EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Object Detection and Segmentation of Power Equipment in Infrared Images via Improved YOLOv8 and Prompt-Optimized SAM

Bing Xue[1,*], Zehui Liu[1], Zhanhong Wang[1], Wenyuan Zhou[1], Baoning Wang[1] and Xukun Yang[1]

[1]State Grid Shaanxi Electric Power Company Limited Weinan Power Supply Company, 110 Qianjin Road, Linwei District, Weinan City, Shaanxi Province, 714000, China

Abstract

To achieve automated infrared monitoring of power equipment in substations, this paper proposes an object detection and segmentation method based on improved YOLOv8 and Prompt-Optimized SAM (Segment Anything Model). Firstly, to address the issues of poor resolution and strong background interference in infrared images, the small object feature extraction capability and bounding box regression accuracy of YOLOv8 are improved by introducing a multi-scale feature extraction module, a robust feature downsampling module, and an improved loss function. The Spatial Pyramid Pooling Fast module is improved using large-kernel depthwise separable convolution, enhancing the extraction capability for both global and local features. Secondly, to improve segmentation accuracy, this paper proposes a method that converts detection boxes into prompt points. GrabCut, combined with colour saliency and a superpixel algorithm, is used to segment high-confidence target regions. Zero-shot prompt point segmentation for SAM is achieved by performing clustering on the regions. Experimental validation on an infrared dataset covering seven types of power equipment shows that the improved object detection model achieves an mAP@0.5 of 95.7%, which is 2.3% higher than the original model, with a detection speed of 107.5 FPS. The proposed segmentation method achieves higher accuracy in complex backgrounds than both bounding box-prompted SAM and GrabCut. This study lays a foundation for the precise processing of infrared images of substation power equipment.

Keywords: Infrared Images, Power Equipment, Object Detection, Image Segmentation, YOLOv8.

Received on 28 October 2025, accepted on 11 December 2025, published on 12 March 2026

doi: 10.4108/eetsis.10727

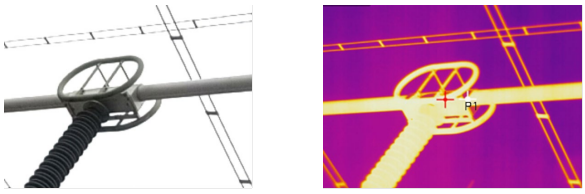
1. Introduction

As a commonly used condition monitoring method in substations, thermographic inspection can reveal temperature anomalies in power equipment from a safe distance based on thermal radiation [1]. Temperature anomalies can be caused by equipment failures, and may also lead to equipment failures. Therefore, during routine inspections, examining infrared images can identify problems in equipment, including insulator aging, bushing insulation defects, etc. Since infrared images themselves only reflect thermal characteristics and have limited colour resolution, manual inspection of infrared images is inefficient, and its accuracy is susceptible to lighting conditions and subjective factors. Compared to manual experience, computer vision-based processing and analysis of infrared images can locate heating areas in equipment more promptly and accurately, and determine whether a fault exists. As crucial steps in image processing, object detection and segmentation enable the critical analysis of operational conditions for various

*Corresponding author. Email: m18961007117@163.com

apparatus. In daily inspection, high-precision equipment detection capability can avoid false detection and missed detection caused by subjective judgment and environmental interference in manual detection [2]. Meanwhile, the efficient segmentation performance provides a clear basis for regional division in equipment condition assessment involving thermal defect analysis [3].

However, the power equipment in the substation can effectively exchange heat with the surrounding environment during normal operation. As shown in Figure 1, which presents visible light and infrared images of a disconnecting switch, the infrared image does not effectively represent other characteristics of the object, such as texture and colour; algorithms suitable for visible light images exhibit reduced performance when processing infrared images. Furthermore, as power equipment in substations is relatively concentrated, with numerous power lines and metal supports present, these backgrounds pose strong interference to object detection and segmentation. These factors collectively increase the difficulty of object detection and segmentation.



(a) (b)

Figure 1. Visible light and infrared images of a disconnecting switch. (a) Visible light image; (b) Infrared image

Traditional infrared image object detection and segmentation methods for power equipment are mainly designed for specific research objects. Manually selected features struggle to cover multiple types of power equipment, and require tuning a large number of parameters. Image segmentation methods primarily achieve segmentation of regions of interest through quantitative analysis of the thermal characteristics of power equipment [4]. In complex environments, quantitative analysis is limited by the quality of infrared images and is prone to over-segmentation.

With the introduction of deep learning technology, its excellent image analysis capability has significantly improved the processing efficiency and accuracy. Ou et al. [5] removes some high-level convolutional layers from Faster R-CNN, achieving fast detection of several power equipment. Zhang et al. [6] employs infrared and visible light fusion combined with YOLOv4 (You Only Look Once version 4) to achieve high-precision object detection in power inspection. The segmentation of power equipment mainly includes semantic segmentation and instance segmentation, and the latter provides accurate contour for each equipment. In [7], a dual-stream encoder is proposed to extract multi-modal features, thereby segmenting more complete power equipment. In [8], a novel instance segmentation framework guided by reconstruction error is proposed to automatically extract insulators from infrared images, with the model achieving high computational speed and accuracy. Although deep learning boosts both the efficiency and the accuracy of automating infrared image analysis during routine inspections, it is still limited by the poor resolution of infrared images. For equipment with complex backgrounds and significant scale variations, object detection and segmentation are prone to failure. Furthermore, when training segmentation algorithms, all training data requires pixel-level annotation.

To address the aforementioned issues, an object detection and segmentation method using improved YOLOv8 with Prompt-Optimized SAM (Segment Anything Model) [9] is proposed. This method is applicable to multiple primary power equipment in substations, and has strong anti-interference ability for infrared images with complex background. The main contributions of this paper are as follows:

- The small object feature extraction capability and bounding box regression accuracy of YOLOv8 are improved by integrating the multi-scale feature extraction module, the robust feature downsampling module, and the Wise-MPDIoU loss function. The SPPF module is improved with large-kernel depthwise separable convolution (LKDC), enhancing the extraction capability for both global and local features;
- We design a prompt optimization method that converts prompt boxes into prompt points and utilizes SAM to accomplish segmentation. After object detection is completed, it combines saliency detection, foreground extraction, and superpixel optimization to initially segment the detection boxes, then uses the cluster centres of the masks as prompts for SAM to accomplish the segmentation;
- Experimental results demonstrate that the improved YOLOv8 has high accuracy for the seven primary categories of substation equipment. When dealing with complex scenes, the segmentation accuracy of Prompt-Optimized SAM is higher than that of SAM prompted with detection boxes and GrabCut.

2. Related Works

2.1. Advances in YOLOv8

YOLO series algorithms take advantage of its end-to-end advantages and have fast detection speed. The accuracy of the algorithm is gradually improved by introducing optimization strategies such as Bag-of-Freebies and architectural reparameterization. As an advanced version of the YOLO series, YOLOv8 adopts a variety of improvement strategies, including anchor-free design and Distribution Focal Loss (DFL) [10], which enables fast object detection in various industrial scenes and have stable performance. However, YOLOv8 still struggles with accurately detecting small and

multi-scale targets at a long distance. In order to solve these problems, many studies have been carried out to improve YOLOv8.

Wu et al. [11] introduce Efficient Multi-Scale Attention into the SPPR module, enhancing the object detection capability under complex backgrounds. Ma et al. [12] replace the original strided convolution module with Space-to-Depth convolution, reducing the degradation of small targets' features during downsampling. Wang et al. [13] developed a high-speed feature processing module using FasterNet, which effectively lowers the rate of missed detections for small objects in UAV-captured images. The above improvements to YOLOv8 mainly focus on preserving more information, enhancing the feature extraction ability, and optimizing the Head component for specific tasks, so as to improve the performance of the algorithm in complex scenes. Although more complex feature extraction will result in a slight increase in algorithm speed, considering the significantly improved accuracy, this is acceptable.

2.2. Prompt-based Segmentation

Segmenting based on prompt information can reduce the computational cost of algorithms blindly searching for targets, and precise prompts can mitigate the degree of over-segmentation or under-segmentation. Benefiting from object detection, image segmentation can utilize detection boxes as prompts. GrabCut, as a traditional algorithm for segmentation based on prompts, demonstrates good segmentation performance on visible light images [14]. GrabCut achieves foreground segmentation by classifying pixels, and performs well when image clarity is high. SAM is an interactive segmentation model proposed by Meta, trained on a large-scale dataset, capable of segmenting based on prompt boxes or prompt points. Dan et al. [15] extracts high-confidence points through local feature matching to prompt SAM for segmenting small infrared targets. Li et al. [16] compares prompt points, prompts boxes, and automatic segmentation, with results indicating that a suitable prompts box can achieve segmentation in agricultural visible light images. Although both GrabCut and SAM can achieve zero-shot segmentation, for infrared images of power equipment, the prompt boxes often include background regions with thermally similar characteristics. When these background regions are adjacent to the foreground, within the prompt boxes, the prompt information for the background and foreground is the same, leading to lower segmentation accuracy. Using prompt points to indicate the foreground can better locate the target and distinguish between the background and foreground.

3. Proposed Method

It can be seen from Figure 2 that the method presented in this research can be separated into two individual stages: object detection using improved YOLOv8 and image segmentation using Prompt-Optimized SAM. This section first introduces the modified components of YOLOv8, then elaborates on the specific procedures of segmentation.

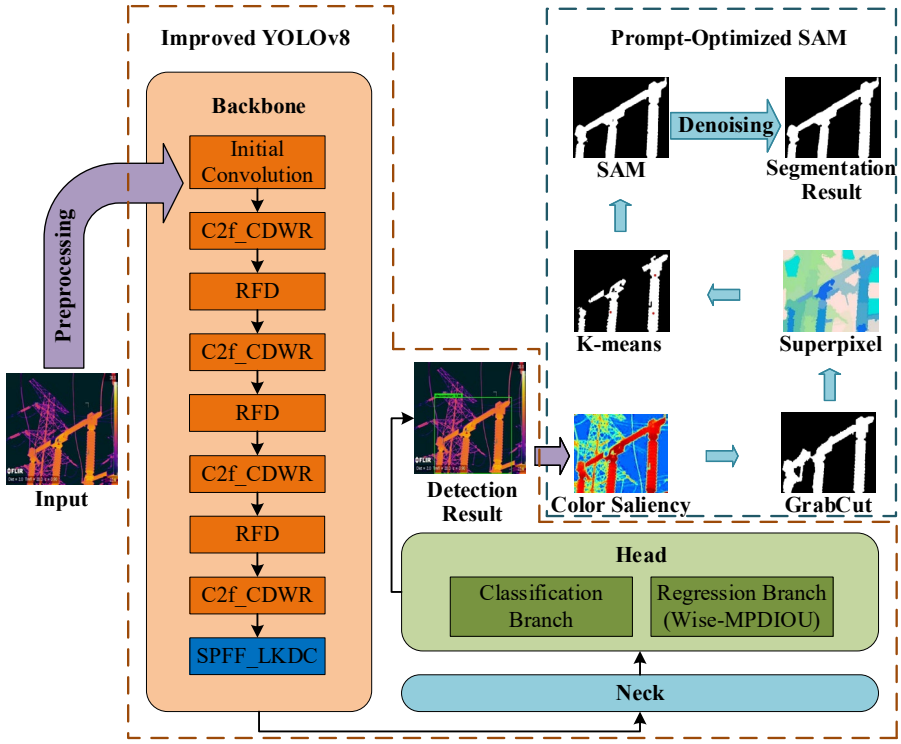


Figure 2. Architecture of the proposed method

3.1. Improved YOLOv8

Multi-Scale Feature Extraction Module

In infrared images of power equipment, target scales vary significantly. For example, the scale difference between a disconnector and a bushing may reach a factor of 10, and small targets are likely to be concealed by background noise because of their low contrast. Although the traditional C2f module extracts scale features through multi-branches, it suffers from insufficient feature fusion and the tendency of small-target features to be diluted in deep networks.

In order to enhance the algorithm's performance, we adopt the C2f_CDWR module [17], which integrates Coordinate Attention (CA) based on the C2f multi-branch framework and enhances multi-scale feature extraction via a synergistic mechanism. Specifically, it retains short-connection branches to transmit low-level details while stacking multiple groups of residual branches: shallow-layer structures concentrate on the fine-grained features of small targets, while the deep layers emphasize the global contours and context information of large targets, thereby remedying the defect of unbalanced scale feature learning. At the output terminals of the residual branches, 1×1 convolution branch weight coefficients are introduced to replace the traditional equalized concatenation and addition methods, making it possible for the model to boost small-target feature weights automatically according to the target scale distribution. For input feature map $X \in \mathbb{R}^{C\times H\times W}$, horizontal and vertical coordinate information is preserved via parallel 1D pooling:

$$z_c^h(h) = \frac{1}{W}\sum_{i=0}^{W-1} x_c(h,i), \quad z_c^w(w) = \frac{1}{H}\sum_{j=0}^{H-1} x_c(j,w). \tag{1}$$

These are concatenated and processed through a shared 1×1 convolution with reduction ratio r:

$$f = \delta\left(\mathrm{F}_1\left(\left[z_h, z_w\right]\right)\right). \tag{2}$$

where δ denotes non-linear activation. The feature map f is split into $f_h \in \mathbb{R}^{C/r\times H}$ and $f_w \in \mathbb{R}^{C/r\times W}$, then transformed to final weight coefficients:

$$g_h = \sigma\left(\mathrm{F}_h\left(f_h\right)\right), \quad g_w = \sigma\left(\mathrm{F}_w\left(f_w\right)\right). \tag{3}$$

where σ is Sigmoid activation. The enhanced feature map is computed as:

$$Y = X \otimes \left(g_h \otimes g_w\right). \tag{4}$$

After feature fusion, the CA module is embedded; through separate pooling (i.e., independent modelling in the height and width directions), attention weights are produced to specially enhance the response of small-target regions and restrain background noise, thereby resolving the issue of insufficient spatial focusing. Figure 3 illustrates the structure of the multi-scale feature extraction module utilized in this research. Experimental tests verify that this module can efficiently boost the capacity for extracting multi-scale features and enhancing the feature reaction of small power equipment targets.

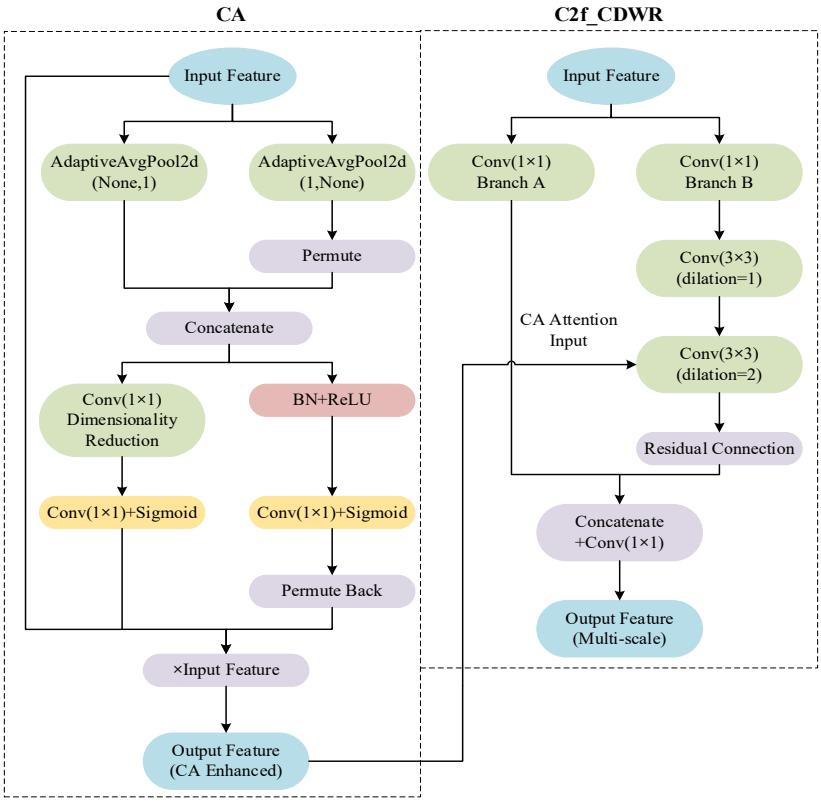


Figure 3. Multi-Scale Feature Extraction Module

Robust Feature Downsampling Module

In the target detection task, feature downsampling serves as a key step for extracting high-level semantic features. Nevertheless, this process often leads to losses of local details and global information, a situation that is detrimental to the detection of power equipment's infrared images. Such images show large changes in the target scale, and local tiny texture features are easily weakened by repeated downsampling. To address the feature loss issue of traditional downsampling, this study introduces a Robust Feature Downsampling (RFD) module [18], which enhances the capability of feature retention through multi-branch fusion.

The central concept of RFD is the complementary fusion of different features extracted by parallel branches, and its structure comprises three key components: by using a 3×3 convolution kernel and setting the stride to 2, the convolution branch conducts downsampling to acquire edge and texture details; the pooling branch conducts downsampling with 2×2 average pooling and a stride of 2 to retain global contours and avoid local noise; the feature fusion layer first processes the output of the pooling branch via channel truncation to match dimensions, it then combines this with the output channels of the convolution branch, and after that adjusts the channel count by means of 1×1 convolution to output a downsampled feature map with consistent dimensions. The structure of the Robust Feature Downsampling Module is illustrated in Figure 4.

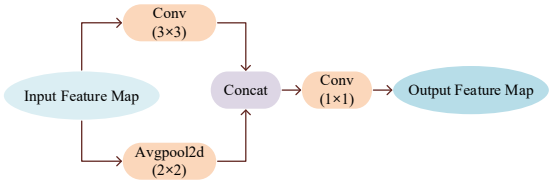


Figure 4. Robust Feature Downsampling Module.

Improved SPPF Module

In object detection tasks, the multi-scale aggregation of high-level features affects the accuracy of object detection. Traditional SPPF (Spatial Pyramid Pooling Fast) exhibits a limited capacity for capturing global features of large-scale targets, thereby restricting its ability to satisfy the cross-scale requirements of infrared detection for power equipment, particularly the coexistence of large-sized equipment and small components, alongside low image contrast. To enhance the algorithm's performance, we propose the integration of the SPPF_LKDC module, which
adopts large-kernel convolution for widening the receptive field and boosting the ability to collaboratively extract global and local features.

The SPPF_LKDC retains the advantage of SPPF's fast pooling and consists of four key components: the Dimensionality Reduction Convolution Layer employs a 1×1 convolution (cv1) for halving the input channels (i.e., c_ = c1//2), reducing computational complexity and compressing redundancy; the Large-Kernel Depthwise Separable Convolution Layer adopts an 11×11 convolution with padding=5 and the number of groups set to c_, expanding the receptive field to capture the global contours of equipment while avoiding parameter explosion. The selection of the 11×11 kernel in the SPPF_LKDC module is predicated on balancing receptive field expansion with computational efficiency. This specific dimension effectively captures the global structural contours of power equipment by enlarging the receptive field, while simultaneously mitigating the computational redundancy and overfitting risks associated with excessively large kernels. The relationship between the receptive field R and the kernel size k is governed by:

$$R_l = R_{l-1} + (k-1) \times \prod_{i=1}^{l-1} s_i . \tag{5}$$

where s_i denotes the stride of the i-th layer. Consequently, the 11×11 convolution within the deep backbone layers yields a receptive field exceeding 200×200 pixels, which is sufficient to encompass the holistic structure of large-scale equipment in infrared imagery.

The Multi-Scale Pooling Layer performs 1, 2, and 3 rounds of max pooling on the output of the large-kernel convolution respectively to acquire local details such as equipment heating spots; the Feature Fusion Layer uses a 1×1 convolution (cv2) to concatenate the original features with the three rounds of pooled features and outputs aggregated features with unified dimensions. Figure 5 depicts the structure of the Improved SPPF Module.

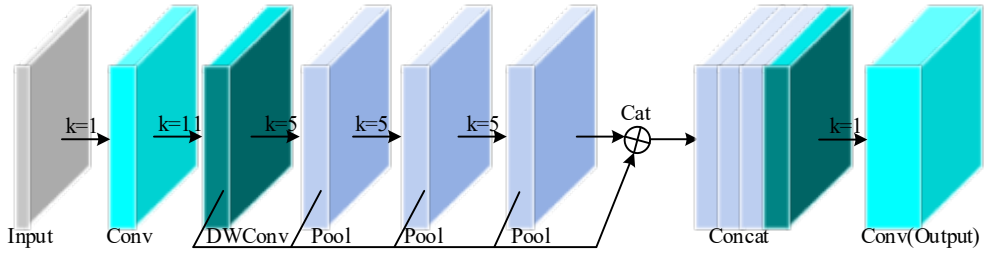


Figure 5. Improved SPPF Module

Improvement of Loss Function

The bounding box regression loss function exerts a direct impact on the model's predictive accuracy regarding the position and shape of targets. Though traditional IoU and its modified variants are able to quantify the overlapping area of the detected box and the ground truth box, they show two limitations for the scenario of power equipment infrared images: first, they exhibit low sensitivity to alignment errors at the corners of bounding boxes, resulting in insufficient edge fitting accuracy for rectangular equipment such as disconnectors and lightning arresters; second, they do not consider the dynamic weight of target centre offset, making it difficult to adapt to the target edge blurriness caused by thermal imaging characteristics in infrared images. For the purpose of solving this problem, the current study suggests introducing the Wise-MPDIoU loss function, which boosts the bounding box regression accuracy for power equipment via the combination of a position penalty term and a dynamic weight mechanism.

The core concept underlying the Wise-MPDIoU loss function involves introducing the bounding box corner position penalty term (MPDIoU) built upon IoU, and dynamically tuning the loss weight via the target centre distance (i.e., the Wise mechanism), in this way making the model concentrate more on hard-to-regress samples during training. The overall loss function is formulated as follows:

$$L_{\text{Wise-MPDIoU}} = 1 - r \cdot \text{MPDIoU}. \tag{6}$$

In the formula, MPDIoU denotes the IoU incorporating corner position penalty, and r denotes the dynamic weight coefficient based on centre distance.

Denote the coordinates of the predicted bounding box as $(p_{x1}, p_{y1}, p_{x2}, p_{y2})$ and those of the ground-truth bounding box as $(t_{x1}, t_{y1}, t_{x2}, t_{y2})$; then the IoU is formulated as:

$$\text{IoU} = \frac{S_{\text{inter}}}{S_{\text{union}} + \varepsilon}. \tag{7}$$

Within this formula, S_{inter} denotes the overlap area of the detected box and the ground-truth box:

$$\begin{aligned} S_{\text{inter}} = {} & \max\left(0, \min\left(p_{x2}, t_{x2}\right) - \max\left(p_{x1}, t_{x1}\right)\right) \times \\ & \max\left(0, \min\left(p_{y2}, t_{y2}\right) - \max\left(p_{y1}, t_{y1}\right)\right). \end{aligned} \tag{8}$$

S_{union} denotes the union area:

$$\begin{aligned} S_{\text{union}} = {} & \left(p_{x2} - p_{x1}\right)\left(p_{y2} - p_{y1}\right) + \\ & \left(t_{x2} - t_{x1}\right)\left(t_{y2} - t_{y1}\right) - S_{\text{inter}}. \end{aligned} \tag{9}$$

ε is a smoothing term to avoid division by zero. To enhance the constraint on the corner alignment of bounding boxes, a corner distance penalty term is defined as follows:

$$\text{MPDIoU} = \text{IoU} - \frac{d_1 + d_2}{c_w^2 + c_h^2 + \varepsilon}. \tag{10}$$

Within this formulation, d_1 and d_2 denote the squared Euclidean distances between the top-left corners as well as the bottom-right corners of the predicted bounding box and the ground-truth bounding box, respectively:

$$\begin{aligned} d_1 &= \left(p_{x1} - t_{x1}\right)^2 + \left(p_{y1} - t_{y1}\right)^2, \\ d_2 &= \left(p_{x2} - t_{x2}\right)^2 + \left(p_{y2} - t_{y2}\right)^2. \end{aligned} \tag{11}$$

c_w and c_h denote the size of the bounding boxes (width and height):

$$\begin{aligned} c_w &= \max\left(p_{x2}, t_{x2}\right) - \min\left(p_{x1}, t_{x1}\right), \\ c_h &= \max\left(p_{y2}, t_{y2}\right) - \min\left(p_{y1}, t_{y1}\right). \end{aligned} \tag{12}$$

To allow the model to prioritize hard-regression samples with substantial centre offsets, the present work proposes a dynamic weight r through leveraging the centre distance between the ground-truth bounding box and the predicted bounding box, which is defined as follows:

$$r = \exp\left(-\frac{\rho}{c_2 + \varepsilon}\right). \tag{13}$$

Within this formulation, ρ represents the squared Euclidean distance between the centre of the predicted bounding box and that of the ground-truth bounding box:

$$\rho = \left(\frac{p_{x1} + p_{x2}}{2} - \frac{t_{x1} + t_{x2}}{2}\right)^2 + \left(\frac{p_{y1} + p_{y2}}{2} - \frac{t_{y1} + t_{y2}}{2}\right)^2. \tag{14}$$

$c_2 = c_w^2 + c_h^2$ denotes the square of the diagonal of the minimum enclosing rectangle, which is used to normalize the centre distance. When the centre offset is large, ρ increases and r approaches 0, resulting in an increased loss value and a strengthened penalty; conversely, r approaches 1, and the loss value is dominated by IoU.

3.2. Prompt-Optimized SAM

Because the pixel information of infrared images differs significantly from that in visible light, in order to ensure the accuracy of SAM, the prompt information needs to be able to fully focus on the foreground subject. As shown in Figure 6, This section will introduce how to use GrabCut to obtain high-confidence foreground, and then use clustering to get the prompt point coordinates for input into SAM. To mitigate the computational burden, the input is constrained to the local region identified by the target detector. Let the input image be denoted as $I : \Omega \to R_3$, where $\Omega \subset Z_2$ is the image domain, and every pixel $(x, y) \in \Omega$ has colour values in the RGB space.

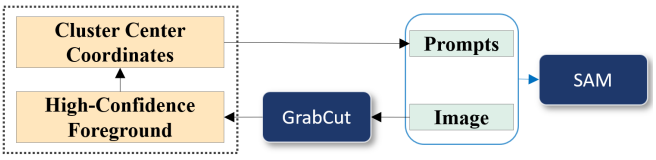


Figure 6. Schematic diagram of GrabCut enhanced with SAM

GrabCut Segmentation Based on Colour Saliency

When the foreground contains a local region that differs in colour from the background but has low saturation and brightness—such as nameplates or status indicators on power equipment—the cluster centre in subsequent clustering may converge within this local area, potentially leading to segmenting the local region. To address this, colour saliency is utilized in this study to initialize the probable foreground and background regions in the GrabCut algorithm. This approach also prevents GrabCut from incorrectly classifying visually inconspicuous background areas near the centre as part of the foreground.

Since regions with higher temperatures are generally brighter, this study performs saliency computation in the HSV colour space, which effectively represents image brightness. Particular emphasis is placed on the saliency of the Value channel in infrared images. In HSV colour space, the saliency weight $I_{\text{saliency}}(x,y)$ for each pixel (x,y) is computed via the formula below:

$$I_{\text{saliency}}(x,y)=\lambda_1\cdot\frac{S(x,y)-S_{\min}}{S_{\max}-S_{\min}}+\lambda_2\cdot\frac{V(x,y)-V_{\min}}{V_{\max}-V_{\min}}. \quad (15)$$

where $S(x,y)$ denotes the saturation channel value of pixel (x,y), and $V(x,y)$ denotes the value of the value channel. $S_{\min}$, $S_{\max}$, $V_{\min}$, and $V_{\max}$ denote the minimum and maximum values of the S and V channels, respectively. λ_1 and λ_2 serve as the saliency scaling coefficients for the S and V channels.

After obtaining the saliency weights of all pixels, the spatial consistency between pixels is enhanced by Gaussian smoothing, and it is determined in the following way:

$$I_{\text{smooth}}(x,y)=G(x,y)*I_{\text{saliency}}(x,y). \quad (16)$$

In which $G(x,y)$ represents a 2D Gaussian kernel function.

The processed pixels are classified as foreground or background, and every pixel is marked as follows:

$$\alpha_i=\begin{cases}0, \text{ if } I_{\text{smooth}}(x,y)>T_{\text{saliency}}\\ 1, \text{ if } I_{\text{smooth}}(x,y)<T_{\text{saliency}} \text{ and } (x,y)\in B.\\ 2, \text{ otherwise}\end{cases} \quad (17)$$

where 0 represents probable foreground, 1 represents definite background, 2 represents probable background, T_{saliency} is the saliency weight threshold, and B refers to the image boundary region, with a boundary width of 5 pixels.

Colour samples are collected from the background and foreground regions to construct Gaussian Mixture Models (GMMs) for both. Each GMM comprises 5 Gaussian components. After initializing the parameters of GMMs, each pixel is linked to the Gaussian component g_i that generates the highest probability for that pixel. g_i is computed via the following formula:

$$g_i=\arg\max_{g\in\{1,\ldots,5\}}[\pi(\alpha_i,g)\cdot N(z_i|\mu(\alpha_i,g),\Sigma(\alpha_i,g))]. \quad (18)$$

where z_i denotes the colour vector, $\pi(\alpha_i,g)$ stands for the weight related to the g-th Gaussian component, $N(z_i|\mu(\alpha_i,g),\Sigma(\alpha_i,g))$ stands for the probability density function belonging to the multivariate Gaussian distribution, $\mu(\alpha_i,g)$ serves as the mean vector of the Gaussian component, while $\Sigma(\alpha_i,g)$ acts as the covariance matrix of that Gaussian component.

After the pixel assignment, the parameters of each Gaussian distribution in GMMs are updated using maximum likelihood estimation. This update enables the model to capture the global distribution pattern of pixels. Subsequently, the probability that a pixel is assigned to the foreground or background class is calculated. The following energy function is minimized via the max-flow/min-cut algorithm:

$$g_i=\arg\max_{g\in\{1,\ldots,5\}}[\pi(\alpha_i,g)\cdot N(z_i|\mu(\alpha_i,g),\Sigma(\alpha_i,g))]. \quad (19)$$

where C denotes the set of neighbouring pixel pairs, and γ, β are the smoothing term parameters.

GrabCut achieves the convergence of the segmentation mask by iterating the GMMs parameters and minimizing the energy function. The final mask is denoted as M_{grab}.

Superpixel Optimization

Compared with ensuring the integrity of foreground segmentation, the subsequent clustering algorithm requires stronger assurance that the masked region contains minimal background. To further reduce background content along the edges of the mask, the SLIC [19] is applied to perform coarse superpixel segmentation on image I, with the number of superpixels denoted as w. The region of M_{grab} is treated as the foreground area, and the foreground ratio of each superpixel region R_s is calculated. Superpixels with a low foreground ratio are labelled as definite background. The label for all pixels contained in the s-th superpixel is determined as follows:

$$\alpha(R_s)=\begin{cases}\text{Keep Original, if } r_s\geq T_{\text{super}}\\ 1, \text{ if } r_s<T_{\text{super}}\end{cases}. \quad (20)$$

where r_s denotes the foreground pixels proportion of the s-th superpixel, and T_{super} is the foreground ratio threshold.

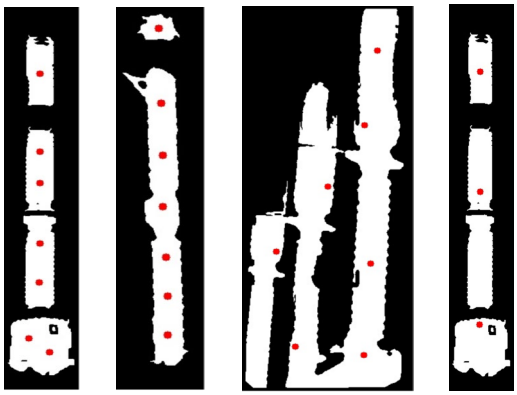
After mask optimization, noise is removed by converting small mask regions (a 5% area threshold) into background, resulting in the final mask M_{result}.

Prompt Point-Based SAM Segmentation

To prevent the SAM prompt points from being too close to edges and potentially causing incorrect segmentation of the background, the M_{result} mask is pre-processed using morphological erosion with an 11×11 square structural element. This operation shrinks the original white regions inward by 5 pixels. K-means [20] clustering is performed on the pre-processed binary image, and the resulting cluster centres are employed as prompt points for SAM. In this study, the number of prompt points for general power equipment is set to 3 to avoid excessive or insufficient points that could cause SAM to overfocus on local details. However, for elongated multi-segment equipment, such as PT and arrester, M_{result} may consist of several connected regions that are relatively far apart. To ensure that all connected regions contain prompt points, the number of prompt points for this type of equipment is increased to 7. Furthermore, since the object detection in this study localizes the three-phase circuit breaker as a whole, the detection bounding box contains three independent pillars. Prompt points alone are insufficient to direct SAM's attention to the entire assembly; therefore, the number of prompt points is increased to 7 to ensure comprehensive segmentation. The clustering results are presented in Figure 7. The figure shows the clustering results of voltage transformer, arrester and breaker. These devices have obvious discontinuous white-value regions, which is caused by uneven temperature distribution of the device itself or coating that unfavourable for infrared detection. If the number of clustering points is less than the number of discontinuous white-value regions, some areas will not be noticed, as shown in (d) of the figure. The number of clusters mentioned above ensures that SAM will not recognize discontinuous white-value regions as independent devices. Since the basic models of the same type of equipment in the substation are the same, the specified cluster count has broad applicability. The K-means algorithm functions by iteratively updating centroids to minimize the following objective function:

$$J = \sum_{n=1}^{m} \sum_{in \in Cn} \left\| in - \mu n \right\|^2. \quad (21)$$

where m represents the specified cluster count, C_n denotes the n-th cluster set, i_n represents a sample point belonging to cluster C_n, and μ_n is the centroid of cluster C_n.



(a) (b) (c) (d)

Figure 7. Binary mask K-means clustering results: (a) Voltage transformer, (b) Arrester, (c) Breaker, (d) Insufficient prompts

The cluster centres are used as prompt points for SAM segmentation. Morphological opening and closing operations are then applied to the resulting mask to remove noise, ultimately achieving accurate segmentation of the power equipment within the detection bounding box.

4. Experiments and Results

4.1 Experimental Dataset

In the current research, the dataset employed includes two components: proprietary infrared images acquired during routine patrols of a power grid company, and public images retrieved from the Roboflow database (https://app.roboflow.com). It encompasses 7 typical categories of substation equipment, including insulators, bushings, disconnectors, breakers, arresters, voltage transformers, and current transformers, with a total of 2198 infrared images. Table 1 summarizes the quantity of equipment samples.

Table 1. Number of equipment samples in the infrared image dataset

Parameter	Insulator	Bushing	Disconnector	Breaker	Arrester	Voltage Transformer	Current Transformer
Number	565	254	414	261	244	199	261

For object detection, power equipment in the original infrared images was annotated using the image annotation function of the Roboflow platform. In order to balance the adequacy of model training and the credibility of model evaluation, we employed a random 7:2:1 split for the training, validation, and test sets. Besides, to address the problem that insufficient samples of diverse power equipment types in the infrared image dataset tend to bring about model overfitting, diversified random operations were performed on the annotated infrared images via data augmentation. This measure aims to expand the dataset size and enhance model robustness. Specifically, each training sample generates 2 output images: the original image is randomly subjected to horizontal or vertical flipping, or rotation in different directions; meanwhile, parameters such as hue, saturation, brightness, and exposure are adjusted within a certain range.

4.2 Experimental Environment

Within the Windows operating system environment, all experiments were performed, where the GPU model employed was RTX 3090. The experimental environment was configured as CUDA 12.9 + Anaconda + Python 3.10 + PyTorch 2.6.0.

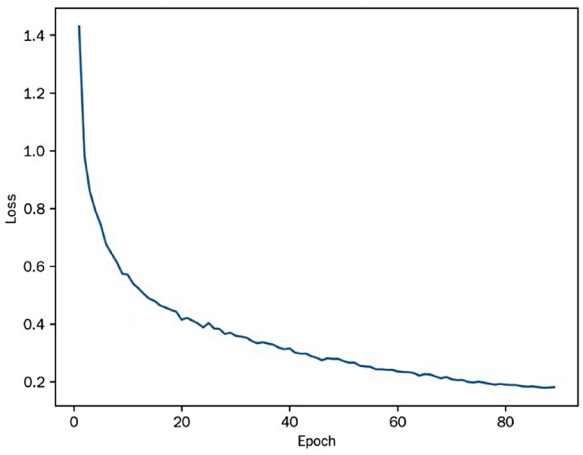


Figure 8. Loss Decline Curve

During the improved YOLOv8 model training process, the AdamW optimizer was adopted, in which the batch size was configured to 12, the initial learning rate was set at 1000, and the training epochs count was 100. Meanwhile, additional configurations included a warm-up epoch of 3, an early stopping strategy (i.e., training stops when no performance improvement is observed for 10 consecutive epochs), and cosine learning rate decay. It can be observed from the total loss function curve in Figure 8 that the total loss value drops drastically during the early stage of training. Along with the growing number of training epochs, the loss curve gradually stabilizes and finally converges to approximately 0.18. This demonstrates that the proposed method incorporating the Wise-MPDIoU loss function features a stable training process and achieves relatively satisfactory training results.

4.3 Object Detection Result

Evaluation Metrics

Each category's Average Precision (AP) and the mean Average Precision (mAP), and detection speed serve as the evaluation metrics for the model. Their calculation methods are as follows respectively:

$$P_{\text{precision}} = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{22}$$

$$R_{\text{recall}} = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{23}$$

$$P_{AP} = \int_0^1 P_{\text{precision}} \left(R_{\text{recall}} \right) \mathrm{d}R_{\text{recall}}, \tag{24}$$

$$P_{mAP} = \frac{\sum_{k=1}^{K} P_{AP}(k)}{K}. \tag{25}$$

Where: $P_{\text{precision}}$ represents Precision; R_{recall} signifies Recall; N_{TP} stands for the count of positive samples correctly identified; N_{FN} stands for the count of positive samples incorrectly classified as negative; N_{FP} stands for the count of negative samples incorrectly classified as positive; and K is the total number of categories. The model adopted an IoU threshold of 0.5.

Result and Analysis

In order to verify the merits of the method put forward, Faster R-CNN, RetinaNet, SSD, and YOLOv8 are adopted as comparative methods, and experiments are carried out on the self-built dataset. The specific findings are provided in Table 2, Figure 9, Table 3, and Table 4.

Table 2. Test results of the proposed method for different power equipment

Electric Power Equipment	mAP@0.5/%	Speed/（frame/s）
Insulator	90.8	
Bushing	98.8	
Disconnector	94.1	
Breaker	95.2	
Arrester	96.7	107.5
Voltage transformer	98.5	
Current transformer	95.5	
mAP /%	95.7	

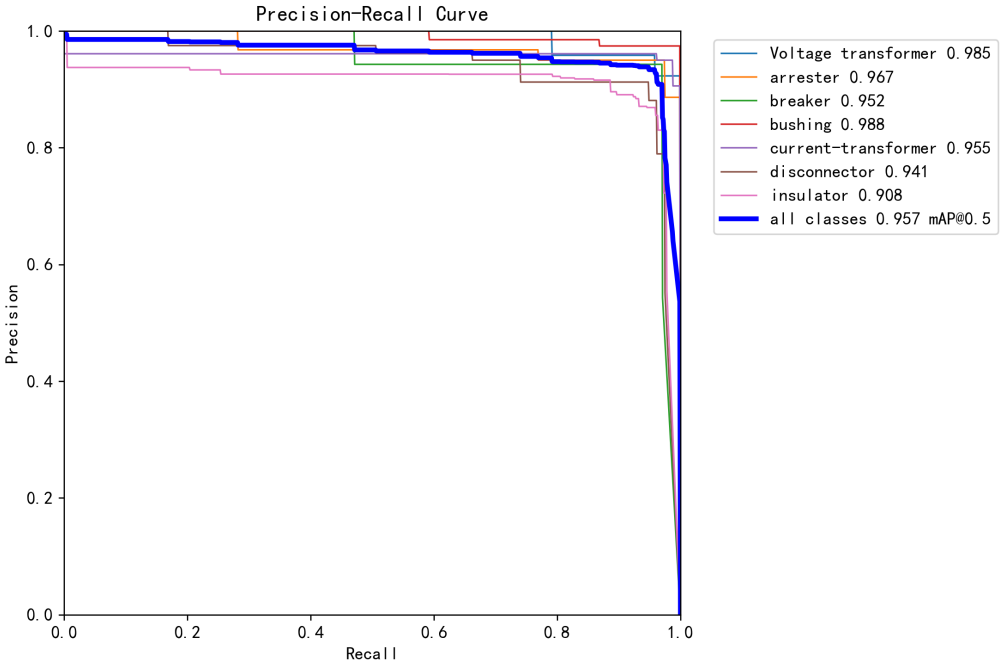


Figure 9. PR curve of the proposed method

As indicated by the PR curve in Figure 9 and the test results of the method we proposed regarding diverse power equipment in Table 2, the improved YOLOv8 achieves good detection performance for various power equipment such as voltage transformers, lightning arresters, and bushings. In the PR curve, the curves for all types of equipment can still retain a high level of precision as the recall rate is enhanced. In Table 2, the mAP@0.5 for bushings reaches 98.8%, and that for voltage transformers reaches 98.5%, both demonstrating excellent classification accuracy. The overall average mAP@0.5 is 95.7%, this demonstrates that our proposed approach works effectively in the object detection of infrared images for power equipment. As for the detection rate, the method we put forward attains 107.5 frame/s when processing power equipment infrared images, and this can fulfil the real-time needs of infrared image detection for power equipment.

Table 3. Comparison of detection performance of different methods for insulators and bushings

Method	Insulator	Bushing
Faster R-CNN	88.9	96.5
SSD	86.0	93.7
YOLOv7	87.4	95.1
YOLOv8	88.5	96.1
Proposed method	**90.8**	**98.8**

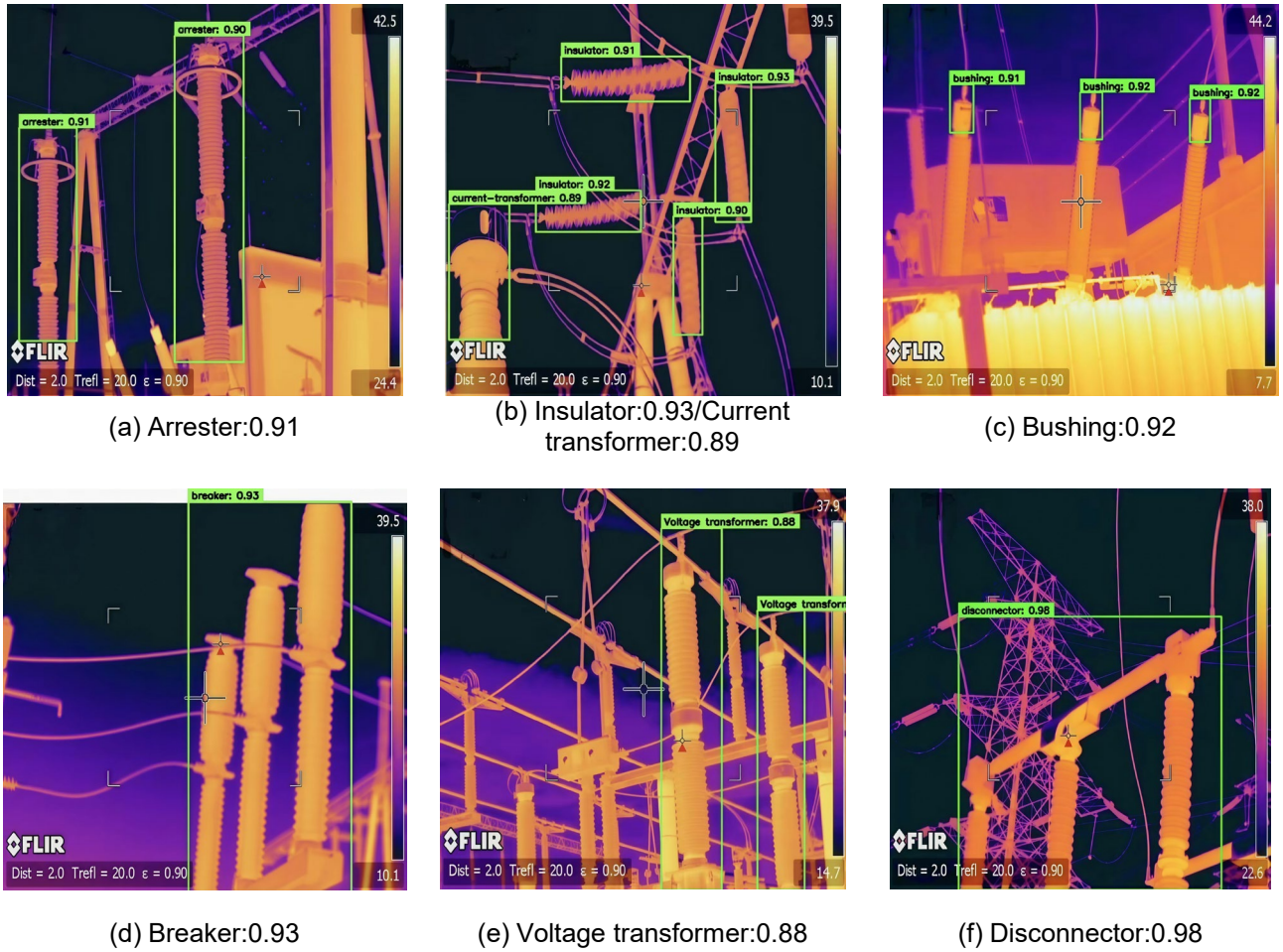
Table 4. Performance comparison of different methods

Method	mAP@0.5/%	Speed/（frame/s）
Faster R-CNN	93.7	17
SSD	91.1	41
YOLOv7	92.4	103
YOLOv8	93.4	110
Proposed method	**95.7**	**107.5**

Table 3 and Table 4 demonstrate that the improved YOLOv8 is significantly superior to the comparison methods (Faster R-CNN, SSD, YOLOv7, and YOLOv8) in the detection of small-scale targets in power equipment infrared images. In Table 3, the mAP@0.5 for insulators reaches 90.8%, which is 2.3 percentage points higher than that of YOLOv8 (88.5%); the mAP@0.5 for bushings reaches 98.8%, which is 2.7 percentage points higher than that of YOLOv8 (96.1%). These results fully demonstrate the accurate detection capability of the proposed method for small-scale power equipment targets. In terms of overall performance, Table 4 shows that the average mAP@0.5 of the proposed method is 95.7%, which is 2.3 percentage points higher than that of YOLOv8 (93.4%); its detection speed is 107.5 frame/s, which has a slight adjustment due to algorithm improvements. However, considering the comprehensive performance of precision and speed, the presented method

possesses higher application value in the object detection task for power equipment infrared images.

Figure 10 displays the outcomes of object detection for the improved YOLOv8, which shows that all power equipment can be well detected, and the detection confidence is high.



(a) Arrester:0.91

(b) Insulator:0.93/Current transformer:0.89

(c) Bushing:0.92

(d) Breaker:0.93

(e) Voltage transformer:0.88

(f) Disconnector:0.98

Figure 10. The recognition results of the proposed object detection method

4.4 Segmentation Result

The key parameters of the proposed method are listed in Table 5. In addition, the sensitivity analysis results of relevant parameters are shown in Figure 11. All other parameter values in the table were kept constant while only the value of the parameter of interest was altered. If the significance proportionality coefficient is set too high or too low, the image will be over-segmented. If the number of superpixels is too large, the denoising effect is not obvious, and if it is too small, it is easy to be over-segmented. Saliency weight threshold has almost no effect on the results, but it can avoid reverse segmentation of the image. If the superpixel ratio threshold is too large, it is easy to be under-segmentation, and if it is too small, it is easy to be over-segmented. In general, each parameter has a wide applicable range.

Table 5. Parameter Settings.

Parameter	Value
Significance proportionality coefficient λ_1	0.3
Significance proportionality coefficient λ_2	0.7
Number of superpixels w	70
Saliency weight threshold T_{saliency}	0.65
Superpixel ratio threshold T_{super}	0.5

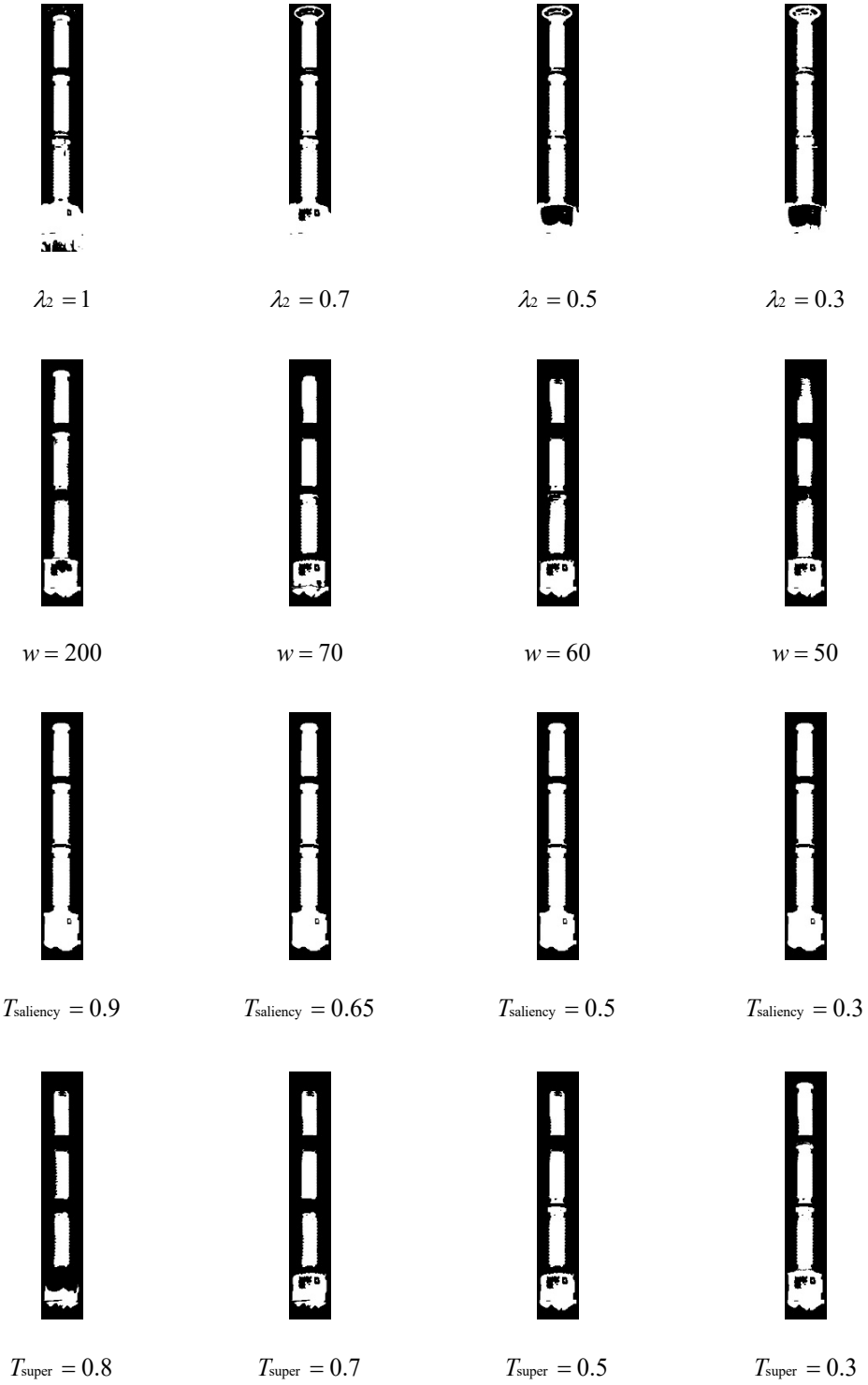


Figure 11. Parameter sensitivity analysis

To confirm the advantage of the segmentation method put forward, experiments were carried out on various equipment in diverse scenarios. A comparative analysis was performed against bounding box-based GrabCut and SAM segmentation methods, using IoU as the criterion to assess the segmentation performance of the different algorithms. The comparison results are presented in Figure 12.

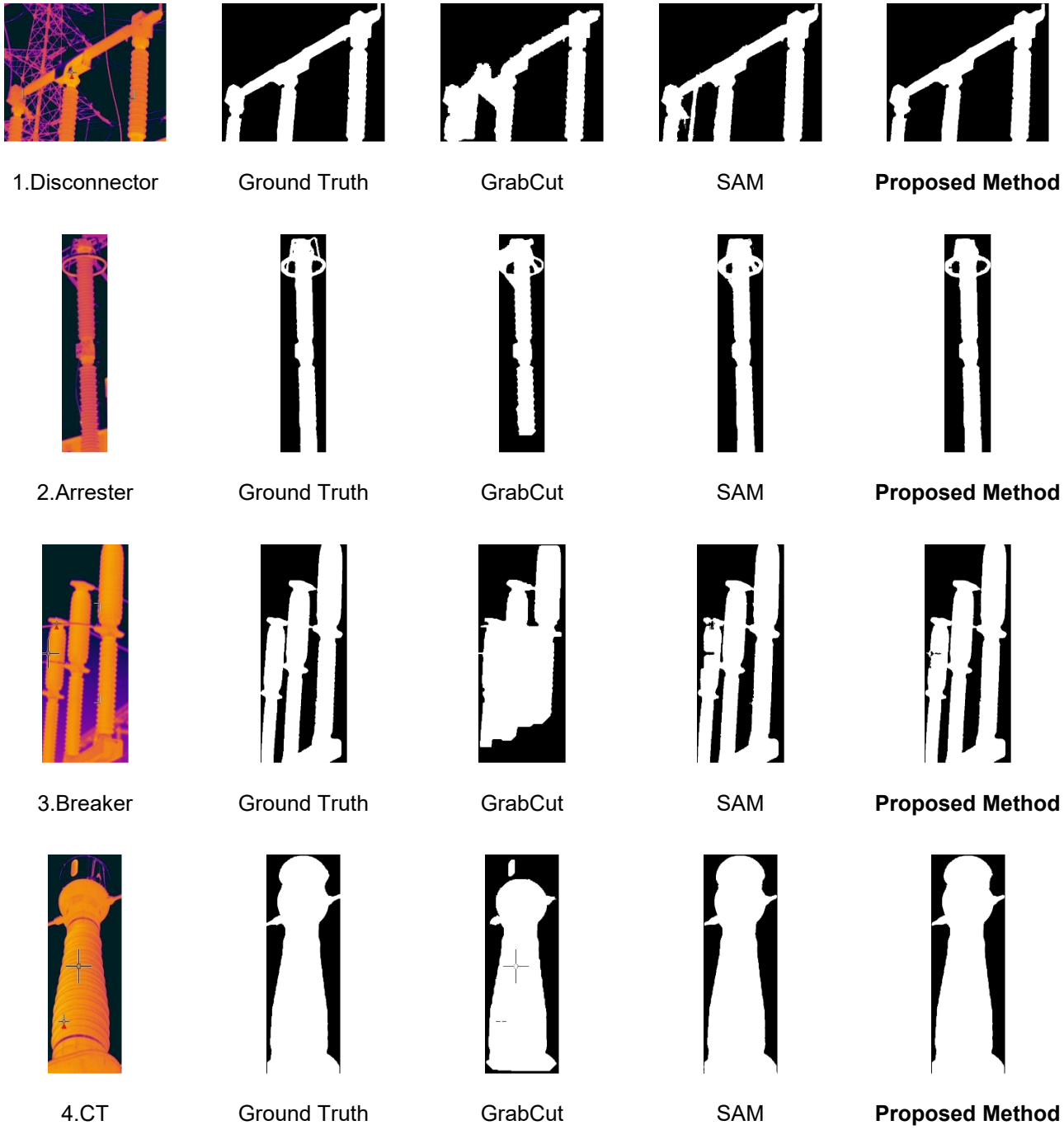


Figure 12. Infrared image segmentation results comparison

Table 6. IoU values of different methods.

Image	GrabCut	SAM	**Proposed Method**
1	0.7730	0.9125	**0.9963**
2	0.8272	0.9363	**0.9832**
3	0.6832	0.9757	**0.9955**
4	0.8711	0.9947	**0.9971**

IoU values fall within the interval [0, 1], and values that are nearer to 1 mean the segmentation result is more precise. As shown in Figure 12 and Table 6, the proposed segmentation method significantly outperforms the bounding box-based GrabCut algorithm. Moreover, it achieves better segmentation performance than the bounding box-based SAM approach when the detection bounding box contains a substantial amount of background. However, since cluster

centres tend not to be located at the edge regions of the mask, segmentation performance slightly declines when the temperature at the ends of the power equipment is relatively low. The dataset used in this study retains area markers and scale markers from the infrared camera, which introduced some interference in the segmentation process. Nonetheless, the resulting findings prove that our proposed approach shows powerful robustness in the face of such disturbances.

Within the task of segmenting substation equipment's infrared images, when the bounding box contains background regions with temperatures similar to the foreground, even a precisely localized bounding box may cause interfering regions within the box to be mistakenly segmented as foreground. The bounding box-based GrabCut algorithm essentially clusters pixels with significant colour differences, making it difficult to accurately segment infrared images where temperature differences are often subtle. In contrast, SAM leverages both global image context and prompt information. However, due to the complex background typical in power equipment scenes, point prompts allow SAM to focus more effectively on the main body of the equipment compared to prompt boxes, resulting in improved segmentation performance.

Although the proposed method achieves accurate segmentation for structurally intact power equipment, it struggles in cases with heavy occlusion or significant overlap between objects of the same type. Therefore, the method has some dependence on the viewing angle during infrared image acquisition. Furthermore, for equipment with low surface emissivity, point-based segmentation may fail to outline the complete object. In such scenarios, simultaneously utilizing both the bounding boxes and prompt points can help avoid under-segmentation and ensure the entire object is captured.

5. Conclusions

We propose a novel framework for object detection and segmentation to analyse infrared images of power equipment in substations. Our main innovations are designed to deal with the difficulties of small-target recognition and complex background interference. The conclusions of this study are as follows:

- We introduce three enhancements to the YOLOv8 backbone, including Coordinate Attention for better focus on small targets, Robust Feature Downsampling to reduce the loss of important information, and Large-Kernel Depthwise Separable Convolution to enlarge the receptive field. In addition, a bounding box corner position penalty term is incorporated to improve localization accuracy for irregularly shaped equipment;
- We fuse colour saliency, GrabCut and superpixel to generate the high-confidence foreground mask, and then guide SAM via clustered prompt points. The original approach makes prompts focus on foreground regions, suppressing the over-segmentation in complex scenes. Meanwhile, this segmentation method does not need pixel level data training;
- Experimental results show that the improved YOLOv8 has better performance than Faster R-CNN, SSD, YOLOv7, and the original YOLOv8 when detecting small-scale targets like insulators and bushings. Compared with prompt box-based SAM and traditional GrabCut, this algorithm has higher IoU in complex backgrounds.

The future research will further enhance the detection ability on different datasets, and improve the segmentation accuracy by detecting the small parts of power equipment.

References

[1] Usamentiaga R, Fernandez MA, Villan AF, Carus JL. Temperature Monitoring for Electrical Substations Using Infrared Thermography: Architecture for Industrial Internet of Things. IEEE Trans Ind Inform. 2018;14(12):5667–5677.

[2] Liu ZQ, Fu H, Li YJ, Zhang GJ, Hu CB, Zhang ZH. Infrared Image Power Equipment Detection Based on Mask-RCNN Transfer Learning. J Data Acquis Process. 2021;36(1):176–183.

[3] Liu X, Zhang Z, Hao Y, Zhao H, Yang Y. Optimized OTSU Segmentation Algorithm-Based Temperature Feature Extraction Method for Infrared Images of Electrical Equipment. Sensors. 2024;24(4):1126.

[4] Jadin MS, Taib S. Recent Progress in Diagnosing the Reliability of Electrical Equipment by Using Infrared Thermography. Infrared Phys Technol. 2012;55(3):236–245.

[5] Ou JH, Wang JG, Xue J, Wang JP, Zhou X, She LG, Fan YD. Infrared Image Target Detection of Substation Electrical Equipment Using an Improved Faster R-CNN. IEEE Trans Power Del. 2023;38(1):387–396.

[6] Zhang L, Kuang J, Teng Y, Xiang S, Li L, Zhou Y. A Lightweight Infrared and Visible Light Multimodal Fusion Method for Object Detection in Power Inspection. Processes. 2025;13(9):2720.

[7] Xu C, Li Q, Jiang X, Yu D, Zhou Y. Dual-Space Graph-Based Interaction Network for RGB-Thermal Semantic Segmentation in Electric Power Scene. IEEE Trans Circuits Syst Video Technol. 2023;33(4):1577–1592.

[8] Wang B, Dong M, Ren M, Wu ZY, Guo CX, Zhuang TX, Pischler O, Xie JC. Automatic Fault Diagnosis of Infrared Insulator Images Based on Image Instance Segmentation and Temperature Analysis. IEEE Trans Instrum Meas. 2020;69(8):5345–5355.

[9] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R. Segment Anything. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 October 1-6; Paris, France. Piscataway, NJ: IEEE; 2023. p. 3992–4003.

[10] Ultralytics LLC. YOLOv8 Documentation Release 8.0. San Francisco, CA, USA: Ultralytics LLC; 2023.

[11] Wu T, Dong Y. YOLO-SE: Improved YOLOv8 for Remote Sensing Object Detection and Recognition. Appl Sci. 2023;13(24):12977.

[12] Ma M, Pang H. SP-YOLOv8s: An Improved YOLOv8s Model for Remote Sensing Image Tiny Object Detection. Appl Sci. 2023;13(14):8161.

[13] Wang G, Chen Y, An P, Hong H, Hu J, Huang T. UAV-YOLOv8: A Small-Object-Detection Model Based on

Improved YOLOv8 for UAV Aerial Photography Scenarios. Sensors. 2023;23(16):7190.
[14] Rother C, Kolmogorov V, Blake A. GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts. ACM Trans Graph. 2004;23(3):309–314.
[15] Dan B, Li M, Tang T, Zhang J. One Shot Is Enough for Sequential Infrared Small Target Segmentation. In: Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 April 6-11; Hyderabad, India. Piscataway, NJ: IEEE; 2025. p. 1–5.
[16] Li Y, Wang D, Yuan C, Li H, Hu J. Enhancing Agricultural Image Segmentation with an Agricultural Segment Anything Model Adapter. Sensors. 2023;23(18):7884.
[17] Ma X, Li Y. Edge-Aided Multiscale Context Network for Infrared Small Target Detection. IEEE Geosci Remote Sens Lett. 2023;20(1):1–5.
[18] Wang T, Zhang J, Ren B, Liu B. MMW-YOLOv5: A Multi-Scale Enhanced Traffic Sign Detection Algorithm. IEEE Access. 2024;12(1):148880–148892.
[19] Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Trans Pattern Anal Mach Intell. 2012;34(11):2274–2282.
[20] Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. J R Stat Soc Ser C (Appl Stat). 1979;28(1):100–108.