression			✓	✓
Matching Sensitive Loss				✓
AP ₅₀	80.8	78.9	85.9	88.6
AP ₇₅	52.4	50.4	57.7	67.6

Table 1: Effects of each component in our method on HRSC2016 dataset.

γ	α	mAP	γ	α	mAP	γ	α	mAP
3	0.2	84.1	4	0.2	88.1	5	0.2	87.3
	0.3	88.3		0.3	88.2		0.3	88.6
	0.5	86.2		0.5	85.5		0.5	88.4
	0.7	84.1		0.7	77.9		0.7	88.1
	0.9	70.1		0.9	75.5		0.9	83.5

Table 2: Analysis of different hyperparameters on HRSC2016 dataset.

and 444 images, respectively. DOTA (Xia et al. 2018) is the largest public dataset for object detection in remote sensing imagery with oriented bounding box annotations. It contains 2806 aerial images with 188,282 annotated instances, there are 15 categories in total. Note that images in DOTA are too large, we crop images into 800×800 patches with the stride set to 200. UCAS-AOD (Zhu et al. 2015) is an aerial aircraft and car detection dataset, which contains 1510 images. We randomly divide it into training set, validation set and test set as 5:2:3. The ICDAR 2015 dataset is used for Incident Scene Text Challenge 4 of the ICDAR Robust Text Detection Challenge (Karatzas et al. 2015). It contains 1500 images, including 1000 training images and 500 test images.

Implementation Details

For the experiments, we build the baseline on RetinaNet as described above. For HRSC2016, DOTA, and UCAS-AOD, only three horizontal anchors are set with aspect ratios of $\{1/2, 1, 2\}$. For ICDAR, only five horizontal anchors are set with aspect ratios of $\{1/5, 1/2, 1, 2, 5\}$. All images are resized to 800×800. We use random flip, rotation, and HSV colour space transformation for data augmentation.

The optimizer used for training is Adam. The initial learning rate is set to 1e-4 and is divided by 10 at each decay step. The total iterations of HRSC2016, DOTA, UCAS-AOD, and ICDAR 2015 are 20k, 30k, 15k, and 40k, respectively. We train the models on RTX 2080Ti with batch size set to 8.

Ablation Study

Evaluation of different components We conduct a component-wise experiment on HRSC2016 to verify the contribution of the proposed method. The experimental results are shown in Table 1. For variant with output IoU, α is set to 0.8 for stable training, even so, detection performance

still drops from 80.8% to 78.9%. It indicates that the output IoU is not always credible for label assignment. With the suppression of regression uncertainty, the prior space alignment and posterior feature alignment can work together effectively on label assignment, and thus performance is dramatically improved by 4.8% higher than the baseline. Furthermore, the model using the matching sensitivity loss function achieves mAP of 88.6%, and the proportion of high-precision detections is significantly increased. For example, AP₇₅ is 9.9% higher than the variants with uncertainty suppression, which indicates that the matching degree guided loss effectively distinguishes anchors with differential localization capability, and pay more attention to high matching degree anchors to improve high-quality detection results.

Hyper-parameters To find suitable hyperparameter settings, and explore the relationship between parameters, we conduct parameter sensitivity experiments, and the results are shown in Table 2. In the presence of uncertainty suppression terms, as the α is reduced appropriately, the influence of feature alignment increases, and the mAP increases. It indicates that the feature alignment represented by the output IoU is beneficial to select anchors with high localization capability. However, when α is extremely large, the performance decreases sharply. The possible reason is that most potential high quality samples are suppressed by uncertainty penalty item when the output IoU can hardly provide feedback information. In this case, weakening the uncertainty suppression ability, that is, increasing γ helps to alleviate this problem and make anchor selection more stable.

Experiment Results

Comparison with other sampling methods The experimental results are shown in Table 3. Baseline model conducts label assignment based on input IoU. ATSS (Zhang et al. 2020b) has achieved great success in object detection using horizontal box. When applied to rotation object detection, there is still a substantial improvement, which is 5.3% higher than the baseline model. As for HAMBox (Liu et al. 2020), since the samples mined according to the output IoU are likely to be low-quality samples, too many mining samples may cause the network to fail to diverge, we only compensate one anchor for GT that do not match enough anchors. It outperforms 4.6% than baseline. The proposed DAL method significantly increases 7.8% compared with baseline model. Compared with the popular method ATSS, our approach considers the localization performance of the regression box, so the selected samples have more powerful localization capability, and the effect is 2.5% higher than it, which confirms the effectiveness of our method.

Results on DOTA We compare the proposed approach with other state-of-the-art methods. As shown in Table 4, we achieve the mAP of 71.44%, which outperforms the baseline model by 3%. Integrated with DAL, even the vanilla RetinaNet can compete with many advanced methods. Besides, we also embed our approach to other models to verify its universality. S²A-Net (Han et al. 2020) is an advanced rotation detector that achieves the state-of-the-art performance on DOTA dataset. It can be seen that our method further

Baseline	(Zhang et al. 2017)	HAMBox(Liu et al. 2020)	ATSS(Zhang et al. 2020b)	DAL(Ours)
80.8	82.2	85.4	86.7	88.6

Table 3: Comparisons with other label assignment strategies on HRSC2016.

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
(Xia et al. 2018)	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
(Yang et al. 2018)	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
(Jiang et al. 2017)	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
(Ma et al. 2018)	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
(Azimi et al. 2018)	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.20	53.64	62.90	67.02	64.17	50.23	68.16
(Ding et al. 2019)	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
(Pan et al. 2020)	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
(Wei et al. 2019)	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
(Yang et al. 2019b)	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
(Yang et al. 2019a)	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
(Yang and Yan 2020)	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
Baseline	88.67	77.62	41.81	58.17	74.58	71.64	79.11	90.29	82.13	74.32	54.75	60.60	62.57	69.67	60.64	68.43
Baseline+DAL	88.68	76.55	45.08	66.80	67.00	76.76	79.74	90.84	79.54	78.45	57.71	62.27	69.05	73.14	60.11	71.44
(Han et al. 2020)	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
(Han et al. 2020) +DAL	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

Table 4: Performance evaluation of OBB task on DOTA dataset. R-101 denotes ResNet-101(likewise for R-50), and H-104 stands for Hourglass-104.

Methods	Backbone	Size	NA	mAP
<i>Two-stage:</i>				
(Jiang et al. 2017)	ResNet101	800×800	21	73.07
(Liu et al. 2017b)	VGG16	-	-	75.70
(Ma et al. 2018)	ResNet101	800×800	54	79.08
(Zhang et al. 2018)	VGG16	-	24	79.60
(Ding et al. 2019)	ResNet101	512×800	5	86.20
citexu2020gliding	ResNet101	512×800	5	88.20
<i>Single-stage:</i>				
(Liao et al. 2018)	VGG16	384×384	13	84.30
(Yang et al. 2019a)	ResNet101	800×800	21	89.26
(Lin et al. 2017b)	ResNet101	800×800	121	89.18
Baseline	ResNet50	416×416	3	80.81
Baseline+DAL	ResNet50	416×416	3	88.60
Baseline+DAL	ResNet101	416×416	3	88.95
Baseline+DAL	ResNet101	800×800	3	89.77

Table 5: Comparisons with state-of-the-art detectors on HRSC2016. NA denotes the number of preset anchor at each location of feature map.

improves the performance by 2.83%, reaches the mAP of 76.95%, achieving the best results among all compared models. Some detection results on DOTA are shown in Figure 4.

Results on HRSC2016 HRSC2016 contains lots of rotated ships with large aspect ratios and arbitrary orientation. Our method achieves the state-of-the-art performances on HRSC2016, as depicted in Table 5. Using ResNet-101 as the backbone and the input image is resized to 800×800, our method has reached the highest mAP of 89.77%. Even

Methods	car	airplane	AP ₅₀	AP ₇₅
(Xia et al. 2018)	86.87	89.86	88.36	47.08
(Ding et al. 2019)	87.99	89.90	88.95	50.54
Baseline	84.64	90.51	87.57	39.15
Baseline+DAL	89.25	90.49	89.87	74.30

Table 6: Detection results on UCAS-AOD dataset.

if we use a lighter backbone ResNet50 and a smaller input scale of 416×416, we still achieve the mAP of 88.6%, which is comparable to many current advanced methods.

It is worth mentioning that our method uses only three horizontal anchors in each position, but outperforms the frameworks with a large number of anchors. This shows that it is critical to effectively utilize the predefined anchors and select high-quality samples, and there is no need to preset a large number of rotated anchors. Besides, our model is a one-stage detector, and the feature map used is $P_3 - P_7$. Compared with the $P_2 - P_6$ for two-stage detectors, the total amount of positions that need to set anchor is less, so the inference speed is faster. With input image resized to 416×416, our model reaches 34 FPS on RTX 2080 Ti GPU.

Results on UCAS-AOD Experimental results in Table 6 show that our model achieves a further improvement of 2.3% compared with baseline. Specifically, the detection performance of small vehicles is significantly improved, indicating that our method is also robust to small objects. Note that the DAL method dramatically improves AP₇₅, which reveals that the matching degree based loss function helps to pay more attention to high-quality samples and effectively distinguish

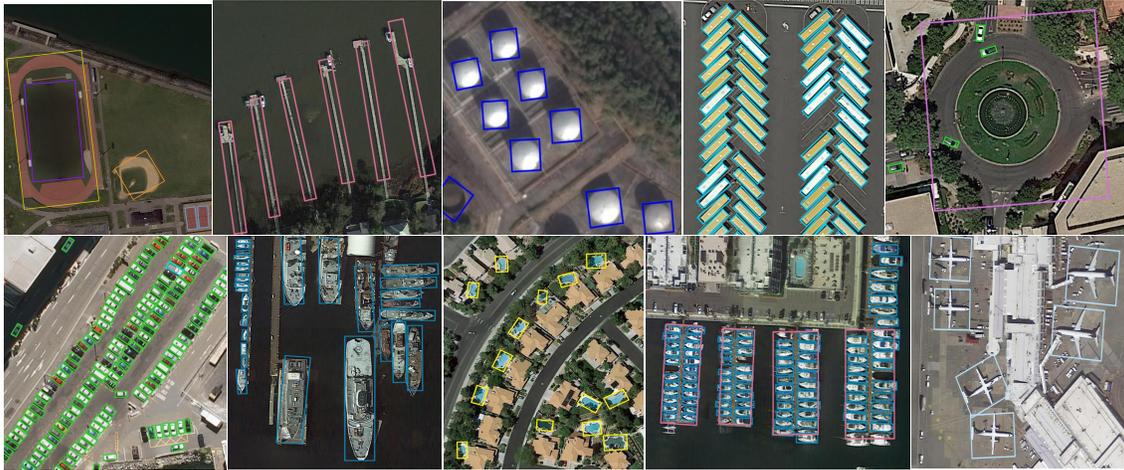


Figure 4: Visualization of detection results on DOTA with our method.

Methods	P	R	F
(Tian et al. 2016)	74.2	51.6	60.9
(Shi, Bai, and Belongie 2017)	73.1	76.8	75.0
(Ma et al. 2018)	82.2	73.2	77.4
(Yang et al. 2019b)	81.3	78.9	80.1
(Liao et al. 2018)	85.6	79.0	82.2
(Liao et al. 2020)	91.8	83.2	87.3
Baseline	77.2	77.8	77.5
Baseline+DAL	83.7	79.5	81.5
Baseline+DAL*	84.4	80.5	82.4

Table 7: Comparisons of different methods on the ICDAR 2015. P, R, F indicate recall, precision and F-measure respectively. * means multi-scale training and testing.

them to achieve high-quality object detection.

Results on ICDAR 2015 To verify the generalization of our method in different scenarios, we also conduct experiments on the scene text detection dataset. The results are shown in Table 7. Our baseline model only achieved an F-measure of 77.5% after careful parameters selection and long-term training. The proposed DAL method improves the detection performance by 4%, achieves an F-measure of 81.5%. With multi-scale training and testing, it reaches 82.4%, which was equivalent to the performance of many well-designed text detectors. However, there are a large number of long texts in ICDAR 2015 dataset, which are often mistakenly detected as several short texts. Designed for general rotation detection, DAL does not specifically consider this situation, therefore, the naive model still cannot outperform current state-of-the-art approaches for scene text detection, such as DB (Liao et al. 2020).

Experiments on Object Detection with HBB When localize objects with horizontal bounding box(HBB), label assignment still suffers from the uneven discriminative feature. Although this situation is not as severe as the rotated objects,

Dataset	Backbone	mAP/F
ICDAR 2013	RetinaNet	77.2
	RetinaNet+DAL	81.3
NWPU VHR-10	RetinaNet	86.4
	RetinaNet+DAL	88.3
VOC 2007	RetinaNet	74.9
	RetinaNet+DAL	76.1

Table 8: Performance evaluation of HBB task on ICDAR 2013 and NWPU VHR-10.

it still lays hidden dangers of unstable training. Therefore, our method is also effective in general object detection using HBB. The experimental results on ICDAR 2013 (Karatzas et al. 2013), NWPU VHR-10 (Cheng, Zhou, and Han 2016) and VOC2007 (Everingham et al. 2010) are shown in the Table 8. It can be seen that DAL achieves substantial improvements for object detection with HBB, which proves the universality of our method.

Conclusion

In this paper, we propose a dynamic anchor learning strategy to achieve high-performance arbitrary-oriented object detection. Specifically, matching degree is constructed to comprehensively considers the spatial alignment, feature alignment ability, and regression uncertainty for label assignment. Then dynamic anchor selection and matching-sensitive loss are integrated into the training pipeline to improves the high-precision detection performance and alleviate the gap between classification and regression tasks. Extensive experiments on several datasets have confirmed the effectiveness and universality of our method.

References

Azimi, S. M.; Vig, E.; Bahmanyar, R.; Körner, M.; and Reinartz, P. 2018. Towards multi-class object detection in

- unconstrained remote sensing imagery. In *Asian Conference on Computer Vision*, 150–165. Springer.
- Cheng, G.; Zhou, P.; and Han, J. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 54(12): 7405–7415.
- Choi, J.; Chun, D.; Kim, H.; and Lee, H.-J. 2019. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, 502–511.
- Choi, S.; Lee, K.; Lim, S.; and Oh, S. 2018. Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 6915–6922. IEEE.
- Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, 379–387.
- Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2849–2858.
- Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2): 303–338.
- Feng, D.; Rosenbaum, L.; and Dietmayer, K. 2018. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3266–3273. IEEE.
- Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2020. Align Deep Features for Oriented Object Detection. *arXiv preprint arXiv:2008.09397*.
- He, Y.; Zhu, C.; Wang, J.; Savvides, M.; and Zhang, X. 2019. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2888–2897.
- Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 784–799.
- Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; and Luo, Z. 2017. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*.
- Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1156–1160. IEEE.
- Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, 1484–1493. IEEE.
- Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.
- Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; and Shi, J. 2019. Consistent optimization for single-shot object detection. *arXiv preprint arXiv:1901.06563*.
- Li, B.; Liu, Y.; and Wang, X. 2019. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8577–8584.
- Li, H.; Wu, Z.; Zhu, C.; Xiong, C.; Socher, R.; and Davis, L. S. 2020. Learning from noisy anchors for one-stage object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10588–10597.
- Liao, M.; Shi, B.; and Bai, X. 2018. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing* 27(8): 3676–3690.
- Liao, M.; Wan, Z.; Yao, C.; Chen, K.; and Bai, X. 2020. Real-Time Scene Text Detection with Differentiable Binarization. In *AAAI*, 11474–11481.
- Liao, M.; Zhu, Z.; Shi, B.; Xia, G.-s.; and Bai, X. 2018. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5909–5918.
- Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.
- Liu, W.; Ma, L.; and Chen, H. 2018. Arbitrary-oriented ship detection framework in optical remote-sensing images. *IEEE Geoscience and Remote Sensing Letters* 15(6): 937–941.
- Liu, Y.; Tang, X.; Han, J.; Liu, J.; Rui, D.; and Wu, X. 2020. HAMBox: Delving Into Mining High-Quality Anchors on Face Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13043–13051. IEEE.
- Liu, Z.; Hu, J.; Weng, L.; and Yang, Y. 2017a. Rotated region based CNN for ship detection. In *2017 IEEE International Conference on Image Processing (ICIP)*, 900–904. IEEE.
- Liu, Z.; Yuan, L.; Weng, L.; and Yang, Y. 2017b. A high resolution optical satellite image dataset for ship recognition

- and some new baselines. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, volume 2, 324–331.
- Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; and Xue, X. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia* 20(11): 3111–3122.
- Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; and Xu, C. 2020. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11207–11216.
- Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; and Lin, D. 2019. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 821–830.
- Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.
- Shi, B.; Bai, X.; and Belongie, S. 2017. Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2550–2558.
- Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 761–769.
- Song, G.; Liu, Y.; and Wang, X. 2020. Revisiting the sibling head in object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11563–11572.
- Tian, Z.; Huang, W.; He, T.; He, P.; and Qiao, Y. 2016. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*, 56–72. Springer.
- Wei, H.; Zhou, L.; Zhang, Y.; Li, H.; Guo, R.; and Wang, H. 2019. Oriented objects as pairs of middle lines. *arXiv preprint arXiv:1912.10694*.
- Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3974–3983.
- Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; and Yu, G. 2019a. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*.
- Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; and Guo, Z. 2018. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sensing* 10(1): 132.
- Yang, X.; and Yan, J. 2020. Arbitrary-Oriented Object Detection with Circular Smooth Label. *arXiv preprint arXiv:2003.05597*.
- Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019b. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, 8232–8241.
- Zhang, H.; Chang, H.; Ma, B.; Wang, N.; and Chen, X. 2020a. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. *arXiv preprint arXiv:2004.06002*.
- Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020b. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9759–9768.
- Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; and Li, S. Z. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, 192–201.
- Zhang, X.; Wan, F.; Liu, C.; Ji, R.; and Ye, Q. 2019. Freeanchor: Learning to match anchors for visual object detection. In *Advances in Neural Information Processing Systems*, 147–155.
- Zhang, Z.; Guo, W.; Zhu, S.; and Yu, W. 2018. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters* 15(11): 1745–1749.
- Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; and Jiao, J. 2015. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing (ICIP)*, 3735–3739. IEEE.