technologies

MDPI

Article

IISD-YOLO: Infrared Detection of Insulator Strings for Transmission Lines Based on Improved YOLOv11

Chen-Hao Zhao, Yi-Feng Ren *, Long-Kun Cao and Hong-Yu Wang

School of Electrical and Control Engineering, North University of China, Taiyuan 030051, China; s202415011@st.nuc.edu.cn (C.-H.Z.); s202415009@st.nuc.edu.cn (L.-K.C.); sz202415013@st.nuc.edu.cn (H.-Y.W.)
* Correspondence: 19890612@nuc.edu.cn

Abstract

In the area of transmission line inspection, one of the prominent areas of research has been to unite Unmanned Aerial Vehicles (UAVs) with neural network object detection algorithms. This area of research is challenging because of high computational resource consumption and poor infrared detection capabilities. In this study we propose an infrared image detection algorithm, named IISD-YOLO, using a modified version of the YOLOv11 network, to detect infrared transmission line insulator strings. Firstly, the original object detection layer was removed and replaced with the ShuffleNetv2 network to achieve the goal of a lightweight model; subsequently, based on the original feature extraction module C3k2, the Manhattan Self-Attention (MaSA) mechanism was introduced to design a new feature extraction module, C3k2-MaSA, which enhances the feature extraction capability for infrared objects; finally, the bidirectional feature pyramid network (Bi-FPN) is used to replace the original feature fusion module, enhancing the network's ability to process and fuse information at different scales. The comparative experiments show that compared with the mainstream YOLO models, IISD-YOLO has improved by 4.5, 6.1, and 4.8 percentage points respectively on mAP@50 over YOLOv5, YOLOv8, and YOLOv10; furthermore, this model outperforms advanced models including YOLO-CIR, FA-YOLO, YOFIR, and RT-DETR, with improvements of 2.9, 9.1, 5.0, and 1.1 percentage points respectively on mAP@50. The ablation study shows that each improvement effectively enhances the overall performance. Compared with the original YOLOv11, the IISD-YOLO has increased its mAP@50 by 3.5 percentage points, while reducing the number of Params by 1.1 million and the computational GFLOPs by 2 G. These results confirm the superior performance of IISD-YOLO in infrared insulator string detection.

Keywords: deep learning; drone; electrical engineering; infrared imaging; insulator string

Check for updates

Academic Editor: Spyridon Nikolaidis

Received: 23 March 2026
Revised: 12 May 2026
Accepted: 13 May 2026
Published: 18 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the wide application of drone technology in various applications such as agriculture, healthcare, and military [1], the demand for image recognition algorithms for drones is increasing. Due to their unique features such as light weight and high degree of freedom, drones are increasingly widely used in the inspection of transmission lines. Especially those equipped with infrared cameras, they play a crucial role in detecting power equipment prone to thermal defects in transmission lines, such as insulator strings and conductors.

In the field of infrared detection of power equipment, scholars have conducted extensive research: G. Wu et al. [2] proposed a grayscale-based template matching algorithm, which compares the template image with the target image to determine the position and

extract the region. This design is applied to the online detection of power equipment; S. Han et al. [3] proposed a deep convolutional neural network (DCNN) based on the fast region of interest (ROI) with hotspot sensitivity of infrared images. This model is applied in embedded devices such as cameras or robots. However, most of the existing algorithms for infrared detection are designed for fixed embedded devices and have relatively high requirements for the computing power of the devices. Therefore, designing a simple and efficient infrared detection algorithm for UAV plays a crucial role in enhancing the inspection capabilities of UAV on power transmission lines.

During the operation of electrical equipment, due to the Joule effect, a certain amount of heat is generated. Long-term operation is prone to heat accumulation, forming the thermal characteristics of electrical equipment. The detection system based on the infrared imaging principle has high sensitivity to the thermal characteristics of power equipment [4], and can accurately locate the overheated power equipment in low-light and night environments. This greatly enhances the detection ability of power transmission line inspectors for power equipment in complex environments and has strong practical value.

Drones are a special type of platform that can carry various devices. They can be equipped with different kinds of sensors for different application scenarios to meet the diverse needs of users. Under the inspection conditions in outdoor environments, drones equipped with infrared thermal imagers can accurately capture the thermal characteristics of power equipment [5]. Based on this function, a target detection algorithm based on deep learning is installed on the UAV to achieve the detection of power equipment.

There are two types of object detection algorithms based on deep learning: one is the two-stage detection algorithm, and the other is the one-stage detection algorithm. Take the R-CNN (Region-based Convolutional Neural Network) series [6–10] as an example. The two-stage detection algorithm has high accuracy, but it brings extremely high computational costs and has high requirements for the computing power of the platform. In contrast, taking the YOLO (You Only Look Once) series [11–15] as an example, one-stage detection algorithms maintain a certain level of accuracy while having a very fast inference speed and consuming less platform computing power, making them suitable for deployment on platforms with limited computational resources.

The general principle of RGB image capture is to capture the visible light reflected by objects, consisting of three channels, each encoded with 8-bit integers. Under sufficient lighting conditions, it provides rich color and texture information. In contrast, infrared images measure the heat emitted by objects, representing the temperature distribution. They are usually single-channel pseudo-color images with higher bit depth. The original infrared data is generally in the form of 16-bit floating-point numbers and normalized into a heat map. When the temperature difference is small, the contrast between the infrared target and the background is low, lacking the color and texture features for target recognition.

In addition, the thermal characteristics of insulators change significantly with load and health conditions. Defective insulators will present local hotspots, and the intra-class differences in these thermal features increase the difficulty of feature learning; at the same time, in the images captured by drones, the scale, direction, and distance of insulator strings are variable, requiring powerful multi-scale processing capabilities.

In view of the unique value and challenges of UAV equipped with infrared cameras in the field of power transmission line inspection, this study proposes an improved deep learning object detection algorithm based on the YOLO11 framework, named IISD-YOLO. Specifically designed for the detection of infrared insulator strings, IISD-YOLO aims to maintain a certain level of accuracy while reducing the consumption of computational re-

sources. Through the optimization of the algorithm, IISD-YOLO has achieved the detection of insulator strings in power transmission lines.

The contributions of this paper are as follows:

(1) To address the issue of excessive computational resources caused by the large number of parameters in the backbone network of the YOLO11 model and facilitate its deployment on UAV, the backbone network is optimized. This study removed the object detection layer and replaced it with the ShuffleNetv2 network as the object detection layer in the backbone network, aiming to achieve a lightweight goal.

(2) To enhance the performance of the feature extraction module of YOLO11, especially for the feature extraction of infrared pseudo-color objects, this study introduces a novel feature extraction module named C3k2-MaSA which is based on the original C3k2 module and the MaSA mechanism. This module can adjust the corresponding weights based on the relative relationships at different positions on the image, achieving prior perception of the image space. The enhanced model has a greater ability to pay attention to low-contrast targets and complex features, thereby significantly improving the overall performance and detection accuracy of the model.

(3) To address the issue that the Path Aggregation Feature Pyramid Network (PAFPN) in the YOLO11 model has a relatively weak ability to handle information of different scales at different levels and lacks adaptability to objects in complex environments, in this study, the PAFPN in the feature fusion layer was removed and replaced by the Bi-FPN network to perform multi-scale feature fusion, achieving better integration of color and texture features while enhancing the efficiency of feature integration and operation efficiency.

(4) By integrating the improved modules with the YOLO11, this study has developed an object detection model for insulator strings of power transmission lines in infrared images from the perspective of UAV.

The IISD-YOLO algorithm proposed in this paper is specifically designed for the infrared image detection of insulator strings on transmission lines captured by UAV. It is only applicable to scenarios with limited computing resources and requiring real-time inference. Compared with existing methods, this algorithm achieves a better balance between accuracy and efficiency. However, this algorithm still has limitations. Due to the lack of public datasets, this experiment uses an expanded dataset to simulate complex environments. However, this cannot fully represent the real environment. Its performance may decline under adverse weather conditions. At the same time, the current model only focuses on insulator strings and cannot simultaneously detect other components of transmission lines, reducing the application scope of this algorithm. Future work will overcome these problems by expanding the dataset and designing better networks.

The structure of this paper is as follows:

(1) The Introduction section reviews the research conducted by other researchers in the field of power equipment detection for transmission lines, and introduces the possibility of using UAV and infrared imaging technology. By introducing related detection algorithms, it leads to the algorithm of this study.

(2) The Related Works section reviews the related work based on YOLO for object detection and infrared image algorithms, and compares their features, advantages, and limitations.

(3) The Model Architecture of IISD-YOLO section details the proposed IISD-YOLO architecture, including the ShuffleNetv2 backbone network, the C3k2-MaSA module, and the Bi-FPN feature fusion layer.

(4) The Simulations section describes the experimental setup, dataset, evaluation metrics, and presents the results and discussions of comparative experiments and ablation experiments.
(5) The Conclusion section summarizes the entire paper and looks forward to future research directions.

2. Related Work

Currently, in the field of image recognition, many scholars are using the YOLO series of algorithms as the basic framework for improving object detection algorithms to meet various requirements in object detection tasks. Among them, there is the MC-YOLOv8 model for power equipment fault detection [16]; there is the improved YOLO algorithm for diagnosing equipment faults in substations [17]; and there is the electrical equipment classification algorithm using the improved YOLOv10 [18]. The YOLO series of algorithms strikes a balance between detection accuracy and inference speed. Their excellent performance enables them to operate effectively in various environments, which is why they are widely studied and used.

2.1. YOLOv5 Network Model

The YOLOv5 algorithm, proposed by G. Jocher [19] in 2020, is a one-stage deep learning object detection algorithm. Compared with traditional two-stage detection algorithms, the YOLOv5 algorithm achieves more efficient object detection while maintaining a certain level of detection accuracy. The network structure of YOLOv5 is composed of four parts: input, backbone, neck, and head.

The input stage of the YOLOv5 algorithm adopts the Mosaic data augmentation method to simulate multiple scenarios and increase data diversity. In the backbone network, through the Cross-Stage Partial (CSP) structure, the use of dense residual blocks for processing and the direct concatenation of the two parts for output significantly reduces the computational burden. In the neck network, a combination of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) is adopted for multi-scale feature fusion, which compensates for the problem of target information loss after using FPN alone. The prediction head adopts a coupled detection approach to simultaneously address classification and regression problems.

2.2. YOLOv8 Network Model

The YOLOv8 algorithm was proposed by G. Jocher [20] in 2023 and is an improved version of the YOLOv5 algorithm. The YOLOv8 algorithm uses an innovative C2f module to replace the C3 module in YOLOv5, optimizing the feature extraction process. In the neck network, YOLOv8 integrates the SPPF module at the end of the backbone network, rather than merely at the beginning of the neck network, which enhances the fusion effect of the subsequent bidirectional feature pyramid structure. The prediction head of YOLOv8 differs from that of YOLOv5 in that it adopts decoupled detection, completing classification and regression tasks in parallel to enhance the detection accuracy of the model. YOLOv8 transitions from the anchor-based design of YOLOv5 to an anchor-free design, eliminating the reliance on predefined anchor boxes and reducing the waste of computational resources. YOLOv8 incorporates the Distribution Focal Loss (DFL) function, which, combined with the CIoU loss function, enhances detection accuracy.

2.3. YOLOv10 Network Model

YOLOv10 algorithm [21] is an improved version of YOLOv8. In its backbone network, it not only retains the excellent modules such as C2f and SPPF from YOLOv8 but also innovatively introduces modules like SCDown and PSA, significantly enhancing the efficiency

and accuracy of feature extraction of the model. The YOLOv10 algorithm adopts the CIB and SCDown modules in the neck network, organically integrating the detailed features from the early stage and the semantic-rich features from the later stage, reducing the loss of feature information and enhancing the detection accuracy. The YOLOv10 algorithm innovatively adopts a dual label assignment strategy in the prediction head. During the training stage, it uses a one-to-many label assignment strategy, enabling the model to learn more feature information and enhance the accuracy of detection. In the inference stage, a one-to-one label assignment strategy is adopted, which avoids redundant predictions and eliminates the need for Non-Maximum Suppression (NMS), achieving efficient object detection.

2.4. YOLOv11 Network Model

The YOLOv11 [22] model is an improved version of YOLOv10. In the main network, the C3K2 module is used instead of the C2f module in YOLOv10. By using feature map segmentation and employing more small-scale convolution kernels, the internal information flow of the network becomes more efficient, and a more efficient feature representation can be achieved with fewer parameters. The C2PSA module integrates Pointwise Spatial Attention (PAS) with C2f; by combining the multi-head attention mechanism and the feedforward neural network, it enhances its ability to extract important features and identify key targets. In the detection head, YOLOv11 adopts a decoupled head structure, using two independent networks to handle the classification and regression tasks separately, effectively improving the independence of tasks and the overall detection performance. Finally, the YOLOv11 model introduces an Anchor-Free structure, reducing the parameter quantity while enhancing the model's generalization ability.

2.5. RT-DETR Network Model

The backbone network of the RT-DETR model [23] is based on the ResNet architecture. Through a series of convolutional operations, it generates feature maps with a three-level pyramid structure, corresponding to the output features of different levels in the network, thereby achieving multi-level feature extraction and representation learning. In terms of parameter initialization, the model loads pre-trained weights from ImageNet to accelerate convergence, and adopts a transfer learning strategy to fine-tune the end network layers, thereby retaining the general visual representation ability while optimizing the feature extraction performance for specific target detection tasks. In response to the computational explosion problem caused by the drastic increase in sequence length when the Transformer encoder processes multi-scale features in the traditional DETR model, RT-DETR proposes an efficient hybrid encoder. By introducing a hierarchical feature processing mechanism, it decouples and optimizes the intra-scale interaction and cross-scale fusion processes of image features, significantly improving the feature encoding efficiency. The adaptive decoder adopts an uncertainty minimization query selection mechanism and a dynamic layer adjustment strategy to achieve the optimal balance between detection quality and inference efficiency.

2.6. Infrared Image Target Detection Algorithm Based on YOLO Network

In recent years, with the development and wide application of infrared camera technology and target detection algorithms, an increasing number of researchers have chosen to use the YOLO network as the basic detection algorithm and combine it with the excellent ideas of other algorithms for target detection in infrared images. This research direction not only broadens the coverage of the target detection field but also introduces new innovation points for the development of the YOLO network.

For instance, J. Zhou et al. [24] proposed an infrared image object detection algorithm named YOLO-CIR based on the YOLO and ConvNext networks. Firstly, in order to adapt to the high bit width characteristics of infrared images, an infrared image enhancement algorithm was proposed, and the image preprocessing algorithm without bit width compression was improved. Subsequently, in order to adapt to low-resolution infrared images, a multi-scale feature extraction network was constructed based on the ConvNext network. Finally, attention modules are respectively introduced into the ConvNext network and the neck network to focus on the target and enhance the feature fusion capability.

S. Du et al. [25] proposed an occluded infrared target detection algorithm for complex backgrounds based on the YOLOv4 network, named FA-YOLO. To reduce the influence of complex backgrounds, a negative sample focus mechanism was adopted, concentrating on training confusing negative samples, suppressing false alarms and improving detection accuracy. To enhance the detection of infrared occluded target features, a dilated convolution block attention module (dilated CBAM) was added to the backbone network of YOLOv4 to achieve the detection of infrared targets.

M. Wen et al. [26] proposed an infrared image enhancement and detection algorithm based on YOLOv8 and FasterNet networks. Infrared images often suffer from blurriness or rough edges of targets due to their unique imaging principles. To address this issue, an algorithm specifically designed for infrared images, YOFIR, is proposed. Firstly, apply the CHALE, Auto Gamma, histogram equalization and bilateral filtering methods to process the infrared images respectively, and then fuse the results together with different weights to enhance the image quality. Then, the improved FasterNet network is used to replace the backbone network in YOLOv8, enhancing the detection accuracy while reducing the model's parameter count by adopting a new convolutional module, GSConv. Finally, a novel feature extraction module, Efficient Multi-Scale Group Conv (EMSGC), was designed to replace the original C2f module in the network to enhance the feature fusion capability. Additionally, the self-attention mechanism, DyHead block, was adopted as the detection head of the model to improve the attention to small targets.

To provide a structured comparison of the aforementioned mainstream YOLO algorithms, the RT-DETR algorithm, and their variants, Table 1 summarizes their main features, advantages, and limitations.

Table 1. Comparison of mainstream detection algorithms.

Model	Key Characteristics	Advantages	Limitations
YOLOv5	CSPNet backbone; FPN + PAN neck; coupled detection head	Efficient target detection, maintaining a certain level of accuracy; the computational burden is small	limited multi-scale fusion
YOLOv8	C2f module; SPPF module at backbone end; coupled detection head; DFL + CIoU loss function	Improvement in parallel processing efficiency; the accuracy has improved.	Increased parameters and FLOPs
YOLOv10	SCDown, PSA, CIB module; dual-label assignment strategy	The efficiency and accuracy of feature extraction have been significantly improved; NMS-free	Poor infrared adaptation
YOLOv11	C3K2, C2PSA module	Fewer parameters, efficient feature representation; enhance the recognition of key targets; good generalization ability	Poor infrared adaptation

Table 1. *Cont.*

Model	Key Characteristics	Advantages	Limitations
RT-DETR	ResNet backbone Network ImageNet pre-training; high-efficiency hybrid encoder; adaptive decoder	Solve the computational explosion problem of traditional DETR; high encoding efficiency; the balance between detection quality and reasoning efficiency	Increased parameters and FLOPs
YOLO-CIR	Based on YOLO + ConvNext; high-bit-width infrared image preprocessing; multi-scale feature extraction network; introduces the attention module	Adapt to low-resolution infrared images; enhance the feature fusion capability	Increased parameters and FLOPs
FA-YOLO	Negative sample focusing mechanism; dilated convolutional block attention module	Reduce interference from complex backgrounds; suppress false alarms; improve the detection of occluded targets	Low precision
YOFIR	Based on YOLOv8 + FasterNet; weighted fusion of multiple image enhancements; GSConv module; EMSGC module; DyHead self-attention detection head	Improve the quality of infrared images; reduce the number of model parameters; enhance the focus on small goals	Increased parameters and FLOPs

Although the above algorithms have been optimized for infrared image target detection, they enhance the performance of target detection in complex environments like infrared images by increasing computational load, which makes them unsuitable for deployment on edge devices such as UAV. Due to the limited computing power of edge devices such as drones, the detection models deployed on drones must be made lightweight in terms of their architecture, requiring fewer parameters and lower computational complexity.

UAV inspection of power lines has been widely applied. Among them, the detection of insulator strings plays a crucial role in the positioning of inspection personnel. Especially since insulator strings have certain thermal characteristics, the use of infrared target detection is very appropriate. It is necessary to further study and improve the detection performance in this field.

3. The Model Architecture of IISD-YOLO

In order to better solve the problems encountered by the infrared images of insulator strings in the perspective detection of unmanned aerial vehicles, this paper improves the YOLOv11 object detection algorithm and proposes the IISD-YOLO model, an infrared insulator string detection algorithm for unmanned aerial vehicles. Firstly, to address the issue of the huge parameter size of the traditional YOLOv11, this study replaces the object detection layer in the backbone network with the ShuffleNetv2 network [27]. This network achieves a light weight by using methods such as restricted element-wise operations and channel splitting. Secondly, this study proposes a new feature extraction module, C3k2-MaSA. This module introduces the MaSA mechanism [28] on the basis of the existing feature extraction module C3k2 in the original YOLOv11 network, and establishes a two-dimensional bidirectional spatial attenuation matrix based on the Manhattan distance between tokens, which is used to measure the relationship size between different positions

in the image and enhance the attention to the target position, thereby improving the overall detection ability of the network. Finally, this paper replaces the original feature fusion layer with the BiFPN [29]. This network adopts an efficient bidirectional cross-scale connection method and weighted feature fusion, which improves the ability to fuse information from different layers and different scales while achieving higher fusion efficiency. The architecture of IISD-YOLO is shown in Figure 1.

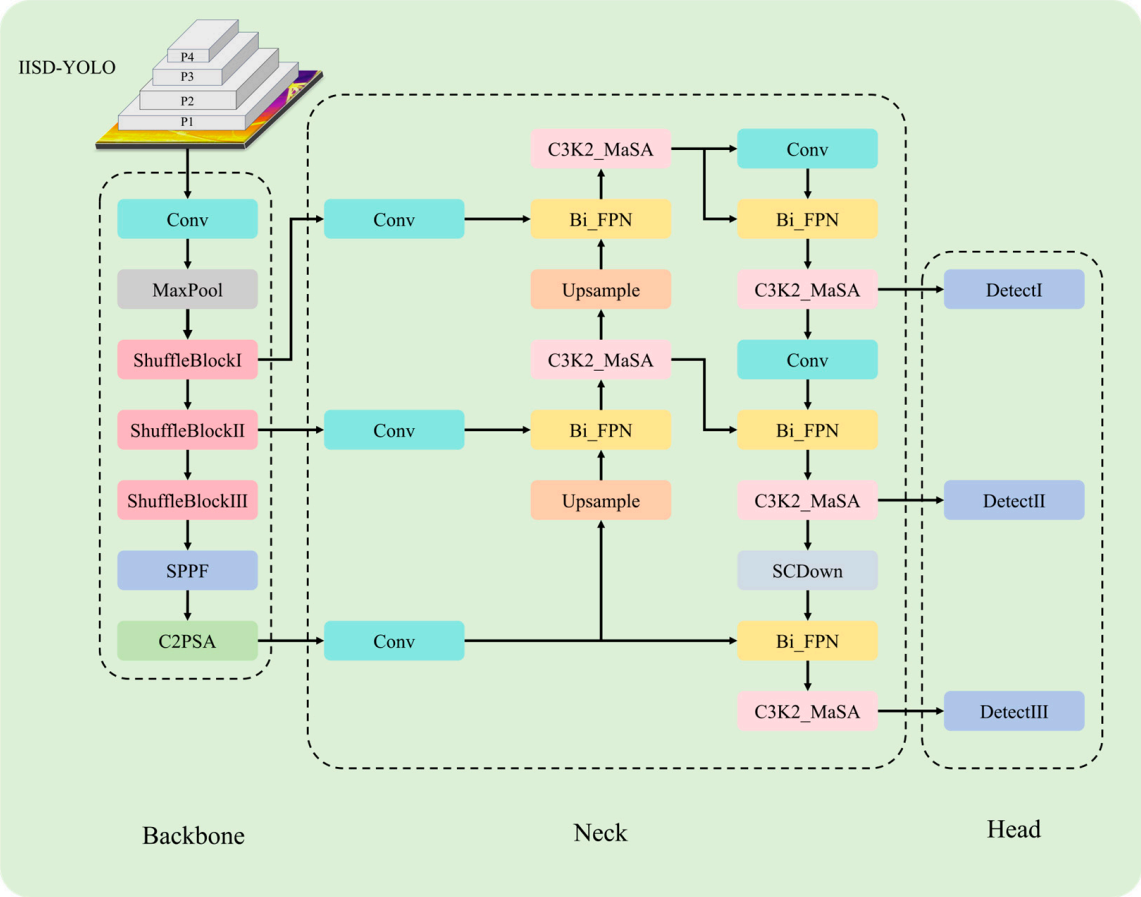


Figure 1. The network structure of IISD-YOLO.

3.1. Backbone Network Improvement: ShuffleNetv2 Network

In the YOLOv11 object detection network, the backbone network serves to extract the features of the objects in the image, enhancing the feature expression ability and detection performance of the detection network. However, it is accompanied by issues such as a large number of model parameters, large volume, and high computational load, which makes the YOLOv11 model more suitable for deployment on GPUs. The goal of this paper is to deploy the target detection model on edge devices such as UAV. To address this issue, the object detection section of the ShuffleNetv2 network will be adopted as the backbone network for the object detection layer, aiming to achieve the purpose of a lightweight network.

The ShuffleNetv2 network was proposed by N. Ma et al. [27] as a lightweight network model, achieving a balance between running speed and accuracy. The core of the network consists of two basic units, as shown in Figure 2. The target detection process is mainly divided into three stages. In each stage, the feature map spatial size is halved through the downsampling module (**a**), and then feature extraction is achieved through n cycles of the feature extraction module (**b**).

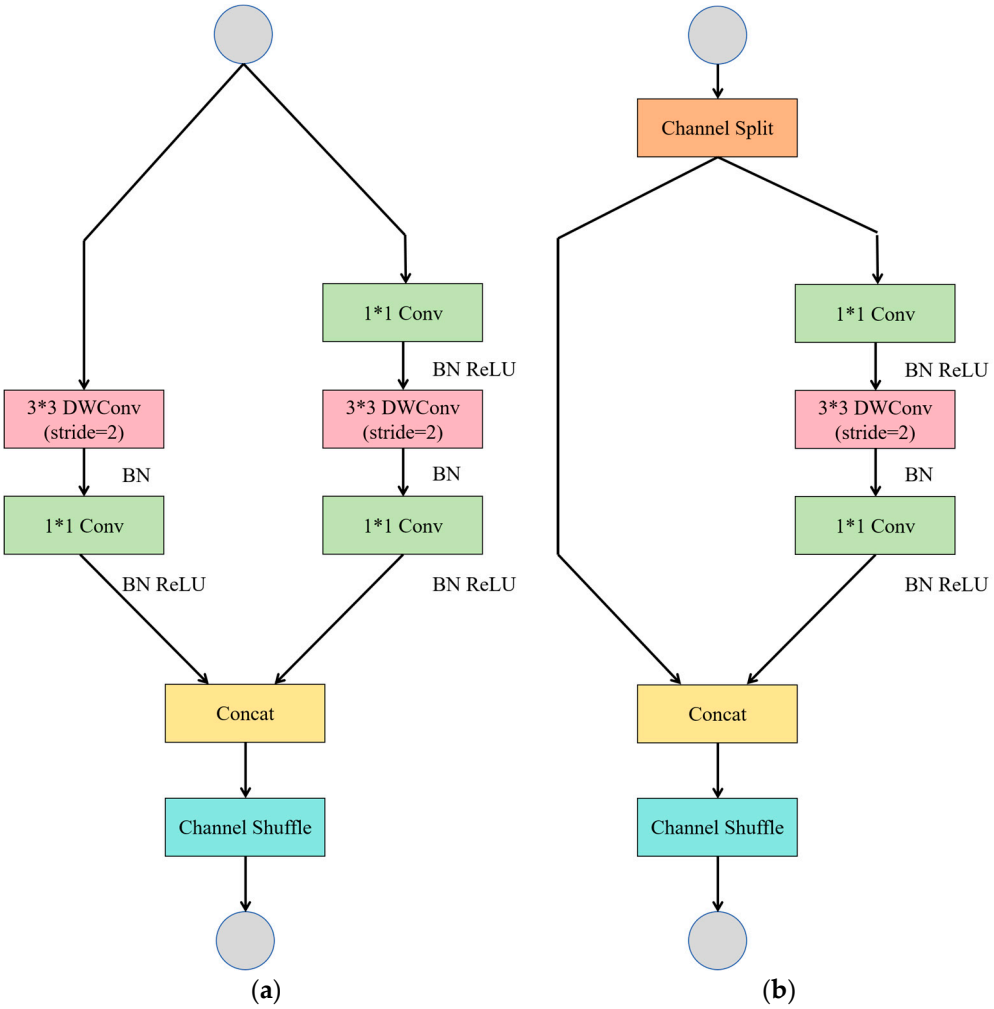


Figure 2. The part of ShuffleNetv2 Network. (**a**) The downsampling module. (**b**) The feature extraction module.

For the feature extraction module, its foundation is **an** inverted residual structure [30]. Specifically, it uses a $1 * 1$ convolution to expand the low-dimensional feature map to a high-dimensional space, providing a spatial area rich in information for the Depthwise convolution. On the expanded high-dimensional feature map, a $3 * 3$ Depthwise convolution is applied to enable it to extract rich and discriminative spatial features. Again, a $1 * 1$ convolution is used to project the high-dimensional feature map back to a low-dimensional space, making the number of output channels match that of the input channels; finally, a residual connection is made between the original input and the output.

The core innovation of the feature extraction module is the introduction of the "Channel Split" operation, which divides the input c channels into two branches, one with $c^{\wedge}$ channels and the other with $c - c^{\wedge}$ channels. In this paper, $c^{\wedge} = \frac{c}{2}$ is set, and one of the branches is used as an identity mapping and directly undergoes the final concatenation operation. The "Channel Split" operation, as a substitute for the ungrouped operation, significantly reduces the computational load.

At the end of the feature extraction module, the output of the convolutional branch and the output of the identity mapping branch are concatenated to achieve the same number of input and output channels. To facilitate information exchange between the two branches, a channel shuffle operation is applied after concatenation to ensure that the features of the two branches can be fully integrated, providing rich input for the next layer.

For the downsampling module, it removes the channel splitting operation on the basis of the feature extraction module, and replaces the original identity mapping branch with a $3 * 3$ Depthwise convolution with a stride of 2, followed by a $1 * 1$ convolution to achieve downsampling. Finally, we concatenate the two branches to make the number of output channels twice that of the input channels.

In the ShuffleNetv2 network, the input and output channel numbers of the feature extraction module are the same, meeting the criterion of balanced convolution to reduce memory access costs; channel splitting operations replace group convolution, reducing the use of group convolution; the feature extraction module has only two branches, with a simple structure and high parallelism; element-wise operations are restricted to one branch, reducing the use of element-wise operations. As can be seen from the above, the ShuffleNetv2 network meets the four efficiency criteria and achieves the goal of being lightweight.

In this study, the ShuffleNetv2 network is used as the backbone network to process the input infrared images as the target detection layer, generating feature maps of three scales. These feature maps are then passed into the neck network to be fused with features of different scales.

3.2. Feature Extraction Module Improvement: C3k2-MaSA Module

In the YOLO11 object detection network, C3k2 is adopted as its feature extraction module to enhance the overall feature expression ability and detection performance of the model. Although the C3k2 module has its unique advantages in feature representation, it does not have a good solution for the problem of unclear targets and large background noise interference in infrared pseudo-color images, and lacks the ability to capture complex and multi-scale features of the target. To solve this problem, a MaSA mechanism will be added on the basis of the C3k2 module to enhance the feature extraction ability of the target.

Vision Transformer (ViT) [31] is a computer vision architecture with strong modeling capabilities. Its core mechanism is the self-attention mechanism. However, it has two limitations: it lacks spatial prior information and the computational cost is proportional to the square of the number of tokens, which is extremely high. To address these issues, Q. Fan et al. [28] combined a distance-dependent time decay matrix from the Retentive Network (RetNet) [32] in the field of natural language processing and extended this time decay matrix to the spatial domain. They proposed the Retentive Networks Meet Vision Transformers (RMT) network. The core of the network is a two-dimensional bidirectional spatial decay matrix based on the Manhattan distance between tokens, providing different attention levels for tokens of different distances. This attention mechanism is named Manhattan Self-Attention (MaSA). At the same time, to solve the huge computational cost brought by the self-attention mechanism, the authors proposed a method for image decomposition along the two axes of the image, enabling MaSA to model global information without compromising the spatial decay matrix. It can be seen that the structure of the MaSA module mainly focuses on establishing and decomposing the attention calculation based on the spatial decay matrix.

For the establishment of the spatial decay matrix, the author starts from the retention mechanism in the RetNet network and transforms the unidirectional one-dimensional time decay into a bidirectional two-dimensional time decay. The parallel and loop forms of the retention mechanism address the high inference cost and computational complexity issues caused by the reliance on the self-attention mechanism in traditional Transformer models, introducing a decay mechanism to the Transformer model.

In the RetNet network, a unidirectional retention mechanism is adopted, where each token only focuses on the previous token. This approach is not suitable for the field of image recognition. Therefore, the authors extended the unidirectional form to a bidirectional form, which is expressed as

$$\begin{cases} BiRetention(X) = (QK^T \odot D^{Bi})V \\ D^{Bi}_{nm} = \gamma^{|n-m|} \end{cases} \tag{1}$$

In the formula, X represents the input sequence. In visual tasks, it denotes the embeddings of a series of image patches; QK^T is the standard attention score matrix, where each element $(QK^T)_{nm}$ represents the dot product of the nth query vector Q and the mth key vector K; V represents the value matrix, which contains the actual information content at each position; D^{Bi} is the bidirectional attenuation matrix, which is the core of the entire mechanism; D^{Bi}_{nm} is an element in matrix D^{Bi}; γ is its attenuation factor, a pre-defined hyperparameter that controls the rate at which attention decays with distance; $|n-m|$ represents the absolute distance between the nth and mth positions in the sequence.

It can be seen from the formula that after the standard attention calculation is completed, the attention scores are decayed according to the different distances between tokens. When the two positions are very close, that is, when $|n-m|$ is very small, $\gamma^{|n-m|} \approx 1$. This means that the attention scores between tokens at close distances will hardly be attenuated, and the model can freely focus on them. When the two positions are far apart, $|n-m|$ is large and $\gamma^{|n-m|}$ becomes very small, which means that the attention scores between tokens at a long distance will be severely attenuated.

In the RetNet network, the focus is on the one-dimensional aspect of text processing and is not applicable to two-dimensional image processing. Based on the Bi-Retention bidirectional modeling, the author has extended the one-dimensional attenuation mechanism to two dimensions. Let the position of the nth token be (x_n, y_n) and then the Manhattan distance matrix between the corresponding tokens is expressed as

$$\begin{cases} MaSA(X) = (Softmax(QK^T) \odot D^{2d})V \\ D^{2d}_{nm} = \gamma^{|x_n - x_m| + |y_n - y_m|} \end{cases} \tag{2}$$

In the formula, $Softmax(QK^T)$ is the probability distribution obtained by normalizing the standard attention score matrix through the Softmax function; D^{2d} represents the two-dimensional spatial decay matrix, and D^{2d}_{nm} is an element in D^{2d}, which is expanded into two-dimensional space based on the element and uses the Manhattan distance $|x_n - x_m| + |y_n - y_m|$ to represent the distance between two tokens. As can be seen from the formula, when the distance between two tokens is greater, the Manhattan distance is larger and the value of D^{2d}_{nm} is smaller, which means that the attention weight of tokens far away will be attenuated. In this way, the model pays more attention to tokens that are spatially adjacent, thereby introducing explicit spatial priors.

To address the issue of high computational cost in self-attention modeling due to a large number of tokens in the early layers of the complete MaSA, the author introduces a simple decomposition method. This method requires calculating the attention scores in both the horizontal and vertical directions separately. First, vertical attention is computed to obtain the aggregated information for each column; then, horizontal attention is calculated to obtain the aggregated information for each row; finally, the two matrices are multiplied to obtain the global information, which is expressed as

$$\begin{cases} Attn_H = Softmax(Q_H K_H^T) \odot D^H \\ Attn_W = Softmax(Q_H K_W^T) \odot D^W \\ MaSA = Attn_H (Attn_W V)^T \end{cases} \tag{3}$$

The one-dimensional attenuation matrices are $D_{nm}^H = \gamma^{|y_n - y_m|}$ and $D_{nm}^W = \gamma^{|x_n - x_m|}$ respectively, representing the horizontal and vertical directions.

Figure 3 shows the calculation method of the receptive field for each token and the final size of the receptive field.

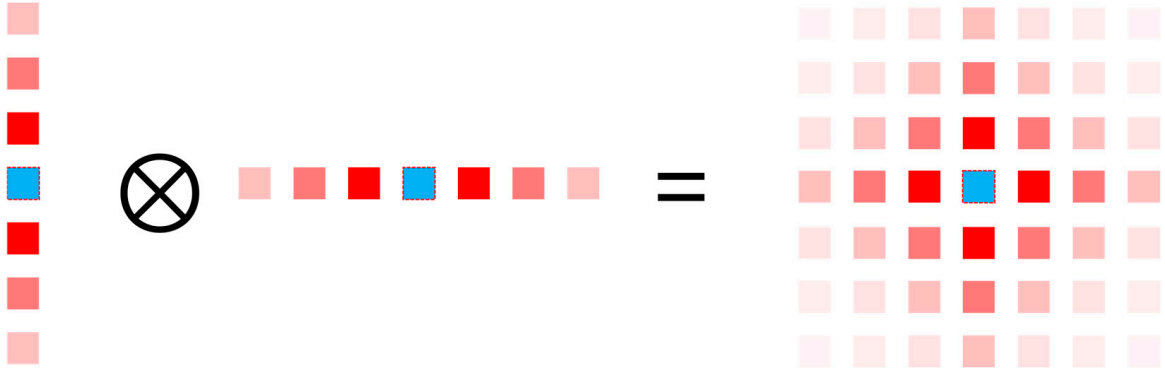


Figure 3. The structure of the Decomposed Manhattan Self-Attention.

Among them, the left matrix in the matrix multiplication represents the attention weights $Attn_H$ calculated along the vertical direction, while the right matrix represents the attention weights $Attn_M$ calculated along the horizontal direction. Through matrix multiplication, the final matrix $MaSA$ is obtained, which represents the receptive field shape of each token. Different deep red blocks represent different distances from the center and thus different attention magnitudes.

By adjusting the magnitudes of the attenuation factors γ in the spatial attenuation matrices D^H and D^W, as well as the weights of the linear layers, the model can adaptively learn the degree of spatial attenuation based on the Manhattan distance between tokens. In this way, the feature extraction capability of the model is enhanced.

By introducing the MaSA mechanism, the C3k2-MaSA module enhances the model's ability to focus on insulator strings and suppress background noise. Without introducing the quadratic complexity of self-attention, it captures the local details and global spatial dependencies of the target, improving the model's detection capability for low-contrast infrared insulator string targets. The C3k2-MaSA module in the neck network receives the feature maps output from the previous layer, extracts features with stronger target representation and sends them to the feature fusion network in the next layer.

3.3. Feature Fusion Layer Improvement: Bi-FPN Module

In the YOLOv11 network, there is a feature fusion layer. It combines the Path Aggregation Network (PANet) and FPN together to form a bidirectional fusion structure called PAFPN. This structure can solve the problem of losing shallow information in the feature extraction network. However, the ability of PAFPN to handle multi-scale information and fuse them is insufficient. This limits the model's adaptability in complex environments. To solve this problem and make the model more lightweight and improve deployment efficiency, this study replaced the original PAFPN with BiFPN for multi-scale feature fusion.

The EfficientDet network was proposed by M. Tan et al. [29]. It is a lightweight network model. The authors of this model introduced a weighted bidirectional feature pyramid network (BiFPN) to achieve multi-scale feature fusion. It adopts the efficient

bidirectional cross-scale connection and weighted feature fusion ideas, which solves the problem that the PANet network requires a large number of parameters and computational resources, and improves the model efficiency.

As shown in Figure 4, due to the insufficient ability of single input nodes in the feature network to fuse different features, the Bi-FPN eliminates the nodes with only one input in the PANet network; then, the author adds a new edge between the input and output nodes of the intermediate layer to fuse multiple features without adding more computational costs; finally, unlike the PANet network which only contains two top-down and bottom-up feature enhancement paths, the author achieves multiple repetitions of the same layer through cross-layer connections, treating each bidirectional path as a feature network to achieve more advanced feature fusion.

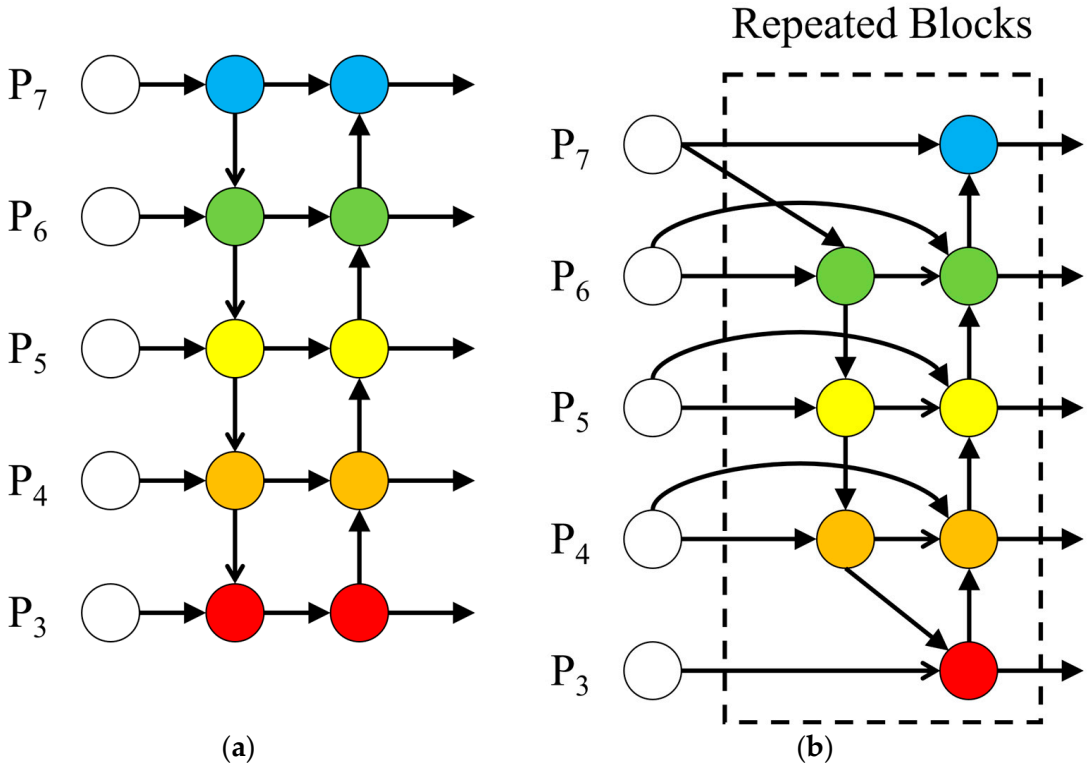


Figure 4. The structure of the PANet and the Bi-FPN. (**a**) The structure of the PANet. (**b**) The structure of the Bi-FPN.

When traditional networks fuse features of different resolutions, they often adjust them to the same resolution and then fuse them in the same way. In the past, each input feature was treated equally. However, input features with different resolutions usually contribute differently to the output features. To address this issue, the authors add different weights before each input feature and use a fast normalization-based fusion method. The formula is expressed as

$$O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} \cdot I_i \tag{4}$$

In the formula, O represents the output feature map, and I_i represents the ith input feature map; is a learnable weight, and each w_i uses the ReLU activation function to ensure $w_i > 0$; the minimum value $\varepsilon = 0.0001$ is used to avoid numerical instability. The entire normalized weight range is between 0 and 1.

The feature fusion of each layer is expressed as

$$\begin{cases} P_6^{td} = Conv(\frac{w_1 \cdot P_6^{in} + w_2 \cdot \text{Resize}(P_7^{in})}{w_1 + w_2 + \varepsilon}) \\ P_6^{out} = Conv(\frac{w_1' \cdot P_6^{in} + w_2' \cdot P_6^{td} + w_3' \cdot Resize(P_5^{out})}{w_1' + w_2' + w_3' + \varepsilon}) \\ P_5^{td} = Conv(\frac{w_4 \cdot P_5^{in} + w_5 \cdot Resize(P_6^{td})}{w_4 + w_5 + \varepsilon}) \\ P_5^{out} = Conv(\frac{w_4' \cdot P_5^{in} + w_5' \cdot P_5^{td} + w_6' \cdot Resize(P_4^{out})}{w_4' + w_5' + w_6' + \varepsilon}) \\ P_4^{td} = Conv(\frac{w_7 \cdot P_4^{in} + w_8 \cdot Resize(P_5^{td})}{w_7 + w_8 + \varepsilon}) \\ P_4^{out} = Conv(\frac{w_7' \cdot P_4^{in} + w_8' \cdot P_4^{td} + w_9' \cdot Resize(P_3^{out})}{w_7' + w_8' + w_9' + \varepsilon}) \end{cases} \quad (5)$$

In the formula, P_4^{td}, P_5^{td} and P_6^{td} are the middle features of the 4th, 5th and 6th layers respectively on the top-down path, while P_4^{out}, P_5^{out} and P_6^{out} are the output features of the 4th, 5th and 6th layers respectively on the bottom-up path, and w is the corresponding weight. $Resize(\cdot)$ represents the size changes of the feature map and aligns with the target size.

In this study, the feature fusion layer is replaced with the Bi-FPN, achieving a more efficient multi-scale fusion operation. Its efficient bidirectional cross-scale connection method balances the contributions of shallow and deep features to the network, playing a crucial role in detecting insulator strings of different scales and distances in infrared images. At the same time, its efficient weighted fusion method helps to maintain the reduction in computational costs.

4. Simulations

4.1. Hardware Environment

The hardware environment used in this experiment is as follows: the CPU is Intel Core i7-12700, the GPU is NVIDIA RTX 4090, and 16GB of memory is utilized. The software environment is as follows: the operating system is Ubuntu 24.04 LTS, CUDA version 11.8, PyTorch version 2.7.1, and Python version 3.11.9.

The training parameters used in this experiment are all set to the initial values, including Image Size of 640, Batch Size of 16, Momentum of 0.937, Weight Decay of 0.0005, Learn Rate of 0.01, and Epoch of 500.

4.2. Construction of the Detection Dataset

A total of 500 infrared insulator string images of transmission lines were collected from public platforms and private centralized sources. Meanwhile, some original visible light images are introduced for comparison. To enhance the diversity of the experiment and meet the training volume requirements of the model, the dataset was expanded to 6500 images through rotation, scaling, and data augmentation. The dataset was divided into a training set, a validation set, and a test set in a ratio of 8:1:1 (5200 images for the training set, 650 images for the validation set, and 650 images for the test set).

The original infrared images were collected in a limited scene, resulting in a relatively narrow distribution of insulator string orientations, scales and directions. Training the model directly on a small dataset would lead to overfitting and poor generalization ability. Therefore, we mainly used techniques such as rotation, scaling and the introduction of simulated environmental noise for data augmentation. These transformations simulate the changes in the actual scene, enriching the training distribution. Through data regularization methods, overfitting can be reduced without saturating the model.

4.3. Simulations Evaluation Indicators

To objectively evaluate the performance of the IISD-YOLO algorithm for infrared insulator string detection based on UAV in this study, this paper adopts the mean average precision (mAP) as the main evaluation index, whose calculation method integrates two evaluation indicators: precision (P) and recall rate (R) [33]. The formulas are expressed as

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$P = \frac{TP}{TP + FP} \tag{7}$$

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

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

In the formula, TP represents the positive samples of the positive class, indicating the samples that the model correctly predicted as the target; FP represents the negative samples of the positive class, indicating the non-target samples that the model wrongly predicted as the target; FN represents the positive samples of the negative class, indicating the target samples that were not recognized by the model. Average Precision (AP) is calculated by integrating the precision curve at different recall rates; mAP is the average value of AP for all categories. In this experiment, the mAP value is equal to the AP value.

To evaluate the hardware consumption and the computational cost required for inference of the model, this paper also introduces the model parameter count (Params) and computational complexity (GFLOPs) as auxiliary indicators.

4.4. Comparison

This experiment uses the same dataset and identical training parameters to compare the IISD-YOLO model proposed in this study with existing detection methods. The comparative results are shown in Table 2.

Table 2. Results of comparison.

Name	P	R	mAP@50	mAP@50:95	Params	GFLOPs
YOLOv5	0.869	**0.771**	0.767	0.427	**1.76**	**4.1**
YOLOv8	0.840	0.740	0.751	0.407	3.01	8.1
YOLOv10	0.851	0.766	0.764	0.414	2.27	6.5
YOLO-CIR	0.845	**0.777**	0.783	0.431	3.59	9.2
FA-YOLO	0.809	0.670	0.721	0.384	2.42	6.8
YOFIR	0.857	0.767	0.762	0.412	9.03	21.7
RT-DETR	**0.872**	0.768	**0.801**	**0.459**	19.88	56.9
IISD-YOLO	**0.870**	0.726	**0.812**	**0.478**	**1.48**	**4.3**

Bolded data indicates the best or second-best data.

As shown in Table 2, different detection models are compared and analyzed based on several evaluation metrics including mAP@50, mAP@50:95, Params, and GFLOPs. The IISD-YOLO model performs well in multiple metrics.

In terms of detection accuracy, the mAP@50 of IISD-YOLO reaches 0.812 and the mAP@50:95 reaches 0.478, both outperforming models such as YOLOv5, YOLOv8, YOLOv10, YOLO-CIR, FA-YOLO, YOFIR, and RT-DETR. This indicates that the model can accurately identify targets and maintains good consistency under different detection stringencies.

In terms of computational expense, IISD-YOLO has 1.48 million Params and 4.3 G GFLOPs, which is significantly lower than that of the RT-DETR model and also outperforms other YOLO models and YOLO-based improved models. This indicates that the IISD-YOLO model is suitable for deployment on devices with limited resources, achieving a balance between model performance and efficiency.

Through in-depth analysis and comparison of the comparative results, it can be observed that the IISD-YOLO model has significant advantages in deployment on edge devices such as drones. This is because neither the YOLO series algorithms nor other algorithms have been optimized for deployment on devices with limited computing resources. In this context, the IISD-YOLO model, without significantly compromising accuracy, replaces the feature detection layer in the YOLOv11 model with the lightweight ShuffleNetv2 network, reducing the number of Params and computational GFLOPs of the model. In addition, compared with the traditional feature extraction module, the IISD-YOLO model proposes the C3k2-MaSA module, which adds a MaSA mechanism on the basis of the original C3k2 module. It introduces different attention sizes for targets at different distances, optimizing the model's attention to less distinct infrared targets and reducing the interference of background noise in detection, thereby enhancing the overall detection capability. Finally, this study introduces the Bi-FPN module as the feature fusion layer of the model, addressing the issue of the lack of fusion capability for information at different scales in different layers in the YOLO series models. This not only enhances the model's adaptability but also improves the fusion efficiency.

In conclusion, the IISD-YOLO model achieves a reduction in computational resource consumption while maintaining accuracy, and is better able to balance the relationship between resource consumption and detection accuracy, thus demonstrating superior performance and practicality.

4.5. Results of Ablation Study

To verify the influence of each module on the entire network, an ablation study was conducted on the YOLOv11 model for each module in this paper. The results of the ablation study are shown in Table 3.

Table 3. Results of ablation study.

YOLOv11	ShuffleNetv2	C3k2-MaSA	Bi-FPN	P	R	mAP@50	mAP@50:95	Params	GFLOPs
√				**0.887**	0.681	0.777	0.413	2.58	6.3
√	√			0.808	0.694	0.740	0.379	1.71	**4.1**
√		√		0.865	0.726	0.794	0.428	2.58	6.4
√			√	**0.911**	0704	0.806	0.454	2.31	6.3
√	√	√		0.852	0.730	0.781	0.426	1.83	4.6
√	√		√	0.839	**0.738**	0.770	0.413	**1.42**	**3.8**
√		√	√	0.854	**0.729**	**0.824**	**0.484**	2.61	6.6
√	√	√	√	0.870	0.726	**0.812**	**0.478**	**1.48**	4.3

Bolded data indicates the best or second-best data. √ indicates whether to introduce the corresponding module.

Table 3 presents the results of the ablation study conducted using the YOLOv11 model as the base model, gradually replacing different modules. By replacing the feature detection layer with the ShuffleNetv2 network, the overall computational cost of the model has significantly decreased. The Params have dropped from 2.58 to 1.71, GFLOPs have decreased from 6.3 to 3.1, and the detection accuracy has also decreased to some extent. By adding the C3k2-MaSA object detection module, the infrared target detection capability has been enhanced, with mAP@50 increasing to 0.794. Meanwhile, Params and GFLOPs have only increased by 0.12 and 0.5 respectively. The introduction of the feature fusion

module Bi-FPN has enhanced the model's detection capability while reducing the number of parameters. The mAP@50 has increased to 0.812, and Params and GFLOPs have decreased to 1.48 and 4.3 respectively. We found that when both the C3k2-MaSA module and the Bi-FPN module are replaced simultaneously without changing the backbone network, the mAP@50 of the model increases by 4.7 percentage points, but it also brings higher parameters and computational complexity. The Params and GFLOPs have risen to 2.61 and 6.6 respectively. Only when all three modules are added can the model's mAP@50 be reduced while ensuring the improvement in detection accuracy, meeting the requirements of lightweight design for deployment on unmanned aerial vehicles with limited onboard processing capabilities.

In conclusion, the proposed improvements have all contributed to the performance of the IISD-YOLO model, enabling the IISD-YOLO model to achieve excellent performance in detecting insulator strings in infrared images.

4.6. Example of Test Results

To evaluate the generalization ability of the IISD-YOLO model on different datasets, we conducted experiments using independent datasets. Figure 5 shows some of the detection results.

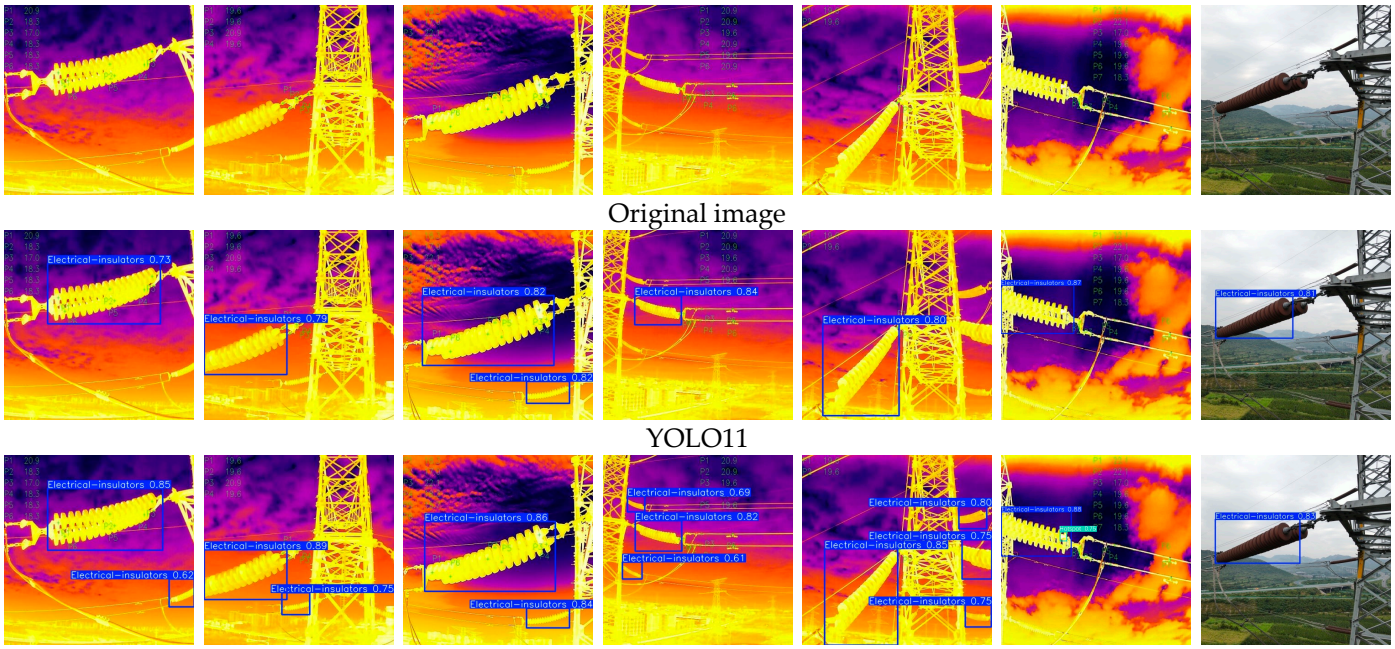


Figure 5. Dataset detection effect display.

Different datasets are used for testing. The IISD-YOLO model performed quite well. Compared with YOLOv11, it has stronger detection capabilities, no missed detections, and no false detections. However, YOLOv11 performs poorly in detecting insulator strings at long distances. Additionally, if there are hotspots on the insulator strings, IISD-YOLO can still detect a portion of them. With the same data, YOLOv11 fails to detect these hotspots. In the visible light images, both models can accurately locate the targets. From this perspective, the performance of IISD-YOLO is better under infrared conditions. Overall, in infrared images, this model can accurately locate the insulator string targets, has a certain detection ability for hotspots, and has high robustness and high accuracy.

5. Conclusions

This study addresses the problem of infrared insulator string detection in UAV transmission line inspection, proposing a network-based IISD-YOLO for infrared insulator string detection. This model achieves superior detection results by integrating a lightweight backbone network, ShuffleNetv2, C3k2-MaSA feature extraction module with Manhattan self-attention, and an efficient multi-scale feature fusion network, Bi-FPN, on the basis of YOLOv11.

Compared with the benchmark model YOLOv11, this model has an increase of 3.5 percentage points in mAP@50 and 6.5 percentage points in mAP@50:95; meanwhile, the parameter size has decreased by 1.1 and the computational cost has reduced by 2. Compared with existing advanced methods such as YOLOv5, YOLOv8, YOLOv10, YOLO-CIR, FA-YOLO, YOFIR, and RT-DETR, IISD-YOLO has the highest mAP@50 and mAP@50:95, which are 0.812 and 0.478 respectively, and the computational cost is very low.

Although the results of this study are quite excellent, there are still some limitations. Although our dataset has been expanded, it still fails to fully simulate adverse weather conditions, and the robustness of the model has not been verified; secondly, the current model design only focuses on insulator strings and fails to simultaneously detect other key transmission line components. Future work will continue to collect and annotate infrared images under different weather conditions and heat loads to improve the model's generalization ability; at the same time, for the features of various power line components, the detection ability of the model will continue to be optimized.

Author Contributions: C.-H.Z.: conceptualization, data curation, formal analysis, software, validation, writing—original draft. Y.-F.R.: conceptualization, supervision, writing—review, methodology and editing. L.-K.C.: data curation, investigation. H.-Y.W.: project administration, data curation, investigation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to privacy and ethical restrictions, the research data will not be made public.

Acknowledgments: During the preparation of this manuscript, the author used Qwen 3.5 for the purposes of English polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
MaSA	Manhattan Self-Attention
Bi-FPN	Bidirectional Feature Pyramid Network
DCNN	Deep Convolutional Neural Network
ROI	Region Of Interest
R-CNN	Region-based Convolutional Neural Network
YOLO	You Only Look Once
PAFPN	Path Aggregation Feature Pyramid Network

CSP	Cross-Stage Partial
FPN	Feature Pyramid Network
PAN	Path Aggregation Network
DFL	Distribution Focal Loss
NMS	Non-Maximum Suppression
PSA	Pointwise Spatial Attention
dilated CBAM	dilated Convolution Block Attention Module
EMSGC	Efficient Multi-Scale Group Conv
ViT	Vision Transformer
RetNet	Retentive Network
RMT	Retentive Networks Meet Vision Transformers
mAP	mean Average Precision
P	Precision
R	Recall rate
AP	Average Precision

References

1. Ayamga, M.; Akaba, S.; Nyaaba, A.A. Multifaceted applicability of drones: A review. *Technol. Forecast. Soc. Change* **2021**, *167*, 120677. [CrossRef]
2. Wu, G.; Yu, M.; Shi, W.; Li, S.; Bao, J. Image recognition in online monitoring of power equipment. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881419900836. [CrossRef]
3. Han, S.; Yang, F.; Yang, G.; Gao, B.; Zhang, N.; Wang, D. Electrical equipment identification in infrared images based on ROI-selected CNN method. *Electr. Power Syst. Res.* **2020**, *188*, 106534. [CrossRef]
4. Cong, S.; Pu, H.; Yao, F. Review on application of infrared detection technology in state detection of electrical equipment. In *Proceedings of the 16th Annual Conference of China Electrotechnical Society: Volume II*; Springer Nature: Singapore, 2022; pp. 1254–1261.
5. Jeong, S.; Kim, M.G.; Kim, J.H.; Oh, K.Y. Thermal monitoring of live-line power transmission lines with an infrared camera mounted on an unmanned aerial vehicle. *Struct. Health Monit.* **2023**, *22*, 3707–3722. [CrossRef]
6. Yuan, X.; Zheng, Z.; Li, Y.; Liu, X.; Liu, L.; Li, X.; Hou, Q.; Cheng, M.-M. Strip R-CNN: Large strip convolution for remote sensing object detection. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 12259–12267. [CrossRef]
7. Su, X.; Wang, J.; Wang, Y.; Zhang, D. Steel roll eye pose detection based on binocular vision and mask R-CNN. *Sensors* **2025**, *25*, 1805. [CrossRef]
8. Sagar, A.S.; Chen, Y.; Xie, Y.; Kim, H.S. MSA R-CNN: A comprehensive approach to remote sensing object detection and scene understanding. *Expert Syst. Appl.* **2024**, *241*, 122788. [CrossRef]
9. Sharifuzzaman, S.A.; Tanveer, J.; Chen, Y.; Chan, J.H.; Kim, H.S.; Kallu, K.D.; Ahmed, S. Bayes R-CNN: An uncertainty-aware Bayesian approach to object detection in remote sensing imagery for enhanced scene interpretation. *Remote Sens.* **2024**, *16*, 2405. [CrossRef]
10. Bai, T.; Luo, J.; Zhou, S.; Lu, Y.; Wang, Y. Vehicle-type recognition method for images based on improved faster R-CNN model. *Sensors* **2024**, *24*, 2650. [CrossRef]
11. Yu, L.; Feng, X.; Zeng, Y.; Guo, W.; Yang, X.; Zhang, X.; Tan, Y.; Sun, C.; Lu, X.; Sun, H. MDB-YOLO: A Lightweight, Multi-Dimensional Bionic YOLO for Real-Time Detection of Incomplete Taro Peeling. *Electronics* **2025**, *15*, 97. [CrossRef]
12. Nguyen, P.T.; Huynh, D.C.; Ho, L.D.; Tran, H.T.; Dunnigan, M.W. InsDD-YOLO: Detection of Transmission Line Insulator Damage Based on the Improved YOLOv13 Model. *IEEE Access* **2025**, *13*, 216596–216610. [CrossRef]
13. Ji, H.; Chen, X.; Bai, J.; Gong, C. Lightweight SCL-YOLOv8: A High-Performance Model for Transmission Line Foreign Object Detection. *Sensors* **2025**, *25*, 5147. [CrossRef]
14. Maksimović, V.; Jakšić, B. Automatic Pollution Detection on High-Voltage Isolators Using a Two-Phase Approach. *ICCK Trans. Electr. Power Netw. Syst.* **2026**, *2*, 7–21. [CrossRef]
15. Souza, B.J.; Stefenon, S.F.; Singh, G.; Freire, R.Z. Hybrid-YOLO for classification of insulators defects in transmission lines based on UAV. *Int. J. Electr. Power Energy Syst.* **2023**, *148*, 108982. [CrossRef]
16. Liu, M.; Kari, T.; Yimamu, A.; Zhou, Y.; Ma, X. MC-YOLOv8: A Hybrid Attention and Adaptive Convolution Model for Power Equipment Fault Inspection. *Signal Image Video Process.* **2025**, *19*, 653. [CrossRef]
17. Liu, T.; Li, G.; Gao, Y. Fault diagnosis method of substation equipment based on You Only Look Once algorithm and infrared imaging. *Energy Rep.* **2022**, *8*, 171–180. [CrossRef]
18. Gao, X.; Du, J.; Liu, X.; Jia, D.; Wang, J. Object Detection Based on Improved YOLOv10 for Electrical Equipment Image Classification. *Processes* **2025**, *13*, 529. [CrossRef]

19. JOCHERG. Yolov5[EB/OL]. [2020-08-10]. Available online: https://github.com/ultralytics/yolov5 (accessed on 15 November 2025).
20. JOCHERG. Yolov8[EB/OL]. [2023-01-10]. Available online: https://github.com/ultralytics/ultralytics (accessed on 15 November 2025).
21. 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.
22. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
23. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
24. Zhou, J.; Zhang, B.; Yuan, X.; Lian, C.; Ji, L.; Zhang, Q.; Yue, J. YOLO-CIR: The network based on YOLO and ConvNeXt for infrared object detection. *Infrared Phys. Technol.* **2023**, *131*, 104703. [CrossRef]
25. Du, S.; Zhang, B.; Zhang, P.; Xiang, P.; Xue, H. FA-YOLO: An Improved YOLO Model for Infrared Occlusion Object Detection under Confusing Background. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1896029. [CrossRef]
26. Wen, M.; Li, C.; Xue, Y.; Xu, M.; Xi, Z.; Qiu, W. YOFIR: High precise infrared object detection algorithm based on YOLO and FasterNet. *Infrared Phys. Technol.* **2025**, *144*, 105627. [CrossRef]
27. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
28. Fan, Q.; Huang, H.; Chen, M.; Liu, H.; He, R. Rmt: Retentive networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5641–5651.
29. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
30. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
31. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Sun, Y.; Dong, L.; Huang, S.; Ma, S.; Xia, Y.; Xue, J.; Wang, J.; Wei, F. Retentive network: A successor to transformer for large language models. *arXiv* **2023**, arXiv:2307.08621. [CrossRef]
33. Allen, K.; Berry, M.M.; Luehrs, F.U., Jr.; Perry, J.W. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *Am. Doc. (Pre-1986)* **1955**, *6*, 93.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.