SARFA-Net: Shape-Aware Label Assignment and Refined Feature Alignment for Arbitrary-Oriented Object Detection in Remote Sensing Images

Yan Dong, Minghong Wei, Guangshuai Gao, Chunlei Li, and Zhoufeng Liu

***Abstract*—Arbitrary-oriented object detection in remote sensing images has witnessed significant progress in recent years. Numerous excellent detection models perform promising results, however, there are two main tough challenges hinder their performances. On the one hand, current label assignment strategies suffer from an imbalance between positive and negative samples, particularly for large aspect ratio and small-scale objects, leading to the Insufficient High-quality Samples. On the other hand, fixed convolution kernels and coarse sampling positions are not well suited for adapting to rotating objects in complex remote sensing scenes, resulting in Feature Misalignment. To alleviate the above issues, in this article, a novel SARFA-Net is proposed, incorporating a Shape-Aware Label Assignment (SALA) strategy and Refined Feature Alignment module (RFAM). Specifically, SALA is proposed to mitigate the problem of insufficient sampling for extremely shaped objects, the core of which is the Shape-Aware Sampling module, to meticulously select more high-quality positive samples within elliptical regions. To further enhance SALA at extremely limited scales and large aspect ratios, a Threshold Compensation Module is designed, which further utilizes the shape characteristics of the objects. Furthermore, RFAM is developed to adaptively align features by adjusting the sampling positions of the convolution kernels based on the refined anchors. Extensive experiments conducted on five large-scale datasets, DIOR-R, DOTA-v1.0, HRSC2016, FAIR1M-v1.0, and UCAS-AOD achieved mAPs of 68.90%, 80.09%, 90.40%, 46.34%, and 90.01%, respectively, demonstrating the effectiveness of the proposed approach and the superiority compared with state-of-the-arts. Compared with the baseline S^2A-Net, we have improved by 1.30, 1.57, 0.23, 5.92, and 0.37 points, respectively, without additional data augmentation.**

Received 11 September 2024; revised 18 November 2024 and 4 January 2025; accepted 16 January 2025. Date of publication 20 January 2025; date of current version 7 April 2025. This work was supported in part by the NSFC under Grant 62301623, Grant 62472463, and Grant 61873293, in part by the Leading Talents of Science and Technology in the Central Plain of China under Grant 234200510009, in part by Henan Key Research and Development Projects under Grant 241111220700, in part by Henan Province Key Science and Technology Research Projects under Grant 232102211002 and Grant 232102211030, in part by Key Scientific Research Project of Colleges and Universities in Henan Province under Grant 25A620001, and in part by the Incubation Program for Young Master Supervisor of Zhongyuan University of Technology under Grant SD202416. *(Corresponding author: Guangshuai Gao.)*

Yan Dong is with the School of Information and Communication Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China, and also with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China (e-mail: dy@zut.edu.cn).

Minghong Wei, Guangshuai Gao, Chunlei Li, and Zhoufeng Liu are with the School of Information and Communication Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China (e-mail: 2022006189@zut.edu.cn; 6911@zut.edu.cn; lichunlei1979@zut.edu.cn; lzhoufeng62@163.com).

The code will be available at https://github.com/H1d30nbu3h/SARFA-Nethttps://github.com/H1d30nbu3h/SARFA-Net.

Digital Object Identifier 10.1109/JSTARS.2025.3532039

***Index Terms*—Arbitrary-oriented object detection (AOOD), feature misalignment (FM), insufficient high-quality samples, label assignment, remote sensing images (RSI).**

I. Introduction

ARBITRARY-ORIENTED object detection (AOOD) in remote sensing images (RSIs) uses oriented bounding boxes (OBBs) to locate and identify objects of interest. In comparison to horizontal bounding boxes (HBBs), OBBs offer more precise object boundaries and retain directional information, which is widely used in both civilian and military domains, such as urban management, agricultural monitoring, topographic mapping, and military investigation [1], [2], [3], [4], [5], [6], [7], [8], [9].

Numerous excellent related studies [10], [11], [12] have made remarkable advancements and attracted wide attention from researchers. However, AOOD in RSIs still remains two critical challenges that hinder the performances.

A. Insufficient High-Quality Samples (IHS)

Label assignment, mapping candidate samples to their corresponding objects, is fundamental to construct a sample space that accurately reflects the shapes and orientations of interested objects. Generating high-quality positive samples during this process is crucial for training robust object detection models. However, remote sensing images (RSIs) often contain oriented objects with complex distribution patterns, characterized by large aspect ratios and significant scale changes. The preset anchors struggle to effectively match these irregularly distributed objects, leading to the *IHS* issue.

Classic detectors [11], [12], [13], [14] often rely on fixed label assignment strategy (e.g., MaxIoU) as the matching metric, as illustrated in Fig. 1(a). This strategy mainly finds the appropriate matching relationship between anchors and ground truth (GTs) based on a fixed IoU threshold (e.g., 0.5), which has become the most mainstream choice due to its simplicity and intuitiveness. However, it treats all positive samples indiscriminately without considering the shape information of objects.

Instead, dynamic label assignment strategy uses an anchor quality score threshold to dynamically divide positive and negative samples, which takes the shape information of each object into account. Some strategies directly use IoU to evaluate the quality of anchors, Ming et al. [15] proposed adaptively

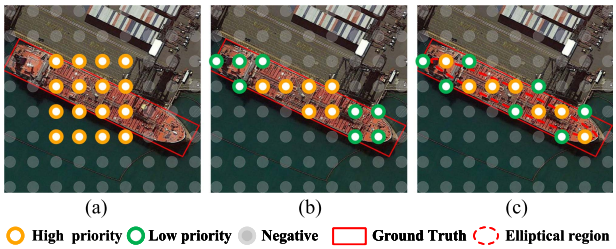


Fig. 1. Different strategies for the priority and selection region of positive samples. Only center points of anchors for simplification. (a) MaxIoU uses a fixed IoU threshold for sampling. (b) ATSS dynamically adjusts the threshold based on the center distance. (c) SALA strategy uses a dynamic threshold and prioritizes selecting high-quality samples from elliptical regions.

weight anchors based on the prior and predicted IoU. Zhang et al. [16] proposed Adaptive Training Sample Selection (ATSS) to measure dynamic thresholds through statistical characteristics, as shown in Fig. 1(b). Inspired by ATSS, some follow-up approaches [17], [18], [19] aim to enhance the quality of positive samples that focus on the central region of the object. These methods set different priorities according to the center point distance between the anchor and the GTs, effectively improving the selection quality of positive samples. The above methods are more effective than fixed ones, however, they are susceptible to parameter sensitivity.

GWD-based methods [20], [21] introduced a 2-D Gaussian distribution and select positive samples by adjusting the priority of anchors (e.g., evaluating the similarity between distributions), which can adapt to oriented objects with various scales and angles. However, its complex network structure and complex nature lead to increased resource space occupation. Therefore, this strategy is not considered for use in our framework.

According to the above analysis, motivated by the adaptive, elegant, and efficient dynamic label assignment strategy, a novel Shape-Aware Label Assignment (SALA) strategy is proposed, which integrates the merits of both IoU and center point metrics to select high-quality positive samples, as shown in Fig. 1(c).

Specifically, regarding objects with large aspect ratios, relying on the center point constraint in the label assignment strategy may cause foreground samples to converge in the circular area at the center of GT, leading to the anchors close to the object's boundary being treated as negative samples. As shown in Fig. 2(a) (left side), the gray anchor boxes (center points to simplify) at both ends are considered as negative samples. To alleviate this problem, a Shape-Aware Sampling (SAS) strategy is proposed to sample within an elliptical region, which is beneficial for edge feature extraction of objects with large aspect ratios. As shown in Fig. 2(a) (right side), the corresponding blue anchors are considered positive samples.

In addition, for small-sized objects, even slight pixel deviations can lead to a significant drop in IoU [22], [23], making it challenging to match positive samples. As depicted in Fig. 2(b), the center point of *anchor box 1* falls outside the designated central area and gets excluded, while *anchor box 2* doesn't meet the threshold criteria and is discarded. However, both anchors have the potential to detect the object. To address this issue, a Threshold Compensation Module (TCM) is further proposed

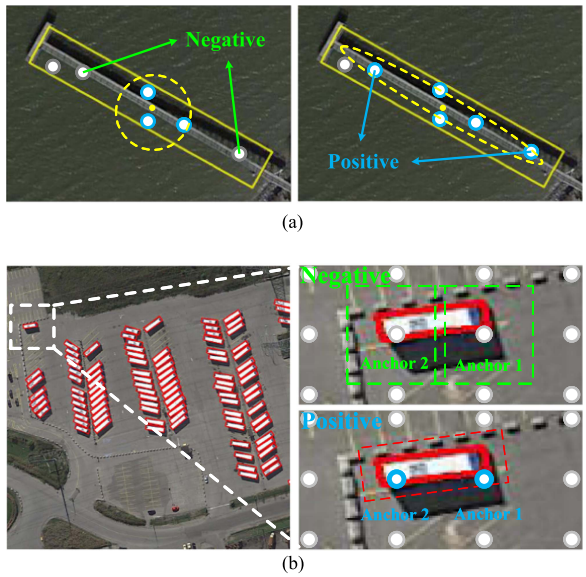


Fig. 2. Illustration of the sampling strategy for extreme scales and aspect ratio objects. Only center points of anchors for simplification. (a) Different regions for sampling. (b) Insufficient small-scale sampling.

to rescue potentially valuable samples previously discarded by existing methods (e.g., [19], [24]) by expanding sampling ranges and adaptively adjusting thresholds. As illustrated in Fig. 2(b), TCM reclassifies two previously negative samples (green dashed boxes) as positive samples (blue anchors) within the sampling range (red dashed range).

B. Feature Misalignment (FM)

A significant spatial misalignment often exists between the feature sampling points of standard convolution kernels and remote sensing objects, leading to incomplete feature coverage and the inclusion of noise information. This phenomenon is referred as the *FM* issue in this article, which primarily stems from two factors: First, the fixed receptive field of standard 2-D convolution kernels conflicts with the varying shapes of objects in remote sensing images. Axis-aligned convolutional features struggle to align with arbitrarily oriented bounding boxes, especially given the wide range of object scales and aspect ratios. Second, the effective receptive field during convolution is often smaller than the theoretical one [25], further exacerbating the misalignment problem.

Some two-stage detectors [26], [27], [28] use the RoI operator to extract fixed-length features within horizontal RoIs, which can only approximately represent the oriented object and introduce additional noise information. RSIs often contain densely packed and oriented objects, leading to horizontal RoIs that frequently encompass multiple objects. A common approach is to set numerous anchor boxes with varying angles, scales, and aspect ratios, which however can be computationally expensive and memory-intensive.

Several studies [29], [30], [31] employ Deformable Convolutional Network (DCN) [26] to learn an offset for improving the spatial accuracy of sampling locations. This offset, however, is learned from the anchor box by convolution without any rotation

processes, leading to instability and potential issues with sampling from background, particularly for oriented objects with complex distributions in RSIs. In contrast, recent works [32], [33], [34] leverage AlignConv [10] to infer the offset from the anchors, which directly offer a more stable initial offset and introduce directional constraints on the sampling points, addressing the limitations of DCN in the context of RSIs. However, AlignConv still has limitations in aligning densely packed and large aspect ratio objects. Based on this insight, we propose a Refined Feature Alignment Module (RFAM) that adaptively adjusts sampling positions based on refined anchors, enabling the extraction of aligned and intact features. Besides, a group of additional sampling points is further appended to expand the refined anchors to mitigate the effects of complex background noise, which can avoid the inappropriate sampling caused by the coarsely regressed refined anchors.

In summary, a novel Shape-Aware Label Assignment and Refined Feature Alignment Network (SARFA-Net) is proposed incorporating SALA strategy and RFAM. Specifically, the proposed SALA strategy dynamically selects more high-quality positive samples from multilevel feature maps, of which the critical component is SAS to avoid insufficient and low-quality sampling. In addition, TCM is proposed to further deal with small objects. Moreover, to relieve the feature misalignment, RFAM is proposed to extract more robust and effective features to adapt the varied objects' shapes.

The contributions of this work are summarized as follows.

1) A simple yet effective SALA strategy is proposed to dynamically select high-quality positive samples on multilevel feature maps, of which the SAS module is proposed to select more high-quality samples within the ellipse region while TCM is designed for threshold compensation of objects with extreme scales.
2) RFAM is proposed to extract aligned features through adaptive sampling points and alleviate the impact of background noise.
3) We report 68.90%, 80.09%, 90.40%, and 46.34% mAPs on five challenging AOOD remote sensing datasets, DIOR-R [35], DOTA-v1.0 [36], HRSC2016 [37], and FAIR1M-v1.0 [38], respectively, which demonstrate the effectiveness of our method and the superiority compared with state-of-the-arts.

The rest of this article is organized as follows. The related work is reviewed in Section II. Section III provides a detailed introduction to the proposed model. In Section IV, experiments and discussions are provided. Finally, Section V concludes this article.

II. Related Work

A. Arbitrary-Oriented Object Detection in Remote Sensing Images

AOOD has developed rapidly due to its wide range of application scenarios. Early methods [39], [40], [41] relied on densely preset anchors with varying scales, aspect ratios, and angles to achieve better regression, coming at the cost of massive parameters.

To address the above challenges, Ding et al. [12] proposed the RoI Transformer to transform horizontal RoI into oriented RoI. Han et al. [10] and Yang et al. [11] generated high-quality OBBs by simply presetting single-scale HBBs through their respective feature alignment modules.

For more accurate OBB predictions, some methods adopt additional OBB expressions. Xu et al. [42] designed a sliding vertex representation method for OBBs, which directly predicts the four-tuple set of each OBB. Guo et al. [43] proposed a convex hull representation method to optimize predicted box regression. Yang et al. proposed GWD [44] and KLD [45] based on the distance between the predicted OBB and the Gaussian distribution generated by GT. In addition, Yang et al. [46] designed CSL to predict oriented objects through angle classification.

Although significant advancements have been made, these methods still suffer from insufficient positive samples and feature misalignment, which hinder substantial detection performance.

B. Label Assignment in Object Detection

Label assignment is a crucial part of a detector, which involves identifying positive and negative samples for training. Nevertheless, traditional label assignment methods relying on fixed metrics (e.g., MaxIoU) encounter difficulties when handling objects with large aspect ratios and scale variations. To alleviate the above problems, Flex-MCFNet [47] designed a flexible-mixup (FlexMix) data augmentation strategy that increases the label's weight with the input image's proportion. However, data augmentation has limited effect on improving sample imbalance. Ming et al. [15] proposed an adaptive weight method based on prior IoU and predicted IoU, aiming to achieve a better matching degree. FSDet [33] utilized a soft assignment mechanism to assign weights to training samples for stable optimization. ATSS [16] dynamically adjusted the IoU threshold based on object statistics. In addition, some methods [17], [19] proposed the center-based sampling strategy to improve the quality of positive samples. Sun et al. [17] explored how sample distribution influences sample assignment. Guan et al. [19] proposed an elliptical distribution-aided label assignment strategy to select positive samples among all feature levels dynamically. While more efficient of the dynamic strategy, they still suffered from the problem of parameter sensitivity.

Hence, some methods [21], [44], [48], [49] defined transformation distances based on GWD-based methods. GGHL [49] fitted each instance with a single 2-D Gaussian heatmap and dynamically reweighted the samples. DCFL [21] proposed a dynamic prior coarse-to-fine assigner for dynamic label assignment, utilizing coarse prior matching and fine posterior constraints. While Gaussian heatmap-based methods [21], [49] effectively captured shape and directional features of arbitrarily oriented objects, their complex representations hindered accurate distance measurement between square objects.

Although the above strategies alleviate the issue of insufficient samples, they require setting parameters beforehand, using complex functions, and introducing low-quality samples. Therefore, SALA strategy is proposed to ensure sufficient positive samples

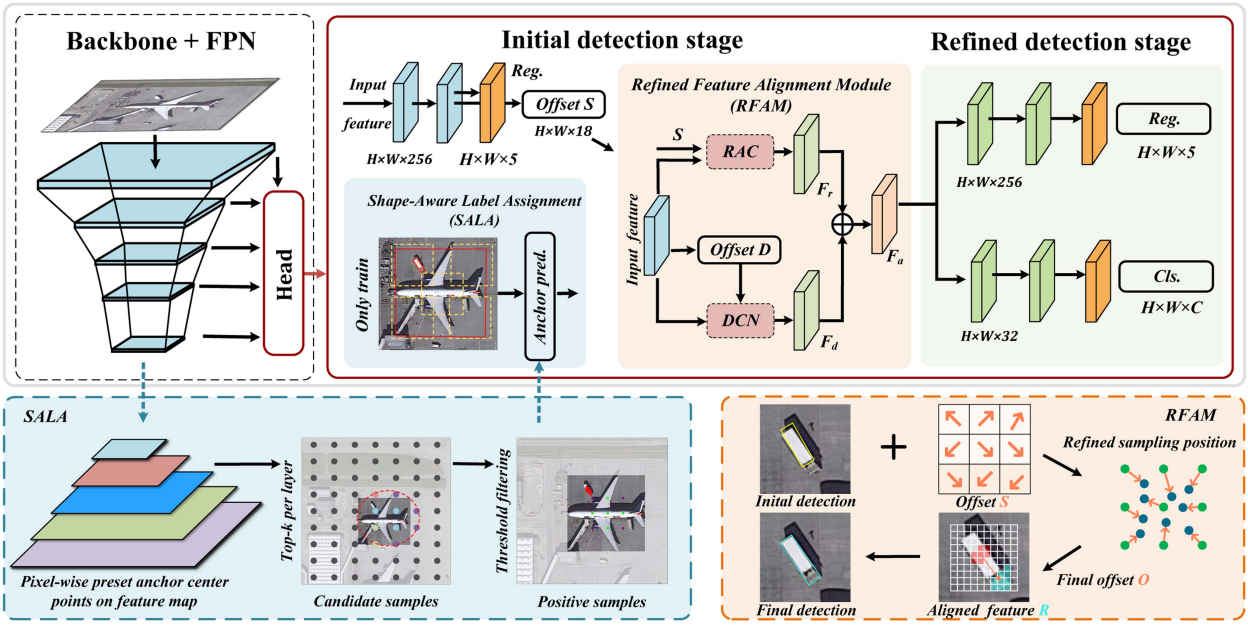


Fig. 3. Pipeline of the proposed method. The upper block represents the architecture of SARFA-Net proposed based on baseline S^{2}A-Net-C [10], where the initial detection stage ignores the classification ($cls.$) branch for simplification. The lower two blocks represent Shape-Aware Label Assignment (SALA) strategy and Refined Feature Alignment module (RFAM), respectively. The SALA strategy is implemented in the initial detection stage to ensure sufficient high-quality positive samples for objects with various shapes and arbitrary orientations. RFAM extracts aligned features by dynamically adjusting the sampling positions of anchors.

for extreme aspect ratios and scale objects while mitigating the impact of low-quality samples on the model.

C. Feature Misalignment

Feature misalignment usually refers to the misalignment between matched RoIs/anchors and convolutional features, which seriously affects the detection accuracy of detectors.

Two-stage detectors typically employed regions of interest (RoI) operations to extract fixed-length feature representations within RoIs. RoIPooling [27] rounded the floating-number boundaries of each RoI to the nearest integer, leading to misalignment issues between features and RoIs. To address the quantization issue of pooling operations, RoIAlign [28] used bilinear interpolation to calculate the precise sampling position in each sub-region (cut from each RoI), which showed significant positioning advantages. Deformable RoIPooling [26] further enhanced feature extraction by introducing learnable offsets. However, these RoI-based operations are computationally complex due to extensive region-wise calculations, e.g., feature warping and interpolation. RoI Transformer [12] addressed this issue by transforming horizontal RoIs into rotated RoIs, thereby reducing the need for numerous anchors. However, it still relies on precisely defined anchors and complex ROI operations.

To alleviate the above problems, Zhang et al. [30] introduced deformable convolutions [26], learning offsets to correct feature misalignment. Chen et al. [31] extended this approach by computing offsets based on refined anchors. For handling oriented objects, Han et al. [10] proposed alignment convolutions to mitigate the misalignment between features and oriented objects. The methods mentioned above present simpler and more lightweight network architecture, however, the alignment effect remains limited in complex remote sensing scenes.

In this work, we propose RFAM that leverages the geometric information of oriented bounding boxes to adaptively align features and flexibly handle contextual information.

III. Proposed Method

This section provides a detailed introduction to the proposed network SARFA-Net with SALA strategy and RFAM. We equip our proposed modules with the baseline method, S^{2}A-Net-C [10], as illustrated in the pipeline of Fig. 3. The network architecture of SARFA-Net comprises a backbone network (i.e., ResNet50), a feature pyramid network (FPN), an initial detection stage, and a refined detection stage. During the initial detection stage, SAS and TCM constitute the main components of SALA, which are designed to strategically select sufficient high-quality samples. In Algorithm 1, we provide a detailed introduction to the sampling process included in the SALA strategy. Afterward, RFAM is designed to reduce the impact of feature misalignment by obtaining adaptively refined sampling points during training.

A. Shape-Aware Label Assignment Strategy

As shown in Fig. 3, given a prior anchor set $A = \mathbb{R}^{H\times W\times C}$ ($H \times W$ is the feature map size, C is the number of shape information) on each layer $\mathcal{L}$ of the feature map of FPN. For simplicity, each feature point has one prior anchor. SALA strategy is proposed to find a proper match between the anchor set A and the ground truth set G, and assign pos/neg to supervise network learning. In Algorithm 1, we show the sampling process of the SALA strategy in detail. SAS is the core component of

Algorithm 1: Shape-Aware Label Assignment Strategy.

Input:
The set of ground truth bboxes for current batch, $\mathcal{G}$;
The set of preset anchor boxes for current batch, $\mathcal{A}$;
The set of each level in the pyramid layers, $\mathcal{L}$;
The initial sampling number, k;
The sampling number, $top\text{-}k$;
Output:
The set of positive samples $\mathcal{P}$ and negative samples $\mathcal{N}$;

Compute the center points of the anchor box, $Points$;
Compute the set of flags for the elliptical region with $\mathcal{G}$:
$F = \text{CheckPointsInEllipse}(\mathcal{G}, Points)$ (1);
for each level $l \in \mathcal{L}$ **do**
  Build an empty set for candidate samples: $\mathcal{C} \leftarrow \varnothing$
  $S_i \leftarrow$ Select k anchors from $\mathcal{A}$ whose center are closest to the center of $\mathcal{G}_i$ when $F_i = \text{True}$;
  if $k < top\text{-}k$ **then**
    $E_i \leftarrow$ Select $top\text{-}k$ anchors from $\mathcal{A}$ whose center are closest to the center of $\mathcal{G}_i$ when $F_i = \text{False}$;
    $S_i = E_i \cup S_i$
  end if
  $C_i = C_i \cup S_i$
end for
Compute threshold for each ground truth:
$\mathcal{T} = \text{ComputThreshold}(C_i, g_i)$ (4);
for each candidate $c \in C_i$ **do**
  if $IoU(c, \mathcal{G}) \geq \mathcal{T}_i$ **then**
    $\mathcal{P} = \mathcal{P} \cup c$
  end if
end for
$\mathcal{N} = \mathcal{A} - \mathcal{P}$
return $\mathcal{P}, \mathcal{N}$;

SALA and is responsible for selecting the sampling range of positive samples. We first obtain the indicator function $Flag$ and select the center point of $\mathcal{A}_i$ as far as possible within the ellipse range represented by $\mathcal{G}_j$ ($Flag = true$) as the candidate sample. Specifically, for the prior anchors of each layer $\mathcal{L}$, we select the top-k anchors with the closest center distance for sampling. If the initial actual sampling number k does not meet $top\text{-}k$, it is supplemented in the box of $\mathcal{G}_j$. The TCM module further selects positive samples from candidate samples through a dynamic threshold designed by statistical information, and finally divides the candidate positive samples that meet the threshold $\mathcal{T}_i$ into positive samples P and the rest into negative samples N.

B. Shape-Aware Sampling (SAS) Module

Most center-based label assignment strategies select samples within the bounding box or its central region. However, these methods overlook the distinctive characteristics of remote sensing objects, such as large aspect ratios and arbitrary directions, leading to insufficient or low-quality sampling. To address this, we propose the SAS module, which incorporates shape information of oriented objects, including aspect ratios and angles.

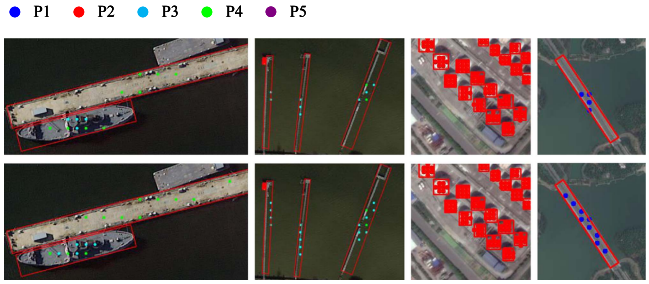


Fig. 4. Illustration of ATSS strategy (top row) and SALA strategy (bottom row) for selecting positive samples from each level of the feature pyramid (p1-p5). Only the center points of anchors are visualized for simplification.

Specifically, we select candidate samples within the constructed dynamic ellipse region.

Following the baseline (S^2A-Net), we adopt the long side definition to represent an oriented object using five parameters (x, y, w, h, θ). These parameters represent the center point coordinates, width, height, and angle, respectively. Hence, given a GT $g_i(x_i, y_i, w_i, h_i, \theta_i)$ and one anchor box $a_j(x_j, y_j, w_j, h_j, \theta_j)$ it matched, the elliptical region can be formulated as

$$F(\cdot) = \begin{cases} \text{true}, & \frac{a^2}{(0.5w_i)^2} + \frac{b^2}{(0.5h_i)^2} < \eta \\ \text{false}, & \text{otherwise} \end{cases} \tag{1}$$

where the ratio factor η acts as an adaptive threshold based on the object's shape. It controls the range of the elliptical distribution, ensuring the selection of high-quality samples with less background noise. Here, η is calculated as follows:

$$\eta = 1 - 0.5/r \tag{2}$$

where $r \in [1, +\infty)$ is the ratio of the long side $max(h_i, w_i)$ and short side $min(h_i, w_i)$ of the object. We will provide a detailed explanation of η in Section IV. Parameters a and b are calculated by the offset of the center point coordinates between g_i and a_j and the angle of g_i, which can be respectively formulated as

$$\begin{aligned} a &= x_j \cos\theta_i + y_j \sin\theta_i \\ b &= x_j \sin\theta_i - y_j \cos\theta_i. \end{aligned} \tag{3}$$

In this way, the sampling distribution can be adjusted dynamically according to the shapes of objects. As shown in Fig. 4, the sampling range tends to be a compact circular distribution when the shape of GT is close to a square. When the GTs are with extremely large aspect ratios, the sampling range will approximate that of an inner tangent ellipse, which is better suited for such objects' shapes.

C. Threshold Compensation Module (TCM)

As shown in Fig. 2, sampling within the inner region leads to insufficient samples for small objects due to the limited scales. To this end, we further develop a novel TCM to dynamically adjust the positive sample threshold and sampling region for small-scale objects during the sample assignment stage.

Specifically, regarding the issue of *anchor box 2* mentioned in Fig. 2(b), we set a monotonical decreasing function as a

weighting factor to obtain an adaptive IoU threshold function $\mathcal{T}_i$. For the given ground truth g_i, $\mathcal{T}_i$ can be defined as

$$\mathcal{T}_i = f(scale;\varphi)\cdot IOU_i^{\text{init}} \tag{4}$$

where φ defaults to 5, indicating a pivotal point where sample count variations become abrupt. In order to adapt to different datasets, $scale$ is obtained by applying a logarithmic function, which is defined as follows:

$$scale = ln(h\cdot w) \tag{5}$$

$$IOU_i^{\text{init}} = \mu + \sigma \tag{6}$$

where μ and σ represent the mean and standard deviation of the IoU between the ground truth g_i and candidate samples it matches, which are defined as

$$\mu = \frac{1}{N}\sum\nolimits_{j=1}^{N} I_{i,j},\ \sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(I_{i,j}-\mu)} \tag{7}$$

where N is the number of candidate samples, which is set to 9 by default and has been proven to be robust in previous work [16]. $I_{i,j}$ is the IoU value between the ith ground-truth box and the jth predict box it matches.

Studies conducted by [22] and [23] demonstrate that the IoU metric is highly sensitive to variations in oriented object aspect ratios and scales. Even minor changes in these attributes can lead to rapid IoU decreases, especially for small and slender oriented objects. Hence, a scale-based weighting function is designed as follows:

$$f(scale;\varphi) = \begin{cases} S_0\cdot e^{\frac{-|scale-\varphi|}{\gamma}}, & scale<\varphi \\ 1, & \text{otherwise.} \end{cases} \tag{8}$$

S_0 is a compensation factor that ensures sufficient samples for small-scale objects by maintaining the stability of the weighted threshold, which is defined as

$$S_0 = e^{-\frac{6}{\gamma}} \tag{9}$$

where γ represents the balancing factor utilized for normalizing the scale size, which is set to 15 based on the maximum $scale$ value. In the experimental Section IV, objects with a scale of 6, typical in remote sensing optical images, will serve as the basis for analysis.

Regarding the issue of example *anchor box 1* mentioned in Fig. 2(b), we compensate for objects with extreme scales (objects with $scale$ less than γ). Specifically, we identify the above cases in the candidate positive sample selection stage and then enlarge the sampling ranges of these extreme-scale objects by 1.1 times.

D. Refined Feature Alignment Module (RFAM)

To achieve high-performance object detection, anchors must be spatially and scale-aligned with the features. Previous work [50] has pointed out that discriminative localization features have arbitrary positions and complex deformations, making it difficult to accurately capture them using standard convolutions, especially for slender and small remote sensing objects. Han et al. [10] proposed a feature alignment module (FAM), which uses spatial feature reconstruction strategy to solve the problem of spatial misalignment, so that the features collected by each refined anchor box are aligned with the object in space. Inspired by this design, we propose RFAM, which introduces the DCN [26] to spatially align anchors and their extracted features.

Specifically, to extract features of the object, we regularly sample $n = k\times k$ feature points in each refined anchor box. For each sampling position p, the output of DCN can be defined as follows:

$$Y(p) = \sum_{i\in\mathcal{I};o\in\mathcal{O}}^{n} W(i)\cdot X(p+i+o) \tag{10}$$

where $\mathcal{I} = \{(i_x, i_y)\}$ is a set of regular grids used in standard convolution, and $\mathcal{O}$ is the offset field of DCN. However, DCN simply changes the sampling points by the offset field $\mathcal{O}$ without considering the direction, which will lead to incorrect sampling.

In this section, RFAM is proposed to generate richer and more distinctive refined alignment features by combining contextual information, which consists of two processes, *refined sampling position generation* and *refined aligned feature generation*. As shown in Fig. 5, the core component of RFAM, Refined Aligned Convolution (RAC), dynamically adjusts the receptive field, allowing for extracting more comprehensive object features, which is denoted as F_r. In addition, a separate DCN branch generates features F_d, which together with F_r ultimately generate robust aligned features denoted as F_a.

1) Refined Sampling Position Generation: The sampling process overview of RAC is depicted in Fig. 5. To refine a given initial prediction, we start by adjusting the sampling position based on the initial prediction box [Fig.] and the corresponding learnable offset field S. Next, as depicted in Fig. 5(b), the refined anchor box is enlarged by α times to obtain the refined sampling position (the four corner points, the center points of the four sides, and the center point of the refined anchor box). The final offset field O_n for each position of DCN is calculated in the refined feature sampling process [Fig.]. Finally, O_n is fed to the DCN to extract the refined alignment feature F_r. Hence, we have

$$\overrightarrow{PP_n} + \overrightarrow{P_nP_n^*} = \overrightarrow{P\widehat{A}_c} + \overrightarrow{\widehat{A}_cP_n^*} \tag{11}$$

$$I_n + O_n^* = \widehat{A}_{\Delta xy} + \widehat{I}_n = \widehat{A}_{\Delta xy} + \widehat{A}_{wh}\cdot I_n\cdot\beta \tag{12}$$

where the hyperparameter β controls the sampling position within each region. Specifically, we evenly partition the region centered around $\widehat{A}_c$ into $k\times k$ grids and sample features from the center of each grid cell. Consequently, β is set to $1/k$. $\widehat{A}_{\Delta xy}$ and $\widehat{A}_{wh}$ are offsets between center points and shapes of the refined anchor after encoding, respectively, providing spatial information to guide feature reconstruction and alignment. The refined sampling positions P_n^* are derived from the initial sampling positions P_n.

Unlike the precalculated offset O_n^* obtained from FAM [10], we generate a set of learnable offsets S_n based on the predicted anchor boxes. These learnable offsets are then combined to generate offset S_n^*, ultimately obtaining offset O_n. Hence, we have

$$\overrightarrow{P_nR_n} = \overrightarrow{P_nP_n^*} + \overrightarrow{P_n^*R_n} \tag{13}$$

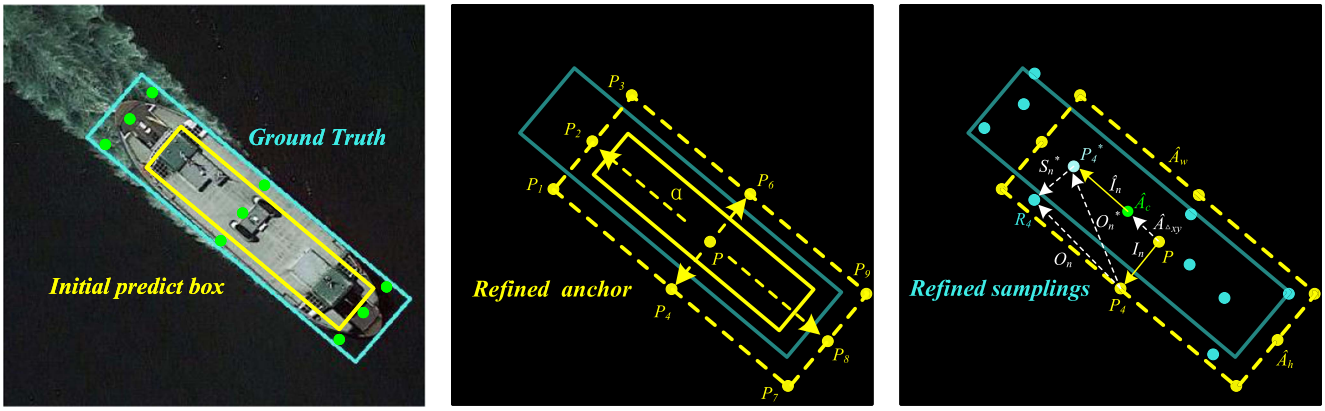


Fig. 5. Illustration of the sampling points of the RAC. The yellow box and yellow dashed box are the initial prediction box and the refined anchor box magnified by α times, respectively, while the blue box is the ground truth. Green and yellow points indicate FAM and initial sampling positions, respectively. Dark blue points indicate refined sampling positions extracted by RAC. Better view in color. (a) Initial predict box. (b) Refined anchor box. (c) RAC.

$$O_n = O_n^* + S_n^* \tag{14}$$

where R_n is the final refined sampling position.

2) Refined Aligned Feature Generation: For each initial prediction bounding box, RFAM constructs a set of learnable refined sampling positions to learn aligned features. This process can be formulated as follows:

$$P = \{p_i\}_{i=1}^{n}, n = 9 \tag{15}$$

where n is the number of feature sampling positions, which is set to 9 by default in our work. Refined feature sampling positions can be obtained based on the initial sampling positions and the corresponding learnable offset field S, which is calculated as follows:

$$p_i^r = (x_i^{init} + \widehat{A}_w * \Delta x_i, y_i^{init} + \widehat{A}_h * \Delta y_i) \tag{16}$$

where x_i^{init} and y_i^{init} are the initial sampling positions. Fig. 5(a) shows the detail of initial sampling position generation. $\widehat{A}_w$ and $\widehat{A}_h$ are the width and height of the refined anchor box, respectively, which are multiplied by the offset to normalize the scale difference. As shown in the network structure of Fig. 3, an additional branch in the initial detection stage is added to generate the learnable offset field $(\Delta x_i, \Delta y_i)$, which can be expressed as follows:

$$S = \delta(conv_1(conv_0(F))) \tag{17}$$

where F denotes the input feature of initial detection stage. $conv_1$ and $conv_0$ are two consecutive standard convolution operations, which is used to obtain S. δ represents the activation function. $S \in \mathbb{R}^{H\times W\times 18}$ represents the offset coordinates of the nine sampling positions. Therefore, for each initial position p, the offset field O_n can be calculated by the refined sampling position

$$O_n = \{RM \cdot p_i^r - p - i\}_{p_i \in S, i \in \mathcal{I}} \tag{18}$$

$$RM = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{19}$$

where RM is a rotation matrix to transform the refined feature sampling position p_i^r into the offset field of DCN. The refined features after RAC resampling can be expressed as

$$F_r = DCN(F, O) \tag{20}$$

where F_r represents the refined features generated by the RAC module. F_d represents the output feature of another DCN module, which is used to enhance the robustness of the sampling features. This feature can be formulated as follows:

$$F_d = DCN(F, D) \tag{21}$$

$$F_a = conv(F_r + F_d) \tag{22}$$

where $D = \mathbb{R}^{H\times W\times 18}$ is the offset learned from the input feature F. The refined aligned feature $F_a = \mathbb{R}^{H\times W\times 256}$ is fed into the refined detection stage, which can be obtained by fusing F_r and F_d.

Compared to the standard convolution, this flexible sampling kernel better adapts to oriented objects with various shapes, such as ship, bridge, and harbor, which often pose challenges for traditional square convolutions due to difficulties in capturing complete and aligned features.

E. Loss Functions

In this article, long side definition (dle) is adopted to represent an oriented bounding box through five parameters (x, y, w, h, θ). The loss of the proposed method is defined as follows:

$$L = L_{reg}^i + L_{cls}^i + L_{reg}^r + L_{cls}^r. \tag{23}$$

L_*^i and L_*^r represent the losses in the initial stage and the refined stage, respectively. The classification loss adopts Focal Loss [51], the regression loss is formulated as

$$L_{reg}^* = \frac{1}{N}\sum_{n=1}^{N} t'_n \sum_{j\in\{x,y,w,h,\theta\}} L_{reg}(v'_{nj}, v_{nj}) \tag{24}$$

TABLE I
INFORMATION ABOUT EACH EXPERIMENTAL DATASET

Datasets	Instances	Images	Categories	Annotation	Image format	Fine-grained
DIOR-R [35]	192 472	23 463	20	OBB	JPG	N
DOTA-v1.0 [36]	188 282	2806	15	OBB	PNG	N
HRSC2016 [37]	2976	1070	1	OBB	BMP	N
FAIR1M-v1.0 [38]	1.02 million	15 266	37	OBB	TIFF	Y
UCAS-AOD [52]	14.596	1510	2	OBB	PNG	N

where N indicates the number of positive sample, t'_n is a indicator function ($t'_n = 1$ for foreground and $t'_n = 0$ for background). v'_{nj} represents the offset vectors of the predicted box, v_{nj} represents the offset vector of GT. The regression loss L_{reg} adopts smooth L1 loss, which as defined in [27].

IV. EXPERIMENTS

A. Datasets

In order to comprehensively evaluate the performance of the proposed SARFA-Net, multiple public datasets with different scenes and image types are used. The detailed properties of the datasets are given in Table I.

DIOR-R [35], based on DIOR [2] dataset, consists of 23 463 images and 192 518 instances annotated with oriented bounding boxes. It covers 20 common categories: Airplane (APL), Airport (APO), Baseball Field (BF), Basketball Court (BC), Bridge (BR), Chimney (CH), Expressway Service Area (ESA), Expressway Toll Station (ETS), Dam (DAM), Golf Field (GF), Ground Track Field (GTF), Harbor (HA), Overpass (OP), Ship (SH), Stadium (STA), Storage Tank (STO), Tennis Court (TC), Train Station (TS), Vehicle (VE), and Windmill (WM).

DOTA-v1.0 [36] comprises 2806 aerial images with varying resolutions (800×800 to 4000×4000 pixels) collected from diverse sensors and platforms. It contains 188 282 instances across 15 categories: Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Small vehicle (SV), Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer ball field (SBF), Roundabout (RA), Harbor (HA), Swimming pool (SP), and Helicopter (HC). In our experiments, we divide images into 1024x1024 sub-images with a 200-pixel overlap and apply random horizontal, vertical, and diagonal flips during training.

HRSC2016 [37] comprises 1061 high-resolution images collected from six major ports and annotated with oriented bounding boxes. It is divided into training (436 images), validation (181 images), and test (444 images) sets.

FAIR1M-v1.0 [38] is a large-scale remote sensing dataset. It consists of 15 266 high-resolution images with resolutions ranging from 300 to 800 from different platforms and more than 1 million instances for fine-grained object recognition. All instances in the FAIR1M-1.0 dataset are annotated with 5 categories and 37 subcategories by oriented bounding boxes. All original images are divided into training set and test set in a ratio of 4:1, and we divide the images into 1024× 1024 with a stride of 200.

TABLE II
ABLATIVE EXPERIMENTS AND EVALUATIONS OF THE PROPOSED METHOD ON THE **DIOR-R** DATASET AND **FAIR1M-V1.0** DATASET

Dataset	SAS	TCM	RFAM	AP_{50}	AP_{75}	$mAP_{50:95}$
DIOR-R	–	–	–	62.50	32.40	34.76
	✓	–	–	63.60	35.80	36.63
	✓	✓	–	63.90	35.80	36.77
	✓	✓	✓	**66.20**	**39.10**	**39.24**
FAIR1M-v1.0	–	–	–	63.70	43.30	40.42
	✓	–	–	66.70	46.00	42.90
	✓	✓	–	67.10	46.40	43.80
	✓	✓	✓	**68.10**	**51.40**	**46.34**

The best result is highlighted in bold.

UCAS-AOD [52] is an aerial aircraft and car detection dataset with 1510 images. We randomly divide it into a training set and a test set at a ratio of 7:3.

B. Experimental Details

In our experiments, we employ S^2A-Net-C as the baseline, a variant of S^2A-Net that replaces aligned convolution with standard convolution. For simplicity and efficiency, we employ a ResNet-50 backbone pre-trained on ImageNet and an FPN neck, unless otherwise specified. Each level of the feature pyramid is preset with a single square anchor in its matched position. SGD optimization with weight decay of 1.0×10^{-2}, momentum of 0.9, and weight decay of 0.1 is used. All experiments are trained for 36 epochs using a single NVIDIA L40 GPU with a batch size of 8. The initial learning rate is 6.25×10^{-2}, reduced by a factor of 10 at epochs 24 and 32.

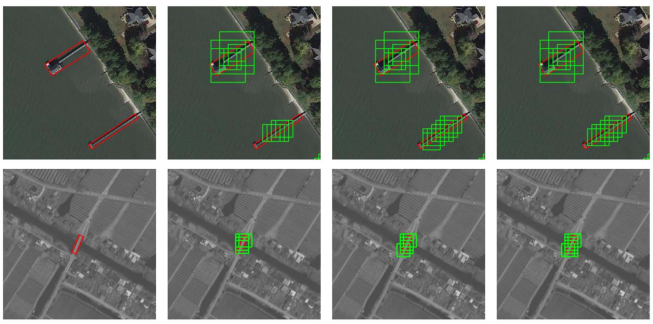
C. Ablation Study

In this section, a series of ablative experiments are conducted on the DIOR-R dataset and FAIR1M-v1.0 dataset to illustrate the advantages of each proposed component. Here, the components proposed are indicated in abbreviated form, i.e., "-S" indicates SAS module, "-T" means TCM, and "-R" means RFAM. The overall results of the ablative experiments are presented in Table II. Specifically, the first row represents the results of our baseline detector. The ablation results in AP_{50} and AP_{75} are 63.90% and 35.80%, respectively, obtained by replacing the MaxIoU assignment strategy with our SALA strategy. Specifically, compared to the baseline, AP_{50} and AP_{75} increase by 1.10% and 3.40%, respectively, by using SAS module. By adopting the TCM module, the AP_{50} further increases by 0.3% based on SAS.

TABLE III
ABLATIVE EXPERIMENTS AND EVALUATIONS OF THE PROPOSED METHOD ON THE **DIOR-R** DATASET

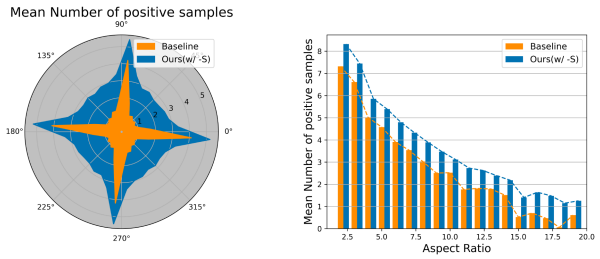
Method	APL	APO	BF	BC	BR	CH	ESA	ETS	DAM	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP
Baseline	67.70	36.30	**76.50**	81.50	35.70	72.60	77.70	65.10	22.50	77.00	80.10	43.50	51.90	80.60	66.90	67.00	81.40	54.30	46.70	64.50	62.50
Ours (w/o -S)	63.00	43.40	71.80	**81.60**	38.20	76.10	79.70	67.40	27.60	**78.70**	80.20	45.30	52.80	80.90	68.00	68.60	81.50	55.70	47.50	64.70	63.60
Ours (w/o -R)	63.10	46.50	72.00	81.50	38.40	72.60	79.50	67.40	27.80	78.20	79.80	45.80	55.00	80.90	70.10	69.50	81.50	54.50	47.80	65.80	63.90
Ours	**69.60**	**51.30**	74.90	81.50	**42.60**	**77.60**	**80.20**	**70.90**	**31.90**	77.70	**80.80**	**47.50**	**57.30**	**80.90**	**72.10**	**70.30**	**81.50**	**59.20**	**50.00**	**66.00**	**66.20**

All methods adopt "3x" training schedule and use R-101 as backbone. The best result is highlighted in bold.



(a) (b) (c) (d)

Fig. 6. Visualization comparison of sampling for extreme scale and aspect ratio objects with different label assignment strategies. (a) Ground-truth. (b) ATSS. (c) SALA (w/o -T). (d) SALA.



(a) (b)

Fig. 7. Illustration of sampling results of MaxIoU and SALA strategy. (a) Imbalanced angles sampling. (b) Imbalanced aspect ratios sampling.

When combining all components, we can achieve 66.20% and 39.10% AP_{50} and AP_{75}, as shown in the last row of Table II. It is worth noting that $mAP_{50:95}$ achieves optimal performance, with 4.48% improvements compared to the baseline. Similarly, we conduct ablation experiments on the FAIR1M-v1.0 dataset to fully demonstrate the robustness of each module. In addition, in order to introduce the improvements of different categories, more detailed experimental results for each category are provided in Table III. The improvement contributions of each component are discussed in detail as follows.

1) Effect of SAS: The core of our proposed SALA is the SAS, which resolves the problem of insufficient sampling of objects, especially those with extreme aspect ratios. Fig. 6 provides the assignment results of different label assignment strategies. Specifically, Fig. 6(a) illustrates example objects used for model training in the remote sensing images. As shown in Fig. 6(b), samples of the ATSS strategy are only clustered in the central area of the object. As a result, the model ignores edge features at the ends of the large aspect ratio objects, leading to poor performance in angle and width/height regression. Therefore, the two end features of extreme aspect ratio objects are ignored by the model, resulting in poor performance in the regression process of orientation and aspect ratio. As shown in Fig. 6(c), the proposed SAS method controls the sampling area by dynamically fitting the shape of the elliptical region, making it more suitable for oriented object detection tasks.

As shown in Table III, the proposed SAS strategy improves mAP by 1.10% compared to the baseline method, especially for objects with extreme aspect ratios such as APO, DAM, HA, and OP, which increase by a large margin. Meanwhile, for multiscale objects such as BR and CH, the improvements are 2.50% and 3.50%, respectively. However, although the SAS strategy can achieve better distribution under extreme shape conditions, it damages the sampling effect of small objects like APL and slightly reduces accuracy. We speculate that the SAS strategy restricts the sampling region when selecting candidate samples, resulting in insufficient positive samples for small-scale objects, which may hinder the model's detection accuracy.

In terms of quantity, as shown in Fig. 7, the SALA strategy effectively alleviates the problem of positive sample imbalance in sampling tasks with arbitrary orientation and extreme aspect ratio. Regarding angles, as the model adopts long edge definition, the angle $\theta \in [-\frac{\pi}{4}, \frac{3\pi}{4}]$. Hence, as shown in Fig. 7(a), we count the number of positive samples for each angle of the MaxIoU and SAS strategies (rounded up every 5 degrees interval), and then expand the number of positive samples for the corresponding angle. As shown in Fig. 7(b), when the aspect ratio is greater than 10, there is a significant improvement in the number of positive samples.

2) Effect of TCM: Based on SAS module, the proposed TCM module alleviates the problem of insufficient sampling for object with extreme scales. Although the SAS module alleviates the sampling imbalance problem of arbitrary orientation and extreme aspect ratios, the inherent flaw of this strategy leads to unfriendly detect performance towards small objects.

As shown in Fig. 6(d), the proposed TCM method dynamically weights the threshold of small objects to increase learning positive samples. In terms of quantity, as shown in Fig. 8(a), when the scale metric benchmark $scale$ is less than 5, the number of positive samples sharply decreases. It is worth noting that the metric scale is processed by logarithmic function to be applied to different datasets. The above issues have been discussed in Section I, and our optimized results are shown in Fig. 8(b). In Table III, it can be seen that compared to the SAS module, the TCM module has improved the detection accuracy of small-scale objects in VE, STO, and WM by 0.30%, 0.90%, and 0.90%, respectively.

3) Effect of Anchor Settings in the SALA: Anchor settings significantly influence the effectiveness of label assignment strategies. To assess this impact, we conduct experiments

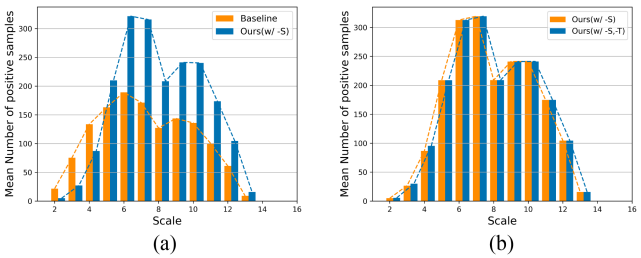


Fig. 8. Illustration of SAS strategy being unfriendly to small objects. (a) Imbalanced scale sampling. (b) Comparison of sampling results with or without the TCM module.

TABLE IV
RESULTS OF DIFFERENT BASELINE AND ANCHOR SETTINGS ON THE **HRSC2016** DATASET

Baseline	MaxIoU	ATSS	SALA	Anchor	mAP
Rotated RetinaNet [51]	✓			**9**	61.90
		✓		**9**	85.30
			✓	**9**	85.60
			✓	**1**	88.90
S^2A-Net-C [10]	✓			**1**	80.00
		✓		**1**	89.10
			✓	**1**	89.70

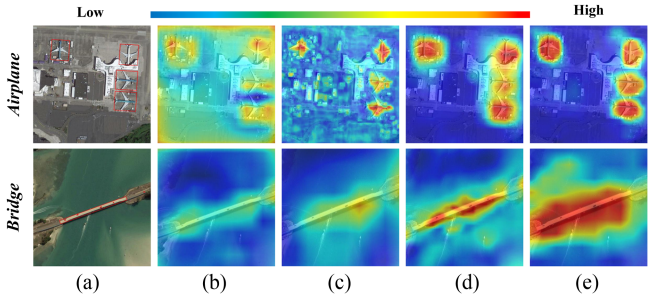


Fig. 9. Heatmaps visualization with different label assignment strategies. (a) Ground Truth. (b) Baseline. (c) ATSS. (d) SALA. (e) SARFA-Net.

on the HRSC2016 dataset using Rotated RetinaNet [51] and S^2A-Net-C [10] as baselines. To ensure a fair comparison, all experiments are conducted under identical conditions. As shown in Table IV, under the Rotated RetinaNet with MaxIoU and 9 anchors per position, we achieve the mAP of 61.90%. Notably, ATSS and SALA, with the same settings, demonstrate accuracies of 85.30% and 85.60%, respectively. It is worth mentioning that when reducing the number of anchors to one per position, SALA's accuracy improves by 3.30%. Experiments show the importance of strategically utilizing predefined anchors and selecting high-quality samples without relying on excessive anchor points.

To further demonstrate the effectiveness of the SALA strategy, we present the class-discriminative heatmap in Fig. 9, which compares the effects of different label assignment strategies. As shown in Fig. 9(a) and (b), ATSS tends to overemphasize the central region of objects, hindering its ability to learn sufficient features and accurately represent orientation information, especially for objects with large aspect ratios, such as harbors

TABLE V
COMPARISON OF DETECTION RESULTS WITH DIFFERENT CONVOLUTION MODULES ON THE **DIOR-R** DATASET

Methods	APL	APO	HA	STO	TC	5-mAP	mAP
Conv	67.70	36.30	43.50	67.00	81.40	59.18	62.50
DCN	63.10	36.20	44.80	68.40	81.50	58.80	62.90
AlignConv	63.00	43.50	46.50	69.90	81.50	60.88	64.60
RFAM (Ours)	**71.10**	**48.50**	**47.20**	**70.30**	**86.90**	**64.80**	**65.80**

5-mAP indicates the performance of the categories listed. The best result is highlighted in bold.

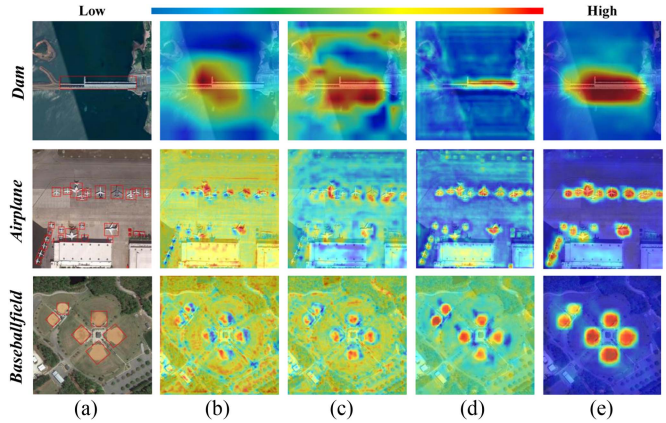


Fig. 10. Heatmaps visualization with different feature alignment strategies.

and bridges. In contrast, Fig. 9(b) and (d) illustrate how SALA effectively focuses on and extracts edge features of objects with extreme scales, making it more suitable for remote sensing oriented objects. When the η value is dynamically adjusted, the model can evidently acquire more comprehensive target features compared to a fixed η of 0.5. By comparing Fig. 9(d) and (e), we observe that the edge information of aircraft becomes clearer, while the edge features of bridges are more prominent. This enhancement reflects a better representation of orientation information and demonstrates the effectiveness of RFAM in extracting more discriminative contextual information.

4) Effect of RFAM: Table V compares the performance of different convolution modules replacing the standard convolution in S^2A-Net-C [10]. Our proposed RFAM achieves the highest accuracy for both 5-mAP and mAP, reaching 64.80% and 65.80%, respectively. Notably, RFAM surpasses all compared methods in detection accuracy across five representative object categories. Compared with baseline, for objects with large aspect ratios like APO and HA, we observe significant improvements of 12.20% and 3.70%, respectively. In addition, for small-scale objects like STO, we achieve gains of 3.30%. Given the inherent background noise of airplanes, flexible sampling points are crucial. Compared to the AlignConv, our approach improves mAP for APL by 8.10% and achieves notable gains of 5.40% and 5.00% for large-scale TC and APO, respectively.

To further demonstrate the effectiveness of the proposed RFAM, we employ a gradient-based heatmap to visualize the feature maps of detectors with different alignment strategies (Fig. 10). As evidenced in the 1 row of Fig. 10(b), the fixed convolution kernel in standard convolution struggles

TABLE VI
EVALUATION OF USING VARIOUS RATIOS FACTOR η IN SALA STRATEGY ON THE **DIOR-R** DATASET

η	1.00	0.75	0.50	0.25	dynamic
mAP	66.00	66.10	65.90	65.60	**66.20**

The best result is highlighted in bold.

TABLE VII
ANALYSIS OF DIFFERENT HYPERPARAMETER α OF RFAM ON THE **DIOR-R** DATASET

α	APL	APO	BR	STO	VE	WM	6-mAP	mAP
0.8	62.90	51.60	40.90	56.30	48.70	65.90	52.55	65.90
0.9	63.10	49.70	41.40	56.60	48.70	65.70	52.52	65.40
1.0	63.10	50.60	41.60	56.70	48.80	65.90	52.78	66.10
1.1	**69.60**	51.30	**42.60**	**57.30**	**50.00**	**66.00**	**54.47**	**66.20**
1.2	63.00	49.10	42.10	56.70	49.00	65.70	52.60	65.50
1.3	67.20	**52.40**	41.60	56.50	48.90	65.60	53.70	66.00

6-mAP indicates the performance of the categories listed. The best result is highlighted in bold.

to capture slender objects' features compared to DCN-based methods [Fig. 10(c)–(e)]. This is because standard convolution can only focus on features at fixed positions, whereas DCN-based methods offer greater flexibility for feature extraction.

While DCN can capture relatively complete object features [Fig. 10(c)], it also introduces background noise (e.g., dam). A key limitation of DCN is its inability to incorporate objects' direction information, leading to suboptimal high-precision positioning performance. Conversely, AlignConv utilizes more flexible feature sampling points [Fig. 10(d)], which reduces background interference while maintaining adequate feature learning. However, small objects (e.g., airplane) benefit from additional contextual features for enhanced learning, whereas slender objects require precise localization to achieve high-precision boundary detection. As shown in Fig. 10(e), RFAM enables robust and aligned localization feature extraction for both slender and small objects.

D. Evaluation of Hyperparameters

1) Ratio Factor η: The ratio factor η is introduced into the SALA strategy to ensure the quality of sample selection by controlling the distribution of samples. As shown in Table VI, different values are tested to seek the optimal value. We can see that when $\eta = 0.25$, the sampling range is small enough to avoid noise, and the mAP reaches 65.60%. As η increases, the detector can learn more information from the selected high-quality samples, thus improving detection performance and reaching a peak of 66.10% when $\eta = 0.75$. Based on the above discussion, we adaptively set the value of η based on the object's shape. From Table VI, it can be observed that mAP achieve its peak values when using the detector with adaptive η, which proves the effectiveness of our proposed adaptive method and demonstrates that the key to improving detection performance through spatial assignment is to obtain more positive samples while reducing the effect of noise.

2) Ratio Factor α: The scale factor α is introduced in the RFAM to control the expansion coefficient to determine the initial sampling positions of the anchor. Because we have added a set of learnable offsets, the sampling positions of the convolution kernels can be adaptively adjusted. As shown in Table VII, different values are tested to find the best value, and the impact of various factors is shown. When $\alpha < 1$, the sampling positions are reduced to focus on internal features, which is not friendly for obtaining edge features of objects with large aspect ratios (e.g., BR). As α increases, the detector can capture more information from higher-quality context features, so the performance increases and reaches its peak when $\alpha = 1.1$. Since invalid background information interference needs to be considered, the performance begins to decline when the threshold is further increased (e.g., STO, VE, and WM). Considering the similarity of the DOTA dataset distribution to that of DIOR-R, it is reasonable to set the weight parameter α of RFAM to 1.1. However, for HRSC2016, which contains more objects with large aspect ratios, setting α to 1.3 is more appropriate.

E. Comparison With State-of-the-Art

In this section, we assess the robustness and superiority of our SARFA-Net by evaluating its performance on five challenging remote sensing datasets for oriented object detection and comparing it with other state-of-the-art models.

1) Results on DIOR-R Dataset: We compare the results of different methods on the DIOR-R dataset in Table VIII. Specifically, based on S^2A-Net [10], our proposed SARFA-Net improves mAP by 2.68%. Our method demonstrates superior detection performance in most categories, especially with extremely small and aspect ratio objects such as APO, TS, and VE, achieving the best results. For challenging categories like APO, HA, and VE, we achieved significant mAP gains of 3.07%, 4.43%, and 1.41%, respectively. Compared with other dynamic label assignment methods, ATSS [16] utilizes statistical results to guide sampling, which is initially developed for object detection in natural scenes. We adapt it for oriented object detection in remote sensing and obtain 1.70% mAP improvements compared with Rotated ATSS [16], as shown in Table VIII. SASM [24] enhances the model's ability to learn objects with large aspect ratios by setting dynamic thresholds. However, it achieves a 6.39% lower mAP compared to our model. Specifically, exemplified by APO, BR, and SH with large aspect ratios, SARFA-Net achieves mAP improvements of 5.27%, 13.19%, and 2.28%, respectively. Some qualitative detection results of SARFA-Net are shown in Fig. 11. From the cases, our method performs well on oriented objects in remote sensing images, even with extreme aspect ratios and scales.

2) Results on DOTA-V1.0 Dataset: The comparison of our method with other state-of-the-art methods is presented in Table IX. As shown, our method achieves 75.66% and 75.69% w.r.t mAP based on R-50 with FPN and R-101 with FPN, respectively, which outperform other state-of-the-art methods. Remarkably, our method also obtains the best results in some very challenging categories, such as BD, TC, and SBF. Compared with the well designed 2-D Gaussian distribution of GGHL [49] (D-53 means DarkNet53 [63]), we only use an intuitive 2-D ellipse to assist positive sample sampling, achieving a mAP

TABLE VIII
COMPARISONS WITH THE ADVANCED ORIENTED DETECTORS ON **DIOR-R** DATASET

Methods	APL	APO	BF	BC	BR	CH	ESA	ETS	DAM	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP
Two-stage																					
Gliding Vertex [42]	62.67	38.56	71.94	81.20	37.73	72.48	78.62	69.04	22.81	77.89	82.13	46.22	54.76	81.03	74.88	62.54	81.41	54.25	43.22	65.13	62.91
Rotated Faster RCNN [53]	66.52	46.80	71.76	81.43	40.81	78.25	79.23	66.63	29.01	78.68	80.19	44.88	57.23	80.91	74.17	68.02	81.48	54.63	47.80	64.41	64.63
RoI-Transformer [12]	63.18	44.33	71.91	81.26	42.19	72.64	79.30	69.67	29.42	77.33	82.88	48.09	57.03	81.18	77.32	62.45	81.38	54.34	43.91	66.30	64.31
ReDet [14]	63.22	44.18	72.11	81.26	43.83	72.72	79.10	69.78	28.45	78.69	77.18	48.24	56.81	81.17	69.17	62.73	81.42	54.90	44.04	66.37	63.81
Oriented RCNN [54]	63.31	43.10	71.89	81.17	44.78	72.64	80.12	69.67	33.78	77.92	83.11	46.29	58.31	81.17	74.54	62.32	81.29	56.30	43.78	65.26	64.53
One-stage																					
Rotated RetinaNet [51]	59.54	25.03	70.08	81.01	28.26	72.02	55.35	56.77	21.26	65.70	70.28	30.52	44.37	77.02	59.01	59.39	81.18	38.43	39.10	61.58	54.83
SASM [24]	61.41	46.03	73.22	82.04	29.41	71.03	69.22	53.91	30.63	70.04	77.02	39.33	47.51	78.62	66.14	62.92	79.93	54.41	40.62	63.01	59.81
S^2A-Net [10]	67.98	44.44	71.63	81.39	42.66	72.72	79.03	70.40	27.08	75.56	81.02	43.41	56.45	81.12	68.00	70.03	87.07	53.88	51.12	65.31	64.50
R3Det [11]	62.55	43.44	71.72	81.48	36.49	72.63	79.50	64.41	27.02	77.36	77.17	40.53	53.33	79.66	69.22	61.10	81.54	52.18	43.57	64.13	61.91
GWD [44]	66.52	46.80	71.76	81.43	40.81	78.25	79.23	66.63	29.01	78.68	80.19	44.88	57.23	80.91	74.17	68.02	81.48	54.63	47.80	64.41	64.63
KLD [45]	69.68	28.83	74.32	81.49	29.62	72.67	76.45	63.14	27.13	77.19	78.94	39.11	42.18	79.10	70.41	58.69	81.52	47.78	44.47	62.63	60.31
Rotated FCOS [55]	62.31	42.18	75.34	81.32	39.26	74.89	77.42	68.67	26.00	73.94	78.73	41.28	54.19	80.61	66.92	69.17	87.20	52.31	47.08	65.21	63.21
Rotated ATSS [16]	62.19	44.63	71.55	81.42	41.08	72.37	78.54	67.50	30.56	75.69	79.11	42.77	56.31	80.92	67.78	69.24	81.62	55.45	47.79	64.10	63.52
CFA [56]	61.10	44.93	77.62	84.67	37.69	75.71	82.68	72.03	33.41	77.25	79.94	46.20	54.27	87.01	70.43	69.58	81.55	55.51	49.53	64.92	65.25
SARFA-Net (Ours)	69.60	47.70	76.30	81.40	41.00	77.90	79.50	66.20	30.20	78.10	81.10	47.20	57.50	88.00	71.80	69.90	81.50	56.00	49.20	65.80	65.80
SARFA-Net* (Ours)	69.60	51.30	74.90	81.50	42.60	77.60	80.20	70.90	31.90	77.70	80.80	47.50	57.30	80.90	72.10	70.30	81.50	59.20	50.00	66.00	66.20
SARFA-Net† (Ours)	71.80	53.60	77.80	89.50	43.90	78.60	87.40	71.20	37.30	79.00	83.30	49.50	58.30	88.40	72.00	70.10	89.70	61.90	49.70	65.70	68.90

All methods default to relying on "3x" training schedule and use ResNet-50 as backbone. * indicates ResNet-101 as backbone. † indicates random rotate data enhancement. Red and blue: Top two performances.

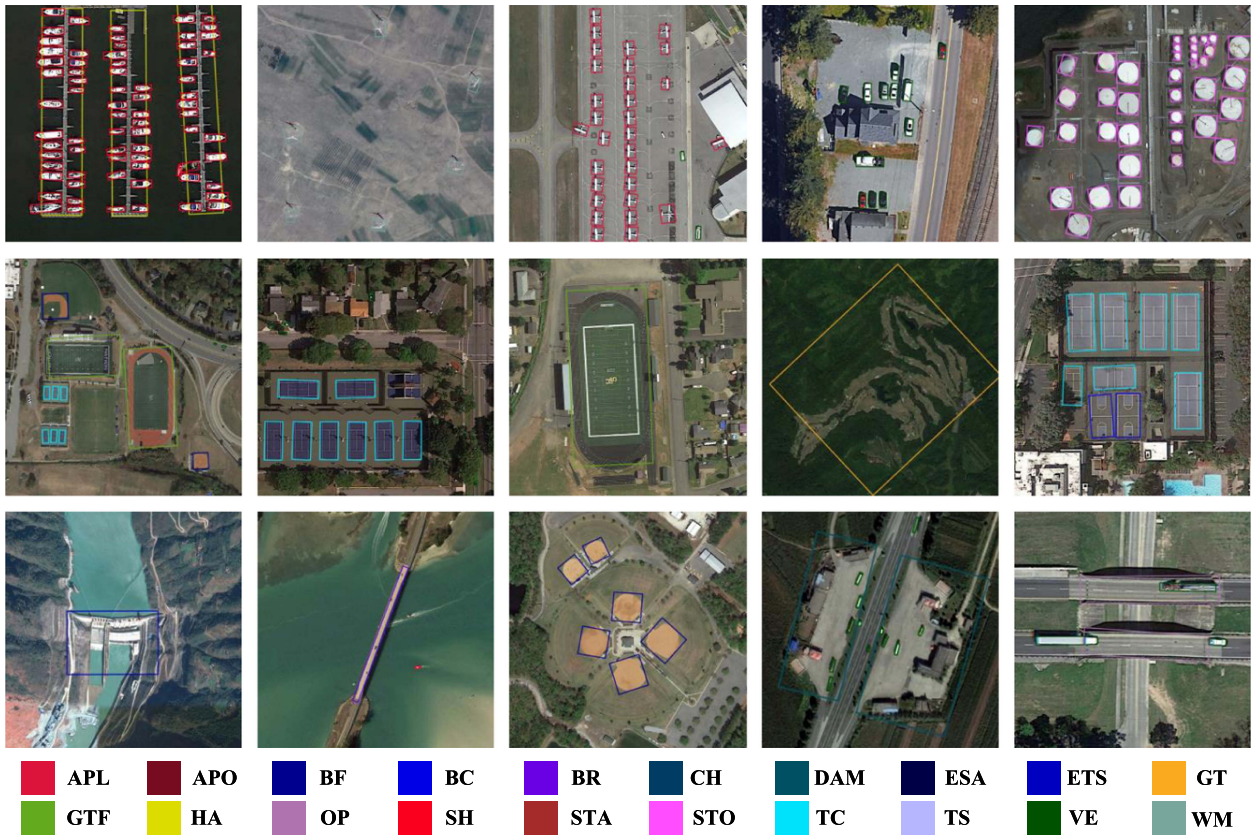


Fig. 11. Visualization of detection results on the **DIOR-R** dataset.

advantage of 3.14%. SCRDet [13] claims that feature information and the number of anchors are two key factors affecting the performance of small object detection. It uses the attention mechanism Multi-Dimensional Attention Network (MDA-Net) and adjusts the sampling step size to more effectively improve the feature extraction of small objects in complex backgrounds. However, for SV and SH, its mAP is 3.24% and 15.31% lower than ours. STD [62] is based on stacked vision Transformer blocks and uses separate network branches to predict the position, size, and angle of the bounding box, which is 2.15% mAP higher than our results. However, it uses HiViT-B [64] as the backbone network, which consumes huge computation and training costs.

Fig. 12 demonstrates our model's ability to detect oriented objects of varying scales and extremely aspect ratios, where denoted by the red or green dotted line is the main region of concern. Specifically, row 1 shows that our SARFA-Net significantly outperforms the baseline in aligning objects with large aspect ratios, which is attributed to enhanced learning the edge features and flexible sampling positions. Compared to

TABLE IX
COMPARISONS WITH THE ADVANCED ORIENTED DETECTORS ON **DOTA-V1.0** DATASET

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-stage																	
Oriented RCNN [54]	R-50	89.23	77.97	54.05	73.38	74.44	78.07	87.93	90.54	77.08	84.68	61.38	65.54	76.27	70.96	51.39	74.19
AOPG [35]	R-50	89.27	83.49	52.50	69.97	73.51	82.31	87.95	90.89	87.64	84.71	60.01	66.12	74.19	68.30	57.80	75.24
RoI Trans. [12]	R-101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
Gliding Vertex [42]	R-101	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
One-stage																	
S^2A-Net [10]	R-50	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
R3Det [11]	R-101	88.76	83.09	50.91	67.27	76.23	80.39	86.72	90.78	84.68	83.24	61.98	61.35	66.91	70.63	53.94	73.79
SASM [24]	R-50	87.51	80.15	51.07	70.35	74.95	75.80	84.23	90.90	80.87	84.93	58.51	65.59	69.74	70.18	42.31	72.47
Rotated ATSS [16]	R-50	88.94	79.89	48.71	70.74	75.80	74.02	84.14	90.89	83.19	84.05	60.48	65.06	66.74	70.14	57.78	73.37
CFC-Net [57]	R-101	89.08	80.41	52.41	70.02	70.02	78.11	87.21	90.89	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50
SCRDet [13]	R-101	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
KLD [45]	R-50	89.13	79.94	51.23	72.56	78.24	78.90	87.10	90.87	85.01	83.81	59.84	64.83	69.92	70.48	55.35	74.48
PSC [58]	R-50	88.27	73.20	44.55	62.29	77.79	77.30	87.04	90.88	78.47	72.01	52.69	61.14	66.36	69.68	58.10	70.65
GGHL‡ [49]	D-53	89.74	85.63	44.50	77.48	76.72	80.45	86.16	90.83	88.18	86.25	67.07	69.40	73.38	68.45	70.14	76.95
H2RBOX-v2‡ [59]	R-50	89.45	80.72	54.29	72.60	81.68	83.98	88.44	90.88	86.11	86.04	64.77	69.42	76.38	79.64	65.08	77.97
RTMDet-tiny† [60]	R-50	89.19	80.01	47.96	69.63	82.06	83.35	88.62	90.90	86.25	86.87	60.06	62.69	74.25	71.93	56.85	75.38
G-Rep [20]	R-50	87.76	81.29	52.64	70.53	80.34	80.56	87.47	90.74	82.91	85.01	61.48	68.51	67.53	73.02	63.54	75.56
YOLO-v8 [61]	D-53	84.90	64.00	38.30	54.80	58.40	73.50	77.20	93.30	67.70	72.10	54.20	57.70	60.50	47.10	47.70	63.40
STD‡ [62]	HiViT-B	89.15	85.03	60.79	82.06	80.90	85.76	88.45	90.83	87.71	87.29	73.99	71.25	85.18	82.17	82.95	82.24
SARFA-Net (Ours)	R-50	89.06	82.92	49.96	71.60	77.56	81.10	87.72	90.88	85.02	86.03	62.05	66.94	75.34	71.45	57.39	75.66
SARFA-Net (Ours)	R-101	89.12	86.11	51.85	70.27	76.60	79.55	87.54	90.90	85.59	85.70	65.48	64.07	76.61	71.07	54.97	75.69
SARFA-Net‡ (Ours)	R-50	89.81	85.24	57.08	77.57	81.06	83.08	88.86	90.86	84.90	88.73	72.11	71.11	79.29	78.19	73.53	80.09
SARFA-Net‡ (Ours)	R-101	89.42	84.59	57.68	79.45	81.16	85.23	89.20	90.79	83.92	88.06	68.81	69.14	79.35	74.26	65.39	79.10

‡ means multiscale training and testing. † indicates random rotate data enhancement. Red and blue: Top two performances.

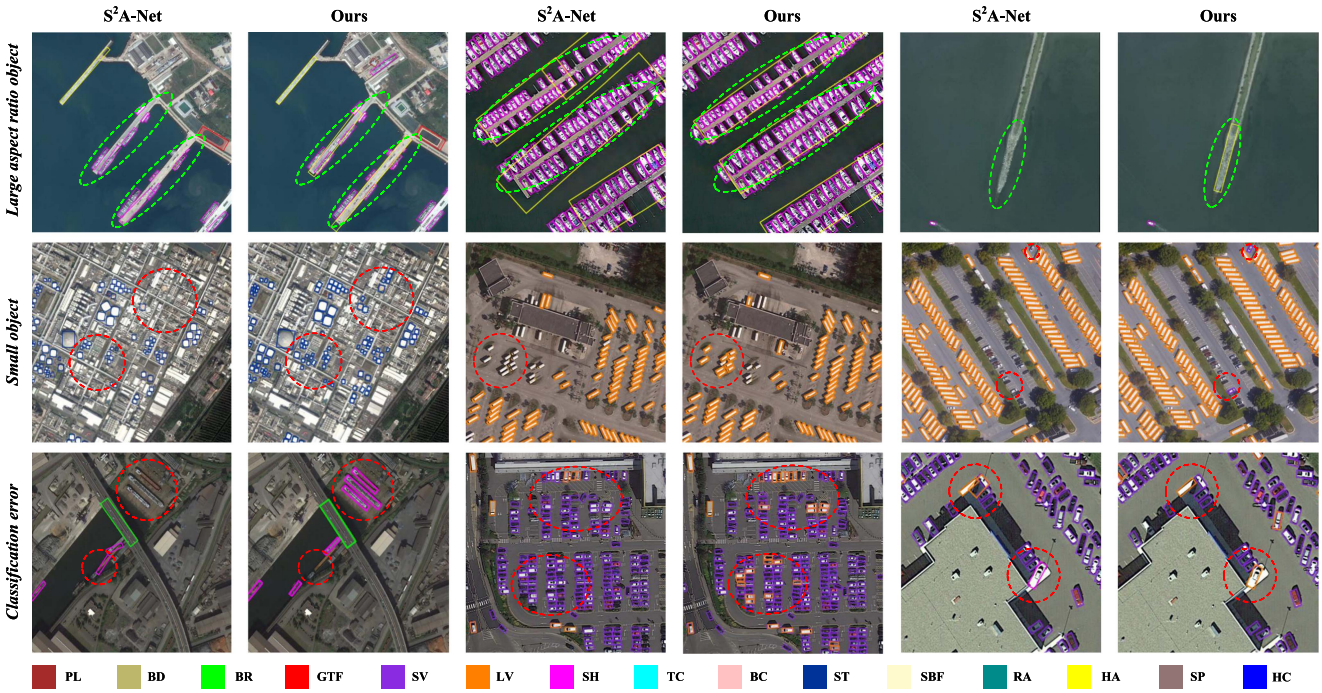


Fig. 12. Detection results are visually compared on the **DOTA-v1.0** dataset. The area surrounded by the red or green dashed line is the focus of attention.

S^2A-Net, we improve the detection performance of objects with extreme aspect ratios (e.g., missed detections and misalignments of SHs and HAs, and a single BR being identified as multiple objects). In addition, rows 2 demonstrates that the detection performance of our method for tiny objects (e.g., missed detection of STs and SVs). Densely packed objects can result in false positives and misclassifications due to the ambiguity of semantic features. Different from some networks [65], [66] that are specifically designed to extract fine-grained features, we use RFAM to reduce the effect of erroneous semantic sampling. As shown in row3, we reduce the false detection of LVs in dense SVs and some misclassification problems (e.g., SHs on the shore, HA is identified as SH). In addition, our method achieves competitive mAP of 80.09% and 79.10% with R-50 and R-101

TABLE X
PERFORMANCE COMPARISON WITH DIFFERENT STATE-OF-THE-ART METHODS ON THE **HRSC2016** DATASET

Methods	RoI Trans. [12]	Gliding Vertex [42]	R3Det [11]	CSL [46]	DAL [15]	S^2A-Net [10]	SASM [24]	SARFA-Net (Ours)
mAP	86.20	88.20	89.26	89.62	89.77	90.17	90.27	**90.40**

The best result is highlighted in bold.

TABLE XI
PERFORMANCE COMPARISON WITH DIFFERENT STATE-OF-THE-ART METHODS ON THE **FAIR1M-V1.0** DATASET

Methods	Gliding Vertex [42]	Rotated RetinaNet [51]	Faster R-CNN [67]	RoI Trans. [12]	Orinented R-CNN [54]	SARFA-Net (Ours)
mAP	29.92	30.67	31.18	35.29	45.60	**46.34**

The best result is highlighted in bold.

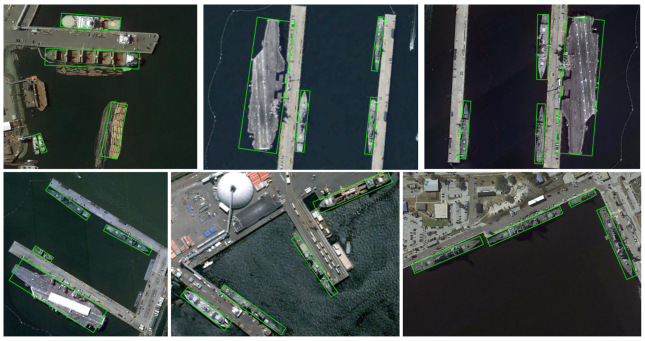


Fig. 13. Visualization of detection results on the **HRSC2016** dataset.

TABLE XII
COMPARISON WITH THE ADVANCED ORIENTED DETECTORS ON THE **UCAS-AOD** DATASET

Methods	Car	Airplane	mAP
Rotated RetinaNet [51]	84.64	90.51	87.57
S^2A-Net [10]	89.30	90.16	89.73
RIDet-Q [68]	88.50	89.96	89.23
RIDet-O [68]	88.88	90.35	89.62
Rotated Faster RCNN [53]	84.64	90.51	87.57
SASM [24]	**89.56**	90.42	90.00
RoI Trans. [12]	88.02	90.02	89.02
DAL [15]	89.25	90.49	89.87
Ours	89.50	**90.60**	**90.10**

The best result is highlighted in bold.

TABLE XIII
EFFICIENCY COMPARISON WITH DIFFERENT METHODS IN MAP (%), FPS (TASK/S), FLOPS (G), AND PARAMETERS (M) ON THE DOTA-V1.0 DATASET

Methods	mAP	GFLOPs / FPS(img/s)	Parameters(M)
Faster R-CNN [67]	69.05	211.30 / 6.4	41.14
RoI Trans. [12]	74.61	225.29 / 4.4	55.13
Rotated RetinaNet [51]	68.43	215.92 / **7.3**	36.42
S^2A-Net [10]	74.12	**172.48** / 7.2	**36.21**
Ours	**75.66**	198.11 / 5.8	39.22

The best result is highlighted in bold.

backbones, respectively, through multiscale training and testing. Generally speaking, the detection performance of R-101-FPN as the backbone network is superior to R-50, but the results are the opposite. We speculate that our network SARFA-Net possess robust detection performance, which enables high-precision detection using the resource-saving R-50.

3) Results on HRSC2016 Dataset: To further evaluate the generalization ability of the proposed method, we further conduct experiments on the HRSC2016 dataset, and the results are detailed in Table X. Notably, our method achieves a remarkable 90.40% mAP, which surpasses the comparative methods by a large margin. The detection results on the HRSC2016 dataset, as depicted in Fig. 13, depicting the powerful ability of our proposed model for the objects with large aspect rations.

4) Results on FAIR1M-V1.0 Dataset: The FAIR1M-v1.0 dataset is a challenging multicategories dataset. As shown in Table XI, the experimental results show that our model achieved the best mAP of 46.34% compared with five other models, which fully demonstrates the superiority of our model in multicategories fine-grained detection tasks.

5) Results on UCAS-AOD Dataset: Cars and airplanes in UCAS-AOD are usually small, distributed in any direction, and surrounded by complex scenes. The results are shown in Table XII. The proposed SARFA-Net achieves 90.10% mAP, verifying the superiority of sample selection and the robustness for small objects.

6) Efficiency Comparison: Comparison results of mAP (%), FPS (img/s), FLOPS (G), and parameters (M) are also performed on the DOTA dataset. All experiments are based on R-50-FPN, using a single L40 GPU with a batch size of 1. The resolution of the input image is 1024× 1024. Table XIII shows the detailed experimental results of different methods. S^2A-Net requires 36.21M Parameters and 172.48 GFLOPs, achieving 74.12% mAP and 7.2 FPS. The proposed SARFA-Net achieves 75.66% mAP and 5.8 FPS. Compared with the baseline, the computational cost of the proposed method is slightly higher (39.22 Parameters versus 36.21 Parameters). It can be seen that the proposed method can achieve better performance with comparable computational cost.

7) Discussion: SALA is proposed to reduce the model's missed detection of specific objects by balancing the number of positive samples. In terms of mAP indicators, although SALA improves detection accuracy, it also has potential risks of misdetection. The edge features of large aspect ratio objects and the contextual features of small objects often carry background features in dense areas. For example, in the second row of

Fig. 12 (left side), it can be clearly seen that some buildings with similar features are still detected as STs; in the third row, although LVs can be effectively detected in dense vehicles, there are also SVs that is misdetected as LVs. Our RFAM can extract complete alignment features through more flexible sampling points, reduce the interference of target background features. However, RFAM is not designed to address the above problems, and the misdetection brought by SALA cannot be completely eliminated. We speculate that the introduction of redundant context information affects the detection effect. In future work, we plan to add some attention mechanisms to obtain appropriate and sufficient context information to further improve the performance.

V. Conclusion

In this article, a novel SARFA-Net is presented for the accurate AOOD in RSIs, which incorporates the SALA and RFAM. SALA, centered on SAS and TCM, alleviates the issue of insufficient positive samples in remote sensing images by dynamically setting thresholds based on suitable, continuous multilevel feature maps. In addition, RFAM addresses the issue of feature misalignment between anchors and GTs. Extensive experimental results demonstrated the effectiveness of our proposed model and the superiority compared with state-of-the-arts. In addition, the oriented object detection model SARFA-Net also has the prospect of being applied to natural scene text detection, lesion detection in medical images, and obstacle detection in intelligent driving.

References

[1] R. Bürgmann, P. A. Rosen, and E. J. Fielding, "Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation," *Annu. Rev. Earth Planet. Sci.*, vol. 28, no. 1, pp. 169–209, 2000.

[2] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[3] C.-A. Liu, Z.-X. Chen, S. Yun, J.-S. Chen, T. Hasi, and H.-Z Pan, "Research advances of SAR remote sensing for agriculture applications: A review," *J. Integrative Agriculture*, vol. 18, no. 3, pp. 506–525, 2019.

[4] A. C. Mondini, F. Guzzetti, K.-T. Chang, O. Monserrat, T. R. Martha, and A. Manconi, "Landslide failures detection and mapping using synthetic aperture radar: Past, present and future," *Earth- Sci. Rev.*, vol. 216, 2021, Art. no. 103574.

[5] B. Zhao et al., "Intermediate domain prototype contrastive adaptation for spartina alterniflora segmentation using multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5401314.

[6] M. Zhang, H. Zheng, M. Gong, Y. Wu, H. Li, and X. Jiang, "Self-structured pyramid network with parallel spatial-channel attention for change detection in vhr remote sensed imagery," *Pattern Recognit.*, vol. 138, 2023, Art. no. 109354.

[7] X. Ma, K. Ji, B. Xiong, L. Zhang, S. Feng, and G. Kuang, "Light-YOLOv4: An edge-device oriented target detection method for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10808–10820, 2021.

[8] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "Bifa-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4209.

[9] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges and perspectives," *Innovation*, vol. 5, 2024, Art. no. 100691.

[10] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.

[11] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3163–3171.

[12] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.

[13] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8231–8240.

[14] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2785–2794.

[15] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2355–2363.

[16] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9756–9765.

[17] P. Sun, Y. Zheng, W. Wu, W. Xu, and S. Bai, "Metric-aligned sample selection and critical feature sampling for oriented object detection," *IEEE Trans. Instrum. Meas.*, 2024.

[18] Y. Li, C. Bian, and H. Chen, "Dynamic soft label assignment for arbitrary-oriented ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1160–1170, 2023.

[19] J. Guan, M. Xie, Y. Lin, G. He, and P. Feng, "EARL: An elliptical distribution aided adaptive rotation label assignment for oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5619715.

[20] L. Hou, K. Lu, X. Yang, Y. Li, and J. Xue, "G-Rep: Gaussian representation for arbitrary-oriented object detection," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 757.

[21] C. Xu et al., "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7318–7328.

[22] M. Gong, H. Zhao, Y. Wu, Z. Tang, K.-Y. Feng, and K. Sheng, "Dual appearance-aware enhancement for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5602914.

[23] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2458–2466.

[24] L. Hou, K. Lu, J. Xue, and Y. Li, "Shape-adaptive selection and measurement for oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 923–932.

[25] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 4898–4906.

[26] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[27] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[29] X. Guo, S. Wang, J. Chang, Z. Chen, and F. Zhao, "SAFE: Simultaneous alignment of features and predictions for dense object detectors," in *Proc. 2023 IEEE Int. Conf. Multimedia Expo*, 2023, pp. 306–311.

[30] H. Zhang, H. Chang, B. Ma, S. Shan, and X. Chen, "Cascade RetinaNet: Maintaining consistency for single-stage object detection," 2019, *arXiv:1907.06881*.

[31] Y. Chen, C. Han, N. Wang, and Z. Zhang, "Revisiting feature alignment for one-stage object detection," 2019, *arXiv:1908.01570*.

[32] X. Qian, J. Zhao, B. Wu, Z. Chen, W. Wang, and H. Kong, "Task-aligned oriented object detection in remote sensing images," *Electronics*, vol. 13, no. 7, 2024, Art. no. 1301.

[33] Y. Yu, X. Yang, J. Li, and X. Gao, "Object detection for aerial images with feature enhancement and soft label assignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624216.

[34] Q. Yang, L. Cao, C. Huang, Q. Song, and C. Yuan, "A2Net: An anchor-free alignment network for oriented object detection in remote sensing images," *IEEE Access*, vol. 12, pp. 42017–42027, 2024.

[35] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411.

[36] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[37] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2017, vol. 2, pp. 324–331.
[38] X. Sun et al., "FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.
[39] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1745–1749, Nov. 2018.
[40] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based CNN for ship detection," in *Proc. 2017 IEEE Int. Conf. Image Process.*, 2017, pp. 900–904.
[41] X. Yang et al., "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 132.
[42] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
[43] Z. Guo, X. Zhang, C. Liu, X. Ji, J. Jiao, and Q. Ye, "Convex-hull feature adaptation for oriented and densely packed object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5252–5265, Aug. 2022.
[44] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.
[45] X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 18381–18394.
[46] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 677–694.
[47] H. Li, E. Tian, W. Zhang, Y. Li, and J. Cao, "Improving remote sensing object detection by using feature extraction and rotational equivariant attention," *Int. J. Remote Sens.*, vol. 45, no. 11, pp. 3789–3806, 2024.
[48] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.
[49] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general Gaussian heatmap label assignment for arbitrary-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1895–1910, 2022.
[50] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607117.
[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2999–3007.
[52] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. 2015 IEEE Int. Conf. Image Process.*, 2015, pp. 3735–3739.
[53] R. Faster, "Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 9199, no. 10.5555, pp. 2969239–2969250.
[54] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3500–3509.
[55] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
[56] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8788–8797.
[57] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605814.
[58] Y. Yu and F. Da, "Phase-shifting coder: Predicting accurate orientation in oriented object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13354–13363.
[59] Y. Yu, G. Zhang, W. Li, X. Wang, Y. Zhou, and J. Yan, "H2rbox: Horizontal box annotation is all you need for oriented object detection," in *Proc. Int. Conf. Learn. Representations*, 2023.
[60] C. Lyu et al., "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.
[61] M. Contributors, "MMYOLO: OpenMMLab YOLO series toolbox and benchmark," 2022. [Online]. Available: https://github.com/open-mmlab/mmyolo
[62] H. Yu, Y. Tian, Q. Ye, and Y. Liu, "Spatial transform decoupling for oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 7, pp. 6782–6790.
[63] J. Redmon, "Yolov3: An incremental improvement," 2018, *arXiv: 1804.02767*.
[64] X. Zhang et al., "HiViT: Hierarchical vision transformer meets masked image modeling," 2022, *arXiv:2205.14949*.
[65] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.
[66] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412.
[67] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[68] Q. Ming, L. Miao, Z. Zhou, X. Yang, and Y. Dong, "Optimization for arbitrary-oriented object detection via representation invariance loss," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8021505.

Yan Dong received the B.S. degree in automation and the M.S. degree in detection technology and automation equipment from Zhengzhou University, Zhengzhou, China.

She is currently an Associate Professor with the School of Information and Communication Engineering, Zhongyuan University of Technology, Zhengzhou. She has authored or coauthored more than 30 academic papers, authorized three patents, and published one book in recent years. Her research interests include artificial intelligence, pattern recognition, and surface defect detection based on machine vision.

Minghong Wei received the B.S. degree in computer science and technology from Tianjin University of Technology, Tianjing, China. He is currently working toward the M.S. degree in signal and information processing with the Zhongyuan University of Technology, Zhengzhou, China.

His research interests include remote sensing, computer vision, and machine learning.

Guangshuai Gao received the B.S. and M.S. degrees in applied physics and signal and information processing from the Zhongyuan University of Technology, Zhengzhou, China, in 2014 and 2017, respectively, and the Ph.D. degree in computer application technology from the School of Computer Science, Beihang University, Beijing, China, in 2022.

He is currently a Lecturer with the School of Information and Communication Engineering, Zhongyuan University of Technology. His research interests include image processing, digital machine learning, and remote sensing imagery interpretation.

Chunlei Li received the B.S. degree in computer science from Zhengzhou University, Zhengzhou, China, in 2001, the M.S. degree from Hohai University, Nanjing, China, in 2004, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2012.

He is currently a Professor with the School of Information and Communication Engineering, Zhongyuan University of Technology, Zhengzhou. He has authored or coauthored more than 50 technical articles and has authored two books in recent years. He has also obtained three patents. His research interests include machine vision for fabric defect detection, pattern recognition, and low-rank representation.

Zhoufeng Liu received the B.S. degree in semiconductor physics from Lanzhou University, Lanzhou, China, in 1982, and the M.S. degree in semiconductor technology from Beijing University of Technology, Beijing, China, in 1985 and the Ph.D. degree in signal and information processing from the Beijing Institute of Technology, Beijing, China, in 2004.

From 1990 to 1995, he was a Lecturer with Zhengzhou University, Zhengzhou, China. From 1998 to 2004, he was an Associate Professor. Since 2004, he has been a Professor with the Zhongyuan University of Technology, Zhengzhou. He was the Technical Leader and the Consultation Committee Member of Henan Province. His research interests include radar image processing and artificial intelligence.