PSTNet: Object Detection in Remote Sensing Images with Point Supervision and Object Templates

Peng Liu [a], Jun Miao [a,*], Yuanhua Qiao [b] and Baixian Zou [c,**]
[a] *School of Computer Science, Beijing Information Science and Technology University, Beijing 102206, China*
[b] *College of Applied Sciences, Beijing University of Technology, Beijing 100124, China*
[c] *College of Applied Arts and Science, Beijing Union University, Beijing 100191, China*
E-mails: jmiao@bistu.edu.cn, zoubx@buu.edu.cn

Abstract. Object detection in remote sensing images has become increasingly critical with the rapid development of fields such as unmanned aerial vehicles. However, while rotated bounding box annotations for such images are costly, the use of low-cost point annotations holds great promise. Nevertheless, the inability of point annotations to provide object size and orientation information poses a significant challenge for precise object localization in models. To address this, we propose PSTNet, a framework integrating Template Overlay Learning, Multi-scale Attention Dilated Block (MADB), and Dynamic K-value Sample Assignment. First, category-specific templates are randomly flipped, scaled, and overlaid as pseudo-labels to teach models size and orientation. Second, MADB replaces FPN with dilated convolutions and attention mechanisms to enhance multi-scale feature fusion for small objects. Third, a dynamic K-value strategy leverages classification scores to adaptively assign positive samples, bypassing IoU dependency. Extensive experiments on four datasets show substantial improvements to the baseline.

Keywords: Object detection, Remote sensing images, Point annotation, Template

1. Introduction

Remote sensing imagery, as a vital carrier of Earth's surface information, plays an irreplaceable role in environmental protection, urban planning and resource management [1]. However, object detection in remote sensing images faces significant challenges. First, the objects in these images are vast in quantity, diverse in categories, and widely distributed. These objects vary significantly in features such as size, shape, color, and texture, increasing the complexity of detection tasks. Second, the quality of remote sensing images is affected by factors like cloud occlusion, shadow interference, and varying capture angles, further complicating detection accuracy. Moreover, due to the nature of aerial imaging, objects in remote sensing scenes are often arbitrarily oriented. Traditional horizontal bounding box detection methods tend to encompass large background regions, reducing localization precision and increasing false positives.

Numerous outstanding research [2–5] achievements using rotated box supervision have effectively addressed the aforementioned challenges by designing novel network architectures, precise sample assignment strategies, and efficient loss functions, achieving remarkable detection accuracy. However, manually annotating fine-grained oriented bounding boxes is time-consuming and costly. According to Google AI Cloud Platform statistics, in the field of object detection, point annotation costs are approximately 36.5% lower than horizontal bounding boxes and 104.8% lower than oriented bounding boxes [6]. The efficiency of point annotation significantly surpasses that of

[*] Corresponding author. E-mail: jmiao@bistu.edu.cn.
[**] Corresponding author. E-mail: zoubx@buu.edu.cn.

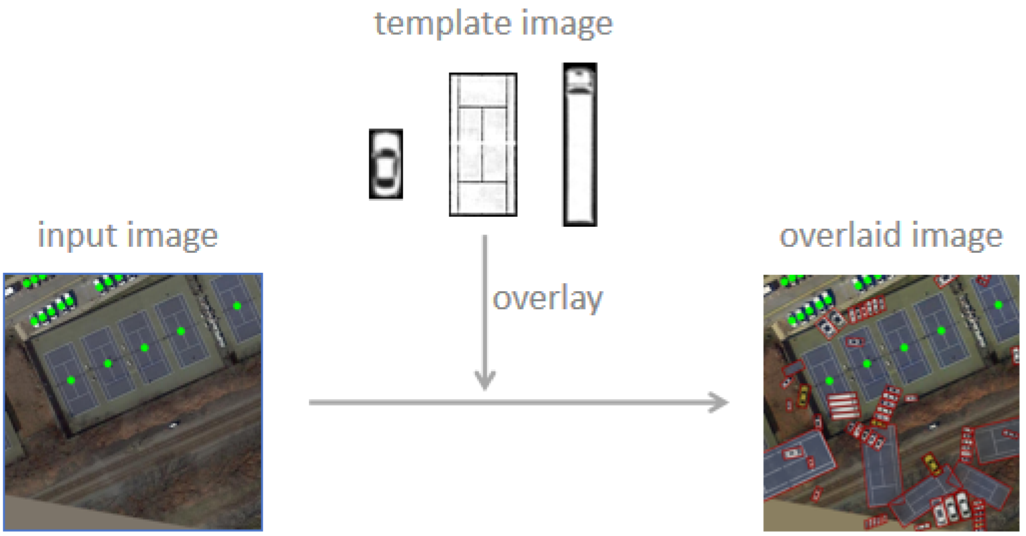


Fig. 1. Template Overlay method.

horizontal and oriented bounding boxes, making point-supervised object detection a more practical approach. Furthermore, most existing object detection datasets are annotated with horizontal bounding boxes. When predicting rotated bounding boxes, re-annotating entire datasets becomes necessary, leading to substantial waste of human resources. Thus, the adoption of point supervision is critically important to reduce annotation burdens and bridge the gap between annotation formats.

Current point-supervised rotated box prediction typically follows indirect or direct approaches. In indirect methods, P2BNet [7] predicts pseudo horizontal bounding boxes from point annotations and employs H2RBox [8] weakly supervised learning method to train rotated object detectors, while Point2Mask [9] generates rotated boxes by predicting object masks. Direct approaches include Point2Rbox [10], which directly predicts rotated boxes by integrating domain knowledge, and PointOBB [11], which designs a sophisticated positive-negative sample assignment strategy to generate category probability maps for rotated box prediction, though with lower accuracy. How to improve the accuracy of point-supervised direct rotated box prediction?

Point annotations cannot provide object size and orientation information, and effectively conveying this information to the model is the critical challenge in detection. Jun et al. proposed a Normal Template Mapping method [12], where a template image is selected for each category in the dataset to learn mapping relationships between original images and template images, provides a promising framework for addressing the challenge. Similarly, for remote sensing datasets, one template image per category can be chosen and randomly overlaid onto original images as pseudo-labels during training, as shown in Fig.1. However, due to the lack of size information, point annotations cannot fully leverage the advantages of FPN [13] for object detection, nor can adopt conventional IoU-based sample assignment strategies. Identifying suitable alternatives to these limitations is pivotal for enhancing detection performance.

In this paper, we introduce PSTNet, a point-supervised and template-integrated object detection network for remote sensing imagery. It achieves template-guided pseudo-label learning under point supervision while compensating for accuracy degradation caused by incompatibility with FPN-based multi-scale detection and IoU-based sample assignment strategies. Specifically, for each category in the remote sensing dataset, we extract a representative image patch as a "template", which is randomly flipped, scaled, and rotated before being overlaid onto the original image as pseudo-labels during training. These pseudo-labels enable the model to learn size and orientation information of objects, while category labels guide the learning of class-specific features. Second, we design a Mutil-scale Attention Dilated Block (MADB) to replace traditional FPN [13], which retains FPN's multi-scale capabilities while integrating dilated convolutions to expand receptive fields and channel-spatial attention mechanisms to enhance feature discriminability. Additionally, for sample assignment, we adopt classification confidence scores instead of IoU metrics and dynamically estimate the number of positive samples based on classification scores and number of candidate points.

In summary, the principal contributions of this paper can be articulated as follows: (1) We design PSTNet, a point-supervised remote sensing image detection network that leverages template-driven pseudo-label learning. (2) We introduce MADB to address the incompatibility between point supervision and FPN in object detection. (3) We propose a dynamic k-estimation sample assignment method guided by classification scores to address the limitations of IoU incompatibility in point-supervised detection. (4) Extensive experiments demonstrate that PSTNet can effectively enhance the accuracy of remote sensing image detection under point supervision.

2. Related Work

2.1. RBox-supervised Oriented Object Detection

Object detection using horizontal bounding boxes faces challenges such as overlapping boxes and high background ratios when detecting multi-oriented objects in remote sensing images. The rotated Region Proposal Network [14] effectively addresses these issues by introducing rotated bounding boxes with angles. These rotating anchors can adapt to objects with varying scales, aspect ratios, and rotation angles. RoI Transformer [15] reduces the number of rotated anchors by incorporating rotated regions of interest and corresponding spatial transformations, enabling better performance in rotation-aware object detection tasks. Oriented R-CNN [16] introduces an oriented proposal generation network to directly produce oriented candidate boxes. By adding two additional regression parameters, each anchor can predict an oriented proposal, overcoming the limitation of traditional RPNs in handling oriented objects. Furthermore, S2A-Net [4], SCRDet [17], and Cad-Net [18] have designed distinct network architectures to enhance feature representation capabilities for oriented object detection.

2.2. Point-supervised Oriented Object Detection

Combining point supervision for predicting horizontal boxes with horizontal box supervision for predicting rotated boxes enables the realization of point supervision for rotated box prediction. P2BNet [7] + H2RBox [8] approach first employs P2BNet to generate pseudo bounding boxes from point annotations, then utilizes these pseudo boxes alongside the weakly supervised learning method of H2RBox to train a oriented object detector. Additionally, several methods directly generate rotated boxes from point supervision. PointOBB [11] designs a positive-negative sample assignment strategy: for positive samples, all points within a fixed radius around each annotated point are selected, while for negative samples, a fixed-scale circle is drawn around each point, with points outside the circle designated as negative samples. This generates a category probability map, followed by Principal Component Analysis to precisely estimate object orientation and boundaries. SPA [19] treats user annotations as starting points for positive samples. It constructs graph elements using a similarity matrix derived from feature maps extracted by a deep learning network, then applies the max-flow algorithm to infer positive sample regions and generate rotated bounding boxes. PointTeacher [20] adopts a teacher-student architecture and decomposes the learning process into a two-stage denoising procedure. The teacher network progressively removes noise from pseudo bounding boxes generated by noisy point annotations to guide the student network.

2.3. Other Weakly-supervised Oriented Object Detection

To reduce annotation costs, H2RBox [8] explores predicting rotated boxes through horizontal box supervision. By designing weakly supervised and self-supervised branches, it learns consistency between the original image and its randomly rotated counterpart to predict object angles, thereby generating rotated boxes from horizontal box annotations. H2RBox-v2 [21] introduces a symmetry-aware learning mechanism that leverages the symmetry of input images to supervise the network. P2BNet [7] proposes an anchor similarity-based approach to construct balanced proposal bags across different objects, avoiding confusion among multiple objects. It also implements a coarse-to-fine pseudo box prediction and refinement process, including a coarse pseudo box prediction stage and a precise refinement stage, narrowing the performance gap between point-supervised and bounding box-supervised object detection. Point2Mask [6] frames pseudo-label generation as an optimal transport problem, where single-point annotations and pixel samples are defined as label suppliers and consumers, respectively. The transportation cost is calculated using task-oriented relationships that emphasize category differences and instance-level distinctions between objects and material objects.

3. Method

Overview. Point supervision can only provide the object's category and approximate location information. Inspired by template-induced learning, overlaying template images onto original images to regress template bounding boxes offers a solution to help models learn object size information. Since FPN with multi-scale detection capability cannot be applied due to the lack of size information in annotations, we design an MAD encoder incorporating multi-scale information, attention mechanisms, and dilated receptive fields to replace FPN, as depicted in Fig. 2. Additionally, we develop a dynamic K-value estimation sample assignment method that uses classification scores as evaluation metrics, effectively solving the problem of being unable to use IoU as an evaluation criterion due to the lack of object size information in annotations.

3.1. Template Overlay Learning

Given that point annotations cannot provide information about object size and orientation, we adopt a template overlay learning method where a category-specific template image is selected for each object class to be detected and randomly overlaid onto the original image, serving as manually annotated rotated bounding boxes for the model to learn size and orientation information, while the original point annotations help the model learn object category information. Specifically, we first crop a standard template image for each class in the dataset, as shown in Fig. 3, and overlay it onto the original image through random flipping, scaling, and rotation. Templates with IoU values exceeding 0.05 between two standard template images are removed to avoid excessive overlap affecting detection. Furthermore, to simulate the real-world scenario of small clustered objects in remote sensing images, we repeatedly stack partial template images (e.g., cars, ships) to create densely arranged clustered template image groups overlaid onto the original image.

The successful practice of H2RBox-v2 [21] demonstrates that self-supervision through vertical flipping and random rotation can effectively learn object rotation angles. To learn scale information from point annotations, we enhance the data augmentation for original images overlaid with randomly transformed template images by incorporating scaling transformations alongside flipping and rotation for scale perception. The augmented images undergo feature extraction, encoding, classification and regression branches through the same network used for the original overlaid images. Mutual supervision is achieved by minimizing the loss between predicted bounding boxes generated from the original overlaid image branch and the augmented image branch.

Since the training data combines real objects with point annotations and template objects with rotated bounding box annotations, their loss functions are computed separately. For real objects with point annotations, the loss primarily focuses on regressing object categories and centers, calculated using Eqs (1) and (2):

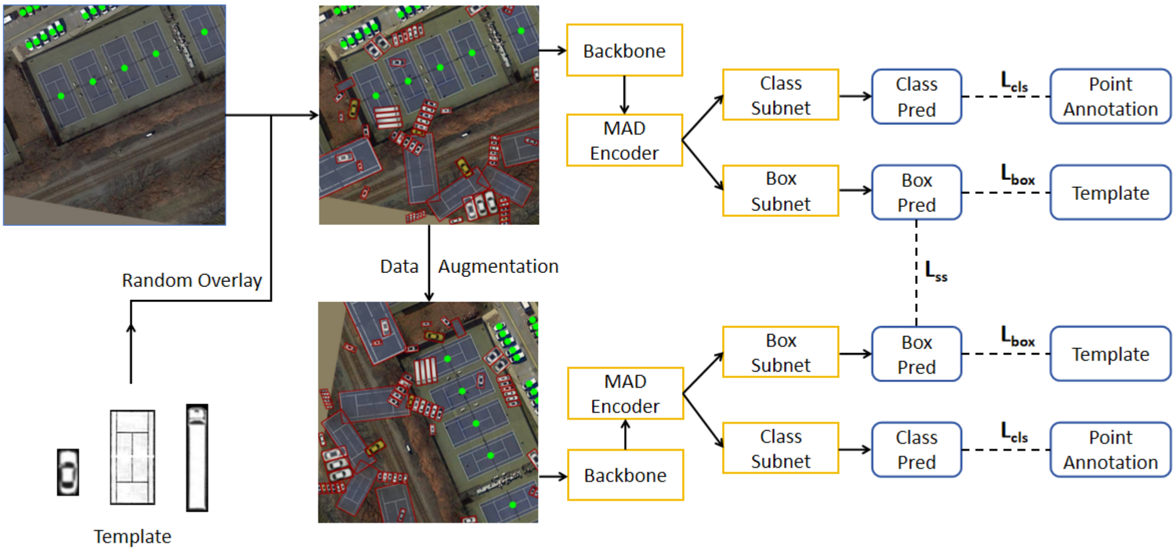


Fig. 2. PSTNet architecture.

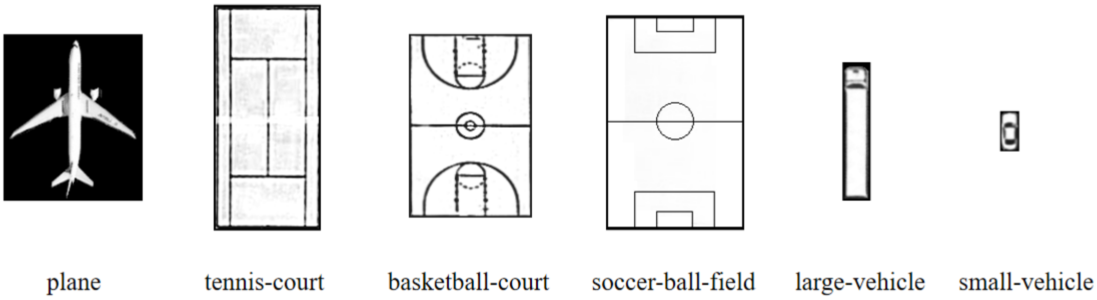


Fig. 3. Examples of template images for partial object categories.

$$L_{cls} = Focalloss(M_{point}c_{pred}, M_{point}c_{gt}) \tag{1}$$

$$L_{cen} = L_1(M_{point}xy_{pred}, M_{point}xy_{gt}) \tag{2}$$

where M_{point} is used to select real objects with point annotations. c_{pred} and c_{gt} represent the classification score of the predicted box and the category of the point annotated object, respectively. xy_{pred} and xy_{gt} denote the center of the predicted box and the center of the ground truth bounding box, respectively. For template objects annotated with rotated bounding boxes, the loss primarily focuses on regressing the rotated bounding boxes, calculated using Eq. (3):

$$L_{box} = CIoU(M_{box}B_{pred}, M_{box}B_{pred}) \tag{3}$$

where M_{box} is used to select template objects annotated with rotated bounding boxes. B_{pred} represents the predicted bounding box. Additionally, there is a self-supervised loss, calculated using Eq. (4):

$$L_{ss} = L_{flp/rot/sca}(M_{ori} M_{point} B_{pred}, M_{trs} M_{point} B_{pred}) \tag{4}$$

where M_{ori} is used to select objects from the original branch, and M_{trs} is used to select objects from the transformed branch. The loss functions for random flipping, random rotation, and scaling transformations are calculated using Eq. (5), Eq. (6) and Eq. (7), respectively:

$$L_{flp}(B_{ori}, B_{trs}) = smooth_{L1}(mod(\theta_{trs} + \theta_{ori}), 0) \tag{5}$$

$$L_{rot}(B_{ori}, B_{trs}) = smooth_{L1}(mod(\theta_{trs} - \theta_{ori}), R) \tag{6}$$

$$L_{sca}(B_{ori}, B_{trs}) = GIoU(r2h(B_{ori}) \times s, r2h(B_{trs})) \tag{7}$$

where B_{ori} represents the rotated bounding box output from the original branch, B_{trs} represents the rotated bounding box output from the augmented branch, θ_{ori} denotes the angle of the rotated bounding box from the original branch, θ_{trs} denotes the angle of the rotated bounding box from the transformed branch, mod() is the modulo function used to constrain angles within [-$\pi/2, \pi/2$], R matches the random rotation angle applied in data augmentation, r2h() is a function converting rotated boxes to horizontal boxes, and s corresponds to the random scaling factor used in data augmentation. The total loss is computed by summing the aforementioned losses, as defined in Eq. (8):

$$L_{total} = L_{cls} + L_{cen} + L_{box} + L_{ss} \tag{8}$$

3.2. Mutil-Scale Attention Dilated Block

As a sophisticated multi-scale feature fusion network architecture, FPN plays a pivotal role in object detection. It constructs a well-defined hierarchical feature pyramid through an intricate combination of top-down and bottom-up feature propagation and fusion mechanisms, significantly enhancing detection stability and accuracy for medium and large objects while markedly improving performance for small objects. However, since point annotations lack object size information, they cannot be assigned to appropriate FPN levels, leading to performance degradation in point-supervised scenarios. To address this, we design an MAD encoder composed of Multi-Scale Attention Dilated Blocks to replace FPN, aimed at improving small object detection accuracy.

To comprehensively capture and preserve multi-scale spatial information with precise structural details, we process the input features by grouping them into C//G×H×W and distributing them across four parallel branches: Branch 1 employs 3×3 dilated convolution to expand the receptive field; Branches 2-3 perform height-wise and width-wise average pooling, concatenate the results, refine them via a 1×1 convolutional layer, and generate attention weights through Sigmoid activation to weight features in Branch 4, followed by group normalization. Multi-scale integration is achieved by applying average pooling to both dilated convolution features and normalized coarse features, using Softmax activation to compute attention weights, and summing the weighted features. Fine-grained attention weights are then generated via Sigmoid and applied to grouped features. Global context is incorporated through global average pooling and a 1×1 convolutional layer, fused with fine-grained features to produce multi-scale, globally-aware attention-weighted features matching the input dimensions, as shown in Fig.4. This design enhances sensitivity to diverse object scales while integrating global context, improving feature richness and detection accuracy.

Additionally, considering the abundance of small objects in remote sensing images, we design the MAD Encoder by stacking the multi-scale attention dilated block four times, with the dilation rates of the dilated convolutions progressively set to 2, 4, 6, and 8. This design aims to preserve spatial information in the image more effectively while ensuring that features are captured at varying scales during each stacking iteration, enhancing the model's ability to detect densely distributed small objects across diverse resolutions.

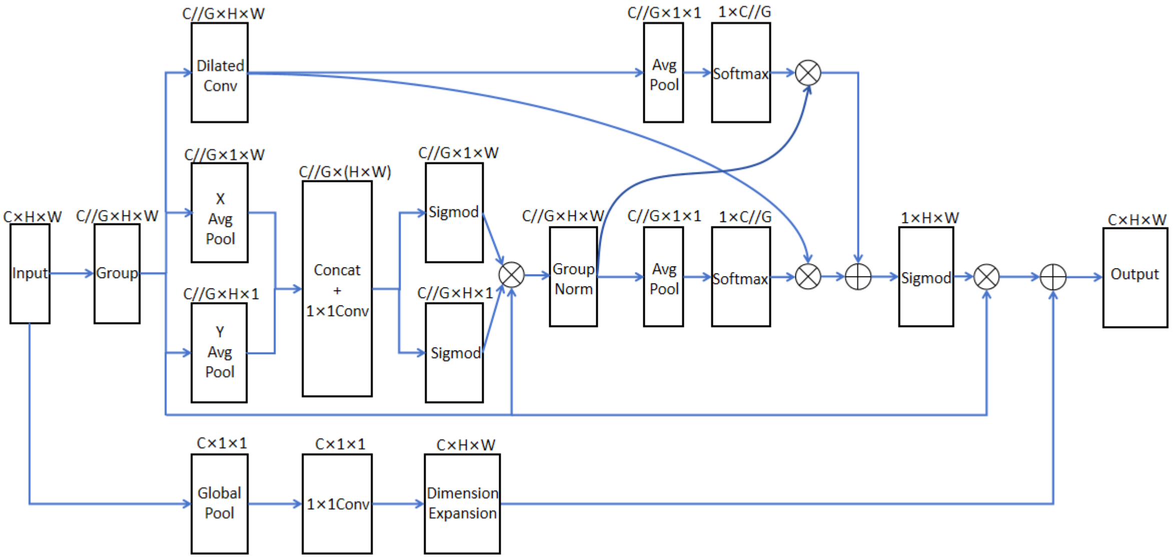
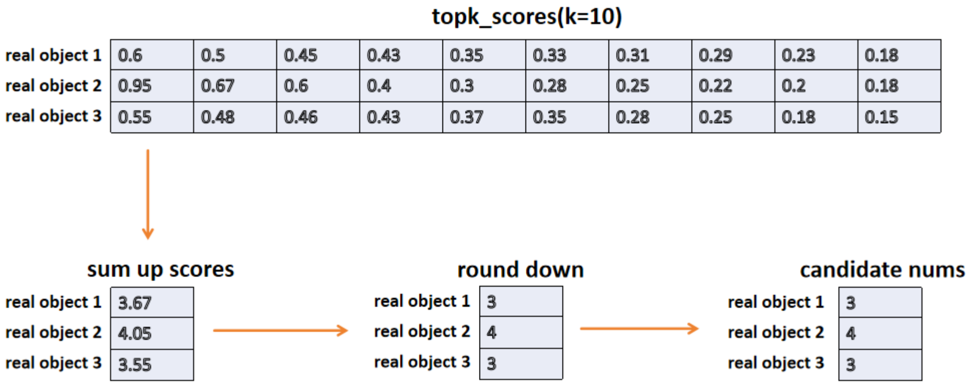


Fig. 4. Mutil-Scale Attention Dilated Block.


topk_scores(k=10)

real object 1	0.6	0.5	0.45	0.43	0.35	0.33	0.31	0.29	0.23	0.18
real object 2	0.95	0.67	0.6	0.4	0.3	0.28	0.25	0.22	0.2	0.18
real object 3	0.55	0.48	0.46	0.43	0.37	0.35	0.28	0.25	0.18	0.15

sum up scores

real object 1	3.67
real object 2	4.05
real object 3	3.55

round down

real object 1	3
real object 2	4
real object 3	3

candidate nums

real object 1	3
real object 2	4
real object 3	3

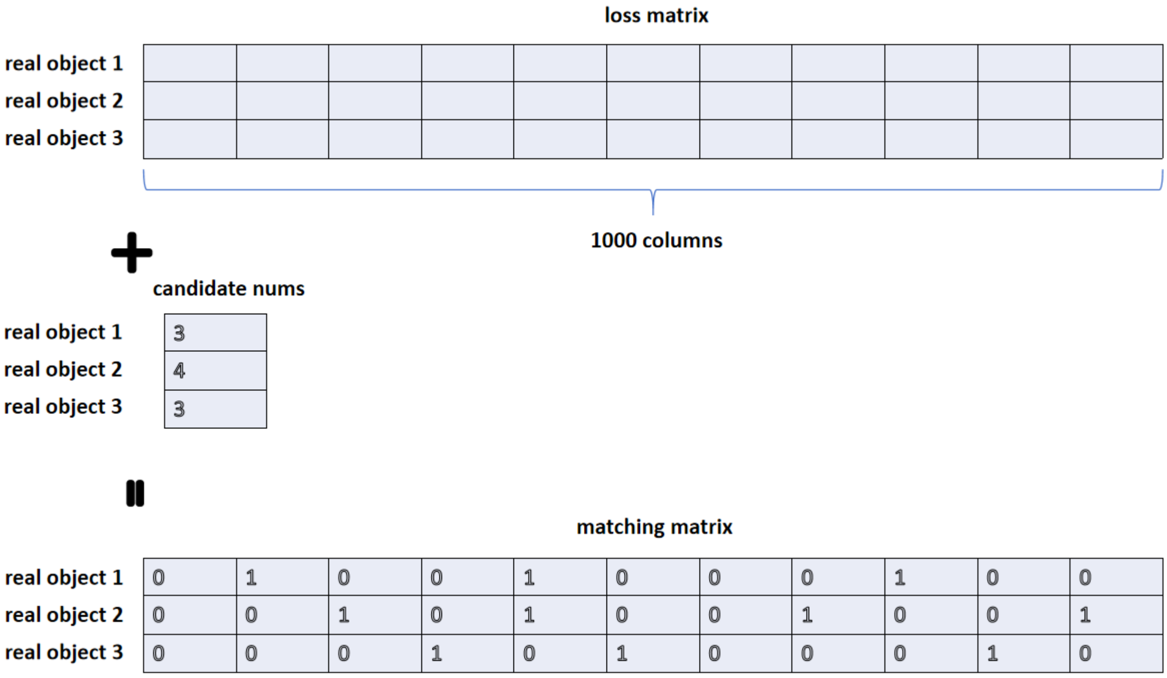
Fig. 5. Estimate the number of candidate points.

3.3. Dynamic K-value Estimation Sample Assignment

Existing object detectors heavily rely on multi-level assignment strategies in FPN and anchor assignment mechanisms, aimed at assigning objects of different sizes to appropriate hierarchical levels or anchors. However, a significant challenge arises since point annotations inherently lack size information, making it impossible to directly apply size-based sample assignment. To address this, we propose a Dynamic K-value Estimation Sample Assignment method that uses classification scores as the primary metric for sample selection, thus addressing the issue of missing size information and enabling effective assignment without relying on IoU-based criteria.

First, generate the loss matrix. Assuming an image contains 3 ground truth objects and the model predicts 1000 bounding boxes, a 3×1000 loss matrix is constructed, where each element corresponds to the classification loss. Additionally, to filter out predicted points that are too distant from the ground truth points, we select 25 candidate points within a circular region around each ground truth point for further processing.

Second, estimate the number of candidate points. Intuitively, since each ground truth object varies in size and occlusion conditions, the number of positive samples k assigned to each object should also differ. To dynamically estimate the k-value for each ground truth object, we first generate a score matrix by subtracting the loss matrix from



real object 1											
real object 2											
real object 3											

real object 1	3
real object 2	4
real object 3	3

real object 1	0	1	0	0	1	0	0	0	1	0	0
real object 2	0	0	1	0	1	0	0	1	0	0	1
real object 3	0	0	0	1	0	1	0	0	0	1	0

Fig. 6. Compute the matching matrix.

1. For each ground truth object, we select the top k candidate points with the highest scores from the score matrix. Subsequently, for each ground truth object, we calculate the sum of scores across its candidate points and apply a floor operation to refine the candidate selection, ensuring a more precise and adaptive allocation of positive samples, as shown in Fig.5.

Third, compute the matching matrix. For each ground truth object, select the candidate points with the lowest loss values from the loss matrix. Based on the dynamically estimated k-value determined earlier, assign the top k candidate points with minimal loss for each object. The matching matrix is then computed by setting positions corresponding to these selected candidate points to 1, while all others are set to 0. This binary matrix ensures precise alignment between predictions and ground truth objects based on adaptive k-value selection, as shown in Fig.6.

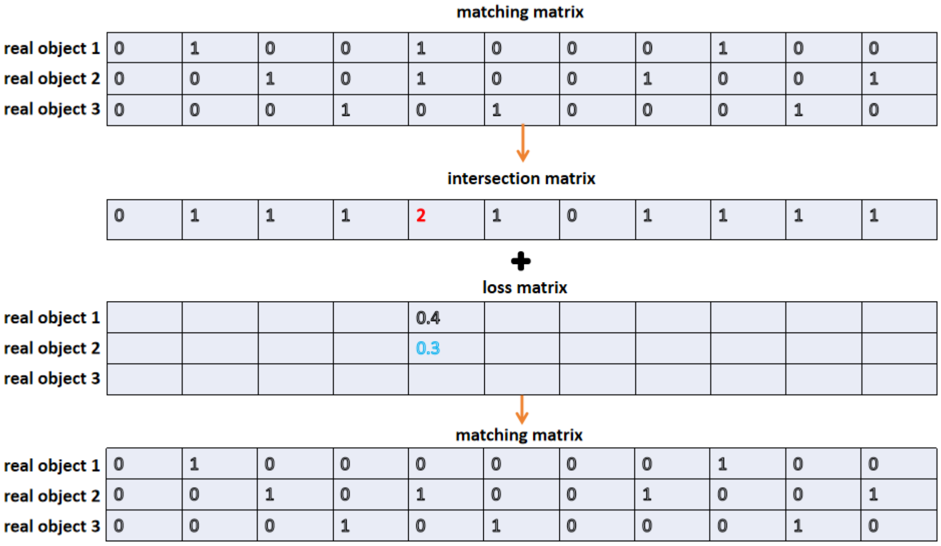
Finally, filter overlapping candidate points. Since the above steps may result in a predicted point being matched to multiple ground truth objects, overlapping candidates need to be resolved. Compute an intersection matrix by summing the columns of the matching matrix. If any column sums to 2 or more, compare the loss values of the conflicting ground truth objects for that predicted point, retain the match with the smallest loss, and filter out others, as shown in Fig.7. This process refines the matching matrix into the final matching matrix, ensuring each predicted point is uniquely assigned to the most suitable ground truth object.

4. Experiment

4.1. Datasets

Our experimental evaluation are done on DOTA [22] series datasets and HRSC dataset, both of which are comprehensive remote sensing image datasets, providing a robust benchmark for evaluating model performance.

DOTA-v1.0 [22], the first large-scale benchmark for aerial image analysis, is specifically designed for advancing object detection research. This dataset contains 2,806 carefully annotated images with 188,282 precisely localized instances across 15 distinct object categories. The multi-scale nature of the imagery, ranging from 800×800 to



matching matrix

real object 1	0	1	0	0	1	0	0	0	1	0	0
real object 2	0	0	1	0	1	0	0	1	0	0	1
real object 3	0	0	0	1	0	1	0	0	0	1	0

intersection matrix

0	1	1	1	2	1	0	1	1	1	1

+

loss matrix

real object 1					0.4						
real object 2					0.3						
real object 3											

matching matrix

real object 1	0	1	0	0	0	0	0	0	1	0	0
real object 2	0	0	1	0	1	0	0	1	0	0	1
real object 3	0	0	0	1	0	1	0	0	0	1	0

Fig. 7. Filter overlapping candidate points.

20,000×20,000 pixels, provides researchers with challenging scenarios to develop robust detection models capable of handling significant scale variations.

DOTA-v1.5 [23] extends its predecessor by introducing two critical enhancements: meticulous annotations for sub-10-pixel objects addressing small object detection challenges, and a novel "container crane" category, resulting in 403,318 precisely annotated instances while maintaining the original 2,806-image corpus from DOTA-v1.0. This strategic augmentation not only elevates the dataset's complexity but also provides a more rigorous evaluation platform for developing robust detection models capable of handling extreme scale variations and fine-grained object discrimination in aerial imagery analysis.

DOTA-v2.0 [23] advances aerial detection benchmarks by integrating Google Earth and GF-2 Satellite imagery, establishing an 18-category taxonomy with specialized classes (e.g., airports, helipads). The expanded dataset contains 11,268 images with 1,793,658 instances (emphasizing sub-50px objects), partitioned into training (1,830 images), validation (593 images), and testing sets to support rigorous evaluation of multi-scale detection algorithms in complex aerial scenarios.

HRSC [24] dataset, a scientifically curated benchmark for maritime object detection in remote sensing, incorporates 2,976 precisely annotated ship instances across 27 sub-categories (grouped into 4 major classes) extracted from six major international ports. Comprising 300×300 to 1500×900 pixel images (with 85% exceeding 1000×600 resolution) captured from Google Earth at 0.4-2m spatial resolution, this dataset features triple annotation modalities: horizontal bounding boxes, oriented bounding boxes, and pixel-level segmentation masks. Its hierarchical taxonomy and multi-perspective annotations establish HRSC as the preeminent benchmark for developing rotation-invariant detection models in complex harbor surveillance scenarios.

4.2. *Implementation Details*

We select the training set of DOTA-v1.0 for model training and its validation set for performance assessment, capitalizing on its diverse representation of remote sensing objects in our ablation study. For methodological benchmarking, we adopt the standardized evaluation protocol by training on the consolidated training-validation splits of HRSC and DOTA series datasets. Specifically, DOTA-v2.0 experiments strictly adhere to the official benchmark specifications with 12 training epochs, ensuring direct comparability with state-of-the-art detectors while maintaining consistency with the community-established evaluation paradigm for remote sensing object detection, HRSC dataset required an extended 72-epoch training regimen to achieve convergence. For a fair comparison, all results are evaluated without multi-scale technique.

Table 1
Effects of different data augmentation strategies

Model	random rotation	random flipping	random scaling	mAP
PSTNet	×	×	×	30.63
PSTNet	✓	×	×	35.94
PSTNet	×	✓	×	35.40
PSTNet	×	×	✓	33.79
PSTNet	✓	✓	×	40.12
PSTNet	✓	×	✓	38.57
PSTNet	×	✓	✓	37.86
PSTNet	✓	✓	✓	**41.20**

All experiments were conducted on a single NVIDIA RTX 2080 Ti GPU using PyTorch-based [25] frameworks, with detection models implemented through MMDetection [26] and rotation-sensitive components developed via MMRotate [27]. We employ ResNet-50 [28] backbone initialized with ImageNet [29] pre-trained weights and maintain methodological consistency by applying standard geometric augmentations such as random horizontal flipping and spatial shifting across all experiments, following established practices in remote sensing image analysis.

4.3. Ablation Study

To rigorously assess the methodological validity, we conduct ablation experiments on the DOTA-v1.0 benchmark, a gold-standard dataset for remote sensing object detection.

Effects of different data augmentation strategies. After overlaying the standard template image onto the original image, three data augmentation strategies—random rotation, random flipping, and random scaling—are applied to the overlaid image. In this section, ablation experiments are conducted to individually assess the impact of these three data augmentation strategies on detection performance. As shown in Table 1, the mAP of the model without any of these data augmentation strategies is 30.63%, which is lower than that of other experimental results. This indicates that employing appropriate data augmentation strategies is necessary for object detection in remote sensing images under point supervision. When only one of the data augmentation strategies—random rotation, random flipping, or random scaling—is applied, the mAP values are 35.94%, 35.40%, and 33.79%, respectively, showing improvements compared to the scenario without any data augmentation. When two of these strategies are employed, the mAP further increases. The highest mAP of 41.20% is achieved when all three data augmentation strategies are applied. This demonstrates that all three data augmentation strategies are essential for object detection in remote sensing images under point supervision.

Effects of MADB. To verify whether the MADB can compensate for the performance loss incurred by the removal of the FPN, ablation experiments are conducted in this section. As shown in Table 2, the mAP is 37.26% when the FPN is removed without any compensatory measures. However, after incorporating the MADB, the mAP increases to 41.20%, demonstrating the effectiveness of MADB in enhancing object detection models for remote sensing images that lack an FPN layer. Additionally, Point2RBox [10] introduces YOLOF-A1 [30], which is compatible with point annotations for comparison. YOLOF-A1 is based on YOLOF but uses only one anchor box size of 64. After removing the FPN and introducing YOLOF-A1, the mAP reaches 40.05%, showing an improvement compared to not introducing any module. Nevertheless, it is still 1.15 percentage points lower than when MADB is introduced. This indicates that MADB achieves similar multi-scale detection advantages as the FPN, effectively enhancing the model's detection performance.

Effects of parameter K. The selection of the K value is crucial in dynamic K-value estimation sample assignment, as it directly affects the number of positive samples, training stability, convergence speed, and model performance. Therefore, this section conducts multiple experiments to select the optimal K value. As shown in Table 3, when K varies from 6 to 14, we find that the mAP reaches its highest value of 41.20% when K equals 10.

Table 2

Effects of different data augmentation strategies

Model	MADB	YOLOF-A1	mAP
PSTNet	×	×	37.26
PSTNet	×	✓	40.05
PSTNet	✓	×	**41.20**

Table 3

Effects of positive samples parameters K

K	mAP
14	40.25
13	40.66
12	41.03
11	40.54
10	**41.20**
9	40.75
8	40.12
7	40.27
6	39.86

Table 4

Comparison with other methods on the DOTA-v1.0 dataset. We follow the official class abbreviations as the DOTA-v1.0 benchmark

Model	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Point2Mask-RBox [6]	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 [7]	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 [7]	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
PointOBB [11]	28.3	**70.7**	1.5	**64.9**	**68.8**	**46.8**	33.9	9.1	10.0	20.1	0.2	47.0	**29.7**	38.2	30.6	33.31
Point2RBox-RC [10]	**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 [10]	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
PSTNet	53.6	59.9	3.8	49.8	27.2	36.5	36.3	**88.3**	**52.4**	40.2	**9.2**	48.2	21.7	**53.6**	**37.6**	**41.20**

4.4. Comparison with other methods

Results on DOTA-v1.0. As systematically quantified in Table 4, PSTNet attains the mAP of 41.20% on the DOTA-v1.0 benchmark, surpassing existing point-supervised counterparts. A granular category-wise analysis reveals exceptional performance in geometrically regular objects: TC, BC, SBF, SP, and HC. While demonstrating competitive performance in most classes, detection accuracy for BR, SV and HA requires further optimization, potentially through enhanced feature aggregation for clustered small objects. Overall, PSTNet fundamentally advances point-supervised detection precision through an FPN-free multi-scale architecture and dynamic k-estimation sample assignment.

Results on DOTA-v1.5. Building upon the experiments conducted on the DOTA-v1.0 dataset, we selecte several models with relatively high mAP for further testing on the DOTA-v1.5 dataset. As shown in Table 5, despite the increased detection difficulty of the DOTA-v1.5 dataset compared to DOTA-v1.0, PSTNet achieves the mAP of 27.58% on the DOTA-v1.0 benchmark, which remains higher than other models' mAP results. This indicates that PSTNet demonstrates strong advantages in both detection accuracy and generalization capability.

Results on DOTA-v2.0. On the DOTA-v2.0 dataset, which contains a higher number of small objects, point-supervised remote sensing image object detection faces greater challenges. Similar to the experiments conducted on the DOTA-v1.5 dataset, we select several models with relatively high mAP for testing on the DOTA-v2.0 dataset. As shown in Table 6, PSTNet achieves the mAP of 22.34% on the DOTA-v2.0 dataset, outperforming other models in detection performance. Nevertheless, there remains room for further improvement.

Table 5

Comparison with other methods on the DOTA-v1.5 dataset

Model	mAP
PointOBB [11]	10.92
Point2RBox-RC [10]	24.31
Point2RBox-SK [10]	26.78
PSTNet	**27.58**

Table 6

Comparison with other methods on the DOTA-v2.0 dataset

Model	mAP
PointOBB [11]	6.29
Point2RBox-RC [10]	14.69
Point2RBox-SK [10]	21.77
PSTNet	**22.34**

Table 7

Comparison with other methods on the HRSC dataset

Model	mAP
RBox-supervised:	
RetinaNet [31]	84.49
GWD [32]	86.67
FCOS [33]	88.99
YOLOF-A5 [30]	**89.44**
YOLOF-A1 [30]	81.14
HBox-supervised:	
H2RBox [8]	7.03
KCR [34]	79.10
H2RBox-v2 [21]	**89.66**
Point-supervised:	
Point2Mask-RBox [6]	29.95
P2BNet+H2RBox-v2 [7]	14.60
Point2RBox-RC [10]	78.77
Point2RBox-SK [10]	79.40
PSTNet	**79.70**

Results on HRSC. On the HRSC dataset, which contains only a single ship category, detection is relatively easier compared to the DOTA series datasets. As shown in Table 7, in the field of point-supervised detection, PSTNet achieves the highest mAP of 79.70%, surpassing other point-supervised detection models. In contrast, the highest detection accuracies for rotated bounding box-supervised and horizontal bounding box-supervised models are 89.44% (YOLOF-A5 [30]) and 89.66% (H2RBox-v2 [21]), respectively. The point-supervised PSTNet is only about 10 percentage points lower, demonstrating that PSTNet achieves detection accuracy close to that of precisely annotated models while significantly reducing annotation costs.

5. Conclusion

This paper addresses the challenge of precise object detection in remote sensing images under point supervision, where the absence of size and orientation information severely limits localization accuracy. To overcome this, we propose PSTNet, which integrates Template Overlay Learning for pseudo-label generation, MADB for enhanced feature fusion, and Dynamic K-value Sample Assignment to bypass IoU dependency. Extensive experiments on DOTA-v1.0/v1.5/v2.0 and HRSC datasets demonstrate PSTNet's superiority, outperforming existing point-supervised methods by significant margins. Nonetheless, limitations persist: detection accuracy for certain categories (e.g., BR, SV, HA) remains suboptimal, likely due to insufficient feature aggregation for densely clustered small objects. Future work will explore incorporating shape priors or keypoint information into pseudo-label generation, refining multi-scale fusion strategies, and extending the framework to handle more complex remote sensing scenarios. These advancements aim to further bridge the gap between point supervision and fully supervised detection performance.

Acknowledgements

This research is partially sponsored by Natural Science Foundation of China (No. 61650201) and Beijing Natural Science Foundation (No. 4202025).

References

[1] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sensing*, vol. 14, no. 4, p. 871, 2022.

[2] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9657–9666.

[3] C. Xu, J. Ding, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7318–7328.

[4] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[5] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–11, 2021.

[6] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *European conference on computer vision*. Springer, 2016, pp. 549–565.

[7] P. Chen, X. Yu, X. Han, N. Hassan, K. Wang, J. Li, J. Zhao, H. Shi, Z. Han, and Q. Ye, "Point-to-box network for accurate object detection via single point supervision," in *European Conference on Computer Vision*. Springer, 2022, pp. 51–67.

[8] X. Yang, G. Zhang, W. Li, X. Wang, Y. Zhou, and J. Yan, "H2rbox: Horizontal box annotation is all you need for oriented object detection," *arXiv preprint arXiv:2210.06742*, 2022.

[9] W. Li, Y. Yuan, S. Wang, J. Zhu, J. Li, J. Liu, and L. Zhang, "Point2mask: Point-supervised panoptic segmentation via optimal transport," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 572–581.

[10] Y. Yu, X. Yang, Q. Li, F. Da, J. Dai, Y. Qiao, and J. 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*, 2024, pp. 16 783–16 793.

[11] J. Luo, X. Yang, Y. Yu, Q. Li, J. Yan, and Y. Li, "Pointobb: Learning oriented object detection via single point supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 730–16 740.

[12] J. Miao, P. Liu, C. Chen, and Y. Qiao, "Normal template mapping: An association-inspired handwritten character recognition model," *Cognitive Computation*, vol. 16, no. 3, pp. 1103–1112, 2024.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[14] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE transactions on multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[15] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2849–2858.

[16] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3520–3529.

[17] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, S. Xian, and K. S. Fu, "Towards more robust detection for small, cluttered and rotated objects. arxiv 2018," *arXiv preprint arXiv:1811.07126*.

[18] G. Zhang, S. Lu, and W. Zhang, "Cad-net: A context-aware detection network for objects in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 10 015–10 024, 2019.

[19] W. Zhao, Z. Fang, J. Cao, and Z. Ju, "Spa: Annotating small object with a single point in remote sensing images," *Remote Sensing*, vol. 16, no. 14, p. 2515, 2024.

[20] H. Zhu, C. Xu, R. Zhang, F. Xu, W. Yang, H. Zhang, and G.-S. Xia, “Tiny object detection with single point supervision,” *arXiv preprint arXiv:2412.05837*, 2024.

[21] Y. Yu, X. Yang, Q. Li, Y. Zhou, F. Da, and J. Yan, “H2rbox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 59 137–59 150, 2023.

[22] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[23] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo *et al.*, “Object detection in aerial images: A large-scale benchmark and challenges,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7778–7796, 2021.

[24] Z. Liu, L. Yuan, L. Weng, and Y. Yang, “A high resolution optical satellite image dataset for ship recognition and some new baselines,” in *International conference on pattern recognition applications and methods*, vol. 2. SciTePress, 2017, pp. 324–331.

[25] A. Paszke, “Pytorch: An imperative style, high-performance deep learning library,” *arXiv preprint arXiv:1912.01703*, 2019.

[26] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, “Mmdetection: Open mmlab detection toolbox and benchmark,” *arXiv preprint arXiv:1906.07155*, 2019.

[27] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu *et al.*, “Mmrotate: A rotated object detection benchmark using pytorch,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7331–7334.

[28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[30] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, “You only look one-level feature,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 039–13 048.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[32] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, “Rethinking rotated object detection with gaussian wasserstein distance loss,” in *International conference on machine learning*. PMLR, 2021, pp. 11 830–11 841.

[33] Z. Tian, C. Shen, H. Chen, and T. He, “Fcos: Fully convolutional one-stage object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[34] T. Zhu, B. Ferenczi, P. Purkait, T. Drummond, H. Rezatofighi, and A. Van Den Hengel, “Knowledge combination to learn rotated detection without rotated annotation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 518–15 527.