information

Article

QYOLO: Contextual Query-Assisted Object Detection in High-Resolution Images

Mingyang Gao [1,2], Wenrui Wang [1], Jia Mao [1], Jun Xiong [3], Zhenming Wang [1] and Bo Wu [1,*]

[1] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] State Grid Fujian Electric Power Co., Xiamen 361005, China
* Correspondence: wubo@sari.ac.cn

Abstract: High-resolution imagery captured by drones can detect critical components on high-voltage transmission towers, providing inspection personnel with essential maintenance insights and improving the efficiency of power line inspections. The high-resolution imagery is particularly effective in enhancing the detection of fine details such as screws. The QYOLO algorithm, an enhancement of YOLOv8, incorporates context queries into the feature pyramid, effectively capturing long-range dependencies and improving the network's ability to detect objects. To address the increased network depth and computational load introduced by query extraction, Ghost Separable Convolution (GSConv) is employed, reducing the computational expense by half and further improving the detection performance for small objects such as screws. The experimental validation using the Transmission Line Accessories Dataset (TLAD) developed for this project demonstrates that the proposed improvements increase the average precision (AP) for small objects by 5.5% and the F1-score by 3.5%. The method also enhances detection performance for overall targets, confirming its efficacy in practical applications.

Keywords: YOLOv8; power transmission line component inspection; multi-scale object detection; GSConv; query-based detector

Citation: Gao, M.; Wang, W.; Mao, J.; Xiong, J.; Wang, Z.; Wu, B. QYOLO: Contextual Query-Assisted Object Detection in High-Resolution Images. *Information* **2024**, *15*, 563. https://doi.org/10.3390/info15090563

Academic Editors: Alessandra Lumini, Vasco N. G. J. Soares, João M. L. P. Caldeira, Bruno Bogaz Zarpelão and Jaime Galán-Jiménez

Received: 5 July 2024
Revised: 22 August 2024
Accepted: 6 September 2024
Published: 12 September 2024

1. Introduction

The ongoing enhancement in camera resolutions mounted on Unmanned Aerial Vehicles (UAVs) [1] captures increasingly detailed information, which is advantageous for detecting densely packed small targets such as screws, thereby aligning with the neural network's requirement for the precise feature extraction of small-scale objects. However, this progression leads to larger model sizes, substantially increased model depth, and computational demands [2]. Consequently, there is a pressing need to improve the existing models' detection accuracy across multiple scales of objects while balancing inference speed and detection precision.

In the context of our custom-built power line inspection dataset, observations from the images captured by UAVs highlight several characteristics: 1. Variability in target morphology is pronounced due to variations in distance and angle between the camera and the high-voltage towers during acquisition. Larger recognition targets such as insulators and vibration dampers are susceptible to obstruction by power lines and the angular steel structures of the towers. 2. When using UAVs for photography with automatic exposure enabled, collection time and weather can have an impact on image quality. Vibrations or sudden changes in the light incident angle can cause the resulting images to be overly bright. Such overexposure increases the challenge of detecting densely packed objects because it amplifies the noise during feature learning in deep learning networks. 3. The dataset comprises high-resolution images in which smaller annotated objects occupy as little as 1600 pixels, representing merely 0.03% of the image area.

Recent years have witnessed relentless innovation in deep learning strategies. Pivotal research [3] strengthens systems' adaptability to environmental changes via an advanced fog injection algorithm for dataset enrichment. Nevertheless, ongoing refinements in datasets are crucial to overcome challenges such as motion blur and variable lighting in diverse operational settings. To enhance the detection of small-scale objects, CNN-based detectors have incorporated Transformer-driven feature fusion [2,4] and pioneering techniques like separable convolutions [5,6], augmenting sensitivity to fine details.

Query-based methods [7] have emerged as powerful techniques to enhance target localization. QueryDet [8] employs a query mechanism to accelerate inference in feature pyramid-based detectors while enabling rough target localization on low-resolution maps before high-resolution refinement for precise outcomes. However, this method requires deeper networks and increased computation, which compromises real-time responsiveness. Similarly, CANet (Context Aggregation Network) [9] utilizes a self-attentive mechanism that incorporates spatial context aggregation. It achieves this by treating features at each pixel location as a query, calculating similarities with other pixel locations, and performing weighted feature aggregation. In object detection, the concept of queries has been expanded by Transformer-based methods like DETR (DEtection TRansformer) [10] to guide the detection process. These queries act as predefined anchors for potential object locations or classes and learn during training to improve localization and classification accuracy.

Despite these advancements, there are still challenges in applying these methods to complex environments, such as high-voltage tower inspections, where substantial depth variations and intricate backgrounds pose significant obstacles. In such scenarios, it is crucial to suppress background noise using strategies such as channel and spatial attention mechanisms that dynamically emphasize important features. However, their effectiveness in environments with densely packed and frequently obstructed targets remains a challenge. The requirement to handle such complex conditions underscores the importance of incorporating cross-semantic context, which is crucial for ensuring the focused detection of a multitude of object types during power line inspections.

This research adopts a practical, scenario-focused power line dataset to enhance the precision in detecting objects of various sizes during power line inspections. The research applies data augmentation strategies involving exposure variations, rotations, and motion blur methodologies [11]. Introducing QYOLOv8, this study targets explicitly to improve the recognition of screws, vibration dampers, and insulators in the power grid infrastructure. The QYOLOv8 network builds upon the foundational YOLOv8 [12,13] architecture and integrates advanced CNN techniques [14] to enhance feature extraction and detection precision. Our work mainly introduces the following contributions:

1. This innovative algorithm integrates GSConv [15] and GSCSP [16], mechanisms that diminish computational demands via grouped computing, optimizing resource utilization.
2. In the feature fusion neck network, features from different layers are extracted and fused by using the rows and columns within each layer as queries. This approach enhances feature flow and interaction within the network.
3. Recognizing that many inspection targets are situated on towering structures, the algorithm optimizes attention mechanisms for efficient long-range feature modeling. Strengthened by enhanced inter-class cross-attention, it fosters stronger correlations among diverse object categories, heightening the overall recognition capability.

The resultant model, QYOLOv8, achieves a harmonious balance of elevated robustness and accuracy, all while maintaining high-performance standards and computational efficiency. It stands as a testament to the refined approach towards augmenting the automation capabilities in power line inspection systems.

2. Background Materials

Current object detectors are primarily categorized into single-stage [11,17], two-stage [18], and Transformer-based multi-head attention mechanism neural networks [4,19].

Two-stage detectors like the R-CNN series [20] generate candidate regions first, followed by classification and localization, which, although precise, come with high computational costs. Ren et al. introduced Faster R-CNN [21], which improves upon its predecessor, Fast R-CNN [22], by integrating a Region Proposal Network (RPN) for efficient object localization and enabling nearly cost-free region proposals. Transformer-based models, such as DETR [10] and Sparse R-CNN [3], leverage self-attention mechanisms to process the global and local contexts of images, achieving end-to-end detection but also having significant computational demands. RT-DETR [23] enhances the inference speed of the original DETR model by introducing efficient feature extraction and a lightweight design but may struggle with generalization and accuracy in complex real-world scenarios. Single-stage detectors directly predict the location and category of objects without the need for candidate regions, offering faster speeds, as seen in YOLO [13,24] and SSD [25]. SSD enhances detection accuracy with multi-scale anchors, but at the cost of higher computational expenses. Recent advancements like YOLOv5 [26] and YOLOv8 [27] have boosted detection speed and accuracy, but still struggle with small object detection. The power inspection workload is substantial, with the need for timely hazard identification, high-quality image collection, and large volumes of data, all of which pose considerable challenges for detectors.

2.1. Object Detector

Power inspection applications involve substantial workloads and intricate objectives, with a critical emphasis on detecting small-scale targets. YOLOv8 [12], an evolution in the YOLO lineage, excels in object detection due to its heightened performance and adaptability. Its architecture, as shown in Figure 1, refines the C2f (cross-stage partial bottleneck with two convolutions) [27] feature extractor for enhanced gradient flow, reducing parameters without compromising efficacy. The model's neck takes PANet [13] instructor and introduces configurable setups, skipping a convolution step for direct upsampling and adaptable layer interconnections, tailored to specific requirements. YOLOv8 flexibly connects either P3~P6 for higher resolution processing or P2~P5 for lower, expanding its operational scope [27].

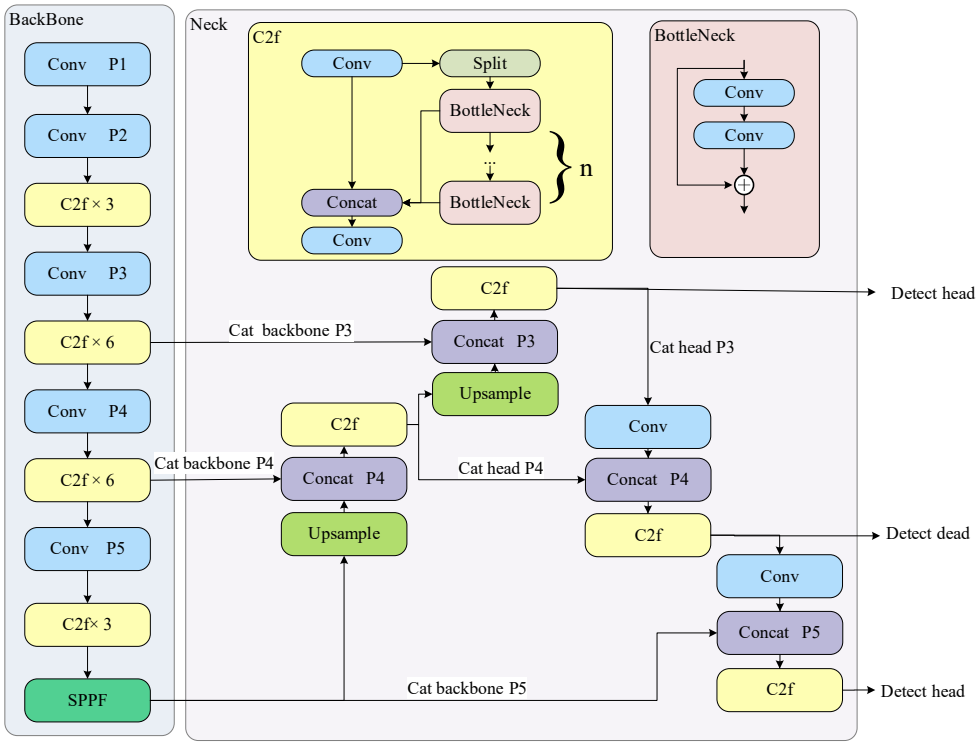

Figure 1. The network structure of YOLOv8 includes the detailed structure of C2f and BottleNeck. Here, n is the number of BottleNecks contained in C2f.

By separating the detection head tasks—classification and localization—it mitigates feature competition and accelerates convergence, thus boosting the overall performance. In bounding box refinement, YOLOv8 integrates CIoU (Complete Intersection over Union) [27] to enhance precision and training stability, particularly beneficial in tackling intricate, imbalanced datasets. Coupled with Focal Loss, this mechanism effectively addresses class imbalance and the detection of small objects, streamlining the anchor-free prediction of object locations and dimensions. Moreover, the classification branch employs the Task-aligned Assigner for assigning positive samples, which aligns targets with proposal boxes using specialized IoU calculations. This advanced matching technique yields finer object localization and excels in challenging power inspection environments, outperforming the conventional IoU-based approaches in accurately defining overlapping areas.

When using YOLOv8 for detecting densely packed small objects, the model's performance is still hindered by complex backgrounds, leading to challenges in accurately identifying small-sized targets and targets that are densely occluded. Further improvements are necessary to enhance detection capabilities under these conditions.

2.2. Query-Based Detector

Query is commonly used in Transformer models as part of the attention mechanism to calculate the attention weights. The attention mechanism in Transformer enables the model to concentrate on various aspects of the sequential data during processing, which is useful in natural language processing and other sequential modeling tasks [5,28]. Transformer introduces a multi-head attention mechanism in which the model consists of multiple independent self-attention heads. Each head has a distinct query which undergoes a linear transformation with keys and values, and then computes different outputs. Finally, these outputs are combined or merged through linear transformations to obtain the final output. In the whole attention mechanism, the role of query is to determine the level of attention of the current position concerning other positions. Different queries lead to distinct attention distributions, enabling the model to capture the relationships between different positions during the processing of the input sequence. This capability enables the Transformer to capture long-range dependencies, making it one of the key mechanisms behind its successful processing of sequential data.

Adamixer [29] introduces a new "Adaptive Mixing" strategy that dynamically adjusts interactions between features and queries, reducing the number of iterations required for model training. This strategy allows for rapid feature-level fusion and adaptive adjustment, enhancing the model's convergence speed and improving detection accuracy and efficiency. SQR [7] offers a method to expedite convergence and significantly reduce computation by selectively forwarding queries to each stage. This approach mitigates the accumulation of inference errors across stages, addressing the issue where category inference errors in one stage propagate and amplify in subsequent stages. Cross-stage interaction [30] introduces a cross-stage interaction mechanism that allows information exchange and reinforcement between features and queries across different stages, enhancing the model's feature learning capabilities. Pairwise Query-Based [31] Detection aimed at human-object interaction detection tasks combines query-based detection with global contextual information, proposing a query-based pairwise detection mechanism that effectively captures interactions between humans and objects. TSCODE [32] introduces a simplified U-Net architecture to merge adjacent upper and lower feature layers, allowing for a decoupled inference that separates classification and localization tasks. BifNet similarly leverages the fusion of feature layers from the backbone, thereby enhancing the robustness and accuracy of object detection.

The extraction and selection of queries play a critical role in object detection, particularly in datasets with complex environments like TLAD. By focusing on the queries derived from self-attention and cross-attention, we integrate the strengths of both. Self-attention excels at modeling intra-feature relationships, and cross-attention effectively captures inter-feature dependencies. This approach not only enhances the model's ability to handle

challenging backgrounds but also improves the overall detection accuracy, making it a robust solution for complex object detection tasks.

3. QYOLOv8 Algorithm

The QYOLO model integrates key technologies, such as Depthwise Separable Convolutions (DSCs) [33], cross-layer feature fusion, and row–column query attention mechanisms, achieving efficient and precise handling of small object detection tasks. This integration provides new technological and methodological support for the application of deep learning in the fields of computer vision and object detection, as shown in Figure 2.

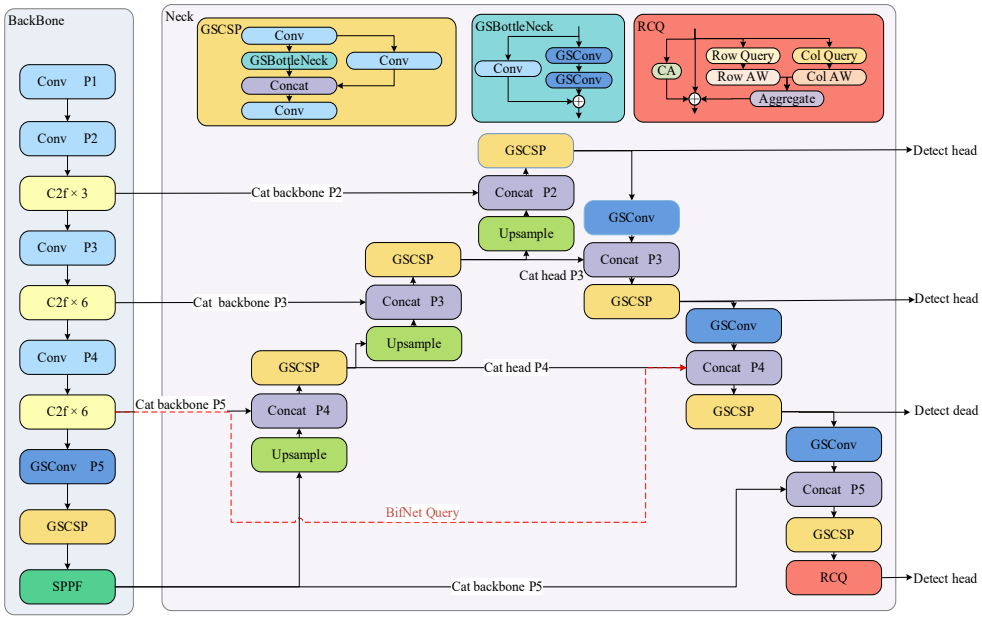

Figure 2. Network structure of QYOLOv8 includes the detailed structure of GSCSP, GSBottleNeck, and RCQ.

3.1. GSConv

In the CNN backbone, spatial details from input images progressively transition into channel information, causing a reduction in the spatial dimensions (width and height) of the feature maps across various stages. This dimensionality compression parallels an increase in channel numbers, which can potentially sacrifice some semantic context [34]. Standard Convolutions (SCs) entail higher computational demands to preserve extensive inter-channel relationships, whereas DSC [35] significantly prune these connections for efficiency. The introduction of GSConv [14,15] seeks to strike a balance between both by maintaining a maximum of these connections while keeping the computational time low, as shown in Figure 3. In the context of convolutional computation, which is commonly measured by FLOPs (Floating Point Operations), the time complexities of different convolution methods are crucial for performance evaluation. SC, DSC, and GSConv have distinct complexities; the time complexity (without bias) is illustrated in Equations (1)–(3):

$$\mathrm{TCsc} = \mathrm{O}(\mathrm{W} \cdot \mathrm{H} \cdot \mathrm{K}_1 \cdot \mathrm{K}_2 \cdot \mathrm{C}_1 \cdot \mathrm{C}_2), \tag{1}$$

$$\mathrm{TC}_{\mathrm{DSC}} = \mathrm{O}(\mathrm{W} \cdot \mathrm{H} \cdot \mathrm{K}_1 \cdot \mathrm{K}_2 \cdot \mathrm{C}_1 + \mathrm{W} \cdot \mathrm{H} \cdot \mathrm{C}_1 \cdot \mathrm{C}_2), \tag{2}$$

$$\mathrm{TC}_{\mathrm{GSConv}} = \mathrm{O}(\mathrm{W} \cdot \mathrm{H} \cdot \mathrm{K}_1 \cdot \mathrm{K}_2 \cdot \mathrm{C}_1 \cdot \mathrm{C}_2/2 + \mathrm{W} \cdot \mathrm{H} \cdot \mathrm{C}_1 \cdot \mathrm{C}_2/2), \tag{3}$$

where W and H denote the width and height of the output feature map, $K_1 \cdot K_2$ is the kernel size, C_1 is the number of input channels per kernel, and C_2 is the number of output channels. The GSConv achieves a remarkable reduction in computational expense, saving approximately 50% of the computation cost [16].

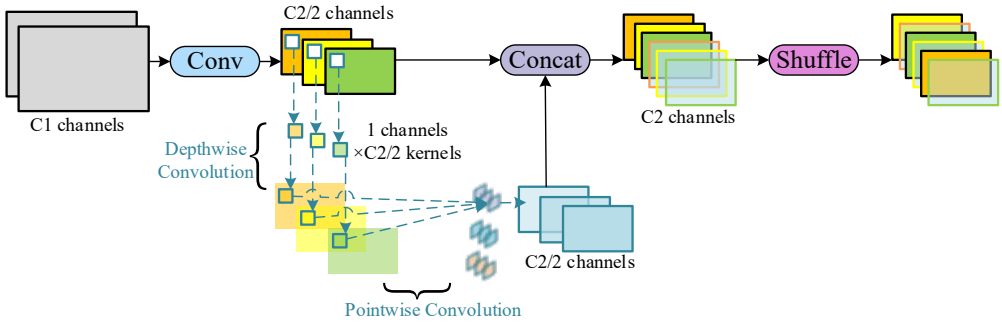

Figure 3. Network structure of GSConv. The bottom branch performs Depthwise Convolution and Pointwise convolution in turn, and the top branch performs SC.

When feature maps exhibit a large number of channels combined with reduced spatial resolution, incorporating GSConv in the neck component is particularly beneficial for processing sequentially connected feature maps. In this context, redundancy and repetition in information are minimized, rendering additional compression unnecessary.

The bottleneck operation in C2f enhances the network's ability to process features with greater precision, allowing it to capture the finer details of small objects. When incorporated into the backbone, the high-resolution feature layers enable a clearer delineation of boundaries and details for small objects in the feature map. The GSConv-based split-path (GSCSP) method offers lower computational complexity, reducing redundancy through feature partitioning. GSCSP effectively handles small objects by minimizing feature blurring and enhancing feature reuse. Its application in low-resolution feature layers and during neck feature fusion improves the overall efficiency.

3.2. Introducing Query Methods

3.2.1. BifNet Query

In BifNet, the direct connection from the backbone network to the neck network, though not strictly a query mechanism in the conventional sense, serves a function akin to that of query mechanisms: it directly extracts and harnesses key features, thereby reinforcing the effect of feature fusion. This design bolsters the expressive power of features and facilitates information flow, enhancing the performance of object detection, particularly excelling in complex scenes and the detection of small objects. Within BifNet [16], the direct linkage from the backbone to the neck network incorporates low-level or mid-level backbone features into the neck, aiding in preserving the details of these foundational features during higher-level feature integration. By circumventing information loss that typically occurs after multiple convolution operations, this direct connection renders the detection process more sensitive to small objects and intricate backgrounds, enabling the upper-level features to better embody the nuances and contextual information present in lower-level features. Consequently, the representational capacity of feature maps is augmented, thereby reinforcing the overall robustness of the detection model.

3.2.2. Row–Column Query

The introduction of the row–column query attention module further optimizes the flow of information and weight allocation between features, improving the detection performance for small objects by focusing on row and column feature information, as shown in Figure 4.

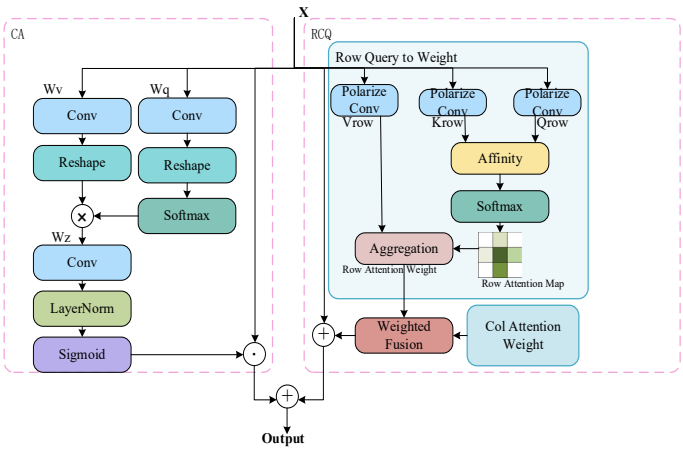

Figure 4. Network structure of RCQ, the left branch attributes channel self-attention and the right branch extracts contextual information in rows and columns.

In object detection, it is crucial to consider the interplay between self-attention and cross-attention. Channel attention [36], a form of self-attention, primarily focuses on understanding the relative positions and relationships between different instances of the same category within an image. For instance, in an image depicting a transmission tower with multiple screws, self-attention aids the model in comprehending the relative positions and potential interactions among these screws. The proposed row–column query module involves associating two distinct sequences or sources of information, constituting a form of cross-attention. This approach emphasizes cross-category relationships, aiding the model in understanding how different object types interact and coexist within the same image space. For example, it enables the model to comprehend the relationship between screws and insulators on a high-voltage transmission tower, despite these being entirely different categories.

The RCQ extraction branch is divided into Row Query extraction and Column Query extraction, both of which are converted into weights. Specifically, taking the Row Query as an example, features are extracted through horizontal convolution in polarized filtering. When distinguishing directional features, polarized filtering minimizes the loss of features in the targeted direction and reduces the parameter count for irrelevant directional features. Similarly, k and v are projected in the row direction with a convolution kernel size of 1×3, and the number of channels is reduced to $1/8$.

As shown in Equation (4), the similarity (attention maps) between Q and K is computed. By weighting the feature V, a feature map incorporating row-direction weighted information is generated, as represented in Equation (5). The row and column attention weights are fused through the adjustment of the weighting operation by a factor γ. Finally, by adding the input feature x, skip connections are established, preserving the original input information and enhancing the feature representation capability, as illustrated in Equation (6).

$$A_{row} = \text{Softmax}(Q_{row} \cdot {K_{row}}^{T}), \tag{4}$$

$$O_{row} = V_{row} \cdot A_{row}, \tag{5}$$

$$y = \gamma(O_{row} + O_{col}) + x, \tag{6}$$

where Q, K, and V are the query, key, and value feature maps respectively. A_{row} is an attention map. The output O_{row} from row attention is obtained by applying the row attention weights to the row value feature map. The final output y integrates the input feature map x with the weighted sum of the row and column outputs. The parameter γ, initialized to zero, is learned during training and modulates the contribution of the attention-based features relative to the original input.

4. Experiment

4.1. Datasets

A comprehensive image dataset was meticulously compiled for the precise detection of transmission tower components. The initial dataset collected 789 high-resolution images, each with a native dimension of 8000 × 6000 pixels, capturing the intricate details essential for meticulous analysis. To counterbalance potential biases and enrich diversity, the initial dataset was augmented through random transformations, including rotations and exposure adjustments, thereby yielding an additional 504 images. Gamma correction with parameters between 1.35 and 1.75 was used to correct overexposed images affected by incident light. Considering the images predominantly capture blue glass insulators, 200 images of red insulators sourced from the well-established SFID [36] dataset were introduced to balance the insulator types and enhance the model's generalizability. The resulting, meticulously curated dataset, designated as the TLAD in Figure 5, encapsulates a total of 1493 images and 8508 annotated instances. These instances are categorized into three primary classes: 1299 insulators, highlighting both common and variant colors; 6667 screws vital for structural integrity assessments; and 542 impact hammers used in maintenance activities. TLAD has been systematically partitioned into a training subset (80%), a testing subset (10%), and a validation (10%) subset.

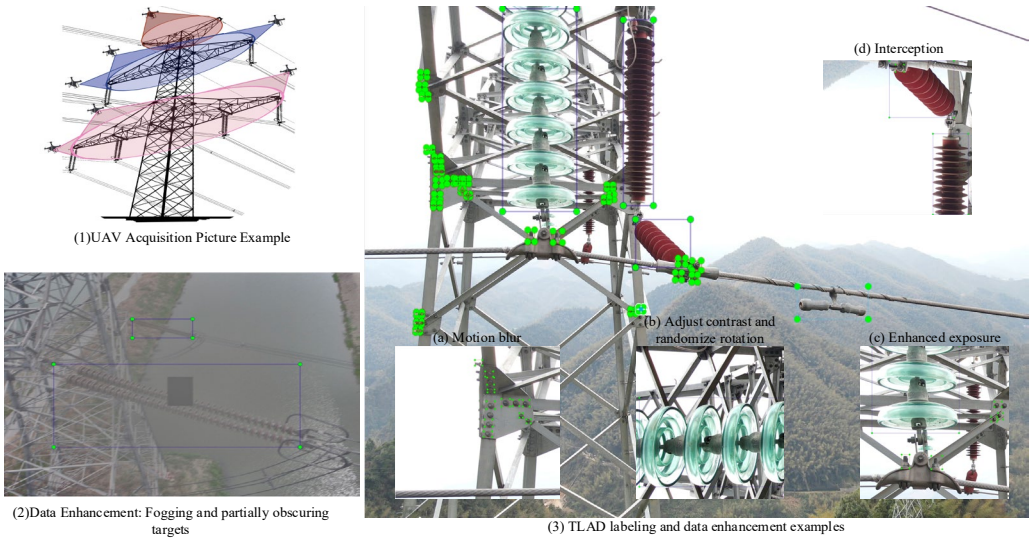

Figure 5. Dataset instruction. Figure (1) shows the position of the drone during image capture, with the coloured area indicating the photographed area. In Figures (2) and (3), the purple bounding boxes mark the insulators, the blue boxes mark the dampers and the brown boxes mark the screws. The green dots represent the corners of the bounding boxes.

To further validate the effectiveness of the QYOLO network in detecting small objects within high-resolution images, we conducted additional tests on the German Traffic Sign Detection Benchmark (GTSDB) [37], an open traffic sign detection dataset. The GTSDB hails from Germany, with a total of 900 images. The image size measures 1360 × 800 pixels, and the traffic sign sizes range from 16 × 16 pixels to 128 × 128 pixels. A notable characteristic of this dataset lies in the rich and variegated marking environments, encompassing diverse viewpoints like frontal, lateral, and oblique shooting angles, as well as conditions such as intense illumination. These multifarious environments augment the complexity and authenticity of the data, presenting significant challenges and practical utilities for the training and evaluation of traffic sign detection models.

4.2. Training Parameters

The experimental TLAD setup encompassing the software and hardware environment configured for the model's training phase is summarized in Table 1, accompanied

by the detailed training model parameters. The model receives inputs of dimensions 1280 × 1280 × 3. Extensive training spanning 100 epochs was administered, during which the learning rate was dynamically adjusted following a cosine annealing schedule. Notably, mosaic data augmentation, a common strategy for enhancing model diversity, was deactivated during the final 10 epochs to refine the model's focus on unaltered image characteristics. The loss curve over the course of 100 training epochs is shown in Figure 6. In QYOLO, the learning rate was set to 0.01 for the initial 90 epochs and reduced to 0.001 for the final 10 epochs. As the training progresses, reducing the learning rate to 0.001 allows the model to fine-tune its parameters with greater precision, improving performance and accuracy through smaller, more precise adjustments.

Table 1. Experimental configurations.

Software configuration	System: Windows 10 Frame: Pytorch1.12.0 Version: CUDA 11.7, cuDNN 8.5.0, Python 3.8
Hardware configuration	CPU: Intel Core I7-13700K GeForce RTX 4090 of GPU Graphics memory: 24 G
	UAV: DJ MAVIC2-ENTERPRISE-ADVANCED, aperture: f/2.8, the equivalent focal length 24 mm, 32× digital zoom, a single shot
Training hyperparameters	Optimizer: SGD Learning rate: 0.01 (initial 90 epochs), 0.001 (final 10 epochs) Weight decay: 0.0005 Learning momentum: 0.937

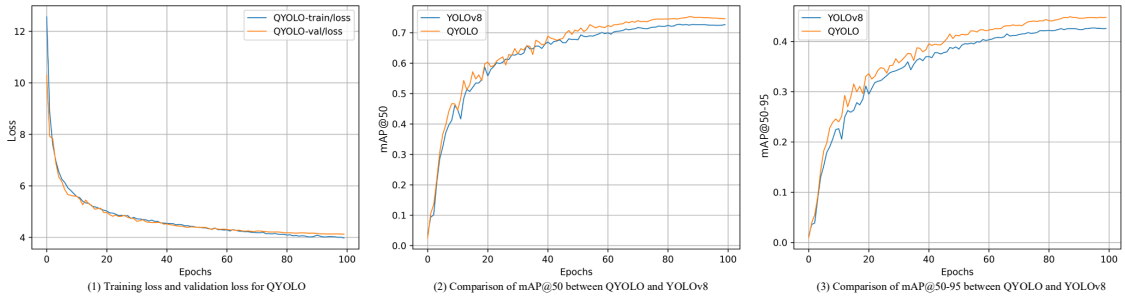

Figure 6. The loss curve, mAP@50, and mAP@50–90 changes during the training process over 100 epochs.

4.3. Evaluation Metrics

In the assessment of the model's performance across subsequent experimental evaluations, the mean Average Precision (mAP) scores at two thresholds, mAP_{50} and $mAP_{50–95}$, served as the principal benchmarks [18]. In Equations (7) and (9), mAP_{50} quantifies the aggregate precision across all the classes when the Intersection over Union (IoU) threshold is set to 0.5; in this context, a prediction is deemed accurate if it overlaps at least 50% with the true object bounding box according to the IoU criterion. Conversely, $mAP_{50–95}$ provides a holistic view of model performance by calculating the weighted mean of mAP across a range of IoU thresholds from 0.5 to 0.95, in increments of 0.05. This broader spectrum analysis offers insights into the model's capability to detect objects with varying degrees of precision and localization accuracy. In Equation (10), the F1-score was added as a reconciled average of precision and recall to provide a composite performance measure.

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \ \text{Recall} = \frac{T_p}{T_p + F_N} \tag{7}$$

$$AP = \int_0^1 P(R)dR, \tag{8}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}, \tag{9}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{10}$$

4.4. Results and Analysis

In Table 2, YOLOv8 [12] is used as a baseline to assess the performance of the improved modules BifNet [38], RCQ [39], GSConv [15], and GSCSP [16] within the QYOLO framework through ablation experiments. The YOLOv8-BifNet-RCQ model achieved an F1-score only second to QYOLO. The introduction of queries into the neck network facilitated a balanced trade-off between precision and recall, yielding robust detection results even with imbalanced positive and negative examples in the dataset. Overall, the $mAP_{50–95}$ across all the categories improved by 1.8%. Moreover, QYOLO integrates GSConv and GSBottleneck into the YOLOv8-BifNet-RCQ model, leading to an overall F1-score improvement of 2.3%.

Table 2. All-category detection results of ablation experiments.

Model	Precision	Recall	F1-Score	mAP_{50}	$mAP_{50–95}$
YOLOv8	77.5	65.9	71.2	73.6	43.1
YOLOv8-BifNet	77.9	66.7	71.9	73.5	43.9
YOLOv8-BifNet-RCQ	78.9	66.8	**72.4**	73.4	43.3
YOLOv8-GSConv-GSCSP	75.6	66.9	71.0	74.6	44.0
QYOLO	**79.3**	**68.5**	**73.5**	**75.2**	**44.9**

In power line inspection applications, the detection of small targets, such as screws, is of particular importance. The ablation experiments in Table 3 further illustrate the enhanced effectiveness of the improved modules in detecting small objects. The YOLOv8-GSConv-GSCSP model achieved a small target detection accuracy of 71.8, with an average precision that closely approaches that of QYOLO. Ultimately, QYOLO improved the mAP_{50} for small object detection by 5.5%.

Table 3. Small object detection results of ablation experiments.

Model	Precision-$_S$	Recall-$_S$	F1-Score-$_S$	$AP_{50\text{-}S}$	$AP_{50–95\text{-}S}$
YOLOv8	70	46.6	56.0	58.3	30.9
YOLOv8-BifNet	70.1	47.6	56.7	58.7	31.0
YOLOv8-BifNet-RCQ	70.6	48.6	**57.6**	59.4	31.2
YOLOv8-GSConv-GSCSP	**71.8**	45.4	55.6	**62.3**	**33.9**
QYOLO	**71.2**	**51.1**	**59.5**	**63.8**	**35.2**

Table 4 compares parameters, inference accuracy, and inference time across the various models. Compared to YOLOv5 [36], YOLOv8 offers inherent advantages in average detection precision and inference time, enhancing small object detection precision by up to 13.3%. Among the query-based YOLOv8 improved models, YOLOv8-C3tr [2] and YOLOv8-CANet [9] incorporate the C3tr and CANet modules, respectively. While these enhancements provided some improvement in average detection precision for small objects without adding additional parameters, their effects were limited. The YOLOv8-DWConv [35] model experienced increases in both parameter count and inference time. In contrast, QYOLO reduced the number of parameters while improving detection performance, with the inference time increasing to 3.9 ms, still less than that of YOLOv5 and YOLOv8-CANet.

Table 4. Experimental results for YOLOv5 and YOLOv8 improved models.

Model	Param.	mAP_{50}	$mAP_{50–95}$	mAP_{50-S}	$mAP_{50–95-S}$	Inference Time
YOLOv5	1.76M	71.5	35.4	46.1	17.6	5.8 ms
YOLOv8	3.01M	73.6	43.1	58.3	30.9	**3.0 ms**
YOLOv8-C3tr	2.72M	72.5	41.4	59.4	31.4	2.6 ms
YOLOv8-CANet	2.81M	73.1	43.1	59.0	31.0	4.0 ms
YOLOv8-DWConv	3.12M	74.5	43.6	62.3	34.2	3.2 ms
QYOLO	**2.92M**	**75.2**	**44.9**	**63.8**	**35.2**	**3.9 ms**

QYOLO achieves a balanced performance in both the overall detection and small object detection, demonstrating significant advantages in adaptability to object scales and robustness in detection. Figure 7 illustrates the detection results of two test images using the YOLO and QYOLO models. The QYOLO model successfully detects more screws, even those partially obscured by the tower structure. Notably, in the areas marked by yellow circles in (1) and (3), QYOLO successfully identified the insulators hidden behind the metal framework, as indicated by the red bounding boxes. Additionally, QYOLO achieves a higher recognition rate of screws that are densely packed and closer to the camera compared to YOLOv8.

Figure 7. Object detection results based on QYOLO and YOLO. The red boxes indicate the recognized targets as insulators, the pink boxes denote screws, and the blue boxes highlight impact hammers. They emphasize the areas of significant difference that both models need to focus on.

In Table 5, the QYOLO model maintains a relatively balanced precision and recall rate, and its F1-score is also relatively high, indicating an improvement in comprehensive performance to further validate QYOLO's enhanced robustness in detecting small objects in high-resolution images under limited lighting conditions.

Table 5. Experimental results of the GTSDB dataset in traffic sign recognition.

Model	Precision	Recall	F1-Score	mAP_{50}	$mAP_{50–95}$	Inference Time
YOLOv8	96.1	90.2	93.06	94.6	78.8	1.7 ms
QYOLO	**97.1**	**92.2**	**94.59**	**94.8**	**79.4**	**2.1 ms**

5. Conclusions

To achieve multi-scale object detection for transmission towers, we propose an enhanced QYOLO algorithm. Our approach integrates GSConv into YOLOv8, incorporating DWConv, which halves the computational load. This improvement is applied to the bottleneck and GSCSP modules containing GSConv, thereby significantly enhancing the network's capacity to learn image features. Inspired by BifNet, we introduce a query-based enhancement method in the neck network, combining BifNet with RCQ to further improve feature learning. The experimental comparisons demonstrate that the QYOLO algorithm markedly enhances the accuracy of small object detection, reduces missed detections, and increases the overall detection robustness. Compared to YOLOv8, the proposed method improves the average precision for small objects by 5.5% and the F1-score by 3.5%. The proposed method effectively identifies maintainable components' locations, pinpointing critical parts for inspection and providing essential maintenance information, which is crucial for improving inspection efficiency in power line patrols by accurately recognizing screws and other key components. Additionally, it can be validated on a traffic sign detection dataset, which resulted in further improvements.

Author Contributions: Conceptualization, W.W. and Z.W.; Methodology, M.G., J.M. and B.W.; Validation, W.W.; Formal analysis, J.M. and Z.W.; Investigation, J.X.; Resources, B.W.; Data curation, J.M.; Writing – original draft, M.G.; Writing – review & editing, M.G. and W.W.; Visualization, Z.W.; Supervision, J.M. and J.X.; Project administration, W.W., J.X. and B.W.; Funding acquisition, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the High-tech Projects of Shanghai Science and Technology Innovation Action Plan 2022 [22511100600] and the Project of Shanghai Pudong New Area Science and Technology Development Fund for People's Livelihood Research (Research on Online Quality Inspection System of Construction Waste Concrete Recycled Aggregate Based on Machine Vision [PKJ2023-C02]).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: Author Mingyang Gao is affiliated with the institution Shanghai Advanced Research Institute, the Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. Author Jun Xiong was employed by the company State Grid Fujian Electric Power Co. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Lu, L.; Dai, F. Accurate Road User Localization in Aerial Images Captured by Unmanned Aerial Vehicles. *Autom. Constr.* **2024**, *158*, 105257. [CrossRef]
2. 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 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; IEEE/CVF: Piscataway, NJ, USA, 2021; pp. 2778–2788.

3. Du, B.; Huang, Y.; Chen, J.; Huang, D. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE/CVF: Piscataway, NJ, USA, 2023; pp. 13435–13444.
4. Dian, S.; Zhong, X.; Zhong, Y. Faster R-Transformer: An Efficient Method for Insulator Detection in Complex Aerial Environments. *Measurement* **2022**, *199*, 111238. [CrossRef]
5. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE Computer Society: Piscataway, NJ, USA, 2023; pp. 14408–14419.
6. Kontogiannis, S.; Konstantinidou, M.; Tsioukas, V.; Pikridas, C. A Cloud-Based Deep Learning Framework for Downy Mildew Detection in Viticulture Using Real-Time Image Acquisition from Embedded Devices and Drones. *Information* **2024**, *15*, 178. [CrossRef]
7. Chen, F.; Zhang, H.; Hu, K.; Huang, Y.; Zhu, C.; Savvides, M. Enhanced Training of Query-Based Object Detection via Selective Query Recollection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 23756–23765.
8. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13668–13677.
9. Liu, Y.; Li, H.; Hu, C.; Luo, S.; Luo, Y.; Chen, C.W. Learning to Aggregate Multi-Scale Context for Instance Segmentation in Remote Sensing Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, 1–15. Available online: https://ieeexplore.ieee.org/document/10412679 (accessed on 12 September 2023).
10. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE Computer Society: Piscataway, NJ, USA, 2022; pp. 13609–13617.
11. Yin, X.; Yu, Z.; Fei, Z.; Lv, W.; Gao, X. PE-YOLO: Pyramid Enhancement Network for Dark Object Detection. In Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2023, Crete, Greece, 26–29 September 2023; Iliadis, L., Papaleonidas, A., Angelov, P., Jayne, C., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 163–174.
12. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO, version 8.0.0; Computer Software. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 23 August 2023).
13. Soylu, E.; Soylu, T. A Performance Comparison of YOLOv8 Models for Traffic Sign Detection in the Robotaxi-Full Scale Autonomous Vehicle Competition. *Multimed. Tools Appl.* **2023**, *83*, 25005–25035. [CrossRef]
14. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the Computer Vision—ECCV 2018, PT XIV, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer Nature: Heidelberg, Germany, 2018; Volume 11218, pp. 122–138.
15. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-Neck by GSConv: A Better Design Paradigm of Detector Architectures for Autonomous Vehicles. *arXiv* **2022**, arXiv:2206.02424.
16. Ben, Y.; Li, X. Dense Small Object Detection Based on Improved Deep Separable Convolution YOLOv5. In Proceedings of the Image and Graphics, Chongqing, China, 6–8 January 2023; Lu, H., Ouyang, W., Huang, H., Lu, J., Liu, R., Dong, J., Xu, M., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 103–115.
17. Huang, H.; Feng, Y.; Zhou, M.; Qiang, B.; Yan, J.; Wei, R. Receptive Field Fusion RetinaNet for Object Detection. *J. Circuits Syst. Comput.* **2021**, *30*, 2150184. [CrossRef]
18. Zhao, Z.; Zhen, Z.; Zhang, L.; Qi, Y.; Kong, Y.; Zhang, K. Insulator Detection Method in Inspection Image Based on Improved Faster R-CNN. *Energies* **2019**, *12*, 1204. [CrossRef]
19. Cao, X.; Zhang, Y.; Lang, S.; Gong, Y. Swin-Transformer-Based YOLOv5 for Small-Object Detection in Remote Sensing Images. *Sensors* **2023**, *23*, 3634. [CrossRef] [PubMed]
20. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
21. 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.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
22. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
23. Yu, C.; Shin, Y. An Enhanced RT-DETR with Dual Convolutional Kernels for SAR Ship Detection. In Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan, 19–22 February 2024; pp. 425–428.
24. Shi, H.; Yang, W.; Chen, D.; Wang, M. CPA-YOLOv7: Contextual and Pyramid Attention-Based Improvement of YOLOv7 for Drones Scene Target Detection. *J. Vis. Commun. Image Represent.* **2023**, *97*, 103965. [CrossRef]
25. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.

26. Cao, S.; Wang, T.; Li, T.; Mao, Z. UAV Small Target Detection Algorithm Based on an Improved YOLOv5s Model. *J. Vis. Commun. Image Represent.* **2023**, *97*, 103936. [CrossRef]
27. Terven, J.; Córdova-Esparza, D.-M.; Romero-González, J.-A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
28. Lin, X.; Sun, S.; Huang, W.; Sheng, B.; Li, P.; Feng, D.D. EAPT: Efficient Attention Pyramid Transformer for Image Processing. *IEEE Trans. Multimed.* **2021**, *25*, 50–61. [CrossRef]
29. Gao, Z.; Wang, L.; Han, B.; Guo, S. AdaMixer: A Fast-Converging Query-Based Object Detector. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5354–5363.
30. Teng, Y.; Liu, H.; Guo, S.; Wang, L. StageInteractor: Query-Based Object Detector with Cross-Stage Interaction. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 6554–6565.
31. Tamura, M.; Ohashi, H.; Yoshinaga, T. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10405–10414.
32. Zhuang, J.; Qin, Z.; Yu, H.; Chen, X. Task-Specific Context Decoupling for Object Detection. *arXiv* **2023**, arXiv:2303.01047. [CrossRef]
33. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
34. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
35. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE Computer Society: Piscataway, NJ, USA, 2017; pp. 764–773.
36. Zhang, Z.-D.; Zhang, B.; Lan, Z.-C.; Liu, H.-C.; Li, D.-Y.; Pei, L.; Yu, W.-X. FINet: An Insulator Dataset and Detection Benchmark Based on Synthetic Fog and Improved YOLOv5. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–8. [CrossRef]
37. Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; Igel, C. Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark. In Proceedings of the International Joint Conference on Neural Networks, Dallas, TX, USA, 4–9 August 2013.
38. Li, H.; Chen, Y.; Zhang, Q.; Zhao, D. BiFNet: Bidirectional Fusion Network for Road Segmentation. *IEEE Trans. Cybern.* **2022**, *52*, 8617–8628. [CrossRef] [PubMed]
39. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention Mechanisms in Computer Vision: A Survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]