GEOMETRY MEETS SEMANTICS: FRACTIONAL GRADIENT STABILIZATION FOR SEMANTIC-DRIVEN BOUNDING BOX OPTIMIZATION IN VISUAL DETECTION TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bounding boxes are fundamental for object localization in visual detection tasks. Among them, oriented bounding boxes are widely used in visual detection tasks, which provide a more precise directional representation. Generally, IoU-based losses are widely adopted to optimize box regression. However, we observed that IoU-driven box optimization suffers from two key issues: (1) it relies solely on geometric properties while ignoring semantic cues; (2) orientation optimization suffers from unstable gradients, causing oscillations in orientation convergence. In this paper, we propose a Fractional Semantic IoU loss to achieve unified semantic-geometric learning with gradient stabilization. First, we design a semantic similarity metric to guide IoU optimization, building a Semantic IoU loss (SIoU loss) with an adaptive gradient gating mechanism. Then, we revisit the gradient instability issue in oriented box optimization and extend the SIoU loss to a fractional-order formulation to build the **Fr**actional **S**emantic **IoU loss** (FrSIoU loss). The FrSIoU loss accumulates historical IoU states to regularize abnormal gradients during bounding box optimization process. Extensive experiments demonstrate that our approach achieves stable performance gains across different bounding box formulations and diverse visual detection tasks. The code will be available on GitHub.

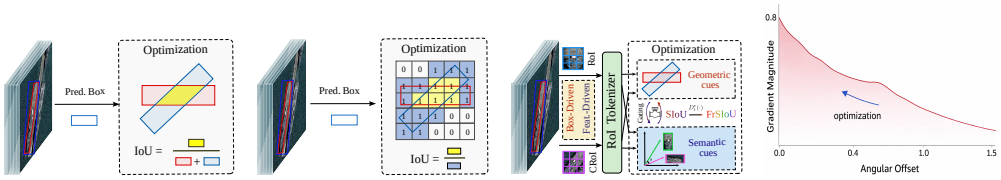
1 INTRODUCTION

Visual detection is fundamental to computer vision and supports applications such as autonomous driving (Zheng et al., 2020a; Ming et al., 2023), remote sensing analysis (Yang et al., 2021c; Ming et al., 2024; Yang et al., 2021b), and robotic interaction (Cheng et al., 2021). It also supports higher-level tasks, including tracking (Yu et al., 2025), segmentation (Ming & Xiao, 2024), and scene understanding (Zuo et al., 2025).

Bounding boxes are the most common geometric representation to localize object instances in detection frameworks. Depending on task requirements, bounding boxes can be categorized into Horizontal Bounding Boxes (HBBs) (Ren et al., 2016; Liu et al., 2016), Oriented Bounding Boxes (OBBs) (Han et al., 2021; Ding et al., 2019), polygon-based representations (Yang et al., 2022b), etc. To achieve accurate bounding box optimization, modern detectors typically rely on regression losses such as L_1 loss, Smooth-L_1 loss (Ren et al., 2016; Liu et al., 2016), or Intersection-over-Union (IoU) based loss functions (Zhou et al., 2019; Rezatofighi et al., 2019; Zheng et al., 2020b; Ming et al., 2024). Among them, IoU-based losses have become a popular metric due to the scale-invariance of the metric and its direct reflection of localization quality. On this basis, a series of IoU-derived losses (e.g., GIoU (Rezatofighi et al., 2019), DIoU (Zheng et al., 2020b), EIoU (Peng & Yu, 2021), GCIoU (Ming et al., 2024)) have been proposed, which improve regression stability and convergence by incorporating geometric factors such as distance, aspect ratio, or overlap consistency into the optimization process.

In recent years, OBB-based visual detection has been widely applied in various scenarios (Zheng et al., 2020a; Ming et al., 2023; Yang et al., 2021c; Ming et al., 2024; Yang et al., 2021b);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Region-based IoU. (b) Pixel-based IoU. (c) Our FrSIoU. (d) Loss gradient curve.

Figure 1: Calculation of different IoU variants (a-c) and visualization of abnormal gradients in IoU-based losses (d). Existing (a) region-based and (b) pixel-based IoUs consider only geometric constraints, while our FrSIoU (c) further incorporates semantic cues into box optimization.

Cheng et al. (2021). As a generalized form of HBBs, OBBs explicitly encode object orientation, which is critical for high-precision detection and recognizing objects with large aspect ratios. For example, in remote sensing imagery, OBBs are commonly used to represent arbitrarily oriented objects such as bridges, ships (Xia et al., 2018; Li et al., 2020); in UAV aerial imagery, OBBs help distinguish densely packed vehicles (Li & Wu, 2023; Chen et al., 2024); and in 3D vision, voxel bounding boxes can be regarded as 2D OBBs extended with height information (Zhou et al., 2019; Sheng et al., 2022). To achieve efficient OBB prediction and optimization, IoU losses have been extended to skew IoU formulations, achieving notable success in both 2D and 3D detection tasks (Ming et al., 2021; Zhou et al., 2019; Sheng et al., 2022; Ming et al., 2023; Zheng et al., 2020a).

However, existing methods remain constrained to the geometric properties of IoU when designing loss functions, which would leave their inherent limitations unresolved. We suggest that the IoU loss design should address two critical issues: (1) **Current IoU-based losses primarily consider geometric constraints while ignoring the semantic content within bounding boxes**. As illustrated in Fig. 1a and Fig. 1b, both region-based methods (e.g., GIoU (Rezatofighi et al., 2019), DIoU (Zheng et al., 2020b)) and pixel-based methods (e.g., PIoU (Chen et al., 2020), PPIoU (Endo et al., 2025)) supervise box optimization purely from geometry with some constraints. Intuitively, features within the box provide explicit semantic cues that could enhance localization accuracy. However, existing methods have not fully exploited this semantic information. (2) **Skew IoU losses suffer from abnormal angular gradient oscillation during box optimization.** Ideally, with the model converging, gradients should diminish to allow fine-grained refinement. However, as shown in Fig. 1d, the optimization process exhibits unexpected gradient surges, which have also been discussed in (Ming et al., 2024). These gradients cause unstable regression and reduce localization accuracy.

Based on the above analysis, we propose the fractional semantic IoU loss to achieve unified semantic-geometric learning and to mitigate abnormal gradient variations for precise OBB optimization in challenging visual detection tasks. First, as illustrated in Fig. 1c, we introduce a semantic similarity metric that extends IoU loss into Semantic IoU loss (SIoU loss). SIoU loss jointly considers both geometric properties and semantic content of bounding boxes, while a gating mechanism dynamically balances the contributions of semantic and geometric similarities in bounding box optimization. Without the carefully designed geometric penalty terms (such as those in GIoU (Rezatofighi et al., 2019), DIoU (Zheng et al., 2020b), WIoU (Tong et al., 2023), etc.), SIoU loss naturally addresses common issues in naive IoU optimization, such as the non-overlapping regression problem (Rezatofighi et al., 2019), achieving more efficient bounding box optimization. To further address gradient instability, we extend SIoU loss into a fractional formulation and develop the Fractional Semantic IoU loss (FrSIoU loss), which stabilizes angular gradients during model convergence and ensures accurate OBB regression. FrSIoU loss uses the historical semantic-geometric similarity to generate modulation weights dynamically, and regularize unstable angular gradients during the optimization. Our main contributions are summarized as follows:

- We revisit the limitation of current IoU losses and conclude that existing methods rely solely on geometric constraints. Then, a Semantic IoU loss (SIoU loss) is proposed for unified semantic-geometric learning. SIoU loss incorporates both geometric and semantic consistency via a dynamic gating mechanism to adaptively guide bounding box regression.
- To address gradient instability in OBB optimization, we further extend SIoU loss into a Fractional Semantic IoU loss (FrSIoU loss). FrSIoU loss stabilizes model convergence via

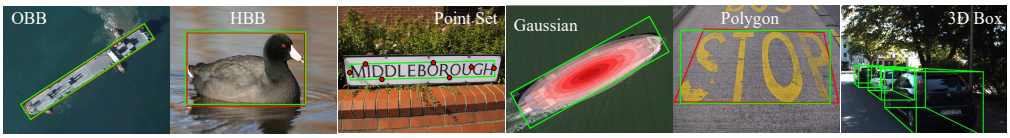


Figure 2: Illustration of different types of bounding box representations. OBB shows strong generality and can be widely applied across various visual detection tasks.

introducing a fractional derivative to capture the historical sIoU, which dynamically adjusts the loss weights to regularize unstable angular gradients.

- We conduct extensive evaluations across multiple detectors and vision detection tasks. Experimental results demonstrate that FrSIoU loss exhibits strong generalization ability, consistently achieving performance gains and offering new insights for future research.

2 Related Work

OBB Detection in Visual Detection Tasks. Oriented bounding boxes (OBBs) generalize horizontal bounding boxes (HBBs) by explicitly modeling object orientation, enabling high-precision detection of arbitrary-oriented objects. They have been widely applied across diverse vision tasks (Cheng et al., 2021; Xia et al., 2018; Chen et al., 2024; Ming et al., 2023; Shi et al., 2018; Endo et al., 2025; Zhu et al., 2025). For example, OBBs are used to represent grasp poses in robotic manipulation (Cheng et al., 2021), detect oriented objects in remote sensing benchmarks (Xia et al., 2018), and distinguish densely arranged vehicles in UAV imagery (Chen et al., 2024). In autonomous driving, 3D bounding boxes can be viewed as OBBs with an additional height dimension (Ming et al., 2023; Sheng et al., 2022). They are also adopted in face detection, scene text detection, and precision agriculture (Shi et al., 2018; Endo et al., 2025; Liao et al., 2018; Zhu et al., 2025). The key advantage of OBBs lies in their generality, as they unify both axis-aligned and oriented representations, with HBB being a special case. Other representations, such as point sets (Li et al., 2022; Song et al., 2023), Gaussian models (Yang et al., 2021c;b), polygons (Yang et al., 2022b; Endo et al., 2025), and vector encodings (Yi et al., 2021), can often be transformed into or approximated by OBBs. This highlights OBBs as a unified, flexible, and fundamental representation for visual detection tasks.

IoU-based Losses for Bounding Box Optimization. Intersection-over-Union (IoU) is the dominant metric for object detection due to its scale-invariance and intuitive geometric interpretation, and IoU-based losses align training objectives with evaluation (Zhou et al., 2019). However, they suffer from slow convergence and zero gradients when predicted and ground-truth boxes do not overlap. To mitigate these issues, many variants introduce auxiliary geometric constraints: GIoU (Rezatofighi et al., 2019) adds an enclosing-box penalty to provide gradients in non-overlapping cases, DIoU and CIoU (Zheng et al., 2020b) incorporate center distance and aspect ratio consistency to accelerate convergence, and WIoU (Tong et al., 2023) adaptively reweights samples to improve robustness against low-quality predictions. Beyond geometric extensions, alternative formulations approximate IoU by modeling boxes as distributions, such as Gaussian Wasserstein Distance (GWD) (Yang et al., 2021b), Kullback-Leibler Divergence (KLD) (Yang et al., 2021c), and Kalman filter-based KFIoU (Yang et al., 2022a). While effective, these approaches still rely on geometric or surrogate supervision, which may partially break the consistency between training loss and evaluation metric. More recently, works have investigated the fundamental gradient behavior of IoU-based losses, revealing vanishing and oscillatory gradients near convergence for HBB (Peng & Yu, 2021), and gradient anomalies with respect to angle and scale in OBB regression (Ming et al., 2024), highlighting the need for more stable and well-behaved optimization strategies.

3 Methodology

3.1 Problem Formulation

Unified OBB Representations Most visual detection tasks represent object instances by encoding object position, scale, and orientation. The 2D OBB is commonly defined as $\mathcal{B}_{\text{OBB}} = (x, y, w, h, \theta)$,

where $(x, y) \in \mathbb{R}^2$ denotes the box center, w, h denote the box extents along the local axes, and $\theta \in \mathbb{R}$ is the rotation angle. We usually normalize the angle to a canonical interval, e.g, $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$ or $\theta \in [0, \pi)$ depending on the application (Han et al., 2021; Ding et al., 2019). Next, we adopt a unified view of these representations and their conversions for analyzing optimization behaviour. (1) *HBB ↔ OBB.* The axis-aligned HBB is widely used in general object detection tasks (Lin et al., 2014; Everingham et al., 2010; Ren et al., 2016; Redmon et al., 2016). HBB is a special case of the OBB without orientation information: $\mathcal{B}_{\mathrm{HBB}} = (x, y, w, h)$ and the corresponding OBB format is $\mathcal{B}_{\mathrm{OBB}} = (x, y, w, h, 0)$. (2) *Point-set ↔ OBB.* A point-set representation describes an object by a finite set of 2D points $\mathcal{P} = \{p_i = (x_i, y_i)\}_{i=1}^{N}$ (Li et al., 2022; Song et al., 2023). Its equivalent OBB can be obtained by computing the minimum-area enclosing rectangle of $\mathcal{P}$ (or of its convex hull). For complex polygons, PCA is the commonly used approximation to obtain the orthogonal direction. (3) *Gaussian representation ↔ OBB.* 2D OBB can be viewed as a Gaussian distribution by mean and covariance, $\mathcal{G} = (\mu, \Sigma)$. More detailed transformation process could be found in (Yang et al., 2021c;b). (4) *Polygon ↔ OBB.* Given a polygon with ordered vertices $\{v_i\}$, the minimum-area bounding OBB can be computed exactly via the rotating-calipers algorithm (Shamos, 1978) applied to the convex hull. We denote this operator by $\mathcal{B}_{\mathrm{OBB}} = \mathrm{MinAreaRect}(\mathrm{ConvexHull}(\{v_i\}))$. (5) *3D bounding box ↔ OBB.* A 3D box $\mathcal{B}_{\mathrm{3D}} = (x, y, z, l, w, h, \phi)$ can be regarded as a 2D OBB with height, since it only has one rotational degree of freedom (yaw). Thus, many 2D OBB regression methods extend directly to 3D detection (Ming et al., 2023; Sheng et al., 2022). As shown in Fig. 2, most representations can be mapped to a unified OBB parametrization, enabling a common optimization perspective across diverse detection formulations.

Anomalous Gradient Analysis for OBB Optimization To ensure efficient and stable OBB optimization, the skew IoU loss is widely employed in visual detection tasks (Zhou et al., 2019; Ming et al., 2024; 2023). Given the predicted box $\mathbf{b}_1 = (x_1, y_1, w_1, h_1, \theta_1)$ and the corresponding target box $\mathbf{b}_2 = (x_2, y_2, w_2, h_2, \theta_2)$, the angular deviation is $\theta = |\theta_1 - \theta_2|$. The classic linear IoU loss is

$$L_{\mathrm{IoU}}(\mathbf{b}_1, \mathbf{b}_2) \;=\; 1 - \mathrm{IoU}(\mathbf{b}_1, \mathbf{b}_2). \tag{1}$$

Previous studies (Ming et al., 2024) have shown that center point convergence in OBB regression is fast and stable, and therefore, the skew IoU computation can be simplified to center-aligned OBBs. In this case, we obtain the loss gradient with respect to θ as follows:

$$\frac{\partial L_{\mathrm{IoU}}}{\partial \theta} = \mathrm{IoU}(\theta)\left(1 + \mathrm{IoU}(\theta)\right)\,\cot\theta\,. \tag{2}$$

The detailed proof can be found in the Appendix. Eq.(2) reveals two mechanisms that produce unstable angular gradients in IoU-driven optimization:

1. *Singularity as* $\theta \to 0$: As the model converges, $\theta \downarrow 0$ and therefore $\cot\theta \to \infty$. The angular derivative grows as the angular error decreases, which produces large gradient magnitudes and leads to loss oscillation in the late stage of optimization. Due to boundary conditions in the skew IoU computation, θ never reaches 0 (see Appendix for proof). Thus, the IoU loss remains convergent. However, its gradient variation during convergence still impedes bounding box refinement.
2. *Dependence on overlap magnitude*: The linear IoU loss gradient contains the multiplicative factor $\mathrm{IoU}(1 + \mathrm{IoU})$, which can increase with IoU and therefore amplify the singular behaviour.

These observations motivate two design goals for a robust localization loss: (1) incorporate semantic cues inside the region to complement purely geometric supervision, and (2) introduce mechanisms to stabilize angular gradients during optimization. In the following sections, we present our semantics-driven bounding box optimization and the fractional IoU loss to achieve the goals.

3.2 Semantics-Driven Bounding Box Optimization

Most IoU-based losses focus exclusively on geometric constraints and neglect the rich semantic cues embedded in deep features. In this section, we will introduce a semantics-aware coupling that complements geometric cues with feature similarity, enabling continuous and adaptive guidance throughout the bounding box optimization process.

Semantic score construction. Let $\mathbf{b} = (x, y, w, h, \theta)$ and $\mathbf{b}^* = (x^*, y^*, w^*, h^*, \theta^*)$ denote a predicted OBB and a ground-truth OBB, and $\mathcal{F}(\cdot)$ be the feature map extracted via the backbone network. We extract two semantic cues:

- *Intra-region feature distance.* Sample N paired locations inside the predicted and GT boxes (after appropriate warping such as RoI Align (He et al., 2017)), yielding feature vectors $\{f_p^{(i)}\}_{i=1}^N$ and $\{f_g^{(i)}\}_{i=1}^N$. Define the mean squared feature distance:

$$d_{\text{intra}} = \frac{1}{N} \sum_{i=1}^{N} \left\| f_p^{(i)} - f_g^{(i)} \right\|_2^2. \tag{3}$$

  Convert distance to a similarity $s_1 \in (0, 1]$ via:

$$s_1 = \exp\big(- \alpha \, d_{\text{intra}} \big), \tag{4}$$

  where $\alpha > 0$ controls sensitivity.

- *Context-region similarity.* Build horizontally-aligned context boxes $\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}^*$ (e.g., axis-aligned expansions of $\mathbf{b}$ and $\mathbf{b}^*$) and extract their corresponding feature vectors $\{\hat{f}_p^{(i)}\}_{i=1}^M$ and $\{\hat{f}_g^{(i)}\}_{i=1}^M$ and compute cosine similarity to obtain a contextual alignment score s_2:

$$s_2 = \frac{1}{2} \left(\frac{1}{M} \sum_{i=1}^{M} \frac{\hat{f}_p^{(i)} \cdot \hat{f}_g^{(i)}}{\|\hat{f}_p^{(i)}\|_2 \|\hat{f}_g^{(i)}\|_2} + 1 \right). \tag{5}$$

Finally, the overall semantic score s is defined as:

$$s = \beta s_1 + (1 - \beta) s_2, \tag{6}$$

where β is generally set to 0.5. It controls the balance between fine-grained object semantics and contextual semantics, so that the final semantic score $s \in [0, 1]$.

In Eq.(6), s_1 reflects the fine-grained semantic consistency between predicted and GT regions, providing a precise, discriminative metric that captures localized appearance similarity. Intuitively, the supervision from s_1 implicitly encodes geometric constraints between predicted boxes and GT objects. For example, if $s_1 \to 1$, the corresponding geometric IoU would naturally reach 1.

However, the precisely rotated RoI features would remove orientation cues, making it difficult for the model to converge with only s_1 supervision (see Appendix for detailed analysis). To address this issue, we introduce cosine similarity to compute s_2. Since contextual semantic information is more flexible, the correlations of context features help to optimize predicted RoI even without perfectly matching the GT region. Therefore, cosine similarity serves as a directional constraint to guide the bounding box optimization.

Semantic IoU loss. Denote the geometric IoU by $u = \text{IoU}(\mathbf{b}, \mathbf{b}^*) \in [0, 1]$. We adopt the coupling

$$\text{sIoU} \;=\; s + (1 - s)\, u, \tag{7}$$

which preserves the property that a perfect geometric overlap ($u = 1$) yields sIoU $= 1$. Then, the complementary residual serves as

$$L_{\text{SIoU}} \;=\; (1 - s)(1 - u). \tag{8}$$

L_{SIoU} is the semantics-driven IoU loss that measures the semantic-geometric misalignment. It extends the geometric-based IoU loss by introducing a feature-driven gating mechanism. When $u = 0$, the loss degenerates to $\mathcal{L}_{sIoU} = 1 - s$, providing non-zero gradients through the semantic term and thus avoiding the vanishing-gradient problem. Therefore, $\mathcal{L}_{sIoU}$ enables feature-level supervision even under complete geometric mismatch. Furthermore, by modulating the geometric term with the semantic score, $\mathcal{L}_{sIoU}$ adaptively emphasizes semantic or geometric alignment depending on the optimization phase.

Analysis of Adaptive Gradient Gating. We next conduct the analysis of the gradient of the SIoU loss. Taking the bounding box parameter θ as an example, we obtain:

$$\frac{\partial L_{\text{SIoU}}}{\partial \theta} \;=\; -(1 - s)\, \frac{\partial u}{\partial \theta} \;-\; (1 - u)\, \frac{\partial s}{\partial \theta}. \tag{9}$$

Eq.(9) exposes the semantic-geometric gating mechanism in SIoU loss during training. When s is large, semantic consistency becomes more confident, and semantic cues are expected to dominate the bounding box optimization. For example, as $s \to 1$, $(1-s) \to 0$, and the overall gradient is dominated by the semantic term $-\partial s/\partial\theta$. Conversely, when s is relatively small while u is large, the semantic score is unreliable, and the model is expected to rely more on geometric cues for refinement. The semantic gradient $-\partial s/\partial\theta$ is scaled by $(1-u)$; as $u \to 1$ and $s \to 1$, the semantic influence is suppressed, thereby avoiding potential interference with geometric alignment.

SIoU loss implements a soft arbitration between semantic and geometric cues. Unreliable semantics increase reliance on geometric information, while confident semantics suppress noisy geometric gradients. Through this adaptive gradient gating mechanism, the SIoU loss dynamically balances semantic and geometric supervision, enabling the network to select the dominant cue based on gradient flow for better bounding box regression.

3.3 Fractional IoU Loss for Gradient Stabilization

Though SIoU loss provides a flexible and adaptive bounding box optimization scheme, it still does not fundamentally resolve the issue of abnormal angular gradient variations. In this section, we will extend the SIoU loss into a fractional-order formulation to further stabilize abnormal gradients.

Fractional IoU Loss. Firstly, a fractional semantic IoU (FrSIoU) loss is proposed to incorporate a memory mechanism of the optimization process. It not only concerns the current loss, but also dynamically adjusts the loss function based on the historical semantic-geometric similarity. Then, it can effectively regularize unstable gradients.

The fractional derivative extends differential operations to non-integer orders and characterizes the nonlocal effects of network training (Khalil et al., 2014). The Grünwald-Letnikov (G-L) approximation is used to compute the v-th derivative at the current epoch (Scherer et al., 2011). Given M historical steps, the G-L derivative is defined as

$$D_t^v(sIoU) = \sum_{k=0}^{M} c_k \cdot sIoU_{t-k}, \tag{10}$$

where c_k is calculated as

$$c_0 = 1, \quad \frac{k-1-v}{k} \cdot c_k = c_{k-1} \quad \text{for } k \geq 1, \tag{11}$$

Taking $M = 2$ as an example, the approximate expression is

$$\begin{aligned} D_t^v(sIoU) \quad &\approx c_0 sIoU_t + c_1 sIoU_{t-1} + c_2 sIoU_{t-2} \\ &= sIoU_t - vsIoU_{t-1} + \tfrac{v(v-1)}{2} sIoU_{t-2}, \end{aligned} \tag{12}$$

The proposed derivative $D_t^v(sIoU)$ represents the trend of semantic-geometric similarity, which further leads to the definition of the modulation weight W_t as

$$W_t = \exp(-\lambda \cdot D_t^v(sIoU)), \tag{13}$$

where λ is a hyperparameter controlling the modulation intensity. Finally, the fractional semantic IoU (FrSIoU) loss is defined as

$$\mathcal{L}_{FrSIoU} = W_t \cdot (1-s)(1-u). \tag{14}$$

Gradient Stabilization of Fractional IoU. The modulated weight W_t uses the fractional derivative of the historical sIoU trajectory to control the optimization step. When the metric improves steadily, $W_t < 1$ suppresses overly large updates near angular boundaries; when the metric drops sharply, $W_t > 1$ strengthens the corrective signal. Thus, the FrSIoU gradient is a weighted sum of historical geometric gradients, which reduces abrupt variation induced by the $\cot\theta$ term. The full derivation is provided in Appendix A.

Table 1: Ablation study of different loss functions on the challenging large-scale DOTA-v1.0 dataset based on Rotated RetinaNet.

Loss terms	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{50}(%)
L_1 loss	87.2	75.2	34.8	63.5	66.4	75.4	85.0	90.3	61.5	60.5	52.4	62.2	60.0	56.8	50.4	65.44
GIoU loss	89.1	77.1	37.2	67.8	67.9	74.3	85.1	90.5	64.3	60.4	48.7	63.1	61.7	56.5	56.7	66.67
SIoU loss	89.2	77.3	37.4	67.0	67.8	74.5	85.4	90.5	61.9	64.3	49.8	65.1	62.0	55.4	54.6	66.82
FrSIoU loss	89.2	76.9	37.6	67.6	67.9	74.6	85.5	90.5	63.5	64.1	48.8	65.2	62.0	56.6	57.6	**67.27**

Table 2: Analysis of hyperparameters in FrSIoU loss.

λ	α	mAP_{50} (%)	λ	α	mAP_{50} (%)	λ	α	mAP_{50} (%)
	0.05	67.21		0.05	67.15		0.05	66.97
0.05	0.1	66.80	0.1	0.1	67.06	0.3	0.1	67.17
	0.3	67.04		0.3	67.09		0.3	**67.27**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We conduct experiments on multiple datasets to verify the generalization and robustness of our approach, including DOTA (Xia et al., 2018), HRSC2016 (Liu et al., 2017) and SSDD (Zhang et al., 2021). DOTA-v1.0 is utilized as the primary benchmark to assess detection performance across diverse categories and complex scenes. As a large-scale aerial dataset, DOTA-v1.0 contains 2,800 images and 188,000 instances across 15 categories, characterized by extreme scale variations and arbitrary orientations. For fine-grained ship detection, HRSC2016 contains 1,061 images annotated with rotated boxes, specificy highlighting objects with large aspect ratios. In the SAR domain, we utilize the SSDD dataset to demonstrate the generalization capability and robustness of our approach across different imaging modalities and environmental conditions. The SSDD dataset provides 1,160 images for ship detection, focusing on detecting small objects within complex coastal environments.

Evaluation Metrics. Following common practice, we use three standard metrics to evaluate detection performance: mAP_{50}, mAP_{75}, and mAP. mAP_{50} and mAP_{75} represent the Average Precision at IoU thresholds of 0.5 and 0.75, respectively, while mAP averages the precision across IoU thresholds from 0.5 to 0.95 with a step of 0.05. Distinct from prior studies that primarily report mAP_{50}, we explicitly include mAP_{75} and mAP to demonstrate the model's robustness in high-precision localization. This stricter evaluation protocol ensures a more rigorous comparison of bounding box regression accuracy.

Implementation Details. For fair comparison, all experiments are implemented based on MMRotate framework (Zhou et al., 2022) using four NVIDIA GeForce RTX 3090 GPUs. We use the SGD optimizer with an initial learning rate of 0.02, momentum of 0.9, and weight decay of 0.0001. The batch size is set to 16 with a 500-iteration linear warm-up. Regarding training schedules, models on the DOTA dataset are trained for 24 epochs, while those on HRSC2016 and SSDD datasets are trained for 72 epochs. Specifically for the DOTA dataset, we adopt single-scale training and testing. Ablation studies are conducted on the validation set. For the main results, models are trained on the combined training and validation sets, and the final performance is obtained by submitting testing results to the official evaluation server.

Controlled Comparison Protocol. Unless otherwise stated, all competing losses use the same detector architecture, training schedule, and evaluation pipeline. We modify only the box regression loss, so the observed performance differences can be attributed to the proposed FrSIoU formulation. For each benchmark, we follow its official evaluation protocol and report the metric appropriate to the corresponding task. This controlled setup ensures that the ablation and cross-dataset comparisons measure localization quality under consistent training and testing conditions.

Table 3: Comparison with state-of-the-arts on the DOTA-v1.0 dataset. The input image size is 1024×1024. The **best** and second-best results are highlighted in **bold** and underlined.

	Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{50}	mAP_{75}	mAP
One Stage	R-RetinaNet (Lin et al., 2017)	**89.65**	82.57	38.16	69.84	77.41	62.76	77.25	90.69	83.80	82.04	59.93	64.83	57.38	64.76	45.56	69.77	37.70	39.65
	CSL (Yang & Yan, 2020)	89.34	79.67	40.99	69.96	77.71	62.08	77.51	90.87	82.87	82.04	60.03	65.27	53.60	64.08	46.61	69.51	40.41	39.69
	R^3Det (Yang et al., 2021a)	89.29	75.21	45.42	69.24	75.54	72.90	79.29	90.89	81.02	83.25	58.82	63.17	63.40	62.22	37.41	69.80	36.58	37.82
	GWD (Yang et al., 2021b)	88.93	77.04	45.85	69.30	72.53	64.05	76.41	90.87	79.19	80.46	57.68	64.37	63.60	64.75	48.25	69.55	38.91	39.50
	KLD (Yang et al., 2021c)	89.21	80.68	48.79	70.56	77.72	75.77	86.34	**90.90**	82.83	83.78	65.70	65.92	66.08	66.42	40.70	72.76	37.83	39.92
	S^2A-Net (Han et al., 2021)	89.30	80.51	50.38	73.23	78.41	77.40	86.80	90.89	85.64	84.23	62.13	65.95	66.61	67.74	53.51	74.18	36.83	39.84
Two Stage	RoI-Trans. (Ding et al., 2019)	88.97	82.12	54.57	**76.27**	79.29	77.96	87.94	**90.90**	**87.19**	85.66	62.18	62.65	74.63	72.43	59.23	76.13	46.88	45.56
	KFIoU (Yang et al., 2022a)	89.05	75.17	49.04	69.67	78.06	75.45	86.69	**90.90**	83.65	84.48	62.21	62.85	66.72	65.95	50.20	72.67	36.01	38.88
	ReDet (Han et al., 2021)	89.20	83.78	52.19	71.05	78.05	82.50	88.24	90.86	87.26	85.97	**65.50**	62.86	75.89	70.05	**66.71**	76.67	48.84	46.71
	O-Former (Zhao et al., 2024)	88.14	79.13	51.96	67.34	**81.02**	**83.26**	**88.29**	**90.90**	85.57	86.25	60.84	66.36	73.81	71.23	56.49	75.37	46.39	45.01
	ACM (Xu et al., 2024)	88.06	73.99	49.08	61.54	79.73	77.23	87.07	90.89	82.17	84.04	56.97	61.21	63.91	67.78	36.16	70.66	42.64	41.26
	Point2RBoxv2 (Yu et al., 2025)	88.21	69.76	23.30	44.22	79.82	76.18	87.09	89.27	44.97	83.26	16.19	41.00	44.95	58.00	49.54	59.72	25.96	30.13
	GSDet (Ding et al., 2025)	88.50	79.02	50.58	71.71	78.53	83.21	88.25	90.84	84.26	80.33	55.87	57.04	74.70	67.87	57.46	73.88	**52.17**	47.69
	GauCho (Marques et al. 2025)	88.56	76.69	49.87	61.59	80.00	79.59	87.46	90.88	83.24	84.19	54.98	64.86	64.12	69.53	51.33	72.46	37.03	40.46
	ReDiffDet (Zhao et al., 2025)	85.45	77.84	50.39	71.63	80.90	84.85	**88.77**	90.88	87.00	86.08	63.30	63.43	76.01	76.03	60.16	76.18	–	–
	FAA (Gu et al., 2026)	89.40	84.15	54.66	72.23	78.42	82.89	87.86	90.89	86.26	85.24	63.95	65.58	75.30	70.44	61.06	76.55	–	–
	O-RCNN (Xie et al., 2021)	89.35	81.41	52.71	75.02	79.03	82.43	87.82	90.89	86.40	85.30	63.37	65.69	68.28	70.48	57.22	75.69	47.18	45.19
	FrSIoUloss (Ours)	89.16	**84.66**	**57.34**	71.42	80.11	82.62	87.98	90.84	80.56	**86.66**	64.75	**70.68**	**77.76**	**73.47**	61.30	**77.28**	50.50	**47.94**

All results use single-scale training and testing; O-RCNN serves as the baseline for our method.

4.2 Ablation Study and Analysis

Effect of Semantic IoU Loss. We first evaluate the impact of introducing semantic guidance into box regression. The baseline model trained with standard L_1 loss achieves 65.44% mAP. While the geometric-based GIoU loss improves this to 66.77% by addressing the non-overlapping issue, it still lacks semantic awareness. By replacing GIoU with our proposed SIoU loss, the performance further improves to 66.82% mAP. Notably, SIoU loss demonstrates superior efficacy in rotation-sensitive categories. For instance, the accuracy for roundabout (RA) increases by 2.0% compared to GIoU from 63.1% to 65.1%. This improvement indicates that the semantic alignment term successfully constrains the optimization direction, enabling more robust regression for objects with complex appearances where purely geometric constraints may fall short.

Effect of Fractional SIoU Loss. Although SIoU loss enhances bounding box optimization flexibility, it does not fundamentally address the problem of angle gradient discontinuity. To further stabilize training, we introduce a Fractional SIoU loss that adaptively decomposes the angle term into periodic components, ensuring smooth gradient propagation near angle boundaries. As presented in Table 1, FrSIoU achieves the best overall performance of 67.27% mAP, outperforming the SIoU loss by 0.45% and the L_1 baseline by 1.83%. Specific gains in categories such as bridge (BR) and small vehicle (SV) confirm that the fractional modulation effectively stabilizes angular gradients, ensuring smooth convergence and high-precision localization.

Effect of Hyperparameters in FrSIoU loss. We further analyze the impact of the modulation intensity λ and the semantic sensitivity parameter α on detection performance, as reported in Table 2. The results demonstrate that FrSIoU is highly robust to hyperparameter variations, with mAP_{50} fluctuations remaining within a narrow range of 0.5%. This stability implies that the proposed FrSIoU effectively regularizes the gradient without requiring delicate tuning. Based on the peak mAP_{50} of 67.27%, we set $\lambda = 0.3$ and $\alpha = 0.3$ as the standard configuration.

4.3 Results and Comparisons

Comparison with SOTA Methods. We validate the proposed FrSIoU loss within the Oriented R-CNN (Xie et al., 2021) framework, chosen for its efficiency and strong baseline performance. Table 3 compares our method with recent state-of-the-art oriented object detection approaches categorized into one-stage methods and two-stage methods. As presented in Table 3, our approach achieves the best overall performance among the compared methods, reaching an mAP of 47.94% and mAP_{50} of 77.28%. Despite using a simple ResNet-50 backbone, our method outperforms recent competitive models, such as ReDet (46.71% mAP) and GSDet (47.69% mAP), which often rely on more complex feature refinement modules.

Table 4: Comparison of different models on HRSC2016 and SSDD datasets.

Model	HRSC2016			SSDD
	mAP(%)	mAP_{50}(%)	mAP_{75}(%)	mAP_{50}(%)
R-RetinaNet	46.63	81.60	48.80	78.32
R-RetinaNet **+ FrSIoU loss**	52.80 (+ 6.17)	83.00 (+1.40)	59.30 (+10.50)	80.49 (+2.17)
O-RCNN	65.81	89.30	79.70	89.45
O-RCNN **+ FrSIoU loss**	66.26 (+ 0.45)	90.50 (+ 1.20)	81.00 (+1.30)	90.07 (+ 0.62)



Figure 3: Visualization results on DOTA, HRSC2016, and SSDD datasets.

Compared with Oriented R-CNN, FrSIoU improves mAP_{75} by 3.32% and mAP by 2.75%, with gains on bridge (BR, +4.63%) and roundabout (RA, +4.99%). These results confirm stable high-precision localization.

Generalization on Ship Detection Datasets. Table 4 shows consistent gains across HRSC2016 and SSDD, detector architectures, and imaging modalities. On HRSC2016, FrSIoU raises Rotated RetinaNet mAP by 6.17% and mAP_{75} by 10.50%; on SSDD, it improves mAP_{50} for both one-stage and two-stage detectors.

Visualization Results. Fig. 3 presents accurate detections on DOTA, HRSC2016, and SSDD across cluttered scenes and sensing modalities.

5 Conclusion

We address two limitations of IoU-based oriented bounding-box optimization: missing semantic supervision and unstable angular gradients. FrSIoU combines semantic-geometric guidance with a fractional-order memory mechanism to stabilize optimization. Experiments across OBB benchmarks, detector architectures, and sensing modalities show consistent improvements, especially on high-precision localization metrics.

AI Use Statement

We used generative AI tools for manuscript revision, language polishing, and LaTeX formatting assistance. These tools were not used to generate or alter experimental data, numerical results, or scientific conclusions. The authors reviewed and verified all AI-assisted outputs and take full responsibility for the final manuscript, including its text, claims, and reported experimental results.

Reproducibility Statement

The mathematical formulation and optimization procedure are specified in Section 3. Section 4 describes the datasets, evaluation metrics, training settings, implementation details, and controlled comparisons. The ablation and cross-dataset results provide supporting evidence for the main claims.

REFERENCES

Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024.

Zhiming Chen, Kean Chen, Weiyao Lin, John See, Hui Yu, Yan Ke, and Cong Yang. PIoU Loss: Towards accurate oriented object detection in complex environments. In *Proc. Eur. Conf. Comput. Vis.*, 2020.

Hu Cheng, Yingying Wang, and Max Q-H Meng. Grasp pose detection from a single rgb image. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2021.

Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning RoI transformer for oriented object detection in aerial images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019.

Zeyu Ding, Jiaqi Zhao, Yong Zhou, Wen-liang Du, Hancheng Zhu, and Rui Yao. Gsdet: Gaussian splatting for oriented object detection. In *Proc. Int. Jt. Conf. Artif. Intell.*, 2025.

Rei Endo, Taro Miyazaki, Takahiro Mochizuki, and Yoshihiko Kawai. Polygon pixel iou: Similarity metric between polygons with different number of vertices for arbitrary-shaped text spotting. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2025.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88(2), 2010.

Changyu Gu, Linwei Chen, Lin Gu, and Ying Fu. Fourier angle alignment for oriented object detection in remote sensing. *arXiv preprint arXiv:2602.23790*, 2026.

J. Han, J. Ding, J. Li, and G. S. Xia. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.*, 2021. doi: 10.1109/TGRS.2021.3062048.

Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. ReDet: A rotation-equivariant detector for aerial object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2021.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017.

Roshdi Khalil, Mohammed Al Horani, Abdelrahman Yousef, and Mohammad Sababheh. A new definition of fractional derivative. *Journal of computational and applied mathematics*, 264, 2014.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.*, 159, 2020.

Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.

Xiaohe Li and Jianping Wu. Developing a more reliable framework for extracting traffic data from a uav video. *IEEE Trans. Intell. Transp. Syst.*, 24(11), 2023.

Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE/CVF Int. Conf. Comput. Vis.*, 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Proc. Eur. Conf. Comput. Vis.* Springer, 2016.

Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proc. Int. Conf. Pattern Recognit. Appl. Methods.*, 2017.

José Henrique Lima Marques, Jeffri Murrugarra-Llerena, and Claudio R Jung. Gaucho: Gaussian distributions with cholesky decomposition for oriented object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025.

Qi Ming and Xiaowu Xiao. Towards accurate medical image segmentation with gradient-optimized dice loss. *IEEE Signal Process. Lett.*, 31, 2024. doi: 10.1109/LSP.2023.3329437.

Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. Dynamic anchor learning for arbitrary-oriented object detection. In *Proc. AAAI Conf. Artif. Intell.*, 2021.

Qi Ming, Lingjuan Miao, Zhe Ma, Lin Zhao, Zhiqiang Zhou, Xuhui Huang, Yuanpei Chen, and Yufei Guo. Deep dive into gradients: Better optimization for 3d object detection with gradient-corrected IoU supervision. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023.

Qi Ming, Lingjuan Miao, Zhiqiang Zhou, Junjie Song, and Aleksandra Pizurica. Gradient calibration loss for fast and accurate oriented bounding box regression. *IEEE Trans. Geosci. Remote Sens.*, 62, 2024.

Hanyang Peng and Shiqi Yu. A systematic iou-related method: Beyond simplified regression for better localization. *IEEE Trans. Image Process.*, 30, 2021.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6), 2016.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019.

Rudolf Scherer, Shyam L Kalla, Yifa Tang, and Jianfei Huang. The grünwald–letnikov method for fractional differential equations. *Computers & Mathematics with Applications*, 62(3), 2011.

Michael Ian Shamos. *Computational geometry*. Yale University, 1978.

Hualian Sheng, Sijia Cai, Na Zhao, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Min-Jian Zhao, and Gim Hee Lee. Rethinking iou-based optimization for single-stage 3d object detection. In *Proc. Eur. Conf. Comput. Vis.* Springer, 2022.

Xuepeng Shi, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Real-time rotation-invariant face detection with progressive calibration networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018.

Junjie Song, Lingjuan Miao, Zhiqiang Zhou, Qi Ming, and Yunpeng Dong. Optimized point set representation for oriented object detection in remote-sensing images. *IEEE Trans. Geosci. Remote Sens.*, 20, 2023.

Zanjia Tong, Yuhang Chen, Zewei Xu, and Rong Yu. Wise-iou: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051*, 2023.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018.

Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *IEEE/CVF Int. Conf. Comput. Vis.*, 2021.

Hang Xu, Xinyuan Liu, Haonan Xu, Yike Ma, Zunjie Zhu, Chenggang Yan, and Feng Dai. Rethinking boundary discontinuity problem for oriented object detection. In *Conf. Comput. Vis. Pattern Recognit.*, 2024.

Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *Proc. Eur. Conf. Comput. Vis.* Springer, 2020.

Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3Det: Refined single-stage detector with feature refinement for rotating object. *Proc. AAAI Conf. Artif. Intell.*, 35(4), May 2021a.

Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *Proc. Int. Conf. Mach. Learn.* PMLR, 2021b.

Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Proc. Adv. Neural Inf. Process. Syst.*, 2021c.

Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The kfiou loss for rotated object detection. *arXiv preprint arXiv:2201.12558*, 2022a.

Yang Yang, Jifeng Chen, Xiaopin Zhong, and Yuanlong Deng. Polygon-to-polygon distance loss for rotated object detection. *Proc. AAAI Conf. Artif. Intell.*, 36(3), 2022b. doi: 10.1609/aaai.v36i3.20214.

Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. Oriented object detection in aerial images with box boundary-aware vectors. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021.

Yi Yu, Botao Ren, Peiyuan Zhang, Mingxin Liu, Junwei Luo, Shaofeng Zhang, Feipeng Da, Junchi Yan, and Xue Yang. Point2rbox-v2: Rethinking point-supervised oriented object detection with spatial layout among instances. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025.

Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18), 2021.

Jiaqi Zhao, Zeyu Ding, Yong Zhou, Hancheng Zhu, Wen-Liang Du, Rui Yao, and Abdulmotaleb El Saddik. Orientedformer: An end-to-end transformer-based oriented object detector in remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 2024.

Jiaqi Zhao, Zeyu Ding, Yong Zhou, Hancheng Zhu, Wen-Liang Du, and Rui Yao. Rediffdet: Rotation-equivariant diffusion model for oriented object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025.

Yu Zheng, Danyang Zhang, Sinan Xie, Jiwen Lu, and Jie Zhou. Rotation-robust intersection over union for 3D object detection. In *Eur. Conf. Comput. Vis.*, 2020a.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proc. AAAI Conf. Artif. Intell.*, 2020b.

Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. Iou loss for 2d/3d object detection. In *Proc. Int. Conf. 3D Vis.* IEEE, 2019.

Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, Wenwei Zhang, and Kai Chen. Mmrotate: A rotated object detection benchmark using pytorch. In *Proc. ACM Int. Conf. Multimedia*, 2022.

Tianxiao Zhu, Zhonghua Miao, Chunjiang Zhao, Liping Chen, and Ya Xiong. Five-degree-of-freedom strawberry pose estimation using key points and oriented bounding box detection. *Computers and Electronics in Agriculture*, 239, 2025.

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *Int. J. Comput. Vis.*, 133(2), 2025.

A Full Gradient Derivation for Fractional IoU

For the angular parameter θ, we treat the semantic score s as constant with respect to geometric regression. The FrSIoU loss is $\mathcal{L}_{\text{FrSIoU}} = W_t L_{\text{base}}$, where $L_{\text{base}} = (1-s)(1-u)$ and $W_t = \exp(-\lambda D_t^{\nu}(\text{sIoU}))$. Its gradient is

$$\frac{\partial \mathcal{L}_{\text{FrSIoU}}}{\partial \theta} = \frac{\partial W_t}{\partial \theta} L_{\text{base}} + W_t \frac{\partial L_{\text{base}}}{\partial \theta}. \tag{15}$$

Since $\partial L_{\text{base}}/\partial \theta = -(1-s_t)\partial u_t/\partial \theta$ and

$$\frac{\partial D_t^{\nu}(\text{sIoU})}{\partial \theta} = \sum_{k=0}^{M} c_k (1-s_{t-k}) \frac{\partial u_{t-k}}{\partial \theta}, \tag{16}$$

we have $\partial W_t/\partial \theta = -\lambda W_t \partial D_t^{\nu}(\text{sIoU})/\partial \theta$. Therefore,

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{FrSIoU}}}{\partial \theta} &= W_t \left(\frac{\partial L_{\text{base}}}{\partial \theta} - \lambda L_{\text{base}} \frac{\partial D_t^{\nu}(\text{sIoU})}{\partial \theta} \right) \\ &= W_t \left[-(1-s_t)\frac{\partial u_t}{\partial \theta} - \lambda L_{\text{base}} \sum_{k=0}^{M} c_k (1-s_{t-k}) \frac{\partial u_{t-k}}{\partial \theta} \right]. \end{aligned} \tag{17}$$

This expression shows that the loss is regularized by historical states and stabilizes high-precision localization.

B Analysis of Gradient w.r.t. Angular Error with Skew IoU loss

We now derive closed-form expressions for the angular gradient of IoU-based losses under the parallelogram approximation (for two centre-aligned OBBs) and demonstrate why these gradients can grow and oscillate during optimization.

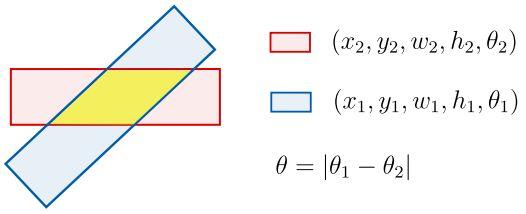


Figure 4: Visualization of the relationship of centre-aligned OBBs.

Derivative of the intersection area. From Fig.4, we have

$$I(\theta) \equiv A_{\cap}(\theta) = \frac{h_1\, h_2}{\sin\theta}, A = w_1 h_1 + w_2 h_2 \tag{18}$$

Differentiating with respect to θ yields

$$\frac{\mathrm{d}I}{\mathrm{d}\theta} = -\frac{h\, h^* \cos\theta}{\sin^2\theta} = -I(\theta)\cot\theta. \tag{19}$$

Derivative of IoU. Writing $\text{IoU}(\theta) = I(\theta)/(A - I(\theta))$, a straightforward application of the quotient rule gives

$$\frac{\mathrm{d}\,\text{IoU}}{\mathrm{d}\theta} = \frac{A\,\dfrac{\mathrm{d}I}{\mathrm{d}\theta}}{(A-I(\theta))^2}. \tag{20}$$

Substituting Eq. equation 19 into Eq. equation 20 yields

$$\frac{\mathrm{d}\,\text{IoU}}{\mathrm{d}\theta} = -\frac{A\, I(\theta)\cot\theta}{(A-I(\theta))^2}. \tag{21}$$

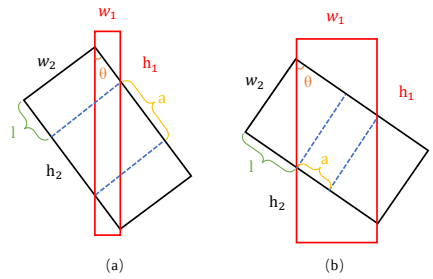


Figure 5: Illustration of two kinds of boundary cases.

Angular gradient of the linear IoU loss. For the linear loss $L_{\text{IoU}} = 1 - \text{IoU}$, we obtain

$$\frac{\partial L_{\text{IoU}}}{\partial \theta} = -\frac{\mathrm{d}\,\text{IoU}}{\mathrm{d}\theta} = \frac{A\,I(\theta)\cot\theta}{(A - I(\theta))^2}. \tag{22}$$

Noting the algebraic identity arising from $\text{IoU} = I/(A - I)$, namely

$$\frac{A\,I}{(A-I)^2} \;=\; \text{IoU}\,(1 + \text{IoU}),$$

Eq. equation 22 can be compactly written as

$$\frac{\partial L_{\text{IoU}}}{\partial \theta} = \text{IoU}\,\big(1 + \text{IoU}\big)\,\cot\theta\,. \tag{23}$$

C Boundary Condition for OBB Optimization

The boundary condition refers to the minimum angular error θ_0 that makes the intersection of two centre-aligned oriented bounding boxes a parallelogram. Fig. 5 shows two possible cases, and we next calculate θ_0 for both cases.

Obviously, $\theta_0 > 0$. For the case (a):

$$\begin{cases} \tan\theta = \dfrac{w_2}{l + a}, \\ \sin\theta = \dfrac{w_1}{l}, \\ 2l + a = h_2. \end{cases} \tag{24}$$

From the equations, we know:

$$\theta_0 = \arctan\left(\frac{w_2 \sin\theta}{h_2 \sin\theta - w_1}\right). \tag{25}$$

Similarly, the geometric relationship in case (b) is as follows:

$$\begin{cases} \tan\theta = \dfrac{w_2}{l}, \\ \sin\theta = \dfrac{w_1}{l + a}, \\ 2l + a = h_2. \end{cases} \tag{26}$$

Then,

$$\theta_0 = \arctan\left(\frac{w_2 \sin\theta}{h_2 \sin\theta - w_1}\right). \tag{27}$$

Therefore, the minimum angular errors θ_0 are the same under both boundary conditions.

Although the IoU loss exhibits gradient singularities with respect to θ, it remains usable for model training since $\theta_0 > 0$, preventing θ from reaching zero in practice. When $\theta < \theta_0$, the IoU loss is free from severe gradient oscillations. However, for arbitrary oriented bounding boxes, θ often exceeds θ_0, causing the model to experience persistent unstable gradients before reaching the boundary condition, which results in loss oscillation.

D Analysis of Fractional Gradient Stabilization

In this section, we firstly analyse the gradient stabilization of the proposed fractional semantic IoU (FrSIoU) loss $\mathcal{L}_{FrSIoU} \in \mathbb{R}$. Then, the fractional modulation weight W_t is proved to be a spectral filter that regularizes the gradient.

D.1 Polynomial Decay of Memory Kernel

With the proposed $v \in (0,1)$ order Grünwald-Letnikov (G-L) operator $\mathcal{D}_t^v$ in Section 3, the memory kernel $\{c_k\}_{k=0}^{\infty}$ shows power-law decay $c_k \sim k^{-(1+v)}$, while the standard momentum methods shows exponential decay $h_k \sim \alpha^k$.

Proof. The fractional derivative of $s_t \equiv \mathrm{sIoU}_t$ is defined as:

$$\mathcal{D}_t^v[s] = \sum_{k=0}^{t} c_k s_{t-k} \tag{28}$$

where the memory kernel $c_k = (-1)^k \binom{v}{k}$.

With the Gamma function $\Gamma(\cdot)$, c_k is

$$c_k = \frac{\Gamma(k-v)}{\Gamma(-v)\Gamma(k+1)} \tag{29}$$

Then, applying the approximation $\Gamma(z) \approx \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z$ as $k \to \infty$:

$$\frac{\Gamma(k-v)}{\Gamma(k+1)} \approx \frac{(k-v)^{k-v-1/2} e^{-(k-v)}}{(k+1)^{k+1/2} e^{-(k+1)}} \approx k^{-v-1} \tag{30}$$

Thus, the polynomial decay of the memory kernel is:

$$\lim_{k\to\infty} c_k = \frac{1}{\Gamma(-v)} k^{-(1+v)} \tag{31}$$

□

Finally, FrSIoU include gradients from historical states to stabilize the training process against local noise $\eta_t \sim \mathcal{N}(0, \sigma^2)$.

$$\left| \frac{\partial \mathcal{L}_{FrSIoU}}{\partial s_{t-k}} \right| \propto k^{-(1+v)} \gg \alpha^k \quad \text{(for large } k, \alpha < 1) \tag{32}$$

D.2 Low-Pass Filtering in Frequency Domain

The proposed fractional derivative operator $\mathcal{D}^v$ is a linear time-invariant filter with lower high-frequency gain than the integer derivative $\mathcal{D}^1$.

With $S(z)$ being the Z-transform of the input loss sequence s_t. The G-L operator is:

$$H_v(z) = (1 - z^{-1})^v \tag{33}$$

By substitution $z = e^{j\omega}$, the magnitude at angular frequency $\omega \in [0, \pi]$ is

$$|H_v(e^{j\omega})| = |1 - e^{-j\omega}|^v = \left| e^{-j\frac{\omega}{2}} (e^{j\frac{\omega}{2}} - e^{-j\frac{\omega}{2}}) \right|^v \tag{34}$$

$$|H_v(\omega)| = \left| 2 \sin\left(\frac{\omega}{2}\right) \right|^v \tag{35}$$

Then the spectral comparison at Nyquist limit ($\omega = \pi$) is as follows:

1. **Integer Derivative ($v = 1$):**

$$|H_1(\pi)| = |2\sin(\pi/2)|^1 = 2$$

Magnitude factor is 2.0, which means maximal noise sensitivity.

2. **Fractional Derivative ($v = 0.5$):**

$$|H_{0.5}(\pi)| = |2\sin(\pi/2)|^{0.5} = \sqrt{2} \approx 1.414$$

The magnitude factor is 1.414, which means reduced noise sensitivity.

Thus, FrSIoU suppresses high-frequency gradient jitter by a magnitude factor of 2^{1-v} compared with the standard momentum difference methods.

$$\forall \omega \in (0, \pi], v \in (0, 1) \implies |H_v(\omega)| < |H_1(\omega)| \tag{36}$$