Electronic Research Archive

http://www.aimspress.com/journal/ERA

ERA, 32(9): 5321–5340.
DOI: 10.3934/era.2024246
Received: 19 July 2024
Revised: 03 September 2024
Accepted: 11 September 2024
Published: 18 September 2024

Research article

Semantic segmentation of substation tools using an improved ICNet network

Guozhong Liu[1,*], Qiongping Tang[1], Changnian Lin[2,3], An Xu[1], Chonglong Lin[2,3], Hao Meng[1], Mengyu Ruan[2,3] and Wei Jin[2,3]

[1] School of Instrument Science and Opto-Electronics Engineering, Beijing Information Science and Technology University, Beijing 100096, China
[2] Beijing Kedong Electric Control System Co., Ltd., Haidian District, Beijing 100192, China
[3] NARI Group Corporation (State Grid Electric Power Research Institute), Nanjing 211106, China

* **Correspondence:** Email: liuguozhong@bistu.edu.cn.

Abstract: In the field of substation operation and maintenance, real-time detection and precise segmentation of tools play an important role in maintaining the safe operation of the power grid and guiding operators to work safely. To improve the accuracy and real-time performance of semantic segmentation of substation operation and maintenance tools, we have proposed an improved, lightweight, real-time, semantic segmentation network based on an efficient image cascade network architecture (ICNet). The network uses multiscale branches and cascaded feature fusion units to extract rich multilevel features. We designed a semantic segmentation and purification module to deal with redundant and conflicting information in multiscale feature fusion. A lightweight backbone network was used in the feature extraction stage at different resolutions, and a recursive gated convolution was used in the upsampling stage to achieve high-order spatial interactions, thereby improving segmentation accuracy. Due to the lack of a substation tool semantic segmentation data set, we constructed one. Training and testing on the data set showed that the proposed model improved the accuracy of tool detection while ensuring real-time performance. Compared with the currently popular semantic segmentation network, it had better performance in real-time and accuracy, and provided a new semantic segmentation method for embedded platforms.

Keywords: ICNet; lightweight; semantic segmentation; tools and instruments; substation operation and maintenance

1. Introduction

At this stage, the construction of the energy internet is constantly upgrading and speeding up. New equipment, new technologies, new systems, and new operation management models are constantly being promoted and applied in substations. It poses new challenges to the level of operational ability and behavioral standardization. A good, real-time, perception ability of tools will help people and tools to interact better, which is conducive to the smooth development of power substation operation and maintenance. Many researchers have conducted related studies. For instance, context-aware recognition methods can identify people, tools, and actions together [1]. Multiview anomaly detection provides complementary information for a single instance [2]. A multigranularity cross-domain alignment framework can enhance the generalization ability of anomaly segmentation [3]. In order to meet multiple functional requirements in the process of substation operation and maintenance at the same time, real-time perception of tool objects is particularly important. Semantic segmentation can assign category labels to each pixel of the image, and can perform scene understanding and object detection. It is an integral part of the perception and recognition of autonomous mobile systems [4]. In the field of power substation operation and maintenance, it can be applied to on-site workpiece defect detection, safety monitoring, recognition of body movements of operation and maintenance personnel, and tool perception. Early traditional semantic segmentation used low-level features such as image grayscale, spatial texture, color, and geometric shape to segment the image into different parts. There are mainly threshold-based [5], edge-based [6], region-based [7,8], and graph-based segmentation methods [9], etc. Although these methods have fast segmentation speed, they need to manually design feature extractors, and the segmentation effect on complex scenes is not good. In contrast, semantic segmentation methods based on deep learning have shown strong feature extraction capabilities. Long et al. [10] proposed a fully convolutional network (FCN) in 2014, replacing the fully connected layer in the convolutional neural network (CNN) with a fully convolutional layer to achieve pixel-level dense prediction, laying the foundation for the rapid development of semantic segmentation [11]. However, an FCN network with a lightweight backbone has insufficient receptive fields on large objects, but has oversized receptive fields on small objects. This will lead to incomplete segmentation of large objects and coarse segmentation of small objects, which would affect the segmentation accuracy [12]. Therefore, subsequent models such as PSPNet and DeepLabv2 introduced pyramid pooling modules and dilated convolutions to extract multiscale context information and improve accuracy [13,14]. U-Net uses an encoder-decoder structure to capture context information and restore the position information of the original image [15], gradually restore object details and image resolution, and improve segmentation accuracy. The network structure of semantic segmentation has high accuracy, but the network parameters are large and the model is complex [16,17]. Therefore, many semantic segmentation models with outstanding performance cannot be well deployed on mobile platforms with limited computing resources. Traditional semantic segmentation cannot meet the real-time perception of images of power substation operation and maintenance tools. Considering the accuracy and real-time performance of the entire network of power substation operation and maintenance tools, it is particularly important to build a lightweight semantic segmentation network [18–20]. Paszke proposed ENet, which uses lightweight convolution blocks and simple upsampling to form an asymmetric codec structure [21]. Although this simple codec structure was fast, its accuracy was poor. Therefore, ERFNet was improved on the basis of Enet. It converted the standard convolution into an asymmetric convolution combination and removed the skip connection between the codec layers, while maintaining high semantic segmentation accuracy, it improves the computational efficiency and robustness of the model [22]. DANet used the self-attention mechanism to improve the model's perception of contextual information at different scales [23]. Li and Wu built a lightweight semantic segmentation model for parallel feature processing of road scenes for autonomous driving, mainly using the MobileNetV2 backbone network, combined with the dual attention mechanism and the empty convolution space pyramid module. The results proved that the the model has both real-time performance and high precision [24]. Therefore, we can conclude that in order to improve the reasoning speed of the semantic segmentation network, a lightweight feature extraction network and efficient convolution can be used to meet the real-time requirements of the model; at the same time, an attention module can be introduced to achieve real-time semantic segmentation balance of accuracy and segmentation speed. ICNet used a pyramid pooling module to fuse multiscale context information, and divided the network structure into three branches: low resolution, medium resolution, and high resolution [25]. It used low resolution to complete semantic segmentation, and a high-resolution strategy to refine the segmentation results and improve the segmentation accuracy of the model. The feature maps extracted by each branch were fused by the CFF module, and finally the segmentation result was obtained by upsampling. In addition, it used cascaded labels to guide the training of each branch, which speeded up the convergence and prediction of the model, improved real-time performance, and had good performance in the segmentation of remote sensing images [26], street view images and lesion images [27], but so far, there is no application in substation operation and maintenance. In this paper, we integrate the attention mechanism into the ICNet network to solve the problem of semantic segmentation of tools for power substation operation and maintenance. The main contributions of this paper are as follows:

- A self-built semantic segmentation dataset of substation electroscope pens. The PP-LCNet [28] lightweight network is used in the low, medium, and high resolution feature extraction parts of ICNet.
- The construction of a semantic segmentation and purification module including spatial attention and channel attention, which is added to the CFF module of ICNet to filter out redundant and conflicting information brought about by the fusion of different scale features.
- The introduction of a recursive gated convolution in the upsampling stage of the ICNet model, which extends the second-order interaction in self-attention to any order, and improves the semantic segmentation effect.

First, we introduce the semantic segmentation architecture and models related to our approach in Section 2. The details of the proposed model and method are then presented in Section 3. Then, we discuss the results in Section 4 and perform the corresponding analysis. Finally, conclusions are drawn in Section 5.

2. Related works

In this section, we introduce the ICNet model related to our method, as well as the relevant content of the efficient convolution and attention modules we adopt.

2.1. ICNet model

ICNet is a lightweight semantic segmentation network with fast detection speed and low memory consumption, which meets the characteristics of strict real-time requirements and low hardware conditions in the segmentation of power substation operation and maintenance tools, and divides the network structure into three branches: low resolution, medium resolution, and high resolution. The input of the low resolution branch was 1/4 of the size of the original image. After passing through the heavy CNN backbone network, the pyramid pooling module was used to increase the receptive field to obtain features whose size was 1/32 of the original image. The image was designed to extract the semantic information of the entire image. The input of the medium-resolution branch was 1/2 of the size of the original image, and the output feature map of the low-resolution branch was fused with the CFF module after passing through the light CNN backbone network to obtain the size of the original image 1/16 feature map. At the same time, the parameters were shared between the middle- and low-resolution image branches to reduce the execution time. The high-resolution branch only needed to focus on fine features. The original image was used as the input of the high-resolution branch, and after passing through the light CNN backbone network, it passed through the CFF module together with the output feature map of the medium-resolution branch to obtain a feature map whose size was 1/8 of the original image. After three times of upsampling, it was expanded to the size of the original image to obtain the final segmented image.

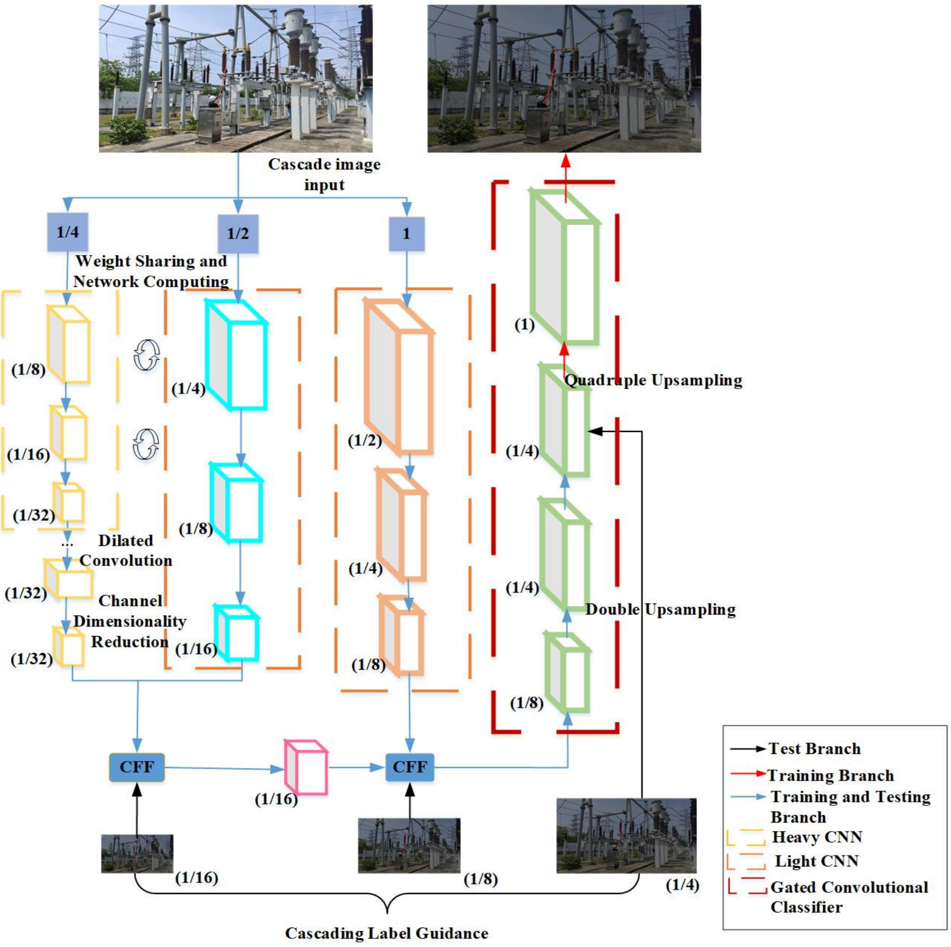


Figure 1. Structure of the ICNet model.

2.2. Efficient convolution

The key point of real-time lightweight semantic segmentation is to achieve the balance between speed and precision, which is usually achieved by efficient convolution. The efficient convolution in common use now includes depth-separable convolution and grouping convolution.

Depth-separable convolution splits the convolution kernel of traditional convolution into channel-by-channel convolution and point-by-point convolution. An *n*n* convolution kernel of channe-by-channel convolution is responsible for just one channel, and a channel can only be convolved by one convolution kernel. The output feature graph of channel-by-channel convolution is taken as the input of point-by-point convolution, and the final output graph is obtained by *N 1*1* convolution kernels. The number of parameters required for a depth-separable convolution is only *1/N + 1/n2* of the standard convolution. Therefore, on the premise of maintaining the accuracy, the number of network parameters and computation amount can be greatly reduced, and the efficiency of the network computation can be improved. While improving the performance, a better balance between precision and speed can be taken into account, which is more conducive to deployment in the embedded platform with low computing power.

Grouping convolution divides the input feature graph and the convolution kernel into *m* groups along the direction of the channel, and each group is convolved separately. The number of parameters required by grouping convolution is only *1/m* of the standard convolution. In addition, the coupling between the results obtained by grouping convolution is low, and the learning ability of the model can be further improved.

2.3. Attention module

To improve the performance of the semantic segmentation model, attention modules can be used. An attention module can be divided into spatial attention and channel attention by weighting the target information in the image. Channel attention mainly focuses on the meaningful information in a single image, while spatial attention mainly focuses on the temporal feature changes between adjacent frames or compared images.

SENet proposes that the SE attention module compresses the spatial dimension of the input feature graph by global averaging pooling to obtain the global features on the channel [29]. Then it gets the weight of different channels according to the two fully connected layers to learn the relationship between each channel. Finally, the normalized weight is weighted to the features of each channel, which helps to focus on the channel with the most information and suppress the unimportant channel information. However, the SE module ignores the spatial information, which can describe the size and position of the object in the input image, and obtain the information where the segmentation target is. This is also crucial to the semantic segmentation task. The CBAM module is composed of a spatial attention module (SAM) and channel attention module (CAM). CAM uses global average pooling and global maximum pooling at the same time to reduce information loss caused by pooling [30]. SAM performs average pooling and maximum pooling operations along the channel axis, and generates spatial attention feature maps through the convolution layer after the two are combined, which conduces to achieve the target positioning.

3. Methods

3.1. Structure of the EICNet module

As shown in Figure 2, the EICNet model in this paper is developed based on the ICNet model. Unlike ICNet, which only uses light CNN for medium- and high-resolution branches, and uses heavy CNN for low-resolution branches. The EICNet model uses a lightweight network PP-LCNet as the backbone network of the three branches of low, medium, and high resolution to reduce the number of parameters of the model. It adds the semantic purification module SCM after the CFF module to solve the contradiction between different semantics and to make up for the lack of CFF, and the convolution of the above sampling part is replaced by recursive gated convolution. Any order of spatial interaction is carried out through gated convolution and recursive design, so as to further improve the accuracy of model segmentation and complete the initial feature extraction of the model.

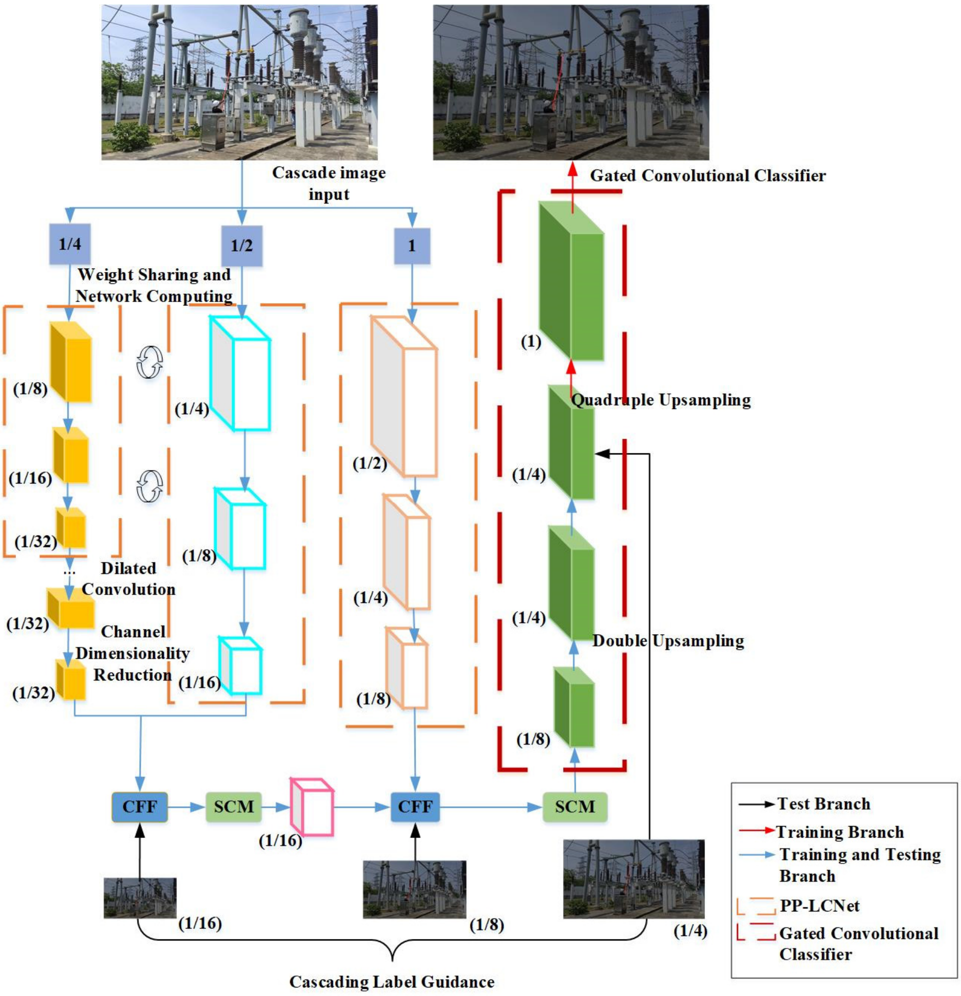


Figure 2. Structure of the EICNet model. "CFF" stands for the cascade feature fusion detailed in ICNet, "SCM" stands for the semantic segmentation and purification module detailed in Section 3.3, and the numbers in parentheses are feature map size ratios to the full resolution input. The gated convolution classifier represents the adoption of the gated convolution mentioned in Section 3.4.

3.2. Initial feature extraction based on the PP-LCNet backbone network

The PP-LCNet backbone network is a lightweight and high-performance convolutional neural network, and its basic module is deep separable convolutional, which has a small number of parameters and a small amount of computation work. The module can be deeply optimized by the Intel CPU acceleration library, and its inference speed is significantly faster than other lightweight convolution modules. Therefore, the EICNet model proposed in this paper is based on the PP-LCNet backbone network to extract initial features, and the working principle is shown in Figure 3. As can be seen from Figure 3, PP-LCNet adopts MobileNetV3 to make the distinction boundary more obvious, and replaces the ReLU activation function with h-swish, thus the fitting is faster and the training effect is significantly improved [31]. In the last layer of the depth-separable convolution module, an SE module is introduced to enhance the splendid features and suppress the unnecessary features. At the end of the model, the 3*3 convolution kernel is replaced by a 5*5 convolution kernel to increase the receptive field. The above operations do not increase much computing overhead, and they greatly reduce the number of operations and delay linearity, so that the model can not only take into account accuracy and real-time performance, but also become more lightweight and more suitable for the real-time perception of substation operation and maintenance equipment.

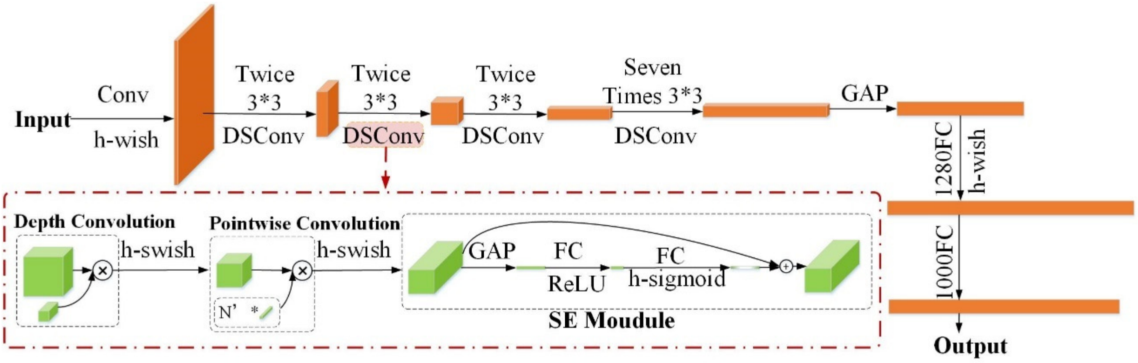


Figure 3. Overall structure of the PP-LCNet model.

3.3. Semantic segmentation and purification module SCM

The semantic difference of different scale features cannot be ignored. As a result, the direct fusion of different scale features will lead to a large amount of redundant and conflicting information, which will reduce the expression ability of multiscale features. In this paper, the semantic segmentation and purification module SCM is introduced after CFF to filter the noise generated after processing at different scales, as shown in Figure 4. SCM consists of two modules: the spatial attention module and the channel attention module. (1) The spatial attention module divides the feature maps into two groups after input. The first group obtains the target location information through the average pooling compression channel dimension. After feature extraction through the convolutional layer, it multiplies the original feature map by the sigmoid nonlinear activation layer to obtain a new feature map. The other group, after normalization, convolution, and nonlinear activation layer operation, is multiplied with the original feature map to obtain a new feature map for optimizing its positioning information [32]. (2) The channel attention module focuses on the importance of each channel and identifies meaningful information in segmentation. After the input feature map is pooled to the maximum and the average, the dimensionality reduction and dimensionality elevation of the attention module of SE are performed. After adding the two results, the new feature map is obtained by multiplying the non-linear activation and the original feature map [33]. In addition, compared with other attention modules, SCM also introduces grouping convolution, which facilitates the model to be additionally lightweight.

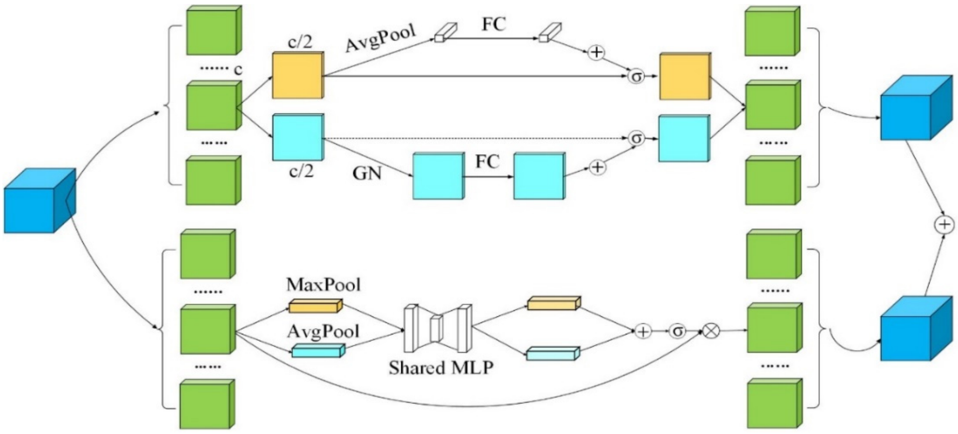


Figure 4. Semantic segmentation and purification module SCM. The upper part of the picture is the spatial attention module, the lower part of the picture is the channel attention module, "c" represents the number of channels, "AvgPool" represents average pooling, "FC" represents full connection, "GN" represents group normalization, "MaxPool" represents the maximum pooling, and "shared MLP" can be seen in SENet [29].

3.4. Gated convolution g_nConv

There are three main advantages of g_nConv. First, it is effective. The convolution-based implementation avoids the quadratic complexity of self-attention. The design of increasing the channel width gradually during the execution space interaction can realize the high-order interaction with limited complexity. Second, it is extensible. g_nConv extends the two-order spatial interaction in self-attention to any order, which can further improve the modeling ability, and is compatible with various scales of convolution kernel and different spatial mixing strategies. Third, it has translation equivariance. g_nConv fully inherits the translation equivariance of standard convolution, introducing beneficial inductive bias to the primary visual task and avoiding the asymmetry caused by local attention. Therefore, in this paper, gated convolution g_nConv is selected for sampling convolution in the EICNet model. As can be seen from Figure 5, g_nConv consists of three parts: a linear projection fully connected layer, a depth-separable convolution layer, and element multiplication, which improves the network capability by realizing high-order spatial interaction [34]. Assuming that the input feature of the gated convolution is $x \in R^{H\times W\times C}$, the implementation of $y = g_nConv(x)$ mainly involves the following four key steps:

Step 1: The dimension of x is increased to n times by the fully connected layer, and $p_0, q_0, ..., q_{n-1}$ is obtained. Slice x by channel through the fully connected layer to get $p_0, q_0, ..., q_{n-1}$, to achieve the purpose of dimension raising.

$$\left[p_0^{H\times W\times C_0}, q_0^{H\times W\times C_0}, \dots, q_{n-1}^{H\times W\times C_{n-1}}\right] = \phi_{in}(x) \in R^{H\times W\times (C_0+\sum_{0\leq k\leq n-1} C_k)} \tag{1}$$

In this formula, H and W represent the characteristic height and width, respectively, C is the number of characteristic channels, and φ_{in} is the input fully connected layer. Then split $p_0, q_0, \dots, q_{n-1}$ into two parts p_0 and $q_0, q_1, \dots, q_{n-1}$.

Step 2: $q_0, q_1, \dots, q_{n-1}$ pass the depthwise separable convolution layer.

$$q_k' = f_k(q_k) \in R^{H\times W\times C_k} \quad k = 0,1,\dots,n-1 \tag{2}$$

Step 3: Dot multiply and p_k, respectively, and the p_k interaction order will be increased by 1 in each step to achieve the purpose of spatial interaction.

$$p_{k+1} = q_k{}' \odot \frac{g_k(p_k)}{\alpha} \in R^{H\times W\times C_k} \;\; k = 0,1,\dots,n-1 \tag{3}$$

In this formula, g_k is the fully connected layer, being used to achieve the purpose of dimension raising. α is the scaling factor. The output features in each operation are scaled according to $1/\alpha$ to contributing to the smooth training.

It is necessary to notice that p_{k+1} after the dot product is still divided by C_k of the last iteration dimension. Therefore, it is needed to raise the dimension to C_{k+1} for the next iteration. The formula of g_k is:

$$g_k = \begin{cases} linear(C_0, C_0) & k = 0 \\ linear(C_{k-1}, C_k) & 1 \leq k \leq n-1 \end{cases} \tag{4}$$

To ensure that the high-order interactive operation does not introduce too much operation cost, the C_k is constrained and its concrete value can be calculated by Eq (5).

$$C_k = \frac{C}{2^{n-1-k}} \tag{5}$$

In this formula, $n-1-k$ represents the dimension of C from small to large.

Step 4: Map the spatial interaction result to the specified dimension through the full connection layer to get the final g_nConv result.

$$y = \phi_{out}(p_{n-1}) \in R^{H\times W\times C_{n-1}} \tag{6}$$

In this formula, ϕ_{out} is the output fully connected layer.

The above design shows that g_nConv performs interactions in a coarse to fine manner. Fewer channels are used to calculate lower-order spatial interactions, and higher-order features need to be mapped to higher-latitude features to learn richer patterns. The overall computational complexity of the gated convolution g_nConv can be divided into three pieces. They are linear projection, depthwise separable convolution, and a recursive gate. Linear projection: mainly ϕ_{in} and ϕ_{out}, because the default is that ϕ_{in} is upgraded to the original $2C$. ϕ_{out} is a map of the same dimension. Therefore, the computational complexity of the linear projection is shown in the following formula:

$$FLOPs(\phi_{in}) = 2 \times H \times W \times C^2 \tag{7}$$

$$FLOPs(\phi_{out}) = H \times W \times C^2 \tag{8}$$

Explained by ϕ_{in}, the linear mapping on each pixel is $2 \times C^2$, and there are a total of $H \times W$ pixels, so the overall computational complexity is $2 \times H \times W \times C^2$.

Depth-separable convolution: The scale of the convolution kernel in depth-separable convolution is K. Then the computational complexity of depth-separable convolution is shown in the following formula:

$$FLOPs(DSConv) = H \times W \times K^2 \times \sum_{k=0}^{n-1} \frac{C}{2^{n-1-k}} = 2 \times H \times W \times C \times K^2 \times (1 - \frac{1}{2^n}) \quad (9)$$

In this formula, K is the convolution core scale in depth-separable convolution. With a recursive gate, the operation includes the dot multiplication and function g_k.

$$FLOPs(g_k) = H \times W \times \sum_{k=0}^{n-2} \frac{C}{2^{n-1-k}} \frac{C}{2^{n-2-k}} = \frac{2}{3} H \times W \times C^2 \times (1 - \frac{1}{4^{n-1}}) \quad (10)$$

$$FLOPs(Mul) = H \times W \times \sum_{k=0}^{n-1} \frac{C}{2^{n-1-k}} = 2 \times H \times W \times C \times (1 - \frac{1}{2^n}) \quad (11)$$

$$FLOPs(Recursivegate) = FLOPs(g_k) + FLOPs(Mul) = H \times W \times C \times (\frac{2}{3} \times C \times (1 - \frac{1}{4^{n-1}}) + 2 - \frac{1}{2^{n-1}}) \quad (12)$$

Ultimately, the computational complexity of g_nConv is,

$$\begin{aligned} FLOPs(gnConv) &= FLOPs(\phi_{in}) + FLOPs(\phi_{out}) + FLOPs(DSConv) + FLOPs(Recursivegate) \\ &= H \times W \times C \times (C \times (\frac{11}{3} - \frac{2}{3} \times \frac{1}{4^{n-1}}) + 2K^2 \times (1 - \frac{1}{2^n}) + 2 - \frac{1}{2^{n-1}}) \\ &< H \times W \times C(2 \times K^2 + \frac{11}{3} \times C + 2) \end{aligned} \quad (13)$$

In this formula, K is the convolution core scale in depth-separable convolution.

Recursive gated convolution can better capture the high-order interactions between features at different levels in dense prediction tasks, thus complementing missing details. For object detection and semantic segmentation tasks, it can significantly reduce the amount of parameters while achieving better segmentation accuracy.

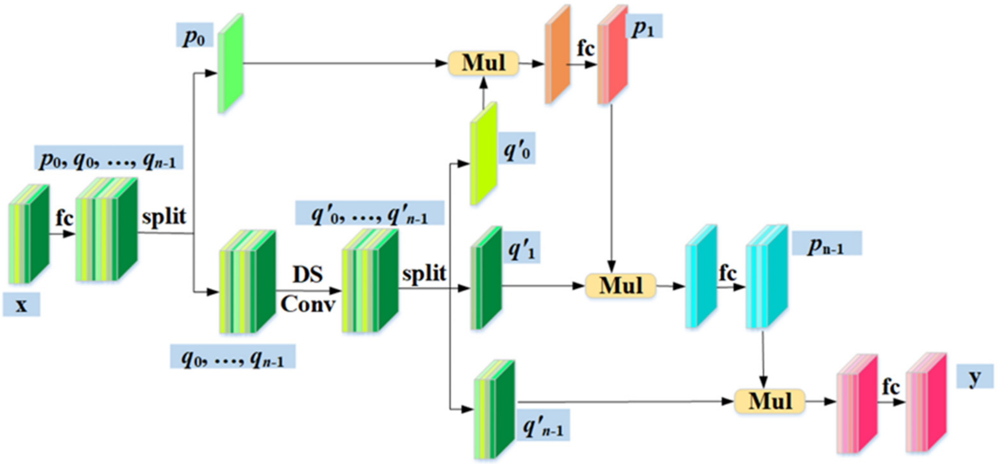


Figure 5. Gated convolutional module. “FC” stands for full connection, “split” stands for segmentation by channel, “DS Conv” stands for depth-separable convolution, and “Mul” stands for point multiplication.

4. Design and analysis of experiments

4.1. Semantic segmentation dataset

To verify the practical performance of the network, this paper utilizes the electroscopic pens as the segmentation target and constructs an on-site, captured, semantic segmentation dataset of substation tools (electroscopic pens). The image collection scenario is the actual substation operation and maintenance outdoor and control room single and multiple working scenes (763 in total), encompassing a wide range of real-world usage scenarios. This dataset is divided into 400, 109, and 254 images for training, validation, and testing, respectively, adequately meeting the requirements for practical experimentation and analysis, and will be expanded gradually as the number of tools to be detected increases. We use two data augmentation techniques during the experiments. The first method is random mirroring, while the second is random scaling by a factor between 0.5 and 2. These techniques aim to enhance the model's generalization capabilities, reduce the risk of overfitting, and improve the performance and robustness of the model. First of all, the original image of the existing target detection dataset is annotated with the semantic segmentation software Labelme to get json files, and then the json files are converted into the corresponding label graph. Part of the original figure and visual label figure of the dataset are shown in Figure 6. It can be seen from Figure 6 that the electroscopic pen accounts for a small proportion of the whole image as the segmentation target. So it can be seen that the image recognition of appliances has a high-precision requirement on the ability of model feature extraction and noise information filtering.

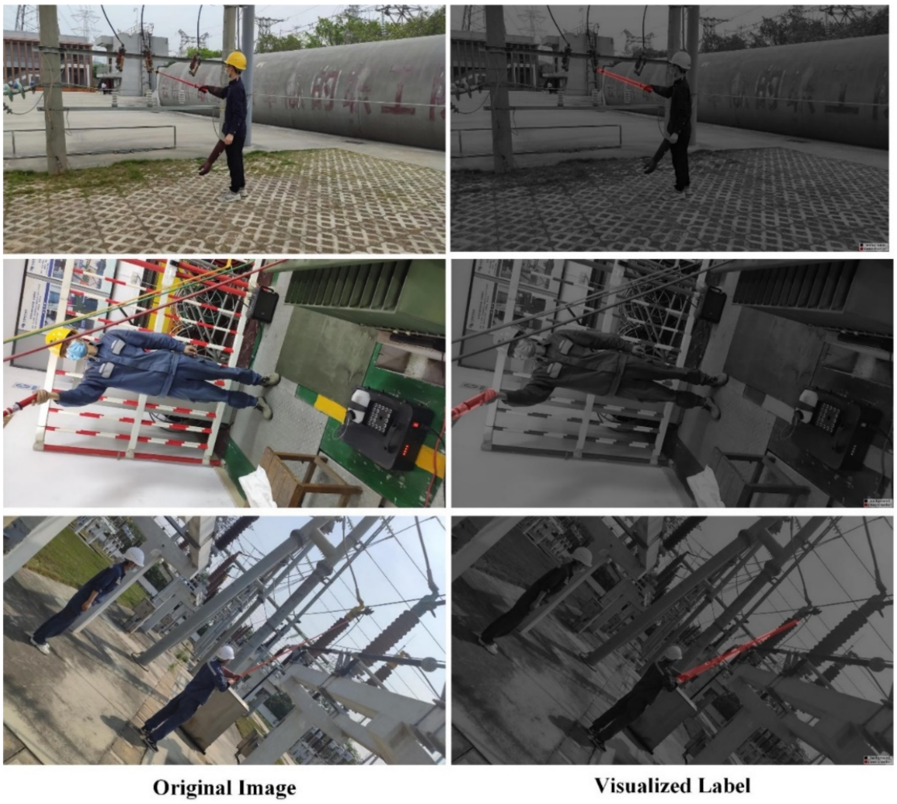


Figure 6. The dataset of tools and instruments for substation operation and maintenance.

4.2. Dataset experimental conditions and evaluation indicators

The operating environment of this experiment was Windows 11 Professional, the processor was Inter(R) Core(TM) i7-12700 CPU @ 2.1 GHz, the graphics card was NVDIA GeForce RTX 3090, and the video memory was 24 GB. The Pytorch deep learning framework is adopted, and the specific version is Pytorch 1.10+Python 3.8+Cuda 11.6. In the training process, the batch size was set as 16 and the basic learning rate as 0.01. A polynomial learning rate attenuation strategy was adopted, the momentum was 0.9, the maximum number of iterations was set as 30K, and the weight attenuation as 0.0001. In order to expand the difference between samples and ensure the generalization ability of the later model training, we randomly mirror and zoom the original image of the substation operation and maintenance tool by 0.5–2 times.

The evaluation indicators involved in the result analysis of this paper are mean intersection over union (mIoU), mean accuracy (mAcc), and frame per second (FPS). mIoU represents the similarity between segmentation results and real labels, mAcc calculates the average accuracy of all categories, and FPS measures the speed of model operation. The larger the three values are, the better the model effect is. The calculation formula is as follows:

$$mIoU = \frac{1}{N+1}\sum_{i=0}^{N} \left(\frac{TP}{TP+FP+FN}\right) \tag{14}$$

$$mAcc = \frac{1}{N+1}\sum_{i=0}^{N} \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \tag{15}$$

$$FPS = \frac{K}{T} \tag{16}$$

In this formula, *N*+1 is the number of categories added to the background. *TP* is a positive sample predicted to be a positive class. *TN* is a negative sample predicted to be a negative class. *FP* is the negative sample predicted to be positive. *FN* is a positive sample predicted to be a negative class. *K* is the number of inference pictures and *T* is the time that inference takes.

4.3. Experimental comparison of semantic segmentation models

4.3.1. Compared with the traditional semantic segmentation model

In order to verify the semantic segmentation effect of the optimization algorithm proposed in this paper on the substation operation and maintenance appliance, the algorithm and the traditional semantic segmentation ANN model [35] were respectively tested on the appliance semantic segmentation dataset, and the test results are shown in Table 1. According to the data in the table, in the EICNet model, mIoU is 69.86% and mAcc is 73.59%. In the ANN model, mIoU and mAccare are, respectively 69.93% and 72.43%. Compared with ANN, the mIoU of the EICNet model only decreases by 0.07% and the mAcc increases by 1.16%. The above indicates that the EICNet model greatly reduces the memory occupied by inference on the premise of basically maintaining the accuracy, and the memory required for inference is only 13.89% of that of ANN. At the same time, the inference speed of the EICNet model reaches 31.12 FPS, which is 8.41 times that of ANN. It can be seen that the adoption of the EICNet model can not only achieve the segmentation similarity of traditional semantic segmentation, but also remarkably reduce the memory occupied by inference and improve the running speed of the model, which is conducive to deployment in embedded platforms and meets the real-time requirements of substation operation and maintenance appliance identification.

Table 1. Comparison of network performance indicators between the ANN model and EICNet.

Algorithm	mIoU/%	mAcc/%	Memory/GB	FPS
ANN	69.93	72.43	9.50	3.70
Proposed	69.86	73.59	1.32	31.12

4.3.2. Compared with the lightweight semantic segmentation model

To verify the efficiency and superiority of the lightweight algorithm in this paper, EICNet and four other lightweight semantic segmentation models (CCNET [36], CGNET [37], DANET, and ERFNET) were further tested on the semantic segmentation dataset of substation operation and maintenance equipment. The results are shown in Table 2. According to the performance data in the table, the performance of the EICNet model is the best, with the largest mIoU and mAcc values, which are 69.86% and 73.59%. Meanwhile, the performance of the CGNet model is the worst, with the smallest mIoU and mAcc values (62.35% and 63.35%, respectively). Compared with the four lightweight semantic segmentation models, the mIoU of EICNet is 7.51% higher than that of CGNET, the memory occupied by inference is reduced by 84.09%, and the inference speed is slightly increased. Compared with ERFNET, it is 3.48% higher than ERFNET's mIoU, and the inference speed is 2.07 times that of ERFNET. In summary, the EICNet model network not only significantly improves accuracy, but also has better inference speed, owning strong competitiveness.

Table 2. Comparison of performance indicators between EICNet and other lightweight networks.

Algorithm	mIoU/%	mAcc/%	Memory/GB	FPS
CCNet	65.12	66.21	10.5	3.32
CGNet	62.35	63.35	8.30	30.51
DANet	66.78	68.17	7.40	2.66
ERFNet	66.38	68.36	6.04	15.28
Proposed	69.86	73.59	1.32	31.12

In order to further visually compare the performance of the EICNet model with other lightweight semantic segmentation models, typical substation operation and maintenance scenarios, such as an outdoor high-climbing operation, outdoor low-climbing operation, and control room inspection, are selected in this paper for different model tests. From Figure 7: (1) It can be seen from the first and third lines that, in the case of dim outdoor light and a complex background, only the EICNet model and CGNet model can accurately identify the area where the electroscope pen is located, while other lightweight models cannot detect the target for segmentation. Among them, the IoU for EICNet is 41.92%, whereas for CGNet, it is only 29.35%. The EICNet model can effectively identify the region from the hand part to the tip of the electroscopic pens. The recognition range of the CGNet model is small, so it can only extract 1/3 of the area of the main pole in the middle part of the electroscopic pens. (2) Lines two and five represent indoor training and actual work scenarios for substation maintenance personnel. Taking line two as an example, it can be seen that the EICNet model fits the area where the electroscopic pen is located, with clear boundaries in edge information segmentation, IoU is 58.68%, and Acc is 96.10%. The CCNet model can only recognize some parts of the electroscopic pen, with unclear segmentation boundaries (53.80%/95.71%). The CGNet model is more accurate than the

CCNet model in recognizing the starting position, but the segmentation area is discontinuous, with an IoU much lower than other models, only 24.2%, while the Acc is 93.16%. The DANet model (57.50%/96.02%) and the ERFNet model (53.10%/93.64%) can accurately recognize the starting position, but their dispersion range is too large. (3) Line four represents an outdoor low-climbing scene for substation maintenance. In this scenario, the EICNet model has more accurate positioning, with an IoU of 28.44% and an Acc of 96.20%. The CCNet model and ERFNet model still have poor segmentation fitting effects and cannot segment the specified area. The DANet model (11.56%/95.33%) can identify the tip and upper part of the electroscopic pen in bright background areas. The CGNet model performs well (21.52%/95.86%), but the starting position is not clear.

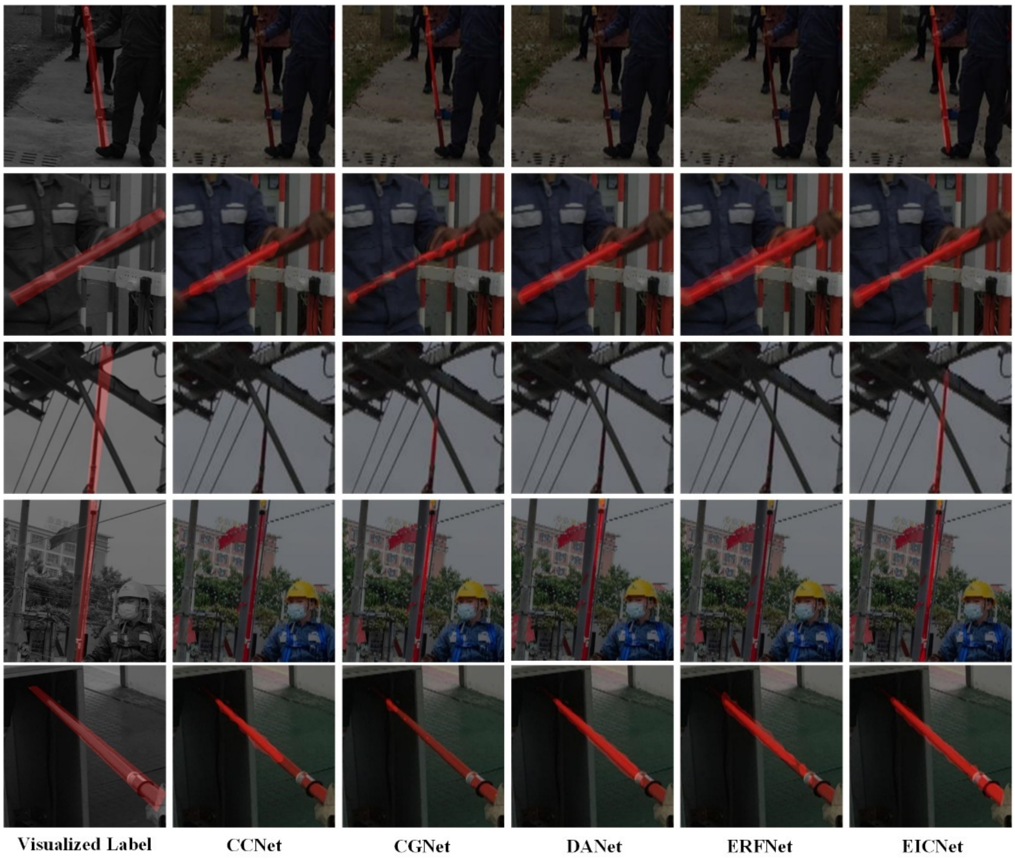


Figure 7. Semantic segmentation results of different networks.

Given all that, the CGNet, ERFNet, and DANet models have high requirements for environmental brightness, and the CGNet and ERFNet models have poor segmentation effect under complex background scenes. Under bright light conditions, the ERFNet and DANet models have nice segmentation effect, but there is a problem that the large dispersion area at the edge of the electroscope pen main pole cannot be defined. The CGNET model can adapt to dark scenes, but it has the problem of discontinuity of the segmentation area. The EICNet model can effectively remove the noise and obtain the target location information through the semantic cleaning module SCM. At the same time, gated convolution g_nConvis is introduced to effectively utilize high-order interactive information and improve the accuracy of the model. So, the EICNet model can adapt to different light conditions and relatively complex environments, so as to obtain a more delicate semantic segmentation effect of appliances.

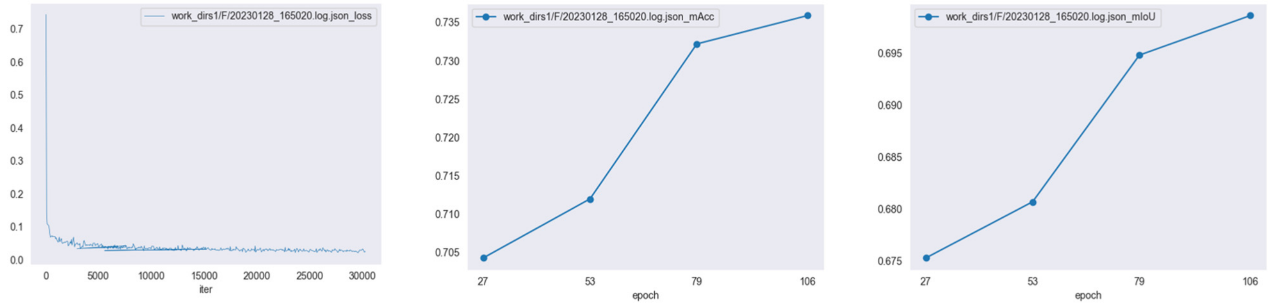
4.4. Ablation experiment

This structure uses the PP-LCNet backbone network as the initial feature extraction network. After the semantic purification module processes the microcontroller, noise is effectively removed to obtain target position information. The semantic purification module SCM is added to resolve conflicts between different semantics. The convolution in the above sampling section is replaced by recursive g_nConv, enabling spatial interactions of any order through g_nConv and recursive design, further improving the segmentation accuracy of the model and completing the initial feature extraction of the model. To validate the effects of these three modules in EICNet, six different module combinations were designed for ablation experiments and tested on a semantic segmentation dataset for substation maintenance equipment. The experimental results are shown in Table 3.

Table 3. Results of ablation experiments.

Method	**PP-LCNet**	**SCM**	**g_nConv**	**mIoU/%**	**mAcc/%**
A (ICNet)				66.75	68.60
B	√			67.06	68.81
C	√	√		67.48	69.86
D	√		√	68.02	70.20
E		√	√	68.68	71.42
F (Ours)	√	√	√	69.86	73.59

In Table 3, method A represents the original ICNet model without modification. Method B is that PP-LCNet is used to replace PSPNet in the feature extraction network. On the basis of method B, method C adds semantic purification module SCM while method D adds gated convolution g_nConv. Method E is to discuss the performance of semantic cleaning module SCM and gated convolutional g_nConv when the backbone network is still PSPNet, and method F is the EICNet model proposed in this paper. (1) Compared with ablation experiments of models A, B, E, and F, it is evident that using PP-LCNet as the feature extraction backbone network not only lightens the network but also enhances accuracy compared to PSPNet. Using PP-LCNet will increase mIoU by 0.31% and mAcc by 0.21%. In the case of using SCM and g_nConv, mIoU increased by 1.18% and mAcc increased by 2.17% after switching to the PP-LCNet backbone network. The overall improvement is approximately 0.745%/1.19%. (2) Comparing the ablation experiments of models B, C, D, and F, it can be concluded that using PP-LCNet, the SCM module improves the model's performance in terms of mIoU and mAcc by approximately 0.42%/1.05%. When using PP-LCNet and g_nConv, the use of the SCM module increases mIoU by 1.84% and mAcc by 3.39%. The overall improvement is approximately 1.13%/2.22%. (3) Comparing the ablation experiments of models B, C, D, and F, it can be concluded that using PP-LCNet, the g_nConv module improves the model's performance in terms of mIoU and mAcc by approximately 0.96%/1.39%. When using PP-LCNet and SCM, the use of the g_nConv module increases mIoU by 2.38% and mAcc by 3.73%. The overall improvement is approximately 1.67%/2.56%. (4) It can be seen from the above results that the enhancement effects of each module align with the design expectations. PP-LCNet, as a lightweight network, has the weakest effect, the SCM module primarily functions as a filter, with a better effect, and gnConv, which leverages higher-order interactive information to improve the model accuracy, has the most obvious effect.



(a) loss

(b) mAcc

(c) mIoU

Figure 8. The curve of loss, mAcc, and mIoU.

The segmentation prediction visualization is shown in Figure 9.

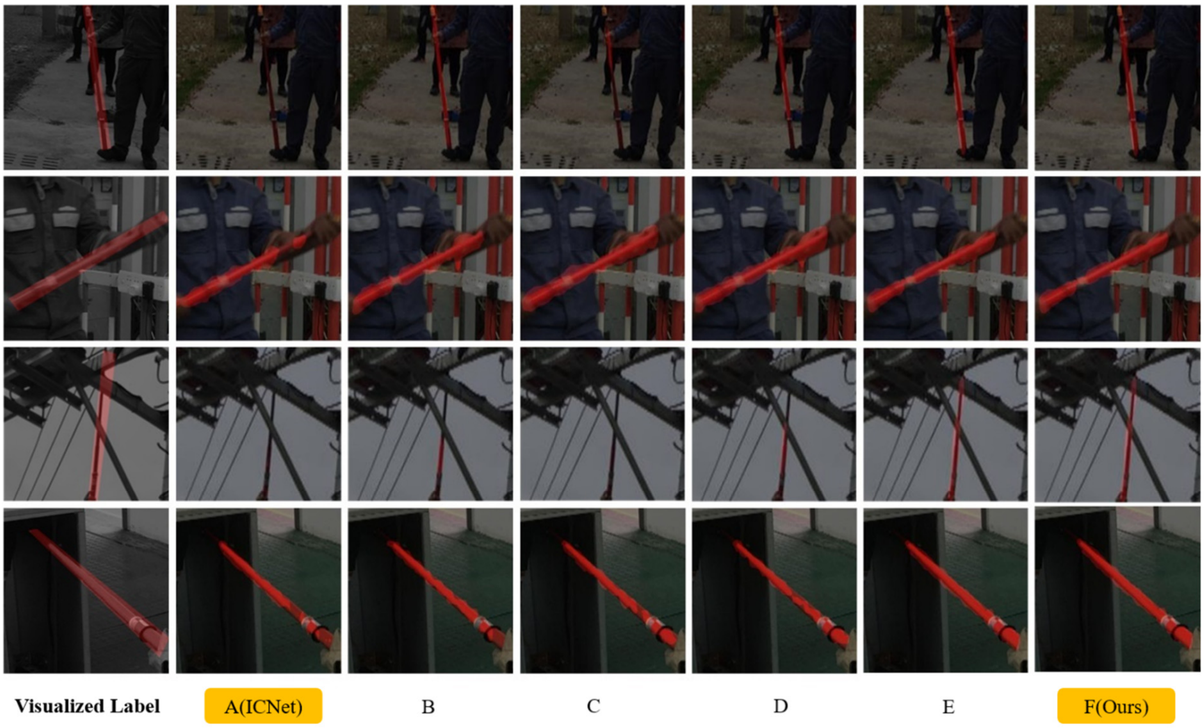


Figure 9. Semantic segmentation results of ablation experiments.

The first and third rows are scenes with dim outdoor light, and the second and fourth rows are scenes with bright indoor light. In the case of dim light and a complex background, ICNet can hardly identify the area where the electroscope is located. Using PP-LCNet instead of PSPNet can better lock the area where the electroscope is located. In bright indoor scenes, ICNet also has block missing phenomenon. The semantic segmentation module and gated convolution are used to repair the large-area segmentation errors existing in the original network, which can well segment out targets that occupy a small area in the image, and it can also distinguish railings that are close to the electrometer, similar in shape, and similar in color. The visualization results can intuitively see that the image semantic segmentation method proposed in this paper can obtain satisfactory image semantic segmentation results and improve the accuracy of image semantic segmentation.

5. Conclusions

In order to improve the accuracy and real-time performance of semantic segmentation for substation operation and maintenance appliances, a lightweight semantic segmentation EICNet model for appliances was designed in this paper. The PP-LCNet backbone network was used as the initial feature extract network in the whole structure, and the noise was effectively removed after processing by semantic purification module SCM, so that target location information could be obtained. Meanwhile, gated convolution g_nConv is introduced to utilize high-order interactive information and improve the accuracy of the model. While ensuring the real-time operation of the whole network, it can effectively reduce the number of network parameters and integrate multilayer feature information to improve the accuracy of the network.

The ablation experiments of different module combinations led to the following conclusions: PPLCNet instead of PSPNet as the feature extraction backbone network had better performance. The semantic purification module SCM can filter out the conflicting information of multiscale feature fusion. Gated convolution g_nConv can enable high-order interaction, thus improving the accuracy of the model. Therefore, when the PP-LCNet backbone network, SCM, and g_nConv module are exerted at the same time, the segmentation model mIoU and mAcc will respectively improve by 2.38% and 3.73% compared with the original ICNet model. The above conclusions show that in the EICNet model, the PP-LCNet backbone network with semantic purification module SCM and gated convolution g_nConv can meet the accuracy requirements of lightweight semantic segmentation of substation operation and maintenance equipment.

The EICNet model and the traditional semantic segmentation ANN model were utilized to segment and identify the single and multiple work image segmentation scenes in the actual substation operation and maintenance outdoor and control room. The comparison results indicated that compared with the ANN model, mIoU only decreased by 0.07% while mAcc increased by 1.16%. On the same hardware platform, we measured the model's inference speed, revealing that the improved model achieves an inference speed of 31.12 frames per second (FPS), which is 9.37 times faster than CCNet and surpasses the relatively efficient lightweight network CGNet (30.51 FPS). Additionally, EICNet outperforms CGNet with a 7.51% higher mIoU and a 10.24% higher mAcc, indicating a significant enhancement in accuracy. In conclusion, the EICNet model demonstrates fast inference speed, low memory footprint, and is capable of meeting real-time detection requirements.

Typical substation operation and maintenance scenarios (an outdoor high-climbing operation, outdoor low-climbing operation, control room inspection, etc.) were selected to test and compare different lightweight semantic segmentation models, and the research results show that the CCNET, CGNET, DANET, and ERFNET models have poor segmentation effect, since they cannot fully identify the electroscopic pen and have high requirements on the brightness of the environment. The EICNet model has the largest mIoU and mAcc values and the highest segmentation accuracy, and as a result, it can effectively identify the area from the hand part to the tip of the electroscopic pen in the substation operation and maintenance equipment scene with different light conditions and complex background environments, reaching the high-precision and real-time effect of semantic segmentation for substation operation and maintenance equipment.

This study has made some achievements in the lightweight semantic segmentation effect of substation operation and maintenance equipment, on the basis that the PP-LCNet backbone network EICNet model can meet the precision requirements of lightweight semantic segmentation of equipment. The future research work will start from small samples, increase the target types of substation operation and maintenance equipment, realize the lightweight semantic segmentation of small samples of digital substation operation and maintenance equipment, and provide a scientific theoretical basis for the subsequent safe operation and reasonable development of substation operation and maintenance.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the Science and Technology Project of State Grid Corporation of China, grant number 52272221001Z.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Z. Q. Cheng, Q. Dai, S. Li, T. Mitamura, A. Hauptmann, Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 3272–3281. https://doi.org/10.1145/3503161.3547943
2. H. Wang, Z. Q. Cheng, J. Sun, X. Yang, X. Wu, H. Y. Chen, et al., Debunking free fusion myth: Online multi-view anomaly detection with disentangled product-of-experts modeling, in *Proceedings of the 31st ACM International Conference on Multimedia*, (2023), 3277–3286. https://doi.org/10.1145/3581783.3612487
3. J. Zhang, X. Wu, Z. Q. Cheng, Q. He, W. Li, Improving anomaly segmentation with multi-granularity cross-domain alignment, in *Proceedings of the 31st ACM International Conference on Multimedia*, (2023), 8515–8524. https://doi.org/10.1145/3581783.3611849
4. S. Gupta, P. Arbeláez, R. Girshick, J. Malik, Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation, *Int. J. Comput. Vision*, **112** (2015), 133–149. https://doi.org/10.1007/s11263-014-0777-6
5. X. M. Zhang, Z. Y. Li, Y. Zheng, Multi-threshold image segmentation based on combining fisher criterion and potential function, *J. Comput. Appl.*, **32** (2012), 2843–2847. https://doi.org/10.3724/SP.J.1087.2012.02843
6. P. Liu, A. M. Yang, A method of region based color image segmentation, *Comput. Eng. Appl.*, **43** (2007), 37–39. https://doi.org/10.3321/j.issn:1002-8331.2007.06.012

7. C. Li, Z. Qu, Review of image edge detection algorithms based on deep learning, *J. Comput. Appl.*, **40** (2020), 3280–3288. https://doi.org/10.11772/j.issn.1001-9081.2020030314
8. J. Song, Y. Yu, Q. Luo, Cross-layer fusion feature based on richer convolutional features for edge detection, *J. Comput. Appl.*, **40** (2020), 2053–2058. https://doi.org/10.11772/j.issn.1001-9081.2019112057
9. S. J. Zhai, *Research on Image Segmentation Based on Optimization Theory*, Ph.D thesis, Hunan Normal University, 2018.
10. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
11. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. https://doi.org/10.1145/3065386
12. J. J. Qiao, Z. Q. Cheng, X. Wu, W. Li, J. Zhang, Real-time semantic segmentation with parallel multiple views feature augmentation, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 6300–6308. https://doi.org/10.1145/3503161.3547786
13. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2881–2890. https://doi.org/10.1109/CVPR.2017.660
14. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2017), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184
15. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Proceedings of the 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
16. C. Peng, T. Tian, C. Chen, X. Guo, J. Ma, Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation, *Neural Networks*. **137** (2021), 188–199. https://doi.org/10.1016/j.neunet.2021.01.021
17. Y. Liu, Z. Zhang, S. Pei, J. H. Wu, L. H. Liang, Z. R. Ma, Faulty insulator segmentation method in infrared image based on deep learning, *Electr. Meas. Instrum.*, **59** (2022), 63–68.
18. Z. Hu, S. Bao, C. Xu, H. Wang, Semantic segmentation algorithm for remote sensing buildings based on DeepLabv3+, *J. Comput. Appl.*, **41** (2021), 71–75.
19. X. Tang, W. Tu, K. Li, J. Cheng, DFFNet: An iot-perceptive dual feature fusion network for general real-time semantic segmentation, *Inf. Sci.*, **565** (2021), 326–343. https://doi.org/10.1016/j.ins.2021.02.004
20. Y. Wang, H. Liu, H. Wang, Y. Qian, Lightweight building semantic segmentation method based on remote sensing images, *Comput. Eng. Design*, **43** (2022), 2646–2653. https://doi.org/10.16208/j.issn1000-7024.2022.09.032
21. A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, preprint, arXiv:1606.02147. https://doi.org/10.48550/arXiv.1606.02147
22. E. Romera, J. M. Alvarez, L. M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, *IEEE Trans. Intell. Transp. Syst.*, **19** (2017), 263–272. https://doi.org/10.1109/TITS.2017.2750080

23. F. Xiong, X. Zhang, X. Han, L. Kuang, H. Liu, J. Jia, Research on improved semantic segmentation of remote sensing, *Comput. Eng. Appl.*, **58** (2022), 185–190. https://doi.org/10.3778/j.issn.1002-8331.2011-0021
24. S. Li, T. Wu, Lightweight semantic segmentation of road scenes for autonomous driving, *Comput. Eng. Appl.*, **59** (2023). https://doi.org/10.3778/j.issn.1002-8331.2206-0433
25. H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, Icnet for real-time semantic segmentation on high-resolution images, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 405–420. https://doi.org/10.1007/978-3-030-01219-9_25
26. S. Liu, H. Ye, K. Jin, H. Cheng, CT-UNet: Context-transfer-UNet for building segmentation in remote sensing images, *Neural Process. Lett.*, **53** (2021), 4257–4277. https://doi.org/10.1007/s11063-021-10592-w
27. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J. Garcia-Rodriguez, A review on deep learning techniques applied to semantic segmentation, preprint, arXiv: 1704.06857. https://doi.org/10.48550/arXiv.1704.06857
28. C. Cui, T. Gao, S. Wei, Y. Du, R. Guo, S. Dong, PP-LCNet: A lightweight CPU convolutional neural network, preprint, arXiv:2109.15099. https://doi.org/10.48550/arXiv.2109.15099
29. K. Zhou, Q. Yang, Y. Wang, J. Zhang, An improved SSD algorithm based on pressure plate status recognition, *Electr. Meas. Instrum,* **58** (2021), 69–76. https://doi.org/10.19753/j.issn1001-1390.2021.01.010
30. Q. Yao, S. Bie, J. Yu, Q. Chen, A bearing fault diagnosis method combining improved inception V2 module and CBAM, *J. Vib. Eng.*, **35** (2022), 949–957. https://doi.org/10.16385/j.cnki.issn.1004-4523.2022.04.019
31. H. Wang, X. Ge, Lightweight DeepLabv3+ building extraction method from remote sensing images, *Remote Sens. Natural Resour.*, **34** (2022), 128–135. https://doi.org/10.6046/zrzyyg.2021219
32. D. Liu, Z. Liang, Y. Sun, Micro-expression recognition method based on spatial attention mechanism and optical flow features, *J. Comput.-Aided Design Comput. Graphics*, **33** (2021), 1541–1552. https://dx.doi.org/10.3724/SP.J.1089.2021.18569
33. Z Lyu, X Xu, F Zhang, Lightweight attention mechanism module based on squeeze and excitation, *J. Comput. Appl.*, **42** (2022), 2353–2360. https://doi.org/10.11772/j.issn.1001-9081.2021061037
34. Y Rao, W Zhao, Y Tang, J Zhou, S. N. Lim, J. Lu, Hornet: Efficient high-order spatial interactions with recursive gated convolutions, preprint, arXiv:2207.1428v3.
35. Y. Liu, F. Zheng, B. Fan, TV news automatic segmentation base on text and audio-visual multi-modal features information, *Comput. Eng. Appl.*, **43** (2007), 190–194. https://doi.org/10.3321/j.issn:1002-8331.2007.35.057
36. P. Wang, L. Liu, H. Zhang, T. Wang, CGNet: A cascaded generative network for dense point cloud reconstruction from a single image, *Knowledge-Based Syst.*, **223** (2021), 107057. https://doi.org/10.1016/j.knosys.2021.107057
37. Q. You, W. Xu, K. Zhang, L. Zhang, X. Yi, D. Yao, C. Wang, et al., ccNET: Database of co-expression networks with functional modules for diploid and polyploid Gossypium, *Nuclc Acids Res.*, **45** (2017), D1090–D1099. https://doi.org/10.1093/nar/gkw910