Hindawi
Complexity
Volume 2022, Article ID 2468431, 14 pages
https://doi.org/10.1155/2022/2468431

Research Article

Insulator Semantic Segmentation in Aerial Images Based on Multiscale Feature Fusion

Zheng Cui, Chunxi Yang, and Sen Wang

Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China

Correspondence should be addressed to Chunxi Yang; ycx2003@163.com

Received 18 April 2022; Revised 15 May 2022; Accepted 26 May 2022; Published 1 July 2022

Academic Editor: Chun Wei

As one of the important components in the transmission line, the insulator is related to the safe and reliable operation of the entire transmission line. Aerial images are characterized by complex backgrounds, multiple pseudotargets, and low signal-to-noise ratios. Rapid and accurate localization of insulators in aerial images is a critical and challenging task in automatic inspection of transmission lines. Most insulator localization methods suffer from the loss of target edge detail information and large amount of model parameters. To solve these problems, this paper adopts an Encoder-Decoder architecture, called ED-Net, to realize end-to-end intelligent and accurate identification of insulators in aerial images. Firstly, Initial Module and CA-Bottleneck which are used to extract features from images to generate finer feature maps are proposed in the Encoder path. Meanwhile, global average pooling is used to preserve the maximum receptive field. Secondly, in the Decoder path, Refinement Boundary Module and Asymmetric Convolution Module are given to perform boundary optimization on the feature map, which are generated by the Encoder path. Finally, the Attention Feature Fusion Module is introduced into the Decoder path to combine high-level features with low-level features better and reduce the gap between features of different levels. The proposed model architecture keeps a suitable balance between the model parameters and insulator segmentation performance on insulator test datasets. Specifically, for a 512 × 512 input image, 95.12% mean intersection over union is achieved on the insulator test datasets with different environments and model parameters size being only 13.61 M. Compared with the current state-of-the-art semantic segmentation methods, the results show that the proposed method has higher efficient and accuracy.

1. Introduction

As an important part of transmission line, the integrity of insulator directly affects the safety and reliability of transmission line. According to statistics, the trip accident of the transmission lines due to insulator fault in Figure 1 accounts for 81.3% [1]. Since insulators are exposed to the natural environment for a long time, they will be affected inevitably by different climate and environmental factors, resulting in defects such as filthiness, corrosion, breakage, and so on, which threaten the safe operation of transmission lines. Therefore, it is necessary to inspect the insulators regularly to eliminate hidden dangers in time to ensure the stable operation of the entire transmission line.

In the early days, the regular inspection of transmission lines was carried out manually, and the insulators on the transmission lines were observed, inspected, and measured manually through eyes or telescopes. This inspection method requires personnel to have rich prior knowledge, which is inefficient and dangerous. In recent years, with the advancement of computer vision and Unmanned Aerial Vehicle (UAV) multimodal information fusion, UAV-based aerial solutions are widely used in the transmission line inspection [2]. Camera is the main way for UAV to perceive external information; technicians use the image information collected by UAV and image processing technology to complete the regular detection of insulators in transmission lines. However, as transmission lines are usually located in different natural environments, aerial images obtained by UAV camera have characteristics of complex background, multiple pseudotargets, and low signal-to-noise ratio, which make it difficult for image

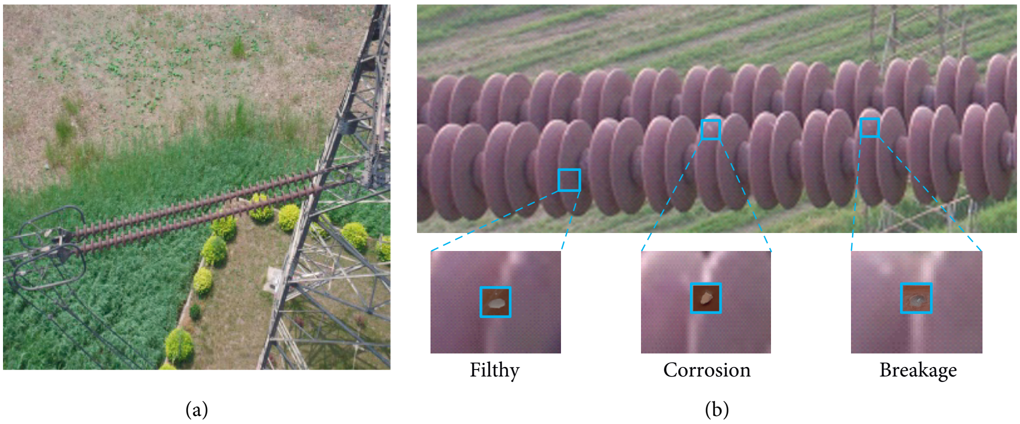


Figure 1: Insulator image. (a) Original aerial image. (b) Common insulator defects.

processing technology to identify insulators from aerial images [3].

In traditional image processing, artificial feature engineering is used to design and extract the feature information (e.g., color, texture, and edge) of the insulator to distinguish the insulator from the background in the aerial image [3]. Reddy et al. [4] first transformed the insulator image into the LAB color space and used the k-means clustering algorithm to fuse similar pixels in the image and then calculated the pixel intensity to draw a bounding box based on a preset threshold. The bounding box features are extracted as the input of the adaptive neurofuzzy inference system (ANFIS), and finally the ANFIS model is trained to detect insulators in the image. The disadvantage of this method is that the model relies on the bounding box generated by the clustering results, which is suitable for the segmentation of insulators with less background interference. Murthy et al. used Hough transform and support vector machines (SVM) to segment the insulators in aerial images and used multiresolution wavelet changes to extract the features of the extracted insulators and then used SVM to judge the state [5]. This method is suitable for aerial images where the background texture and intensity information are quite different from those of insulators. In [6], a method based on orientation angle detection and binary shape prior knowledge (OAD-BSPK) was proposed to realize the localization of insulators with different azimuths in complex aerial survey images. The algorithm first uses the binarization and multicascade morphological methods to obtain the edge of the insulator in the image to initially detect the possible azimuth of the insulator. For possible azimuths, binary shape priors are used to precisely locate the insulators in the image. However, this method needs to set a large number of thresholds to segment the insulators, and the thresholds for different environmental changes need to be reset, which greatly weakens the detection of insulators in the natural environment. In [7], a method was proposed to identify insulators in aerial images by fusing shape, color, and texture information. The line segments in different directions in different images are extracted; thus, the candidate regions are obtained by clustering according to the characteristics of the insulators. Moreover, the insulator regions are obtained based on the saliency features of the color and the prior knowledge model; the texture features are used to detect the insulator shedding defects. In [8], saliency detection was used to determine the position of insulators based on color and gradient features. Furthermore, adaptive morphological methods were used to detect insulator defects. This method is only suitable for the identification of glass insulators. Traditional image processing algorithms are limited to scenes with single background and fixed objects. Although all the above methods can solve some problems in object recognition, the design of artificial features requires more a priori knowledge and makes it difficult to be highly robust in natural environment image of insulators.

In recent years, with the development of deep learning, many models based on convolutional neural networks (CNN) have achieved excellent results in semantic segmentation tasks, such as FCN [9], SegNet [10], U-Net [11], and so on. Pixel-level segmentation can obtain targets in the image and the location of the targets precisely; these models have been widely used in the field of autonomous driving, industry, and embedded devices, solving problems in computer vision and promoting industrial automation. Compared with traditional methods, CNN can automatically extract features from images, which reduces the subjectivity of artificially designed features and provides more flexible solutions. It makes the insulator segmentation and identification process more robust to light, background, material, and other changes. At present, many researchers have studied semantic segmentation network models to perform pixel-level segmentation of insulator images. Gao et al. [12] used the two-stage target detection Faster-RCNN network to detect the insulators in the aerial images, input the detected insulators into the fully convolutional network (FCN) to complete the segmentation, and completed the defect detection of the insulators. But the network is divided into two convolutional neural networks: object detection and semantic segmentation. This method reduces the computational complexity of the semantic segmentation algorithm and improves the segmentation speed. Ling et al. [13] used the object detection network to identify insulators in aerial images and cropped the identified insulators from the images, which were input into the U-Net semantic segmentation network to complete the detection of the insulator defects. The object detection would be unable to satisfy demands for the detection of small defects (e.g., flashover and corrosion). However, semantic segmentation is pixel-level detection, which would improve the accuracy of model detection. It was noted that Up-Net network is given to achieve semantic segmentation of insulator images in [14]. This method does not require an object detection network in advance and directly uses the semantic segmentation method to complete the identification and segmentation of insulators. Alahyari et al. [15] proposed a two-stage convolutional neural network model consisting of segmentation and classification units for fault classification of insulators. Aerial insulator images have the characteristics of complex backgrounds, many pseudotargets, and low signal-to-noise ratios, which bring difficulties for segment insulators. Semantic segmentation networks based on deep convolutional neural networks are general algorithms. Consequently, it is necessary to fine-tune the algorithm to improve its accurate and efficiency for the enhancement of insulator segmentation performance in complex background.

Since the premise of insulator fault detection is to accurately identify the insulator target in the image, it is radical to develop a better insulator identification algorithm. This paper focuses on the identification and segmentation of insulators in complex background, aiming to improve the identification efficiency and segmentation accuracy of insulators in aerial images. In this paper, an insulator semantic segmentation network is proposed, called ED-Net, which is constructed based on an end-to-end Encoder-Decoder structure, which realizes pixel-level segmentation of insulators. The network consists of two parts: Encoder path and Decoder path. In the Encoder path, Initial Module (IM) and CA-Bottleneck Module are proposed to extract the feature of the original image. The Coordinate Attention (CA) mechanism [16] and the depth-wise separable convolution are used in the CA-Bottleneck module to make the network pay more attention to the region of interest in the image and reduce the amount of model parameters and make the network more effective, respectively. The global average pooling layer (GP) is added in the tail of the Encoder path to retain the maximum receptive field. In the Decoder path, we proposed Asymmetric Convolution Module (ASM) and Refinement Boundary Module (RBM) to optimize the insulator boundary information. For reducing the diversity of features, the attention feature fusion module (AFFM) is proposed to fuse the features better in each stage.

Our contributions are summarized as follows:

(1) In the Encoder path, the proposed IM is used to quickly downsample the original image and then add coordinate attention mechanism in the original Bottleneck module, named CA-Bottleneck Module. The module is stacked in the Encoder path to enhance the feature extraction ability of the network. Ultimately, the GP is added at the end of the encoder path to retain the maximum receptive field.

(2) In the Decoder path, this paper introduces ASM and RBM modules for optimizing edge detail information of feature maps. AFFM is thus used to fuse high-level features and low-level features and reduce the diversity of them. In this case, the effectiveness of feature fusion can be ensured.

(3) The impressive results are achieved on the insulator dataset. More specifically, 95.18% mean IOU on the insulator test dataset and the parameter amount of the network of only 13.61 M are obtained.

The rest of this paper is organized as follows: Section 2 reviews the semantic segmentation network model and the work of attention mechanism. Section 3 introduces the proposed semantic segmentation network model based on the encoder-decoder structure and the details of each part of the network. Section 4 introduces the insulator segmentation dataset and discusses the impact of each module in the network model proposed in Section 3 on the overall segmentation accuracy. The experimental results of mainstream semantic segmentation models in the insulator test dataset are compared and analyzed. Section 5 provides the conclusions.

2. Related Work

In recent years, researchers have made a lot of progress on insulator defect detection and state classification. However, in most of these tasks, deep learning-based object detection algorithms are used to obtain the region of the insulator in the image and crop it into the region of interest. The segmentation method is used to obtain the segmentation map of the insulator in the region of interest (RoI), and finally the insulator defect detection is carried out. These algorithms usually require two CNN models, resulting in a large number of model parameters, which cannot be applied to devices with limited computing resources. With the development of semantic segmentation algorithms in deep learning, more and more semantic segmentation models can be used to solve the problems of complex backgrounds and many pseudotargets when segmenting insulator images and have achieved state-of-the-art performance. Most of the current insulator segmentation methods are based on FCN [9] variants or U-Net [11] variants to achieve high performance. However, these methods do not consider applying the CNN model to mobile devices. In semantic segmentation tasks, most networks are designed based on encoder-decoder structures, and to achieve a balance between network speed and accuracy, some methods will be used, such as depth-wise separable convolution and attention mechanisms.

2.1. Encoder-Decoder Structure.

2.1. Encoder-Decoder Structure. The encoder part inherent in the FCN [9] model encodes the feature of different scales. Naturally, some methods combine features of different scales to optimize the final prediction map. These methods mainly consider the loss of target spatial information due to the decrease of spatial resolution caused by continuous pooling and convolution with stride equal to 2 and restore the resolution of the feature map by feature fusion. For example, SegNet [10] restores the resolution of the feature map by saving the index at the time of the maximum pooling operation to reduce the loss of image spatial information, U-Net [11] uses skip joins to fuse high-level semantic information with low-level spatial information to improve the result of the segmentation, GCN [17] uses a large convolution kernel to obtain a larger receptive field and extract more image context information when fusing the feature map in the encoding and decoding structure. However, spatial information lost during image downsampling is difficult to be restored by directly fusing feature maps of different scales [18].

2.2. Feature Fusion in Semantic Segmentation. Feature fusion is a common method in the semantic segmentation task, which is used to fuse feature images extracted at different stages. In DeepLab series [19–22], ASPP modules are proposed to extract multiscale features with different dilation rates and combine them to process targets of different scales. The pyramid pooling module in PSPNet [23] achieves the goal of coding different scales through the feature map of different stages. ParseNet [24] adds global pooling branches to extract multiscale feature.

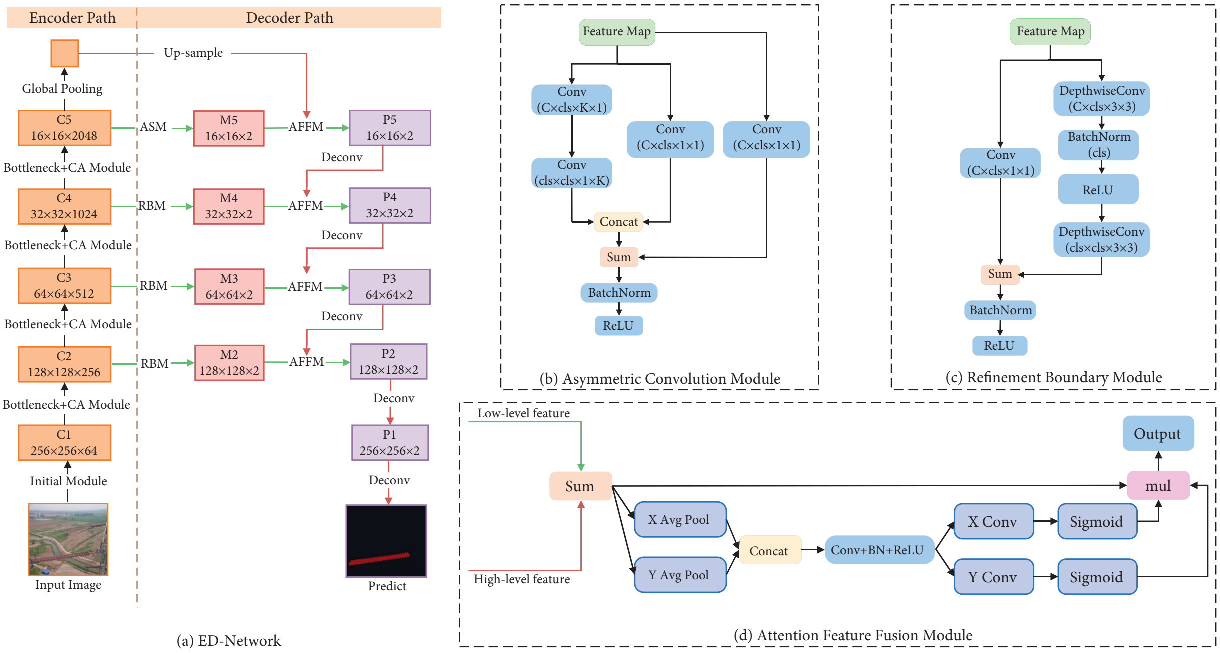
2.3. Context Information. Semantic segmentation requires more image context information to generate more accurate segmentation results. Recently, most methods use the fusion of different scale feature maps or large receptive fields to obtain the context information of the image. DeepLab v2/v3 uses hole convolution with different dilation rates in parallel to extract multiscale context information from the feature map extracted from the backbone network. It is proposed that ASPP module can change the receptive field of convolution kernel by controlling the dilation rate so as to capture richer multiscale features. PSPNet [23] proposed a PSP module using multiscale pooling to captures multiscale information. However, atrous convolution will lose the continuity of image information while increasing the receptive field, which is not conducive to the pixel-level dense prediction task [25].

2.4. Attention Mechanism. The attention module can make the model pay more attention to the region of interest. Attention mechanism is a powerful tool for depth convolution neural network [26]. At present, the most popular attention mechanism is the SE Attention proposed by SENet [27]. It calculates channel attention through 2D global pooling, providing significant performance improvements at a relatively low computational cost. However, the SE module only considers the encoding of information between channels and ignores the importance of location information, which is crucial for semantic segmentation. CBAM [28] combines the spatial attention module and the channel attention module to refine the extracted features and improve the expressive force of the model.

3. Methodology

In daily inspections, UAV can collect hundreds of thousands of aerial images. The network should have less complexity and less parameters while ensuring the segmentation accuracy. Traditional convolutional neural networks (e.g., GoogLeNet [29] and ResNet [30]) have a large number of network model parameters and high computational complexity, making it difficult to run on mobile devices (e.g., UAV). Aiming at the characteristics of the complex background of aerial images, the traditional CNNs is complicated calculation, and the number of model parameters is large. In this section, an alternative network is proposed to solve the above problems, called ED-Net; Figure 2(a) shows the overall framework of the network. The network achieves a balance between the amounts of parameters and the segmentation accuracy. Then, the details of components in these two paths are described. Finally, how we fuse features of different scales in the decoder path and restore the image resolution to obtain the final prediction result is demonstrated.

3.1. Encoder Path. For the insulator semantic segmentation task, ensuring a large receptive field is crucial for the final semantic segmentation result. Since the insulator is only in a small part of the aerial image and the pseudotarget in the background will interfere with the segmentation of the insulator, it is necessary to retain the largest receptive field to ensure the accuracy of insulator segmentation from a global perspective. At present, the mainstream methods of expanding the receptive field use the pyramid pooling module, Atrous Spatial Pyramid Pooling [20], or large kernel [17]. In this paper, attention mechanism and depth-wise separable convolution are introduced in the encoder stage to reduce the number of model parameters and expand the receptive field so that it can obtain more image context information. The encoder path consists of five stages, one of which is the initial module, and the other four stages are modules with the same structure. The details of Encoder path are shown in Figure 3, including IM, CA module, and CA-Bottleneck module. The initial module consists of two branches. One branch uses the depth-wise separable convolution of 3×3 and stride equal to 2 to downsample the original image; the other is divided into max-pooling to downsample the original image and finally fuse and input to the next stage in Figure 3(a). The next four stages consist of stacking different numbers of CA-Bottleneck modules, and only the first CA-Bottleneck module in each stage uses 1×1 conv for input-output feature map scale matching. The rest of the CA-Bottleneck modules have the same structure, as shown in Figure 3(c). It consists of the following four parts: (1) 1×1 convolution layer for dimension decline. (2) Depth-wise separable convolution is used to extract its features. The significance of depth-wise separable convolution layer is to reduce the amount of model parameters. (3) Adding



(a) ED-Network

(b) Asymmetric Convolution Module

(c) Refinement Boundary Module

(d) Attention Feature Fusion Module

Figure 2: Overview of proposed ED-Net architecture (a). It consists of two parts: Encoder path and Decoder path. The feature maps of C1, C2, C3, C4, and C5 are obtained by the Initial Module (IM) and Bottleneck + CA Module. M2, M3, M4, and M5 are the feature maps of the encoder path obtained through Asymmetric Convolution Module (ASM) or Refinement Boundary Module (RBM). P1–P5 are obtained by the Attention Feature Fusion Module (AFFM), and P1 is used for insulator segmentation. The details of ASM, RBM, and AFFM are illustrated in (b), (c), and (d), respectively. The red and black lines represent the upsample and downsample operations, respectively. The green line does not change the size of the feature map, only the number of channels.

Coordinate Attention Mechanism and residual connection to weight the feature map to make the model pay more attention to the region of interest. (4) 1 × 1 convolution layer is used to improve the dimension and fuse with the input feature map to obtain the final output. The Batch Normalization [31] and ReLU are placed in the between of whole convolution operation.

3.2. Decoder Path. In the semantic segmentation task, the final segmentation result of the network is determined by the quality of feature fusion, since the features at different stages have different recognition capabilities. In the shallow stage of the encoder path, the network pays more attention to some low-level features, such as point, line, or edge information. At this time, the feature map has a large resolution, so it contains a lot of spatial information. However, in the high-level stage of the encoder, the features extracted from the network contain more semantic information, but the resolution of the feature map is low at this time, and the spatial information of the image is lost. Therefore, the Decoder path proposed in this paper integrates the high-level semantic information extracted from the network into the low-level features so as to better combine the semantic information and spatial information, thus improving the segmentation result of the network.

In this paper, RBM is proposed to optimize edge details of feature maps. The module is a residual structure composed of depth-wise separable convolution, Batch normalization, and nonlinear activation function, as shown in Figure 2(c). It is worth noting that only the feature maps from stage 2 to stage 4 of the Encoder path need to go through RBM. Second, inspired by the structure of InceptionV3, on $m \times m$ feature maps, where m ranges between 12 and 20, using asymmetric convolutions can improve the classification performance of the network [32]. Therefore, this paper proposes ASM for the feature map of stage 5 output in the Encoder path, as shown in Figure 2(b). ASM uses asymmetric convolution. Compared with the traditional 3×3 convolution, it requires less resource consumption and parameter amount, which is more effective for large kernel sizes. Finally, due to the difference between low-level features and high-level features, these features cannot be simply added together. In the early stage of the network, the network encodes rich spatial information, and in the later stage of the network, it mainly encodes the context information of the image. In other words, the feature map output at the early stage of the network is of low level, and the feature map output at the later stage of the network is of high level. Therefore, the AFFM is proposed to fuse low-level and high-level features in Figure 2(d). This module first sums low-dimensional and high-dimensional features for different levels of features. Next, average pooling is performed on the x-direction and y-direction of the feature map, respectively. After concatenating the feature map, convolution, Batch normalization and ReLU are used to balance the feature map scale. Finally, the weight vectors in the x-direction and y-direction are obtained through the sigmoid activation function, respectively, and the original feature map is reweighted. This module uses high-level semantic

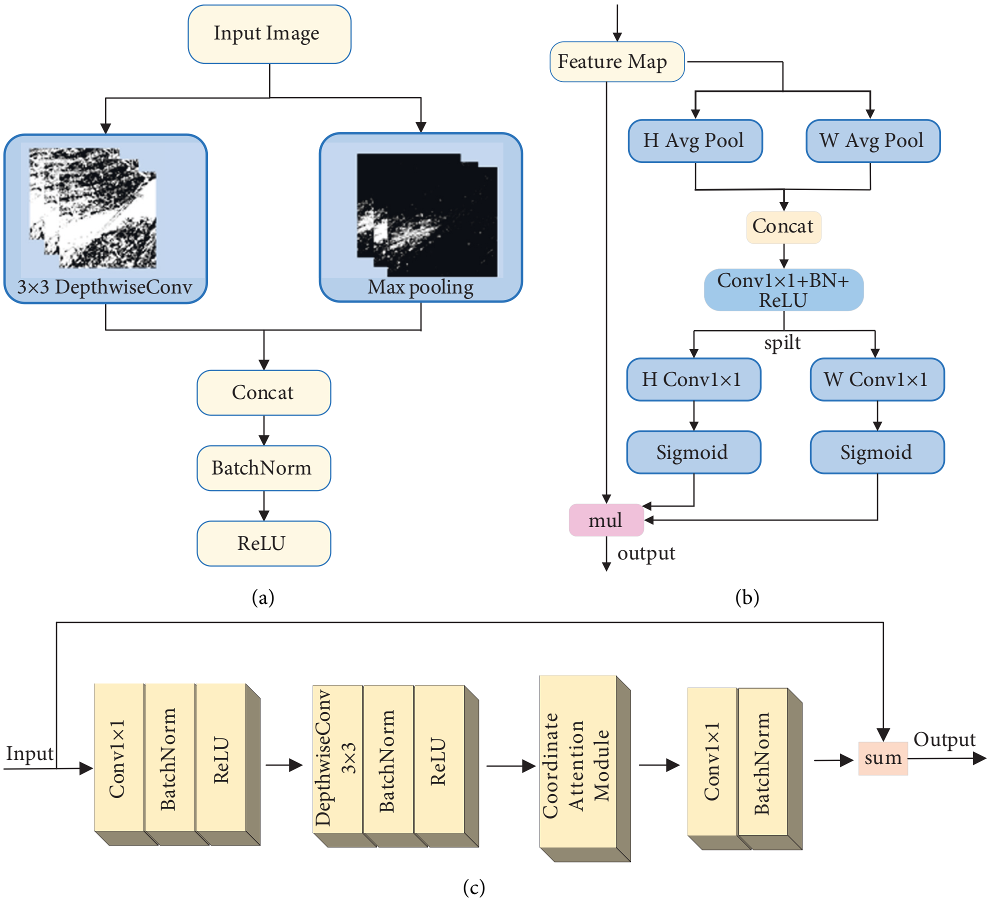


Figure 3: Encoder path components. (a) Components of the Initial Module. (b) Coordinate Attention Module. (c) Overall composition of CA-Bottleneck Module.

features to guide low-level detail features and refines feature maps so as to obtain improved segmentation result.

The proposed Encoder-Decoder Path Network did not use bias terms in any of the projections in order to reduce the number of kernel and overall memory operations, as cuDNN [33] uses separate kernels for convolution and bias addition. This choice did not have any impact on the accuracy. It is noted that each convolution operation make network better for gradient return accompanied by normalization and activation function, reducing the phenomenon of gradient disappearance and gradient explosion.

3.3. Network Architecture. With the Encoder path and Decoder path, a semantic segmentation network based on Encoder-Decoder structure is proposed to segment insulator images, called ED-Net. The overall flow of the network in this paper is shown in Figure 4. The multiscale feature map that is extracted from different stages of the Encoder path is used in the decoder path to generate finally predicted result.

In the Encoder path, the IM and stacked CA-Bottleneck module are used as the backbone network. CA-Bottleneck module adds coordinate attention mechanism with depth-wise separable convolution. The coordinate attention mechanism is effectively applied to semantic segmentation tasks, which enables the convolutional kernel to capture channel, direction, and position information and enables the model to locate the target more accurately. Depth-wise separable convolution is used to reduce the number of model parameters and improve the efficiency of model calculation. Then, the GP is added in the tail of backbone network to get the strongest consistency feature.

In the Decoder path, the output feature maps of Encoder path C2 to C4 are passed through RBM module, and the output of C5 is passed through ASM module to further enhance the consistency of feature. The feature maps of GP are 2 × upsampling and the feature maps of C5 are fused by AFFM. The fused feature map is deconvolved and the output feature map of the previous stage of RBM is fused by AFFM. Through this operation, high-level features and low-level features are fused together to reduce the gap between feature maps so that feature maps contain rich semantic information and spatial information. Finally, the image size is restored through deconvolution to achieve end-to-end training of the network. It is worth noting that the attention mechanism is only performed on the current feature map in the last AFFM.

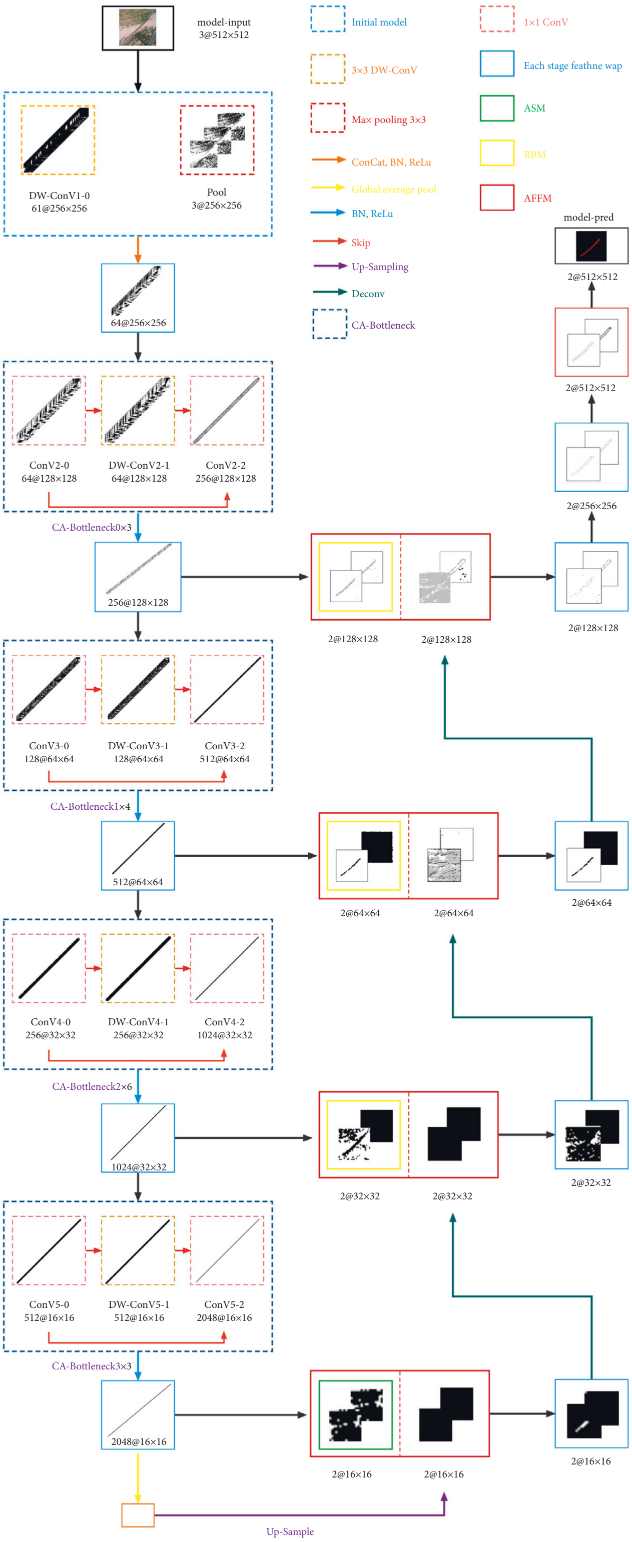


Figure 4: Inference process of the Encoder-Decoder Network.

4. Experiment Results

In this section, the proposed ED-Net network model is evaluated to verify the segmentation effect of the model on insulators in different environments. Firstly, the insulator dataset, implementation, and evaluation metrics used in the experiments are introduced. Then, the impact of each module in the proposed method on the overall model is analyzed. Next, the amounts of parameters, GFLOPs of the network, and the results that are generated by our algorithm on the insulator test dataset are shown, respectively. Finally, comparison with the current mainstream semantic segmentation models is carried out.

4.1. Insulator Dataset and Implementation

4.1.1. Insulator Dataset. Figure 5 shows the acquisition equipment of the aerial insulator image data used in this paper. The data acquisition system is composed of a FC300S camera of CASIO and a DJI Phantom 3 advanced UAV. The camera captures insulator images of transmission lines located in different environments, and the captured images are used as the insulator dataset in this paper. The insulator dataset used in this paper contains 100 training images and 20 testing images, which are resized to 512 × 512. Examples of train sample and label data are shown in Figure 6. The images contain 2 semantic categories, namely, insulator and background.

4.1.2. Training details. During the training time, a batch size of 16 is applied. The standard Adaptive Moment Estimation (Adam) optimizer is used to update the model weight parameters, where β_1 and β_2 are set to 0.9 and 0.99, respectively, weight decay $1e^{-3}$, initial learning rate $1e^{-4}$, and the learn rate is multiplied by 0.1 for every 40 iterations. All experiments are implemented in Python and trained using the PyTorch deep learning framework. All the experiments have been conducted on a Nvidia GeForce GTX 3090 GPU under Ubuntu 16.04.

4.1.3. Evaluation Metrics. In order to better understand the results of the insulator semantic segmentation network, this paper summarizes the different evaluation metrics [9] as follows:

(i) Pixel Accuracy: $(\sum_i n_{ii}/(\sum_i \sum_i n_{ij}))$

(ii) Class Accuracy: $((\sum_i n_{ii} + \sum_j n_{jj})/(\sum_i n_{ij} + \sum_j n_{ji} + \sum_i n_{ii} + \sum_j n_{jj}))$

(iii) Mean Intersection over Union (mean IOU): $(1/n_c(\sum_i n_{ii}/(\sum_i n_{ij} + \sum_j n_{ji} - n_{ii})))$

(iv) Frequency weighted Intersection over Union (f.w. IOU): $((1/\sum_i \sum_j n_{ij})\sum_i n_{ii}/(\sum_i n_{ij} + \sum_j n_{ji} - n_{ii}))$

where n_c represents the number of categories in the dataset, n_{ii} represents the number of pixels whose real pixel class is i predicted to belong to class i, n_{ji} represents the number of pixels whose real category j predicted to belong to class i, and n_{ij} represents the number of pixels whose real category is i predicted to belong to class j.

In the following sections, a series of ablation experiments is to evaluate the effectiveness of proposed method. Then, the full results on Insulator test dataset are reported.

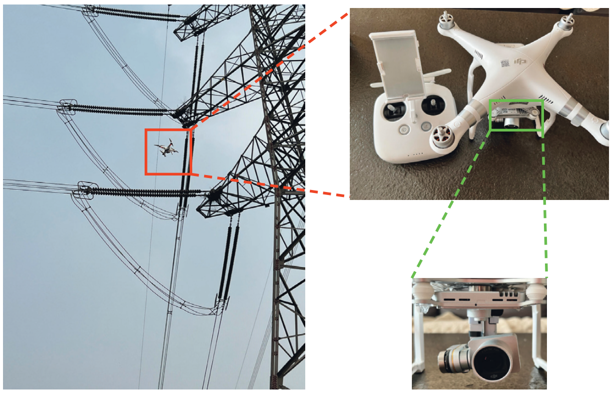


Figure 5: Data acquisition system.

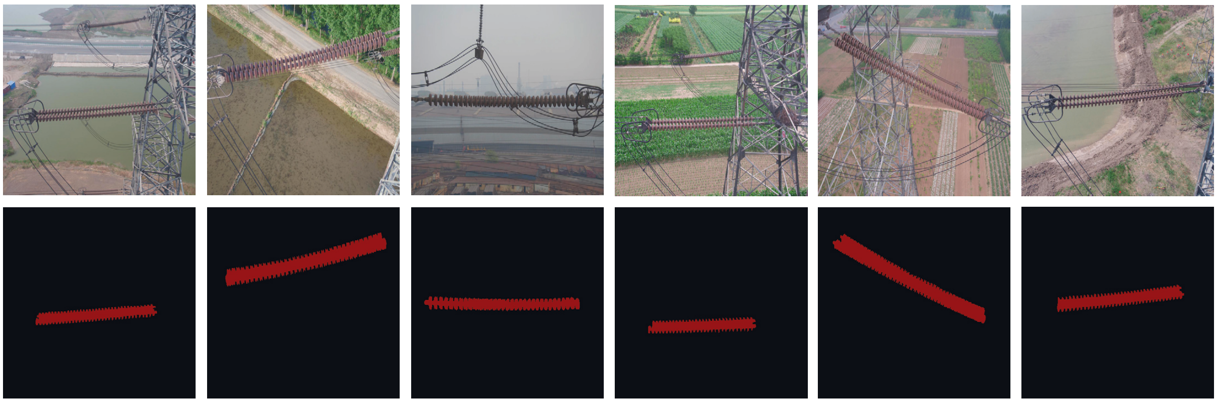


Figure 6: Example of train sample and label data.

4.2. Ablation Study. In this section, the proposed method is broken down step by step to reveal the effect produced by each module. The effectiveness of our method on the insulator dataset is verified using the Base Model as the base network.

Base Model. ResNet-50 is used as the feature extraction module of the Base Model, which is the Encoder Path. The Decoder Path directly deconvolution and fuses the feature maps to obtain the final prediction result, as shown in Figure 7. The Conv-1 in the Base Model uses the convolution of kernel size equal to 7, stride equal to 2 to downsample the image, and then stacks different numbers of Bottlenecks Module to obtain feature maps Res-2, Res-3, Res-4m and Res-5. The pooling operation is not used in the network, and the convolution with stride equal to 2 is used for downsampling instead. Considering the amounts of parameter in the network, all 3 × 3 convolution operations in the network are replaced by depth-wise separable convolutions. The depth-wise separable convolution can greatly reduce the amount of network parameters, reducing the complexity of the model and making it possible to run on devices with limited computing resources, as shown in Table 1. The output feature map of each stage first needs to go through the Trans Module (composed of 1 × 1 convolution), and after reducing the number of channels, deconvolution is performed and the feature map of the previous stage is fused. It is worth noting that the Conv-1 feature map does not participate in the fusion. After obtaining the P2 feature map, the image size is restored by two deconvolutions to obtain the final prediction result. The network prediction results are shown in the second column of Figure 8.

CA-Bottleneck Module. The Encoder part of the ED-Net proposed in this paper uses the CA-Bottleneck module for stacking, and adding an attention mechanism to the Bottleneck makes the network better focus on the global information of the image. The key parameter reduction ratio γ is gradually increased, but the performance of the model decreases when γ is increased to 512, as shown in Table 2. This shows that choosing a suitable reduction ratio γ in the early stage of the model is crucial to improve the performance of the model. Note that only multiples of two are used in the experiments because there is a twofold relationship between feature map size reduction and channel count improvement. In particular, when γ equals to 2, the model outperforms the Base Model by 6.05%. In the Base Model, the size of the feature map at each stage needs to be reduced by 1/2, and the number of corresponding channels needs to be doubled. According to this idea, the corresponding reduction ratio γ in the CA module of Res-3, Res-4, and Res-5 should also be reduced by two times to increase the number of channels. Using the CA-Bottleneck

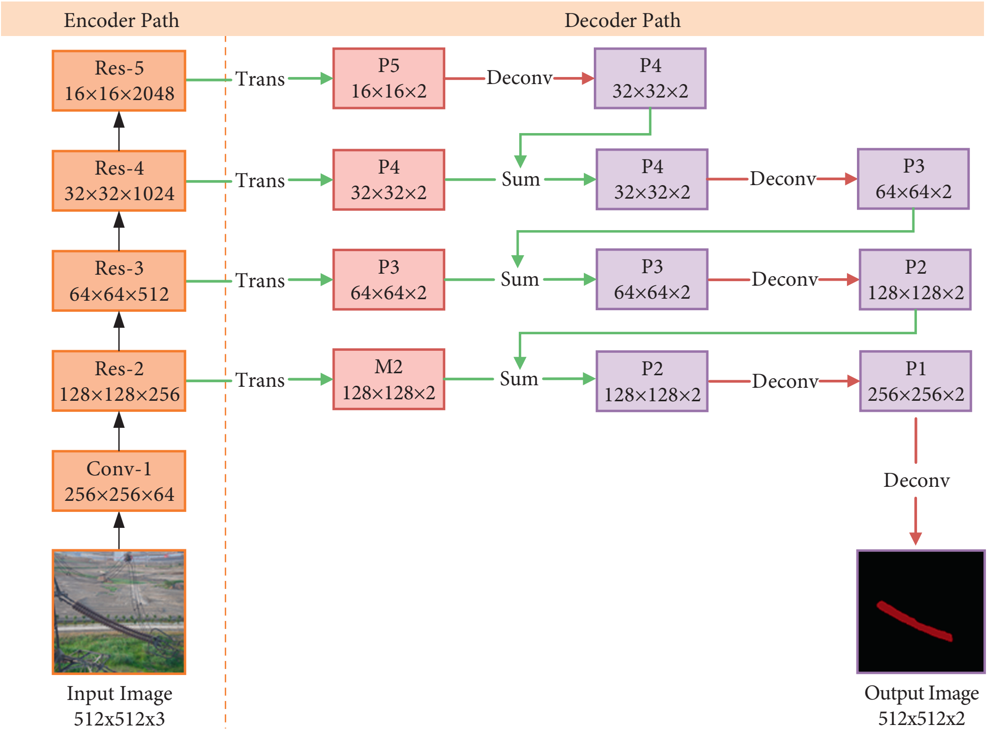


Figure 7: Base Model. Trans is to use 1 $*$ 1 convolution to change the number of channels. Sum is the addition of the corresponding elements of the feature map. Deconv uses the deconvolution of kernel size equal to 4, stride equal to 2, and padding equal to 2 to expand the feature map size.

Table 1: Accuracy and parameter analysis of Base Model on the insulator dataset. ResNet-50 and depth-wise separable convolution (DW)—ResNet-50 refers to the replacement of all 3×3 convolution in the network with depth-wise separable convolutions. GFLOPs estimated model complexity for $3 \times 512 \times 512$ input image.

	Parameter (M)	GFLOPs	mIOU (%)
Base	25.56	21.71	72.57
DW-base	13.48	13.18	72.50

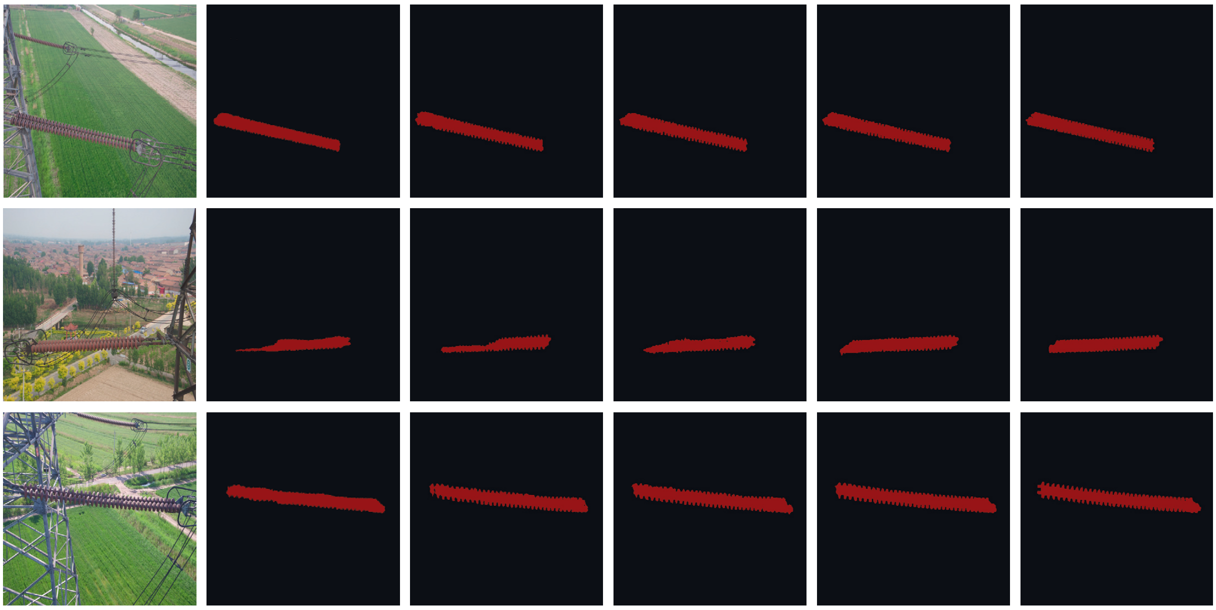


Figure 8: Examples of semantic segmentation results on insulator test dataset. (a) Original input image; (b) the predicted image of Base Module; (c) the predicted image of the former combined with CAM; (d) the predicted image of the former combined with RBM; (e) the predicted image of the former combined with AFFM; and (f) Ground truth.

Table 2: Experimental results of setting different reduction ratios γ using the CA-Bottleneck module in Res-2. The score is evaluated by standard mean IOU (%) on insulator test dataset.

γ	Base	8	16	32	64	128	256	512
mIOU	72.50	73.91	74.88	75.14	75.58	76.89	78.55	77.65

module in ResNet-50, the model mIOU = 81.08%. After adding CA Module, the network prediction results are shown in the third column of Figure 8.

Initial Module. As described in Section 3, the module is proposed which consists of depth-wise separable convolution and max-pooling to reduce model parameters and spatial information loss. The module subjects the input image to 3 × 3 depth-wise separable convolution with stride equal to 2 and overlapping max-pooling, respectively. The two outputs are concatenated in the channel dimension through Batch Normalization and ReLU. This module replaces the convolution of kernel size equal to 7, and the network performance is improved from 81.08% to 82.29%, as shown in Table 3.

4.2.1. Global average Pooling. The layer is used to make the Encoder path provide a larger receptive field. Although the original ResNet-50 network can theoretically cover most of the input image, it is necessary to use GP to further expand the receptive field. In this paper, the GP is added to the tail of the Encoder Path, then the output upsampling of the GP is fused with the feature map of Res-5 in the Encoder path in an additive manner. Model performance ranges from 82.29% to 83.93%, validating the effectiveness of GP, as shown in Table 3.

4.2.2. Refinement Boundary Module. To further improve the performance of the network, the RBM is designed in the Decoder Path. This module contains convolution, Batch normalization, and ReLU unit. Compared with Trans Module, RBM improves the model ability to optimize the target boundary, as shown in Table 3. After adding RBM, the network results are shown in the fourth column of Figure 8.

4.2.3. Asymmetric Convolution Module. Based on the RBM in the Decoder Path, the ASM is proposed to be applied in the stage with the smallest feature map size. ASM uses large convolution kernels to densely connect feature maps. The model performance ranges from 88.51% to 89.57%, which verifies the improvement of the overall performance of the model by ASM, as shown in Table 3.

4.2.4. Attention Feature Fusion Module. Considering the different levels of feature maps generated in different stages of the network, the low-level features generated when the network is shallow, and the high-level features generated by the deep network, the AFFM is proposed to effectively fuse these features. The evaluation of the results that is generated by summing these features directly and proposed AFFM is shown in Table 3. The network prediction results after using AFFM are shown in the fifth column of Figure 8.

Table 3: Detailed performance comparison of each component in our proposed ED-Net.

Module	mIOU (%)
CAM	81.08
CAM + IM	82.29
CAM + IM + GP	83.95
CAM + IM + GP + RBM	88.51
CAM + IM + GP + RBM (ASM)	89.57
CAM + IM + GP + RBM (ASM) + AFFM	95.12

Table 4: Parameter comparison of our method against other state-of-the-art methods on the Insulator test dataset. GFLOPs are estimated for input of 3 × 512 × 512.

Model	Backbone	GFLOPs	Parameters (M)
FCN-8s	VGG16	80.63	20.1
SegNet	VGG16	286.0	29.43
U-Net	\	184.64	34.53
DeepLabV3	Xception	57.06	29.4
GCN	ResNet152	67.96	58.38
Ours	CA-Bottleneck	12.62	13.61

4.3. Comparison of Different Semantic Segmentation Algorithm. In this section, the complexity and parameter quantity of the model in this paper are firstly analyzed, and a comparative analysis is made with the current mainstream semantic segmentation networks, as shown in Table 4. Secondly, the segmentation results of the proposed algorithm and the mainstream algorithm in the insulator test set are compared, as shown in Table 5. Finally, some visual examples of the method in this paper and the mainstream semantic segmentation models are also given, as shown in Figure 9.

As shown in Table 4, the comparison between our proposed method and other methods between GFLOPs and parameter quantities is shown. GFLOPs represent the complexity of the model, and the amounts of parameter represents the number of operations when processing the image. In this paper, the unified input image resolution is 512 × 512. Table 5 shows the accuracy and speed comparison between the different methods in the insulator test dataset. Figure 10 shows the ROC curves generated by ED-Net and mainstream semantic segmentation models on the insulator test dataset. Compared with other mainstream

Table 5: Experimental results of our method against other state-of-the-art methods on the Insulator test dataset comparison of our method against other state-of-the-art methods on the Insulator test dataset.

Model	Pixel Acc	Class Acc	mIOU	f.w. IOU	Time (ms)
FCN-8s	98.85	88.50	86.66	97.75	157
SegNet	99.07	93.01	89.64	98.22	167
U-Net	99.20	93.28	90.67	98.46	181
DeepLabV3	98.90	91.10	87.40	97.89	176
GCN	99.06	92.11	89.07	98.18	348
Ours	99.61	97.44	95.12	99.24	67

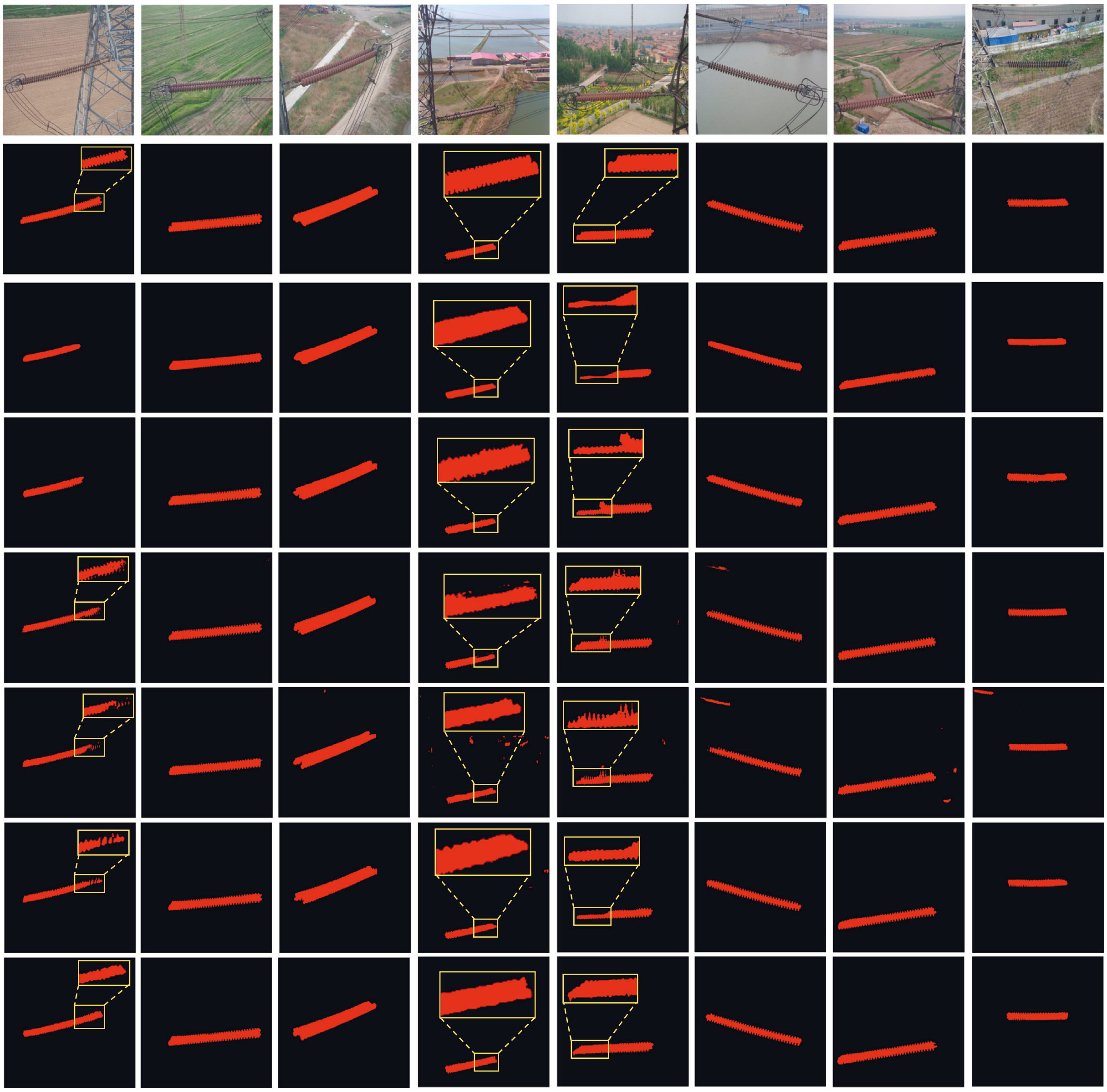


Figure 9: In the insulator test dataset, the ROC curves are generated by ED-Net and mainstream segmentation models.

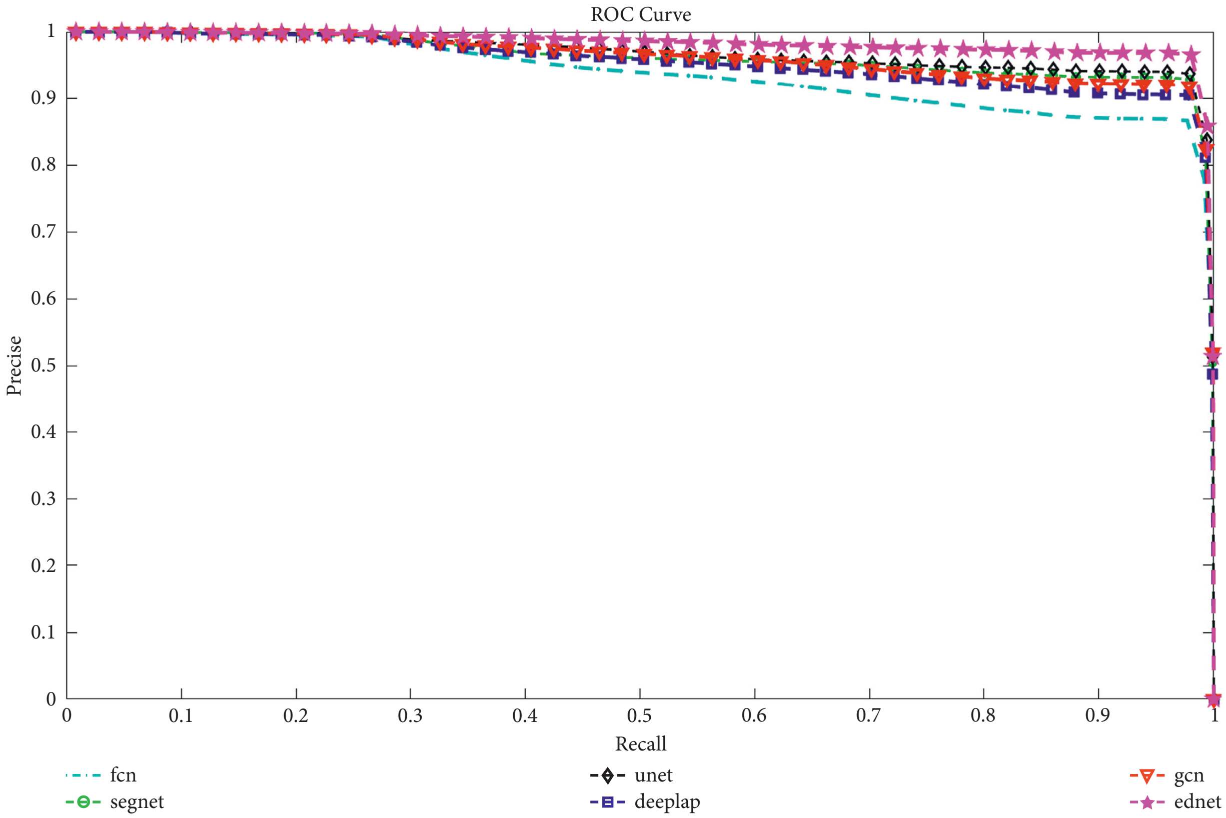


Figure 10: Semantic segmentation results of our method and other state-of-the-art methods on the insulator test set. The figure is the original image and the prediction results of each algorithm. From top to bottom are the original image, Ground Truth, FCN-8s, SegNet-VGG16, U-Net, DeepLabV3+, and GCN-ResNet152 and the last row is the prediction result of the network in this paper. The yellow box shows the details of the segmentation.

semantic segmentation models, our method has achieved great progress in both speed and accuracy.

Figure 9 shows the segmentation results of the proposed segmentation method and other state-of-the-art segmentation methods on the insulator test dataset. This set of images is representative because these images include insulators that are in different shapes, viewing angle, environments, and so on. As shown in Figure 9, semantic segmentation results of insulators can be achieved in most networks, but misclassification and missing segmentation will occur when there is too much background and insulator interference, such as the fourth and sixth images. The robustness of the proposed method is proved.

5. Conclusion

In this paper, an insulator semantic segmentation network is designed to achieve accurate and efficient segmentation insulators in different environments, which is based on Encoder-Decoder structure, called ED-Net. The network architecture consists of two paths: Encoder path and Decoder path. In the Encoder path, in order to improve the feature extraction ability of encoder, the CA Module is added into original Bottleneck to make the network focus on the insulator region. The amount of model parameters is reduced by initial module and depth-wise separable convolution to improve the efficiency of feature extraction. Moreover, GP is used to achieve more semantic information. In the Decoder Path, the RBM optimizes the edge details of the feature map generated by Encoder path. ASM uses large kernel size to obtain rich contexture information and reduce the number of parameters. Attention feature fusion model are proposed to reduce the difference between high-level features and low-level features and improve the accuracy of the model. For aerial images under different environments and lighting conditions, 95.12% mean IOU is obtained in the insulator test dataset and the amount of model parameters is only 13.61 M. In the future work, the insulator dataset needs to be further extended to ensure that the model can obtain accurate insulators in more complex environments. In addition, we will pay more attention to lightweight convolution neural network to obtain real-time segmentation results.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 62063011) and partially supported by Yunnan Fundamental Research Project (Grant no. 202001AV070001).

References

[1] W. Tong, J. Yuan, and B. Li, "Application of image processing in patrol inspection of overhead transmission line by helicopter," *Power System Technology*, vol. 34, no. 12, pp. 204–208, 2010.

[2] G. Qiang, Y. Wu, and L. Qian, "Research on deep belief network layer tendency and its application into identifying fault images of aerial images," *Chinese Journal of Scientific Instrument*, vol. 36, no. 6, pp. 1267–1274, 2015.

[3] W. Chang, G. Yang, J. Yu, and Z. Liang, "Real-time segmentation of various insulators using generative adversarial networks," *IET Computer Vision*, vol. 12, no. 5, pp. 596–602, 2018.

[4] M. J. B. Reddy, B. K. Chandra, and D. K. Mohanta, "A DOST based approach for the condition monitoring of 11 kV distribution line insulators," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 18, no. 2, pp. 588–595, 2011.

[5] V. Murthy, K. Tarakanath, D. Mohanta, and S. Gupta, "Insulator condition analysis for overhead distribution lines using combined wavelet support vector machine (SVM)," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 17, no. 1, pp. 89–99, 2010.

[6] Z. Zhao, N. Liu, and L. Wang, "Localization of multiple insulators by orientation angle detection and binary shape prior knowledge," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 22, no. 6, pp. 3421–3428, 2015.

[7] W. Wang, Y. Wang, J. Han, and Y. Liu, "Recognition and drop-off detection of insulator based on aerial image," in *Proceedings of the International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1, pp. 162–167, Hangzhou, China, December 2016.

[8] Y. Zhai, D. Wang, M. Zhang, J. Wang, and F. Guo, "Fault detection of insulator based on saliency and adaptive morphology," *Multimedia Tools and Applications*, vol. 76, no. 9, Article ID 12051, 2017.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE conference on computer vision and pattern recognition*, vol. 39, pp. 3431–3440, 2015.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, Switzerland, Europe, 2015.

[12] F. Gao, J. Wang, Z. Kong et al., "Recognition of Insulator Explosion Based on Deep learning," in *Proceedings of the International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 79–82, Chengdu, China, December 2017.

[13] Z. Ling, R. C. Qiu, Z. Jin et al., "An accurate and real-time self-blast glass insulator location method based on faster R-CNN and U-net with aerial images," 2018, https://arxiv.org/abs/1801.05143.

[14] C. Sampedro, J. Rodriguez-Vazquez, A. Rodriguez-Ramos, and A. P. Carrio, "Deep learning-based system for automatic recognition and diagnosis of electrical insulator strings," *IEEE Access*, vol. 7, Article ID 101283, 2019.

[15] A. Alahyari, A. Hinneck, R. Tariverdizadeh, and P. David, "Segmentation and defect classification of the power line insulators: a deep learning-based approach," *International Conference on Smart Grids and Energy Systems (SGES)*, pp. 476–481, Perth, Australia, 2020.

[16] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Article ID 13713, Nashville, TN, USA, June 2021.

[17] C. Peng, X. Zhang, G. Yu, and J. Sun, "Large kernel matters--improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4353–4361, Honolulu, HI, USA, July 2017.

[18] C. Wei, Y. Zhao, Y. Zheng, L. Xie, and K. M. Smedley, "Analysis and design of a non-isolated high step-down converter with coupled inductor and ZVS operation," *IEEE Transactions on Industrial Electronics*, vol. 69, pp. 9007–9018, 2021.

[19] L. C. Chen, G. Papandreou, I. Kokkinos, M. Kevin, and L. Y. Alan, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," 2014, https://arxiv.org/abs/1412.7062.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[21] L. C. Chen, G. Papandreou, F. Schroff, and A. Hartwig, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[22] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and A. Hartwig, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Proceedings of the European conference on computer vision (ECCV)*, Springer, Cham, Switzerland, Europe, 2018.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, Honolulu, HI, USA, July 2017.

[24] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," 2015, https://arxiv.org/abs/1506.04579.

[25] D. Xiao, H. Chen, C. Wei, and B. Xiaoqing, "Statistical measure for risk-seeking stochastic wind power offering strategies in electricity markets," *Journal of Modern Power Systems and Clean Energy*, pp. 1–6, 2021.

[26] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," *Proceedings of the European conference on computer vision (ECCV)*, Springer, Cham, Switzerland, Europe, 2018.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, and D. Anguelov, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[31] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," pp. 448–456, 2015, https://arxiv.org/pdf/1502.03167.pdf.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, Las Vegas, NV, USA, June 2016.

[33] S. Chetlur, C. Woolley, P. Vandermersch et al., "Cudnn: Efficient Primitives for Deep learning," 2014, https://arxiv.org/abs/1410.0759.