

Pixel-level Quality Assessment for Oriented Object Detection

Yunhui Zhu¹, Buliao Huang^{2*}

¹School of Computer Science, Nanjing Audit University, Nanjing 211815, China

²School of Computer Science and Engineering, Jinling Institute of Technology, Nanjing 211169, China
yh Zhu@nau.edu.cn, hbl1995@jit.edu.cn

Abstract

Modern oriented object detectors typically predict a set of bounding boxes and select the top-ranked ones based on estimated localization quality. Achieving high detection performance requires that the estimated quality closely aligns with the actual localization accuracy. To this end, existing approaches predict the Intersection over Union (IoU) between the predicted and ground-truth (GT) boxes as a proxy for localization quality. However, box-level IoU prediction suffers from a structural coupling issue: since the predicted box is derived from the detector’s internal estimation of the GT box, the predicted IoU—based on their similarity—can be overestimated for poorly localized boxes. To overcome this limitation, we propose a novel Pixel-level Quality Assessment (PQA) framework, which replaces box-level IoU prediction with the integration of pixel-level spatial consistency. PQA measures the alignment between each pixel’s relative position to the predicted box and its corresponding position to the GT box. By operating at the pixel level, PQA avoids directly comparing the predicted box with the estimated GT box, thereby eliminating the inherent similarity bias in box-level IoU prediction. Furthermore, we introduce a new integration metric that aggregates pixel-level spatial consistency into a unified quality score, yielding a more accurate approximation of the actual localization quality. Extensive experiments on HRSC2016 and DOTA demonstrate that PQA can be seamlessly integrated into various oriented object detectors, consistently improving performance (e.g., +5.96% AP_{50:95} on Rotated RetinaNet and +2.32% on STD).

Extended version —

<https://doi.org/10.48550/arXiv.2511.08186>

Introduction

Oriented object detection, a task focused on accurately classifying and localizing objects with arbitrary orientations in images, has garnered significant attention. Modern oriented object detectors (Yu et al. 2024; Pu et al. 2023b; Hou et al. 2022; Huang et al. 2022; Xie et al. 2021) typically generate a set of bounding boxes, rank them according to their classification scores, and select the top-ranked ones as final detections. To achieve high detection performance, it is

*Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

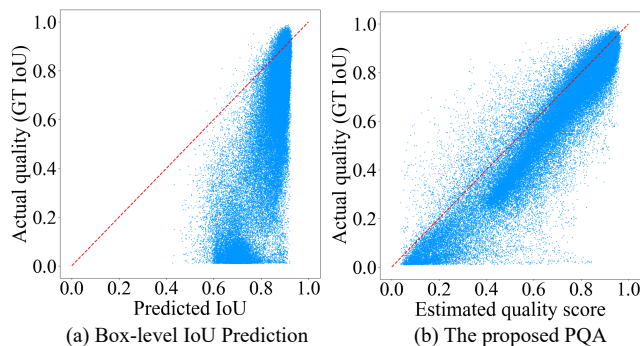


Figure 1: Correlation between the estimated and actual localization quality of predicted oriented boxes.

essential that detections with better localization quality are ranked higher. However, classification scores are primarily optimized for category discrimination and do not necessarily reflect localization accuracy. This discrepancy can lead to suboptimal ranking of predicted boxes, thereby degrading average precision, particularly under high Intersection over Union (IoU) thresholds (Kahraman et al. 2023).

Motivated by this issue, a number of studies have aimed to accurately estimate the localization quality. In general, the actual localization quality of a predicted box is quantified by its GT IoU with the GT box. However, since the GT box is unavailable during inference, existing methods (Xi et al. 2024; Ming et al. 2023; Pu et al. 2023a; Zhang et al. 2021; Li et al. 2021) typically predict the IoU between the predicted and GT boxes to approximate the actual localization quality. Despite its intuitive appeal, this approach may yield unreliable estimates. As illustrated in Figure 1(a), predicted boxes with low GT IoU values may still receive high predicted IoU scores. We attribute this to IoU prediction being performed at the box level, where it measures the similarity between the predicted box and the detector’s internal estimation of the GT box. However, since the predicted box is itself derived from this internal estimation, it naturally shares a high similarity with the estimated GT box. This structural coupling could lead to overestimated IoU scores for poorly localized boxes, undermining the reliability of box-level IoU prediction.

To address this limitation, we propose a novel Pixel-

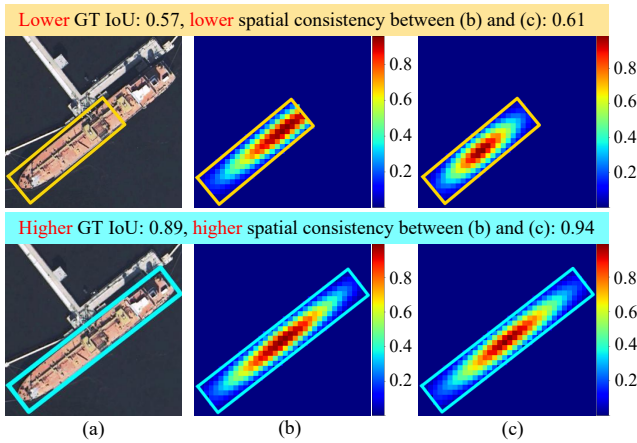


Figure 2: Illustrative examples of how pixel-level spatial consistency correlates with GT IoU. (a) Predicted oriented boxes. (b) Heatmap encoding pixel-wise relative positions to the GT box, where higher values indicate closer proximity to the box center. (c) Heatmap encoding pixel-wise relative positions to the predicted box, represented analogously.

level Quality Assessment (PQA) framework, which replaces box-level IoU prediction with the integration of pixel-level spatial consistency. Spatial consistency is measured as the alignment between a pixel’s relative positions to the predicted box and the GT box. This design is motivated by the observation that, for a well-localized predicted box, each pixel’s relative position to the predicted box should closely match its position relative to the GT box. As shown in Figure 2, the predicted box with a higher GT IoU exhibits stronger pixel-level spatial consistency. By operating at the pixel level, PQA avoids directly comparing the predicted box with the estimated GT box, thereby eliminating the inherent similarity bias in box-level IoU prediction and mitigating the resulting overestimation of localization quality. This advantage is further validated in Figure 1(b), which shows that the proposed PQA framework establishes a robust positive correlation between the estimated localization quality and the actual GT IoU. Additionally, to convert pixel-level spatial consistency within a predicted box into a quantifiable quality score, we introduce a novel integration metric that provides a closer approximation to the GT IoU. This aggregation strategy also improves robustness to potential noise in pixel-level spatial consistency, ensuring a more stable estimation of localization quality. This stands in stark contrast to conventional box-level IoU prediction, which directly regresses a scalar IoU value as the quality score, making it inherently sensitive to prediction errors.

To verify the effectiveness of PQA, we conducted extensive experiments on two popular oriented object detection datasets: HRSC2016 (Liu et al. 2017) and DOTA (Xia et al. 2018). Experiments demonstrate that the proposed PQA can be integrated into various oriented detectors and consistently enhances their performance (e.g., +5.96% $AP_{50:95}$ on Rotated RetinaNet (Lin et al. 2017) and +2.32% on STD (Yu et al. 2024)). The main contributions of this paper are sum-

marized as follows:

- This paper proposes a novel Pixel-level Quality Assessment (PQA) framework that replaces conventional box-level IoU prediction with the integration of pixel-level spatial consistency, effectively eliminating inherent similarity bias and mitigating the overestimation of localization quality.
- This paper introduces a novel integration metric that aggregates pixel-level spatial consistency into a robust and unified quality score, enabling a more accurate approximation of the GT IoU while remaining resilient to prediction noise.
- The proposed PQA framework can be seamlessly integrated into existing oriented object detectors, consistently improving localization quality assessment and overall detection performance.

Related Work

Oriented Object Detection

Based on generic object detection frameworks, recent studies have sought to enhance oriented detection performance from various perspectives. Some methods have concentrated on effective oriented proposal generation (Cheng et al. 2022; Xie et al. 2021; Ding et al. 2019) or more powerful feature extraction (Lee et al. 2024; Pu et al. 2023b; Zhen et al. 2023; Yang et al. 2021a; Han et al. 2021). For example, ARC (Pu et al. 2023b) proposed an adaptive rotated convolution module to extract high-quality features for objects with varying orientations. In addition, several approaches have explored more flexible representations for oriented objects (Yu and Da 2023; Li et al. 2022; Xu et al. 2021), as well as more effective loss functions for localization (Murrugarra-Llerena et al. 2024; Yang et al. 2023b, 2021b). Moreover, various approaches have investigated advanced label assignment strategies (Xu et al. 2023; Hou et al. 2022; Huang et al. 2022). Specifically, SASM (Hou et al. 2022) selected positive training samples based on the shape and distribution information of oriented objects. Furthermore, multiple methods have attempted to enhance the potential of visual transformers (Yu et al. 2024; Zeng et al. 2024; Dai et al. 2023). For example, STD (Yu et al. 2024) divided the prediction process of oriented boxes into multiple stages to improve the performance of visual transformers. Despite the progress made by these diverse approaches, most of them relied on the predicted classification scores to assess the quality of predicted oriented boxes and rarely explored the impact of different quality assessment methods on detection performance.

Localization Quality Assessment

Most existing detectors (Yu et al. 2024; Pu et al. 2023b; Yu and Da 2024; Hou et al. 2022; Xie et al. 2021; Lin et al. 2017) relied solely on the predicted classification scores to evaluate the quality of predicted boxes. However, the inconsistency between classification scores and actual localization quality often impeded the performance of detectors. To address this issue, a number of methods (Xi et al. 2024; Li

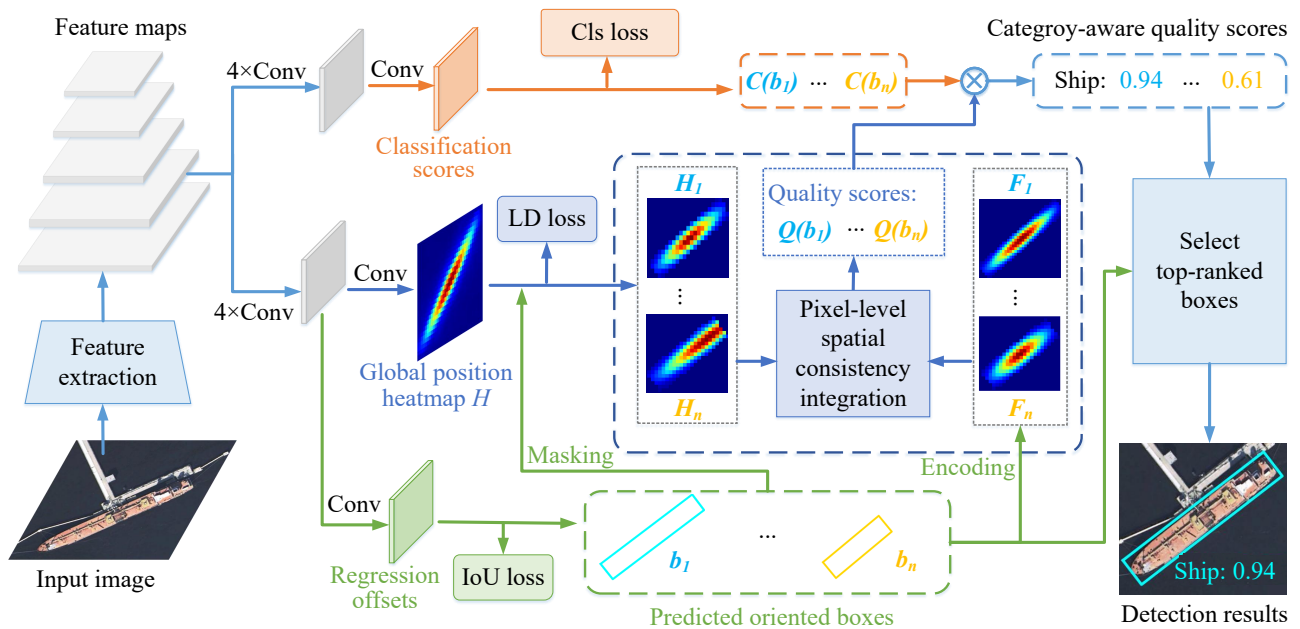


Figure 3: Overall framework of PQA. H_i and F_i provide spatial encodings of pixels’ relative positions to the nearest GT box and to the predicted box b_i , respectively.

et al. 2021; Ge et al. 2021; Kim and Lee 2020; Tian et al. 2019; Jiang et al. 2018) introduced an auxiliary branch to predict IoU or centerness values, which were then combined with classification scores. Additionally, some methods (Ming et al. 2023; Zhang et al. 2021) jointly modeled classification and IoU prediction, enabling the classification scores to reflect localization quality. For instance, TIOE (Ming et al. 2023) introduced a hierarchical alignment label that stratified IoU values into discrete levels to supervise the classifier. Moreover, several works (Kahraman et al. 2023; Oksuz et al. 2021) designed ranking-based loss functions to enforce the predicted classification scores to follow the same order as the localization quality of the corresponding boxes. In contrast to existing approaches, the proposed PQA transforms box-level estimation into the integration of pixel-level spatial consistency, reducing overestimation of localization quality and ensuring a clear positive correlation between the estimated quality score and the GT IoU.

Method

Overview of PQA Framework

The overall framework of PQA is illustrated in Figure 3. For each predicted oriented box b_i , PQA estimates its localization quality $Q(b_i)$ through the integration of pixel-level spatial consistency within b_i . Specifically, PQA first adds a convolutional layer to predict a global position heatmap H , which encodes the relative positions of all pixels with respect to their nearest GT oriented boxes. Next, for each predicted box b_i , PQA applies a spatial mask to H to extract a localized heatmap H_i , which activates only the pixels within b_i . Simultaneously, the relative positions of these activated pixels to b_i itself are encoded as a heatmap F_i .

Finally, PQA introduces an integration metric to aggregate the spatial consistency between H_i and F_i , producing the final quality score $Q(b_i)$. This score effectively captures the pixel-level alignment between the predicted box and the GT box, addressing the structural coupling issue inherent in traditional box-level IoU prediction. The following sections provide detailed descriptions of the prediction of the global position heatmap H and the integration of pixel-level spatial consistency.

Global Position Heatmap Prediction

As shown in Figure 3, PQA incorporates a 3×3 convolution layer in the localization subnet to predict the global position heatmap H that encodes the relative positions of all pixels with regard to their nearest GT oriented boxes. This subsection details the label definition and loss function (LD loss) used for global position heatmap prediction.

The global position heatmap label is designed to satisfy two conditions: 1) pixels outside the GT oriented box are assigned a value of 0, while those inside are assigned values greater than 0; 2) for pixels inside the GT box, the values vary with the distance to the box center, incorporating both the orientation and size of the GT box to encode the pixel’s relative position. Inspired by prior works (Yang et al. 2021b; Huang et al. 2022) that represent oriented boxes as Gaussian distributions, PQA generates the global position heatmap label based on 2D Gaussian functions, providing a geometry-aware representation.

Formally, given an image I with M objects, let $G = \{g_j\}_{j=1}^M, g_j \in \mathbb{R}^5$ be the set of GT oriented boxes for these objects. For any pixel p in the image I , the label $H^*(p)$

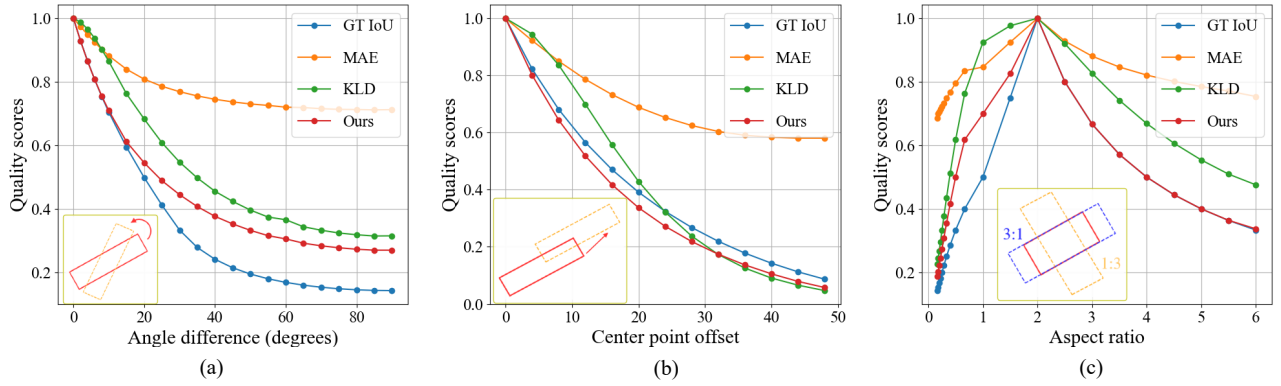


Figure 4: Comparison of quality scores computed using different integration metrics for pixel-level spatial consistency, as the predicted oriented boxes vary in (a) orientation angle, (b) center point offset, and (c) aspect ratio.

global position heatmap at pixel p is defined as:

$$H^*(p) = \max\{\Phi(p|g_j)\}_{j=1}^M. \quad (1)$$

For any $g_j \in G$, $\Phi(p|g_j)$ is calculated as follows:

$$\Phi(p|g_j) = \begin{cases} e^{-\frac{1}{2}(p-g_j^\mu)^T \Sigma_{g_j}^{-1}(p-g_j^\mu)} & \text{if } p \text{ is inside } g_j \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Here, g_j^μ denotes the center point of the object g_j . Σ_{g_j} represents the covariance matrix of the object g_j and is calculated as follows:

$$\begin{aligned} \Sigma_{g_j}^{1/2} &= U \Lambda U^T \\ &= \begin{pmatrix} \cos g_j^\theta & -\sin g_j^\theta \\ \sin g_j^\theta & \cos g_j^\theta \end{pmatrix} \begin{pmatrix} \frac{g_j^w}{4} & 0 \\ 0 & \frac{g_j^h}{4} \end{pmatrix} \begin{pmatrix} \cos g_j^\theta & \sin g_j^\theta \\ -\sin g_j^\theta & \cos g_j^\theta \end{pmatrix}. \end{aligned} \quad (3)$$

Here, U is the rotation matrix and Λ is the eigenvalue matrix. The parameters g_j^θ , g_j^w , and g_j^h denote the angle, width, and height of the GT oriented box g_j , respectively. According to Eqs. (1)-(3), $H^*(p)$ is associated with the relative position of pixel p to the center of the GT box, while accounting for both the box's orientation and size. Notably, beyond the Gaussian-based formulation, $H^*(p)$ can also be defined as the centerness (Tian et al. 2019) of pixel p . In the experimental section, Table 2 compares these two label definitions, showing that PQA consistently outperforms existing methods under both settings, which demonstrates the superiority and robustness of the PQA framework.

The LD loss \mathcal{L}_{ld} is inspired by the Focal Loss framework (Lin et al. 2017), particularly in its treatment of negative samples ($H^*(p) = 0$), where it focuses more on hard negatives while down-weighting the loss from abundant easy negatives. Moreover, to optimize the predicted continuous values for positive samples ($H^*(p) > 0$), we adopt the formulation from GFL (Li et al. 2020). Specifically, the LD loss \mathcal{L}_{ld} is defined as:

$$\begin{aligned} \mathcal{L}_{ld} &= (1/\sum_p H^*(p)) \sum_p L(H(p), H^*(p)), \quad \text{where} \\ L(x, y) &= \begin{cases} -\alpha|y-x|(y \log x + (1-y) \log(1-x)), & y > 0 \\ -(1-\alpha)(1-x)^\beta \log x, & y = 0. \end{cases} \end{aligned} \quad (4)$$

Here, H is the predicted global position heatmap. α and β are set to 0.25 and 2, following (Lin et al. 2017).

Pixel-level Spatial Consistency Integration

This subsection details the integration of pixel-level spatial consistency between H_i and F_i to produce the final quality score $Q(b_i)$.

Let $B = \{b_i\}_{i=1}^N$, $b_i \in \mathbb{R}^5$ be the set of all N predicted oriented boxes. For each predicted box b_i , PQA applies a spatial mask to the global position heatmap H to extract a localized heatmap H_i , which activates only the pixels within b_i . Let $P(b_i)$ represent the set of pixels activated inside b_i . The value $H_i(p)$ at any pixel p is defined as follows:

$$H_i(p) = \begin{cases} H(p) & \text{if } p \in P(b_i) \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Next, PQA encodes the relative positions of these activated pixels to b_i itself into F_i based on Eqs. (2) and (3). $F_i(p)$ is defined as follows:

$$F_i(p) = \Phi(p|b_i). \quad (6)$$

Subsequently, the quality score $Q(b_i)$ for the predicted box b_i is computed by aggregating the pixel-level spatial consistency between H_i and F_i . Since $Q(b_i)$ is derived from all pixels inside b_i , it exhibits robustness to potential noise in pixel-level spatial consistency. In contrast, conventional box-level IoU prediction directly regresses a scalar IoU value as the quality score, making it inherently sensitive to prediction errors. The robustness of our pixel-level integration strategy is further validated in Supplementary (Zhu and Huang 2025). Formally, $Q(b_i)$ is computed as follows:

$$Q(b_i) = S(H_i, F_i), \quad (7)$$

which can be instantiated using various integration metrics $S(\cdot)$. Once the quality score $Q(b_i)$ is calculated, it is combined with the classification score $C(b_i)$ of the predicted box to obtain a category-aware quality score:

$$CQ(b_i) = C(b_i) \times Q(b_i). \quad (8)$$

The most straightforward integration metric $S(\cdot)$ is the Mean Absolute Error (MAE) between H_i and F_i . However, as shown in Figure 4, the resulting quality score Q is overly smooth and lacks sufficient discriminability, making it highly sensitive to classification noise and thereby degrading detection performance. A detailed analysis is provided in Supplementary (Zhu and Huang 2025). To mitigate

this issue, we adopt the widely-used Kullback-Leibler Divergence (KLD) as the integration metric. As illustrated in Figure 4, the KLD-based approach yields more discriminative scores. Nevertheless, KLD inherently emphasizes pixels with higher values in the heatmap, which may not align well with the objectives of our task. A more in-depth discussion is available in Supplementary (Zhu and Huang 2025).

To ensure that the calculated quality score aligns more closely with the GT IoU, we have redesigned the integration metric inspired by the conventional IoU calculation used for detection boxes. The proposed metric treats the two heatmaps H_i and F_i as two three-dimensional objects and computes the volume IoU between them, defined as:

$$S(H_i, F_i) = \frac{\sum_{p \in P(b_i)} \min(H_i(p), F_i(p))}{\sum_{p \in P(b_i)} \max(H_i(p), F_i(p))}. \quad (9)$$

Compared to KLD and MAE, the proposed metric offers better distinction and consistently demonstrates a smaller discrepancy between the calculated quality score and the GT IoU value, as shown in Figure 4. The experimental section also compares the performance of PQA with different integration metrics.

The total training loss of PQA is defined as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{loc} + \lambda \mathcal{L}_{ld}, \quad (10)$$

where \mathcal{L}_{cls} and \mathcal{L}_{loc} represent the classification and localization losses, respectively. In practice, \mathcal{L}_{cls} employs Focal loss (Lin et al. 2017), and \mathcal{L}_{loc} uses Rotated IoU loss (Zhou et al. 2019). λ serves as a scaling factor, and its various values are evaluated in Supplementary (Zhu and Huang 2025).

Experiments

Experimental Settings

Datasets Two widely used oriented object detection datasets, DOTA-v1.0 (Xia et al. 2018) and HRSC2016 (Liu et al. 2017), are employed for performance evaluation. DOTA is a large-scale dataset comprising 2806 images and 188,282 instances. It includes fifteen object classes: 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). The images in DOTA vary in size from 800×800 to 4000×4000 pixels. In line with common practice, we crop the original images into 1024×1024 patches with an overlap of 200. HRSC2016 is a widely used dataset for ship detection, comprising 1680 images with dimensions ranging from 339 to 1333 pixels. The dataset includes a training set of 436 images, a validation set of 181 images, and a test set of 444 images. Following common practice for the two datasets, both the training and validation sets are used for training, while the test set is reserved for testing. For performance evaluation, we report the average precision results under IoU thresholds of 0.5 (AP_{50}) and 0.75 (AP_{75}), as well as the average precision averaged over IoU thresholds from 0.5 to 0.95 in steps of 0.05 ($AP_{50:95}$).

Method	DOTA-v1.0			HRSC2016		
	$AP_{50:95}$	AP_{50}	AP_{75}	$AP_{50:95}$	AP_{50}	AP_{75}
Baseline	36.07	65.17	34.00	62.48	90.09	77.21
Auxiliary Head						
Centerness	36.77	66.42	35.27	62.92	89.83	76.71
IoU	36.79	66.52	35.12	63.16	89.77	75.90
SOOD	36.53	66.21	34.95	63.19	89.90	76.82
IoU-aware Cls.						
TIOE	36.03	65.74	33.88	63.98	90.12	76.71
VFL (Zhang et al. 2021)	36.54	65.97	34.03	62.08	89.69	76.54
Ranking-based Loss						
Correlation Loss	36.06	64.79	34.81	63.61	89.84	77.35
RS Loss	35.91	65.77	33.62	63.05	90.03	76.16
PQA	38.46 ($\uparrow 2.39$)	67.43	36.44	66.28 ($\uparrow 3.80$)	90.17	79.07

Table 1: Comparison with existing quality assessment methods.

Method	Label Definition		Integration Metric			$AP_{50:95}$	AP_{50}	AP_{75}
	Centerness	Gaussian	MAE	KLD	Ours			
Baseline						36.07	65.17	34.00
PQA		✓	✓			37.97	67.22	36.10
		✓		✓		38.04	67.16	36.18
	✓				✓	38.08	67.41	35.78
		✓			✓	38.46	67.43	36.44

Table 2: Performance of PQA with different configurations.

Implementation Details The implementation of PQA is based on the MMRotate toolbox (Zhou et al. 2022). The initial learning rate is set to 2.5×10^{-3} with two images per mini-batch. For the DOTA-v1.0 dataset, the training runs for 12 epochs, with the learning rate reduced by a factor of 10 after 8 and 11 epochs. For the HRSC2016 dataset, training lasts for 72 epochs, with the learning rate decaying after 48 and 66 epochs. During training, only random horizontal, vertical, and diagonal flipping are used for image augmentation, with no additional tricks employed. No data augmentation is applied during testing. Unless otherwise specified, the ablation study is conducted on the DOTA-v1.0 dataset, and results are reported on the validation set. The baseline model is Rotated RetinaNet (Lin et al. 2017) with a ResNet-50 backbone (He et al. 2016). It adopts the Rotated IoU loss (Zhou et al. 2019) instead of the Smooth L_1 loss (Girshick 2015) for localization optimization, and employs SimOTA (Ge et al. 2021) in place of the Max-IoU strategy (Ren et al. 2015) for label assignment. Classification scores are directly used as estimates of localization quality.

Ablation Studies

Compare with Other Quality Assessment Methods Existing quality assessment methods can be broadly categorized into three types: 1) Auxiliary head methods that introduce an additional branch to predict IoU or centerness values; 2) IoU-aware classification methods that use GT IoU as the classification confidence label for positive samples; and 3) ranking-based loss methods that enforce consistency between the rankings of predicted classification scores and GT IoUs. To validate the effectiveness of the proposed PQA

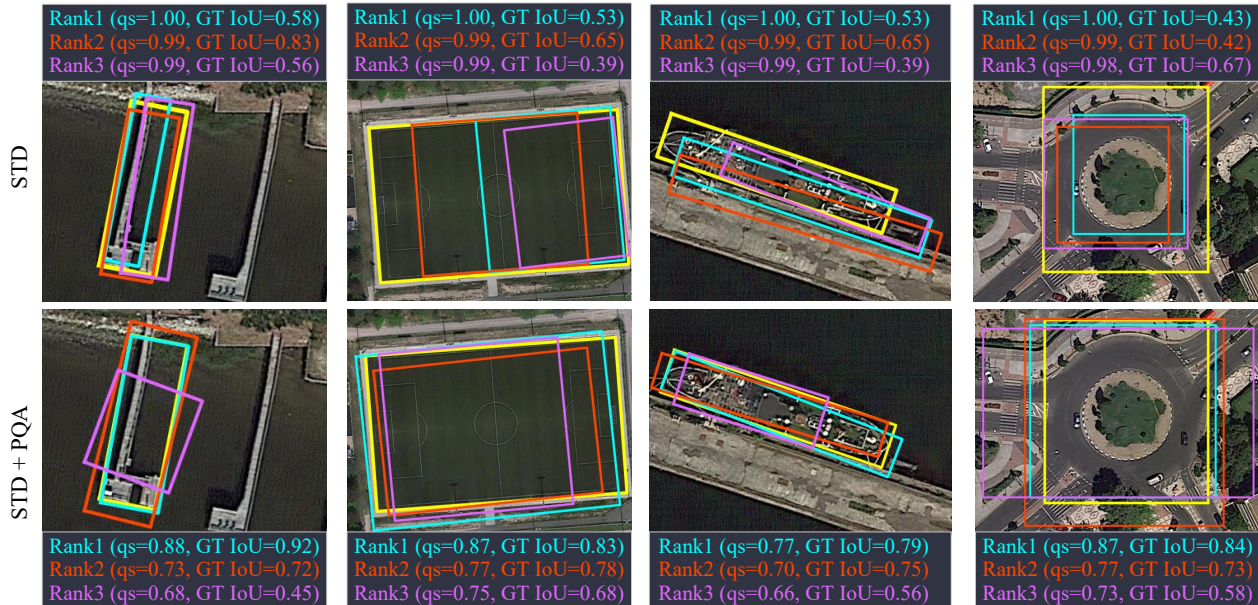


Figure 5: Visualization of top-ranked predicted oriented boxes with their estimated quality scores (qs) and actual localization quality (GT IoU) before (row 1) and after (row 2) integrating PQA into STD. Predicted boxes are ranked based on qs. Yellow denotes GT boxes; cyan, red, and purple indicate the rank-1, rank-2, and rank-3 predicted boxes, respectively.

Method	PQA	DOTA-v1.0			HRSC2016		
		AP _{50:95}	AP ₅₀	AP ₇₅	AP _{50:95}	AP ₅₀	AP ₇₅
SASM		33.61	60.40	32.71	59.89	86.96	68.50
SASM	✓	35.49 (↑1.88)	62.38	34.93	66.61 (↑6.72)	89.99	79.49
Rotated RetinaNet		35.82	64.76	33.79	50.33	85.17	49.61
Rotated RetinaNet	✓	37.28 (↑1.46)	65.82	35.81	56.29 (↑5.96)	87.85	62.34
Rotated FCOS		37.09	67.44	35.16	60.78	89.11	74.08
Rotated FCOS	✓	38.25 (↑1.16)	69.03	35.17	65.49 (↑4.71)	90.14	79.48
STD		45.69	74.87	48.06	70.07	90.49	88.69
STD	✓	46.82 (↑1.13)	74.99	49.97	72.39 (↑2.32)	90.63	89.41

Table 3: Performance of PQA integrated into various oriented object detectors.

Method	AP _{50:95}	AP ₅₀	AP ₇₅	Params (M) ↓	FLOPs (G) ↓	FPS ↑
Baseline	36.07	65.17	34.00	36.05	207.87	54.2
PQA	38.46 (↑2.39)	67.43	36.44	36.07	208.28	46.9
PQA-Lite	37.69 (↑1.62)	65.27	36.48	36.07	208.28	52.0

Table 4: Comparison of full and lightweight PQA variants.

framework, we compare it with recent and representative works from each category, including Centerness (Tian et al. 2019), IoU (Jiang et al. 2018), SOOD (Xi et al. 2024), TIOE (Ming et al. 2023), VFL (Zhang et al. 2021), Correlation Loss (Kahraman et al. 2023), and RS Loss (Oksuz et al. 2021). For fair comparison, all methods share the same detection architecture as PQA, differing only in their quality assessment strategies. As shown in Table 1, PQA achieves significant improvements of 2.39% and 3.80% in AP_{50:95} over the baseline on the DOTA and HRSC2016 datasets, respectively. Furthermore, it outperforms all compared meth-

ods by at least 1.67% and 2.30% on the two datasets, demonstrating the superiority and effectiveness of PQA.

Different Configurations of PQA We evaluate different label definitions for the global position heatmap and various integration metrics for pixel-level spatial consistency. Both centerness (Tian et al. 2019) and the Gaussian-based formulation can be employed as labels for the global position heatmap, while commonly used metrics such as MAE and KLD, together with our proposed metric (Eq. (9)), serve as integration metrics. The comparison results are presented in Table 2. It can be observed that, regardless of the choice of label definition or integration metric, PQA consistently outperforms the baseline by more than 1.9%, demonstrating its robustness and overall superiority. Moreover, our proposed integration metric achieves the best performance, which is consistent with the observation in Figure 4.

Compatibility for Various Oriented Detectors To verify the compatibility of PQA, it is integrated into various oriented detectors: SASM(Hou et al. 2022), Rotated RetinaNet (Lin et al. 2017), Rotated FCOS (Tian et al. 2019), and STD (Yu et al. 2024), encompassing both anchor-based and anchor-free, single-stage and two-stage methods. Table 3 presents the numerical results of these detectors with and without PQA. The results show that PQA consistently brings significant performance improvements across various oriented detectors, particularly in AP₇₅ and AP_{50:95}.

Computational Overhead and Lightweight Variant While PQA yields notable gains in AP, it incurs a modest computational overhead. As presented in Table 4, it introduces only 0.02M additional parameters and 0.41G FLOPs

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP _{50:95}	AP ₅₀	AP ₇₅
Rotated FCOS (Tian et al. 2019)	R-50	89.06	76.97	47.92	58.55	79.78	76.95	86.90	90.90	84.87	84.58	57.11	64.68	63.69	69.38	46.87	39.80	71.88	37.30
Gliding Vertex (Xu et al. 2021)	R-50	89.20	75.92	51.31	69.56	78.11	75.63	86.87	90.90	85.40	84.77	53.36	66.65	66.31	69.99	54.39	39.52	73.22	37.47
S ² A-Net (Han et al. 2021)	R-50	89.25	81.19	51.55	71.39	78.61	77.37	86.77	90.89	86.28	84.64	61.21	65.65	66.07	67.57	50.18	39.05	73.91	35.52
R ³ Det (Yang et al. 2021a)	R-50	89.29	75.21	45.41	69.23	75.53	72.89	79.28	90.88	81.02	83.25	58.81	63.15	63.40	62.21	37.41	37.82	69.80	36.59
SASM (Hou et al. 2022)	R-50	87.51	80.15	51.07	70.35	74.95	75.80	84.23	90.90	80.87	84.93	58.51	65.59	69.74	70.18	42.31	43.01	72.47	44.21
O-Reppoints (Li et al. 2022)	R-50	87.78	77.67	49.54	66.46	78.51	73.11	86.58	90.86	83.75	84.34	53.14	65.63	63.70	68.71	45.91	40.88	71.71	41.39
H2RBox (Yang et al. 2023a)	R-50	88.16	80.47	40.88	61.27	79.78	75.25	84.40	90.89	80.05	85.35	58.91	68.46	63.67	71.87	47.18	41.49	71.77	41.42
KFloU (Yang et al. 2023b)	R-50	89.20	76.40	51.64	70.15	78.31	76.43	87.10	90.88	81.68	82.22	64.65	64.84	66.77	70.68	49.52	41.70	73.37	42.71
AO2-DETR (Dai et al. 2023)	R-50	87.99	79.46	45.74	66.64	78.90	73.90	73.30	90.40	80.55	85.89	55.19	63.62	51.83	70.15	60.04	33.31	70.91	22.60
RoI Trans. (Ding et al. 2019)	Swin-T	88.44	85.53	54.56	74.55	73.43	78.39	87.64	90.88	87.23	87.11	64.25	63.27	77.93	74.10	60.03	47.60	76.49	50.15
PSC (Yu and Da 2023)	R-50	89.65	83.80	43.64	70.98	79.00	71.35	85.08	90.90	84.28	82.51	60.64	65.06	62.52	69.61	54.00	43.98	72.87	46.18
ARS-DETR (Zeng et al. 2024)	Swin-T	87.65	76.54	50.64	69.85	79.76	83.91	87.92	90.26	86.24	85.09	54.58	67.01	75.62	73.66	63.39	47.77	75.47	51.77
Oriented RCNN (Xie et al. 2021)	R-50	89.47	83.84	55.42	72.93	78.71	83.86	88.00	90.90	87.07	84.45	63.26	69.07	76.21	69.03	51.17	46.88	75.67	50.41
Oriented RCNN (Pu et al. 2023b)	ARC-R50	89.46	82.84	54.64	75.57	79.24	84.25	88.16	90.90	86.21	85.69	63.80	65.92	74.70	69.94	60.80	47.63	76.81	51.72
STD (Yu et al. 2024)	ViT-B	89.56	83.18	56.45	70.45	79.39	85.84	88.39	90.88	85.32	86.54	61.56	68.29	76.81	72.92	69.81	48.26	77.68	53.12
STD + PQA	ViT-B	89.54	82.58	55.55	72.89	80.48	85.63	88.28	90.91	86.01	85.16	59.91	70.41	77.79	71.81	73.77	49.47 ($\uparrow 1.21$)	78.05	53.91

Table 5: Performance comparison on the DOTA-v1.0 dataset. R-50, Swin-T, and ViT-B represent ResNet50 (He et al. 2016), Swin-Transformer (Liu et al. 2021), Vision Transformer-Base (Dosovitskiy et al. 2020) respectively.

Method	Backbone	AP _{50:95}	AP ₅₀	AP ₇₅
Rotated RetinaNet (Lin et al. 2017)	R-50	50.33	85.17	49.61
Rotated FCOS (Tian et al. 2019)	R-50	60.78	87.85	62.34
S ² A-Net (Han et al. 2021)	R-50	57.92	90.18	67.11
SASM (Hou et al. 2022)	R-50	59.89	86.96	68.50
AO2-DETR (Dai et al. 2023)	R-50	51.13	87.70	54.90
PSC (Yu and Da 2023)	R-50	67.57	90.06	78.56
ProbloU (Murrugarra-Llerena et al. 2024)	R-50	61.65	90.23	78.46
Oriented RCNN (Xie et al. 2021)	R-50	70.18	90.33	89.09
Oriented RCNN (Pu et al. 2023b)	ARC-R50	71.10	90.47	89.32
STD (Yu et al. 2024)	ViT-B	70.07	90.49	88.69
STD + PQA	ViT-B	72.39 ($\uparrow 2.32$)	90.63	89.41

Table 6: Performance comparison on HRSC2016.

compared to the baseline, with inference speed decreasing from 54.2 to 46.9 FPS. This drop is primarily due to the integration of pixel-level spatial consistency across all predicted boxes and their internal pixels. To alleviate this overhead, we propose a lightweight variant, PQA-Lite, which restricts spatial consistency integration to high-confidence predicted boxes and reduces the number of sampled pixels per box. PQA-Lite achieves performance comparable to the full version while improving inference speed to 52.0 FPS, highlighting the efficiency and adaptability of PQA. Details of PQA-Lite see Supplementary (Zhu and Huang 2025).

Comparison with State-of-the-arts

DOTA-v1.0 Table 5 provides a comparison between PQA and state-of-the-art methods on the DOTA-v1.0 dataset. To ensure a fair evaluation, all methods adopt the same data augmentation strategies and employ single-scale training and testing. When integrated into the advanced detector STD, PQA yields a 1.21% improvement in AP_{50:95}, achieving a competitive AP₅₀ of 78.05% and AP_{50:95} of 49.47% without relying on any test-time enhancements.

HRSC2016 Table 6 presents the evaluation results on the HRSC2016 dataset. Integrating PQA into the detector STD

consistently improves performance, yielding a 2.32% gain in AP_{50:95} and achieving a remarkable AP₅₀ of 90.63% and AP_{50:95} of 72.39%. These results are highly competitive and compare favorably with existing approaches.

Visualization

Figure 5 compares the ranking of predicted oriented boxes before (row 1) and after (row 2) integrating PQA into STD. During inference, box ranking is determined by the estimated quality score (denoted as qs in Figure 5), and a reliable detector should yield consistent rankings with the GT IoU values. Prior to applying PQA, the qs in STD is derived from the classification score, which shows limited variation (e.g., 1.0 vs. 0.99) and insufficient discriminability. As shown in row 1, some rank-1 boxes still have lower GT IoU values, indicating that classification-based qs fails to reflect actual localization quality. In contrast, after incorporating PQA (row 2), the estimated qs closely aligns with the GT IoU, leading to more accurate and consistent rankings. This highlights the effectiveness of PQA in enhancing localization quality assessment.

Conclusion

This paper presents a Pixel-level Quality Assessment (PQA) framework for oriented object detection. PQA replaces conventional box-level IoU prediction with the integration of pixel-level spatial consistency. By operating at the pixel level, PQA effectively eliminates the structural coupling issue and ensures a robust positive correlation between the estimated and actual localization quality. In addition, a novel integration metric is proposed to convert pixel-level spatial consistency into a unified quality score, offering a closer approximation to the GT IoU. Experiments on benchmark datasets demonstrate that PQA can be seamlessly integrated into various oriented object detectors and consistently improve their performance, highlighting the effectiveness and generalizability of the proposed approach.

Acknowledgements

This work was supported by the Basic Science (Natural Science) Research Project of Higher Education Institutions of Jiangsu Province under Grants No.25KJB520021.

References

- Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; and Han, J. 2022. Anchor-free oriented proposal generator for object detection. *TGRS*, 60: 1–11.
- Dai, L.; Liu, H.; Tang, H.; Wu, Z.; and Song, P. 2023. AO2-DETR: Arbitrary-oriented object detection transformer. *TCSVT*, 33(5): 2342–2356.
- Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning RoI transformer for oriented object detection in aerial images. In *CVPR*, 2849–2858.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.
- Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; and Sun, J. 2021. Ota: Optimal transport assignment for object detection. In *CVPR*, 303–312.
- Girshick, R. 2015. Fast R-CNN. In *ICCV*, 1440–1448.
- Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2021. Align deep features for oriented object detection. *TGRS*, 60: 1–11.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.
- Hou, L.; Lu, K.; Xue, J.; and Li, Y. 2022. Shape-adaptive selection and measurement for oriented object detection. In *AAAI*, volume 36, 923–932.
- Huang, Z.; Li, W.; Xia, X.-G.; and Tao, R. 2022. A general gaussian heatmap label assignment for arbitrary-oriented object detection. *TIP*, 31: 1895–1910.
- Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *ECCV*, 784–799.
- Kahraman, F.; Oksuz, K.; Kalkan, S.; and Akbas, E. 2023. Correlation loss: Enforcing correlation between classification and localization. In *AAAI*, volume 37, 1087–1095.
- Kim, K.; and Lee, H. S. 2020. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, 355–371.
- Lee, C.; Son, J.; Shon, H.; Jeon, Y.; and Kim, J. 2024. FRED: Towards a Full Rotation-Equivariance in Aerial Image Object Detection. In *AAAI*, volume 38, 2883–2891.
- Li, W.; Chen, Y.; Hu, K.; and Zhu, J. 2022. Oriented keypoints for aerial object detection. In *CVPR*, 1829–1838.
- Li, X.; Wang, W.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2021. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *CVPR*, 11632–11641.
- Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. volume 33, 21002–21012.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.
- Liu, Z.; Yuan, L.; Weng, L.; and Yang, Y. 2017. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *ICPRAM*, volume 2, 324–331.
- Ming, Q.; Miao, L.; Zhou, Z.; Song, J.; Dong, Y.; and Yang, X. 2023. Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. *ISPRS*, 196: 241–255.
- Murrugarra-Llerena, J.; Kirsten, L. N.; Zeni, L. F.; and Jung, C. R. 2024. Probabilistic Intersection-Over-Union for Training and Evaluation of Oriented Object Detectors. *TIP*, 33: 671–681.
- Oksuz, K.; Cam, B. C.; Akbas, E.; and Kalkan, S. 2021. Rank & sort loss for object detection and instance segmentation. In *ICCV*, 3009–3018.
- Pu, Y.; Liang, W.; Hao, Y.; Yuan, Y.; Yang, Y.; Zhang, C.; Hu, H.; and Huang, G. 2023a. Rank-DETR for high quality object detection. In *NeurIPS*, volume 36, 16100–16113.
- Pu, Y.; Wang, Y.; Xia, Z.; Han, Y.; Wang, Y.; Gan, W.; Wang, Z.; Song, S.; and Huang, G. 2023b. Adaptive rotated convolution for rotated object detection. In *ICCV*, 6589–6600.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 91–99.
- Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 9627–9636.
- Xi, Y.; Lu, T.; Kang, X.; and Li, S. 2024. Structure-adaptive Oriented Object Detection Network for Remote Sensing Images. *TGRS*, 62: 1–13.
- 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 *CVPR*, 3974–3983.
- Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for object detection. In *ICCV*, 3520–3529.
- Xu, C.; Ding, J.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; and Xia, G.-S. 2023. Dynamic coarse-to-fine learning for oriented tiny object detection. In *CVPR*, 7318–7328.
- Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; and Bai, X. 2021. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *TPAMI*, 43(4): 1452–1459.
- Yang, X.; Yan, J.; Feng, Z.; and He, T. 2021a. R3det: Refined single-stage detector with feature refinement for rotating object. In *AAAI*, volume 35, 3163–3171.
- Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; and Yan, J. 2021b. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *NeurIPS*, 34: 18381–18394.

Yang, X.; Zhang, G.; Li, W.; Zhou, Y.; Wang, X.; and Yan, J. 2023a. H2RBox: Horizontal Box Annotation is All You Need for Oriented Object Detection. In *ICLR*.

Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; and Tian, Q. 2023b. The KFIoU Loss for Rotated Object Detection. In *ICLR*.

Yu, H.; Tian, Y.; Ye, Q.; and Liu, Y. 2024. Spatial transform decoupling for oriented object detection. In *AAAI*, volume 38, 6782–6790.

Yu, Y.; and Da, F. 2023. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In *CVPR*, 13354–13363.

Yu, Y.; and Da, F. 2024. On boundary discontinuity in angle regression based arbitrary oriented object detection. *TPAMI*, 46.

Zeng, Y.; Chen, Y.; Yang, X.; Li, Q.; and Yan, J. 2024. ARS-DETR: Aspect Ratio-Sensitive Detection Transformer for Aerial Oriented Object Detection. *TGRS*, 62: 1–15.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021. VarifocalNet: An iou-aware dense object detector. In *CVPR*, 8514–8523.

Zhen, P.; Wang, S.; Zhang, S.; Yan, X.; Wang, W.; Ji, Z.; and Chen, H.-B. 2023. Towards accurate oriented object detection in aerial images with adaptive multi-level feature fusion. *ACM TOMM*, 19(1): 1–22.

Zhou, D.; Fang, J.; Song, X.; Guan, C.; Yin, J.; Dai, Y.; and Yang, R. 2019. Iou loss for 2d/3d object detection. In *3DV*, 85–94.

Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. 2022. Mmrotate: A rotated object detection benchmark using pytorch. In *ACM MM*, 7331–7334.

Zhu, Y.; and Huang, B. 2025. Pixel-level Quality Assessment for Oriented Object Detection. arXiv:2511.08186.