Pattern Recognition 172 (2026) 112740

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

MSDP-Net: Multi-scale distribution perception network for rotating object detection in remote sensing ⋆

Ke Liu [a], Jian Zou [a], Wei Zhang [b], Qiang Li [a,*], QiWang [a,*]

[a] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China
[b] School of Computer Science, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China

ARTICLE INFO

Keywords:
Rotating object detection
Remote sensing images
Multi-scale perception
Kullback-Leibler divergence
Spatial adaptive feature modulation

ABSTRACT

The accurate detection of arbitrarily oriented objects in remote sensing imagery is a fundamental yet challenging task, critical for applications like urban planning and maritime surveillance. Devising a robust detector is primarily challenged by two factors. First, the deep downsampling in standard convolutional networks leads to severe feature loss for small, densely packed objects, making them difficult to distinguish from complex background clutter. Second, conventional angle-based regression suffers from inherent flaws, such as boundary discontinuity and a mismatch between the optimization loss and the final detection metric, which destabilizes training and degrades localization accuracy. To address these specific challenges, we propose MSDP-Net, a Multi-scale Distribution Perception Network. To counteract the information bottleneck caused by network downsampling, we introduce a Spatial Feature Enhancement Module (SFEM). This module establishes an explicit detail-preservation pathway, adaptively enhancing the weak features of small objects while suppressing background noise. Furthermore, to resolve the limitations of angle-based regression, we propose the RepPoints Gaussian Kullback-Leibler Divergence (RP-GKLD) loss. By modeling the bounding box as a 2D Gaussian distribution, this loss function provides a representation that is free from angular periodicity and boundary issues. Extensive experiments and ablation studies on challenging public benchmarks, including DIOR-R, DOTA-v1.0, DOTA-v1.5, and DOSR, validate the effectiveness of our design. The results demonstrate that MSDP-Net achieves a superior balance of accuracy and efficiency, outperforming many existing methods and establishing itself as a robust and highly competitive solution.

1. Introduction

The advancement of Earth observation technologies, particularly from high-resolution satellites and aerial platforms, has made remote sensing imagery a vital source of information for numerous applications [1]. A key enabling technology in this field is oriented object detection, which focuses on accurately localizing objects that exhibit arbitrary rotations [2]. In overhead imagery, objects such as ships, aircraft, and vehicles rarely align with the image axes. Consequently, traditional horizontal bounding box detectors struggle to precisely capture their spatial and geometric characteristics. This limitation necessitates the use of specialized rotating object detection techniques, which provide superior localization accuracy and are crucial for subsequent analysis [3].

Despite significant progress, rotating object detection in remote sensing is hindered by several persistent challenges that critically limit performance in real world scenarios. First, the detection of small and densely packed objects remains a primary hurdle. Due to the vast coverage area of remote sensing images, objects often have small pixel footprints, resulting in weak features that are susceptible to being lost in deep networks. In congested scenarios such as ports or large vehicle depots, severe occlusion and feature interference among adjacent objects drastically reduce recall and can lead to catastrophic missed detections. Second, interference from complex backgrounds and variable imaging conditions poses another significant obstacle; cluttered backgrounds can contain textures that mimic target features, making object-background separation difficult and leading to a high rate of false positives. Finally, the robust estimation of orientation and shape is fraught with difficulty. This is not merely a geometric inconvenience; the periodic nature of angle parameters introduces fundamental ambiguity, while the high diversity of target aspect ratios complicates the regression task.

⋆ This work was supported by the National Natural Science Foundationof China under Grant 62471394 and U21B2041.
* Corresponding authors.
E-mail addresses: liuke990203@mail.nwpu.edu.cn (K. Liu), zoujian@mail.nwpu.edu.cn (J. Zou), zhangwei707@mail.nwpu.edu.cn (W. Zhang), liqmges@gmail.com (Q. Li), crabwq@gmail.com (Q. Wang).

https://doi.org/10.1016/j.patcog.2025.112740
Received 24 July 2025; Received in revised form 11 November 2025; Accepted 11 November 2025
Available online 14 November 2025

The architectural development of rotating object detection has evolved alongside advancements in general object detection, representing the first wave of attempts to address these challenges. The breakthroughs of the deep learning era originated with two-stage CNN detectors, exemplified by the R-CNN [4] series, which established a foundation for high-accuracy detection. Subsequently, to meet real-time demands, single-stage detectors like YOLO [5,6] were introduced, unifying the detection process. The initial approaches to rotating object detection were direct extensions of these paradigms, with the core idea being the introduction of rotated anchors. However, this anchor-based strategy faced a common bottleneck: to cover all orientations and scales, the number of anchors increased exponentially, leading to immense computational overhead and sensitive hyperparameters.

To break free from this reliance on dense anchors, anchor-free methods emerged as a key turning point. These approaches, by defining objects through keypoints or regressing vertices directly, greatly enhanced model flexibility and efficiency. In recent years, the Transformer architecture has further revolutionized the field with its end-to-end, query-based paradigm, eliminating post-processing steps like non-maximum suppression (NMS). This was quickly adapted for rotation, in pioneering works like AO2-DETR [7]. Subsequent models like FFRViT [8] enhanced accuracy by refining multi-scale features, while others introduced novel matching strategies, such as the Hausdorff distance, and adaptive query denoising to further boost performance [9].

While these architectural evolutions significantly improved detector efficiency and flexibility, they did not fully resolve the difficulties inherent in the regression task itself. In fact, they exposed a deeper, underlying issue: the challenge of orientation regression, in particular, is exacerbated by fundamental flaws in traditional loss function designs. Losses such as Smooth ℓ_1 suffer from the Periodicity of Angle (PoA) and boundary discontinuity, which creates a fundamental inconsistency between the regression loss and the evaluation metric (IoU), thereby destabilizing the optimization process. To address this mismatch, researchers introduced loss functions based on Kullback-Leibler (KL) divergence [10]. However, these methods revealed critical drawbacks in practice, notably numerical instability and high computational complexity. Their fundamental limitation is their continued implicit reliance on the discontinuous angle parameter θ for distribution modeling, which simply propagates angular instability into the loss calculation and hinders their robustness and efficiency.

These persistent limitations in regression modeling clearly indicate that while detector architectures have matured, the underlying mechanism has not. The fundamental instability of the regression loss, a problem shared by anchor-based, anchor-free, and Transformer-based models alike, remains a significant bottleneck. The clear need for a more robust and stable solution to this core regression problem thus forms the primary motivation for our work. Therefore, developing a truly continuous and stable regression mechanism, one that is free from the instabilities of explicit angle parameters, is essential to finally overcome the limitations hindering high-precision rotated object detection.

To address these critical issues, we introduce **MSDP-Net (Multi-scale Distribution Perception Network)**, a framework engineered for robust rotating object detection in challenging remote sensing environments. MSDP-Net is designed to significantly improve the detection of small, oriented objects and ensure precise localization even in cluttered conditions. To achieve this, we focus on the following two aspects:

On one hand, to directly address the feature-level difficulties (small objects and background noise), we introduce the **Spatial Feature Enhancement Module (SFEM)**, which establishes a detail-preservation pathway from shallow to deep layers. Through cross-layer feature modulation, SFEM reinforces the feature representations of small objects while simultaneously suppressing interference from complex backgrounds, thereby significantly boosting overall feature robustness. On the other hand, to resolve the core regression instability (boundary discontinuity and angle ambiguity), we propose the **RepPoints Gaussian Kullback-Leibler Divergence (RP-GKLD) Loss**. By modeling the RepPoints set as a 2D Gaussian distribution and employing the Kullback-Leibler (KL) divergence as the similarity metric, our approach naturally circumvents the angle periodicity problem. This method also facilitates a more flexible and accurate representation of irregularly shaped target contours. Furthermore, its distribution-based modeling is inherently more robust to challenging cases such as truncated objects.

The main contributions of this work are summarized as follows:

- **MSDP-Net Framework**: We propose a novel and highly efficient network architecture tailored for the precise detection of rotating objects. By synergistically integrating our new feature enhancement and regression modules, MSDP-Net achieves a superior balance of accuracy and efficiency. It establishes state-of-the-art or highly competitive performance on challenging benchmarks while using significantly fewer parameters and computational resources than competing methods.
- **SFEM**: We propose a cross-level feature interaction mechanism designed to combat feature loss and enhance discriminability. By capturing fine-grained details and suppressing background noise, it is particularly effective for challenging scenarios involving blurred, densely packed, or weakly represented objects.
- **RP-GKLD Loss**: We propose a reformulated loss function that resolves the inherent flaws of conventional angle-based regression, such as boundary discontinuity and training instability. Our loss models the set of RepPoints as a 2D Gaussian distribution and employs the Kullback-Leibler (KL) divergence as a similarity metric. This approach evaluates matching in a continuous probability space, which naturally circumvents angle periodicity problems and improves the robustness of orientation prediction, particularly for irregularly shaped or truncated objects.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 details our proposed MSDP-Net and its core innovations: the Spatial Feature Enhancement Module (SFEM) and the RP-GKLD loss. In Section 4, we present extensive experiments, ablation studies, and comparisons with state-of-the-art methods to validate our approach. Finally, Section 5 summarizes our contributions and suggests future research directions.

2. Related work

This section reviews the evolution from general object detection to the more complex domain of rotated object detection. We then focus our analysis on three representative research directions: advanced feature enhancement, geometric modeling, and specialized loss functions.

2.1. *From axis-aligned to rotated object detection*

Deep learning has reshaped object detection, with early frameworks using axis-aligned bounding boxes (AABBs) achieving remarkable performance. Classic single-stage detectors such as the YOLO series [5] and two-stage methods like R-CNN [4] and Faster R-CNN [11] laid the foundation of modern detection pipelines. Subsequent improvements incorporated hierarchical feature aggregation [12] and refined loss formulations such as Focal Loss [13] to mitigate class imbalance and enhance robustness across scales.

However, the intrinsic limitation of AABBs lies in their inability to capture object orientation and spatial extent, especially in remote sensing imagery where elongated or arbitrarily oriented objects are common. The misalignment between object geometry and bounding box orientation often leads to localization errors and background interference. To address these issues, rotation-aware detection emerged as an extension of general object detection. Early attempts, such as R2CNN [14] and RRPN [15], integrated rotation proposals into two-stage frameworks, achieving higher localization precision but retaining structural complexity. These developments encouraged research toward

more direct and efficient approaches that better balance accuracy and computational cost.

2.2. Key technologies in rotated object detection

Progress in rotated object detection has been driven by advances in three key areas: (1) the design of enhanced feature representation to better capture spatial and semantic cues; (2) refined geometric modeling strategies for stable orientation estimation; and (3) tailored loss functions to achieve more consistent optimization with rotation-aware metrics.

2.2.1. Feature enhancement and representation

Feature representation plays a central role in detecting small, densely distributed, and occluded objects in aerial and remote sensing imagery, a challenge that extends to downstream tasks like multi-object tracking where handling occlusions and maintaining identity consistency is critical [16]. Similar challenges appear in related vision tasks such as multi-view stereo reconstruction, where visual consistency is important [17]. Early global modeling methods, including non-local operations, can capture long-range dependencies but are constrained by computational cost [18]. The introduction of adaptive convolutions, for example deformable convolution in RepPoints, improved spatial flexibility by sampling features according to object geometry, which is beneficial for oriented objects [19].

Recent work has concentrated on backbone and neck designs that improve scale adaptation and context integration. Large-kernel networks such as LSKNet-S and strip-based convolutions expand receptive fields to gather broader context [20,21], while lightweight attention and feature-decoupling techniques aim to raise efficiency without degrading accuracy [22,23]. This focus on enhancing critical features is explored in related domains as well; for example, in video saliency prediction, single feature enhancement pathways combined with temporal recurrence have been used to isolate salient information [24]. In remote sensing specifically, adaptive downsampling and scale-enhanced detection heads have been proposed to strengthen tiny-object representation and balance multi-scale features [25]. PolyBuild [26] further advances polygonal building contour extraction by introducing an end-to-end vertex-sequence modeling approach that ensures more accurate vectorization and eliminates post-processing overhead. Likewise, methods from multimodal and temporal fusion-for instance asymmetric decoding for RGB-thermal perception and pyramid-structured multi-scale fusion-highlight that flexible fusion and adaptive aggregation can improve robustness in complex scenes [27,28].

Compared with these multi-scale fusion strategies, SFEM emphasizes position-adaptive alignment between feature levels. Conventional fusion approaches often rely on hierarchical concatenation, summation, or channel-wise recalibration, which can treat spatial locations uniformly and introduce redundant information flow. SFEM instead applies spatially adaptive weighting and cross-level modulation to align local details with global context, allowing the network to strengthen weak small-object responses while suppressing background interference. This targeted alignment makes multi-scale interactions more precise and helps maintain feature consistency under varying orientations and scales.

Although these advances have improved general feature extraction, explicit rotation-aware representations remain an active direction. Approaches such as ReDet and RP-Net incorporate rotation-equivariant designs or polar transformations to embed geometric priors into features [29,30]. SFEM is complementary to these efforts: by improving spatial alignment and multi-scale detail preservation, it can provide stronger inputs for rotation-aware layers or polar-transformed representations, thereby supporting more stable detection of oriented objects without relying solely on heavy augmentation.

2.2.2. Geometric modeling and orientation estimation

Accurate geometric modeling is essential for describing object rotation and boundary alignment. Early methods extended horizontal detectors by adding rotated anchors [31]. Although this approach was intuitive, the number of anchors increased rapidly, raising computation and sensitivity to hyperparameter choices. To reduce redundancy, refinement-based approaches such as R3Det [32] and the RoI Transformer [33] introduced progressive feature alignment and transformation modules to better fit object geometry. Building on this, subsequent work further improved the RoI Transformer by incorporating multi-scale feature fusion strategies to enhance detection performance in remote sensing images.

Subsequent research explored anchor-free paradigms that predict object orientation directly through keypoints or vertex regression. Examples include CenterMap [34], PARDet with dynamic point set alignment [35], and Gliding Vertex [36], all of which improved detection flexibility. However, these methods faced challenges with angular periodicity, where small rotation errors could lead to inconsistent loss responses. To address this, some studies reformulated rotated boxes as two-dimensional Gaussian distributions [37], while linear Gaussian modeling [38] further stabilized regression by aligning distribution parameters. More recently, transformer-based frameworks such as AO2-DETR [7] adopted a set-prediction paradigm, removing hand-crafted components like anchor generation and Non-Maximum Suppression (NMS), which simplified the detection pipeline while maintaining rotation sensitivity.

2.2.3. Optimized loss functions

Loss design directly influences the accuracy and stability of rotated object regression. Traditional L1 or L2 losses are less effective because they do not reflect the discontinuities and angular periodicity of rotated boxes, causing misalignment with IoU-based evaluation metrics. Early formulations such as Modulated Angle Loss [39] alleviated these issues by smoothing angle transitions during training.

To further align optimization with geometric overlap, several studies introduced probabilistic or distribution-based loss functions. The Gaussian Wasserstein Distance Loss [40] and Probabilistic IoU Loss [41] reformulated regression as the alignment of probabilistic representations. Following this idea, Kullback-Leibler Divergence (KLD)-based regression [10] provided a more nuanced measure of distribution discrepancy, improving both optimization stability and localization accuracy.

Subsequent works have continued to refine loss design to better capture the geometric and statistical characteristics of rotated targets. Xu et al. [42] introduced a multiattention-based feature aggregation mechanism combined with dual focal loss to enhance the discrimination and robustness of rotation-sensitive features. Qin et al. [43] further optimized the loss landscape by proposing an adaptive symmetric loss, effectively mitigating instability during optimization. Li et al. [44] explored adaptive regulation of loss functions within a unified learning framework, enhancing model generalization in uncertain scenarios. More recently, FPDIoU [45] advanced rotated object detection by introducing a four-point distance-based loss function that ensures more accurate bounding box regression and faster convergence.

This continuous evolution of loss design, spanning from angle modulation to probabilistic modeling and geometric distance optimization, establishes a coherent theoretical foundation for our proposed RP-GKLD loss.

3. Methodology

Fig. 1 illustrates the overall architecture of our proposed **MSDP-Net**, a model specifically designed for robust rotating object detection in challenging overhead or varied-viewpoint imagery. The network is engineered to enhance the detection of small objects, address challenges in cluttered or dense scenes, and optimize oriented bounding box angle

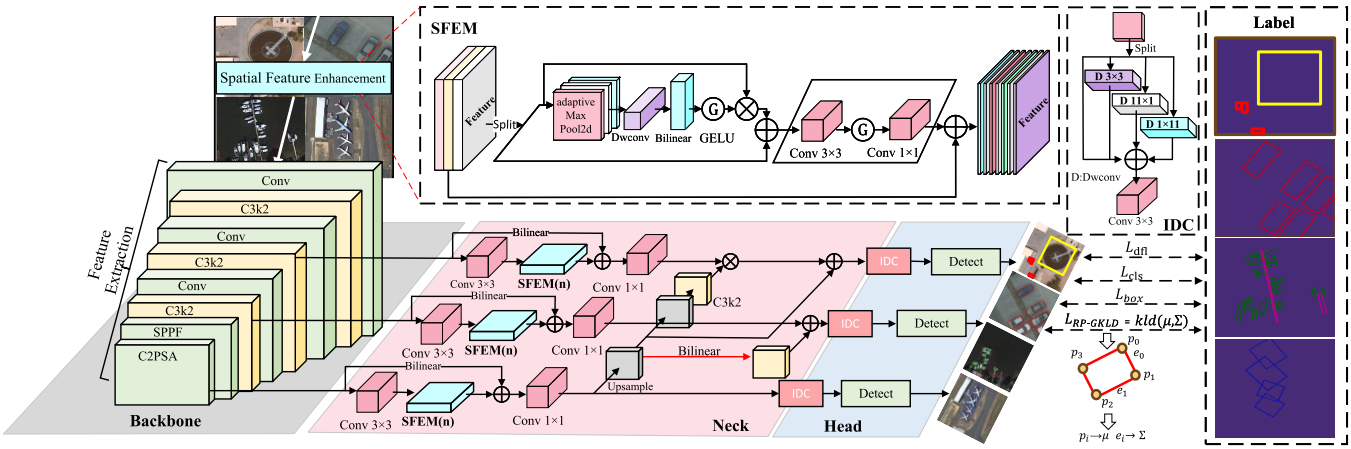

Fig. 1. Overview of the proposed MSDP-Net architecture. The Spatial Feature Enhancement Module (SFEM) interacts with feature extraction layers to perform small-target feature enhancement and mitigate background interference. The network is further optimized with the RepPoints Gaussian KL-Divergence (RP-GKLD) loss, which establishes a probabilistic distribution matching framework for robust rotation angle prediction.

regression. This is achieved by integrating multi-scale feature enhancement, efficient convolutional structures (such as the Inception Depthwise Convolution (IDC) module[46], which aids in efficient feature extraction, particularly for densely packed objects), and distribution-aware loss mechanisms. The primary innovations of this work, which will be detailed in subsequent sections, are SFEM for superior small object feature retention and background interference isolation, and RP-GKLD Loss for precise and stable orientation estimation. These components synergistically address key challenges in detecting rotating objects, including small object feature loss and unstable angle regression.

3.1. Spatial feature enhancement module

Remote sensing imagery presents numerous challenges. These include variable resolution, complex backgrounds, and degradations in imaging quality. Such factors hinder the detection of diverse objects, particularly small objects. This is especially true for objects in dense arrangements or those with blurred edges, which severely impacts detection accuracy and robustness. While traditional methods often struggle to effectively extract and fuse multi-scale spatial details from such imagery, techniques inspired by super-resolution offer a promising direction by mining details through refined spatial feature modulation [47,48]. Drawing from this approach, we propose SFEM. Its structure and integration within our MSDP-Net are illustrated in Fig. 1. SFEM integrates spatial adaptive feature modulation with convolutional channel mixing to mitigate these issues. This design significantly improves detail recovery and enhances discriminative performance for small objects found in complex scenes.

SFEM employs a dual-branch residual structure designed to synergistically preserve low-frequency information while concurrently restoring high-frequency details, thereby boosting small target detection. Given an input feature map $F \in \mathbb{R}^{C\times H\times W}$, the first branch performs 2× bilinear upsampling to generate the residual-connected baseline feature, F_0. This F_0 branch is crucial for retaining the fundamental low-frequency content of the original image:

$$F_0 = \text{Upsample}_{\text{bilinear}}(F)^{\epsilon}. \tag{1}$$

This baseline feature F_0 serves as a foundational reference for the subsequent enhancement performed by the second branch.

Recognizing that conventional bilinear interpolation is often insufficient for restoring high-frequency details, the second branch of SFEM module is dedicated to intensive spatial detail enhancement. This process begins when a 3×3 convolution projects the input feature F into an intermediate representation, F_{proj}. Subsequently, a series of n stacked Spatial Context Enhancement Blocks (SCEBs) process this representation.

Each SCEB consists of an Edge-aware Spatial Module (ESM) followed by a channel mixing stage. ESM focuses on multi-scale feature extraction and adaptive spatial attention. It first splits its input feature channel-wise into two parts, F_{E1} and F_{E2}, based on a factor σ. One part, F_{E2}, undergoes resolution reduction via adaptive max pooling to a target resolution of $h/8\times w/8$. This is followed by feature modulation using a depthwise convolution, a step that yields the multi-scale contextual feature F_{E3}. This contextualized feature is then upsampled and used to modulate the other feature part, F_{E1}, through element-wise multiplication after a GELU activation. Finally, the resulting modulated features and the original F_{E1} are concatenated and refined by a 1×1 convolution to produce the ESM output, F_{ESM}, which enhances spatial details and texture recovery.

To complement the spatial focus of the ESM, the subsequent channel mixing stage processes the feature map F_{ESM} to strengthen global semantic information and inter-channel correlations. This is achieved by expanding channel dimensions with a 3×3 convolution and a GELU activation, followed by channel reduction via a 1×1 convolution. This step yields the final SCEB-enhanced features F_{SCEB}.

To provide a clearer and more intuitive overview of this process, the complete structure of the SCEB is summarized in Algorithm 1.

Algorithm 1 Spatial context enhancement block (SCEB).

Require: Input feature map $F_{in} \in \mathbb{R}^{C\times H\times W}$
Ensure: Output feature map $F_{out} \in \mathbb{R}^{C\times H\times W}$
Parameters: Channel split ratio σ, target pooling size $(h_p, w_p) = (H/8, W/8)$
Edge-aware Spatial Module (ESM):
$F_{E1}, F_{E2} \leftarrow \text{ChannelSplit}(F_{in}, \text{ratio} = \sigma)$
$F_{E3\_pooled} \leftarrow \text{AdaptiveMaxPool}(F_{E2}, \text{target\_size} = (h_p, w_p))$
$F_{E3} \leftarrow \text{DepthwiseConv}(F_{E3\_pooled})$
$F_{E3\_upsampled} \leftarrow \text{Upsample}(F_{E3}, \text{target\_size} = \text{shape}(F_{E1}))$
$F_{modulated} \leftarrow \text{GELU}(F_{E3\_upsampled}) \times F_{E1}$
$F_{concat} \leftarrow \text{Concat}(F_{modulated}, F_{E1})$
$F_{ESM} \leftarrow \text{Conv1x1}(F_{concat})$
Channel Mixing Stage:
$F_{expanded} \leftarrow \text{GELU}(\text{Conv3x3}(F_{ESM}))$
$F_{out} \leftarrow \text{Conv1x1}(F_{expanded})$
return F_{out}

After passing through n stacked SCEB blocks, the deep features from this second branch, denoted F^n_{SCEB}, encapsulate the enhanced high-

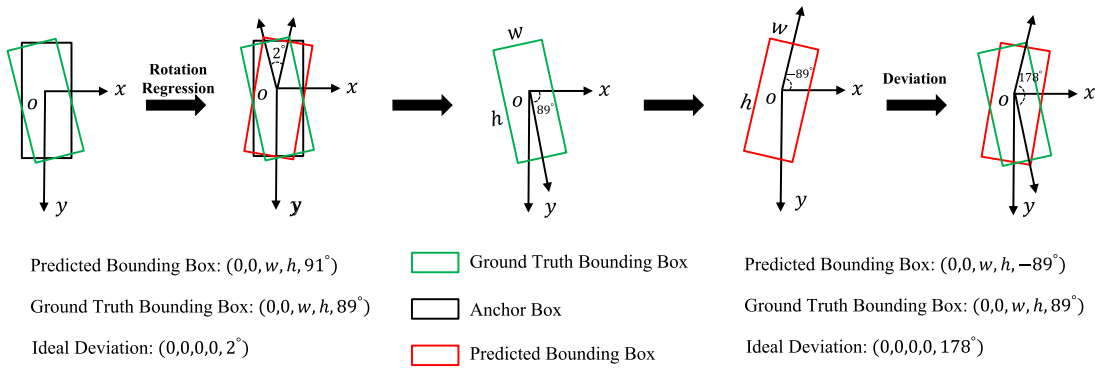

Fig. 2. Conceptual illustration of the boundary discontinuity in traditional angle regression.

frequency details. These features are then processed by a 1×1 convolution and subsequently upsampled using the PixelShuffle operation to match the resolution of F_0. Finally, this refined high-resolution component is combined with the baseline F_0 from the first branch via residual addition to produce the SFEM-enhanced feature map, F_{SFEM}:

$$F_{\text{SFEM}} = F_0 + \text{PixelShuffle}(\text{Conv}_{1\times1}(F^n_{\text{SCEB}})). \tag{2}$$

This dual-branch architecture, with its detailed multi-level feature modulation within the SCEBs, ensures comprehensive feature enhancement critical for robust object detection in challenging imagery.

3.2. RepPoints Gaussian Kullback-Leibler divergence loss

In dense rotated object detection, regression-based methods often suffer from principal axis confusion and distribution misalignment, primarily caused by directional ambiguity and feature coupling. Consequently, traditional loss functions, such as Smooth ℓ_1 and its IoU-based variants, perform suboptimally.

This directional ambiguity stems from the Periodicity of Angle (PoA) and the resultant boundary discontinuity. As illustrated in Fig. 2, this problem is particularly acute under angle normalization conventions, such as the long-edge definition which constrains the range to $[-90°, 90°)$. A minuscule angular change can lead to a drastic jump in the parameter space. For instance, a predicted box at 89° and a ground-truth box at −89° are virtually identical visually, as indicated by a high IoU, yet their ℓ_1 loss incurs a disproportionately large penalty of 178°. This fundamental inconsistency between the regression loss, such as ℓ_1, and the evaluation metric (IoU) creates a highly non-convex and discontinuous loss landscape at the angular boundaries. This misalignment, where high visual overlap still incurs an extreme loss penalty, destabilizes the optimization process and hinders the model from converging to high-accuracy localization.

To resolve this mismatch, researchers introduced loss functions based on Kullback-Leibler (KL) divergence [10] as a significant conceptual advancement. Theoretically, by modeling bounding boxes as 2D Gaussian distributions, the distributions represented by 89° and −89° are highly similar, yielding a very low KL divergence, which aligns with the high IoU. However, practical implementations of these methods revealed critical drawbacks, notably numerical instability and high computational complexity. Their fundamental limitation is their continued reliance on the discontinuous angle parameter θ to construct the covariance matrix Σ (i.e., $\Sigma = R(\theta)\Lambda R(\theta)^T$). This discontinuous θ remains part of the computation graph, and its instability propagates directly into the KLD calculation.

Our core motivation is to completely sever the dependency on the explicit angle parameter θ, thereby achieving a truly continuous, end-to-end probabilistic distribution regression to mitigate all the aforementioned limitations. To this end, we propose RP-GKLD Loss.

Specifically, RP-GKLD Loss adopts a probability distribution alignment-based geometric representation optimization framework. It maps the RepPoints-predicted polygon $P = [P_0, P_1, P_2, P_3] \in \mathbb{R}^{4\times2}$ and the ground-truth rotated bounding box $P_t \in \mathbb{R}^{4\times2}$ to the Gaussian distribution space, leveraging the Kullback-Leibler divergence to measure the distribution difference between them. This enables fine-grained supervision of rotation-sensitive features. The computational process includes three stages: geometric representation transformation, distribution difference measurement, and loss function optimization. The method first maps the polygon predicted by RepPoints and the ground-truth rotated bounding box to the Gaussian distribution space. For any set of polygon vertices, the mean vector of the Gaussian distribution is calculated based on the geometric center:

$$\mu = \frac{1}{4}\sum_{i=0}^{3} P_i. \tag{3}$$

Subsequently, a direction-sensitive covariance matrix is constructed based on the adjacent edge vectors $e_1 = P_1 - P_0$ and $e_2 = P_2 - P_1$. Through principal axis decomposition, a rotation matrix R and a scaling matrix Λ are constructed:

$$R = \begin{bmatrix} \frac{e_1}{\|e_1\|} & \frac{e_1^\perp}{\|e_1\|} \end{bmatrix}, \quad \Lambda = \frac{1}{4L^2}\text{diag}(\| e_1 \|^2, \| e_2 \|^2), \tag{4}$$

where $e_1^\perp$ represents the orthogonal component, and $L = 3$ is the normalization coefficient used to control the proportional relationship between the Gaussian distribution range and the actual target size. Then, a numerically stable covariance matrix is generated through affine transformation:

$$\Sigma = R\Lambda R^T + 10^{-6}I. \tag{5}$$

This single matrix simultaneously encodes three key properties. First, it represents the rotation angle of the target, as determined by the direction basis of R. Second, it captures scale information, which is characterized by the diagonal elements of Λ. Finally, it ensures numerical stability by incorporating a small identity matrix, $10^{-6}I$, to avoid singularity. To further enhance computational robustness, a symmetry constraint is applied during the construction of the covariance matrix:

$$\Sigma \leftarrow \frac{1}{2}(\Sigma + \Sigma^T). \tag{6}$$

This operation enforces the symmetry of the matrix, avoiding asymmetry issues caused by floating-point errors. After mapping the predicted distribution (μ_p, Σ_p) and the ground-truth distribution (μ_t, Σ_t), the Kullback-Leibler divergence is used to quantify their difference:

$$\text{KLD} = \frac{1}{2}\left[\text{tr}\left(\Sigma_t^{-1}\Sigma_p\right) + \left(\mu_t - \mu_p\right)^T\Sigma_t^{-1}\left(\mu_t - \mu_p\right) - 2 + \ln\frac{\det\Sigma_t}{\det\Sigma_p}\right], \tag{7}$$

where the inverse term $\text{tr}\left(\Sigma_t^{-1}\Sigma_p\right)$ reflects scale differences, the quadratic term $(\mu_t - \mu_p)^T\Sigma_t^{-1}\left(\mu_t - \mu_p\right)$ measures the center point offset, and the determinant logarithmic term $\ln\frac{\det\Sigma_t}{\det\Sigma_p}$ evaluates directional consistency. To avoid numerical instability during matrix inversion, a regularization

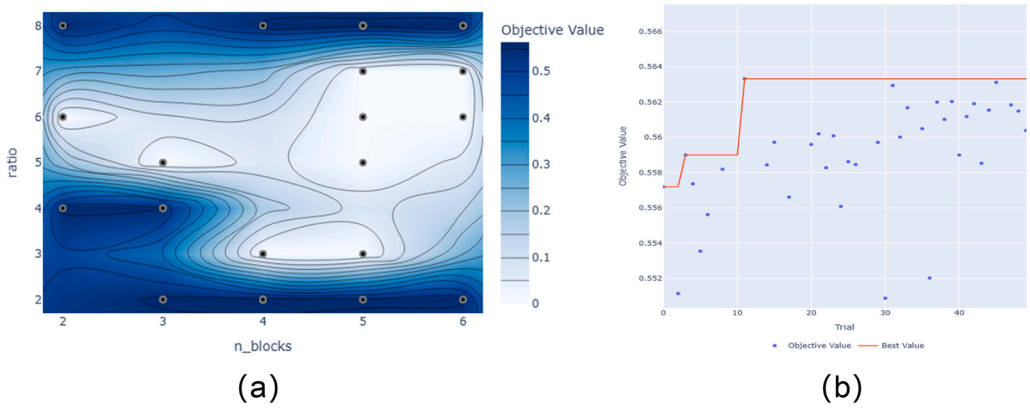

Fig. 3. Bayesian Optimization for SFEM hyperparameters (n: *n_blocks*, $\sigma = 1/ratio$). (a) Contour plot of the mAP landscape. (b) Optimization history showing convergence to n_blocks = 4, ratio = 2.

strategy is adopted:

$$\Sigma^{-1} = (\Sigma + 10^{-6} I)^{-1}. \tag{8}$$

Finally, the KLD value is transformed into an optimizable loss through a nonlinear mapping function:

$$L_{\text{RP-GKLD}} = 1 - \frac{1}{2 + \sqrt{\text{KLD}}}. \tag{9}$$

This function compresses the divergence value into the interval $[0, 1]$. When the distributions are perfectly aligned (KLD $\to 0$), the loss approaches zero; conversely, for significant deviations (KLD $\to \infty$), the loss saturates at 1. This property effectively makes the model more sensitive to minor geometric deviations.

The overall training objective for MSDP-Net combines our proposed $L_{\text{RP-GKLD}}$ with standard baseline losses from the underlying detection framework. The total loss function, L_{total}, is formulated as a weighted sum of four components: the baseline bounding box regression loss (L_{box}), classification loss (L_{cls}), distribution focal loss (L_{dfl}), and our proposed RepPoints Gaussian KL-Divergence loss ($L_{\text{RP-GKLD}}$):

$$L_{\text{total}} = \lambda_{\text{box}} L_{\text{box}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{dfl}} L_{\text{dfl}} + \lambda_{\text{RP-GKLD}} L_{\text{RP-GKLD}}, \tag{10}$$

here, L_{box}, L_{cls}, and L_{dfl} and their respective hyperparameters λ_{box}, λ_{cls}, and λ_{dfl} are adopted from the baseline model and retain their original values. The hyperparameter $\lambda_{\text{RP-GKLD}}$ for our proposed loss term was determined to be 0.2 through Bayesian hyperparameter optimization, effectively balancing its contribution with the other loss components during training.

4. Experiment

In this section, we experimentally validate our proposed MSDP-Net. We compare its performance against several advanced methods on several challenging public datasets. Furthermore, we conduct in-depth ablation studies to verify the effectiveness of our core components: the SFEM and the RP-GKLD Loss.

4.1. Dataset

To evaluate the performance of the proposed modules and the overall network, we selected several popular and challenging datasets for rotated object detection. The evaluation is grounded on large-scale benchmarks such as DIOR-R [49], which contains 23,463 images and over 192,000 object instances across 20 common classes. We further test our model on the demanding DOTA benchmark, utilizing both version 1.0 and 1.5. DOTA-v1.0 [50] consists of 2806 high-resolution images with 188,282 instances across 15 categories, presenting significant variations in object size, orientation, and aspect ratio. The more challenging DOTA-v1.5 [50] extends this by adding a new container crane category and substantially increasing the number of small objects (under 10 pixels), making it ideal for assessing performance in dense scenes. To specifically evaluate fine-grained recognition capabilities, the DOSR dataset [51] was also included, which provides 1066 images focused on distinguishing between different ship classes. Collectively, this suite of datasets provides diverse and challenging scenarios to thoroughly test the robustness and accuracy of our proposed method.

4.2. Implementation details

Our MSDP-Net utilizes the backbone from YOLOv11 for feature extraction, initializing the network with its publicly available pretrained weights. All models were trained for a maximum of 40 epochs, with an early stopping strategy implemented to prevent overfitting; specifically, training is terminated if the validation mAP does not improve for 20 consecutive epochs. We utilized a cosine annealing schedule for the learning rate (initial 0.01, minimum learning rate $1e-4$), with a weight decay of 0.0005 and momentum of 0.9. Data augmentation techniques included RandAugment and random erasing (probability 0.4). All experiments were conducted on a single NVIDIA RTX 3090 GPU with 24 GB of memory. Due to the large size of images in the employed datasets, data cropping was performed: for DOTA-v1.0, multiscale $(0.5, 1.0, 1.5)$ training and testing involved cropping images to 1024×1024 pixels with a 500-pixel overlap. For DOTA-v1.5, single-scale processing was used, with images cropped to 1024×1024 patches using a 200-pixel overlap. Batch sizes were set as follows: 4 for DIOR-R, 4 for DOTA-v1.0, 8 for DOTA-v1.5, and 8 for DOSR. Mean Average Precision (mAP) was the primary evaluation metric.

To determine the optimal configuration for SFEM, we performed hyperparameter optimization for the number of stacked SCEB blocks (n) and the channel split ratio (σ). We employed Bayesian Optimization for its efficiency in exploring the parameter space. The search spaces were $n \in \{2, 4, 6, 8\}$ and ratio $= 1/\sigma \in \{1, \ldots, 10\}$, with the objective to maximize mAP on the DIOR-R validation set. Each configuration was evaluated after 10 epochs of training. As visualized in Fig. 3, this process identified the optimal parameters as $n = 4$ and $ratio = 2$ ($\sigma = 1/2$), achieving a validation mAP of 0.563. These values were used in our main experiments.

4.3. Ablation study

In this section, we conduct ablation experiments on the DIOR-R dataset to systematically evaluate the effectiveness of each key proposed component within our MSDP-Net. The DIOR-R dataset, while a remote

Table 1
Ablation studies of each component in MSDP-Net on the DIOR-R dataset. In each column, **bold** indicates the best result, while underlined values denote the second-best results.

Baseline	SFEM	IDC	RP-GKLD	APL	APO	BF	BC	BR	CH	DAM	ESA	ETS	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP
✓				**92.05**	59.33	89.55	92.49	50.24	75.96	43.74	76.90	61.83	81.18	73.33	61.80	60.32	94.28	65.50	91.72	94.50	70.12	64.85	85.45	74.26
✓	✓			90.51	61.87	89.22	92.47	54.03	77.59	38.41	85.12	65.61	82.38	75.63	62.01	65.41	94.96	66.85	91.96	94.49	66.47	**65.38**	89.00	75.50
✓		✓		90.48	60.96	88.87	92.28	52.37	75.96	40.90	83.13	63.45	84.33	77.12	58.18	63.49	93.84	66.74	89.19	93.79	70.17	62.29	89.71	74.80
✓			✓	90.59	64.16	89.50	92.71	53.44	76.78	48.91	81.88	65.13	83.71	**78.57**	61.89	63.92	94.59	63.72	**92.03**	93.74	70.37	64.45	87.50	75.88
✓	✓	✓		89.74	61.61	88.56	92.84	53.44	78.45	47.61	81.90	69.27	86.48	75.98	59.14	65.40	94.08	68.74	89.55	94.25	70.26	62.50	90.09	75.99
✓	✓		✓	90.56	**66.71**	**89.83**	92.63	**56.54**	77.55	34.81	83.12	68.62	83.83	76.10	**62.73**	**66.61**	**95.22**	66.36	91.85	**94.52**	**72.63**	65.36	89.90	76.27
✓		✓	✓	91.65	61.78	89.11	92.23	53.66	76.18	46.31	83.38	68.38	**86.57**	78.11	58.39	65.99	93.39	**72.67**	90.22	93.52	67.13	61.53	**90.13**	76.03
✓	✓	✓	✓	91.31	60.67	88.32	**93.10**	53.19	**79.65**	**52.37**	**85.87**	**72.31**	86.49	77.10	60.44	66.46	94.31	70.24	89.25	94.11	71.79	62.00	89.98	**76.95**

Table 2
Performance comparison with prominent contemporary methods on the DOTA-v1.0 dataset. Results are reported as mean Average Precision (mAP), per-category Average Precision (AP), number of parameters (Params), and FLOPs. In each column, **bold** indicates the best result, while underlined values denote the second-best results.

Method	FPS↑	Params(M)↓	FLOPs(G)↓	PL↑	BD↑	BR↑	GTF↑	SV↑	LV↑	SH↑	TC↑	BC↑	ST↑	SBF↑	RA↑	HA↑	SP↑	HC↑	mAP↑
EMO2-DETR [52]	–	74.3	304.0	87.99	79.46	45.74	66.64	78.90	73.90	73.30	90.40	80.55	85.89	55.19	63.62	51.83	70.15	60.04	70.91
CenterMap [34]	–	41.1	198.0	89.02	80.56	49.41	61.98	77.99	74.19	83.74	89.44	78.01	83.52	47.64	65.93	63.68	67.07	61.59	71.59
AO2-DETR [7]	–	74.3	304.0	86.01	75.92	46.02	66.65	79.70	79.93	89.17	90.44	81.19	76.00	56.91	62.45	64.22	65.80	58.96	72.15
SCRDet [53]	–	41.9	–	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
R3Det [32]	–	41.9	336.0	89.50	81.20	50.05	66.10	70.90	78.70	78.20	90.80	85.30	84.20	61.80	63.80	68.20	69.80	67.20	73.70
Roi Trans. [33]	–	55.1	225.3	89.01	77.48	51.64	72.07	74.43	77.55	87.76	90.81	79.71	85.27	58.36	64.11	76.50	71.99	54.06	74.05
G.V. [54]	–	41.1	198.0	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
ReDet [29]	–	31.6	–	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25
LSKNet-S [20]	24.5	31.0	161.0	89.66	**85.52**	57.72	75.70	74.95	78.69	88.24	90.88	86.79	86.38	66.92	63.77	77.77	74.47	64.82	77.49
DecoupleNet-D2 [23]	21.8	23.3	142.4	89.37	83.25	54.29	75.51	79.83	84.82	88.49	90.89	87.19	86.23	66.07	65.53	77.23	72.34	72.34	78.04
PKINet-S [55]	10.3	30.8	184.6	89.72	84.20	55.81	77.63	80.25	84.45	88.12	90.88	87.57	86.07	66.86	70.23	77.47	73.62	62.94	78.39
RTMDET-FPDIoU [45]	26.6	24.67	99.76	88.50	83.50	60.32	78.28	**80.50**	84.50	87.50	89.50	85.50	86.50	**72.75**	**71.92**	78.99	79.50	75.25	78.63
FFRViT [8]	–	–	–	88.93	84.49	60.75	**81.57**	79.18	86.06	89.53	89.50	85.18	**87.82**	69.56	70.32	77.45	**79.68**	**79.79**	78.63
LWGANet-L2 [22]	17.4	29.2	159.1	89.68	82.82	55.17	77.48	80.13	84.93	88.22	90.90	87.82	85.64	65.48	67.49	77.10	73.12	73.57	**78.64**
MSDP-Net(Ours)	**31**	**6.63**	**18.4**	**95.03**	82.75	**64.42**	74.99	72.29	**86.85**	**90.48**	**94.61**	64.38	83.02	67.61	71.37	**85.94**	74.83	70.73	78.62

sensing benchmark, includes various objects such as vehicles and ships, making it relevant for assessing performance on these common object types. The results are summarized in Table 1.

1) Analysis of MSDP-Net: MSDP-Net, which integrates our SFEM and RP-GKLD Loss with an enhanced backbone architecture, is designed for efficient localization and precise modeling of rotating small objects in challenging overhead imagery. As demonstrated in Table 1, this complete MSDP-Net architecture achieves a peak mAP of 76.95 % on the DIOR-R dataset, affirming the efficacy of the overall design.

2) Analysis of SFEM: SFEM is introduced to improve feature representation, particularly for **densely arranged** objects and those in cluttered scenes. It employs a cross-level feature interaction mechanism with two parallel pathways: one for global context-awareness and another for local detail enhancement. This design allows for the synergistic fusion of texture and semantic information across the feature pyramid. As demonstrated in Table 1, integrating SFEM into the baseline model boosts the mAP from 74.26 % to 75.50 %. This significant improvement underscores the ability of SFEM to mitigate adverse background interference and enhance feature robustness for these challenging dense objects.

3) Analysis of RP-GKLD Loss: To address persistent challenges in achieving precise and stable orientation estimation for rotated objects, we proposed RP-GKLD Loss. This loss function reframes oriented bounding box regression by leveraging the geometric invariance of Gaussian distributions and a probabilistic matching paradigm, thereby effectively mitigating issues related to angle parameter boundary sensitivity. When RP-GKLD Loss is incorporated into SFEM-enhanced baseline (Backbone + SFEM), a further increase in mAP from 75.50 % to 76.95 % is observed (Table 1). This result validates the significant contribution of our probabilistic modeling strategy towards enhancing orientation prediction accuracy and overall detection performance.

4.4. Quantitative analyses

This section provides a comprehensive quantitative evaluation to validate the efficacy of the proposed MSDP-Net and to offer deep insights into its performance characteristics. We benchmark MSDP-Net against a range of established and leading-edge models on three challenging remote sensing datasets: DOTA-v1.0, DOTA-v1.5, and DOSR. The analysis focuses on key metrics including mean Average Precision (mAP), model parameters (Params), and computational complexity (FLOPs), aiming to underscore the architectural superiority and practical value of our design.

Our benchmark results reveal that MSDP-Net achieves an exceptional balance between accuracy and efficiency. On the widely-used DOTA-v1.0 dataset (Table 2), it achieves a highly competitive mAP of 78.62 %, which is nearly on par with top-scoring models like LWGANet-L2 and FFRViT, demonstrating excellent performance.The core advantage of MSDP-Net lies in its unparalleled computational efficiency. Our model has only 6.63M parameters, 18.4G FLOPs, and achieves an inference speed of 31 FPS; all three key efficiency metrics rank first among all compared methods. Compared to the slightly higher-scoring LWGANet-L2, MSDP-Net reduces parameter count and computational load by approximately 4.4 times and 8.6 times, respectively. This remarkable efficiency stems from a lightweight yet powerful architecture that utilizes efficient convolutional structures for potent feature extraction without a heavy computational burden.

A detailed analysis of per-category performance shows that MSDP-Net excels at detecting densely arranged objects and those with regular geometric features. It achieves the highest AP scores in a total of six categories, including Plane (PL), Bridge (BR), Large Vehicle (LV), Ship (SH), Tennis Court (TC), and Harbor (HA). Our model shows a significant lead, particularly in categories often characterized by dense arrangements, such as Large Vehicle (LV), Ship (SH), and Plane (PL). This success is attributed to our innovative SFEM , which effectively filters

Table 3
Performance comparison (mAP%) with prominent contemporary methods on the DOTA-v1.5 dataset. **Bold** indicates the best result, while underlined values denote the second-best results.

Metric	ReDet [29]	FSDet [56]	LSKNet-S [20]	SOOD [57]	PKINet-S [55]	LWGANet [22]	Strip-CNN-S [21]	MSDP-Net(Ours)
mAP	66.86	68.37	70.26	70.39	71.47	71.72	72.27	**72.39**

Table 4
Performance comparison with prominent contemporary methods on the DOSR dataset. **Bold** indicates the best result, while underlined values denote the second-best results.

Metric	R2CNN [14]	RRPN [15]	SCRDet [53]	RetinaNet-O [13]	R3Det [32]	SCRDet++ [58]	RSDet [39]	ReDet [29]	Yolov12-s [59]	EIRNet [51]	MSDP-Net (Ours)
mAP (%)	45.58	50.58	54.29	36.70	52.66	56.33	45.15	57.32	60.00	61.39	**66.05**
Speed (s)	0.3513	0.3703	0.2602	0.2686	0.2043	0.3513	0.3612	0.0833	0.0797	0.2708	**0.0776**

background clutter and enhances the core features of these objects, often found in complex or dense environments. Furthermore, this enhanced feature representation is leveraged by our RP-GKLD Loss function, which enables precise angular prediction to accurately capture the orientation of structured objects like Ships (SH) and Bridges (BR).

However, this detailed analysis also allows us to objectively identify the model's shortcomings, which are consistent with the trade-offs of our efficiency-focused design. The model's performance is less competitive in two types of categories: first, extremely small-scale Small Vehicles (SV), and second, Baseball Courts (BC), which exhibit high intra-class similarity and are easily confused with the background. In these two categories, our model's scores show a significant gap compared to the best-performing methods for those specific classes. This suggests that while our lightweight backbone and SFEM are highly effective for general and structured objects, achieving top-tier performance on these specific, challenging sub-categories would likely require the heavier computational overhead of other models. This outcome aligns with our primary goal: to achieve a balanced, excellent-level overall mAP at a fraction of the computational cost, rather than optimizing for every individual category.

The architectural strengths of MSDP-Net are further validated on the more demanding DOTA-v1.5 dataset (Table 3), which is known for its higher prevalence of extremely small and densely packed objects. In this challenging scenario, MSDP-Net surpasses all competitors, achieving a top-ranked mAP of 72.39 %. This leading performance provides definitive evidence for the effectiveness of SFEM. By adeptly enhancing fine-grained details and suppressing background clutter, SFEM empowers the model to maintain robust detection capabilities even when objects are diminutive and clustered.

Beyond general object detection, MSDP-Net also demonstrates effective performance in specialized, fine-grained recognition tasks. On the DOSR ship dataset (Table 4), the model achieves a mAP of 66.05 %, the highest among the compared methods. This result indicates that the model can effectively learn discriminative features to distinguish between subtle variations in object appearance. Regarding operational speed, the model achieves an inference time of 0.0776 s per image, ranking fastest among the listed methods. This balance of precision and efficiency confirms that MSDP-Net is a practical solution for applications where real-time performance is a critical requirement.

In conclusion, the extensive evaluations conducted across diverse and challenging datasets validate MSDP-Net as a highly effective and efficient framework. The model consistently delivers state-of-the-art or competitive accuracy while substantially reducing parameters and computational cost. This favorable balance between precision and efficiency suggests that MSDP-Net is a promising framework for a range of demanding oriented object detection tasks.

4.5. Visualization analyses

In this section, we provide qualitative visual results to further demonstrate the effectiveness of the proposed MSDP-Net and to offer intuitive insights into its operational mechanisms. We present typical detection results on challenging imagery and visualize feature heatmaps to highlight the contributions of key components in our network.

1) Detection Results Visualization: To qualitatively assess the capabilities of MSDP-Net in challenging scenarios, Fig. 4 presents a visual comparison between our model, MSDP-Net, and several other methods, namely LWGANet, LEGNet-S, and Strip-R-CNN. The examples are drawn from the DOTA-v1.0 dataset. These scenes were specifically selected to illustrate comparative performance in several demanding contexts, including: the detection of small objects; the discernment of objects within cluttered backgrounds; the effective localization of partially occluded objects; the handling of truncated objects; and the precise regression of objects with a large aspect ratio. These visualizations intuitively demonstrate the robustness of MSDP-Net and its superior performance in addressing specific detection challenges.

2) Feature Heatmap Analysis: To offer deeper insight into the internal mechanisms of MSDP-Net, we present comparative feature heatmaps in Fig. 5 to visualize the impact of our proposed components on feature representation. This figure provides a visual ablation study, comparing feature heatmaps from a standard baseline model against two enhanced versions: one with only the SFEM module added, and the other being our full MSDP-Net. This comparison provides qualitative evidence of the advantages offered by our method, clearly demonstrating superior feature focus and background suppression for challenging objects, such as those in dense arrangements, with small scales, or with large aspect ratios.

Fig. 5 provides a comparative analysis focusing on challenging scenarios, such as densely packed vehicles and small objects within complex backgrounds. The visualizations, which include magnified insets of select small objects, clearly demonstrate that MSDP-Net enhances feature responses for these difficult-to-detect objects and suppresses surrounding clutter more effectively than the baseline. Subsequently, Fig. 5 compares MSDP-Net with the baseline in handling objects characterized by extreme aspect ratios (e.g., large vehicles or bridges). These heatmaps showcase the pronounced effectiveness of MSDP-Net in substantially reducing background noise while maintaining strong, precise feature localization on such elongated objects. This highlights the robust geometric feature representation achieved by the model. Collectively, these comparative visualizations reinforce our quantitative results, visually confirming that MSDP-Net more effectively focuses on informative object regions and exhibits superior adaptability to complex geometries compared to the baseline approach.

In summary, the visualization analyses provide compelling qualitative support for the quantitative results. The detection examples underscore the robustness and accuracy of MSDP-Net in complex visual environments, particularly for small, dense, and arbitrarily oriented objects. Furthermore, the feature heatmaps offer valuable insights into how different components within MSDP-Net contribute to enhanced perception, thus validating our design choices.

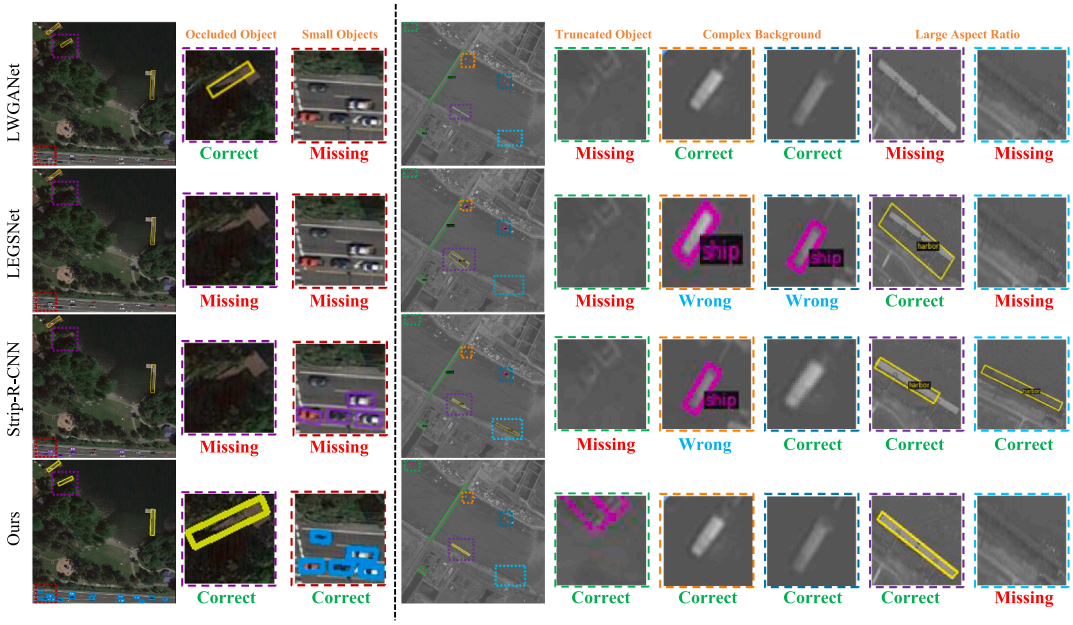

Fig. 4. Comparative detection results of **MSDP-Net (Ours)** against LWGANet, LEGNet-S, and Strip-R-CNN on challenging DOTA-v1.0 scenarios. The visualizations highlight performance on various challenging cases, including small objects, objects in complex backgrounds, partially occluded objects, truncated objects, and objects with a large aspect ratio.

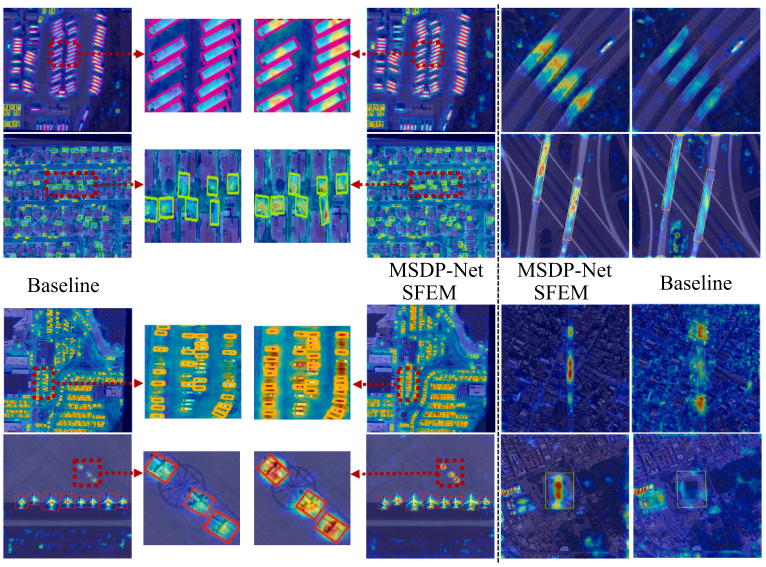

Fig. 5. Visualizing the impact of SFEM and MSDP-Net on feature representation. This figure compares feature heatmaps from a baseline model against two enhanced versions: one with only the SFEM module added, and the other being our full MSDP-Net. The results clearly show superior feature focus and background suppression for challenging objects, such as those in dense arrangements, with small scales, or with large aspect ratios.

5. Conclusion

We present MSDP-Net, a Multi-scale Distribution Perception Network for robust rotating object detection in remote sensing imagery. Our network introduces two key innovations: a Spatial Feature Enhancement Module (SFEM) and a RepPoints Gaussian Kullback-Leibler Divergence (RP-GKLD) Loss. SFEM enhances feature representation for small, densely packed objects, while the RP-GKLD Loss provides a probabilistic framework for accurate and stable orientation estimation. Comprehensive evaluations on challenging public benchmarks, including DOTA-v1.0, DOTA-v1.5, and DOSR, validate that MSDP-Net achieves an excellent balance between performance and efficiency.

Despite its strong performance, we recognize several limitations that open avenues for future research. Firstly, the added computational complexity of our proposed modules may pose challenges for real-time, on-device deployment. Secondly, the model's performance, while robust, can degrade in scenarios with extremely dense and heavily occluded objects. Finally, its generalization capability to disparate domains without significant fine-tuning remains an open question.

To address these points, our future work will focus on several key directions. To tackle the computational overhead, we plan to investigate

model compression techniques, such as quantization and pruning, to develop a lightweight version of MSDP-Net. To better handle complex scenes, we will explore incorporating relationship modeling to explicitly reason about interactions between overlapping instances. Furthermore, we aim to enhance the model's adaptability by leveraging unsupervised domain adaptation and self-supervised learning paradigms. Extending the framework to 3D object detection for advanced spatial analysis also remains a promising long-term goal.

CRediT authorship contribution statement

Ke Liu: Writing – original draft; **Jian Zou:** Writing – review & editing; **Wei Zhang:** Writing – review & editing; **Qiang Li:** Writing – review & editing; **QiWang:** Supervision, Funding acquisition.

Data availability

The data that has been used is confidential.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Li, H. Taubenböck, X.X. Zhu, Identification of the potential for roof greening using remote sensing and deep learning, Cities 159 (2025) 105782.
[2] Z. Chen, H. Wang, X. Wu, J. Wang, X. Lin, C. Wang, K. Gao, M. Chapman, D. Li, Object detection in aerial images using DOTA dataset: a survey, Int. J. Appl. Earth Obs. Geoinform. 134 (2024) 104208.
[3] X. Yang, A.S.A. Mohamed, Gaussian-based R-CNN with large selective kernel for rotated object detection in remote sensing images, Neurocomputing 620 (2025) 129248.
[4] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.
[5] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.
[6] R. Varghese, M. Sambath, YOLOv8: a novel object detection algorithm with enhanced performance and robustness, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2024, pp. 1–6.
[7] L. Dai, H. Liu, H. Tang, Z. Wu, P. Song, AO2-DETR: arbitrary-oriented object detection transformer, IEEE Trans. Circuits Syst. Video Technol. 33 (5) (2022) 2342–2356.
[8] L. Fu, W. Liu, G. Li, W. Huang, FFRViT: frequency feature refinement vision transformer for remote sensing object detection, IEEE Trans. Geosci. Remote Sens. 63 (2025), 1–13.
[9] H. Lee, M. Song, J. Koo, J. Seo, Hausdorff distance matching with adaptive query denoising for rotated detection transformer, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 1872–1882.
[10] X. Yang, X. Yang, J. Yang, Q. Ming, W. Wang, Q. Tian, J. Yan, Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 34 (2021) 18381–18394.
[11] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 39 (2017) 1137–1149.
[12] A. Farhadi, J. Redmon, Yolov3: an incremental improvement, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 1804, 2018, pp. 1–6.
[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2980–2988.
[14] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, Z. Luo, R2CNN: rotational region CNN for arbitrarily-oriented scene text detection, 2018, pp. 3610–3615.
[15] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, IEEE Trans. Multimed. 20 (11) (2018) 3111–3122.
[16] J. Zou, W. Zhang, Q. Li, Q. Wang, MOSAIC-tracker: mutual-enhanced occlusion-aware spatiotemporal adaptive identity consistency network for aerial multi-object tracking, ISPRS J. Photogramm. Remote Sens. 229 (2025) 138–154.
[17] W. Zhang, Q. Li, Y. Yuan, Q. Wang, Visual consistency enhancement for multi-view stereo reconstruction in remote sensing, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–11.
[18] Y. Zhao, Y. Liu, R. Song, M. Zhang, Extended non-local means filter for surface saliency detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, 633–636.
[19] Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, RepPoints: point set representation for object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9657–9666.
[20] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, X. Li, Large selective kernel network for remote sensing object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 16748–16759.
[21] X. Yuan, Z. Zheng, Y. Li, X. Liu, L. Liu, X. Li, Q. Hou, M.-M. Cheng, Strip R-CNN: large strip convolution for remote sensing object detection, (2025). arXiv preprint arXiv:2501.03775
[22] W. Lu, X. Yang, S.-B. Chen, LWGANet: addressing spatial and channel redundancy in remote sensing visual tasks with light-weight grouped attention, in: AAAI Conference on Artificial Intelligence, 2026.
[23] W. Lu, S.-B. Chen, Q.-L. Shu, J. Tang, B. Luo, DecoupleNet: a lightweight backbone network with efficient feature decoupling for remote sensing visual tasks, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–13.
[24] Y. Zhang, Y. Xiao, Y. Zhang, T. Zhang, Video saliency prediction via single feature enhancement and temporal recurrence, Eng. Appl. Artif. Intell. 160 (2025) 111840.
[25] Y. Zhang, T. Liu, J. Zhen, Y. Kang, Y. Cheng, Adaptive downsampling and scale enhanced detection head for tiny object detection in remote sensing image, IEEE Geosci. Remote Sens. Lett. 22 (2025) 1–5.
[26] Y. Zhang, J. Zhang, G. Wang, J. Deng, H. Sheng, Y. Muhammad, S. Wei, PolyBuild: an end-to-end method for polygonal building contour extraction from high-resolution remote sensing images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 18 (2025) 12531–12544.
[27] Y. Zhang, S. Wang, Y. Zhang, P. Yu, Asymmetric light-aware progressive decoding network for RGB-thermal salient object detection, J. Electron. Imaging 34 (1) (2025) 013005–013005.
[28] Y. Zhang, P. Yu, Y. Xiao, S. Wang, Pyramid-structured multi-scale transformer for efficient semi-supervised video object segmentation with adaptive fusion, Pattern Recognit. Lett. 194 (2025) 48–54.
[29] J. Han, J. Ding, N. Xue, G.-S. Xia, ReDet: a rotation-equivariant detector for aerial object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 2786–2795.
[30] H. Kaewkorn, L. Zhou, W. Li, RP-Net: a robust polar transformation network for rotation-invariant face detection, Pattern Recognit. 158 (2025) 111044.
[31] M. Liao, B. Shi, X. Bai, TextBoxes++: a single-shot oriented scene text detector, IEEE Trans. Image Process. 27 (8) (2018) 3676–3690.
[32] X. Yang, J. Yan, Z. Feng, T. He, R3Det: refined single-stage detector with feature refinement for rotating object, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 364–369.
[33] J. Ding, N. Xue, Y. Long, G.-S. Xia, Q. Lu, Learning RoI transformer for oriented object detection in aerial images, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2844–2853.
[34] J. Wang, W. Yang, H.-C. Li, H. Zhang, G.-S. Xia, Learning center probability map for detecting objects in aerial images, IEEE Trans. Geosci. Remote Sens. 59 (5) (2021) 4307–4323.
[35] Y. Xu, J. Shen, M. Dai, W. Yang, PARDet: dynamic point set alignment for rotated object detection, Pattern Recognit. 153 (2024) 110534.
[36] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, X. Bai, Gliding vertex on the horizontal bounding box for multi-oriented object detection, IEEE Trans. Pattern Anal. Mach. Intell. 43 (4) (2020) 1452–1459.
[37] L. Hou, K. Lu, X. Yang, Y. Li, J. Xue, G-Rep: gaussian representation for arbitrary-oriented object detection, Remote Sens. 15 (3) (2023) 757.
[38] Z. Zhou, Y. Ma, J. Fan, Z. Liu, F. Jing, M. Tan, Linear Gaussian bounding box representation and ring-shaped rotated convolution for oriented object detection, Pattern Recognit. 155 (2024) 110677.
[39] W. Qian, X. Yang, S. Peng, J. Yan, Y. Guo, Learning modulated loss for rotated object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 35, 2021, pp. 2458–2466.
[40] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, Q. Tian, Rethinking rotated object detection with Gaussian wasserstein distance loss, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., PMLR, 2021, pp. 11830–11841.
[41] Z. Chen, K. Chen, W. Lin, J. See, H. Yu, Y. Ke, C. Yang, PIoU loss: towards accurate oriented object detection in complex environments, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Springer, 2020, pp. 195–211.
[42] Y. Xu, S. Li, X. Yan, J. He, Q. Ni, Y. Sun, Y. Wang, Multiattention-based feature aggregation convolutional networks with dual focal loss for fault diagnosis of rotating machinery under data imbalance conditions, IEEE Trans. Instrum. Meas. 73 (2024) 1–11.
[43] G. Qin, K. Zhang, X. Lai, Q. Zheng, G. Ding, M. Zhao, Y. Zhang, An adaptive symmetric loss in dynamic wide-kernel ResNet for rotating machinery fault diagnosis under noisy labels, IEEE Trans. Instrum. Meas. 73 (2024) 1–12.
[44] J. Li, X. Zhang, K. Yue, J. Chen, Z. Chen, W. Li, An auto-regulated universal domain adaptation network for uncertain diagnostic scenarios of rotating machinery, Expert Syst. Appl. 249 (2024) 123836.
[45] S. Ma, Y. Xu, FPDIoU loss: a loss function for efficient bounding box regression of rotated object detection, Image Vis. Comput. 154 (2025) 105381.
[46] W. Yu, P. Zhou, S. Yan, X. Wang, InceptionNext: when inception meets convnext, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2024, pp. 5672–5683.
[47] L. Sun, J. Dong, J. Tang, J. Pan, Spatially-adaptive feature modulation for efficient image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 13190–13199.
[48] Q. Li, Y. Yuan, Q. Wang, Multiscale factor joint learning for hyperspectral image super-resolution, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–10.
[49] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, J. Han, Anchor-free oriented proposal generator for object detection, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–11.

[50] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, DOTA: a large-scale dataset for object detection in aerial images, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3974–3983.
[51] Y. Han, X. Yang, T. Pu, Z. Peng, Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–18.
[52] Z. Hu, K. Gao, X. Zhang, J. Wang, H. Wang, Z. Yang, C. Li, W. Li, EMO2-DETR: efficient-matching oriented object detection with transformers, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–14.
[53] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, K. Fu, SCRDet: towards more robust detection for small, cluttered and rotated objects, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8231–8240.
[54] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, X. Bai, Gliding vertex on the horizontal bounding box for multi-oriented object detection, IEEE Trans. Pattern Anal. Mach. Intell. 43 (4) (2021) 1452–1459.
[55] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, Y. Yao, Poly kernel inception network for remote sensing detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2024, pp. 27706–27716.
[56] X. Wang, T.E. Huang, T. Darrell, J.E. Gonzalez, F. Yu, Frustratingly simple few-shot object detection 119 (2020) 9919–9928.
[57] Y. Xi, T. Lu, X. Kang, S. Li, Structure-adaptive oriented object detection network for remote sensing images, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–13.
[58] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, T. He, SCRDet++: detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2023) 2384–2399.
[59] Y. Tian, Q. Ye, D. Doermann, YOLOv12: attention-centric real-time object detectors, arXiv:2502.12524 (2025).

Ke Liu is pursuing a master degree in the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing rotated object detection and image generation.

Jian Zou is pursuing a Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and computer vision.

Wei Zhang is pursuing a Ph.D. degree in computer science and technology at the School of Computer Science and the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, remote sensing, and 3D reconstruction.

Qiang Li is a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/).