DIFFERENT INSTANCES, DIFFERENT SUPERVISION: INSTANCE-ADAPTIVE POINT-SUPERVISED ORIENTED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Point annotations reduce labeling costs for oriented object detection but provide neither object scale nor orientation. Complementary mask-based supervision sources provide extent cues, yet their relative reliability varies with the local spatial configuration of each instance. A patch-wise source choice can therefore be unsuitable for individual instances when isolated and densely clustered objects coexist in remote-sensing imagery. We find that local spatial configuration provides an effective cue for instance-level supervision selection, yet spatial cues alone cannot fully determine which source is most suitable for each instance. Based on this observation, we propose **InstaPoint**, an instance-adaptive framework that combines spatially guided initialization with prediction-guided supervision assignment. **Instance Density Calibration (IDC)** provides prediction-independent instance-specific initial assignments, while **Instance-wise Supervision Assignment (ISA)** subsequently selects, for each instance, the candidate supervision most consistent with the detector's current prediction and retains the selected source for the remainder of training. ISA revises the initial supervision source for 47.62% of instances on DIOR-R and 42.99% on DOTA-v1.0, indicating substantial instance-wise reassignment beyond the spatial initialization. InstaPoint changes how extent supervision is assigned while retaining the baseline detector architecture and other training objectives. Under the end-to-end and two-stage pseudo-labeling protocols, InstaPoint achieves 42.49/47.10 AP_{50} on DIOR-R and 62.53/67.74 AP_{50} on DOTA-v1.0, surpassing the strongest point-supervised baseline by 0.79/0.50 points on DIOR-R and 2.92/1.65 points on DOTA-v1.0.

1 INTRODUCTION

Oriented Object Detection (OOD) has been widely studied in remote sensing, with applications including urban monitoring, maritime surveillance, and traffic analysis. In aerial and satellite imagery, objects often appear at arbitrary orientations and in dense spatial layouts, making rotated bounding boxes (RBoxes) suitable for accurately representing their locations and orientations (Wen et al., 2023). However, fully supervised OOD requires precise RBox annotations for every instance, resulting in substantial annotation costs (Xia et al., 2018; Yang et al., 2023a). To alleviate this burden, Weakly Supervised Oriented Object Detection (WS-OOD) has been explored using weaker annotations such as horizontal bounding boxes (Yang et al., 2023a; Yu et al., 2023; Lee et al., 2025) and points (Cao et al., 2023; Luo et al., 2024; Ren et al., 2025; Yu et al., 2024; Zhang et al., 2024; 2025; Liu et al., 2025; Yu et al., 2025a; Zhang et al., 2026). In particular, point supervision provides an efficient annotation scheme but does not explicitly provide object extent or orientation, making the recovery of geometric information a key challenge in point-supervised OOD.

Recent point-supervised OOD methods recover missing geometric information from sparse point annotations using various auxiliary cues (Cao et al., 2023; Luo et al., 2024; Ren et al., 2025; Yu et al., 2024; Zhang et al., 2024; 2025; Liu et al., 2025; Yu et al., 2025a; Zhang et al., 2026). More recently, Point2RBox-v3 (Zhang et al., 2026) combines Segment Anything Model (SAM) (Kirillov et al., 2023) and watershed (Yu et al., 2025a) masks as complementary supervision for recovering object geometry. SAM and watershed exhibit different characteristics depending on the surrounding object distribution. Watershed exploits spatial relationships among annotated points and becomes

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.

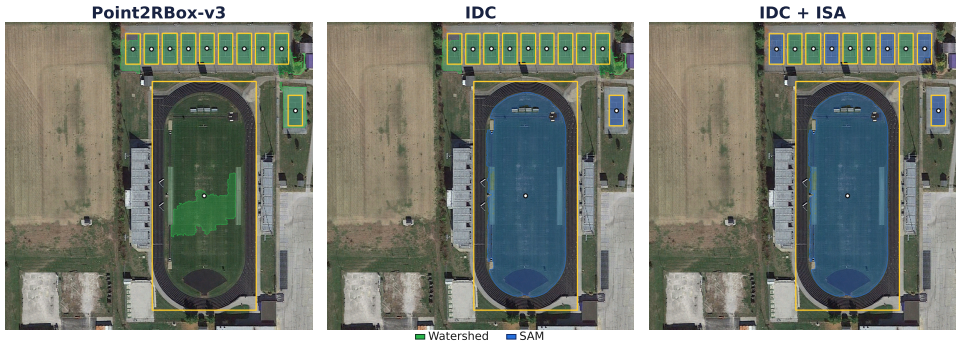


Figure 1: Qualitative comparison of supervision assignment. Point2RBox-v3 (Zhang et al., 2026) uses a shared patch-wise supervision source, while our Instance Density Calibration (IDC) and Instance-wise Supervision Assignment (ISA) perform instance-wise source assignment between SAM (blue) and watershed (green) supervision. Yellow boxes denote the RBoxes and white dots denote point annotations.

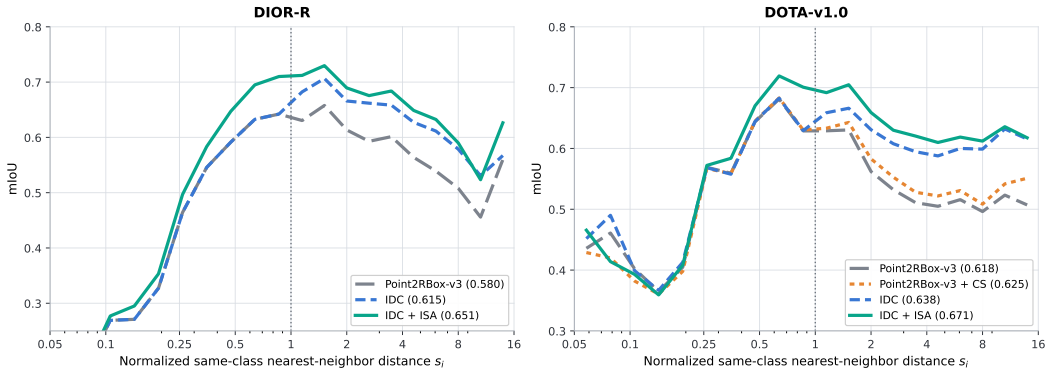


Figure 2: Comparison of supervision assignment strategies on DIOR-R and DOTA-v1.0. We report the mean IoU between the GT RBox and the RBox constructed from the selected SAM mask or watershed region, centered at the annotated point and oriented by the GT angle. Instances are binned by s_i; the dashed line denotes the IDC boundary at $s_i = 1$. CS denotes the Class-Specific Watershed.

less reliable for isolated objects where neighboring constraints are limited, whereas SAM can provide informative masks for isolated objects but may merge adjacent instances in densely populated regions (Zhang et al., 2026). This complementarity raises an important question: *which supervision should be assigned to each individual instance?*

However, the patch-wise supervision assignment used in Point2RBox-v3 (Zhang et al., 2026) cannot fully capture such instance-wise variation. Remote-sensing images often contain heterogeneous layouts in which dense clusters and isolated objects coexist within the same scene. As shown in Fig. 1, watershed supervision can appropriately constrain instances within densely populated regions, while an isolated instance in the same image may be better represented by SAM. This mismatch motivates shifting the assignment granularity from the image to the individual instance.

We further analyze this mismatch quantitatively in Fig. 2. We group instances by the normalized same-class nearest-neighbor distance s_i defined in equation 6, where larger values indicate greater local isolation. For this analysis, we fix the center and orientation of each mask-derived RBox to the annotated point and the GT angle, respectively, so that its IoU with the GT RBox primarily reflects the quality of the recovered object extent. The differences among assignment strategies are relatively small in dense neighborhoods, whereas instance-wise spatial assignment yields more accurate extent targets for locally isolated instances. On DOTA-v1.0 (Xia et al., 2018), 88.0% of instances with $s_i > 1$ are located in split patches containing more than four objects and therefore receive watershed supervision under the patch-wise rule despite being locally isolated. Detailed overall and class-wise statistics are provided in Appendix Table 5. These results highlight the mismatch between patch-

wise source selection and instance-wise spatial configurations, and identify local point spacing as a useful cue.

However, local spacing describes the arrangement of neighboring annotations rather than directly assessing which candidate mask better represents the object extent. We therefore propose InstaPoint, an instance-adaptive supervision assignment framework that uses spatial cues for initialization and detector predictions for subsequent refinement. Specifically, Instance Density Calibration (IDC) uses class-calibrated local spacing to assign SAM or watershed supervision to each instance during warm-up without relying on detector predictions. After a fixed warm-up period, Instance-wise Supervision Assignment (ISA) compares the extent targets derived from both sources with the detector's current prediction and performs a one-time reassignment by selecting the source with the lower regression loss. Importantly, ISA changes the supervision source for 47.62% of instances on DIOR-R and 42.99% on DOTA-v1.0 (in Appendix Table 4), demonstrating substantial refinement of the spatial initialization. The selected source is then retained for the remainder of training. Our proposed InstaPoint achieves consistent gains in detection performance without modifying the detector architecture or introducing additional training objectives. These gains highlight the importance of obtaining complementary supervision and assigning the appropriate source to each instance.

Our main contributions are summarized as follows:

- We identify a mismatch between patch-wise supervision assignment and instance-wise spatial configurations. Our analysis of mask-derived extent targets shows that accounting for local point spacing improves supervision quality, particularly for locally isolated instances.
- We propose InstaPoint, an instance-adaptive assignment framework that combines prediction-independent spatial initialization with a one-time prediction-guided reassignment. Instance Density Calibration (IDC) initializes source selection using class-calibrated local spacing, while Instance-wise Supervision Assignment (ISA) refines this selection after warm-up by comparing candidate extent targets with the detector's current prediction.
- Extensive experiments on DOTA-v1.0 and DIOR-R demonstrate that InstaPoint achieves state-of-the-art performance under both end-to-end and two-stage pseudo-labeling protocols, reaching 62.53/67.74 AP_{50} on DOTA-v1.0 and 42.49/47.10 on DIOR-R.

The code will be publicly released to facilitate reproducibility and further research.

2 Related Work

2.1 Fully Supervised Oriented Object Detection

Oriented object detection (OOD) localizes arbitrarily oriented objects using rotated bounding boxes (RBoxes). Fully supervised OOD has been extensively studied through orientation-aware feature learning and specialized rotated detectors (Ding et al., 2019; Yang et al., 2019; Han et al., 2021a; Yang et al., 2021a; Han et al., 2021b; Xie et al., 2021). Rotated box regression has also been improved through specialized regression objectives (Yang et al., 2021b;c; 2023b). These methods require RBox annotations that specify object location, extent, and orientation. Obtaining such precise geometric annotations is labor-intensive, especially for large-scale remote-sensing datasets. This motivates weakly supervised OOD using cheaper forms of supervision.

2.2 Weakly Supervised Oriented Object Detection

Weakly supervised OOD reduces annotation requirements by replacing RBox annotations with coarser supervision, such as horizontal bounding boxes (HBoxes) or points. HBoxes retain axis-aligned object extent but omit orientation, whereas point annotations provide only object locations. Recent studies further extend weakly supervised OOD to more flexible supervision settings, including diversified and partially available weak annotations (Yu et al., 2025b; Liu et al., 2026a).

HBox-Supervised Oriented Object Detection H2RBox (Yang et al., 2023a) learns object orientation through geometric consistency across transformed views. H2RBox-v2 (Yu et al., 2023) further exploits object symmetry for orientation learning, while EIE-Det (Wang et al., 2025) introduces explicit and implicit box equivariance under HBox supervision. More recently, ABBSPO (Lee et al.,

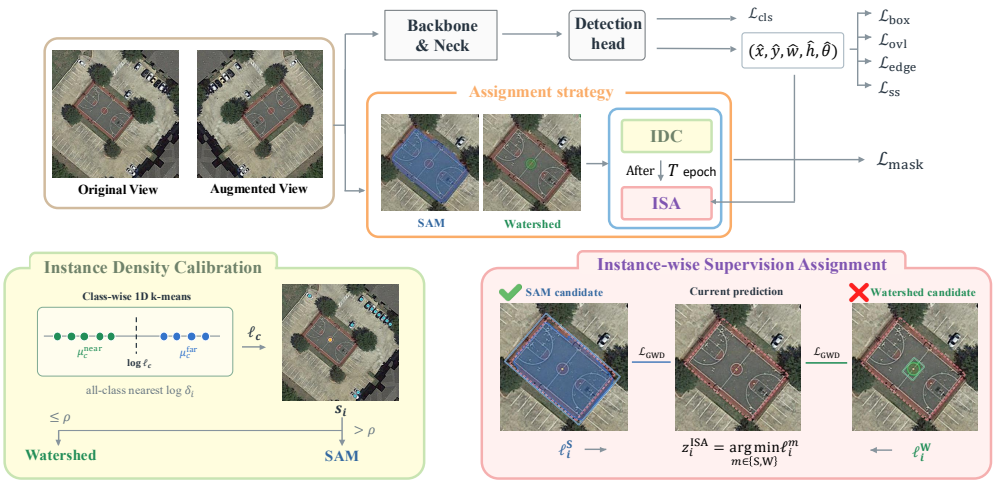


Figure 3: Overview of InstaPoint. IDC initializes instance-wise assignments to SAM or watershed using class-calibrated point spacing. ISA then makes a one-time selection of the source with the lower GWD-based size loss relative to the detector's current prediction.

2025) combines adaptive box scaling with symmetry-prior-based orientation learning. HBoxes provide axis-aligned spatial bounds but omit object orientation.

Point-Supervised Oriented Object Detection With only object locations available, point-supervised OOD must recover both object extent and orientation from additional visual or spatial cues. Point2RBox (Yu et al., 2024) combines synthetic visual patterns with transformation-based self-supervision for end-to-end oriented detection. PointOBB (Luo et al., 2024) and PointOBB-v3 (Zhang et al., 2025) exploit multi-view learning for scale and orientation estimation, whereas PointOBB-v2 (Ren et al., 2025) derives pseudo RBoxes from probability maps.

Segmentation priors offer another way to estimate object extent, with SAM-based approaches exploiting point-prompted masks to derive pseudo-box supervision (Cao et al., 2023; Zhang et al., 2024; Liu et al., 2025; Lu & Bie, 2025). However, SAM-based supervision becomes less reliable in densely packed scenes (Liu et al., 2025; Zhang et al., 2026). Point2RBox-v2 (Yu et al., 2025a) exploits spatial relationships among annotated instances through Gaussian-overlap and Voronoi-watershed supervision, while recent approaches exploit instance density for sample assignment and spatial partitioning for both sample assignment and pseudo-label generation (Chen et al., 2026; Liu et al., 2026b). Yet watershed-based spatial supervision becomes less reliable for isolated instances with limited neighboring geometry (Zhang et al., 2026). Together, these complementary failure modes motivate adapting supervision to the local spatial configuration of individual instances.

Specifically, Point2RBox-v3 (Zhang et al., 2026) combines SAM and watershed supervision through PGDM-Loss, but performs supervision-source assignment in a patch-wise manner based on the instance count. Such patch-wise assignment cannot account for spatial heterogeneity within the same patch. In contrast, InstaPoint performs instance-wise supervision-source assignment; to the best of our knowledge, it is the first point-supervised OOD framework to do so.

3 Method

3.1 Overview

As illustrated in Fig. 3, **InstaPoint** assigns complementary SAM and watershed supervision to individual instances through spatial initialization followed by a one-time prediction-guided reassignment. We build on Point2RBox-v3 (Zhang et al., 2026) and replace its patch-wise source selection with instance-adaptive assignment.

During warm-up, **Instance Density Calibration (IDC)** uses class-calibrated local point spacing to select a supervision source for each instance, without relying on detector predictions. After warm-

up, **Instance-wise Supervision Assignment (ISA)** compares the extent targets derived from both sources with the detector's current prediction and selects the source yielding the lower regression loss. This source choice is made once per instance and retained for the remainder of training.

In both phases, the selected mask is converted into a width-height target using the current predicted orientation to supervise object extent. Although the source is fixed after reassignment, its target continues to be recomputed as the predicted orientation changes. InstaPoint thereby adapts extent supervision to individual instances while preserving the baseline detector architecture, mask-to-target construction, and remaining training objectives.

3.2 Instance Density Calibration

To initialize supervision assignment without relying on early detector predictions, IDC uses the locations and class labels of point annotations. Instance Density Calibration (IDC) therefore provides an initial supervision-source assignment using only the spatial distribution of point annotations. Specifically, IDC estimates a class-specific spacing reference from the training annotations and uses it to calibrate the local spacing of each instance. This allows the same raw point spacing to be interpreted relative to the typical spacing of each class.

Let $\mathbf{p}_i$ and c_i denote the point location and class label of instance i, respectively. All point-spacing measurements are computed within the same detector input image, corresponding to an input patch when patch-based training is used.

We first collect point-spacing statistics from the training set. For each annotation point $\mathbf{p}_i$, we compute the distance to its nearest neighboring annotation in the same image, regardless of class:

$$\delta_i = \min_{j \neq i} \|\mathbf{p}_i - \mathbf{p}_j\|_2 . \tag{1}$$

Repeating this computation over the training set, we collect valid log-transformed distances according to the class of the corresponding instance:

$$\mathcal{D}_c = \{\log \delta_i \mid c_i = c,\ 0 < \delta_i < \infty\} . \tag{2}$$

Thus, for each class c, we apply one-dimensional k-means clustering with $K = 2$ to $\mathcal{D}_c$, obtaining two cluster centers in log-distance space:

$$\mu_c^{\text{near}} < \mu_c^{\text{far}}. \tag{3}$$

We define the class-specific spacing reference by taking the midpoint between these centers and transforming it back to the original distance scale:

$$\ell_c = \exp\left(\frac{\mu_c^{\text{near}} + \mu_c^{\text{far}}}{2}\right). \tag{4}$$

This reference provides a class-specific scale for normalizing local point spacing. For classes with insufficient valid samples for class-wise clustering, we use a fallback spacing reference estimated from the pooled point-spacing statistics of the entire training set.

Given the class-specific reference, IDC measures each instance's separation from neighboring annotations of the same class. For instance i, we compute

$$d_i = \min_{j \neq i,\ c_j = c_i} \|\mathbf{p}_i - \mathbf{p}_j\|_2 , \tag{5}$$

where the search is restricted to the same image, and $d_i = \infty$ if no other instance of the same class exists in the image. We then calibrate the local spacing by the corresponding class-specific reference:

$$s_i = \frac{d_i}{\ell_{c_i}}. \tag{6}$$

A larger s_i indicates greater separation from same-class neighbors relative to the class-specific reference, whereas a smaller value indicates a more closely spaced neighborhood. IDC assigns the initial supervision source as

$$z_i^{\text{IDC}} = \begin{cases} \text{SAM}, & \text{if } s_i > \rho, \\ \text{Watershed}, & \text{if } s_i \leq \rho, \end{cases} \tag{7}$$

where ρ controls the assignment boundary and is set to 1 by default.

3.3 Instance-wise Supervision Assignment

IDC initializes supervision assignment using local point spacing, but this spatial criterion does not directly assess the extent targets derived from the candidate masks. After T warm-up epochs, Instance-wise Supervision Assignment (ISA) performs a one-time reassignment by comparing the SAM and watershed targets with the detector's current prediction.

Orientation-aligned extent targets. For instance i, let $\hat{w}_i, \hat{h}_i$, and $\hat{\theta}_i$ denote the predicted width, height, and orientation, respectively. For a candidate mask M_i^m, where $m \in \{\mathrm{SAM}, \mathrm{Watershed}\}$, we derive an orientation-aligned width-height target using the current predicted orientation:

$$\begin{bmatrix} w_i^m \\ h_i^m \end{bmatrix} = 2 \max_{\mathbf{x} \in M_i^m} \left| R(\hat{\theta}_i)^\top (\mathbf{x} - \mathbf{p}_i) \right|, \tag{8}$$

where $\mathbf{x}$ denotes a foreground pixel coordinate in the candidate mask, $R(\cdot)$ denotes a two-dimensional rotation matrix, and the absolute value and maximum are applied element-wise. The target width and height are obtained by doubling the maximum absolute foreground offsets from the annotated point $\mathbf{p}_i$ along the predicted box axes. The resulting target is detached from the computation graph, preventing gradients from propagating to $\hat{\theta}_i$ through target construction. The same construction is used for the source assigned by IDC during warm-up.

Prediction-guided source selection. We then compare each candidate target with the predicted width and height using Gaussian Wasserstein Distance (GWD) loss (Yang et al., 2021b):

$$\ell_i^m = \mathcal{L}_{\mathrm{GWD}} \left((\hat{w}_i, \hat{h}_i), (w_i^m, h_i^m) \right). \tag{9}$$

This criterion evaluates candidate extent targets under the same regression objective used to supervise the detector. ISA assigns the source yielding the lower loss:

$$z_i^{\mathrm{ISA}} = \underset{m \in \{\mathrm{SAM}, \mathrm{Watershed}\}}{\arg\min} \ell_i^m. \tag{10}$$

The assigned source z_i^{ISA} is fixed for the remainder of training. Its corresponding width-height target, however, is recomputed at each iteration using the current predicted orientation, so that the target remains aligned with the evolving detector prediction.

3.4 Training Objective

During the first T epochs, IDC determines the supervision source for each instance. After warm-up, ISA performs a one-time reassignment, and the resulting source assignments are retained for the remainder of training. Accordingly, the source used for instance i at epoch $e \geq 1$ is

$$z_i(e) = \begin{cases} z_i^{\mathrm{IDC}}, & \text{if } e \leq T, \\ z_i^{\mathrm{ISA}}, & \text{if } e > T. \end{cases} \tag{11}$$

For an input image with N annotated instances, the mask supervision loss is

$$\mathcal{L}_{\mathrm{mask}} = \frac{1}{N} \sum_{i=1}^{N} \ell_i^{z_i(e)}, \tag{12}$$

where each regression loss is evaluated using the current prediction and the orientation-aligned extent target from the assigned source.

The overall training objective follows the baseline formulation:

$$\begin{aligned} \mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{cls}} \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{box}} \mathcal{L}_{\mathrm{box}} + \lambda_{\mathrm{mask}} \mathcal{L}_{\mathrm{mask}} \\ + \lambda_{\mathrm{ovl}} \mathcal{L}_{\mathrm{ovl}} + \lambda_{\mathrm{edge}} \mathcal{L}_{\mathrm{edge}} + \lambda_{\mathrm{ss}} \mathcal{L}_{\mathrm{ss}}. \end{aligned} \tag{13}$$

Here, $\mathcal{L}_{\mathrm{cls}}$, $\mathcal{L}_{\mathrm{box}}$, $\mathcal{L}_{\mathrm{ovl}}$, $\mathcal{L}_{\mathrm{edge}}$, and $\mathcal{L}_{\mathrm{ss}}$ denote the classification, box regression, overlap, edge, and geometric self-consistency losses, respectively. IDC and ISA determine which source provides the

extent target for $\mathcal{L}_{\text{mask}}$, while the remaining loss formulations are unchanged. Following the baseline configuration, we set

$$(\lambda_{\text{cls}}, \lambda_{\text{box}}, \lambda_{\text{mask}}, \lambda_{\text{ovl}}, \lambda_{\text{ss}}) = (1, 5, 5, 10, 1). \tag{14}$$

We set $T = 6$ by default. The edge loss follows the baseline schedule with $\lambda_{\text{edge}} = 0$ for $e \leq 6$ and $\lambda_{\text{edge}} = 0.3$ for $e > 6$.

4 Experiments

4.1 Setup

Datasets. We evaluate on two widely used oriented object detection benchmarks. DOTA-v1.0 (Xia et al., 2018) contains 2,806 high-resolution aerial images across 15 categories, with large variations in object scale and spatial density, including many small, densely distributed instances. DIOR-R (Cheng et al., 2022) comprises 23,463 remote-sensing images across 20 categories with diverse object scales and scene types. For both datasets, we train on the trainval split using only point annotations and evaluate on the test set.

Evaluation and Implementation Details. We report AP_{50} under the end-to-end and two-stage protocols; the latter trains FCOS on predicted pseudo RBoxes. We build on Point2RBox-v3 (Zhang et al., 2026), with MobileSAM (Zhang et al., 2023) and Class-Specific Watershed (CS) (Zhang et al., 2026) as complementary extent supervision. IDC statistics are computed on the preprocessed training images (split patches for DOTA-v1.0). All models are trained for 12 epochs with batch sizes of 2 (point-supervised model) and 4 (FCOS); other settings follow Point2RBox-v3.

4.2 Results

Quantitative Results. Tables 1 and 2 summarize the quantitative results on DOTA-v1.0 and DIOR-R. InstaPoint achieves 62.53/67.74 AP_{50} on DOTA-v1.0 and 42.49/47.10 AP_{50} on DIOR-R under the end-to-end/two-stage protocols, respectively. Compared with the strongest point-supervised baselines, InstaPoint improves performance by 2.92/1.65 AP_{50} on DOTA-v1.0 and 0.79/0.50 AP_{50} on DIOR-R. The consistent improvements across both datasets and evaluation protocols demonstrate the effectiveness of instance-adaptive supervision assignment.

Qualitative Results. Figure 4 presents qualitative comparisons with the fully supervised Rotated FCOS (Tian et al., 2019), the HBox-supervised H2RBox-v2 (Yu et al., 2023), and the point-supervised Point2RBox-v3 (Zhang et al., 2026). Despite using only point supervision, InstaPoint produces more accurate and better-aligned RBoxes. On DIOR-R, it detects the aircraft in the lower-right region missed by all baselines and better aligns the surrounding aircraft, with more consistent orientation and box extent. On DOTA-v1.0, InstaPoint accurately localizes the highlighted object, whereas the baselines either miss it or produce a noticeably misaligned RBox.

Ablation Study. Table 3 analyzes the proposed two-phase supervision assignment framework, the assignment threshold ρ, and the transition timing from IDC to ISA. As shown in Table 3(a), IDC-only applies IDC throughout training, while ISA-only uses the baseline assignment for the first six epochs and activates ISA from epoch 7 onward. The full model instead uses IDC during the first six epochs before transitioning to ISA. ISA-only consistently improves upon the baseline, whereas IDC alone yields inconsistent gains. However, using IDC to initialize the assignments before ISA achieves the best performance on both datasets. Compared with ISA-only, the full IDC–ISA configuration improves E2E/FCOS AP_{50} by 1.90/1.08 points on DIOR-R and 2.30/1.73 points on DOTA-v1.0. These results indicate that IDC serves as a more effective initialization for the subsequent prediction-guided ISA stage than the baseline assignment.

Intuitively, local spacing describes the arrangement of neighboring annotations, whereas the detector provides an image-conditioned estimation of object extent. When this estimate approximately captures an individual object, it can help distinguish a suitable candidate from one that substantially overextends or includes neighboring instances. Consistent with this intuition, Fig. 2 shows

Table 1: Per-category AP_{50} results on DOTA-v1.0. $\mathcal{S}_R$, $\mathcal{S}_H$, and $\mathcal{S}_P$ denote RBox-, HBox-, and point-supervised settings, respectively; ✓/× denote end-to-end/two-stage protocols. P2RBox, PMS-SAM-RSD, and Point2RBox-v3 use a pre-trained SAM model, while Point2RBox+SK uses one-shot class sketches. BoxInst-RBox and Point2Mask-RBox obtain RBoxes by minimum-area rectangle fitting. Existing results are taken from Zhang et al. (2026), except Point2RBox-v3+CS†, which is our reproduction with Class-Specific Watershed.

	Methods	E2E	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP_{50}
$\mathcal{S}_R$	RepPoints (2022)	✓	86.7	81.1	41.6	62.0	76.2	56.3	75.7	90.7	80.8	85.3	63.3	66.6	59.1	67.6	33.7	68.45
	RetinaNet (2017)	✓	88.2	77.0	45.0	69.4	71.5	59.0	74.5	90.8	84.9	79.3	57.3	64.7	62.7	66.5	39.6	68.69
	GWD (2021b)	✓	89.3	75.4	47.8	61.9	79.5	73.8	86.1	90.9	84.5	79.4	55.9	59.7	63.2	71.0	45.4	71.66
	FCOS (2019)	✓	89.1	76.9	50.1	63.2	79.8	79.8	87.1	90.4	80.8	84.6	59.7	66.3	65.8	71.3	41.7	72.44
	S^2A-Net (2021a)	✓	89.2	83.0	52.5	74.6	78.8	79.2	87.5	90.9	84.9	84.8	61.9	68.0	70.7	71.4	59.8	75.81
$\mathcal{S}_H$	Sun et al. (2021)	×	51.5	38.7	16.1	36.8	29.8	19.2	23.4	83.9	50.6	80.0	18.9	50.2	25.6	28.7	25.5	38.60
	BoxInst-RBox (2021)	×	68.4	40.8	33.1	32.3	46.9	55.4	56.6	79.5	66.8	82.1	41.2	52.8	52.8	65.0	30.0	53.59
	H2RBox (2023a)	✓	88.5	73.5	40.8	56.9	77.5	65.4	77.9	90.9	83.2	85.3	55.3	62.9	52.4	63.6	43.3	67.82
	EIE-Det (2025)	✓	87.7	70.2	41.5	60.5	80.7	76.3	86.3	90.9	82.6	84.7	53.1	64.5	58.1	70.4	43.8	70.10
	H2RBox-v2 (2023)	✓	89.0	74.4	50.0	60.5	79.8	75.3	86.9	90.9	85.1	85.0	59.2	63.2	65.2	70.5	49.7	72.31
$\mathcal{S}_P$	Point2Mask-RBox (2023)	×	4.0	23.1	3.8	1.3	15.1	1.0	3.3	19.0	1.0	29.1	0.0	9.5	7.4	21.1	7.1	9.72
	P2BNet+H2RBox (2023a)	×	24.7	35.9	7.1	27.9	3.3	12.1	17.5	17.5	0.8	34.0	6.3	49.6	11.6	27.2	18.8	19.63
	P2BNet+H2RBox-v2 (2023)	×	11.0	44.8	14.9	15.4	36.8	16.7	27.8	12.1	1.8	31.2	3.4	50.6	12.6	36.7	12.5	21.87
	P2RBox (2023)	×	87.8	65.7	15.0	60.7	73.0	71.7	78.9	81.5	44.5	81.2	41.2	39.3	45.5	57.5	41.2	59.04
	PointOBB (2024)	×	26.1	65.7	9.1	59.4	65.8	34.9	29.8	0.5	2.3	16.7	0.6	49.0	21.8	41.0	36.7	30.08
	Point2RBox (2024)	✓	62.9	64.3	14.4	35.0	28.2	38.9	33.3	25.2	2.2	44.5	3.4	48.1	25.9	45.0	22.6	34.07
	Point2RBox+SK (2024)	✓	53.3	63.9	3.7	50.9	40.0	39.2	45.7	76.7	10.5	56.1	5.4	49.5	24.2	51.2	33.8	40.27
	Point2RBox+SK (2024)	×	66.4	59.5	5.2	52.6	54.1	53.9	57.3	90.8	3.2	57.8	6.1	47.4	22.9	55.7	40.5	44.90
	PointOBB-v2 (2025)	×	64.5	27.8	1.9	36.2	58.8	47.2	53.4	90.5	62.2	45.3	12.1	41.7	8.1	43.7	32.0	41.68
	PMS-SAM-RSD (2025)	✓	69.0	39.5	6.7	44.8	64.7	71.9	79.6	79.8	2.7	60.0	12.1	32.6	39.6	44.8	42.5	46.00
	PointOBB-v3 (2025)	✓	30.9	39.4	13.5	22.7	61.2	7.0	43.1	62.4	59.8	47.3	2.7	45.1	16.8	55.2	11.4	41.29
	PointOBB-v3 (2025)	×	52.9	54.4	21.3	52.7	65.6	44.9	67.8	87.2	26.7	73.4	32.6	53.3	39.0	56.4	10.2	49.24
	Point2RBox-v2 (2025a)	✓	78.4	52.7	8.3	40.9	71.0	60.5	74.7	88.7	65.5	72.1	24.4	26.1	30.1	50.7	21.0	51.00
	Point2RBox-v2 (2025a)	×	88.0	72.6	8.0	46.2	79.6	76.3	86.9	89.1	79.7	82.9	26.2	45.3	45.8	66.3	46.3	62.61
	Point2RBox-v3 (2026)	✓	86.5	53.4	35.7	37.5	78.8	75.3	86.3	86.6	66.0	80.2	29.7	49.9	36.7	59.1	32.5	59.61
	Point2RBox-v3 (2026)	×	89.0	72.5	41.6	45.1	79.4	76.4	87.7	85.1	75.7	84.8	44.4	55.4	45.7	59.8	48.8	66.09
	Point2RBox-v3+CS† (2026)	✓	85.5	58.3	38.7	39.1	78.8	74.6	86.7	86.6	64.0	79.4	18.9	46.2	42.4	58.5	34.5	59.49
	Point2RBox-v3+CS† (2026)	×	88.7	73.9	42.0	47.9	79.4	76.4	87.5	87.7	69.2	84.6	27.2	52.3	51.7	60.0	48.3	65.11
	InstaPoint (Ours)	✓	83.8	58.6	35.5	43.9	78.9	75.5	86.8	87.9	75.9	81.0	35.8	43.6	43.0	63.2	44.6	**62.53**
	InstaPoint (Ours)	×	88.4	72.5	36.2	50.7	79.5	77.1	87.8	87.7	79.7	84.5	48.8	54.1	53.4	66.1	49.7	**67.74**

Table 2: Per-category AP_{50} results on DIOR-R. $\mathcal{S}_R$, $\mathcal{S}_H$, and $\mathcal{S}_P$ denote RBox-, HBox-, and point-supervised settings, respectively; ✓/× denote end-to-end/two-stage protocols. Point2RBox-v3 and InstaPoint use pre-trained SAM, while Point2RBox+SK uses one-shot class sketches. * denotes Official checkpoint/log results; † denotes our reproduction of Point2RBox-v3 with both the original PGDM-Loss and Class-Specific Watershed, where PGDM routes patches with $J \leq 4$ to SAM, and J denotes the number of instances within a patch. – indicates unavailable class-wise results.

	Methods	E2E	APL	APO	BF	BC	BR	CH	ESA	ETS	DAM	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	AP_{50}
$\mathcal{S}_R$	RetinaNet (2017)	✓	59.8	19.3	69.7	81.3	17.2	72.7	68.7	49.4	18.4	69.5	71.3	33.3	34.1	75.8	67.1	59.6	81.0	44.1	38.0	62.5	54.64
	FCOS (2019)	✓	62.1	37.9	74.6	81.2	32.9	72.1	75.3	61.8	27.4	69.1	78.7	34.4	50.6	80.1	68.6	68.1	81.3	49.1	43.4	64.5	60.66
	Oriented R-CNN (2021)	✓	63.0	36.7	71.9	81.6	41.1	72.6	77.8	65.5	24.8	72.9	82.1	40.9	56.5	81.2	73.4	62.4	81.5	53.3	43.3	65.6	62.41
	GWD (RetinaNet) (2021b)	✓	61.5	23.6	73.6	81.1	17.4	72.7	68.3	47.2	20.7	71.2	73.2	33.9	34.3	77.6	64.7	57.5	80.9	42.1	39.7	60.2	55.07
	KLD (RetinaNet) (2021c)	✓	57.8	22.6	71.5	81.2	16.9	72.7	68.9	52.1	20.6	73.5	71.0	33.7	33.2	77.1	68.9	59.9	80.9	43.9	39.1	60.9	55.32
	KFIoU (RetinaNet) (2023b)	✓	60.6	36.6	73.6	80.9	27.0	72.6	73.4	56.5	25.4	73.9	72.0	32.9	45.8	75.8	65.2	57.6	80.0	48.0	40.1	58.8	57.84
$\mathcal{S}_H$	H2RBox (2023a)	✓	57.1	14.4	72.2	82.6	17.5	71.2	56.5	55.2	14.0	67.7	77.9	31.0	40.7	76.3	66.2	63.4	81.5	50.4	38.0	57.6	54.57
	H2RBox-v2 (2023)	✓	55.5	17.8	76.9	80.5	27.7	72.2	63.0	58.6	24.4	73.9	80.3	33.9	47.2	77.4	58.7	60.9	81.4	48.1	41.1	53.9	56.67
	ABBSPO (2025)	✓	69.5	15.7	76.2	87.5	29.9	72.3	75.3	61.2	28.1	74.1	81.7	34.7	48.2	79.3	67.4	61.4	81.5	54.7	41.5	53.8	59.70
$\mathcal{S}_P$	PointOBB (2024)	×	58.4	17.1	70.7	77.7	0.1	70.3	64.7	4.5	7.2	0.8	74.2	9.9	9.1	69.0	38.2	49.8	46.1	16.8	32.4	29.6	37.31
	Point2RBox+SK (2024)	✓	41.9	9.1	62.9	52.8	10.8	72.2	3.0	43.9	5.5	9.7	25.1	9.1	21.0	24.0	20.4	25.1	71.7	4.5	16.1	16.3	27.26
	PointOBB-v3 (2025)	✓	53.1	11.0	70.4	79.0	0.1	70.1	53.0	1.6	0.9	45.7	69.4	3.1	3.4	63.2	44.5	57.3	58.1	1.3	30.3	36.5	37.60
	PointOBB-v3 (2025)	×	37.1	13.0	70.2	79.4	0.1	71.5	44.7	50.8	4.5	52.9	74.7	2.5	0.5	69.6	46.1	58.0	58.9	3.0	31.8	34.4	40.18
	Point2RBox-v2 (2025a)	✓										–											34.70
	Point2RBox-v2 (2025a)	×										–											44.45
	Point2RBox-v3* (2026)	✓	66.7	10.8	71.8	81.0	14.2	64.3	36.0	42.1	12.8	10.3	57.9	3.5	32.2	76.7	49.3	59.2	80.9	20.2	27.9	16.5	41.70
	Point2RBox-v3* (2026)	×	70.3	12.7	70.4	81.2	16.3	71.2	49.1	48.6	20.0	19.3	62.9	9.4	35.0	79.8	59.5	68.0	81.1	28.1	29.1	20.2	46.60
	Point2RBox-v3† (2026)	✓	60.7	5.2	69.8	80.5	16.4	66.1	29.9	45.5	6.9	11.0	59.3	9.6	34.0	76.2	23.4	58.5	80.9	12.2	29.7	17.8	39.67
	Point2RBox-v3† (2026)	×	69.4	10.7	71.2	81.2	19.5	71.9	47.1	56.1	19.3	18.9	64.3	10.7	39.5	80.1	29.3	68.9	81.1	15.8	31.5	20.6	45.36
	InstaPoint (Ours)	✓	65.7	8.7	72.9	80.5	17.1	69.9	31.8	45.6	10.4	11.0	61.8	9.1	35.4	78.1	33.0	66.6	81.0	16.6	32.8	21.8	**42.49**
	InstaPoint (Ours)	×	70.5	11.8	76.0	81.3	20.7	72.1	48.5	58.6	17.3	17.8	66.2	5.5	40.7	80.4	45.6	69.5	81.2	22.2	33.1	23.4	**47.10**

that IDC+ISA yields higher mean target IoU than IDC alone on both datasets. We further reverse the ISA criterion by selecting the candidate with the larger regression loss instead of the smaller one. This reversed assignment reduces E2E/FCOS AP_{50} to 40.54/46.12 on DIOR-R and 59.55/64.92 on DOTA-v1.0, further supporting prediction–supervision consistency as a useful signal for assignment. These results are consistent with the analyses presented in Sec. 1: Fig. 1 provides a qualitative example of instance-wise source assignment, while Fig. 2 shows that ISA improves the mean IoU

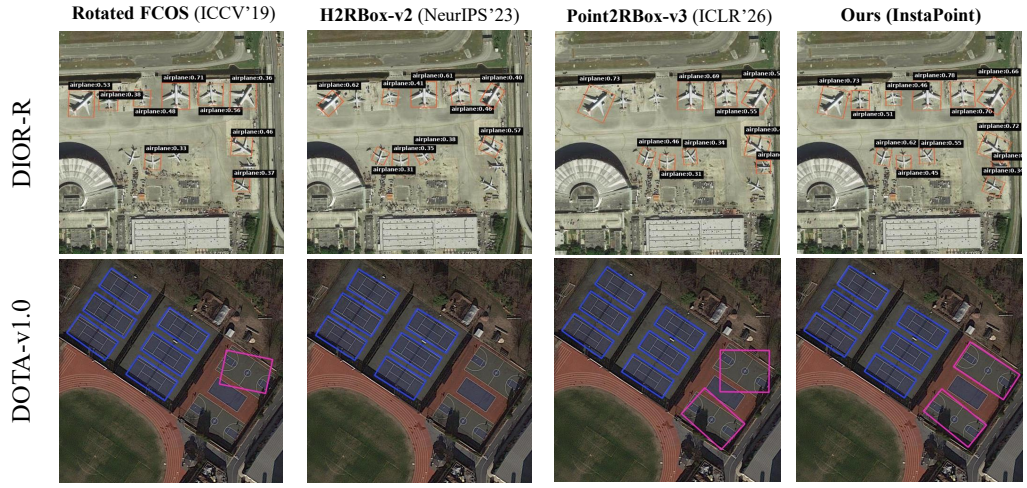


Figure 4: Visual comparison on DIOR-R and DOTA-v1.0. For point-supervised methods, results are obtained from the second-stage FCOS detector trained with pseudo RBoxes.

Table 3: Ablation studies of InstaPoint on DIOR-R and DOTA-v1.0. (a) Module ablation; (b) sensitivity to the assignment threshold ρ; and (c) sensitivity to the ISA switch epoch.

Modules		DIOR-R		DOTA-v1.0	
IDC	ISA	E2E	FCOS	E2E	FCOS
		39.67	45.36	59.49	65.11
✓		40.84	45.35	58.88	66.00
	✓	40.59	46.02	60.23	66.01
✓	✓	**42.49**	**47.10**	**62.53**	**67.74**

(a) Module Ablation

ρ	DIOR-R		DOTA-v1.0	
	E2E	FCOS	E2E	FCOS
0.5	41.66	46.32	59.57	65.28
1	**42.49**	**47.10**	**62.53**	**67.74**
2	42.41	46.84	60.95	67.12
3	41.97	46.99	60.02	66.27

(b) ρ Threshold

Epoch	DIOR-R		DOTA-v1.0	
	E2E	FCOS	E2E	FCOS
0	41.77	46.43	61.12	66.82
3	41.51	45.57	60.08	65.11
6	**42.49**	**47.10**	**62.53**	**67.74**
9	40.41	45.77	61.19	66.85
12	40.84	45.35	58.88	66.00

(c) ISA Switch Epoch

of the selected extent targets. Together with the detection gains from combining IDC and ISA in Table 3(a), these findings demonstrate the effectiveness of the proposed assignment strategy.

Table 3(b) studies the effect of the assignment threshold ρ. The default setting $\rho = 1$ yields the best performance across both datasets and evaluation protocols, consistent with the class-specific spacing boundary used by IDC to separate locally dense and sparse instances. Finally, Table 3(c) examines the transition from IDC to ISA. Switching at epoch 6 consistently achieves the best performance, whereas applying ISA from the beginning or delaying the transition leads to lower results. This trend supports our two-phase design: spatial cues provide useful prediction-independent supervision early in training, while prediction-guided assignment becomes more effective once detector predictions provide a sufficiently reliable basis for directly comparing candidate supervision sources.

5 Conclusion

We presented **InstaPoint**, an instance-adaptive framework for assigning complementary extent supervision in point-supervised oriented object detection. We observe that patch-wise source selection can be misaligned with the local spatial configuration of individual instances, motivating supervision assignment at the instance level. InstaPoint uses IDC to initialize assignments from local point spacing, followed by ISA for a one-time prediction-guided reassignment after warm-up.

Across DIOR-R and DOTA-v1.0, InstaPoint consistently improves end-to-end and two-stage pseudo-labeling performance, reaching 42.49/47.10 AP_{50} on DIOR-R and 62.53/67.74 AP_{50} on DOTA-v1.0. Compared with the strongest point-supervised baseline, these results correspond to gains of 0.79/0.50 AP_{50} on DIOR-R and 2.92/1.65 AP_{50} on DOTA-v1.0 under the end-to-end/two-stage protocols, respectively. Our code will be publicly released. These results highlight the importance of instance-level supervision assignment and its timing in point-supervised OOD.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring mandatory disclosure. Additionally, we used generative AI tools to aid and polish the writing of this paper, including improving grammar, clarity, and readability. All AI-assisted writing was reviewed and revised by the authors to ensure that it accurately reflects the intended technical content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Guangming Cao, Xuehui Yu, Wenwen Yu, Xumeng Han, Xue Yang, Guorong Li, Jianbin Jiao, and Zhenjun Han. P2rbox: Point prompt oriented object detection with sam. *arXiv preprint arXiv:2311.13128*, 2023.

Hui Chen, Tianbo Chen, and Haoyang Zhang. Pointood: Point-supervised remote sensing oriented object detection via canonical feature fusion and dynamic radius sampling. *Digital Signal Processing*, 168:105665, 2026. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2025.105665. URL https://www.sciencedirect.com/science/article/pii/S1051200425006876.

Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. doi: 10.1109/TGRS.2022.3183022.

Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2849–2858, 2019.

Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE transactions on geoscience and remote sensing*, 60:1–11, 2021a.

Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2786–2795, 2021b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Woojin Lee, Hyugjae Chang, Jaeho Moon, Jaehyup Lee, and Munchurl Kim. Abbspo: Adaptive bounding box scaling and symmetric prior based orientation prediction for detecting aerial image objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8848–8858, June 2025.

Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1829–1838, 2022.

Wentong Li, Yuqian Yuan, Song Wang, Jianke Zhu, Jianshu Li, Jian Liu, and Lei Zhang. Point2mask: Point-supervised panoptic segmentation via optimal transport. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 572–581. IEEE, 2023.

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

Mingxin Liu, Peiyuan Zhang, Yuan Liu, Wei Zhang, Yue Zhou, Ning Liao, Ziyang Gong, Junwei Luo, Zhirui Wang, Yi Yu, et al. Partial weakly-supervised oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27644–27654, 2026a.

Nanqing Liu, Xun Xu, Yongyi Su, Haojie Zhang, and Heng-Chao Li. Pointsam: Pointly-supervised segment anything model for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

Xinyuan Liu, Hang Xu, Zirui Chen, Yike Ma, Chenggang Yan, and Feng Dai. Semantic-decoupled spatial partition guided point-supervised oriented object detection. *Pattern Recognition*, 180: 114079, 2026b.

Xuran Lu and Zhisong Bie. Semantic segmentation everything model for point-prompted oriented object detection. *Electronics Letters*, 61(1):e70254, 2025.

Junwei Luo, Xue Yang, Yi Yu, Qingyun Li, Junchi Yan, and Yansheng Li. Pointobb: Learning oriented object detection via single point supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16730–16740, 2024.

Botao Ren, Xue Yang, Yi Yu, Junwei Luo, and Zhidong Deng. PointOBB-v2: Towards simpler, faster, and stronger single point supervised oriented object detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=R22JPTQYWV.

Yongqing Sun, Jie Ran, Feng Yang, Chenqiang Gao, Takayuki Kurozumi, Hideaki Kimata, and Ziqi Ye. Oriented object detection for remote sensing images based on weakly supervised learning. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2021.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5439–5448. IEEE, 2021.

Linfei Wang, Yibing Zhan, Xu Lin, Baosheng Yu, Liang Ding, Jianqing Zhu, and Dapeng Tao. Explicit and implicit box equivariance learning for weakly-supervised rotated object detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(1):509–521, 2025. doi: 10.1109/TETCI.2024.3398020.

Long Wen, Yu Cheng, Yi Fang, and Xinyu Li. A comprehensive survey of oriented object detection in remote sensing images. *Expert Systems with Applications*, 224:119960, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.119960. URL https://www.sciencedirect.com/science/article/pii/S0957417423004621.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3520–3529, 2021.

Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3163–3171, 2021a.

Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International conference on machine learning*, pp. 11830–11841. 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. *Advances in Neural Information Processing Systems*, 34:18381–18394, 2021c.

Xue Yang, Gefan Zhang, Wentong Li, Yue Zhou, Xuehui Wang, and Junchi Yan. H2RBox: Horizontal box annotation is all you need for oriented object detection. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=NPfDKT9OUJ3.

Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, XIAOPENG ZHANG, and Qi Tian. The KFIou loss for rotated object detection. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=qUKsCztWlKq.

Yi Yu, Xue Yang, Qingyun Li, Yue Zhou, Feipeng Da, and Junchi Yan. H2rbox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection. *Advances in Neural Information Processing Systems*, 36:59137–59150, 2023.

Yi Yu, Xue Yang, Qingyun Li, Feipeng Da, Jifeng Dai, Yu Qiao, and Junchi Yan. Point2rbox: Combine knowledge from synthetic visual patterns for end-to-end oriented object detection with single point supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16783–16793, 2024.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19283–19293, 2025a.

Yi Yu, Xue Yang, Yansheng Li, Zhenjun Han, Feipeng Da, and Junchi Yan. Wholly-wood: Wholly leveraging diversified-quality labels for weakly-supervised oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4438–4454, 2025b.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

Peiyuan Zhang, Junwei Luo, Xue Yang, Yi Yu, Qingyun Li, Yue Zhou, Xiaosong Jia, Xudong Lu, Jingdong Chen, Xiang Li, et al. Pointobb-v3: Expanding performance boundaries of single point-supervised oriented object detection. *International Journal of Computer Vision*, 133(9):6108–6128, 2025.

Shun Zhang, Jihui Long, Yaohui Xu, and Shaohui Mei. Pmho: Point-supervised oriented object detection based on segmentation-driven proposal generation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024.

Teng Zhang, Ziqian Fan, Mingxin Liu, Xin Zhang, Xudong Lu, Wentong Li, Yue Zhou, Yi Yu, Xiang Li, Junchi Yan, and Xue Yang. Point2RBox-v3: Self-bootstrapping from point annotations via integrated pseudo-label refinement and utilization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9vlS8PSGG7.

A Appendix

Table 4: Class-wise mask assignment changes between IDC and the final ISA selection. W and S denote Watershed and SAM, respectively. Transition percentages are computed over all instances in each class, and ΔS is the change in SAM assignment rate in percentage points.

DIOR-R

Class	N	W→S	S→W	ΔS
PL	1,888	14.14	36.97	-22.83
APO	662	0.00	80.51	-80.51
BF	2,384	2.56	38.34	-35.78
BC	1,077	10.86	39.18	-28.32
BR	1,367	8.85	41.62	-32.77
CH	649	16.95	32.51	-15.56
ESA	1,080	9.63	51.48	-41.85
ETS	610	2.46	23.77	-21.31
DAM	512	0.78	50.59	-49.80
GF	511	0.78	60.08	-59.30
GTF	1,162	1.46	28.40	-26.94
HA	2,364	1.86	60.32	-58.46
OP	1,330	5.79	53.08	-47.29
SH	27,351	31.03	16.48	+14.55
STA	595	0.00	59.33	-59.33
STO	3,042	11.41	25.61	-14.20
TC	4,898	15.15	34.75	-19.60
TS	501	0.40	53.09	-52.69
VE	13,724	26.19	18.16	+8.04
WM	2,365	26.22	21.10	+5.12
All	68,072	21.65	25.97	-4.32

DOTA-v1.0

Class	N	W→S	S→W	ΔS
PL	18,788	24.46	19.71	+4.75
BD	1,087	22.54	19.60	+2.94
BR	4,183	5.69	52.14	-46.45
GTF	721	0.28	21.22	-20.94
SV	58,819	22.87	16.27	+6.60
LV	43,075	39.44	9.78	+29.66
SH	76,153	24.68	14.15	+10.53
TC	5,923	21.88	31.72	-9.84
BC	1,180	17.37	47.71	-30.34
ST	13,670	20.49	18.07	+2.42
SBF	794	0.00	36.40	-36.40
RA	973	11.31	24.56	-13.26
HA	15,468	19.25	32.87	-13.62
SP	3,836	41.94	10.01	+31.93
HC	1,189	52.06	4.54	+47.52
All	245,859	26.00	16.99	+9.01

A.1 Reassignment Behavior of ISA

Table 4 analyzes how ISA revises the assignments initialized by IDC. On DIOR-R, 21.65% of instances change from watershed to SAM and 25.97% change in the opposite direction. On DOTA-v1.0, the corresponding proportions are 26.00% and 16.99%. These substantial bidirectional transitions indicate that ISA does not simply bias the assignments toward one supervision source, but instead revises the initialization differently across instances. The class-wise differences further suggest that the preferred supervision source depends on the characteristics of individual instances rather than following a uniform dataset-level tendency.

A.2 Class-wise Analysis of Supervision Assignment

Figure 5 provides a class-wise breakdown of the comparison in Fig. 2. The relative effectiveness of the assignment strategies varies across object categories. Overall, IDC improves the mask-to-GT box IoU for several categories, while ISA further improves the results for many classes by revising the initial spatially guided assignments.

A.3 Mismatch between Local Density and Patch-wise Assignment

Table 5 further illustrates the mismatch between local instance configurations and patch-wise assignment. On DOTA-v1.0, 88.01% of locally isolated instances occur in patches with at least five annotated objects, while the corresponding proportion is 60.08% on DIOR-R. Thus, an instance can be locally isolated even when it belongs to a globally crowded image or patch, showing why the patch-wise instance count does not always reflect the spatial configuration of each instance.

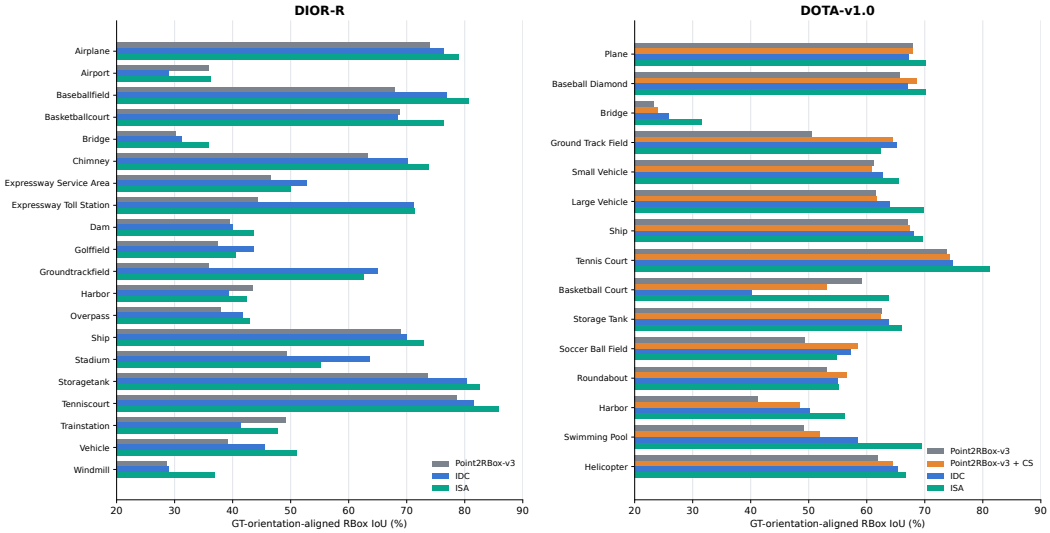


Figure 5: Class-wise mask-to-GT oriented box IoU on DIOR-R and DOTA-v1.0. For each instance, we fix the box center to its annotated point and the orientation to the GT angle, then derive its width and height from the selected mask. Each bar reports the mean IoU with GT boxes within a class. The GT angle is used only for this analysis, not for training supervision.

Table 5: Class-wise proportions of isolated instances ($s_i > 1$) that occur in images or split patches with at least five annotated objects. N_{iso} denotes the number of isolated instances, and $N_{J\geq 5}$ counts those with $J \geq 5$; the percentage is $100N_{J\geq 5}/N_{\text{iso}}$. DIOR-R uses trainval images and DOTA-v1.0 uses train+val split patches. These statistics describe scene composition rather than actual SAM selection.

DIOR-R

Class	N_{iso}	$N_{J\geq 5}$	(%)
All	**37,992**	**22,825**	**60.08**
PL	1,373	461	33.58
APO	662	45	6.80
BF	2,315	686	29.63
BC	691	196	28.36
BR	1,019	309	30.32
CH	482	89	18.46
ESA	912	447	49.01
ETS	587	129	21.98
DAM	504	9	1.79
GF	499	47	9.42
GTF	1,142	350	30.65
HA	2,259	2,194	97.12
OP	1,159	380	32.79
SH	8,625	7,625	88.41
STA	593	99	16.69
STO	2,120	1,335	62.97
TC	3,660	2,691	73.52
TS	499	64	12.83
VE	7,655	5,522	72.14
WM	1,236	147	11.89

DOTA-v1.0

Class	N_{iso}	$N_{J\geq 5}$	(%)
All	**85,981**	**75,676**	**88.01**
PL	8,353	5,901	70.65
BD	724	415	57.32
BR	3,254	1,136	34.91
GTF	719	291	40.47
SV	19,863	19,521	98.28
LV	9,926	9,611	96.83
SH	19,799	18,722	94.56
TC	3,664	3,490	95.25
BC	729	567	77.78
ST	5,868	5,499	93.71
SBF	794	463	58.31
RA	811	197	24.29
HA	9,669	8,376	86.63
SP	1,565	1,361	86.96
HC	243	126	51.85