A Comprehensive Framework for Fine-Grained Object Recognition in Remote Sensing

Xin Chi
Beijing Research Institute of Uranium Geology
Beijing, China
National Key Laboratory of Uranium Resources Exploration-Mining and Nuclear Remote Sensing
Beijing, China
13664284913@163.com

Yu Sun
sunyutectonics@163.com

Yingjun Zhao
zhaoyingjun@briug.cn

Donghua Lu
alexgreat@126.com

Jun Yang
17856790828@163.com

Yiting Zhang
bwfyzwy224@126.com

Abstract

Fine-grained object recognition represents a practical requirement for intelligent interpretation of high-resolution remote sensing imagery. Existing research primarily concentrates on the detection of stumpy targets. Nevertheless, slender objects characterized by a height significantly exceeding their length or width are also common in practical applications. Current research is inadequate to tackle the challenges presented by slender targets, and there is an urgent need for effective methodologies to address this issue. To this end, this paper proposes a model, named Generalized Adaptive Rotation Faster R-CNN (GA-RFRCNN). The GA-RFRCNN optimizes feature representation across multiple scales by integrating selective enhancement feature pyramid network (SE-FPN). Besides, it introduces an enhanced rotation region proposal network (ERRPN) to enhance the object localization. Furthermore, a dynamically adjusted training process is used to handle difficult-to-detect samples by introducing the adaptive slide loss (ASLoss). We conduct extensive experiments on the transmission tower custom dataset (TT-OBB) and the HRSC2016 dataset, and the results show that our model achieves significant improvements in recognition accuracy and oriented bounding box detection.

Keywords: ***Fine-grained object recognition, high-resolution remote sensing imagery, oriented bounding box detection, varied object structures***

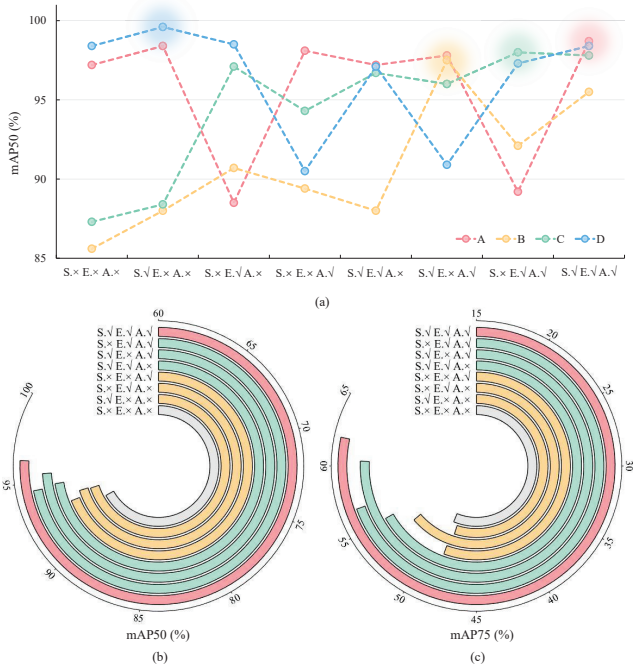


Figure 1. Detection performance of different network configurations on the TT-OBB dataset, with ablation analysis for (b) mAP50 and (c) mAP75. The highlighted section in (a) shows the scenario with the highest category accuracy. S. represents SE-FPN, E. represents ERRPN, and A. represents ASLoss.

1. Introduction

With advancements in remote sensing technology and the widespread adoption of deep learning, high-resolution

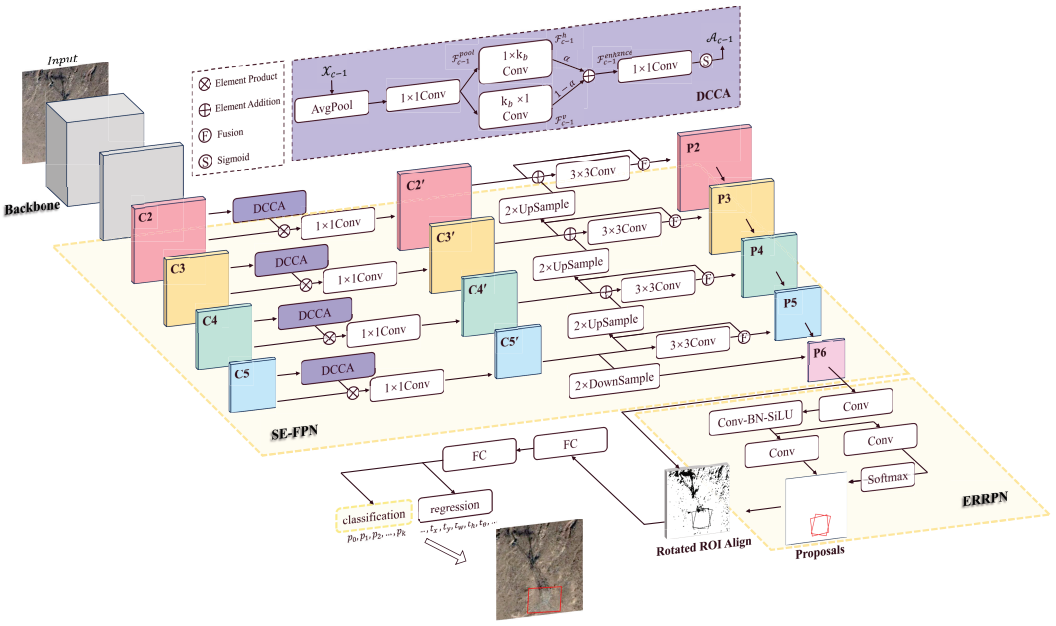


Figure 2. Overview of GA-RFRCNN. The SE-FPN integrates the innovative DCCA module along with enhanced feature fusion mechanisms. ERRPN and ASLoss are also essential components, with detailed descriptions provided in later sections.

remote sensing images have significantly enhanced the accuracy of earth observation and information extraction. Accurate recognition of targets in remote sensing imagery not only improves our understanding of terrestrial phenomena but also has important applications in areas such as disaster prevention [1, 27, 3, 19, 18], infrastructure management [8, 9], and smart city development [48, 22, 35].

However, as application requirements evolve, traditional detection models exhibit limitations in addressing fine-grained recognition [32]. These constraints stem from substantial intra-class variation and minimal inter-class differences [6]. Targets within the same category may exhibit considerable differences in appearance due to factors such as viewing angles, lighting conditions, occlusion, and background clutter. Conversely, targets from different categories often share structural similarities that complicate differentiation. This complexity hinders the model's ability to extract stable features necessary for distinguishing fine-grained categories. Moreover, targets in remote sensing images are often embedded in complex backgrounds and exhibit significant scale variations, further increasing the difficulty of detection and classification. Some progress has been made in addressing challenges related to complex backgrounds and angular rotations [31, 39, 4, 30, 14].

Despite this progress, most current research focuses on fine-grained recognition challenges associated with stumpy targets such as ship and aircraft [6, 12, 29, 4, 11, 14, 21]. In contrast, slender targets like high-voltage transmission towers, wind power towers, and industrial chimneys receive less attention, despite their practical significance. Their slender morphology, with a height significantly exceeding their length and width, makes them difficult to detect. In satellite remote sensing imagery, the perspective often compresses these vertical objects, resulting in a smaller pixel representation. Additionally, the imaging process is influenced by various factors such as lighting conditions, atmospheric interference, and satellite viewing angles, leading to substantial differences in imaging quality [44, 43, 17, 15, 16]. These factors, combined with the frequent appearance of slender targets in complex backgrounds at varying angles, further challenge traditional detection methods, often resulting in failure to capture their complete structure and essential features.

Environmental factors and geographical variations can alter the structure, scale, and orientation of slender targets in imagery. Traditional horizontal bounding box (HBB) detection techniques often include excessive background information, complicating the accurate representation of an object's orientation and shape. Therefore, the introduction of oriented bounding boxes (OBBs), aligned with the actual orientation of the object, has become a critical strategy for improving detection and classification performance of slender targets.

To address these challenges and expand the application scope of fine-grained recognition in remote sensing, this paper proposes the Generalized Adaptive Rotation Faster R-CNN (GA-RFRCNN) model. This model is specifically designed to accommodate a broader range of fine-grained

recognition tasks by integrating several key innovations aimed at improving accuracy and robustness.

- Transmission Tower Oriented Bounding Box (TT-OBB) Datasets: To enable a fine-grained analysis of slender targets in remote sensing imagery, this study presents a specialized dataset for transmission tower detection. The dataset validates model performance on slender targets, complementing public datasets that primarily focus on stumpy targets. This approach ensures a comprehensive evaluation of the model across various target types.

- Network Architecture Optimization: Modifications to GA-RFRCNN include optimized feature extraction techniques, enhanced region proposal mechanisms, and dynamically adjusted loss functions. These enhancements aim to improve both accuracy and recall in complex environments. The model is particularly effective for objects with varying sizes and orientations, especially slender targets.

- Comprehensive Validation: Extensive evaluations were conducted on both custom datasets and the publicly available HRSC2016 dataset. These experiments demonstrate the model's effectiveness in fine-grained recognition and high-precision oriented bounding box detection.

The rest of this article is organized as follows: Section 2 surveys related work, Section 3 elaborates on network architecture and its enhancements, Section 4 details the experimental setup and outcomes, and Section 5 synthesizes the main findings and proposes future research directions.

2. Related work

2.1. Fine-Grained object recognition in remote sensing imagery

Fine-grained recognition in remote sensing imagery is a critical research area. It focuses on distinguishing targets with highly similar structures and appearances, which pose challenges for traditional detection methods. The diversity of objects and subtle visual differences complicate accurate recognition. The detailed capture of ground features by high-resolution optical satellite imagery enables precise detection and monitoring of infrastructure. In this context, deep learning methodologies play a key role in enhancing recognition accuracy. These methods help distinguish objects with minimal inter-class variation and significant intra-class differences, which are often influenced by environmental and observational factors.

Several studies have addressed the challenges of fine-grained object recognition, particularly for diverse and complex targets. For instance, Osswald-Cankaya and Mayer [26] proposed a method for fine-grained recognition in satellite images based on task separation and orientation normalization. Their approach separates detection and classification tasks and normalizes object orientations, leading to improved accuracy on the FAIR1M dataset. Guo et al. [12] introduced MSRIP-Net, a multi-scale rotation-invariant network, which excels at fine-grained aircraft identification without additional annotations. Zhao et al. [46] developed a two-stage CNN architecture using Sentinel-2 imagery for detecting and localizing reservoirs across China, addressing the crucial challenge of accurate identification for water management and flood control.

To further improve detection accuracy across various fine-grained targets, researchers have focused on techniques specifically designed for slender objects. These techniques leverage auxiliary information, such as shadows and imaging parameters, to improve localization. Huang et al. proposed SI-STD [15] and IPC-Det [17], which use shadow information and imaging parameters to improve the localization of slender targets, like transmission towers. However, these methods not only rely heavily on image-specific parameters, such as solar altitude and satellite viewing angles, but also increase the need for labor-intensive manual labeling, limiting their scalability for large-scale fine-grained recognition tasks.

2.2. Dataset limitations and challenges

The constraints of publicly available datasets have significantly impacted the development of robust fine-grained recognition models in remote sensing. Many widely used datasets rely on HBBs for object annotation. While HBBs are effective for conventional detection, they struggle to represent objects with arbitrary orientations. This limitation hampers the model's ability to fully utilize spatial and angular information, which is critical for detecting and classifying complex objects with rotational variance.

Moreover, existing fine-grained recognition datasets, though beneficial, have notable scope limitations. Datasets like HRSC2016 [23], RarePlanes [28], FGSC-23 [45], and FGSCR-42 [6] mainly focus on stumpy object types such as ships, aircraft, and vehicles. However, less attention has been given to objects with distinct slender features. Expanding the diversity of objects in fine-grained recognition datasets could significantly enhance model performance across a broader range of remote sensing detection tasks.

3. Methodology

The GA-RFRCNN model, as shown in Figure 2, includes three main innovations designed to enhance fine-grained recognition in complex remote sensing images. SE-FPN boost the model's object recognition capabilities by using a channel attention mechanis m and an adaptive fusion

strategy. This approach facilitates more effective handling of intricate geometric shapes against complex backgrounds. ERRPN improves target localization by incorporating high-level convolutional layers and multi-layer feature fusion, which accurately captures the rotational characteristics of objects, especially those appearing in various orientations. ASLoss fine-tunes training by dynamically adjusting loss weights, focusing more on hard-to-classify targets. This targeted optimization improves the model's detection accuracy for slender objects commonly missed in complex backgrounds.

In summary, these innovations in GA-RFRCNN strengthen its ability to classify and detect a broader range of objects, with an emphasis on improving performance for the slender targets that traditional methods often overlook.

3.1. Selective enhancement feature pyramid network (SE-FPN)

Fine-grained recognition in remote sensing imagery presents significant challenges due to the structural and appearance variations of multi-scale objects. To address this issue, this paper introduces the SE-FPN, which integrates channel attention mechanisms and adaptive feature fusion to improve object detection accuracy. SE-FPN selectively enhances key features and integrates multi-scale information, more effectively, overcoming the limitations of traditional Feature Pyramid Networks (FPNs).

SE-FPN extracts multi-scale feature maps C2, C3, C4, C5 from the backbone network. Its core innovation is the Dual-Channel Convolutional Attention (DCCA) module, enhancing feature expressiveness by capturing spatial dependencies and emphasizing critical features. This approach is inspired by prior research[2, 20, 34]. For a given feature map, the DCCA applies average pooling and 1×1 convolutions to extract local features:

$$\mathcal{F}_{c-1}^{pool} = \mathrm{Conv}_{1\times1}(\mathcal{P}_{\mathrm{avg}}(\mathcal{X}_{c-1})) \tag{1}$$

where $\mathcal{P}_{\mathrm{avg}}$ represents the average pooling operation. Horizontal and vertical convolutions are then applied to capture directional context:

$$\mathcal{F}_{c-1}^{h} = \mathrm{Conv}_{1\times k_b}(\mathcal{F}_{c-1}^{pool}) \tag{2}$$

$$\mathcal{F}_{c-1}^{v} = \mathrm{Conv}_{k_b\times1}(\mathcal{F}_{c-1}^{pool}) \tag{3}$$

where k_b represents the kernel size. A learnable parameter α balances these features, and the combined representation is:

$$\mathcal{F}_{c-1}^{\mathrm{enhance}} = \alpha\mathcal{F}_{c-1}^{h} + (1-\alpha)\mathcal{F}_{c-1}^{v} \tag{4}$$

where σ is the Sigmoid activation function. The attention weight is computed as:

$$\mathcal{A}_{c-1} = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathcal{F}_{c-1}^{\mathrm{enhance}}\right)\right) \tag{5}$$

The DCCA module also utilizes adaptive padding to maintain consistent feature map dimensions. Afterward, the feature fusion module of SE-FPN aggregates multi-scale features through both up-sampling and down-sampling techniques. This fusion process is guided by a weighted combination [33] of aligned features:

$$\mathcal{X}'_{c-1} = \mathrm{Conv}_{1\times1}\left(\mathcal{A}_{c-1}\cdot\mathcal{X}_{c-1}\right) \tag{6}$$

$$\mathcal{X}_{p-1}^{\mathrm{td}} = \mathrm{Conv}_{3\times3}\left(X'_{c-1} + \mathrm{Resize}(X'_{c})\right) \tag{7}$$

$$\mathcal{X}_{p-1} = \mathrm{Conv}_{1\times1}\left(\frac{\omega_1\cdot\mathcal{X}'_{c-1} + \omega_2\cdot\mathcal{X}_{p-1}^{\mathrm{td}}}{\omega_1+\omega_2+\epsilon}\right) \tag{8}$$

where $\mathcal{X}'_{c-1}$ represents the enhanced feature map obtained after combining the feature map $\mathcal{X}_{c-1}$ with its corresponding attention weight $\mathcal{A}_{c-1}$. $\mathcal{X}_{p-1}^{\mathrm{td}}$ denotes the intermediate feature at the p-th level in the top-down pathway. The learnable weights ω_i balance each input feature, and ϵ prevents division by zero.

SE-FPN overcomes the limitations of traditional FPNs, improving the detection of multi-scale and complex objects. Its ability to capture subtle structural variations makes it highly effective for fine-grained recognition tasks.

3.2. Enhanced rotation region proposal network (ERRPN)

Accurate localization of rotated objects is essential for fine-grained recognition in optical remote sensing. ERRPN enhances this process by reengineering the traditional Rotated Region Proposal Network (RRPN) [25] through advanced convolutional layers. These enhancements significantly improve its ability to detect objects with arbitrary orientations, a prevalent challenge in remote sensing imagery.

As illustrated in Figure 3, ERRPN's proposal generation is centered around a series of convolutional layers. It commences with a shared 3×3 convolutional block (padding=1). This step augments feature representation by transforming the input channels into intermediate channels, which is critical for accurate proposal generation.

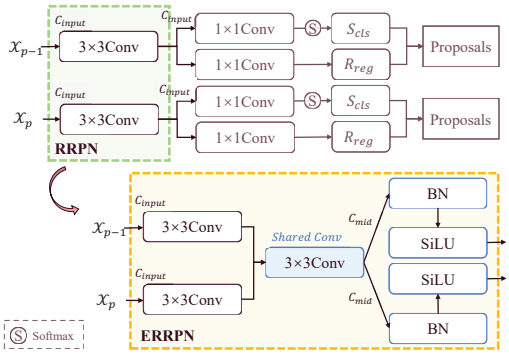


Figure 3. Structure comparison between RRPN and ERRPN.

Subsequently, a Batch Normalization (BN) layer combined with SiLU activation functions to smooth gradients, stabilizing training and enhancing convergence. The network then diverges into two branches: one for classification and one for regression.

Classification Head: Generates a score map S_{cls}, which estimates the probability of each anchor containing an object:

$$S_{\text{cls}} = \text{Conv}_{1\times1}(C_{\text{mid}}, N \times \mathcal{C}_{\text{cls}}) \tag{9}$$

where C_{mid} is the transformed middle channel, N is the number of anchors per location, and $\mathcal{C}_{\text{cls}}$ represents the classification output (object vs. background).

Regression Head: Outputs five parameters—x, y, w, h, and θ—for each anchor, predicting rotated bounding boxes:

$$R_{\text{reg}} = \text{Conv}_{1\times1}(C_{\text{mid}}, N \times 5) \tag{10}$$

These parameters are decoded to recover final bounding box coordinates in the original image space. Non-Maximum Suppression (NMS), adapted for rotation angles, is then applied to refine proposal selection.

3.3. Adaptive slide loss (ASLoss)

ASLoss enhances fine-grained recognition by dynamically adjusting classification loss based on the Intersection over Union (IoU) between predicted and ground truth bounding boxes. In contrast to Slide Loss [40], which rely solely on IoU, ASLoss incorporates a confidence factor that allows for more precise and adaptable loss adjustments. This innovation helps the model better handle challenging samples, particularly complex targets, by refining predictions that are nearly accurate but lack confident classification.

The key innovation of ASLoss lies in its use of a modulation weight β, which is determined by the IoU, $confidence$, and a threshold τ. This combination allows ASLoss to dynamically adjust the classification loss, placing more emphasis on difficult or ambiguous predictions. The IoU between a ground truth box A and a predicted box B is defined as:

$$\text{IoU} = \frac{A \cap B}{A \cup B} \tag{11}$$

The modulation weight β is computed as:

$$\beta = \begin{cases} 1 & \text{IoU} \leq \tau - 0.1 \\ e^{(1-\tau)} & \tau - 0.1 < \text{IoU} < \tau \\ e^{(2-\text{confidence})} & \text{IoU} \geq \tau \end{cases} \tag{12}$$

where $confidence$ refers to the predicted confidence score of the bounding box, typically ranging between 0 and 1. This formulation ensures that predictions near the threshold τ, which are often more challenging, receive greater emphasis.

The total loss for GA-RFRCNN is composed of two components: the classification loss L_{cls} and the regression loss L_{reg}. These components jointly optimize the model for both accurate class prediction and precise bounding box localization.

The classification loss is defined as:

$$L_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N}\left[\beta_i\left(\log\left(1+e^{-p_i}\right) - y_i p_i\right)\right] \tag{13}$$

where N is the number of samples, y_i is the ground truth label, p_i is the predicted probability and β_i is the modulation weight.

The regression loss measures the difference between predicted and ground truth bounding boxes:

$$L_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} F_{\text{reg}}(t_i, t_i^*) \tag{14}$$

where t_i is the predicted bounding box parameter (coordinates, width, height, and angle) and t_i^* is the ground truth. F_{reg} is the Smooth L1 loss.

The total loss function is a weighted sum of the classification and regression losses:

$$L = L_{\text{cls}} + \lambda L_{\text{reg}} \tag{15}$$

where λ balances the importance of classification and regression components, ensuring optimal model performance across various tasks.

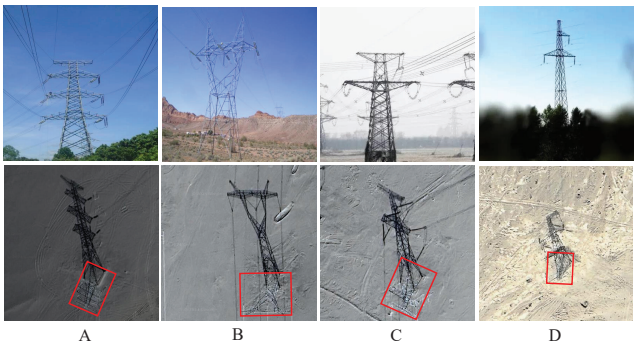


Figure 4. Example transmission tower categories (A, B, C, D) in natural scene images and remote sensing images.

4. Experiments

4.1. Datasets

In this study, we developed the TT-OBB dataset, specifically for detecting transmission towers in optical satellite remote sensing imagery. Transmission towers were chosen because of their diverse designs and complex structures, making them ideal slender targets for fine-grained recognition studies. The dataset contains 1804 high-resolution

satellite images, primarily sourced from Google Earth and SuperView satellites. These images were partitioned into training, validation, and test sets in an 8:1:1 ratio. Each image has a resolution of 1024×1024 pixels, with a spatial resolution ranging from approximately 0.3 to 0.6 meters per pixel. Focusing on the northwestern region of China, which is characterized by extensive transmission infrastructure, the dataset provides a diverse array of examples essential for the development and evaluation of detection algorithms.

Notably, the TT-OBB dataset employs the OBB annotation method. This approach accommodates the varied orientations and perspectives of transmission towers in satellite imagery, substantially enhancing the accuracy and robustness of detection. Moreover, the dataset meticulously categorizes transmission towers into four common types, as shown in Figure 4. These categories encompass the most prevalent transmission tower designs, laying a robust foundation for advancing fine-grained recognition research.

To further evaluate the effectiveness and generalizability of the models, the publicly available HRSC2016 dataset was incorporated into the experiments. This dataset is widely utilized for ship detection in remote sensing applications and comprises 1,061 aerial images with dimensions ranging from 300×300 to 1500×900 pixels. Although HRSC2016 includes 29 fine-grained categories, we followed the methodology in [14], testing on 19 categories including Nimitz (Nim.), Enterprise (Ent.), Arleigh Burke (Arl.), WhidbeyIsland (Whi.), Perry (Per.), Sanantonio (San.), Ticonderoga (Tic.), Admiral (Adm.), Austen (Aus.), Tarawa (Tar.), Container (Con.), Command ship (Com.), Car carrier A (CarA.), Container ship A (ConA.), Submarine (Sub.), Lute-shaped warship (War.), Medical ship (Med.), Car carrier B (CarB.) and Midway (Mid). Combining this evaluation with the TT-OBB dataset provided a comprehensive framework, significantly enhancing the reliability of the findings.

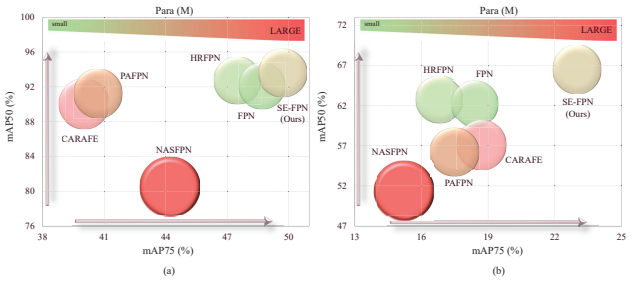


Figure 5. Performance comparison of multi-level feature fusion methods on (a) TT-OBB and (b) HRSC2016 datasets. The size and color of the circles represents the number of parameters (Para) in each method, with larger and darker circles indicating higher parameter counts. The position of each circle indicates the trade-off between model accuracy (mAP) and complexity.

4.2. Implementation details

For the training of remote sensing object detectors, experiments were conducted using the MMRotate [47] framework. Models were trained on both training and validation sets before testing on the testing set. The models were trained for 36 epochs on HRSC2016 and 20 epochs on TT-OBB, utilizing the AdamW [24] optimizer with an initial learning rate of 0.0001 and weight decay of 0.0005. Training was performed on eight RTX3090 GPUs with a batch size of eight.

Model performance was primarily evaluated using mean average precision at IoU thresholds of 50% (mAP50) and 75% (mAP75). While mAP50 provides a general measure of detection accuracy, it may not be stringent enough for tasks requiring precise angle estimation. In contrast, mAP75 demands tighter alignment between predicted bounding boxes and ground truth boxes, including accurate angle predictions, making it a more reliable performance indicator for oriented object detection tasks [42]. Additionally, we considered model parameters to evaluate detection efficiency and complexity.

4.3. Ablation study

Comparison of multi-level feature fusion strategies. To assess the effectiveness of SE-FPN in multi-scale feature fusion for fine-grained object recognition, we have compared it with advanced methods such as FPN, HRFPN, PAFPN, NASFPN, and CARAFE across two challenging datasets. As illustrated in Figure 5(a), SE-FPN enhances the mAP50 by 1.5% and mAP75 by 1.0% on the TT-OBB dataset compared to FPN. This demonstrates its capability for cross-scale feature integration.

SE-FPN shows significant advantages over other methods, achieving a maximum mAP50 of 93.6% in categories with substantial structural variation. While there is a slight increase in computational cost, SE-FPN effectively balances performance and efficiency, making it suitable for high-precision remote sensing applications. Despite the inherent challenges of transmission tower detection, SE-FPN continues to exhibit robust performance improvements.

On the HRSC2016 dataset, as shown in Figure 5(b), SE-FPN achieves 66.5% mAP50 and 23.0% mAP75, surpassing FPN within complex categories. Its consistent performance across various object types highlights its ability to manage diverse scales and complexities.Results from both datasets confirm SE-FPN's high flexibility in fine-grained recognition tasks, especially when detecting objects with varying shapes and complexities.

Comparison of different region proposal networks. Experiments on the TT-OBB and HRSC2016 datasets demonstrate significant improvements in region proposal accuracy with ERRPN. An analysis of 200 randomly selected region proposals indicates that ERRPN markedly en-

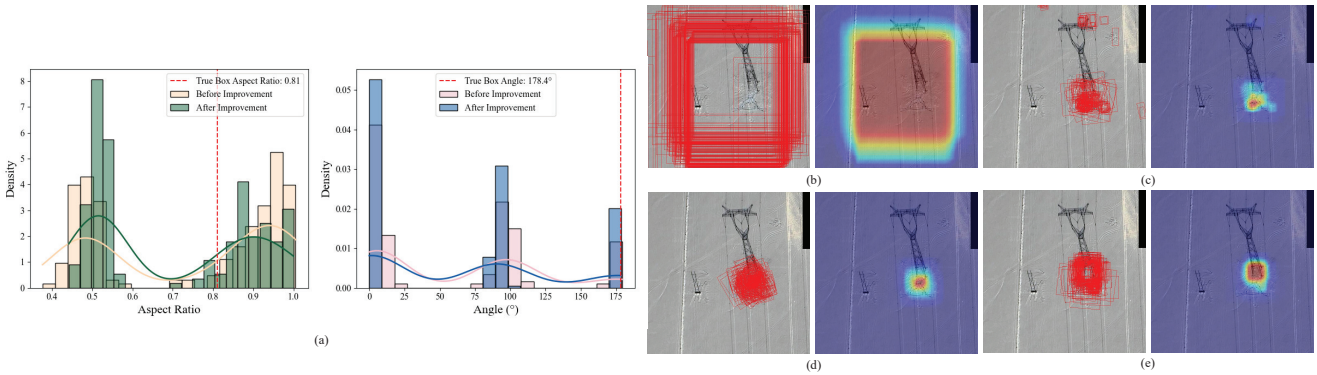


Figure 6. (a) Aspect ratio and angle distribution before and after ERRPN improvement on a sample from the TT-OBB dataset; Region proposal visualizations of different RPN variants on the same sample: (b) RPN, (c) RRPN, (d) Oriented RPN [36], and (e) ERRPN.

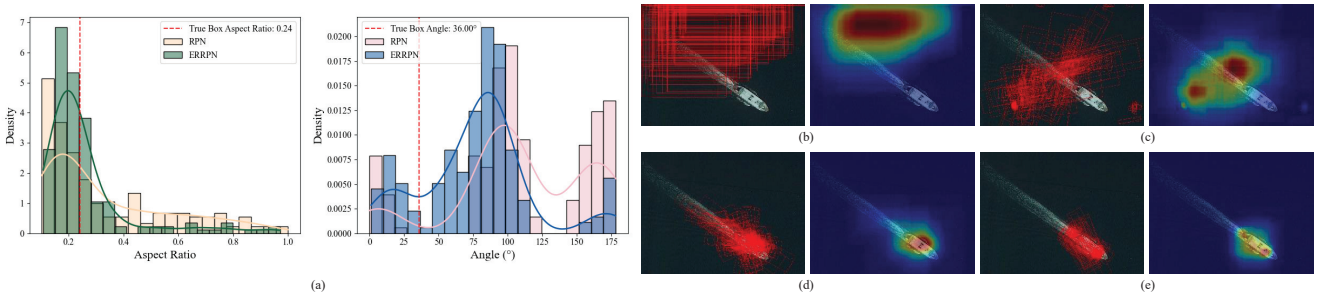


Figure 7. (a) Aspect ratio and angle distribution before and after ERRPN improvement on a sample from the HRSC2016 dataset; Region proposal visualizations of different RPN variants on the same sample: (b) RPN, (c) RRPN, (d) Oriented RPN, and (e) ERRPN.

hances both aspect ratio and angle prediction, as illustrated in Figure 6 and Figure 7.

In aspect ratio prediction, ERRPN produces results closely aligned with true object aspect ratios, exhibiting minimal deviation. This contrasts with RRPN, which displays a broader and less accurate distribution of predictions. In angle prediction, ERRPN also shows clear advancements, aligning predicted angles more closely with actual object orientations.

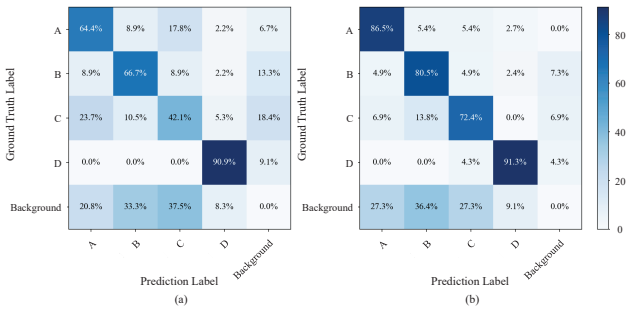


Figure 8. Comparison of normalized confusion matrices: before improvement - CrossEntropy Loss (a) and after improvement - ASLoss (b) on the TT-OBB dataset.

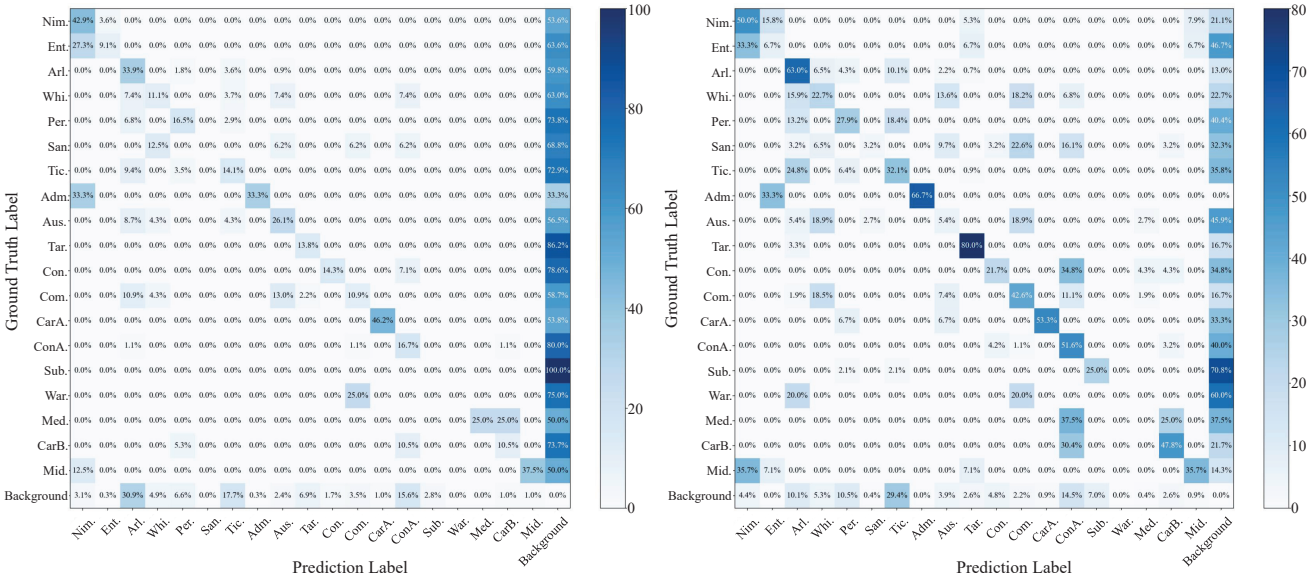
The impact of ERRPN is particularly evident in visualizations of region proposals. Before the improvement, RRPN generated scattered and overlapping proposals, which led to lower localization precision. After the improvement, the proposals become more focused and better aligned with object locations, reducing noise and improving accuracy—critical for detecting complex shapes and orientations. Furthermore, the heat map shows that ERRPN exhibits higher response values in central areas compared to the more dispersed responses of Oriented RPN.

Performance evaluation of ASLoss. Results from the TT-OBB dataset validate ASLoss's efficacy in enhancing classification accuracy, especially for objects with arbitrary orientations in remote sensing imagery. As shown in Table 2, mAP50 consistently increases as the threshold τ is raised from 0.3 to 0.7, while mAP75 reaches its peak at 48.4% when τ is set to 0.5. This indicates that a threshold of 0.5 may enable the model to more effectively manage lower IoU samples, thereby improving its performance in complex scenarios characterized by imprecise object boundaries.

The confusion matrix analysis (Figure 8 and Figure 9) for both the TT-OBB and HRSC2016 datasets demonstrates significant performance enhancements with ASLoss. On

S.	E.	A.	Nim.	Ent.	Arl.	Whi.	Per.	San.	Tic.	Adm.	Aus.	Tar.	Con.	Com.	CarA.	ConA.	Sub.	War.	Med.	CarB.	Mid.	mAP50 (%)	mAP75 (%)
✗	✗	✗	90.4	97.4	55.4	51.2	68.4	58.7	37.7	69.2	55.8	69.5	50.0	72.9	52.7	32.3	30.7	65.2	86.5	58.2	81.8	62.2	20.4
✓	✗	✗	51.6	80.5	68.9	74.1	66.9	**81.5**	**57.0**	47.2	52.1	78.8	64.5	66.7	97.7	55.3	44.4	41.5	72.5	80.8	82.6	66.5	23.0
✗	✓	✗	82.3	64.7	62.2	87.3	65.7	70.3	50.1	82.3	58.2	50.4	53.4	79.6	55.6	**70.7**	31.9	57.7	**92.0**	55.6	80.5	65.8	22.5
✗	✗	✓	57.8	89.6	61.1	73.5	60.0	69.7	45.8	57.4	50.6	70.9	49.4	65.9	86.8	68.7	55.6	44.2	81.3	53.5	73.4	64.0	20.0
✓	✓	✗	87.7	63.6	69.9	78.2	78.4	81.1	51.6	76.8	54.5	68.7	65.4	71.3	54.5	49.8	38.4	67.5	90.9	**81.8**	71.5	67.3	27.7
✓	✗	✓	68.4	81.6	72.6	**91.6**	**79.8**	75.6	42.1	74.8	62.7	53.5	**70.0**	72.5	56.6	64.9	**61.6**	51.4	91.4	56.6	75.5	68.6	24.2
✗	✓	✓	91.3	75.7	70.9	89.1	64.9	74.8	42.5	**87.5**	**67.4**	**82.3**	66.3	**81.2**	61.7	65.5	49.7	43.8	80.9	71.7	82.7	69.5	25.6
✓	✓	✓	**96.4**	**98.6**	**76.5**	86.7	76.2	61.7	49.1	79.1	44.1	67.1	61.2	67.7	**98.6**	57.5	38.8	**70.3**	74.7	59.2	**89.5**	**71.2**	**33.5**

Table 1. Ablation study results on the HRSC2016 dataset. S. represents SE-FPN, E. represents ERRPN, and A. represents ASLoss.



(a) (b)

Figure 9. Comparison of normalized confusion matrices: before improvement - CrossEntropy Loss (a) and after improvement - ASLoss (b) on the HRSC2016 dataset.

the TT-OBB dataset, accuracy for category A increases from 64.4% to 86.5%, while category C rises from 42.1% to 72.4%. Notably, there is a marked reduction in misclassifications between similar categories, particularly between A and C. Similarly, on the HRSC2016 dataset, ASLoss contributes to enhanced accuracy in categories such as ConA. and Tar., achieving accuracies of 51.6% and 80.0%, respectively. Moreover, the reduced occurrence of missed detections, particularly in the background category, highlights ASLoss's ability to better capture and classify previously overlooked objects. This significantly decreases the rate of false negatives.

τ	mAP50 (%)	mAP75 (%)
0.3	92.6	47.3
0.5	93.1	**48.4**
0.7	**93.5**	48.1

Table 2. Results of different threshold τ.

Overall, across both datasets, ASLoss consistently outperforms the previously used CrossEntropy loss method by dynamically adjusting the classification loss. This approach mitigates confusion among categories and enhances the model's capability to differentiate objects exhibiting subtle inter-class variations. These advancements render ASLoss particularly suitable for fine-grained classification tasks where substantial variation in object orientation and structure presents challenges for conventional methods.

Ablation study analysis. Ablation studies on the TT-OBB and HRSC2016 datasets reveal significant performance enhancements due to the integration of SE-FPN, ERRPN, and ASLoss, as depicted in the Figure 1 and Table 1. These improvements are particularly pronounced in fine-grained recognition tasks, with Rotated Faster R-CNN serving as the baseline model.

On the TT-OBB dataset, the comprehensive model that incorporates SE-FPN, ERRPN, and ASLoss attains an impressive mAP50 of 96.3% and an mAP75 of 62.0%. In con-

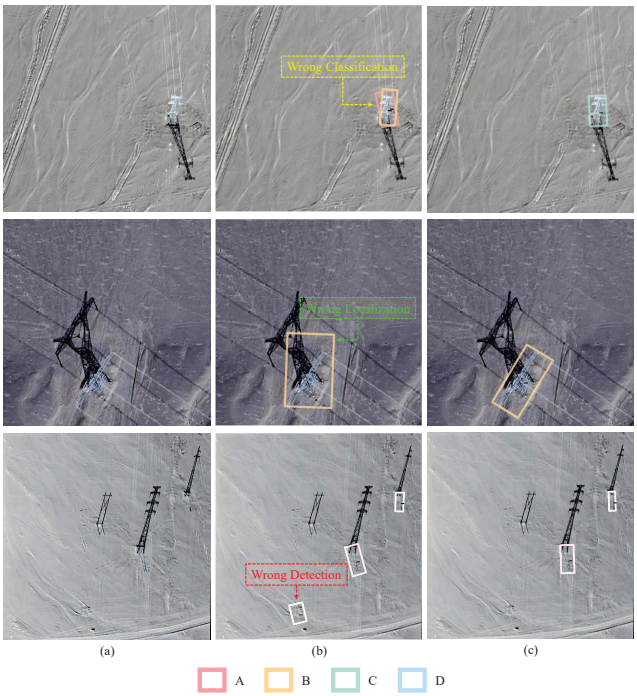


Figure 10. Visual results on TT-OBB dataset: (a) input image; (b) baseline; (c) GA-RFRCNN.

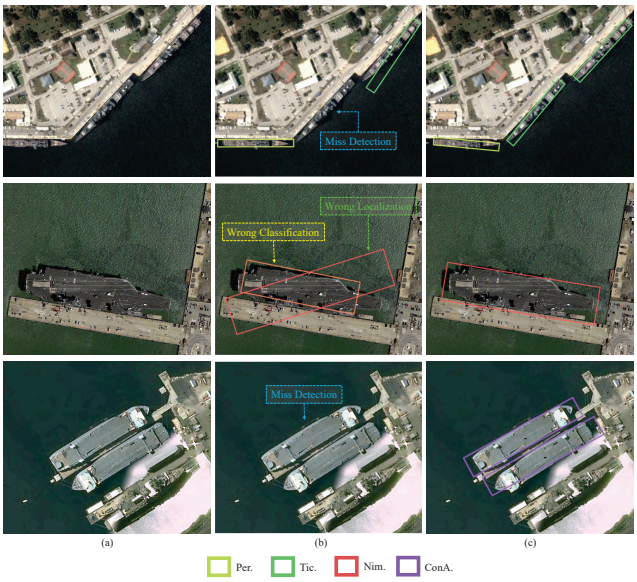


Figure 11. Visual results on HRSC2016 dataset: (a) input image; (b) baseline; (c) GA-RFRCNN.

trast, the baseline Rotated Faster R-CNN records an mAP50 of 92.1% and an mAP75 of 48.7%. The advancements are especially notable in challenging categories such as B and C.

On the HRSC2016 dataset, the full model achieves an mAP50 of 71.2% and an mAP75 of 33.5%, significantly surpassing the performance of the baseline Rotated Faster R-CNN (mAP50 of 62.2% and mAP75 of 20.4%). The improvements are most evident in fine-grained categories including CarA., Whi., ConA., and Arl..

These enhancements enable the model to more effectively capture subtle variations in object appearance, orientation, and scale, ultimately leading to improved performance and robustness across both datasets.

4.4. Comparison with other methods

Among the compared detectors, GA-RFRCNN achieves the highest mAP75, as shown in Table 3 and Table 4. It shows significant improvements over traditional models, particularly for objects with complex geometries and diverse forms. While two-stage methods such as RoI Transformer and O-RCNN exhibit commendable performance, they still fall short of GA-RFRCNN's efficacy. This underscores the effectiveness of the enhancements introduced, including SE-FPN, ERRPN, and ASLoss, in fine-grained recognition tasks. These improvements are especially evident when detecting slender structures and objects with intricate shapes and orientations. GA-RFRCNN also proves highly competitive with popular detectors like DETR, showing strong performance on both the custom TT-OBB dataset and publicly available datasets.

In visual comparisons on the TT-OBB and HRSC2016 datasets, the baseline Rotated Faster R-CNN frequently misclassifies and mislocalizes objects, particularly within cluttered environments. In contrast, GA-RFRCNN significantly enhances both classification accuracy and bounding box alignment. It reduces missed detections and errors in challenging scenarios when detecting transmission towers and ships, as shown in Figure 10 and Figure 11. These results confirm GA-RFRCNN's superior performance in remote sensing applications.

Model	A	B	C	D	mAP50 (%)	mAP75 (%)
DETR-based						
AO2-DETR [5]	87.4	89.4	87.1	87.7	86.9	45.6
ARS-DETR [41]	**99.8**	90.0	97.4	97.2	95.0	52.4
One-stage						
R3Det [38]	97.5	94.0	87.2	86.1	92.9	45.9
S2ANet [13]	98.7	**97.5**	86.9	87.4	94.5	45.1
YOLOv8x [10]	54.4	97.4	95.7	80.2	82.1	40.3
Two-stage						
G.V. [37]	93.3	80.8	88.9	93.5	89.1	44.3
RoI Trans [7]	98.5	95.9	93.6	98.8	95.2	47.1
O-RCNN [36]	88.1	84.7	93.3	94.1	90.1	53.8
GA-RFRCNN	98.7	95.5	**97.8**	**98.4**	**96.3**	**62.0**

Table 3. Experimental results on TT-OBB dataset.

Model	Nim.	Ent.	Arl.	Whi.	Per.	San.	Tic.	Adm.	Aus.	Tar.	Con.	Com.	CarA.	ConA.	Sub.	War.	Med.	CarB.	Mid.	mAP50 (%)	mAP75 (%)
DETR-based																					
AO2-DETR	74.6	89.7	66.1	81.3	79.9	67.4	51.2	82.5	48.3	79.1	55.3	71.6	70.1	54.3	51.6	77.1	61.2	40.1	84.8	51.3	23.1
ARS-DETR	76.3	93.7	68.4	84.2	**82.1**	69.7	53.1	**91.5**	55.7	**81.7**	56.8	78.4	72.5	56.1	**53.1**	**79.9**	63.4	41.8	87.1	58.9	**40.7**
One-stage																					
R3Det	66.3	86.0	53.3	74.5	73.1	58.5	41.1	70.9	38.1	69.6	45.4	62.5	61.4	39.5	41.6	71.0	52.4	30.4	76.9	57.6	21.7
S2ANet	73.3	91.2	69.3	80.5	78.4	63.0	50.9	82.7	44.0	75.8	53.0	70.1	69.4	56.1	51.2	77.4	60.3	38.9	83.1	55.5	27.2
YOLOv8x	64.9	70.4	60.2	71.3	68.3	55.2	40.8	70.5	35.7	65.8	42.5	62.7	58.3	45.0	38.1	69.1	50.1	30.0	72.9	49.7	18.9
Two-stage																					
G.V.	63.8	53.9	33.2	79.4	49.8	73.2	**59.7**	50.4	45.4	59.1	**74.1**	67.8	68.2	52.9	27.7	65.0	65.9	74.2	66.9	61.1	24.5
RoI Trans	79.4	74.9	76.9	45.3	74.2	65.2	55.6	77.0	**76.4**	39.8	39.4	**86.6**	61.2	**67.9**	33.1	31.5	68.1	40.0	87.6	62.2	32.6
O-RCNN	76.4	59.4	**86.7**	41.6	39.3	**76.2**	57.9	73.8	39.5	73.2	32.3	74.7	60.9	66.5	51.5	34.9	39.7	**74.2**	87.2	62.8	30.5
GA-RFRCNN	**96.4**	**98.6**	76.5	**86.7**	76.2	61.7	49.1	79.1	44.1	67.1	61.2	67.7	**98.6**	57.5	38.8	70.3	**74.7**	59.2	**89.5**	**71.2**	33.5

Table 4. Experimental results on HRSC2016 dataset.

5. Conclusion

In this study, we investigated the challenges associated with fine-grained object recognition in remote sensing imagery, emphasizing the intricate geometric structures, diverse forms, and challenging environments inherent to these tasks. To tackle these challenges, we propose GA-RFRCNN, which incorporates SE-FPN, ERRPN, and ASLoss to enhance detection accuracy and robustness across various datasets. These innovations not only demonstrate their efficacy in managing complex details and various object types within fine-grained recognition tasks but also underscore significant advancements in detecting vertically structured objects.

Although our methods were tested on satellite imagery with large viewing angles, which introduced certain detection constraints, future work will aim to incorporate additional contextual information, such as shadow features. This could further enhance object localization and category predictions, especially in environments with complex lighting conditions and subtle visual cues.

Acknowledgement

This study was funded by National Key Laboratory of Uranium Resources Exploration-Mining and Nuclear Remote Sensing (6142A012301).

References

[1] R. Ahmad. Smart remote sensing network for disaster management: an overview. *Telecommunication Systems*, 87:213–237, 2024. 2

[2] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao. Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27706–27716, 2024. 4

[3] N. Casagli, E. Intrieri, V. Tofani, G. Gigli, and F. Raspini. Landslide detection, monitoring and prediction with remote-sensing techniques. *Nature Reviews Earth & Environment*, 4(1):51–64, 2023. 2

[4] G. Cheng, Q. Li, G. Wang, X. Xie, L. Min, and J. Han. Sfrnet: Fine-grained oriented object recognition via separate feature refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–10, 2023. 2

[5] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5):2342–2356, 2022. 9

[6] Y. Di, Z. Jiang, and H. Zhang. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing*, 13(4):747, 2021. 2, 3

[7] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 9

[8] N. Ejaz and S. Choudhury. Computer vision in drone imagery for infrastructure management. *Automation in Construction*, 163:105418, 2024. 2

[9] J. Fan and M. A. Saadeghvaziri. Applications of drones in infrastructures: Challenges and opportunities. *International Journal of Mechanical and Mechatronics Engineering*, 13(10):649–655, 2019. 2

[10] J. Glenn. Yolov8, 2023. 9

[11] Q. Guan, Y. Liu, L. Chen, S. Zhao, and G. Li. Aircraft detection and fine-grained recognition based on high-resolution remote sensing images. *Electronics*, 12(14):3146, 2023. 2

[12] Z. Guo, B. Hou, X. Guo, Z. Wu, C. Yang, B. Ren, and L. Jiao. Msrip-net: Addressing interpretability and accuracy challenges in aircraft fine-grained recognition of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 2, 3

[13] J. Han, J. Ding, J. Li, and G.-S. Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 9

[14] Y. Han, X. Yang, T. Pu, and Z. Peng. Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021. 2, 6

[15] Z. Huang, F. Wang, H. You, and Y. Hu. Shadow information-based slender targets detection method in optical satellite images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 2, 3

[16] Z. Huang, F. Wang, H. You, and Y. Hu. Stc-det: A slender target detector combining shadow and target information in

optical satellite images. *Remote Sensing*, 13(20):4183, 2021. 2

[17] Z. Huang, F. Wang, H. You, and Y. Hu. Imaging parameters-considered slender target detection in optical satellite images. *Remote Sensing*, 14(6):1385, 2022. 2, 3

[18] K. Kaku. Satellite remote sensing for disaster management support: A holistic and staged approach based on case studies in sentinel asia. *International Journal of Disaster Risk Reduction*, 33:417–432, 2019. 2

[19] M. Kucharczyk and C. H. Hugenholtz. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sensing of Environment*, 264:112577, 2021. 2

[20] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li. Large selective kernel network for remote sensing object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16794–16805, 2023. 4

[21] W. Liang, J. Li, W. Diao, X. Sun, K. Fu, and Y. Wu. Fgatr-net: Automatic network architecture design for fine-grained aircraft type recognition in remote sensing images. *Remote Sensing*, 12(24):4187, 2020. 2

[22] Y. Liu. Application of remote sensing technology in smart city construction and planning. *Journal of Physics: Conference Series*, 2608(1):012052, 2023. 2

[23] Z. Liu, L. Yuan, L. Weng, and Y. Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, pages 324–331, 2017. 3

[24] I. Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[25] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018. 4

[26] M. Osswald-Cankaya and H. Mayer. Fine-grained airplane recognition in satellite images based on task separation and orientation normalization. In *2023 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6545–6548, 2023. 3

[27] S. Panda, V. S. Yadav, and V. K. Tripathi. Application of remote sensing in natural resource management. *Sustainable Development and Geospatial Technology*, 2:173–180, 2024. 2

[28] J. Shermeyer, T. Hossler, A. Van-Etten, D. Hogan, R. Lewis, and D. Kim. Rareplanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 207–217, 2021. 3

[29] S. Song, R. Zhang, M. Hu, and F. Huang. Fine-grained ship recognition based on visible and near-infrared multimodal remote sensing images: dataset, methodology and evaluation. *Computers, Materials & Continua*, 79(3):5243–5271, 2024. 2

[30] P. Sun, Y. Zheng, W. Wu, W. Xu, and S. Bai. Metric-aligned sample selection and critical feature sampling for oriented object detection. *arXiv preprint arXiv:2306.16718*, 2023. 2

[31] P. Sun, Y. Zheng, W. Wu, W. Xu, S. Bai, and X. Lu. Learning critical features for arbitrary-oriented object detection in remote sensing optical images. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2024. 2

[32] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu, M. Weinmann, S. Hinz, C. Wang, and K. Fu. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 2

[33] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 4

[34] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang. Ghostnetv2: Enhance cheap operation with long-range attention. *Advances in Neural Information Processing Systems*, 35:9969–9982, 2022. 4

[35] Y. Xi, Y. Liu, T. Li, J. Ding, Y. Zhang, S. Tarkoma, Y. Li, and P. Hui. A satellite imagery dataset for long-term sustainable development in united states cities. *Scientific Data*, 10(1):866, 2023. 2

[36] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3520–3529, 2021. 7, 9

[37] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1452–1459, 2020. 9

[38] X. Yang, J. Yan, Z. Feng, and T. He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3163–3171, 2021. 9

[39] Y. Yu, X. Yang, Q. Li, Y. Zhou, F. Da, and J. Yan. H2rbox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection. *Advances in Neural Information Processing Systems*, 36:1–14, 2024. 2

[40] Z. Yu, H. Huang, W. Chen, Y. Su, Y. Liu, and X. Wang. Yolo-facev2: A scale and occlusion aware face detector. *arXiv preprint arXiv:2208.02019*, 2022. 5

[41] Y. Zeng, Y. Chen, X. Yang, Q. Li, and J. Yan. Ars-detr: Aspect ratio-sensitive detection transformer for aerial oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. 9

[42] Y. Zeng, Y. Guo, and J. Li. Recognition and extraction of high-resolution satellite remote sensing image buildings based on deep learning. *Neural Computing and Applications*, 34(4):2691–2706, 2022. 6

[43] W. Zha, L. Hu, C. Duan, and Y. Li. Semi-supervised learning-based satellite remote sensing object detection method for power transmission towers. *Energy Reports*, 9:15–27, 2023. 2

[44] W. Zha, L. Hu, Y. Sun, and Y. Li. Engd-bifpn: A remote sensing object detection model based on grouped deformable convolution for power transmission towers. *Mul-*

timedia Tools and Applications, 82(29):45585–45604, 2023. 2

[45] X. Zhang, Y. Lv, L. Yao, W. Xiong, and C. Fu. A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1271–1285, 2020. 3

[46] G. Zhao, P. Yao, L. Fu, Z. Zhang, S. Lu, and T. Long. A deep learning method based on two-stage cnn framework for recognition of chinese reservoirs with sentinel-2 images. *Water*, 14(22):3755, 2022. 3

[47] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu, W. Zhang, and K. Chen. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7331–7334, 2022. 6

[48] S. Zhu and K. Yang. Deep remote sensing object detection for smart city applications. In *2024 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII)*, pages 167–171, 2024. 2