remote sensing

MDPI

Article

YOLOv9-GDV: A Power Pylon Detection Model for Remote Sensing Images

Ke Zhang [1,2,3], Ningxuan Zhang [1], Chaojun Shi [1,2,3,*], Qiaochu Lu [4], Xian Zheng [4], Yujie Cao [1], Xiaoyun Zhang [1] and Jiyuan Yang [1]

[1] Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China; zhangkeit@ncepu.edu.cn (K.Z.); znx@ncepu.edu.cn (N.Z.); cyj@ncepu.edu.cn (Y.C.); zxy@ncepu.edu.cn (X.Z.); yjy@ncepu.edu.cn (J.Y.)
[2] Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding 071003, China
[3] Hebei Engineering Research Center of Intelligent Technology for Power Internet of Things, North China Electric Power University, Baoding 071003, China
[4] Department of Electrical and Electoinic Engineering, North China Electric Power University, Beijing 102206, China; lqc@ncepu.edu.cn (Q.L.); zhengxian0815@ncepu.edu.cn (X.Z.)
* Correspondence: scj@ncepu.edu.cn

Abstract

Under the background of continuous breakthroughs in the spatial resolution of satellite remote sensing technology, high-resolution remote sensing images have become a frontier data source for intelligent inspection research of power infrastructure. To address existing issues in remote sensing image application algorithms such as difficulties in power target feature extraction, low detection accuracy, and false positives/missed detections, this paper proposes the YOLOv9-GDV power tower detection algorithm specifically for power tower detection in high-resolution satellite remote sensing images. Firstly, under high-resolution imaging conditions where transmission tower features are prominent, a Global Pyramid Attention (GPA) mechanism is proposed. This mechanism enhances global representation capabilities, enabling the model to better understand object–background relationships and effectively integrate multi-scale spatial information, thereby improving detection accuracy and robustness. Secondly, a Diverse Branch Block (DBB) is embedded in the feature extraction–fusion module, which enriches the feature space by enhancing the representation capability of single-convolution operations, thereby improving model feature extraction performance without increasing inference time costs. Finally, the Variable Minimum Point Distance Intersection over Union (VMPDIoU) loss is proposed to optimize the model's loss function. This method employs variable input parameters to directly calculate key point distances between predicted and ground-truth boxes, more accurately reflecting positional differences between detection results and reference targets, thus effectively improving the model's mean Average Precision (mAP). On the Satellite Remote Sensing Power Tower Dataset (SRSPTD), the YOLOv9-GDV algorithm achieves an mAP of 80.2%, representing a 4.7% improvement over the baseline algorithm. On the multi-scene high-resolution power transmission tower dataset (GFTD), the algorithm obtains an mAP of 94.6%, showing a 2.3% improvement over the original model. The significant mAP improvements on both datasets validate the effectiveness and feasibility of the proposed method.

Keywords: remote sensing imagery; power tower; YOLOv9-GDV; Global Pyramid Attention; Diverse Branch Block; VMPDIoU

Academic Editors: Lingyan Ran, Tao Zhuo, Shizhou Zhang and Fa-Ming Fang

Received: 20 May 2025
Revised: 20 June 2025
Accepted: 26 June 2025
Published: 29 June 2025

Citation: Zhang, K.; Zhang, N.; Shi, C.; Lu, Q.; Zheng, X.; Cao, Y.; Zhang, X.; Yang, J. YOLOv9-GDV: A Power Pylon Detection Model for Remote Sensing Images. *Remote Sens.* **2025**, *17*, 2229. https://doi.org/10.3390/rs17132229

1. Introduction

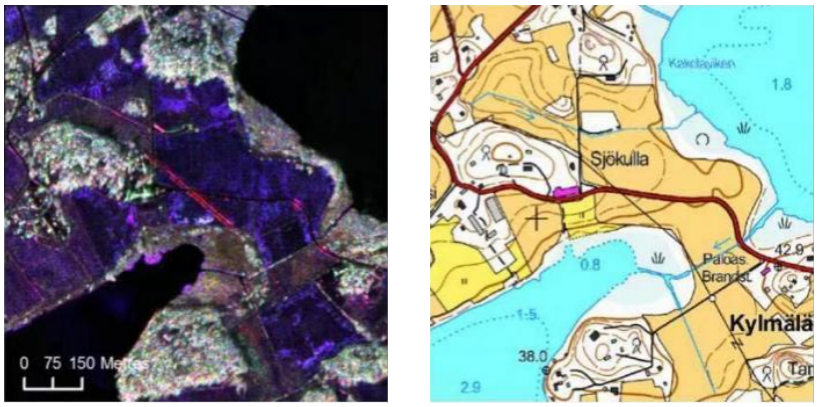
With the rapid advancement of satellite remote sensing technology, high-resolution satellite imagery has been extensively adopted in land planning, natural disaster risk assessments, environmental monitoring, and various domains including territorial resource management, urban planning, and agricultural resource surveys. Over the past five years, satellite remote sensing imagery has provided enhanced guidance solutions for power line inspection [1]. For power grid inspection applications, satellite remote sensing data primarily originates from two sources: optical remote sensing satellites and synthetic aperture radar (SAR) satellites [2]. Remote sensing images are further categorized into high-, medium-, and low-resolution based on their spatial resolution. High-resolution images are more widely utilized due to their superior precision and shorter revisit times. Each data source exhibits distinct characteristics in acquired high-resolution remote sensing imagery.

Optical remote sensing satellites operate similarly to human vision by passively capturing reflected sunlight from Earth's surface. The optical satellite image is shown in Figure 1. While their imagery offers intuitive visualization and clear interpretability, it suffers from significant meteorological constraints, such as obstruction by clouds, fog, or darkness. In power line inspection, sub-meter-level high-resolution optical satellite imagery enables the precise identification of typical objects near transmission lines. This capability facilitates object extraction and distance measurement between targets and power towers, thereby supporting safety hazard assessments.

(**a**) Optical satellite image (**b**) Topographic map data

Figure 1. (**a**) RapidEye optical satellite image with a pixel resolution of 8 m × 8 m. It includes material from © (2013) BlackBridge, LLC. All rights reserved. The high-voltage power line corridor within the forested environment is discernible in the optical satellite image. (**b**) Topographic map data © National Land Survey of Finland, 2015.

Compared to optical satellites, synthetic aperture radar (SAR) satellites represent a later-developed active remote sensing modality [3]. They operate by transmitting electromagnetic waves multiple times toward Earth's surface and imaging through echo signal processing, enabling the indirect measurement of target characteristics. The SAR image is shown in Figure 2. SAR satellites demonstrate distinct advantages in all-weather and day–night operational capabilities, coupled with partial ground penetration capacity, thereby compensating for limitations inherent in optical and infrared remote sensing systems. However, conventional SAR satellites typically operate on a weekly revisit cycle with lower data refresh rates, making them particularly suitable for long-term large-scale surveys. Their technical strengths prove advantageous for monitoring vast or remote geographical areas [4].



(**a**) Airborne L-band SAR image (**b**) Topographic map data

Figure 2. (**a**) E-SAR airborne L-band SAR image with a pixel resolution of 1 m × 1 m. Original image data © DLR and Astrium GmbH 2000, image processing by the Finnish Geospatial Research Institute (FGI). Partial power line infrastructure, represented by red linear features in the central region, is discernible in the SAR image. (**b**) Topographic map data sourced from the National Land Survey of Finland (2015).

Power towers, as critical components of power transmission systems, require the precise detection of their spatial distribution and operational status to ensure grid stability and maintain national energy security. Remote sensing-based power tower detection technology provides innovative approaches for monitoring tower distribution patterns and structural variations, offering substantial support for transmission line inspection workflows. This methodology enhances field inspection efficiency by delivering systematic decision-making assistance to maintenance personnel through automated feature extraction and spatial analysis.

Traditional power tower monitoring methods primarily rely on manual inspections and low-resolution image analysis, which are constrained by inefficiency, high operational costs, and limited accuracy. Manual inspections are not only labor-intensive but also restricted by geographical obstacles and meteorological conditions, often resulting in delayed or incomplete data acquisition. Furthermore, low-resolution imaging fails to meet modern requirements for refined tower management, demonstrating insufficient capability in precise localization and condition assessments. With the industry's transition from digital to intelligent inspection frameworks, unmanned aerial vehicle (UAV)-based inspection has become the dominant paradigm due to its high efficiency, operational safety, and deployment flexibility [5]. UAVs enable real-time acquisition of high-resolution imagery and multispectral data, significantly improving inspection accuracy and throughput. However, technical limitations such as vulnerability to meteorological conditions and limited endurance restrict UAVs' scalability for large-scale power line monitoring, introducing potential risks to grid stability during long-term surveillance operations [6].

With the continuous improvement of the spatial resolution of satellite remote sensing images and the shortening of revisit cycles, researchers have found that it is now possible to detect and recognize large-scale power infrastructure such as power poles and substations from sub-meter optical satellite remote sensing images. Compared to drone inspections, satellite-based intelligent inspection of power lines can achieve large-scale, business-oriented inspections of power corridors, greatly improving the efficiency and specificity of inspections. However, most current research has focused on using satellite remote sensing images for tasks such as vegetation erosion and disaster monitoring in power corridors, with less attention paid to the identification of targets such as power poles. Therefore, deep learning-based power pole detection technology for high-resolution

satellite remote sensing images is expected to break through the bottlenecks of traditional detection methods and provide a better option for power inspections.

This paper conducts in-depth research and improves the transmission line power pole detection task based on the YOLOv9 model. By applying the GPA attention mechanism in the YOLOv9 network, embedding the Diversity Branch Block (DBB) in the core module RepNCSPELAN, and using VMPDIoU to improve the model's loss function, we design an efficient and robust power pole detection model for transmission lines. The main contributions of this paper include the following:

(1) A power pole detection model for transmission lines based on YOLOv9 is proposed, where improvements are made to the network architecture and attention mechanism of the baseline model YOLOv9. A new attention mechanism, GPA, is introduced into the YOLOv9 network, which brings global representation capabilities to the model.

(2) A Diversity Branch Block (DBB) is embedded in the core module RepNCSPELAN of YOLOv9 to enhance the representation capability of individual convolutions, enrich the feature space, and improve the model's feature extraction ability.

(3) The VMPDIoU is proposed to improve the model's loss function. By varying inputs and directly calculating the key point distances between the predicted boxes and ground-truth boxes, this method can more accurately reflect the differences between the predicted and true boxes, thus enhancing the model's mean accuracy.

2. Related Works

Over the past few years, satellite remote sensing imagery has provided valuable research insights for power line inspection. Scholars have explored numerous new approaches and methodologies for power line inspection by utilizing satellite remote sensing data. Regarding the application of satellite remote sensing data in power line inspection, it primarily involves two data sources: optical remote sensing satellites and synthetic aperture radar (SAR) satellites.

For the application of optical remote sensing satellite imagery in power line inspection, notable examples include the following: Mikhalap et al. (2019) [7] adopted the NDVI threshold as a vegetation encroachment detection method, inspecting overhead power line corridors spanning approximately 550 km in the Pskov region of Russia. Through GIS systems, they identified that 84% of the power line corridors required management. The study by [8] combined NDVI with GIS information analysis. First, satellite images were classified into different regions based on NDVI values. Then, the stereo matching method was employed to estimate tree heights around transmission towers. Figure 3 shows two distinct left and right images with positional coordinates (x, y, z), both corresponding to the same ground point P(x, y, z). This method generates a depth map and integrates GIS information technology. Since GIS data contains the geographic coordinates of transmission towers, analyzing the GIS information of the images enables vegetation height estimation.

Another method for detecting vegetation encroachment in power line corridors is the object detection-based (ODB) approach, where the targets are typically power line corridors or transmission tower bases. The study by [9] implements the ODB method on Google Map images to extract and detect power line corridors. The process involves loading target images containing all relevant information, converting them to grayscale, and applying filtering. To eliminate irrelevant data, a transmission tower library is created by separately extracting the average histogram of each tower. Finally, targets are extracted pixel by pixel and matched with the tower library. Bounding boxes are then generated around the detected transmission towers, and paths between towers are plotted to extract the corridor from the image.

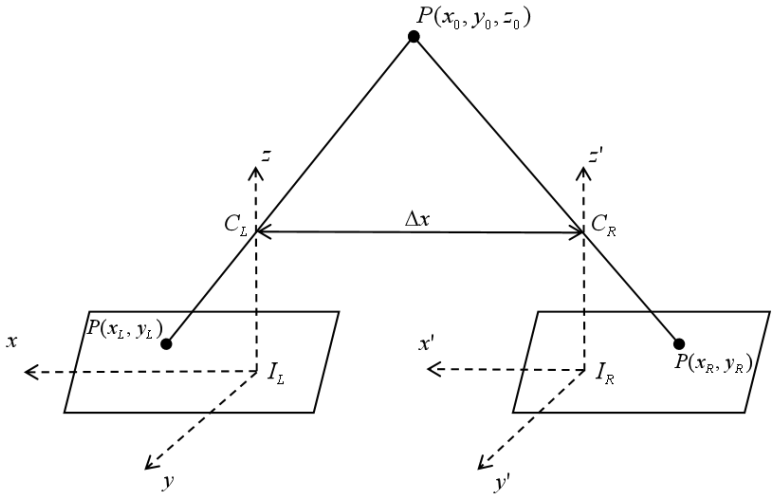


Figure 3. Schematic diagram of stereo matching technology (Adapted with permission from Ref. [8]. Copyright year 2014, copyright owner's name: Abdul Qayyum).

The applications of SAR imagery primarily focus on two aspects: deformation detection of power towers using SAR data [10] and disaster monitoring around power lines. Prior to 2015, SAR satellites were primarily employed to investigate the scattering characteristics of transmission towers. As early as 2000, Sarabandi et al. [11] proposed a statistical polarization detection algorithm that significantly improves the signal-to-clutter ratio. This algorithm uses the coherence between co-polarized and cross-polarized backscatter components as the detection parameter. In addition to using SAR satellites for studying the scattering characteristics of transmission line towers, SAR satellite imagery has also been applied for the indirect monitoring of infrastructure safety, including power lines. In 2025, the study by [12] proposed a hybrid method that combines large-scale vision models and improved small models for few-shot detection of collapsed transmission towers.

Traditional optical satellite remote sensing image target detection primarily relies on machine learning and image processing technologies, with a workflow that includes region selection, feature extraction, and classifiers. The feature extraction phase involves information such as image grayscale values, textures, and ground object spectral data, all of which play crucial roles in target detection. Machine learning algorithms are currently being applied in power line inspection based on satellite remote sensing imagery. Uehara et al. [9] utilized higher-order local autocorrelation image features to extract spatial and spectral relationships from multispectral satellite images while employing AdaBoost as a classifier for patch images, achieving approximately 90% precision and 80% recall. Prakash et al. [13] used an active learning technique on 150,000 square kilometers of multispectral imagery from World-View2 in southeastern Australia, employing supervised pre-labeled data to detect transmission towers, ultimately achieving 80% precision and 50% recall for the transmission tower detector. Rohrer et al. [14] proposed an automated method that estimates power line locations in images using indirect indicators of electrical infrastructure such as MODIS land cover data and nighttime light data. However, traditional optical satellite remote sensing target detection algorithms are only suitable for specific environments and backgrounds, demonstrating poor generalization capability. They are mainly used for detecting larger targets like transmission towers with little coverage of distribution towers and struggle to effectively address challenges in satellite remote sensing imagery such as high noise levels, complex backgrounds, and varying target scales.

With the rapid development of deep learning, researchers have found that applying deep learning-based target detection methods can better achieve the extraction of electrical targets in satellite remote sensing imagery. Deep learning-based target detection methods

automatically extract features from images through convolutional neural networks (CNNs), enabling hierarchical and high-dimensional feature representations.

Current mainstream target detection methods are mainly divided into two categories. The first category consists of region-based two-stage object detection algorithms represented by R-CNN [15], Fast R-CNN [16], and Faster R-CNN [17]. Their core idea involves first generating region proposals using selective search methods, followed by regression and classification on these proposals. The second category includes one-stage detection algorithms represented by SSD [18], RetinaNet [19], and the YOLO series [20–28]. These approaches formulate the detection problem as a regression task, generating anchor boxes of varying sizes and aspect ratios at each position on specific feature maps to predict target class probabilities and locations. The YOLO algorithm has evolved through multiple generations, from the initial YOLOv1 to the YOLOv9 used in this paper, offering richer functionality, higher classification accuracy, and faster detection speeds.

The application of deep learning-based target detection methods in satellite remote sensing imagery has achieved significant progress. Hu et al. [29] constructed four power tower subsets from different geographic locations based on the power transmission and distribution infrastructure image dataset [30] and explored the performance differences of Faster R-CNN, YOLOv2, and RetinaNet in automatically detecting power towers from satellite remote sensing images. RetinaNet ultimately demonstrated the best performance, achieving 47% precision and 60% recall. The study also revealed that the satellite imagery resolution must reach at least 0.3m to detect at least half of the power towers. Haroun F. et al. [31] employed the RetinaNet deep learning model to detect transmission tower locations in satellite imagery, achieving a mean Average Precision (mAP) of 72.45% at an IoU threshold of 0.5. They further developed a routing algorithm that extracts power corridor areas by creating virtual paths between each pair of detected adjacent transmission towers. Current deep learning-based power tower detection algorithms in satellite remote sensing imagery primarily leverage the advantages of typical detection models but still fail to adequately address challenges such as small target sizes, multi-scale characteristics, and complex backgrounds specific to power towers in satellite imagery.

The features of optical satellite remote sensing images can generally be divided into high-level features and low-level features. High-level features represent abstract semantic information, while low-level features contain detailed information such as spectral and texture characteristics. In CNN networks, the distribution of features across different layers is related to the scale of the targets. In satellite remote sensing imagery of power towers, the significant scale differences between distribution towers and transmission towers make it challenging to synchronize their feature propagation to deeper network layers. Additionally, deep learning-based target detection models primarily use the final layer features extracted by convolutional neural networks for classification and localization, resulting in reduced small-target information in the final network output. This leads to lower detection accuracy for small targets like distribution towers. Therefore, leveraging multi-scale feature maps and designing multi-scale feature fusion modules becomes particularly critical.

Hou et al. [32] designed a feature fusion strategy through cascading to integrate high-level semantic information with low-level detailed information, enhancing multi-scale feature representation and mitigating the information loss of small targets during CNN propagation. Fu et al. [33] proposed a feature fusion architecture to generate multi-scale feature hierarchies, incorporating top-down and bottom-up pathways to blend shallow-layer features with high-level features. Zhu et al. [34] introduced the TPH-YOLO model, which enhances YOLOv5's prediction network using transformers and self-attention mechanisms to achieve effective multi-scale target detection. Zhou et al. [35] developed an attention multi-hop graph and multi-scale convolution fusion network (AMGCFN), which includes

a full multi-scale CNN and a multi-hop GCN, to extract hierarchical information from hyperspectral images.

Another challenge in object detection of power tower satellite remote sensing images lies in the interference from complex backgrounds. Power towers are highly susceptible to false detection due to the color, shape, and other characteristics of similar objects. Hong et al. [36] proposed a high-resolution domain adaptation network (HighDAN), which captures multi-scale image representations from parallel high-to-low resolution subnetworks. This network efficiently generates repetitive information across resolutions and bridges differences between remote sensing images under varying backgrounds through adversarial learning, effectively mitigating background interference. To better handle multi-source remote sensing data, Hong et al. [37] further designed a universal remote sensing foundation model, SPectralGPT. This model adapts to input images with diverse sizes, resolutions, temporal sequences, and regions through a progressive training strategy while employing multi-objective reconstruction to capture spectral sequential patterns for the comprehensive utilization of remote sensing data across scenarios.To address issues such as insufficient feature representation and background confusion, Zhang et al. [38] proposed an efficient algorithm called FFCA-YOLO. It improves the network's capabilities in local area awareness, multi-scale feature fusion, and global association across channels and space, enhancing the feature representation of small objects in remote sensing images and suppressing confusing background interference.To address complex background interference, Zhang et al. [39] proposed a context-aware detection network (CAD-Net). This network integrates attention-modulated features with global and local contextual information to adapt to environmental variations around targets. Wang et al. [40] developed a representation-enhanced state replay network, which jointly optimizes parameters across different branches to enhance the interactive fusion of information between heterogeneous remote sensing images.

In addition, remote sensing image generation models have gradually become a widely recognized research direction. In 2025, Yang et al. [41] proposed a multi-class and multi-scale object (MMO) image generator called MMO-IG, based on deep generative models (DGMs). This model is capable of generating remote sensing images with supervised object labels from both global and local perspectives. MMO-IG achieves precise modeling, rational spatial distribution, and global guidance of object instances through the collaborative use of ISIM, SCDKG, and SODI. This not only expands the dataset size but also significantly enhances the effectiveness and representational quality of the data, thereby improving the performance of remote sensing object detection.

In summary, multi-scale detection and complex background interference hinder the development of power tower object detection algorithms in satellite remote sensing images. Most existing power tower detection methods still rely on basic single-stage or two-stage mainstream object detection models, which perform poorly in small object detection and fail to adequately address challenges such as multi-scale feature fusion and model generalization in complex backgrounds.

3. Methods

To address the aforementioned challenges, prior to the formal commencement of this study, we compared and evaluated various mainstream object detection models. The baseline YOLOv9 model demonstrated relatively strong performance in terms of accuracy, recall, and robustness. As the latest iteration of the YOLO series, YOLOv9 inherits the efficiency and accuracy of its predecessors while introducing several innovations in model architecture. Therefore, this study ultimately selected the YOLOv9 network as the foundational framework to implement modifications tailored for power tower targets in remote

sensing imagery, aiming to further enhance its detection performance under the specified research conditions.

3.1. Architecture of the YOLOv9 Model

The YOLOv9 model's core concept is to predict object categories, locations, and confidence scores across multi-scale feature maps through a single forward pass. The model incorporates the Programmable Gradient Information (PGI) framework to address the diverse variations required by deep networks for multi-objective optimization. PGI provides complete input information for target tasks to compute loss functions, thereby deriving reliable gradient information for updating network weights. Additionally, based on gradient path planning, the model introduces a novel lightweight network architecture—the generalized efficient layer aggregation network (GELAN). Its design validates that PGI achieves superior results on lightweight models.

YOLOv9's architecture draws inspiration from YOLOv5, YOLOv6, YOLOv7, and YOLOv8. Its core module, RepNCSPELAN, integrates the CSPNet Block from YOLOv5, the Rep module from YOLOv6, and the ELAN module from YOLOv7. For training, the task-aligned assigner from YOLOv8 is adopted for positive/negative sample selection, while the DFLLoss function is referenced for loss computation.

3.2. YOLOv9-GDV Model Architecture

This paper presents an improved version of the YOLOv9 baseline model, resulting in the enhanced YOLOv9-GDV network architecture shown in Figure 4. The improvements include (1) the integration of the GPA attention mechanism into the YOLOv9 baseline network to enhance global representation capabilities; (2) the incorporation of the Diverse Branch Block (DBB) into the core RepNCSPELAN module; and (3) the adoption of the VMPDIoU loss to refine the model. By directly computing key point distances between predicted and ground-truth bounding boxes, this loss function more accurately reflects their discrepancies, thereby improving the model's average accuracy.

3.3. Global Pyramid Attention

For remote sensing image object detection models, introducing global representation capabilities is crucial. In remote sensing images, detection targets are typically small, and the detection scenarios are often complex, with occlusions, overlaps, or multi-category objects. Without understanding and distinguishing background information, numerous interference factors may arise. Global representation capability helps the model comprehend the relationship between objects and the background, enabling the effective separation and identification of different targets in complex visual environments, thereby improving detection accuracy. Models with global representation capabilities can better integrate information from different scales and locations, enhancing detection performance and robustness.

Given high-resolution satellite remote sensing images of power towers, this study improves the network architecture and attention mechanism of the baseline model YOLOv9 by designing an efficient multi-scale attention mechanism module. This introduces global representation capability to the model, helping it understand the relationship between objects and the background. As a result, the model can better process remote sensing image data and separate detection targets from complex backgrounds. The specific structure is shown in Figure 5.

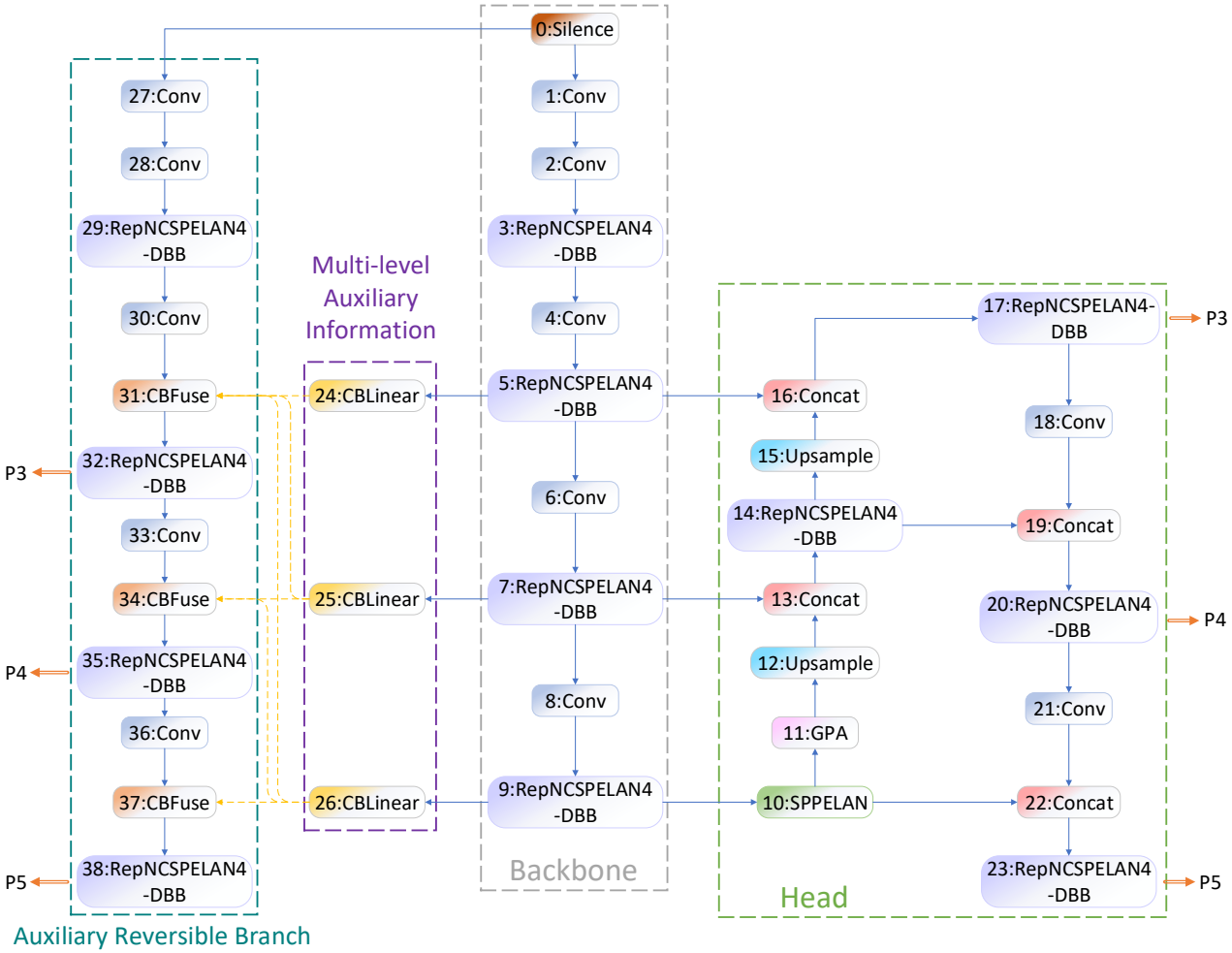


Figure 4. YOLOv9-GDV network architecture diagram.

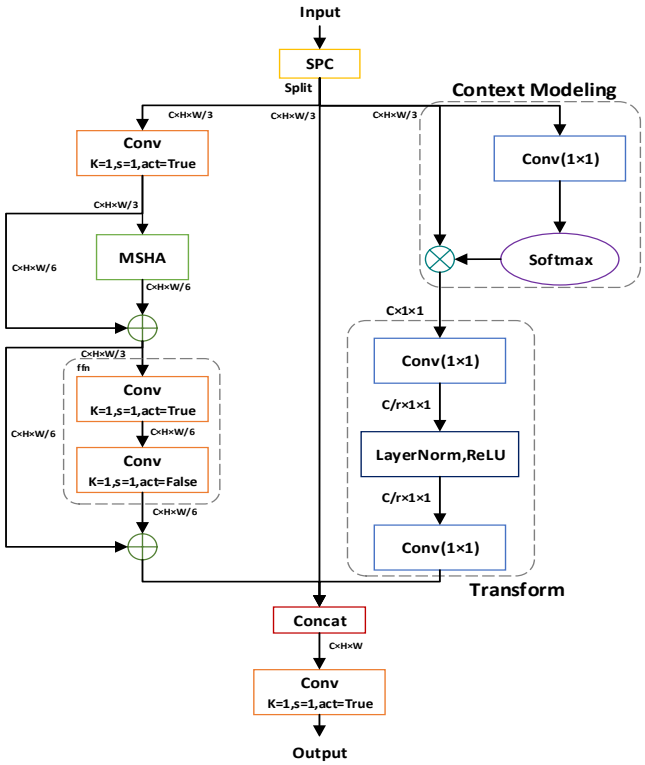


Figure 5. Global Pyramid Attention.

We name this partial self-attention mechanism Global Pyramid Attention (GPA), which can be applied to tasks such as image classification and object detection. GPA works by concatenating convolution results from kernels of different sizes to form a pyramid-shaped feature map, then applying attention mechanisms on this feature map. Simultaneously, a global context module captures global information by performing global average pooling on the original input feature map, enhancing its representational capacity to produce a globally weighted feature map. The two weighted feature maps are then connected to the original feature map via residual connections to extract richer feature information.

The GPA module is implemented through the following steps:

1. Feature Partitioning: Use the SPC module to uniformly divide the input features into three parts via a 1 × 1 convolution.

2. Take one partitioned feature and apply global average pooling to obtain a global context vector. Perform feature transformation: Reduce channel dimensions via a 1 × 1 convolution, introduce non-linearity with ReLU activation, restore channels with another 1 × 1 convolution, and generate channel attention weights using a Sigmoid function. Enhance features by multiplying the weights with the input features to produce a weighted feature map.

3. Feed another partitioned feature into the NPSA block, which consists of a multi-head self-attention (MHSA) module and a feed-forward network (FFN).

4. Concatenate the three partitioned features and fuse them via a 1 × 1 convolution to generate the final output.

The PSA module is only placed after SPPELAN to avoid excessive computational overhead caused by the quadratic complexity of self-attention. This design introduces global representation learning to the YOLO model at a low computational cost, enhancing model capability and performance.

Through these precision-driven designs, the YOLO model's performance is improved without significantly increasing computational costs.

The specific derivation process is as follows:

1. First, let the input feature map be $X \in \mathbb{R}^{C\times H\times W}$. The input feature map is partitioned into three components by the channel-splitting SPC module.

$$X = [X_1, X_2, X_3], \quad X_i \in \mathbb{R}^{\frac{C}{3}\times H\times W} \tag{1}$$

Part 1 performs local feature enhancement. Part 2 preserves the original information to avoid feature shift, and Part 3 conducts global context modeling. Finally, the three feature branches are concatenated and fused through a 1 × 1 convolution to produce the output.

2. The input to the local feature enhancement branch is $X_1 \in \mathbb{R}^{\frac{C}{3}\times H\times W}$. First, convolutional dimensionality reduction is performed by applying a X_1 convolution to a 1 × 1 convolution, with the ReLU activation function:

$$X_1' = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times 1}(X_1))) \in \mathbb{R}^{\frac{C}{3}\times H\times W} \tag{2}$$

The MSHA module (multi-scale attention) fuses multi-scale features by dividing the feature map into two levels and applying two different scale-specific feature processing functions f_1, f_2:

$$X_1^{(1)} = f_1(X_1'), \quad X_1^{(2)} = f_2(X_1') \tag{3}$$

After concatenating the feature maps of the two scales, a sigmoid function is applied to obtain the attention map:

$$\alpha = \sigma(\mathrm{Concat}[X_1^{(1)}, X_1^{(2)}]) \tag{4}$$

Then it is applied to the original feature map:

$$X_{\text{MSHA}} = \alpha \odot X_1' \tag{5}$$

The feed-forward neural network (FFN) consists of two 1×1 convolutions:

$$F_1 = \text{ReLU}(\text{Conv}_{1\times1}(X_{\text{MSHA}})) \tag{6}$$

$$F_2 = \text{Conv}_{1\times1}(F_1) \tag{7}$$

Adding a residual connection, we obtain

$$X_{\text{MSHA+FFN}} = X_{\text{MSHA}} + F_2 \tag{8}$$

3. The input to the original information branch is $X_2 \in \mathbb{R}^{\frac{C}{3}\times H\times W}$. It is directly fed into the Concat operation to ensure that while the model focuses on attention regions, it also retains basic features such as original textures and edges, maintaining semantic balance between the branches.

4. The input to the global context modeling branch is $X_3 \in \mathbb{R}^{\frac{C}{3}\times H\times W}$. First, a 1×1 convolution is applied to obtain the attention logits, which are then passed through a Softmax function to produce the spatial attention map A:

$$A = \text{Softmax}(\text{Conv}_{1\times1}(X_3)), \quad A \in \mathbb{R}^{\frac{C}{3}\times H\times W} \tag{9}$$

Softmax normalization is performed over the spatial dimensions $H \times W$ to obtain global receptive field context weights. The attention map is used to perform global context weighting on the input features:

$$X_{\text{ctx}} = A \odot X_3 \tag{10}$$

Channel modeling is performed through two layers of 1×1 convolutions followed by LayerNorm and ReLU nonlinear mappings:

$$F_c = \text{Conv}_{1\times1}(\text{ReLU}(\text{LayerNorm}(\text{Conv}_{1\times1}(X_{\text{ctx}})))) \tag{11}$$

5. For concatenation and fusion, the three feature branches are

$$(X_{\text{MSHA+FFN}}, \quad X_2, \quad F_c) \in \mathbb{R}^{\frac{C}{3}\times H\times W} \tag{12}$$

After concatenation, the features are fused and output through a 1×1 convolution:

$$X_{\text{concat}} = \text{Concat}(X_{\text{MSHA+FFN}}, X_2, F_c) \in \mathbb{R}^{C\times H\times W} \tag{13}$$

$$Y = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(X_{\text{concat}}))) \tag{14}$$

3.4. RepNCSPELAN-DBB Feature Extraction and Fusion Module

During the experimental phase, the feature extraction performance for power towers in remote sensing imagery was largely unsatisfactory, exhibiting insufficient feature learning effectiveness. To enhance the applicability of the designed object detection model for power transmission line equipment, this study improves the RepNCSPELAN feature extraction–fusion module in the YOLOv9 network architecture. A Diverse Branch Block (DBB) is introduced to reconstruct the module structure, forming a novel feature extraction–fusion module named RepNCSPELAN-DBB. This enhancement aims to improve the feature extraction network's capability in capturing distinctive features of power tower targets.

A Diverse Branch Block (DBB) is a universal convolutional neural network (ConvNet) building block that enhances performance without increasing inference-time costs. The DBB strengthens the representational capacity of a single convolution by incorporating diverse branches with varying scales and complexities, enriching the feature space through components such as convolutional sequences, multi-scale convolutions, and average pooling. After training, the DBB can be equivalently converted into a single convolutional layer for deployment. Unlike advances in novel ConvNet architectures, the DBB complicates the microstructure during training while preserving the macroarchitecture, enabling it to serve as a plug-and-play replacement for standard convolutional layers in any architecture. This approach allows models to achieve higher performance during training and then revert to the original inference-time structure for deployment.

The core principles of the DBB involve increasing the complexity of convolutional layers during the training phase by introducing branches of different sizes and structures to enrich the network's feature representation capabilities. These principles can be summarized as follows:

1. Diverse Branch Structures: The DBB enhances the complexity of convolutional layers during training by introducing branches with varying kernel sizes and architectures (e.g., different-sized convolution kernels and average pooling). This diversifies the feature representation capabilities of a single convolutional layer.

2. Training–Inference Decoupling: During the training phase, DBB employs complex multi-branch structures to enrich feature learning. During the inference phase, these branches are equivalently converted into a single convolutional layer, ensuring efficient deployment without the computational overhead.

3. Macroarchitecture Preservation: The DBB acts as a drop-in replacement for standard convolutional layers, enabling seamless integration into existing networks without altering their macroarchitecture.

Figure 6 illustrates the structure of the Diverse Branch Block (DBB) during training (left); the DBB consists of convolutional layers of varying sizes and average pooling layers arranged in parallel in a complex configuration, which are merged to produce the final output; after training, these intricate structures are equivalently converted into a single convolutional layer for the inference phase (right), preserving deployment efficiency. This transformation enables the DBB to enhance microstructural complexity during training while maintaining the macroarchitecture.

The concept of training–inference decoupling refers to using the complex DBB structure during the model's training phase and converting it into a simplified convolutional structure during inference. This design allows the model to leverage the diversity of the DBB to enhance feature extraction and learning capabilities during training while maintaining efficiency during inference by reducing the computational load. As a result, the model achieves high performance while ensuring operational speed and resource efficiency.

Figure 7 illustrates how different convolutional combinations (e.g., the 1 × 1 and K × K convolutions shown in the figure) are employed during the training phase and how these combinations are equivalently converted into a simplified structure (e.g., the concatenation operation represented by Transformation I in the figure) during the inference phase.

(A) Group-wise Convolution: It divides the input into multiple groups, each processed with distinct convolution kernels.

(B) 1×1-K×K Structure During Training: It first applies a 1 × 1 convolution (to reduce feature dimensionality), followed by grouped K × K convolutions.

(C) From the Perspective of Transformation I: This illustrates the merging of outputs from multiple grouped convolutions. Here, feature maps after group-wise convolutions are first processed through 1 × 1 convolutions, then concatenated.

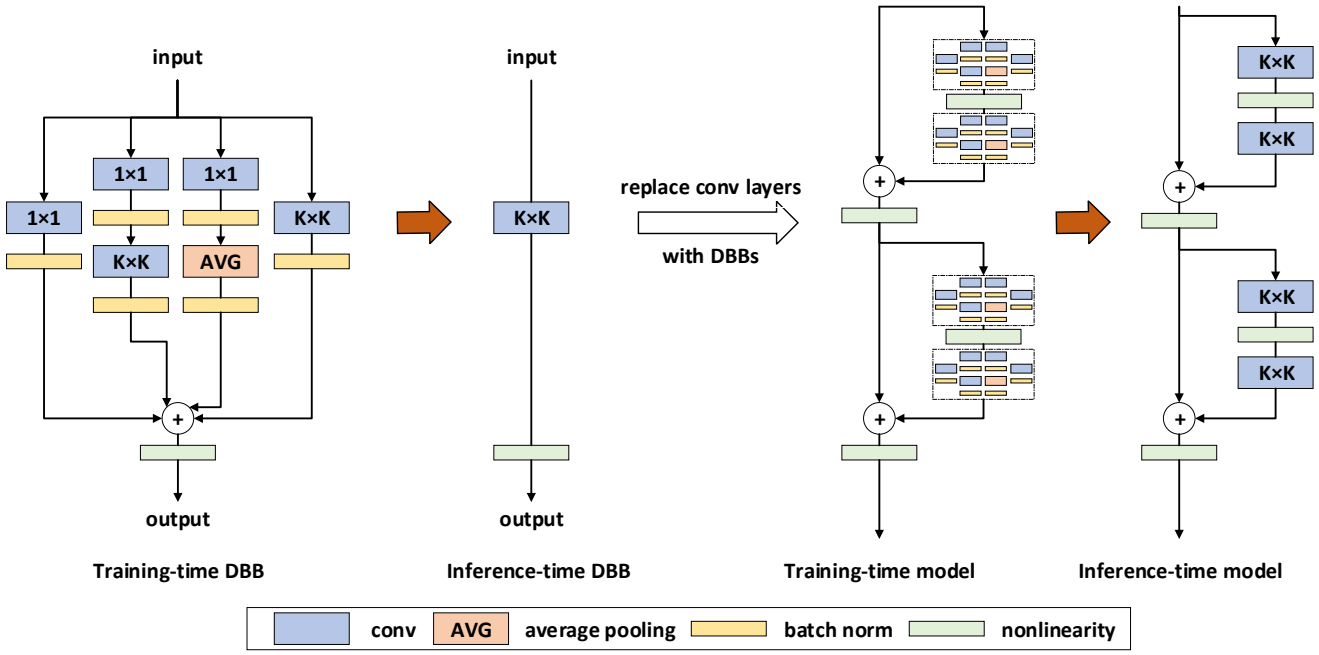


Figure 6. Diverse Branch Block diagram.

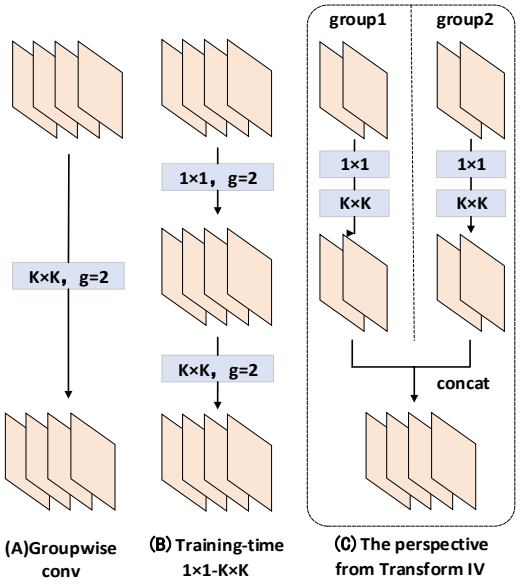


Figure 7. Example operations of the DBB in training and inference.

3.5. VMPDIoU Loss

In object detection, the bounding box regression loss function is critical as it quantifies the discrepancy between ground-truth and predicted boxes. Traditional bounding box regression (BBR) loss functions struggle to optimize scenarios where predicted and ground-truth boxes share the same aspect ratio but differ in specific dimensions. To address this limitation, this study introduces MPDIoU_Loss and proposes an enhancement tailored for remote sensing imagery by defining width (w) and height (h) as multiples of the ground-truth box's dimensions, naming this improved loss VMPDIoU (Variable Minimum Point Distance Intersection over Union).

MPDIoU (Minimum Point Distance Intersection over Union [42]) is a novel bounding box similarity metric based on the minimum point distance of axis-aligned rectangles, which integrates overlap area, center point distance, and width/height deviations into a unified measure while simplifying the computational process.

The Figure 8 illustrates two distinct bounding box regression results. The green boxes represent ground-truth bounding boxes, while the red boxes represent predicted bounding boxes. In the left image, the predicted box has a length of 4 units and the ground-truth box has a length of 2 units, resulting in a ratio of 2:1. In the right image, the predicted box has a length of 1 unit and the ground-truth box has a length of 2 units, resulting in a ratio of 1:2. In both cases, traditional loss functions (e.g., Generalized Intersection over Union [43], Distance Intersection over Union [44], Complete Intersection over Union [44], and Efficient Intersection over Union [45]) yield identical loss values, whereas the MPDIoU method produces distinct loss values. This demonstrates that traditional approaches may fail to differentiate between certain prediction scenarios, while MPDIoU more accurately reflects discrepancies between predicted and ground-truth boxes. This highlights MPDIoU's superiority in bounding box regression, particularly in distinguishing boxes with identical aspect ratios but differing sizes or positions. By directly calculating the critical point distances between predicted and ground-truth boxes, MPDIoU provies a more precise loss metric for such cases.

Figure 8. Two different bounding box regression results. The green boxes represent ground-truth bounding boxes with a length of 2 units. The red boxes represent predicted bounding boxes, with a length of 4 units in the left image and 1 unit in the right image.

Inspired by the geometric properties of bounding boxes, MPDIoU compels each predicted bounding box to converge toward its ground-truth counterpart during training by minimizing the loss function. It employs the coordinates of the four corner points to represent all components of existing bounding box regression loss functions.

The schematic diagram is shown in Figure 9. The blue borders indicate the annotation boxes, while the red borders represent the prediction boxes. Assuming the input image has a width of a and a height of b, the top-left and bottom-right coordinates of the ground-truth box and the predicted box are represented as (x_1^A, y_1^A), (x_2^A, y_2^A), (x_1^B, y_1^B), and (x_2^B, y_2^B). Then, the distances between the top-left corners and the bottom-right corners of the two bounding boxes can be expressed as follows:

$$d_1^2 = \left(\mathrm{x}_1^B - \mathrm{x}_1^\mathrm{A}\right)^2 + \left(y_1^B - y_1^A\right)^2 \tag{15}$$

$$d_2^2 = \left(\mathrm{x}_2^\mathrm{B} - \mathrm{x}_2^\mathrm{A}\right)^2 + \left(y_2^B - y_2^A\right)^2 \tag{16}$$

where d_1 represents the distance of the top-left corner and d_2 represents the distance of the bottom-right corner. MPDIoU and MPDIoU_Loss can be expressed as

$$\text{MPDIoU} = \frac{d_1^2}{a^2+b^2} - \frac{d_2^2}{a^2+b^2} \tag{17}$$

$$L_{\text{MPDIoU}} = 1 - MPDIoU \tag{18}$$

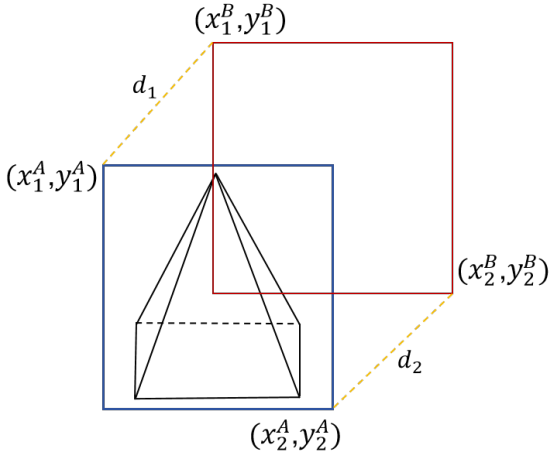


Figure 9. Factors of MPDIoU.

Therefore, all components of existing bounding box regression loss functions can be derived from the coordinates of the four corner points. The conversion formulas are as follows:

$$\begin{aligned} |\mathrm{C}| &= L1 \times L2 \\ L1 &= \left(\max\left(\mathrm{x}_2^A, \mathrm{x}_2^B\right) - \min\left(\mathrm{x}_1^A, \mathrm{x}_1^B\right)\right) \\ L2 &= \left(\max\left(\mathrm{y}_2^A, \mathrm{y}_2^B\right) - \min\left(y_1^A, y_1^B\right)\right) \end{aligned} \tag{19}$$

$$\begin{aligned} \mathrm{x}_c^A &= \frac{\mathrm{x}_1^A + \mathrm{x}_2^A}{2}, y_c^A = \frac{y_1^A + y_2^A}{2}, \\ y_c^B &= \frac{y_1^B + y_2^B}{2}, \mathrm{x}_c^B = \frac{\mathrm{x}_1^B + \mathrm{x}_2^B}{2} \end{aligned} \tag{20}$$

$$\begin{aligned} w^A &= \mathrm{x}_2^A - \mathrm{x}_1^A, h^A = y_2^A - y_1^A, \\ w^B &= \mathrm{x}_2^B - \mathrm{x}_1^B, h^B = y_2^B - y_1^B \end{aligned} \tag{21}$$

$|C|$ represents the area of the minimum enclosing rectangle that covers both A and B. (x_c^A, y_c^A) and (x_c^B, y_c^B) represent the coordinates of the center points of the ground-truth box and the predicted box, respectively. w^A and h^A represent the width and height of the ground-truth box. w^B and h^B represent the width and height of the predicted box.

In Equation (17), d_1 represents the distance between the top-left corners of the two bounding boxes, while d_2 represents the distance between their bottom-right corners. a denotes the width of the input image, and b denotes its height. In remote sensing images, power tower targets are relatively small, and both the annotated ground-truth boxes and the predicted boxes occupy only a small portion of the image. Therefore, d_1 and d_2 are very small compared to a and b. As a result, when the position of the predicted box changes, the computed MPDIoU value changes only slightly, making it difficult to clearly reflect the deviation between the predicted box and the ground-truth box.

Moreover, most remote sensing datasets are processed to highlight the target objects, often through image cropping. This creates an issue when using the original input image

width a and height b in the MPDIoU calculation. Suppose the predicted box remains unchanged: after cropping the image, the distances d_1 and d_2 between the two boxes remain the same, but the input image's width and height change, which leads to inconsistencies in the computed MPDIoU.

To ensure consistency in MPDIoU_Loss calculations before and after data processing, we have modified the definitions of a and b in the formula. Specifically, a and b are redefined as multiples of the ground-truth box's width w^A and height h^A, respectively. This adjustment ensures that even if the image is cropped, the computed MPDIoU_Loss remains stable and unaffected by the change.

4. Experiment

4.1. Experimental Setup

All experiments in this study were conducted on a computer equipped with an AMD EPYC 7453 CPU, an NVIDIA GeForce RTX 4090 GPU, and a Linux operating system. Python (version 3.9) was used as the programming language, with PyTorch (version 1.11.0) as the deep learning framework and CUDA 12.2 for GPU acceleration. For hyperparameter settings, a batch size of 16 was selected, while training proceeded for 200 epochs, and the learning rate was dynamically adjusted using cosine annealing. The input resolution for the network was set to 512 × 512 pixels. Other settings remained at their default configurations. To evaluate model performance, the metrics mAP@0.5 and mAP@0.5:0.95 were employed to assess detection accuracy.

4.2. Dataset

This paper employs two satellite remote sensing power tower detection datasets to validate the effectiveness of the proposed method. The first is the Satellite Remote Sensing Power Tower Dataset (SRSPTD) curated by Professor Ke Zhang's team at North China Electric Power University, while the second is the multi-scenario high-resolution satellite remote sensing transmission tower dataset (GFTD) developed by Dean Xiaojin Yan's research group at North China Institute Of Aerospace Engineering.

The SRSPTD dataset was constructed based on power transmission and distribution infrastructure images collected by Duke University's Bass Connections Project in the United States. The data sources span six regions across two countries—Arizona (AZ), Connecticut (CT), Kansas (KS), and North Carolina (NC) in the United States, along with Taranga and Dunedin in Austria—encompassing four distinct geographical environments (desert, plains, forests, and coastal areas) and three human settlement density zones (suburban, rural, and urban). Satellite remote sensing images of power towers from these six regions were selected as data subsets. After re-cropping, annotation, and processing, the final dataset comprises 2740 annotated images of satellite remote sensing power towers, including 2760 distribution towers and 284 transmission towers. As illustrated in Figure 10, the dataset includes examples of power transmission and distribution infrastructure across varying human activity density areas and diverse terrain conditions. The dataset is partitioned into an 80% training set, with 10% each allocated to the testing and validation sets.

The GFTD is derived from satellite remote sensing imagery captured by the GaoFen-2 and GaoFen-7 satellites, as illustrated in Figure 11. It primarily focuses on three high-voltage transmission lines: from Zhangjiakou (Hebei) to Beijing West, from Beijing West to Baoding (Hebei), and from Zhalute (Inner Mongolia) to Qingzhou (Shandong). The imagery spans all four seasons of 2023 and 2024, covering six distinct scenarios: built-up environments, coastal areas, deserts, plains, woodlands, and mountainous regions. The dataset was constructed by fusing 4 m resolution multispectral imagery with 1 m resolution panchromatic imagery to achieve 1 m resolution composite images, which were

then cropped to 512 × 512 pixels. The Labelme annotation software was employed for labeling, with transmission towers and their geometrically distinctive shadows annotated as integrated entities to provide richer feature information for the model and lay the groundwork for subsequent tower-type classification. The dataset comprises 3000 images including 2870 towers, which were divided into the training, validation, and test sets at a 7:2:1 ratio. To enhance model robustness across diverse environmental conditions, data augmentation techniques including random cropping, translation, rotation, Mosaic, and Mixup were applied during preprocessing.

Figure 10. SRSPTD sample images.

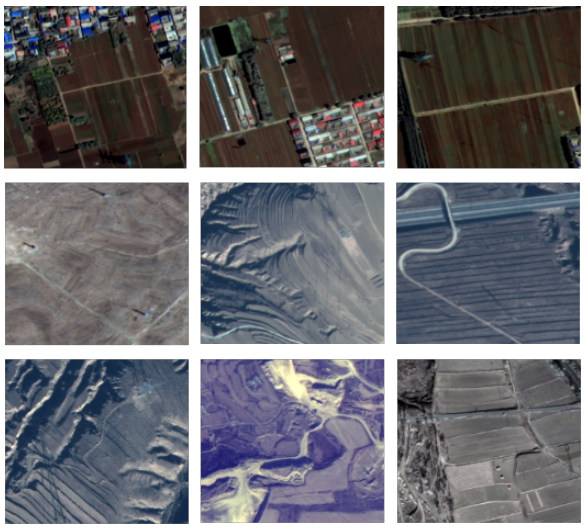


Figure 11. GFTD Dataset sample images.

4.3. Controlled Experiment

In our work, we selected representative baselines based on the following considerations: (1) popularity and wide usage in the remote sensing or small object detection community; (2) performance on our standard benchmarks; and (3) architectural diversity, including both one-stage and two-stage detectors.

To validate the effectiveness of our algorithm, comparative experiments were conducted with mainstream object detection models including SSD, Faster R-CNN, RetinaNet, Deformable-DETR, and various YOLO-series algorithms on both the Satellite Remote Sens-

ing Power Tower Dataset (SRSPTD) and the multi-scenario high-resolution satellite remote sensing transmission tower dataset (GFTD). Additionally, comparisons were incorporated with TPH-YOLO—a specialized model for small object detection—and LSKF-YOLO, the current state-of-the-art (SOTA) model in power tower detection for remote sensing imagery.

The comparative results on the SRSPTD are presented in Table 1. As shown in the table, among various baseline algorithms, the YOLOv9 model achieves a mean Average Precision (mAP@0.5) of 75.5% while maintaining relatively few parameters, making it more suitable as the foundational model for this study. The improved YOLOv9-GDV model demonstrates significant enhancements over baseline algorithms, with mAP@0.5 improvements of 7.6%, 6.3%, 4.7%, 5.4%, 5.5%, 8.0%, 9.7%, and 12.9% compared to YOLOv11, YOLOv10, YOLOv9, Deformable-DETR, YOLOv8, RetinaNet, SSD512, and Faster R-CNN, respectively. Furthermore, it outperforms the domain-specific enhanced models LSKF-YOLO and TPH-YOLO—specialized for power line detection in remote sensing imagery—by 2.8% and 5.3%, respectively. In summary, the YOLOv9-GDV model achieves a state-of-the-art mAP@0.5 of 80.2%.

Table 1. Experimental results of different object detection models on the SRSPTD.

MODEL	AP (DT)	AP (TT)	mAP@0.5 (%)	Params (MB)	Flops (G)
Faster-RCNN (VGG16) [17]	53.6	80.9	67.3	28.3	473.2
SSD512 (VGG16) [18]	56.4	84.6	70.5	24.6	30.8
RetinaNet [19]	57.3	83.4	72.2	80.2	77.6
YOLOv4 [23]	66.2	82.4	74.3	14.8	20.5
YOLOv5-S [46]	65.3	82.2	73.7	13.7	16.2
YOLOv7-Tiny [26]	67.1	82.1	74.6	12.3	15.5
YOLOX-S [24]	66.3	84.5	75.4	21.1	24.8
TPH-YOLO [34]	66.5	83.2	74.9	13.67	16.0
YOLOv8-S [47]	65.2	84.2	74.7	11.13	28.6
Deformable-DETR [48]	-	-	74.8	40.2	173.3
Gold-YOLO-S [49]	67.8	83.6	75.7	21.51	46.03
LSKF-YOLO [50]	68.0	86.9	77.4	23.2	18.5
YOLOv9-S [27]	65.2	84.5	75.5	9.6	38.7
YOLOv10-S [28]	66.2	83.2	73.9	7.2	21.6
YOLOv11-S [51]	65.8	82.8	72.6	9.4	21.5
YOLOv9-GDV	73.1	87.3	80.2	12.6	40.6

In practical testing scenarios, the proposed method demonstrates refined recognition capabilities for power towers in remote sensing images, achieving robust detection performance. Specific visual detection results are illustrated in Figure 12. The figure reveals that even under challenging conditions—where background colors closely resemble target hues and numerous interfering elements are present—the system maintains reliable identification accuracy with minimal occurrences of false positives or missed detections. Furthermore, the prediction boxes exhibit precise alignment with the actual targets.

The comparative results on the GFTD are presented in Table 2. As indicated in the table, the enhanced YOLOv9-GDV model demonstrates superior accuracy and detection performance among baseline algorithms. Compared to baseline models including YOLOv11, YOLOv10, YOLOv9, Deformable-DETR, YOLOv8, RetinaNet, SSD512, and Faster R-CNN, the mAP@0.5 improvements are 1.9%, 3.0%, 2.3%, 4.2%, 5.7%, 14.4%, 16.1%, and 21.4%, respectively. Additionally, it outperforms LSKF-YOLO—a specialized model for power line detection in remote sensing imagery—by 1.7%. While the dataset's limited data diversity and the inclusion of target shadows during annotation contribute to generally elevated detection accuracy across all models, YOLOv9-GDV still achieves significantly higher

precision compared to its counterparts. In summary, the YOLOv9-GDV model attains a state-of-the-art mean Average Precision (mAP@0.5) of 94.6%.

Figure 12. Visualized detection results on the SRSPTD.

Table 2. Experimental results of different object detection models on the GFTD.

MODEL	mAP@0.5 (%)	Params (MB)	Flops (G)
Faster-RCNN (VGG16)	73.2	28.3	473.2
SSD512 (VGG16)	78.5	24.6	30.8
RetinaNet	80.2	80.2	77.6
YOLOv5-S	86.8	13.7	16.2
YOLOv7-Tiny	85.5	12.3	15.5
YOLOX-S	87.6	21.1	24.8
YOLOv8-S	88.9	11.13	28.6
Deformable-DETR	90.4	40.2	173.3
LSKF-YOLO	92.9	23.2	18.5
YOLOv9-S	92.3	9.6	38.7
YOLOv10-S	91.6	7.2	21.6
YOLOv11-S	92.7	9.4	21.5
YOLOv9-GDV	94.6	12.6	40.6

In practical testing, the proposed model achieves detailed recognition of power towers in remote sensing images with satisfactory detection performance. Specific visualized detection results are shown in Figure 13. We selected detection outcomes under multiple complex backgrounds, including variations in terrain, dominant colors, and other conditions. As shown in the figure, our model delivers highly accurate identification of targets, with prediction bounding boxes tightly fitting the detected objects and their shadows, aligning well with the annotation boxes.

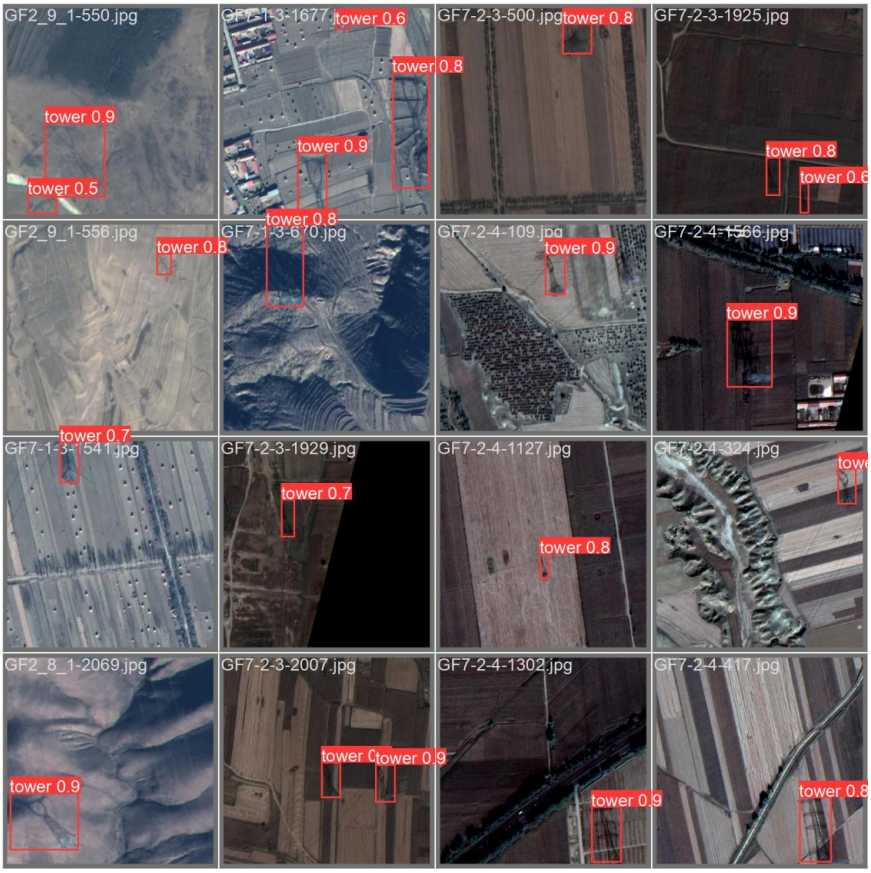


Figure 13. Visualized detection results on the GFTD.

4.4. Ablation Study

To validate the effectiveness of individual modules, this study conducted ablation experiments for comparative analysis. These experiments were designed to investigate the impact of different improvement strategies on model detection performance, providing critical insights for model design and optimization. Throughout the experiments, identical parameter configurations were maintained. The ablation study results are summarized in Table 3, where "Y" indicates the inclusion of a specific enhancement and "N" denotes its exclusion.

The GPA attention mechanism is applied in the network to introduce global representation capability, and the Diversity Branch Block (DBB) is embedded into the core module RepNCSPELAN. Additionally, VMPDIoU is utilized as the model's loss function.

Experimental analysis of Table 3 reveals the following insights: The incorporation of the Global Pyramid Attention (GPA) mechanism into YOLOv9 (Improved Model 1) elevates the mAP@0.5 by 1.2% compared to the baseline, demonstrating that GPA enhances global contextual representation by modeling object–background relationships and integrating multi-scale spatial information. Improved Model 2, which embeds the Diverse Branch Block (DBB) into the RepNCSPELAN core module, achieves a 2.6% mAP@0.5 improvement, attributed to the DBB's ability to strengthen convolutional representational diversity and enrich feature hierarchies.The experimental results demonstrate that Improved Model 3, which employs the enhanced VMPDIoU (Minimum Point Distance IoU) as the loss function, achieves a 2.3% increase in mAP@0.5. This improvement validates that VMPDIoU more accurately quantifies discrepancies between predicted and ground-truth bounding boxes by directly minimizing keypoint distances, thereby refining localization precision. Furthermore, the integrated YOLOv9-GDV model proposed in this study exhibits the most significant performance gain, with a 4.7% enhancement in mAP@0.5 compared to the

original YOLOv9 model, underscoring the synergistic efficacy of the proposed architectural optimizations.

Table 3. Ablation study.

Model	GPA	DBB	VMPDIOU	mAP@0.5 (%)
YOLOv9s	N	N	N	75.5
improved model1	Y	N	N	+1.2
improved model2	N	Y	N	+2.6
improved model3	N	N	Y	+2.3
YOLOv9-GDV	Y	Y	Y	80.2

Based on the comprehensive analysis above, it has been proven that the combination of the proposed improvements effectively enhances the model's detection accuracy. Compared to the original YOLOv9 model, the refined YOLOv9-GDV model demonstrates a significant improvement in detection precision.

4.5. Robustness Experiment

Remote sensing data are often susceptible to factors such as illumination intensity and the degree of haze during the imaging process, which may result in indistinct features of target objects. This issue becomes more pronounced when the targets are small, making feature extraction even more difficult. To evaluate the robustness of YOLOv9-GDV under low-light and hazy conditions, we simulated the degradation of remote sensing images to generate a series of test datasets, each using the same original images but under different degradation settings.

The hazy image test sets were generated using the atmospheric scattering model by setting different atmospheric light parameters A. The mathematical formulation of the atmospheric scattering model is as follows:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{22}$$

where $I(x)$ is the observed hazy image, $J(x)$ is the scene without haze, and A is the atmospheric light parameter, representing the color of the haze or smoke. As A increases, the hazy effect becomes more pronounced, resulting in a more blurred and distorted visual appearance. The term $t(x)$, known as the transmission, indicates the extent to which light travels through the atmosphere. As the distance increases, more light is scattered and absorbed. The transmission can be expressed as

$$t(x) = e^{-\beta d(x)} \tag{23}$$

Here, $d(x)$ represents the depth information of the scene, and β is the scattering coefficient, which controls the density of the haze.

The low-light image test sets were generated using OpenCV by adjusting different brightness values b based on the following formula:

$$g(i,j) = af(i,j) + b \tag{24}$$

where $g(i,j)$ is the transformed pixel value, $f(i,j)$ is the original pixel value, a is the contrast gain factor, and b is the brightness offset.

We selected YOLOv9-GDV and YOLOv9s for robustness testing. The experimental results indicate that both algorithms exhibit a certain degree of robustness to hazy and

low-light images. As shown in Table 4, YOLOv9-GDV model performs slightly better than YOLOv9s.

Table 4. Robustness experiment.

Fog (A)	Light Intensity (b)	mAP@0.5 (%) (YOLOv9-GDV)	mAP@0.5 (%) (YOLOv9s)
-	-	80.2	75.5
0.05	-	79.6	74.1
0.10	-	78.9	72.8
0.15	-	77.0	71.3
-	−5	80.1	74.9
-	−10	79.6	74.2
-	−15	79.0	73.8

Although both algorithms demonstrate a certain degree of robustness to hazy and low-light images, accuracy degradation and an increase in false positives and false negatives inevitably occur during training. To mitigate these issues, we recommend applying image preprocessing to some extent before detecting small targets in remote sensing images to alleviate the effects of haze and low illumination.

5. Conclusions

With the advancement of satellite remote sensing technology, the application of high-resolution satellite remote sensing data in power line inspection has gradually become a critical research focus. This paper proposes a power pylon detection algorithm for high-resolution satellite remote sensing images, named YOLOv9-GDV. First, we improve the network architecture and attention mechanism of the baseline model YOLOv9 by integrating the global position-aware (GPA) attention mechanism. This enhancement introduces global representation capabilities to the model, enabling it to better understand the relationships between objects and backgrounds, integrate multi-scale and positional information, and thereby improve detection accuracy and robustness. Second, we embed the Diverse Branch Block (DBB) into the RepNCSPELAN4 module, constructing a novel feature extraction and fusion module called RepNCSPELAN4-DBB. This modification enhances the representational capacity of individual convolutions, enriches the feature space, and improves the model's feature extraction performance without significantly increasing the computational overhead. Finally, we employ VMPDIoU as the improved loss function, which directly calculates the keypoint distances between predicted and ground-truth bounding boxes to more accurately reflect their discrepancies, further boosting the model's mean Average Precision.

Experiments on the Satellite Remote Sensing Power Pylon Dataset (SRSPTD) and the general-scene high-resolution satellite remote sensing transmission tower dataset (GFTD) demonstrate the effectiveness of YOLOv9-GDV. The model achieves significant improvements in mean Average Precision (mAP), validating the feasibility of the proposed enhancements.

However, the algorithm has limitations in dataset diversity, such as the lack of remote sensing images under extreme weather conditions (e.g., fog, rain, or cloud occlusion), leading to suboptimal detection performance in such scenarios. Future work should focus on expanding the dataset to enhance the model's generalization ability. Through this research, we believe that applications of remote sensing image data will continue to advance, and deep learning-based power pylon detection technologies will play an increasingly vital role in the intelligent operation and maintenance of power transmission lines, providing robust support for ensuring the safety and stability of power grids.

Author Contributions: Conceptualization, K.Z. and N.Z.; methodology, K.Z. and N.Z.; software, N.Z.; validation, Q.L., Y.C., X.Z. (Xiaoyun Zhang) and J.Y.; resources, X.Z. (Xian Zheng); data curation, X.Z. (Xian Zheng); writing—original draft preparation, N.Z.; writing—review and editing, C.S.; visualization, N.Z.; supervision, K.Z. and C.S.; project administration, K.Z. and C.S.; funding acquisition, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62206095; Hebei Provincial Natural Science Foundation Beijing Tianjin Hebei Basic Research Cooperation Special Project under Grant F2024502017; and in part by the Fundamental Research Funds for the Central Universities under Grant 2023JG002 and Grant 2024MS117.

Data Availability Statement: This study utilizes two datasets: SRSPTD and GFTD. The Satellite Remote Sensing Power Tower Dataset (SRSPTD), curated and developed by Ke Zhang's team at North China Electric Power University, is publicly accessible at https://github.com/ZX815/LSKF-YOLO/tree/main accessed on 18 May 2025. In contrast, the multi-scene high-resolution satellite remote sensing transmission tower dataset (GFTD), compiled by Dean Xiaojin Yan's team at North China Institute of Aerospace Engineering, is a proprietary dataset and thus not publicly available due to data sharing restrictions.

Acknowledgments: The authors acknowledge all the technical support of those who helped in conducting the study. The authors would like to acknowledge the academic editor and anonymous reviewers in advance for accepting to review earlier versions of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Yang, L.; Fan, J.; Liu, Y.; Li, E.; Peng, J.; Liang, Z. A review on state-of-the-art power line inspection techniques. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9350–9365. [CrossRef]
2. Li, J.; Wang, L.; Shen, X. Unmanned aerial vehicle intelligent patrol-inspection system applied to transmission grid. In Proceedings of the 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, China, 20–22 October 2018; pp. 1–5.
3. Meng, L.; Yan, C.; Lv, S.; Sun, H.; Xue, S.; Li, Q.; Zhou, L.; Edwing, D.; Edwing, K.; Geng, X.; et al. Synthetic aperture radar for geosciences. *Rev. Geophys.* **2024**, *62*, e2023RG000821. [CrossRef]
4. Li, W.; Ma, P.; Wang, H.; Fang, C. SAR-TSCC: A novel approach for long time series SAR image change detection and pattern analysis. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5203016. [CrossRef]
5. Li, K.; Yan, X.; Han, Y. Multi-mechanism swarm optimization for multi-UAV task assignment and path planning in transmission line inspection under multi-wind field. *Appl. Soft Comput.* **2024**, *150*, 111033. [CrossRef]
6. Luo, Y.; Yu, X.; Yang, D.; Zhou, B. A survey of intelligent transmission line inspection based on unmanned aerial vehicle. *Artif. Intell. Rev.* **2023**, *56*, 173–201. [CrossRef]
7. Mikhalap, S.; Trashenkov, S.; Vasilyeva, V. Study of overhead power line corridors on the territory of pskov region (russia) based on satellite sounding data. In Proceedings of the International Scientific and Practical Conference, Ekaterinburg, Russia, 21–22 March 2019; Volume 1, pp. 164–167.
8. Qayyum, A.; Malik, A.S.; Saad, M.N.M.; Iqbal, M.; Abdullah, M.; Abdullah, T.A.R.B.T.; Ramli, A.Q. Power LinesVegetation enchroachment monitoring based on Satellite Stereo images using stereo matching. In Proceedings of the 2014 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Kuala Lumpur, Malaysia, 25–25 November 2014; pp. 1–5.
9. Uehara, K.; Sakanashi, H.; Nosato, H.; Murakawa, M.; Miyamoto, H.; Nakamura, R. Object detection of satellite images using multi-channel higher-order local autocorrelation. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 1339–1344.
10. Liu, Y.; Guo, Y.; Ma, S.; Yang, Z.; Chen, C.; Wu, G.; Liu, W.; Ren, W.; Sang, X.; Li, T. Monitoring ultra high-voltage transmission tower deformation and corridor subsidence using China C-band Synthetic Aperture Radar and Sentinel-1 data. *Geocarto Int.* **2025**, *40*, 2462482. [CrossRef]
11. Sarabandi, K.; Park, M. Extraction of power line maps from millimeter-wave polarimetric SAR images. *IEEE Trans. Antennas Propag.* **2000**, *48*, 1802–1809. [CrossRef]

12. Cheng, H.; Gu, Y.; Xi, M.; Zhong, Q.; Wei, L. A Few-Shot Collapsed Transmission Tower Detection Method Combining Large and Small Models in Remote Sensing Image. *IEEE Access* **2025**, *13*, 41670–41681.
13. Prakash, T.; Kak, A.C. Active learning for designing detectors for infrequently occurring objects in wide-area satellite imagery. *Comput. Vis. Image Underst.* **2018**, *170*, 92–108. [CrossRef]
14. Rohrer, B.; Lerner, A.; Gershenson, D. A New Predictive Model for More Accurate Electrical Grid Mapping. 2019. Available online: https://engineering.fb.com/2019/01/25/connectivity/electrical-grid-mapping/ (accessed on 18 May 2025).
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Computer Vision ECCV2016 14th European Conference Proceedings, Amsterdam, The Netherlands, 11–14 October 2016; Part I, pp. 21–37.
19. Ross, T.Y.; Dollár, G. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 2980–2988.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
22. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
23. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
24. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
25. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
26. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
27. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 1–21.
28. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
29. Hu, W.; Alexander, B.; Cathcart, W.; Hu, A.; Nair, V.; Zuo, L.; Malof, J.; Collins, L.; Bradbury, K. Mapping electric transmission line infrastructure from aerial imagery with deep learning. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 2229–2232.
30. Bradbury, K.; Han, Q.; Nair, V.; Pathirathna, T.; You, X. Electric transmission and distribution infrastructure imagery dataset. *Accessed Aug.* **2018**, *14*, 2020.
31. Haroun, F.M.E.; Deros, S.N.M.; Din, N.M. Detection and monitoring of power line corridor from satellite imagery using RetinaNet and K-Mean clustering. *IEEE Access* **2021**, *9*, 116720–116730. [CrossRef]
32. Hou, L.; Lu, K.; Xue, J.; Hao, L. Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
33. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
34. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
35. Zhou, H.; Luo, F.; Zhuang, H.; Weng, Z.; Gong, X.; Lin, Z. Attention multihop graph and multiscale convolutional fusion network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5508614. [CrossRef]
36. Hong, D.; Zhang, B.; Li, H.; Li, Y.; Yao, J.; Li, C.; Werner, M.; Chanussot, J.; Zipf, A.; Zhu, X.X. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sens. Environ.* **2023**, *299*, 113856. [CrossRef]

37. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [CrossRef] [PubMed]
38. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for small object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [CrossRef]
39. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
40. Wang, J.; Li, W.; Zhang, M.; Tao, R.; Chanussot, J. Remote-sensing scene classification via multistage self-guided separation network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615312. [CrossRef]
41. Yang, C.; Zhao, B.; Zhou, Q.; Wang, Q. MMO-IG: Multi-Class and Multi-Scale Object Image Generation for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5616412. [CrossRef]
42. Ma, S.; Xu, Y. Mpdiou: A loss for efficient and accurate bounding box regression. *arXiv* **2023**, arXiv:2307.07662.
43. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
44. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
45. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
46. Jocher, G.; Chaurasia, A.; Borovec, J.; Nana-Boateng; Wong, A.; Kwon, Y.; Chang, C.Y.; Laughing; Fang, J.; Hogan, A.; et al. YOLOv5 by Ultralytics. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 18 May 2025).
47. Jocher, G.; Chaurasia, A.; Qiu, J. YOLOv8 by Ultralytics. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 18 May 2025).
48. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
49. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 51094–51112.
50. Shi, C.; Zheng, X.; Zhao, Z.; Zhang, K.; Su, Z.; Lu, Q. LSKF-YOLO: Large selective kernel feature fusion network for power tower detection in high-resolution satellite remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5620116. [CrossRef]
51. Jocher, G.; Qiu, J. YOLOv11 by Ultralytics. 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 18 May 2025).