

LEGNet: A Lightweight Edge-Gaussian Network for Low-Quality Remote Sensing Image Object Detection

Wei Lu¹, Si-Bao Chen¹, Hui-Dong Li¹, Qing-Ling Shu¹, Chris H. Q. Ding², Jin Tang¹, Bin Luo¹

¹Anhui University, ²The Chinese University of Hong Kong (Shenzhen)

{2858191255, 2563489133}@qq.com, {sbchen, tangjin, luobin}@ahu.edu.cn,
e23301341@stu.ahu.edu.cn, chrisding@cuhk.edu.cn

Abstract

Remote sensing object detection (RSOD) often suffers from degradations such as low spatial resolution, sensor noise, motion blur, and adverse illumination. These factors diminish feature distinctiveness, leading to ambiguous object representations and inadequate foreground-background separation. Existing RSOD methods exhibit limitations in robust detection of low-quality objects. To address these pressing challenges, we introduce LEGNet, a lightweight backbone network featuring a novel Edge-Gaussian Aggregation (EGA) module specifically engineered to enhance feature representation derived from low-quality remote sensing images. EGA module integrates: (a) orientation-aware Scharr filters to sharpen crucial edge details often lost in low-contrast or blurred objects, and (b) Gaussian-prior-based feature refinement to suppress noise and regularize ambiguous feature responses, enhancing foreground saliency under challenging conditions. EGA module alleviates prevalent problems in reduced contrast, structural discontinuities, and ambiguous feature responses prevalent in degraded images, effectively improving model robustness while maintaining computational efficiency. Comprehensive evaluations across five benchmarks (DOTA-v1.0, v1.5, DIOR-R, FAIR1M-v1.0, and VisDrone2019) demonstrate that LEGNet achieves state-of-the-art performance, particularly in detecting low-quality objects. The code is available at [here](#).

1. Introduction

Remote sensing object detection (RSOD) is pivotal for diverse real-world applications, yet it contends with persistent challenges from the inherent complexities of aerial and satellite images. A critical challenge is the prevalence of low-quality objects, factors including limited sensor resolution, atmospheric interference, motion blur, variable illumination (e.g., shadows, low-light), and occlusions by natural or man-made structures. These degradation factors com-

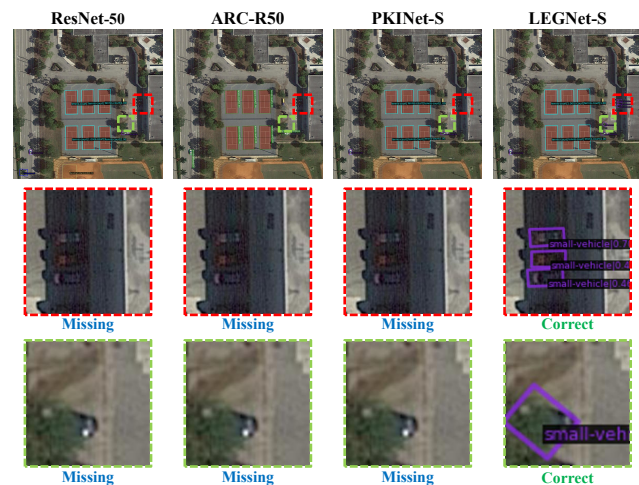


Figure 1. Visualization results on the DOTA-v1.0 test set [36]. All models were built with Oriented R-CNN [37] detector. Our LEGNet demonstrates robust detection under challenging conditions such as occlusion (e.g., objects obscured by trees) and low-light (e.g., building shadows), surpassing previous state-of-the-art methods in both accuracy and robustness for low-quality objects.

promise feature discriminability, leading to: (1) diminished contrast, making foreground objects difficult to distinguish from complex backgrounds; (2) fragmented structural details and edge information, hindering precise localization; and (3) ambiguous or diminished feature responses, particularly when objects are partially obscured or under challenging lighting. Consequently, the robustness and deployability of detection models is undermined, especially for these challenging-to-discern, low-quality objects.

Numerous approaches have been developed to enhance RSOD performance, particularly for multi-scale object detection. For instance, methods employing large kernel convolutions (e.g., LSKNet [19]) or multi-scale feature extraction strategies (e.g., PKINet [1]) have shown promise in capturing contextual information and handling object size variations. However, despite their success in multi-scale

detection, these methods often struggle to robustly represent features from low-quality or degraded objects. Consequently, the subtle yet crucial details of such low-quality objects are frequently lost, distorted, or overwhelmed by noise during feature propagation in deep networks. This problem is exacerbated in lightweight architectures, which are increasingly important for deployment on resource-constrained platforms (e.g., UAVs, satellites) but often sacrifice representational capacity for efficiency, further exhibiting difficulty in preserving subtle characteristics of degraded objects. This performance gap underscores the pressing need for lightweight networks that can specifically enhance features from low-quality RS images without incurring prohibitive computational costs.

To address this critical deficiency, we introduce LEGNet, a Lightweight Edge-Gaussian Network, designed to enhance the representation of low-quality objects in RS images while maintaining model efficiency. At the core of LEGNet is our novel Edge-Gaussian Aggregation (EGA) module. As visualized in Fig. 1, LEGNet exhibits robust performance in detecting challenging objects under conditions like occlusions and low-light environments, outperforming existing state-of-the-art (SOTA) methods, especially for objects with compromised visual quality.

The EGA module combines classical image processing techniques with modern deep learning techniques to specifically address feature degradation. It comprises two key components: First, an orientation-aware edge enhancement mechanism leveraging Scharr operators. These operators preserve fine edge structures with rotational invariance, which is critical for identifying objects with degraded boundaries caused by resolution limits, motion, or weather effects. This explicit edge-awareness helps counteract the structural discontinuity issue. Second, a Gaussian-prior-based feature modeling component. This part refines features, particularly those with low confidence or high ambiguity, by applying Gaussian smoothing with fixed-parameter convolutional kernels. This not only aids in attenuating noise and irrelevant background clutter but also enhances the saliency of true foreground objects, especially under varying illumination or partial occlusions, thereby addressing ambiguous feature responses and improving contrast. The EGA module, by integrating these complementary strategies, effectively bridges traditional image processing insights with learnable deep representations. LEGNet provides a conceptually sound framework for robust feature extraction from degraded visual data, which is well-suited for RSOD tasks in challenging environments.

To validate the efficacy and efficiency of LEGNet, comprehensive experiments are conducted on four challenging public RSOD benchmarks (DOTA-v1.0 [36], DOTA-v1.5 [36], DIOR-R [4], FAIR1M-v1.0 [31]) and a UAV-view dataset (VisDrone2019 [8]). LEGNet consistently

achieves SOTA performance across all datasets while maintaining a lightweight architecture. Specifically, it attains an mAP of 80.03% on DOTA-v1.0, 72.89% on DOTA-v1.5, 68.40% on DIOR-R, 48.35% on FAIR1M-v1.0, and 55.0% mAP50 on VisDrone2019. These results demonstrate that LEGNet improves RSOD performance while remaining computationally efficient, making it suitable for practical, resource-constrained RS applications.

The main contributions are summarized as follows:

- We introduce the novel EGA module, which combines traditional image processing operators (orientation-aware Scharr edge enhancement and Gaussian-prior-based feature modeling) with learnable deep features to specifically tackle feature degradation in low-quality RS images.
- We propose LEGNet, a lightweight network built upon the EGA module, designed to efficiently and effectively improve the detection of challenging objects (e.g., low-quality, blurred, or occluded) while maintaining computational efficiency suitable for edge deployment.
- Extensive experiments on five challenging RSOD benchmarks demonstrate that LEGNet establishes new SOTA performance, particularly excelling on low-quality objects, validating its effectiveness and practical applicability for real-world, resource-constrained RS scenarios.

2. Related Work

2.1. Oriented Object Detection Methods

Horizontal Bounding Box (HBB) methods often struggle in RS scenarios due to their inability to represent object rotations accurately. To address this, Oriented Bounding Box (OBB) detectors have been proposed. Early methods like RRPN [42] introduced rotated anchors to improve object coverage. Subsequent approaches such as RoI Transformer [7], R³Det [44], and S²ANet [10] refined feature alignment and proposal generation. Gliding Vertex [39] and Oriented RepPoints [17] further enhanced geometric flexibility. To tackle the challenges of angular periodicity and regression discontinuity, CSL [41] reformulated angle prediction as classification. Methods like SCRDet [43] introduced smooth IoU-based losses. More recently, Gaussian-based metrics such as GWD [45], KLD [46], and KFIoU [47] have provided differentiable formulations for skewed IoU landscapes, improving optimization stability. Transformer-based methods have also gained traction in RSOD. AO2-DETR [5] proposed oriented proposals to enhance feature interaction, while FPNformer [32] combined FPNs with Transformer decoders to capture multi-scale and rotation-aware features.

2.2. Backbone Networks for RSOD

Backbone design plays a critical role in RSOD performance. Recent efforts include ARC [28], which em-

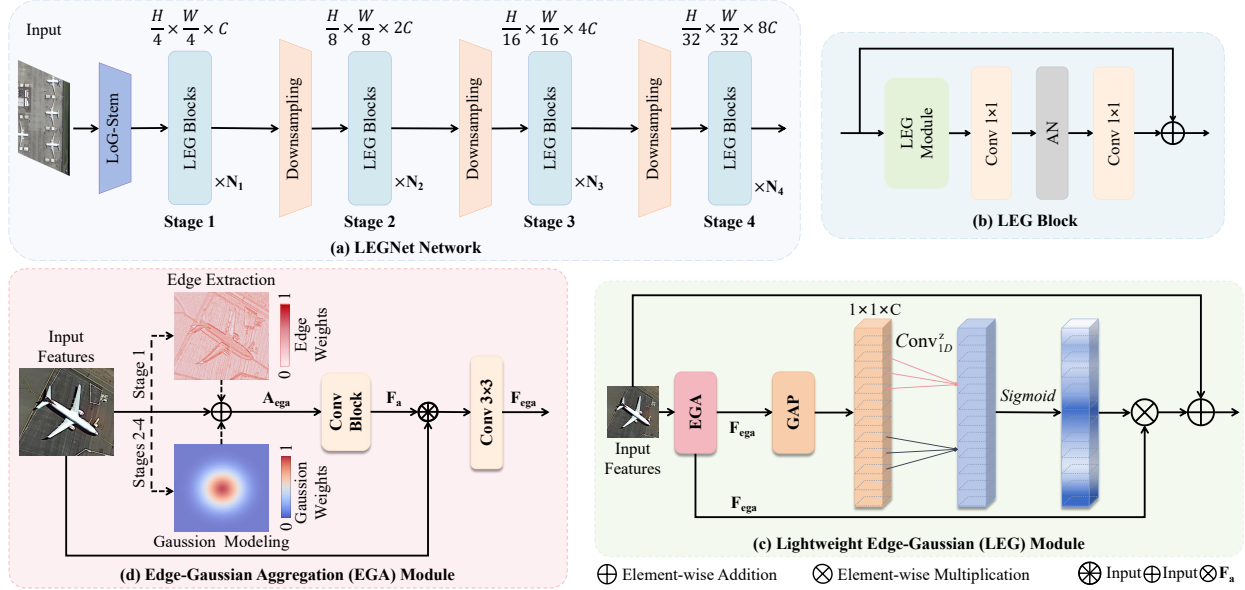


Figure 2. Overview of LEGNet architecture. It consists of 4 stages with input resolutions downsampled by factors of 4, 8, 16, and 32. AN, Conv, GAP, and z denote the activation–normalization layer, convolutional layer, global average pooling, and 1D Conv size, respectively.

ployed adaptive rotation convolutions, and LSKNet [19], which leveraged large kernels to expand receptive fields. PKINet [1] integrated multi-scale kernels, while DecoupleNet [25] focused on preserving fine object details. LWGANet [26] demonstrated competitive results through efficient multi-scale attention within a lightweight architecture.

Although recent methods have effectively addressed challenges related to scale, density, and orientation, RSOD continues to be hindered by degraded feature quality due to sensor limitations and adverse imaging conditions. Nevertheless, robust representation learning under such degradations remains insufficiently explored.

To address this challenge, we propose LEGNet—a lightweight edge-Gaussian network tailored for low-quality RSOD. To the best of our knowledge, LEGNet is the first RSOD backbone to integrate traditional image processing techniques (edge detection and Gaussian filtering) into a modern deep learning framework, offering a novel approach to robust feature extraction in degraded scenarios.

3. Method

This section introduces the architecture and core components of the proposed LEGNet backbone network, which addresses RSOD challenges in low-quality images by leveraging edge cues and Gaussian modeling.

3.1. LEGNet Architecture Overview

As shown in Fig. 2(a), LEGNet adopts a four-stage architecture that progressively extracts multi-scale features with downsampling ratios of 1/4, 1/8, 1/16, and 1/32. It com-

prises three main components: a LoG-Stem layer, downsampling based on DRFD module [24], and lightweight edge-Gaussian (LEG) blocks.

The input image $I \in \mathbb{R}^{H \times W \times 3}$ first passes through the LoG-Stem layer, where a Laplacian of Gaussian (LoG) filter [15] performs initial downsampling while capturing edge features. This produces a feature map $F_{\text{stem}} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, where C set to 32 for LEGNet-Tiny and 64 for LEGNet-Small. Stage 1 processes F_{stem} using N_1 LEG blocks to refine 1/4-scale features. Stages 2 to 4 begin with DRFD modules, which combine convolution and max pooling to preserve small-scale features while reducing resolution. They are followed by N_2 , N_3 , and N_4 LEG blocks, which further refine features at scales of 1/8, 1/16, and 1/32. LEG block enhances feature representation through either edge cues (in shallow stages) or Gaussian modeling (in deeper stages) to improve robustness against noise and shape variations. Finally, outputs from all four stages are aggregated by the RSOD decoder to generate detection results. Detailed configurations are provided in the Appendix.

3.2. LoG-Stem Layer

To alleviate the challenges of noisy and degraded RS images, we introduce the LoG-Stem layer as a key component of LEGNet. It enhances edge features at the initial stage by leveraging the LoG filter’s dual capability of noise suppression and edge detection—critical for RS images where the deep learning feature extractors are often inadequate.

The LoG filter combines Gaussian smoothing with a Laplacian operator, suppressing noise while highlighting re-

gions of rapid intensity change. In our implementation, the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ is first processed by a 7×7 convolution to extract initial features, followed by a LoG filter with kernel size 7×7 and $\sigma = 1.0$, enabling edge-aware feature representation. The $2D$ LoG filter at position $\mathbf{x} = (i, j)$ is defined as:

$$\text{LoG}_{\sigma}^{k \times k}(\mathbf{x}) = \frac{1}{\pi \sigma^4} \left(1 - \frac{i^2 + j^2}{\sigma^2} \right) e^{-\frac{i^2 + j^2}{2\sigma^2}}, \quad (1)$$

with the Gaussian filter $G_{\sigma}^{k \times k}(\mathbf{x})$ given by:

$$G_{\sigma}^{k \times k}(\mathbf{x}) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2 + j^2}{2\sigma^2}}, \quad (2)$$

where $k \times k$ denotes the kernel size, and σ is the standard deviation, set to 1.0 in our case for optimal edge detection.

The output of the LoG filter is activated + normalized (AN), and added to the input via a residual connection: $\mathbf{F}_{LoG} = \text{Norm}(\mathbf{I} + \text{AN}(\text{LoG}_{1.0}^{7 \times 7}(\text{Conv}_{2D}^{7 \times 7}(\mathbf{I}))))$. This residual-like design preserves image details and promotes stable gradient flow. \mathbf{F}_{LoG} is then passed through two 3×3 convolutional layers, with the second using stride 2 for downsampling: $\mathbf{F}_1 = \text{Conv}_{2D}^{3 \times 3}(\text{Conv}_{2D}^{3 \times 3}(\mathbf{F}_{LoG}))$, reducing spatial size to $\frac{H}{2} \times \frac{W}{2}$.

To further enhance multi-scale features, \mathbf{F}_1 is processed by Gaussian filters with kernel sizes 9×9 and 5×5 , both using $\sigma = 0.5$. The outputs are summed, normalized, and passed through the DRFD module to obtain the final 1/4-resolution feature map: $\mathbf{F}_{LoG-STEM} = \text{DRFD}(G_{0.5}^{5 \times 5}(\text{Norm}(G_{0.5}^{9 \times 9}(\mathbf{F}_1) + \mathbf{F}_1)))$.

This process enriches edge features and integrates multi-scale context, providing a robust foundation for downstream detection. The resulting feature map $\mathbf{F}_{LoG-STEM} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ is fed into the Stage 1 LEG blocks.

3.3. LEG Block

As illustrated in Fig. 2(b), the LEG block enhances low-quality features via the EGA module. First, we apply the LEG module to the input feature \mathbf{F}_{in} , then a 1×1 convolution expands channels from C to $2C$, followed by an activation-normalization (AN) layer to yield $\mathbf{F}_{mid} = \text{AN}(\text{Conv}_{2D}^{1 \times 1}(\text{LEG}(\mathbf{F}_{in})))$. Next, a second 1×1 convolution reduces channels back to C , followed by normalization and dropout (rate 0.1). Its output is added to \mathbf{F}_{in} to produce the final output $\mathbf{F}_o = \mathbf{F}_{in} + \text{Norm}(\text{Drop}(\text{Conv}_{2D}^{1 \times 1}(\mathbf{F}_{mid})))$.

3.3.1. Edge Extraction

At Stage 1 (feature size 1/4 of the input), edge and high-frequency details remain salient. We use the Scharr filter—an improved Sobel variant with better rotation invariance—to extract robust edges. Its kernels are

$$S_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}, \quad S_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix},$$

where S_x and S_y denote the parameters of the respective convolution filters. We fix these weights in two convolutions $\text{Conv}_{2D}^{S_x}$ and $\text{Conv}_{2D}^{S_y}$, then combine their outputs via $\mathbf{A}_{edge} = \sqrt{(\text{Conv}_{2D}^{S_x}(\mathbf{F}_{in}))^2 + (\text{Conv}_{2D}^{S_y}(\mathbf{F}_{in}))^2}$.

3.3.2. Gaussian Modeling

In Stages 2-4, edges are blurred with decreasing resolution, but object features tend to exhibit a Gaussian distribution centered around a specific point. To better capture these object features, Gaussian convolution kernels based on a Gaussian spatial prior are employed, which help to smooth feature variations and suppress noise, thereby preserving critical object structures and preventing their degradation during the forward propagation of the network. As described in Eq. (2), the Gaussian filter assigns higher weights to pixels closer to the center of the kernel, emphasizing prominent features in the image while suppressing background noise. This is effective for stabilizing feature representations, enabling the network to focus on the most informative regions, and minimizing artifacts from high-frequency noise. We apply a fixed 5×5 Gaussian kernel via depthwise convolution: $\mathbf{A}_{gauss} = G_{1.0}^{5 \times 5}(\mathbf{F}_{in})$, which smooths each channel, emphasizing central pixels and suppressing noise.

3.3.3. LEG Module

As shown in Fig. 2(c), LEG module processes the input feature \mathbf{F}_{in} through an EGA submodule, which yield the features \mathbf{F}_{ega} . To emphasize critical channels, we adopt the ECA strategy [34]. The ECA is represented as:

$$\mathbf{F}_{temp} = \text{Sigmoid}(\text{Conv}_{1D}^z(\text{GAP}(\mathbf{F}_{ega}))),$$

$$\mathbf{F}_o = \text{Norm}((\mathbf{F}_{temp} \otimes \mathbf{F}_{ega}) + \mathbf{F}_{in}),$$

where Conv_{1D}^z is an adaptive one-dimensional convolution with kernel size z proportional to the channels C , GAP is channel-wise global average pooling, and \otimes indicates element-wise multiplication.

3.3.4. EGA Module

As shown in Fig. 2(d), EGA module processes the input \mathbf{F}_{in} through a selection mechanism, which applies either edge extraction or Gaussian modeling depending on the stage:

$$\mathbf{A}_{ega} = \begin{cases} \mathbf{A}_{edge}(\mathbf{F}_{in}), & \text{if Stage} = 1, \\ \mathbf{A}_{gauss}(\mathbf{F}_{in}), & \text{otherwise.} \end{cases}$$

we add \mathbf{A}_{ega} to \mathbf{F}_{in} and pass the sum through a three-layer Conv_Block to enhance feature representation:

$$\mathbf{F}_{temp} = \text{Conv}_{2D}^{3 \times 3}(\text{AN}(\text{Conv}_{2D}^{1 \times 1}(\mathbf{F}_{in} + \mathbf{A}_{ega}(\mathbf{F}_{in}))))),$$

$$\text{Conv_Block}(\mathbf{F}_{in}) = \text{Norm}(\text{Conv}_{2D}^{1 \times 1}(\text{AN}(\mathbf{F}_{temp}))).$$

The Conv_Block result \mathbf{F}_a combines \mathbf{F}_{in} using element-wise multiplication and addition operations, followed by a 3×3 convolution: $\mathbf{F}_{ega} = \text{Conv}_{2D}^{3 \times 3}((\mathbf{F}_{in} \otimes \mathbf{F}_a) + \mathbf{F}_{in})$.

3.4. Macro Designs

Our fine-tuning experiments demonstrated that edge information is most effective when integrated into the shallow layers of the network. This effectiveness can be attributed to the fact that shallow layers are primarily tasked with capturing low-level, high-frequency visual cues such as edges, textures, and local patterns, which are crucial for defining object boundaries—particularly in degraded images where such details are subtle. However, the integration of these low-level edge features into deeper layers has a detrimental effect on performance. This is typically attributed to the introduction of edge interferences, which causes feature redundancy and disrupts the semantic hierarchy in deeper layers, thereby leading to poor convergence and training instability. In contrast, in deeper layers, explicit edge extraction is omitted to avoid redundancy, and Gaussian modeling is applied to deep features, which are represented as Gaussian distributions—an approach that better aligns with the abstract nature of deeper representations.

To reduce noise in RS images, the LoG-Stem module is applied in shallow layers, allowing edge features to be enhanced by suppressing noise interference. To facilitate the effective fusion of information from both the original input and its corresponding edge maps, the LEG module is designed with extensive residual connections. These connections ensure a balanced integration of both components. Additionally, several standard convolutional layers are employed to further refine the fused features.

4. Experiments

This section presents a comprehensive evaluation of the proposed LEGNet on the RSOD task. We conducted extensive experiments on four RSOD benchmark datasets: DOTA-v1.0 and v1.5 [36], DIOR-R [4], and FAIR1M-v1.0 [31], as well as a UAV-based dataset: VisDrone2019 [8]. Results on DOTA-v1.0, DOTA-v1.5, and FAIR1M-v1.0 were obtained through official online testing to ensure fair and reproducible evaluation. These datasets present diverse challenges, including scale variation, poor-quality features, fine-grained categories, and small object detection, offering a thorough test of LEGNet’s robustness and effectiveness.

4.1. Implementation Details

For a fair comparison with existing SOTA methods [1, 19, 28], we pre-trained our backbone on ImageNet-1K [6] for 300 epochs to enhance model robustness on RSOD datasets.

For training and evaluation, we followed the configurations from previous works [1, 19, 25, 28]. Specifically, for DOTA-v1.0, DOTA-v1.5, and DIOR-R, we applied single-scale training/testing with a scale factor of $3\times$. DOTA images were cropped into $1,024 \times 1,024$ patches with 200-pixel overlap, while DIOR-R used 800×800 inputs. For

FAIR1M-v1.0, we followed [19] and adopted multi-scale training/testing (scales: 0.5, 1.0, 1.5) with $1,024 \times 1,024$ crops and 500-pixel overlap. For VisDrone2019, we used single-scale training/testing with a $1\times$ scale and $1,333 \times 800$ input size. The AdamW optimizer [22] was used with a momentum of 0.9 and weight decay of 0.05, an initial learning rate of 0.0002. A cosine learning rate schedule [23] with warm-up was adopted. Random resizing and flipping were applied during training. The test stage used the same resolution as training for consistency.

All experiments were conducted using MMRotate [50] and PyTorch [27] on Ubuntu 20.04. We used 4 NVIDIA RTX 3090 GPUs for both pre-training and RSOD experiments. Unless otherwise specified, LEGNet was integrated into the O-RCNN [37] detector.

4.2. Quantitative Results

4.2.1. Performance on DOTA-v1.0

LEGNet achieves SOTA performance on the DOTA-v1.0 benchmark under single-scale training and testing protocols. As shown in Tab. 1, LEGNet-S with the O-RCNN [37] two-stage detector achieves 80.03% mAP (29.8M parameters), surpassing existing methods while maintaining competitive model complexity. With the S²ANet [10] single-stage detector, LEGNet-S (23.9M parameters) attains 79.37% mAP. Notably, LEGNet variants secure top-two results in 13 of the 15 categories.

Compared to other lightweight models, LEGNet-T sets a new SOTA accuracy-efficiency trade-off, reaching 78.96% mAP with only 20.6M parameters—the most lightweight configuration listed. This is a 0.92% mAP improvement over the previous leading efficient model, DecoupleNet-D2 [25], using 11.6% fewer parameters.

LEGNet demonstrates superior performance on challenging categories, often characterized by low-quality instances: (1) For small vehicles (SV), LEGNet-S (with S²ANet) achieves 81.42% AP, outperforming the previous best (PKINet-S [1] with S²ANet) by 0.61%. (2) For swimming pool (SP) detection, LEGNet-S (with O-RCNN) obtains 82.00% AP, surpassing competitors by at least 1.8%. (3) For helicopter (HC) detection, LEGNet-T achieves 71.39% AP, outperforming PKINet-S (with O-RCNN) by 8.45% with 33.1% fewer parameters. These results demonstrate LEGNet’s strong balance of parameter efficiency and detection accuracy for oriented objects on this benchmark. To our knowledge, this is the first instance of achieving over 80% mAP on DOTA-v1.0 (single-scale protocols), a significant advancement in RSOD. This underscores the importance of effective feature enhancement and extraction for low-quality objects.

As shown in Tab. 2, LEGNet’s effectiveness as a general-purpose RSOD backbone is validated through framework compatibility experiments. Integrated with several main-

Method	Backbone	#P ↓	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP ↑
<i>One-stage</i>																		
R ³ Det [44]	ResNet-50 [12]	41.9M	89.00	75.60	46.64	67.09	76.18	73.40	79.02	90.88	78.62	84.88	59.00	61.16	63.65	62.39	37.94	69.70
SASM [14]	ResNet-50 [12]	36.6M	86.42	78.97	52.47	69.84	77.30	75.99	86.72	90.89	82.63	85.66	60.13	68.25	73.98	72.22	62.37	74.92
O-RepPoints [17]	ResNet-50 [12]	36.6M	87.02	83.17	54.13	71.16	80.18	78.40	87.28	90.90	85.97	86.25	59.90	70.49	73.53	72.27	58.97	75.97
R ³ Det-GWD [45]	ResNet-50 [12]	41.9M	88.82	82.94	55.63	72.75	78.52	83.10	87.46	90.21	86.36	85.44	64.70	61.41	73.46	76.94	57.38	76.34
R ³ Det-KLD [46]	ResNet-50 [12]	41.9M	88.90	84.17	55.80	69.35	78.72	84.08	87.00	89.75	84.32	85.73	64.74	61.80	76.62	78.49	70.89	77.36
S ² ANet [10]	ResNet-50 [12]	38.5M	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
	ARC-R50 [28]	71.8M	89.28	78.77	53.00	72.44	79.81	77.84	86.81	90.88	84.27	86.20	60.74	68.97	66.35	71.25	65.77	75.49
	PKINet-S [1]	24.8M	89.67	84.16	51.94	71.89	80.81	83.47	88.29	90.80	87.01	86.94	65.02	69.53	75.83	80.20	61.85	77.83
	LEGNet-S (Ours)	23.9M	89.22	85.04	55.59	75.73	81.42	84.56	88.32	90.90	87.90	87.02	65.98	70.13	77.56	81.88	69.26	79.37
<i>Two-stage</i>																		
CenterMap [21]	ResNet-50 [12]	41.1M	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
SCRDet [43]	ResNet-50 [12]	41.9M	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
FR-O [29]	ResNet-50 [12]	41.1M	89.40	81.81	47.28	67.44	73.96	73.12	85.03	90.90	85.15	84.90	56.60	64.77	64.70	70.28	62.22	73.17
Roi Trans. [7]	ResNet-50 [12]	55.1M	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. [39]	ResNet-50 [12]	41.1M	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 [11]	ResNet-50 [12]	31.6M	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
O-RCNN [37]	ResNet-50 [12]	41.1M	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87
	ARC-R50 [28]	74.4M	89.40	82.48	55.33	73.88	79.37	84.05	88.06	90.90	86.44	84.83	63.63	70.32	74.29	71.91	65.43	77.35
	LSKNet-S [19]	31.0M	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 [25]	23.3M	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	69.62	78.04
	PKINet-S [1]	30.8M	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
	LEGNet-T (Ours)	20.6M	89.45	86.49	55.76	76.38	80.59	85.40	88.42	90.90	88.72	86.42	65.24	67.81	77.93	73.49	71.39	78.96
LEGNet-S (Ours)	29.8M	89.87	86.32	56.36	78.67	81.34	85.29	88.53	90.89	87.31	86.46	70.38	68.48	78.34	82.00	70.25	80.03	

Table 1. Performance comparison on DOTA-v1.0 test set [36] under single-scale training and testing. Results for other methods are sourced from [1, 25]. “#P” is calculated for backbone + detector. Red and blue indicate the best and second-best results per column, respectively.

Framework	mAP (%) ↑			
	ResNet-50 [12]	PKINet-S [1]	LEGNet-S	
<i>One Stage</i>	R-FCOS [33]	72.45	74.86	76.89
	R ³ Det [44]	69.70	75.89	77.62
	S ² ANet [10]	74.12	77.83	79.37
<i>Two Stage</i>	FR-O [29]	73.17	76.45	77.19
	Roi Trans. [7]	74.05	77.17	79.62
	O-RCNN [37]	75.87	78.39	80.03

Table 2. Comparison of mAP across different detectors using various backbones on the DOTA-v1.0 test set.

stream detection frameworks, LEGNet-S consistently outperforms both ResNet-50 [12] and the recent PKINet-S [1] across these one-stage and two-stage paradigms.

For single-stage detectors, LEGNet-S yields mAP improvements of +7.92% (R³Det [44]), +4.44% (R-FCOS [33]), and +5.25% (S²ANet [10]) over their ResNet-50. LEGNet-S, with S²ANet, achieves 79.37% mAP.

In two-stage detectors, LEGNet-S similarly demonstrates strong performance, with mAP gains of +5.57% for Roi Transformer [7] and +4.16% for O-RCNN [37] compared to their ResNet-50 versions. Particularly, O-RCNN equipped with LEGNet-S achieves 80.03% mAP, surpassing its PKINet-S counterpart by 1.64% while maintaining comparable parameter efficiency.

These comparisons demonstrate that LEGNet enhances feature representation for various detection architectures, highlighting its strong compatibility and learning capacity.

Tab. 3 presents a comparative evaluation of LEGNet in-

Backbone	Backbone only		Speed (FPS) ↑	mAP (%) ↑
	#P ↓	FLOPs ↓		
ResNet-50 [12]	23.3M	86.1G	23.4	75.87
ARC-R50 [28]	N/A	N/A	11.8	77.35
LSKNet-S [19]	14.4M	54.4G	22.5	77.49
DecoupleNet-D2 [25]	6.2M	23.1G	26.9	78.04
PKINet-S [1]	13.7M	70.2G	5.4	78.39
LEGNet-T	3.6M	30.2G	28.3	78.96
LEGNet-S	12.7M	65.4G	20.9	80.03

Table 3. Comparison of mAP and FPS with various backbone networks using the O-RCNN [37] detector on the DOTA-v1.0 test set. Inference speed, with detector, is measured on a single RTX 3090 GPU. “#P” and FLOPs are calculated for backbones only.

tegrated with O-RCNN [37]. LEGNet-S achieves state-of-the-art performance with 80.03% mAP, outperforming all baselines while maintaining moderate complexity (12.7M params, 65.4G FLOPs) and a competitive speed (20.9 FPS), which is 1.77× faster than ARC-R50 [28].

The lightweight variant LEGNet-T offers the best accuracy-efficiency trade-off, reaching 78.96% mAP with only 3.6M parameters and 30.2G FLOPs—41.9% fewer parameters and 0.92% higher mAP than DecoupleNet-D2 [25]. It also outperforms PKINet-S by 0.57% mAP with 73.7% fewer parameters and 42.7% less computation.

Overall, LEGNet provides the largest performance gain among all backbones. It reduces parameters by 84.5% compared to ResNet-50 while retaining high detection accuracy, setting a new efficiency benchmark. These results highlight

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP \uparrow
RetinaNet-O [20]	71.43	77.64	42.12	64.65	44.53	56.79	73.31	90.84	76.02	59.96	46.95	69.24	59.65	64.52	48.06	0.83	59.16
FR-O [29]	71.89	74.47	44.45	59.87	51.28	68.98	79.37	90.78	77.38	67.50	47.75	69.72	61.22	65.28	60.47	1.54	62.00
Mask R-CNN [13]	76.84	73.51	49.90	57.80	51.31	71.34	79.75	90.46	74.21	66.07	46.21	70.61	63.07	64.46	57.81	9.42	62.67
HTC [3]	77.80	73.67	51.40	63.99	51.54	73.31	80.31	90.48	75.12	67.34	48.51	70.63	64.84	64.48	55.87	5.15	63.40
ReDet [11]	79.20	82.81	51.92	71.41	52.38	75.73	80.92	90.83	75.81	68.64	49.29	72.03	73.36	70.55	63.33	11.53	66.86
LSKNet-S [19]	72.05	84.94	55.41	74.93	52.42	77.45	81.17	90.85	79.44	69.00	62.10	73.72	77.49	75.29	55.81	42.19	70.26
SOOD [35]	80.32	84.41	52.59	74.77	58.48	76.90	86.97	90.87	78.62	76.56	62.93	71.16	74.64	76.04	55.97	25.09	70.39
PKINet-S [1]	80.31	85.00	55.61	74.38	52.41	76.85	88.38	90.87	79.04	68.78	67.47	72.45	76.24	74.53	64.07	37.13	71.47
LEGNet-S (ours)	80.48	85.04	55.64	74.86	52.64	82.16	89.11	90.88	84.55	68.91	66.21	74.16	77.44	74.68	65.76	43.75	72.89

Table 4. Comparison with SOTA methods on the DOTA-v1.5 test set [36] using single-scale training and testing. LEGNet-S is developed based on the O-RCNN [37] detector. Results for other methods are sourced from PKINet [1] and SOOD [35].

Method	G. V. [39]	RetinaNet [20]	C-RCNN [2]	FR-O [29]	RoI Trans. [7]	O-RCNN [37]	LSKNet-S [19]	LEGNet-S
mAP(%) \uparrow	29.92	30.67	31.18	32.12	35.29	45.60	47.87	48.35

Table 5. Comparison with SOTA methods on FAIR1M-v1.0 test set [31]. Results for other methods are sourced from LSKNet [19].

LEGNet’s strong balance between representation and efficiency, making it well-suited for real-time RSOD applications on resource-constrained devices.

4.2.2. Performance on DOTA-v1.5

As shown in Tab. 4, LEGNet-S demonstrates superior performance on the DOTA-v1.5 test set under single-scale training and testing. It achieves a top mAP of 72.89%, surpassing PKINet-S (71.47% mAP) and SOOD [35] (70.39% mAP) by 1.42% and 2.50% mAP, respectively. Our method secures top-two AP scores in 14 of 16 categories, including key ones like large vehicle (LV: 82.16%), ship (SH: 89.11%), roundabout (RA: 74.16%), and container crane (CC: 43.75%). LEGNet-S shows 0.01–5.11% AP improvements in different categories, indicating robust generalization across diverse objects.

4.2.3. Performance on FAIR1M-v1.0

On the FAIR1M-v1.0 benchmark (Tab. 5), LEGNet-S achieves a new SOTA mAP of 48.35%. This result is 0.48% mAP higher than LSKNet-S (47.87% mAP) and shows a 2.75% absolute mAP gain over the O-RCNN baseline (45.60% mAP). LEGNet-S consistently outperforms various architectures, with margins of 18.43% mAP over Gliding Vertex [39] (29.92% mAP) and 13.06% mAP over RoI Transformer [7] (35.29% mAP).

4.2.4. Performance on DIOR-R

Experimental results on the DIOR-R test set demonstrate the superior efficiency and accuracy of our LEGNet-S. As summarized in Tab. 6, LEGNet-S achieves SOTA detection performance (68.40% mAP) with the lowest model complexity among the compared methods, utilizing only 29.9M parameters and 118.1G FLOPs. Compared to the previous leading method, PKINet-S (67.03% mAP), our LEGNet-S achieves a 1.37% mAP improvement while reducing the parameters by 0.9M. LEGNet-S achieves these results with computational costs equivalent to PKINet-S (118.1G FLOPs) and surpasses other efficiency-focused methods,

Method	#P (M) \downarrow	FLOPs (G) \downarrow	mAP (%) \uparrow
FR-O [29]	41.1	134.4	59.54
G.V. [39]	41.1	134.4	60.06
RoI Trans. [7]	55.1	148.4	63.87
LSKNet-S [19]	31.0	173.6	65.90
Oriented Rep [17]	36.6	118.8	66.71
DCFL [38]	36.1	-	66.80
PKINet-S [1]	30.8	118.1	67.03
LEGNet-S	29.9	118.1	68.40

Table 6. Experimental results on the DIOR-R test set [4]. #P and FLOPs were tested by 800×800 with backbone + detector.

Method	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
GFL V1 [18]	23.4	38.2	24.2	14.3	34.7	40.1
CMDNet [9]	29.2	49.5	29.8	16.1	35.9	36.5
QueryDet [40]	28.3	48.1	28.7	-	-	-
DINO [48]	26.8	44.2	28.9	17.5	37.3	41.3
RT-DETR [49]	15.9	36.5	10.7	13.8	28.4	23.1
BAFNet [30]	30.8	52.6	30.9	22.4	41.0	43.2
LEGNet-S	32.2	55.0	32.8	24.5	42.5	45.0

Table 7. Results of LEGNet compared with SOTA methods on VisDrone2019 [8]. LEGNet-S is built within BAFNet [30] detector. Other experimental results were referenced from BAFNet.

such as Oriented Rep [17] (66.71% mAP at 118.8G FLOPs) and LSKNet-S (65.90% mAP at 173.6G FLOPs).

4.2.5. Performance on VisDrone2019

Results on the VisDrone2019 benchmark (Tab. 7) further confirm the effectiveness of our LEGNet-S. When integrated with the BAFNet [30] detector, LEGNet-S achieves new SOTA performance across all primary metrics, reaching 32.2% AP—a 1.4% absolute improvement over BAFNet (30.8% AP). Our method excels in small object detection (AP_s), achieving 24.5%, which is 2.1% higher than BAFNet, and surpasses GFL V1 (14.3%) and DINO (17.5%) by 10.2% and 7.0%, respectively. With 55.0% AP₅₀ and 32.8% AP₇₅, LEGNet-S outperforms BAFNet by 2.4% and 1.9%, respectively, indicating superior localization accuracy at multiple IoU thresholds. LEGNet-S maintains advantages across all object scales,

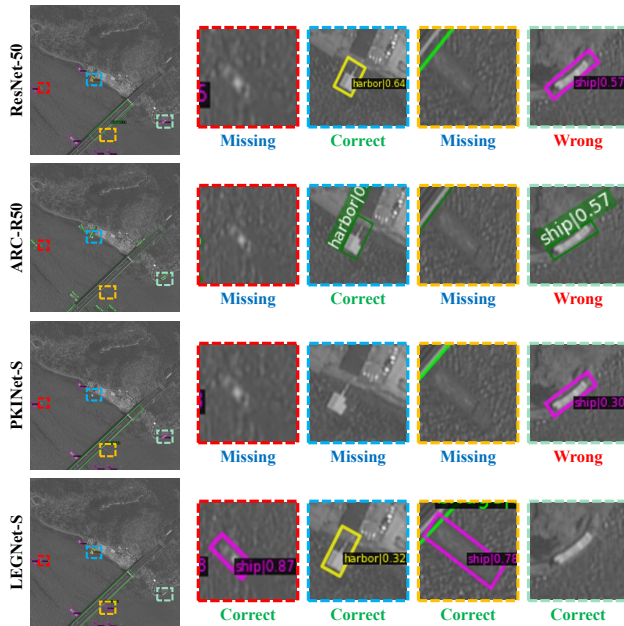


Figure 3. Visualization of detection results on DOTA-v1.0 test set [36]. Input images resolution were $1,024 \times 1,024$.

LoG-Stem	EGA	#P ↓	FLOPs ↓	mAP ↑	AP for each category (%) ↑			
					SV	LV	SP	HC
		20.61	135.6	70.8	59.7	84.0	55.3	46.0
✓		20.65	148.5	71.2	61.5	84.8	61.3	52.7
	✓	20.61	135.6	71.5	59.7	84.5	57.2	61.4
	LEGNet-T	20.65	148.5	72.2	63.2	85.0	59.6	58.3

Table 8. Ablation study of LEGNet on DOTA-v1.0 validation set. Models were pre-trained 100 epoch on Imagenet-1k.

particularly for small objects, and also improves large object detection (45.0% AP_l vs. BAFNet’s 43.2%).

These findings validate LEGNet-S’s enhanced feature representation capabilities, especially when combined with boundary-aware decoders. This demonstrates its effectiveness for drone-captured scenarios characterized by dense small objects and complex backgrounds.

4.3. Qualitative Results

Visual results on the DOTA-v1.0 test set are presented in Fig. 3. We visually compare LEGNet with SOTA backbones designed for RSOD: ARC-R50 [28] and PKINet [1]. As depicted, LEGNet demonstrates superior capability in identifying blurry and low-quality objects, effectively detecting challenging examples such as ships. In contrast, ResNet-50, ARC-R50, and PKINet demonstrate suboptimal performance on such images. Furthermore, these methods exhibit false positives, for example, misclassifying a distinct white object on the shore as a ship, highlighting their limitations in precise object boundary discrimination.

4.4. Ablation Studies

Tab. 8 details an ablation study of LEGNet on the DOTA-v1.0 validation set. All models used a backbone pre-trained for 100 epochs on ImageNet-1K [6], following [1, 19] for efficiency. We compare a baseline (no LoG-Stem or EGA), variants with only LoG-Stem or EGA, and the LEGNet-T. Ablating the EGA module means its fixed kernels for edge/Gaussian attention are replaced by learnable convolution. Ablating the LoG-Stem replaces it with a learnable 4×4 stride-4 convolutional layer. We report mAP and AP for challenging categories: small vehicles (SV), large vehicles (LV), swimming pools (SP), and helicopters (HC).

As shown in Tab. 8, the variant with only LoG-Stem (EGA removed) achieves 71.2% mAP. The variant with only EGA (LoG-Stem removed) yields 71.5% mAP, notably improving HC detection to 61.4%. The full LEGNet-T attains the highest mAP of 72.2%, significantly outperforming the baseline in SV (63.2% vs. 59.7%) and HC detection (58.3% vs. 46.0%). These results confirm that both LoG-Stem and EGA modules enhance detection performance, especially for these challenging classes.

4.5. Limitations and Future Works

The optimal Gaussian kernel size requires careful selection, as its effectiveness is likely dataset-dependent, varying with object characteristics. Furthermore, while LEGNet surpasses current SOTA methods with fewer parameters and FLOPs, its inference speed indicates clear potential for further optimization (less optimized implementations compared to convolution in deep learning framework).

Future work could extend this framework to downstream tasks like semantic segmentation and object tracking. Leveraging its demonstrated robust feature extraction and efficiency, these applications could benefit from edge-Gaussian modeling. Extending our approach broadens its applicability and opens avenues for high performance in diverse vision challenges. Further exploration of task-specific adaptations and further optimization to build efficient and scalable systems for real-world problems.

5. Conclusion

In this paper, we present LEGNet—a lightweight network designed for robust RSOD under low-quality imaging conditions. By bridging traditional image processing (orientation-aware Scharr edge enhancement and Gaussian-prior-based feature modeling) and deep learning, LEGNet enhances low-quality feature representation and object boundary clarity, achieving SOTA performance on multiple benchmarks while maintaining computational efficiency. Future works could extend this framework to additional downstream tasks, providing a versatile foundation for further research in aerial and satellite image analysis.

Acknowledgements: This work was supported in part by NSFC Key Project of International (Regional) Cooperation and Exchanges (No. 61860206004), NSFC Key Project of Joint Fund for Enterprise Innovation and Development (No. U20B2068, U24A20342) and National Natural Science Foundation of China (No. 61976004).

References

- [1] Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *CVPR*, pages 27706–27716, 2024. 1, 3, 5, 6, 7, 8, 2
- [2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 7
- [3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiao-xiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, pages 4974–4983, 2019. 7
- [4] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.*, 60:1–11, 2022. 2, 5, 7, 1
- [5] Linhui Dai, Hong Liu, Hao Tang, Zhiwei Wu, and Pinhao Song. Ao2-DETR: Arbitrary-oriented object detection transformer. *IEEE TCSVT*, 33(5):2342–2356, 2022. 2
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5, 8
- [7] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *CVPR*, pages 2849–2858, 2019. 2, 6, 7
- [8] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *ICCVW*, 2019. 2, 5, 7, 1
- [9] Chengzhen Duan, Zhiwei Wei, Chi Zhang, Siying Qu, and Hongpeng Wang. Coarse-grained density map guided object detection in aerial images. *ICCVW*, pages 2789–2798, 2021. 7
- [10] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.*, 60:1–11, 2021. 2, 5, 6
- [11] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *CVPR*, pages 2786–2795, 2021. 6, 7
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 2
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017. 7
- [14] Liping Hou, Ke Lu, Jian Xue, and Yuqiu Li. Shape-adaptive selection and measurement for oriented object detection. In *AAAI*, pages 923–932, 2022. 6
- [15] Hui Kong, Hatice Cinar Akakin, and Sanjay E Sarma. A generalized laplacian of gaussian filter for blob detection and its applications. *IEEE Trans. Cybern.*, 43(6):1719–1733, 2013. 3
- [16] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.*, 159:296–307, 2020. 1
- [17] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *CVPR*, pages 1829–1838, 2022. 2, 6, 7
- [18] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, pages 21002–21012, 2020. 7
- [19] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection. In *ICCV*, pages 16794–16805, 2023. 1, 3, 5, 6, 7, 8
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 7
- [21] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 14:4205–4230, 2021. 6
- [22] I Loshchilov. Decoupled weight decay regularization. *arXiv*, abs/1711.05101, 2017. 5
- [23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv*, abs/1608.03983, 2016. 5
- [24] Wei Lu, Si-Bao Chen, Jin Tang, Chris H. Q. Ding, and Bin Luo. A robust feature downsampling module for remote-sensing visual tasks. *IEEE Trans. Geosci. Remote Sens.*, 61:1–12, 2023. 3
- [25] Wei Lu, Si-Bao Chen, Qing-Ling Shu, Jin Tang, and Bin Luo. Decouplenet: A lightweight backbone network with efficient feature decoupling for remote sensing visual tasks. *IEEE Trans. Geosci. Remote Sens.*, 62:1–13, 2024. 3, 5, 6
- [26] Wei Lu, Sihao Chen, Chris H. Q. Ding, Jin Tang, and Bin Luo. Lwganet: A lightweight group attention backbone for remote sensing visual tasks. *arXiv*, abs/2501.10040, 2025. 3
- [27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5
- [28] Yifan Pu, Yiru Wang, Zhuofan Xia, Yizeng Han, Yulin Wang, Weihao Gan, Zidong Wang, Shiji Song, and Gao Huang. Adaptive rotated convolution for rotated object detection. In *ICCV*, pages 6589–6600, 2023. 2, 5, 6, 8
- [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2016. 6, 7

- [30] Jingnan Song, Mingliang Zhou, Jun Luo, Huayan Pu, Yong Feng, Xuekai Wei, and Weijia Jia. Boundary-aware feature fusion with dual-stream attention for remote sensing small object detection. *IEEE Trans. Geosci. Remote Sens.*, 63:1–13, 2025. 7
- [31] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. FAIRIM: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.*, 184:116–130, 2022. 2, 5, 7, 1
- [32] Yang Tian, Mengmeng Zhang, Jinyu Li, Yangfan Li, Hong Yang, and Wei Li. Fpnformer: Rethink the method of processing the rotation-invariance and rotation-equivariance on arbitrary-oriented object detection. *IEEE Trans. Geosci. Remote Sens.*, 62:1–10, 2024. 2
- [33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE TPAMI*, 44(4):1922–1933, 2020. 6
- [34] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-Net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, pages 11531–11539, 2020. 4
- [35] Yifan Xi, Ting Lu, Xudong Kang, and Shutao Li. Structure-adaptive oriented object detection network for remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 2024. 7
- [36] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, pages 3974–3983, 2018. 1, 2, 5, 6, 7, 8
- [37] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *ICCV*, pages 3520–3529, 2021. 1, 5, 6, 7, 2
- [38] Chang Xu, Jian Ding, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Dynamic coarse-to-fine learning for oriented tiny object detection. In *CVPR*, pages 7318–7328, 2023. 7
- [39] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE TPAMI*, 43(4):1452–1459, 2020. 2, 6, 7
- [40] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *CVPR*, pages 13658–13667, 2022. 7
- [41] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *ECCV*, pages 677–694, 2020. 2
- [42] Xue Yang, Hao Sun, Kun Fu, Jirui Yang, Xian Sun, Menglong Yan, and Zhi Guo. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.*, 10(1):132, 2018. 2
- [43] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *ICCV*, pages 8232–8241, 2019. 2, 6
- [44] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *AAAI*, pages 3163–3171, 2021. 2, 6
- [45] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *ICML*, pages 11830–11841, 2021. 2, 6
- [46] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. In *NeurIPS*, pages 18381–18394, 2021. 2, 6
- [47] Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The KFIoU loss for rotated object detection. In *ICLR*, 2023. 2
- [48] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 7
- [49] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Dets beat yolos on real-time object detection. *CVPR*, pages 16965–16974, 2024. 7
- [50] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, et al. Mmrotate: A rotated object detection benchmark using pytorch. In *ACM MM*, pages 7331–7334, 2022. 5