Proceedings of Machine Learning Research vol 278:1–13, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

# Research on interpretable methods for detecting elongated objects in power operations

**Miao Li** LM2018216002@163.COM
*Beijing Union University, Beijing, China*

**Yanxia Liu*** YANXIA.LIU@163.COM
*Beijing Union University, Beijing, China*

**Editors:** Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

## Abstract

Power operations take place in high-risk environments, such as high voltage and strong magnetic fields, making standardized procedures crucial. Employing object detection technology to monitor operational compliance enhances electrical safety. However, the presence of numerous elongated objects and background columnar interferences in power operation datasets significantly affects detection accuracy. To address this issue, we explore model structure improvements from an interpretability perspective. Using Grad-CAM heatmap visualization, we analyze the regions where the model focuses on detection targets. We propose a lightweight convolutional attention mechanism, LCA (Lightweight Convolution Attention), which significantly enhances YOLOv7's attention to elongated targets while reducing the impact of columnar interference. This improves both the model's robustness and interpretability. Experimental results show that LCA outperforms classical attention modules such as SE, ECA, and CA, while maintaining a minimal parameter size. Specifically, the mAP of the extremely elongated and challenging sample "operatingbar" increased by 4.4%, and the mAP of the small target "wrongglove" improved by approximately 2%. This makes LCA well-suited for detecting elongated targets in complex power operation environments.

**Keywords:** Elongated Target Detection; Attention Mechanism; Grad-CAM; Power Field Operations; Interpretability Analysis

## 1. Introduction

The power field operation process is complex, and improper operation or lack of protective equipment can easily lead to safety accidents. Power operators are required to perform electrical testing and disconnection operations on-site every day. Manual inspections are not only time-consuming but also prone to oversight. To further reduce labor costs, it is necessary to develop efficient object detection algorithms to monitor the compliance of power field operations, ensuring safety and improving inspection efficiency.

With the advancement of convolutional neural network theory and the improvement of computational power, deep learning-based detection algorithms, known for their superior representational ability and robustness, have been applied in safety-sensitive fields such as power and healthcare. Compared to traditional methods, they offer a better balance between detection speed and accuracy (Gong et al., 2018). Li et al. (2022) improved the MobileNetV2 and SSDLite networks using depthwise separable convolutions to enhance the detection accuracy and speed for small targets. The detection targets include people and ladders in substations. Peng et al. (2023) improved the YOLOv5 network using GhostNetV2 for smoking behavior detection in substations. Xiao et al. (2022) enhanced YOLOv5 using PANet to detect fires, large trucks, and workers at substations.

Zhou et al. (2023) introduced a method called MD-Yolov7, which combines Multi-Layer Feature Fusion (MLFF) and Detection Transformer (DETR) techniques for intelligent detection of electrical equipment in substations. Yang et al. (2024) reduced model parameters by removing large branches in the network detection component, aimed at detecting faults in substations. Wu et al. (2024) introduced the InResNet architecture to replace the residual block structure of ResNet-34 in Faster R-CNN, improving the recognition accuracy of different types of electrical equipment in infrared images. Zhang et al. (2024) proposed a target detection model called YOLO-SD for defect detection in substation equipment, with primary targets including insulators and electric meters.

In existing studies on target detection in the power sector, research subjects primarily include electrical equipment (Zhou et al., 2023), safety helmets (Wang, 2021), ladders (Li et al., 2022), smoking behavior (Peng et al., 2023), and fire, large trucks, and workers in substations (Xiao et al., 2022). These targets exhibit relatively small size variations, with almost no extremely elongated objects featuring high aspect ratios. However, in power operation compliance detection scenarios, elongated objects are quite common. Additionally, the background of power operation datasets collected from substation construction sites is highly complex. It contains numerous cylindrical interfering objects that share similar contour features with the detection targets, such as operating rods and voltage detectors. This poses significant challenges for object detection. Studies indicate that evaluating only elongated objects from the COCO dataset results in a sharp 18.9% drop in mAP, highlighting the difficulty of achieving high-precision detection for elongated targets (Wan et al., 2020). Ou et al. (2023) proposed an improved Faster R-CNN by introducing anchor boxes with aspect ratios of 1:3 and 3:1 to enhance the detection accuracy of elongated equipment. However, its effectiveness for objects with extreme aspect ratios remains unverified. Additionally, existing object detection algorithms applied in the power sector generally suffer from poor interpretability (Kehe, 2020), which is one of the key obstacles hindering the deployment of deep learning algorithms in safety-critical applications such as power operations (Tianjiao et al., 2023).

In summary, the challenges of normative behavior detection in substation construction sites primarily lie in two aspects: (1) The dataset contains extremely elongated targets with large size variations, small inter-class variance, and high similarity between background objects and foreground targets. (2) The poor interpretability of deep learning models. To address these challenges in power operation object detection, this study employs Grad-CAM-based (Selvaraju et al., 2017) class activation heatmaps for interpretability analysis and designs an attention mechanism specifically for detecting elongated objects to enhance the YOLO object detection algorithm.

Heatmap visualization has a wide range of applications. In organ lesion detection, heatmaps highlight the exact location of lesions and help assess whether the model's reasoning process is interpretable (Jiang et al., 2020). In industrial defect detection, heatmaps allow users to identify the model's focus areas and understand its classification criteria (Xia et al., 2020). Overall, interpretability analysis enhances the transparency of deep learning models, mitigates their "black-box" nature, and improves their trustworthiness.

The most common attention mechanisms include self-attention (Vaswani et al., 2017), spatial attention (Woo et al., 2018), and channel attention (Wang et al., 2020; Hou et al., 2021). Although Transformer-based models leveraging self-attention have made significant progress in computer vision (CV), self-attention primarily focuses on capturing global information (Srinivas et al., 2021), making it less effective for detecting extremely elongated, fine-grained targets. To address this, Bhattacharya et al. (2020) introduced a parallel feature extraction mechanism to capture fine details.

Table 1: Power Operation Dataset

| Class | Label | Note |
|---|---|---|
| Guardianship armband | badge | Red armband |
| Person | person | All on-site personnel |
| Insulated gloves | glove | Insulation gloves |
| No gloves | wrongglove | Other gloves or bare hands |
| Operating rods | operatingbar | Yellow, with uniform thickness |
| Voltage detectors | powerchecker | Red or blue, Irregular thickness |

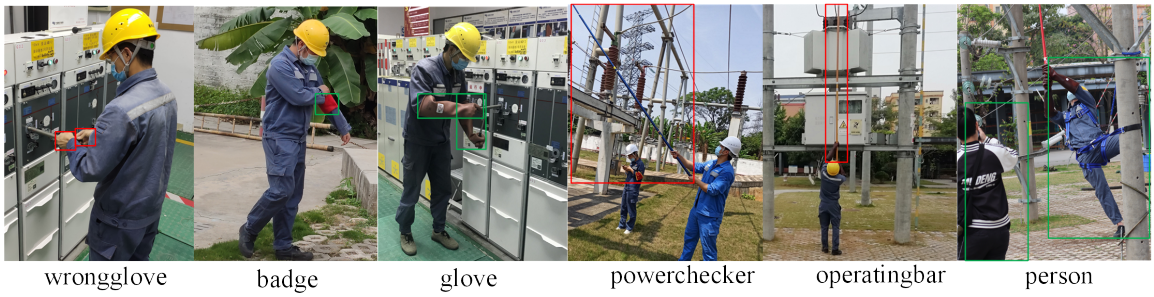


Figure 1: Example of data set.

However, the use of multiple multi-head attention branches slows down detection speed, making it unsuitable for real-time applications.

This study focuses on the power operation dataset and employs Grad-CAM visualization technology to guide attention mechanism improvements from an interpretability perspective, advancing from machine learning to machine teaching. A lightweight attention mechanism, LCA, is proposed specifically for detecting elongated targets. Compared to CA (Hou et al., 2021), SE (Hu et al., 2018), and CBAM (Woo et al., 2018), LCA is more lightweight, with parameter complexity comparable to ECA (Wang et al., 2020). It has minimal impact on YOLOv7's detection speed while offering improved interpretability. To further validate LCA's effectiveness, experiments were conducted on a subset of elongated targets from the COCO public dataset. The results confirmed the model's strong performance in detecting elongated objects.

## 2. Interpretable Target Detection for Substation On-Site Operations

### 2.1. Construction of a Substation Safety Inspection Dataset

This project uses the dataset provided by the Guangdong Power Grid Smart On-site Operations Challenge. Power grid operators are required to perform live-line and power-off operations daily. To ensure safety, inspectors must wear rubber insulating gloves, use operating rods and voltage testers, and have a supervisor present on-site to provide assistance and raise alarms in case of emergencies.

The power operation dataset contains 2,619 images, categorized into six classes, as detailed in Table 1. The dataset includes challenging-to-detect elongated objects ('operatingbar' and 'powerchecker') and small objects ('wrongglove'), as shown in Figure 1. The operating rods and voltage testers have very similar contour shapes, with a small inter-class variance, which poses a significant challenge for target detection and classifies them as difficult samples.

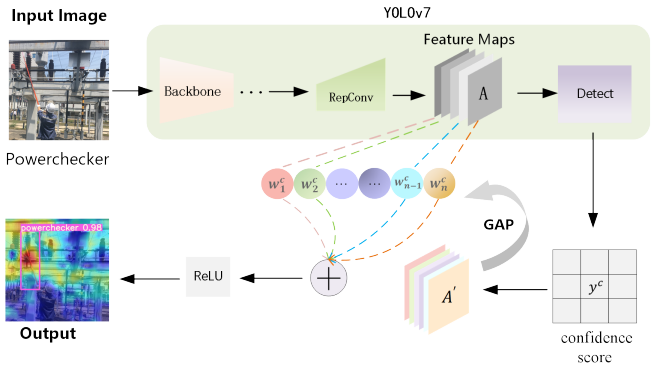


Figure 2: Structure of YOLOv7-GCAM region of interest visualization algorithm.

### 2.2. YOLOv7-GCAM Interpretability Analysis Algorithm Structure

Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) generates visualizations highlighting the regions of focus when a model predicts a specific class, thereby demystifying the "attention" and "learned features" within the model and breaking the "black box" nature of deep learning. By analyzing whether its learned features match reality, it can guide model design and improvement, making 'machine teaching' possible. For instance, we use Grad-CAM to analyze the attention region of YOLOv7 for the "powerchecker" class. The network structure of YOLOv7-GCAM is shown in Figure 2.

First, the confidence score $y^c$ for the "powerchecker" class predicted by the YOLOv7 model is differentiated with respect to each element of the feature map A extracted by the backbone network. This results in the gradient matrix A', which indicates the contribution of each element in A to the "powerchecker" class. A higher value signifies that the network has learned more important features from that region. Next, perform global average pooling on A' across each channel to obtain the importance level of each channel regarding the prediction of the "powerchecker" category. Finally, calculate the weighted sum of the feature map A using these importance scores and apply the ReLU function. This yields the gradient-weighted class activation mapping for the "powerchecker" category, with a similar process for other categories.

## 3. Method for Improving Attention Mechanism Guided by Grad-CAM

Attention mechanisms, including SE, ECA, CBAM, CA, were added after the SPPCSPT module in the neck of YOLOv7. Each experiment introduced only one attention mechanism, and the modified models were named SE-YOLOv7, ECA-YOLOv7, CBAM-YOLOv7, and CA-YOLOv7, respectively.

SE-YOLOv7: As shown in Figure 3, the attention mechanism in SE-YOLOv7 exhibits greater focus compared to the baseline YOLOv7 model. However, substantial red-highlighted regions persist on slender columnar interference objects beyond the detection targets, indicating localization errors and insufficient interpretability.

ECA-YOLOv7: Figure 3 reveals that the heatmap corresponding to ECA demonstrates dispersed attention regions with relatively weak interpretability. Nevertheless, the ECA-YOLOv7 module achieves higher mAP than SE-YOLOv7, particularly showing significant improvement in

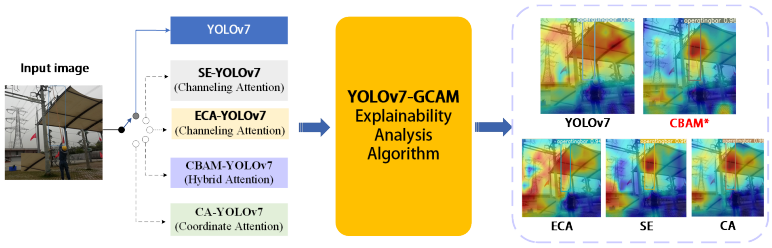


Figure 3: Activation heatmap visualization for slender-object detection in YOLOv7

small object detection (see Section 5.2). As illustrated in Figure 4, the fundamental distinction between ECA and SE architectures lies in their channel interaction mechanisms: SE employs fully connected layers, whereas ECA utilizes 1D convolutional layers. The ECA module replaces SE's fully connected layers with 1D convolution to establish local cross-channel interactions. This design effectively captures channel-wise dependencies and local correlations, thereby enhancing model accuracy and robustness. This validates the effectiveness of ECA's 1D convolution-based local cross-channel interaction mechanism for accuracy enhancement.

CBAM-YOLOv7: As shown in Figure 3, the heatmap's key focus areas do not overlap with other distracting objects, indicating that the CBAM module has strong interpretability. As shown in Figure 4, CBAM consists of a channel attention module and a spatial attention module (SAM), where the spatial attention mechanism significantly enhances the model's interpretability.

CA-YOLOv7: As shown in Figure 3, similar to SE, CA highlights large areas in red on elongated columnar interference objects in addition to the detected targets, indicating lower interpretability. However, the CA attention module significantly improves the mAP(see Section 5.2) of elongated targets, validating the effectiveness of the coordinate attention mechanism in aggregating directional features for elongated objects. As shown in Figure 4, the CA coordinate attention module is essentially based on channel attention mechanisms in the x and y directions. This directional feature aggregation approach is effective at capturing long-range dependencies, which is beneficial for detecting elongated objects.

In summary, the CBAM attention module significantly enhances model interpretability, while the directional feature aggregation in the CA attention mechanism effectively captures long-range dependencies of elongated targets. The 1D convolution in ECA extracts attention between adjacent channels, improving detection accuracy. To leverage the strengths and address the limitations of these attention mechanisms, this study incorporates 1D convolution to improve the CA attention mechanism. Additionally, it adopts the strong interpretability of the SAM module in CBAM to redesign a new attention mechanism, LCA (see Chapter 4), aimed at enhancing feature extraction for elongated targets.

## 4. LCA(Lightweight Convolution Attention)Attention Mechanism

We propose a novel lightweight convolutional attention mechanism, LCA, which consists of a channel attention module (CAM) and a spatial attention module (SAM), as illustrated in Figure 5.

Let the input feature map of the LCA attention mechanism be $F \in \mathbb{R}^{C\times H\times W}$,and the output feature map $F'' \in \mathbb{R}^{C\times H\times W}$ is given by Equation (1). Here, C, H, and W represent the number of channels, height, and width of the input feature map $F$, respectively.$M_c(F) \in \mathbb{R}^{C\times 1\times 1}$ denotes

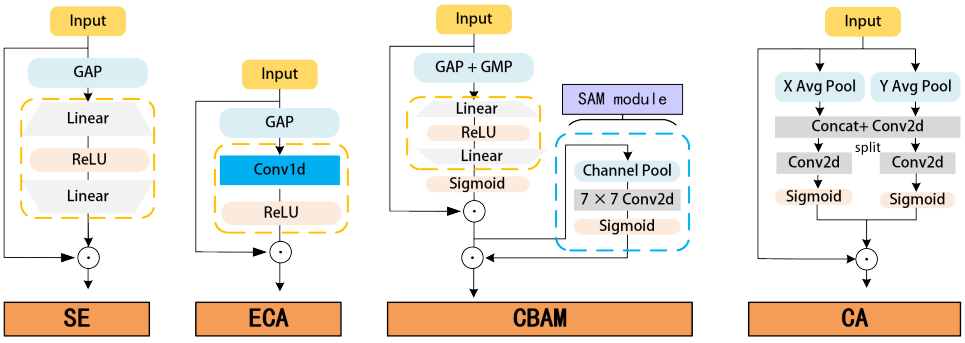


Figure 4: Structural Comparison Diagram of Different Attention Mechanisms

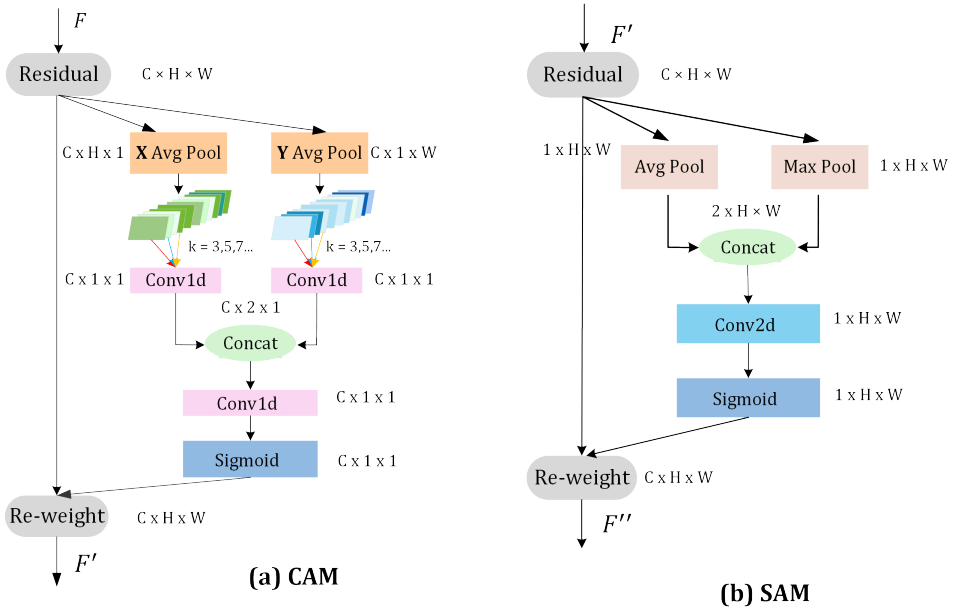


Figure 5: Structure of the LCA Attention Mechanism

the channel attention weights, and $M_s(F') \in \mathbb{R}^{C\times1\times1}$ represents the spatial attention weights.$\otimes$ indicates that $M_c(F)$ is broadcast (replicated) along the H, W, and C dimensions and then element-wise multiplied with the corresponding input feature map $F$.

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \tag{1}$$

### 4.1. Channel Attention

**Directional feature aggregation.** For each channel of the input feature map F, feature aggregation is performed using pooling kernels of size (1,W) along the horizontal axis (X direction) and (H,1) along the vertical axis (Y direction), resulting in a pair of direction-aware feature maps $Z_c^h$ and $Z_c^w$,

as shown in Equations (2) and (3).

$$Z_c^h = \frac{1}{w} \sum_{0 \le i \le w} F_c(h, i) \tag{2}$$

$$Z_c^w = \frac{1}{h} \sum_{0 \le j \le h} F_c(j, w) \tag{3}$$

$F_c(h, i)$ denotes the pixel value at position (h,i) in the c-th channel of the feature map F,and $Z_c^h$ represents the aggregation result along the X-axis for the h-th row of the c-th channel feature map $F_c$. Similarly, $F_c(j, w)$ denotes the pixel value at position (j,w) in the c-th channel of the feature map F, and $Z_c^w$ represents the aggregation result along the Y-axis for the w-th column of the c-th channel feature map $F_c$. Equation (2) aggregates long-range dependencies along the horizontal direction while preserving precise positional information in the vertical direction. Similarly, Equation (3) aggregates long-range dependencies along the vertical direction while maintaining precise positional information in the horizontal direction. Generally, elongated objects tend to have large spans along both the X and Y axes. Therefore, the aforementioned directional feature aggregation is more effective in capturing feature information of elongated objects along both directions.

The above process is applied to a single channel. For the entire feature map F, where $c \in \{1, 2, \cdots C\}$, $h \in \{1, 2, \cdots H\}$, and $w \in \{1, 2, \cdots W\}$ , the feature map $Z^H \in \mathbb{R}^{C \times H \times 1}$ is obtained by aggregating along the x-axis direction using Equation (2); similarly, the feature map $Z^W \in \mathbb{R}^{C \times 1 \times W}$ is obtained by aggregating along the y-axis direction using Equation (3).

**Local cross-channel interactions.** First, a one-dimensional convolution $Conv_{1d}^k$ with kernel depths of $H$ and $W$ is applied to $Z^H$ and $Z^W$, respectively, to achieve local cross-channel interaction. This results in $F^H \in \mathbb{R}^{C \times 1 \times 1}$ and $F^W \in \mathbb{R}^{C \times 1 \times 1}$, as shown in Equations (4) and (5).

$$F^H = Conv_{1d}^k(Z^H) \tag{4}$$

$$F^W = Conv_{1d}^k(Z^W) \tag{5}$$

Here, $Conv_{1d}^k$ represents a one-dimensional convolution, where k is the kernel size. The value of k determines the interaction range between adjacent channels and is an adjustable odd number, typically set to 3, 5, 7, or 9. Next, $F^H$ and $F^W$ are concatenated and processed through a one-dimensional convolution $Conv_{1d}^2$ with a kernel size of 2, followed by a sigmoid function, to obtain the channel attention weights $M_C(F)$ ,as shown in Equation (6).Finally, according to Equation (1), the original feature map F is refined to obtain the intermediate feature map $F'$,which maintains the same dimensions as F.

$$M_c(F) = \left[\sigma(Conv_{1d}^2 \left[F^H; F^W\right])\right] \tag{6}$$

During training, the convolutional kernels of each channel in a CNN are updated independently via gradient descent, gradually adapting to detect specific feature types. In other words, each channel can be regarded as a dedicated feature detector, capturing different aspects of the input image. Therefore, local cross-channel interactions facilitate learning more effective attention weights.

### 4.2. Spatial Attention

The input to the spatial attention mechanism is the intermediate feature map $F'$,First, average pooling and max pooling are applied along the channel dimension, producing $F'_{avg} \in \mathbb{R}^{1 \times H \times W}$ and

$F'_{max} \in \mathbb{R}^{1\times H\times W}$ respectively. These two feature maps are then concatenated along the channel dimension and processed by a 2D convolution with a kernel depth of 2 to extract spatial attention information. The output is passed through a sigmoid function to obtain the spatial attention weights. Finally, using Equation (1), the LCA attention mechanism generates the refined feature map $F''$.

## 5. Experiment

### 5.1. Experimental Environment

The training process was conducted on a server running Ubuntu 20.04, with PyTorch 1.13.1 as the deep learning framework. The hardware setup included an 11th Gen Intel(R) Core(TM) i5-11400 @ 2.60GHz CPU and an NVIDIA GeForce RTX 3060 GPU with 80 GB of memory. The experiments, including comparative and ablation studies, were conducted on a substation power operation dataset, with the training and validation sets split in a 4:1 ratio. Transfer learning was employed, utilizing pre-trained weights from the COCO dataset for all improved models. The training was optimized using the SGD optimizer with a batch size of 64, 100 training epochs, an initial learning rate of 0.001, and a final learning rate of 0.01. The same validation set was used for evaluation after each training epoch.

### 5.2. Ablation and Comparative Experiments

The experimental results of the original YOLOv7 model are shown in Table 2. It is evident that "wrongglove", "operatingbar", and "powerchecker" are challenging samples, with mAP@.5 values below 90%. As shown in Figure 6, the loss function gradually decreases and stabilizes with increasing iterations, eventually reaching a low and steady level, indicating satisfactory model convergence.

SE, ECA, CBAM, CA, and LCA (ours) attention mechanisms were respectively added after the SPPCSPT module in the neck of YOLOv7. The experimental results are shown in Table 3, and the heatmap visualizations are presented in Figure 7. The SE attention mechanism shows limited interpretability and moderate detection performance for elongated objects, making it more suitable for general applications with low safety sensitivity. "YOLOv7+CA" demonstrates strong detection performance for elongated objects but lacks interpretability. In addition to the target objects, the heatmap shows large red-highlighted areas on other elongated, column-like distractors. LCA improved the performance by 1.8% on the small target "wrongglove", 4.4% on the elongated target "operatingbar", and 1.6% on "powerchecker". LCA has fewer parameters than attention modules such as CA and SE, resulting in better speed performance. In the heatmap of the LCA-YOLOv7 model (shown in Figure 7(f)), the red highlights in the important feature areas of elongated targets such as "operatingbar" and "powerchecker" are more concentrated, indicating the network's strong capability in key feature extraction and target localization.

To validate the effectiveness of the proposed LCA, we conducted experiments on its CAM and SAM modules, as shown in Table 4. The mAP@0.5 of "YOLOv7+CAM" reached 89.0%, showing a 1.2% improvement over the original YOLOv7 model. Meanwhile, "YOLOv7+LCA" achieved an mAP@0.5 of 89.3%, outperforming YOLOv7 by 1.5%. Additionally, to demonstrate the generalizability of the model, we conducted experiments on YOLOv5, which also yielded promising results.

Table 2: Evaluation results of the original YOLOv7 model across different categories

| class(label) | images | lable | mAP@.5(%) |
|---|---|---|---|
| glove | 523 | 336 | 95.1 |
| bagde | 523 | 154 | 96.4 |
| person | 523 | 958 | 98.4 |
| wrongglove | 523 | 1069 | 75.7 |
| operatingbar | 523 | 170 | 75.7 |
| powerchecker | 523 | 163 | 85.4 |


Figure 6: Loss function curve

Table 3: Comparison of Different Attention Mechanisms

| CNN model | Param(M) | mAP@.5(%) | Challenging samples mAP@.5(%) | | |
|---|---|---|---|---|---|
| | | | wrongglove | operatingbar | powerchecker |
| YOLOv7 | 36.50 | 87.8 | 75.7 | 75.7 | 85.4 |
| YOLOv7+SE | 36.54 | 88.7 | 77.8 | 77.0 | 85.7 |
| YOLOv7+CBMA | 36.54 | 88.6 | 76.8 | 79.6 | 82.9 |
| YOLOv7+ECA | 36.50 | 88.8 | 78.1 | 76.8 | 86.2 |
| YOLOv7+CA | 36.53 | 88.6 | 76.6 | 77.5 | 86.7 |
| YOLOv7+LCA(ours) | 36.51 | **89.3** | 77.5 | 80.1 | 87.0 |

Table 4: Experimental results of LCA and its components on YOLOv5 and YOLOv7

| CNN model | Param(M) | mAP@.5(%) | Challenging samples mAP@.5(%) | | |
|---|---|---|---|---|---|
| | | | wrongglove | operatingbar | powerchecker |
| YOLOv7 | 36.50 | 87.8 | 75.7 | 75.7 | 85.4 |
| YOLOv7+CAM | 36.54 | 89.0 | 76.6 | 77.8 | 85.7 |
| YOLOv7+SAM | 37.03 | 88.4 | 77.4 | 76.9 | 86.0 |
| YOLOv7+LCA(ours) | 36.51 | **89.3** | 77.5 | 80.1 | 87.0 |
| YOLOv5 | 7.02 | 82.8 | 65.7 | 67.1 | 78.5 |
| YOLOv5+CAM | 6.69 | 83.7 | 65.9 | 68.8 | 81.2 |
| YOLOv5+SAM | 6.69 | 83.7 | 66.2 | 68.4 | 80.8 |
| YOLOv5+LCA(ours) | 6.69 | **84.0** | 66.0 | 68.5 | 82.4 |

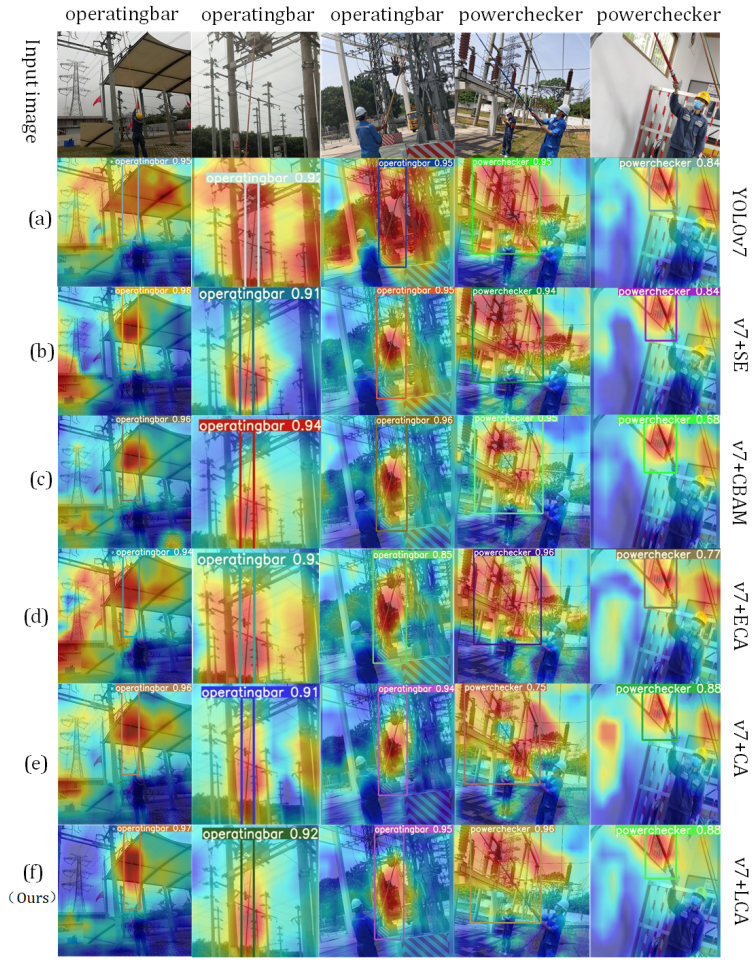


Figure 7: Comparative visualization of the activation heatmaps for elongated targets

Table 5: Evaluation results of LCA on a subset of the COCO dataset

| CNN model | Param(M) | mAP@.5(%) | Challenging samples mAP@.5(%) | | | |
|---|---|---|---|---|---|---|
| | | | fork | skis | baseball bat | snowboard |
| YOLOv7 | 36.50 | 73.7 | 67.9 | 63.9 | 79.1 | 65.4 |
| **YOLOv7+LCA(ours)** | 36.50 | 74.9 | 69.4 | 64.5 | 81.5 | 68.6 |
| YOLOv7+ECA | 36.50 | 74.3 | 69.0 | 62.7 | 81.1 | 66.6 |
| YOLOv7+CA | 36.53 | 73.6 | 67.6 | 64.7 | 76.3 | 67.6 |
| YOLOv7+CBAM | 36.54 | 73.4 | 68.5 | 63.4 | 77.3 | 67.0 |

### 5.3. Validate the Model's Performance on the COCO Dataset

To evaluate the model performance, experiments were conducted on four elongated object categories from the COCO dataset: skis, snowboard, fork, and baseball bat. The comparison of LCA with other attention mechanisms is presented in Table 5. The small target "fork" improved by 1.5%, while the

elongated target "snowboard" improved by 3%. A comparative analysis of attention mechanisms ECA, CA, and CBAM on the COCO dataset showed that LCA achieved the best performance on difficult samples while also attaining the highest overall accuracy.

## 6. Conclusion

In this paper, we address the issue of improving deep learning models for substation power operation by proposing a method that utilizes heatmaps for guidance from an interpretability perspective. The attention mechanism LCA, designed for challenging samples of elongated objects with extreme aspect ratios, resulted in a significant increase of 1.6% and 4.4% in mAP@.5 for the "operatingbar" and "powerchecker" categories, respectively. This clearly demonstrates the effectiveness of this attention mechanism in improving the detection of elongated targets.

## Acknowledgments

We thank a bunch of people and funding agency. This work was supported by Beijing Nature Foundation (No. L221015) , the Academic Research Projects of Beijing Union University (No.ZK20202302) and the Theme Case Project of Degree Center of the Ministry of Education of the People's Republic of China (ZT-231141703).

## References

Gaurab Bhattacharya, Bappaditya Mandal, and Niladri B Puhan. Multi-deformation aware attention learning for concrete structural defect classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3707–3713, 2020.

Xiaojin Gong, Qi Yao, Menglin Wang, and Ying Lin. A deep learning approach for oriented electrical equipment detection in thermal images. *IEEE Access*, 6:41590–41597, 2018.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Hongyang Jiang, Jie Xu, Rongjie Shi, Kang Yang, Dongdong Zhang, Mengdi Gao, He Ma, and Wei Qian. A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1560–1563. IEEE, 2020.

Mo Beibei & Wu Kehe. Research on wearable intelligent detection technology for power operation violations introducing self-attention. *Computers and modernization*, (115-121+126), 2020. ISSN 1006-2475.

Bin Li, Shuang Wu, Shengjie Wang, and Liang Zhang. Automatic inspection of power system operations based on lightweight neural network. In *2022 7th Asia Conference on Power and Electrical Engineering (ACPEE)*, pages 2161–2165. IEEE, 2022.

Jianhua Ou, Jianguo Wang, Jian Xue, Jianping Wang, Xian Zhou, Lingcong She, and Yadong Fan. Infrared image target detection of substation electrical equipment using an improved faster r-cnn. *IEEE Transactions on Power Delivery*, 38(1):387–396, 2023. doi: 10.1109/TPWRD.2022.3191694.

Jishen Peng, Chang Wang, Yang Li, and Hongtian Chen. Substation personnel smoking detection based on ghostnetv2-yolov5. In *2023 6th International Symposium on Autonomous Systems (ISAS)*, pages 1–6. IEEE, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021.

Pu Tianjiao, Qiao Ji, Zhao Zixuan, et al. Research on interpretable methods of machine learning applied in intelligent analysis of power system (part i): basic concept and framework [j/ol]. *Proceedings of the CSEE*, 43(18):7010–7029, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhaoyi Wan, Yimin Chen, Sutao Deng, Kunpeng Chen, Cong Yao, and Jiebo Luo. Slender object detection: Diagnoses and improvements. *arXiv preprint arXiv:2011.08529*, 2020.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

Sheng Wang. Substation personnel safety detection network based on yolov4. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 877–881. IEEE, 2021.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Changdong Wu, Yanliang Wu, and Xu He. Infrared image target detection for substation electrical equipment based on improved faster region-based convolutional neural network algorithm. *Review of Scientific Instruments*, 95(4), 2024.

Chunyang Xia, Zengxi Pan, Zhenyu Fei, Shiyu Zhang, and Huijun Li. Vision based defects detection for keyhole tig welding using deep learning with visual explanation. *Journal of Manufacturing Processes*, 56:845–855, 2020.

Yaohui Xiao, An Chang, Yufeng Wang, Yu Huang, Junsong Yu, and Lihai Huo. Real-time object detection for substation security early-warning with deep neural network based on yolo-v5. In *2022 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*, pages 45–50. IEEE, 2022.

Fayu Yang, Xuejiao Duan, Jia Wang, Yueming Wang, and Hong-an Zhang. Research and application of target detection algorithm for live operation in substation. In *Journal of Physics: Conference Series*, volume 2703, page 012038. IOP Publishing, 2024.

Na Zhang, Gang Yang, Dawei Wang, Fan Hu, Hua Yu, and Jingjing Fan. A defect detection method for substation equipment based on image data generation and deep learning. *IEEE Access*, 2024.

Tao Zhou, Qian Huang, Xiaolong Zhang, and Yong Zhang. Intelligent detection method of substation environmental targets based on md-yolov7. *Journal of Intelligent Learning Systems and Applications*, 15(3):76–88, 2023.