Progress In Electromagnetics Research M, Vol. 119, 117–128, 2023

Detecting Temperature Anomaly at the Key Parts of Power Transmission and Transformation Equipment Using Infrared Imaging Based on SegFormer

Haozhe Wang[1, 2], Dawei Gong[1], Guokai Cheng[3], Jiang Jiong[3], Dun Wu[4], Xinhua Zhu[5, 1], Shengnan Wu[6], Gaoao Ye[2], Lingling Guo[2, *], and Sailing He[1, 6, *]

Abstract—Methods of manual analysis for infrared image and temperature detection of power transmission and transformation equipment typically have problems, such as low efficiency, strong subjectivity, easy to make mistakes and poor real-time feedback. In this paper, a high temperature anomaly detection method based on SegFormer in infrared image of power transmission and transformation equipment is proposed. Many infrared images of power transmission and transformation equipment are collected and preprocessed, and the temperature information of each infrared image is read out using the DJI sdk tool to construct the temperature data matrix. In the segmentation stage, the SegFormer network is used to segment the key parts of the power transmission and transformation equipment to obtain the mask for detection. The maximum values of the temperature data in the mask area are calculated, and the high temperature anomaly detection at the key parts of the power transmission and transformation equipment is realized. The test results on the test set show that the overall performance of the method is the highest as compared to other methods such as FCN, UNet, SegNet, DeepLabV3+, and an automatic temperature recognition can be realized, which has important practical value for the detection of high temperature anomaly at the key parts of power transmission and transformation equipment.

1. INTRODUCTION

Transmission and transformation equipment is an important part of the power system, and plays an important role [1] in the stable and safe operation of the power system. However, the equipment in the operation process will be affected by the voltage and current, which may produce some damage. In the long-term, the damage can make the equipment defective causing the temperature anomalies of the defective parts, which leads to more serious accidents. This will not only affect the stable operation of the power system, but also cause higher equipment maintenance costs. Therefore, timely and accurate detection of these defects in the transmission and transformation equipment is of great importance for the safety and stability of the power system and the reduction of economic losses.

At present, there are many methods for detecting the defects in power transmission and transformation equipment. Among them, infrared imaging technology is widely used in the field of fault diagnosis of power transmission and transformation equipment due to its non-contact, high precision, safety and other advantages. In related studies people mainly determine whether the power transmission

Received 11 August 2023, Accepted 22 September 2023, Scheduled 28 September 2023
* Corresponding authors: Lingling Guo (guoll@zju.edu.cn), Sailing He (sailing@zju.edu.cn).
[1] National Engineering Research Center for Optical Instruments, Centre for Optical and Electromagnetic Research, Zhejiang University, Hangzhou 310058, China. [2] Research Institute of Zhejiang University-Taizhou, Taizhou 318000, China. [3] State Grid Zhejiang Electric Power Co., LTD., Ningbo Power Supply Company, Ningbo 315010, Zhejiang Province, China. [4] Baolue Technology (Zhejiang) Co., Ltd., Ningbo 315000, Zhejiang Province, China. [5] Taizhou Agility Smart Technologies Co., Ltd, Taizhou, Zhejiang Province, China. [6] Ningbo Innovation Center, Zhejiang University, China.

and transformation equipment in [2] the infrared image is in an abnormal state through digital image processing methods and deep learning methods.

Detection methods based on digital image processing can be roughly divided into threshold segmentation [3], edge detection [4] and region segmentation methods [5]. However, the rationality of a specific method largely depends on the expert experience. Since the infrared image background is complex and the shooting angle is different, the specific parameters need to be set according to the characteristics of a single device in practical application, which makes low detection efficiency and strong subjectivity. The reliability and real-time performance of the device state detection method is difficult to improve.

In recent years, the rapid development of smart grid and deep learning technology [6] has brought new ideas for infrared image extraction methods of power transmission and transformation equipment. Convolutional Neural Network (CNN) [7], by simulating the visual cognitive mechanism of biological brain neurons, can establish a mapping relationship between low-level signals and high-level semantics,

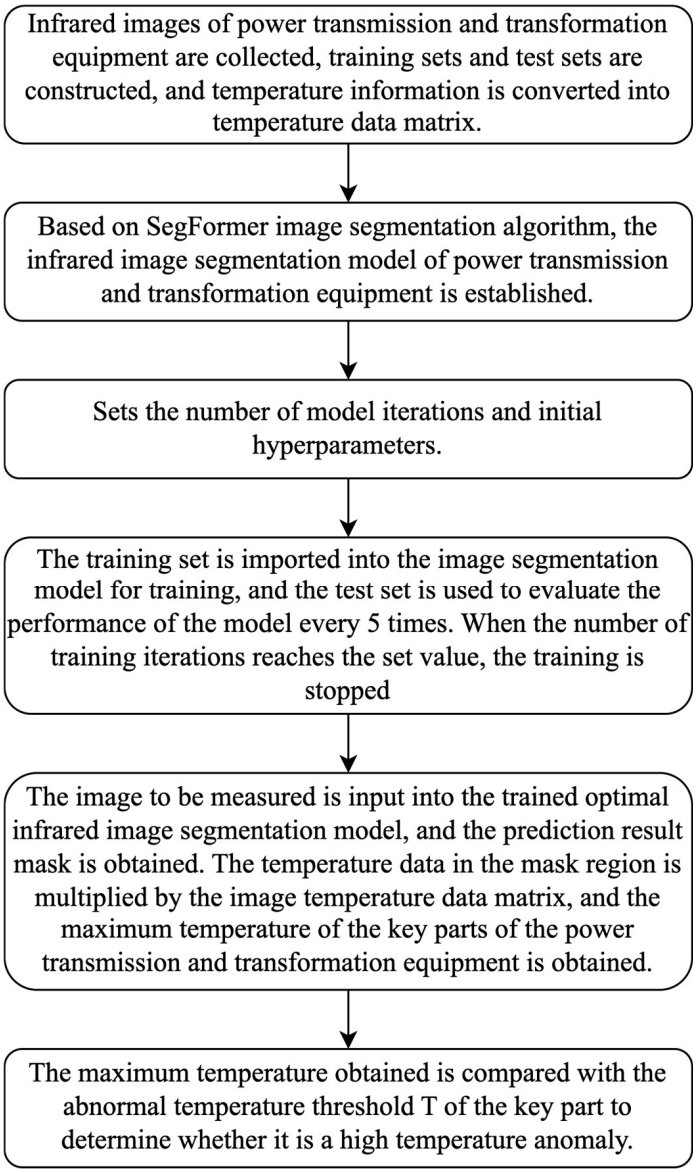


Figure 1. Flowchart of detecting temperature at the key parts of power transmission and transformation equipment.

which makes the field of image segmentation enter a new period of development. Infrared image segmentation models based on deep learning have also been proposed by many scholars. Hou et al. [8] adopted the method of feature extractor plus classifier to identify insulators, but the recognition speed is slow and the accuracy is low. Sampedro et al. [9] used a semantic segmentation method to analyze insulator string. Wang et al. used Mask R-CNN to extract insulator instances in infrared images, and obtained the temperature distribution of each insulator through function fitting. The method realizes the automatic diagnosis of infrared faults in power equipment [10].

However, the traditional CNN mainly focuses on local feature extraction and captures the local information of the image through the convolution operation of the local receptive field. The global context modeling cannot be realized, and the relationship between different positions in the image can not be considered at the same time and the long-distance dependence in the image can not be better captured. Transformer [11] is a kind of deep neural network with global context modeling capabilities based on the self-attention mechanism, which was first used in natural language processing (NLP) field. The great success of Transformer in NLP has also inspired similar approaches in computer vision that have demonstrated similar or better performance [12] than applied convolutional neural networks. As a result, the literature [13, 14] has used ViT (Vision Transformer) for semantic segmentation tasks. The literature [15] points out that although ViT has high performance, there are still some problems, which can only output single-scale features. Therefore, SegFormer architecture was proposed [15], which showed good performance in the semantic segmentation task of public data sets.

Based on the above analysis, in this study, we aim to solve the problem of high temperature anomaly detection through the infrared images at the key parts of power transmission and transformation equipment. To this end, we put forward the following framework, as shown in Fig. 1: First, the infrared images of power transmission and transformation equipment are collected and preprocessed, and the temperature information of each infrared image is read by DJI sdk tool to build the temperature data matrix; Then Segformer network was used to segment the key parts of the power transmission and transformation equipment to get the prediction result mask; Finally, the maximum value of the temperature data in the mask area is calculated, that is, the maximum value of the temperature of the key part of the power transmission and transformation equipment is obtained.

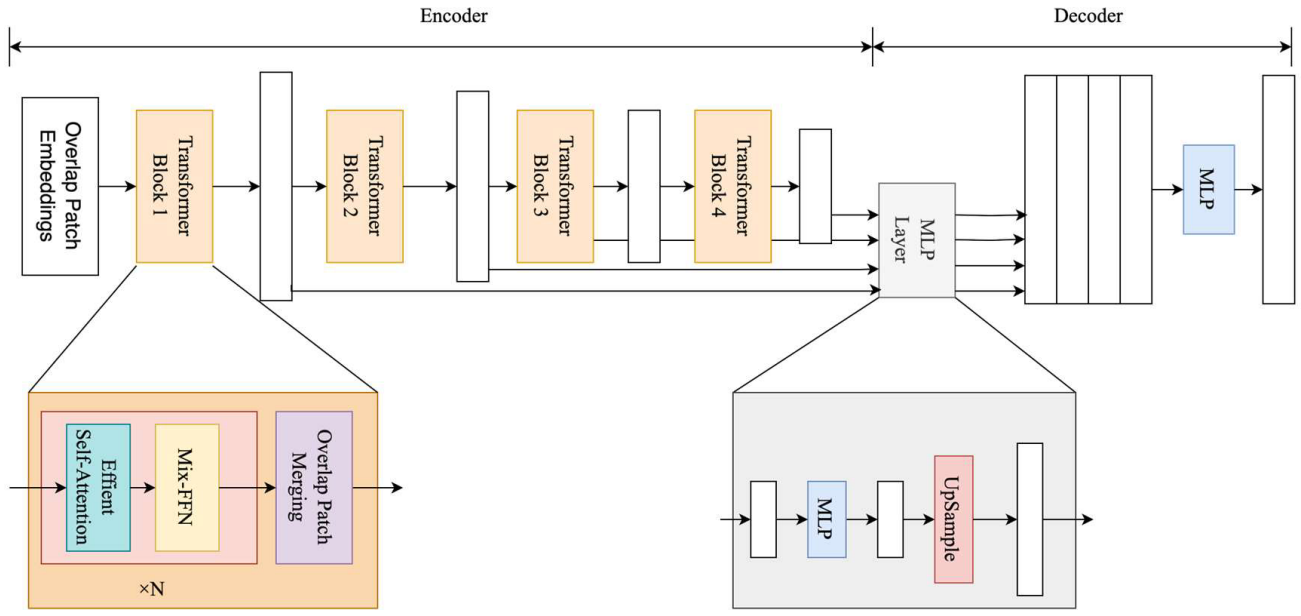


Figure 2. SegFormer network.

2. NETWORK STRUCTURE

As shown in Fig. 2, SegFormer is a simple and efficient semantic segmentation architecture based on Vision Transformer, a framework similar to DeepLabV3+, which is a network with an encoder-decoder structure. However, the encoder module in the SegFormer architecture is a layered Transformer structure consisting of multiple encoder layers to extract multi-scale features, each of which contains a multi-head self-attention mechanism and feedforward neural network. The decoder module is a full multi-layer perceptron (All-MLP), and each decoder layer similarly contains a multi-head self-attention mechanism and feedforward neural network for further processing and enhancement of features. Multiple decoder layers combine local attention and global attention to perform layer-by-layer fusion and upsampling operations of features, aggregate information from different layers, and gradually recover the details of the segmentation results. Finally, the output features of the decoder are passed through a segmentation head for the final segmentation prediction.

2.1. Encoder

SegFormer's encoder (MiT) has four different stages. In MiT's *stage-i*, the input feature map $\tilde{X}_i \in R^{C_i \times H \times W}$ is transformed into a patch embedded in $X_i = \text{PatchMerge}(\tilde{X}_i) \in R^{N \times C_i}$; where C_i, H, W and N are the hidden dimension of the i-th stage, the height of the feature maps, the width of the feature maps and the length of the sequence, respectively. The multi-head self-attention (MHA) for each Transformer layer is then calculated $Q = \text{FC}(X_i) \in R^{N \times C_i}$, $K = \text{FC}(X_i) \in R^{\tilde{N} \times C_i}$ and $V = \text{FC}(X_i) \in R^{\tilde{N} \times C_i}$, where $\text{FC}(\cdot)$ is the linear layer. Both K and V have a small sequence length $\tilde{N} = N/R$, because in the original multi-head self-attention calculation process, Q, K and V of each head have the same dimension $N \times C$, $N = H \times W$ is the length of the sequence, and the time complexity of self-attention calculation process is $O(N^2)$, which is computationally large for high-resolution images. Therefore, a reduction ratio R is used in SegFormer to reduce the length of the sequence, thus reducing the time complexity of self-attention to $O(N^2/R)$. Thus, the output of MHA can be obtained by formula (1) and formula (2)

$$A(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{\tilde{N}}}\right) V \in \mathbb{R}^{N \times C_i} \tag{1}$$

$$X_{mha} = LN\left(\hat{X} + FC(A)\right) \in \mathbb{R}^{N \times C_i} \tag{2}$$

where $LN(\cdot)$ is the layer normalized layer. Then the hybrid feedforward network (Mix-FFN) transforms X_{mha} to X_{ffn}

$$X_{ffn} = LN\left(FC\left(GeLU\left(Conv(FC(X_{mha}))\right)\right) + X_{mha}\right) \tag{3}$$

where $Conv(\cdot)$ represents the convolution operation. SegFormer does not use position encoding in the image patch, taking into account the effect of zero filling on the position information, it directly uses a 3×3 convolution in the FFN.

2.2. Decoder

SegFormer integrates a lightweight decoder that contains only the MLP layer. The key to achieving this simple decoder is that SegFormer's hierarchical Transformer encoder has a larger receptive field than a traditional CNN encoder.

The All-MLP decoder proposed by SegFormer consists of four main steps. First, extract the multilevel feature from the MiT encoder, denoted F_i, and unify the number of channels through the MLP layer, as shown by formula (4). Second, the features $\hat{F}_i$ are upsampled to a quarter size and then concat together, as shown by formula (5). Third, MLP fuses the features after concatenation to get the concatenated feature F, shown by formula (6). Finally, another MLP layer predicts the fused feature F to get a segmentation *Mask* with a resolution of $H/4 \times W/4 \times Num_{cls}$, as shown in formula (7)

$$\hat{F}_i = FC(C_i, C)(F_i), \forall i \tag{4}$$

$$\hat{F}_i = Upsample\left(\frac{W}{4} \times \frac{H}{4}\right)(\hat{F}_i), \forall i \tag{5}$$

$$F = FC(4C, C)\left(Concat(\hat{F}_i)\right), \forall i \tag{6}$$

$$Mask = FC(C, N_{cls})(F) \tag{7}$$

3. EXPERIMENTAL PROCESS AND ANALYSIS

3.1. Collection of Infrared Image Data Set of Power Transmission and Transformation Equipment

To verify the effectiveness of this method, we used the FLIR thermal imager to collect the experimental data set from the equipment of power transmission and transformation stations in Ningbo, Zhejiang Province. The resolution of the collected infrared image was 640×512 pixels, and it was composed of five kinds of power transmission and transformation equipment, including insulators, conductors, tensions, drainage wires and wire clips. Part of the images are shown in Fig. 3. 556 infrared images are selected as the original sample data. Through the image enhancement processing of the samples, such as rotation, Gaussian blur, increasing noise and mirror symmetry, the data is expanded, and finally a infrared image dataset containing 2780 pieces of power transmission and transformation equipment is established. In the data set, the method of random sampling, according to the ratio of $7 : 1.5 : 1.5$, selected 1946 images as training samples, 417 as verification samples, 417 as test samples. The proportion of each device type in the training set, verification set and test set remained the same. The data of the training set, verification set and test set divided by the obtained pictures are shown in Table 1.

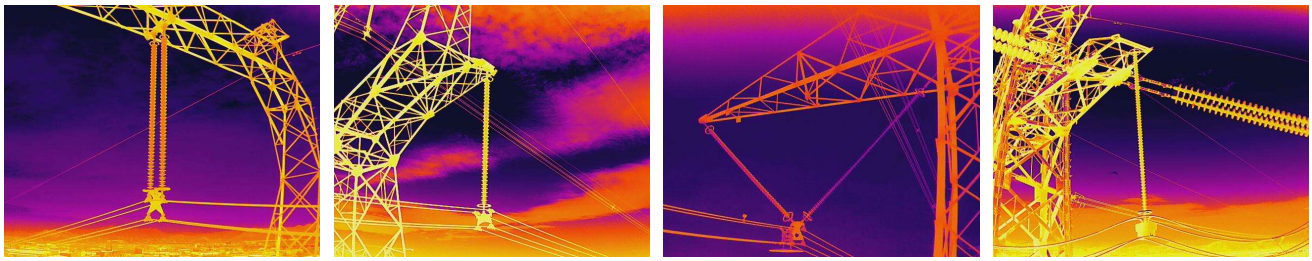


Figure 3. Some infrared images.

Table 1. Partition distribution of our image dataset.

Total number of data samples	Image size	Number of training sets, verification sets, and test sets
2780	640×512	Training set: 1946 Verification set: 417 Test set: 417

3.2. Implementation Details

The experimental platform of this paper is Ubuntu 16.04 operating system, CPU is Intel(R) Xeon(R) Silver 4110, memory is 128 GB, GPU is NVIDIA GeForce RTX 2080Ti, video memory is 11 GB, version is 525.60.11. The CUDA version is 10.1 and cuDNN version is 8.0. The development environment is Python3.7 and Pytorch1.7.

In the training model stage, the processed data set is tested. The training hyperparameters [16] of the proposed algorithm are shown in Table 2. In Table 2, the variable Epoch represents the number of training iterations; Batchsize represents the number of images input to the model during each iteration training. Optimizer represents the optimizer that adjusts the parameters of the model during training; Lr stands for learning rate; Weight_decay is the $L2$ regular term used to prevent overfitting; The learning rate is adjusted by a fixed step. Stepsize indicates the interval for adjusting the learning rate, and Gamma indicates the adjustment multiple. The loss function uses the cross entropy loss function.

Table 2. Hyperparameter settings.

Epoch	Batchsize	Optimizer	Lr	Weight_decay	Stepsize	Gamma
300	8	Adam	0.0005	0.0001	10	0.92

Figure 4 shows the loss of the training set of the model with the increase of the number of iterations under the above hyperparameter setting conditions.

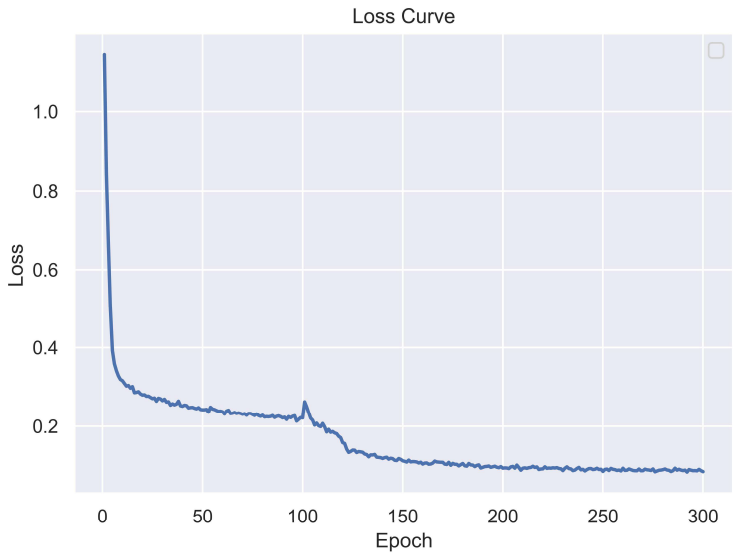


Figure 4. Image segmentation loss curve in the SegFormer model.

3.3. Evaluation Metric

The network is evaluated using six different metrics: Accuracy, Precision, Recall, F1 score (F1), Intersection over Union (IoU), and Mean Intersection over Union, MIoU). These metrics are used in both semantic segmentation and classification problems. All of these metrics are calculated using TP(true positive), FP(false positive), TN(true negative), and FN(false negative).

TP is expressed as the number of pixels that the model predicted to be the target class and are actually the target class; FP represents the number of pixels predicted by the model to be the target class but actually not the target class; FN represents the number of pixels that the model predicts is not the target class, but actually is the target class; TN represents the number of pixels that the model predicts is not the target class, but is actually not the target class.

Accuracy represents the percentage of the total predicted value that the prediction results were correct, as shown in formula (8)

$$Accuracy = \frac{TP}{TP + FN + FP + TN} \tag{8}$$

The accuracy rate is defined in formula (9), which represents the probability that a certain category is predicted correctly in the predicted outcome.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall rate refers to the probability that a category is predicted correctly in true values, as defined in Equation (10).

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

F-measure combines accuracy and recall rate into one measure to comprehensively evaluate the performance of the detection method, which is given in formula (11). In the experiment, the accuracy rate, recall rate, and F1-score are averages of the test set.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

The intersection ratio represents the ratio between the intersection and the union of two sets of predicted values and true values of the model for a certain class. The calculation formula is shown in (12).

$$IoU = \frac{TP}{TP + FN + FP} \tag{12}$$

The average intersection ratio represents the average of the intersection and union ratio between the predicted value and the real value set, which can reflect the degree of overlap between the segmentation result and the real label. The calculation formula is shown in (13).

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{TP + FN + FP} \tag{13}$$

3.4. Comparison of Different Networks

To test the validity of the method used in this paper, FCN [17], U-Net [18], SegNet [19] and DeepLabV3+ [20] semantic segmentation networks are implemented under the same environment, and the comparison experiments with the method in this paper are carried out on the test set. The results are shown in Table 3 and Table 4.

As can be seen from Table 3 and Table 4, compared with other comparison methods, the method in this paper achieves the highest in all indicators except rate Recall. In mIoU, it is 2.06, 0.65, 0.42 and 0.75 percentage points higher than the classical segmentation models in recent years, such as FCN, U-Net, Deeplabv3+ and SegNet, respectively. Accuracy is higher than FCN, U-Net, Deeplabv3+ and SegNet by 8.93, 3.59, 1.7 and 4.47 percentage points, respectively. In Precision, it exceeded FCN, U-Net, Deeplabv3+ and SegNet by 3.94, 1.26, 3.79 and 2.38 percentage points respectively; And beat FCN, U-Net, Deeplabv3+ and SegNet by 6.55, 2.01, 6.36, 2.67 percentage points respectively, in F1. In Recall, although DeepLabV3+ is the highest, DeepLabV3+ is still inferior to SegFormer in terms of overall results. In terms of specific categories, the segmentation results of the proposed method are significantly better than those of other methods in terms of the accurate segmentation of five types of power transmission and transformation equipment, including insulators, conductors, tensioning tubes, drainage wires and wire clips, as well as the background. The segmentation accuracy of the SegFormer model in this paper reached 98.92%, indicating that the method proposed in this paper can effectively segment the key parts of the power transmission and transformation equipment of infrared images, and adapt to the high-precision requirements of actual segmentation. The high performance of the Segformer model on Precision, Recall and F1 shows that it can better distinguish the pixels of the key parts of the equipment from the background pixels. The performance of the network also achieves good results on the main evaluation indexes IoU and MIoU, which also shows that the network has strong generalizability.

In this paper, the experimental results of U-Net model, SegNet model, DeeplabV3+ model and SegFormer model with good performance in the experiment are visualized on the test set, as shown in

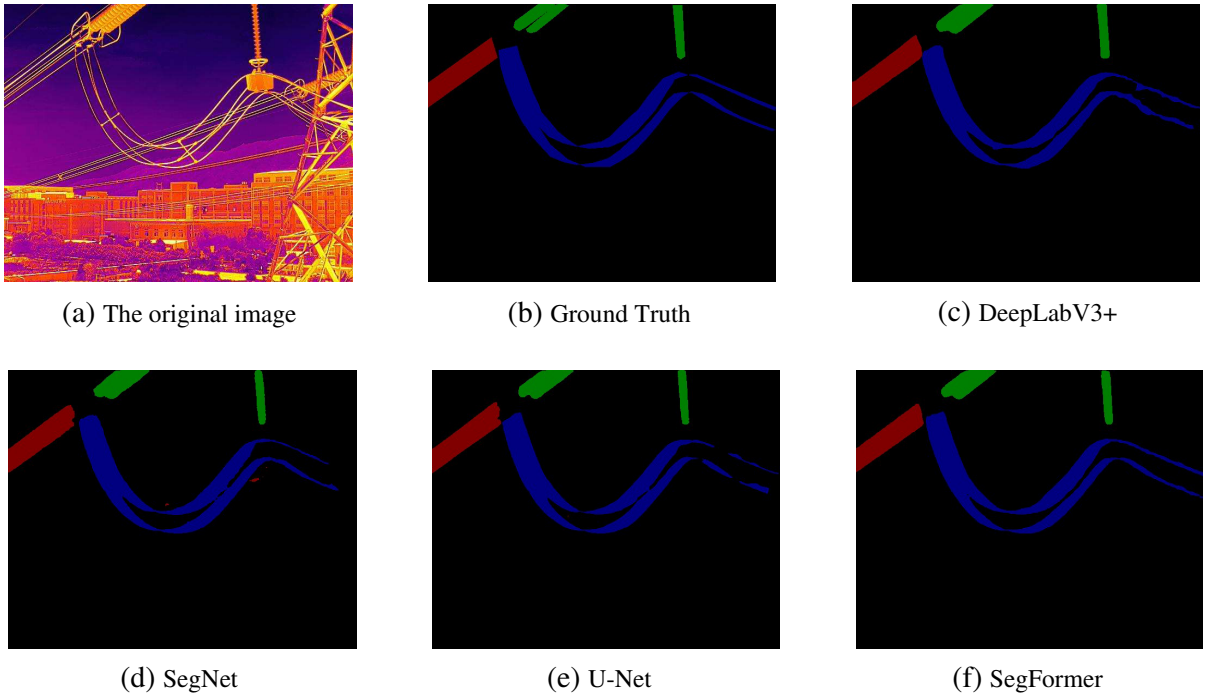
Table 3. Comparison results of different models on IoU and MIoU.

Models	Categories	IoU	MIoU
FCN	Insulator	0.8141	0.7755
	Wire	0.7462	
	Strained tube	0.6555	
	Drain wire	0.5929	
	Clamp	0.8787	
	Background	0.966	
U-Net	Insulator	0.8437	0.8435
	Wire	0.8534	
	Strained tube	0.7230	
	Drain wire	0.7768	
	Clamp	0.8831	
	Background	0.9812	
DeepLabV3+	Insulator	0.86	0.8624
	Wire	0.9043	
	Strained tube	0.7295	
	Drain wire	0.8011	
	Clamp	0.8941	
	Background	0.9853	
SegNet	Insulator	0.843	0.8347
	Wire	0.8463	
	Strained tube	0.7056	
	Drain wire	0.7569	
	Clamp	0.8761	
	Background	0.9801	
SegFormer	Insulator	0.8792	0.8794
	Wire	0.9220	
	Strained tube	0.7553	
	Drain wire	0.8357	
	Clamp	0.8924	
	Background	0.9882	

Table 4. Comparison results of different models in Accuracy, Precision, Recall and F1.

Models	Accuracy	Precision	Recall	F1
FCN	0.9686	0.8359	0.9039	0.8686
U-Net	0.9827	0.9126	0.9155	0.9140
DeepLabV3+	0.9850	0.8873	0.9544	0.8705
SegNet	0.9817	0.9014	0.9136	0.9074
SegFormer	0.9892	0.9252	0.9433	0.9341

Fig. 5, where (a) represents the original test image, (b) represents the ground truth corresponding to the original test image, (c) represents the prediction result of DeepLabV3+ model, (d) represents the prediction result of SegNet model, (e) represents the prediction result of U-Net model, and (f) represents the prediction result of SegFormer model. In the prediction results of each model, the pixel value of



(a) The original image (b) Ground Truth (c) DeepLabV3+

(d) SegNet (e) U-Net (f) SegFormer

Figure 5. Comparison of segmentation results.

0 in the output represents the background, the pixel value of 1 represents wires, the pixel value of 2 represents insulators, and the pixel value of 3 represents drainage lines. For better visualization, we use pseudo colors to represent different pixel values, i.e., black represents the background, red represents wires, green represents insulators, and blue represents drainage lines. As can be seen from the actual test results, the segmentation results of the proposed algorithm are basically consistent with the results of experts' manual labeling. In addition, to pay more attention to the segmentation of the target, the invalid background interference information was also suppressed. Compared with other methods, the present method can retain image details better. From the perspective of the overall visual effect, no matter the background of the aerial image is simple or complex, the SegFormer model used in this paper can not only segment the target clearly, but also segment the target more finely, with better usability and robustness.

3.5. Temperature Detection

The infrared image of transmission and transformation equipment taken by handheld infrared thermal imager is taken as an example to illustrate the automatic detection method proposed in this paper. The shooting scene is the substation of a company of Zhejiang Power Grid, and the instrument used is FLIR T640 infrared thermal imager. By SegFormer deep learning segmentation model, the pixel-level segmentation results of the infrared image are shown in Fig. 6. Wires, insulators and drainage wire are segmented at the pixel level, and the segmentation result figure mask is obtained. In the mask, the neural network divides the input infrared image into four categories, including the background represented by the black color with output pixel value of 0, a wire represented by the red color with output pixel value of 1, an insulator represented by the green color with output pixel value of 2, and a drainage line represented by the blue color with output pixel value of 3. The purpose of this segmentation is to separate the key parts of the equipment from the complex background, in order to reduce the interference from the complex background and measure more accurately the temperature of the key parts in the image. Then the mask is multiplied with the image temperature data matrix obtained from DJI sdk to obtain the temperature data in the mask region. Finally, the maximum temperature of key parts of power transmission and transformation equipment as shown in Fig. 7 is obtained. The

Figure 6. Segmentation mask.

Figure 7. Temperature data image.

obtained maximum temperature of the wire, the insulator and the drainage wire are 37.4°C, 35°C and 39.8°C respectively, all of which are lower than their respective temperature thresholds, that is, the temperature of the equipment is normal.

4. CONCLUSION

To overcome the problem of low efficiency of manual inspection of power transmission and transformation equipment, this paper proposes a method of infrared image high temperature anomaly detection of power transmission and transformation equipment based on SegFormer. The collected infrared images of power transmission and transformation equipment are preprocessed, and the temperature information of each infrared image is read out by using DJI sdk tool to construct the temperature data matrix. In the segmentation stage, SegFormer network was used to segment the key parts of the power transmission and transformation equipment to get the prediction mask. Then, the maximum values of the temperature data in the mask area are calculated, and the high temperature anomaly detection of the key parts of the power transmission and transformation equipment is realized. The test results of this method show that the model proposed in this paper reaches the highest in all indexes except rate Recall, among which MIoU, Accuracy, Precision, Recall and F1 reach 0.8794, 0.9892, 0.9252, 0.9433 and 0.9341 respectively. It also shows that the method proposed in this paper can effectively segment the key parts of the transmission and transformation equipment through the infrared images, and can meet the high precision requirements of practical segmentation. The method proposed in this paper makes it possible to diagnose the fault of transmission and transformation equipment on site, which is of great significance for reducing the monitoring cost of transmission and transformation equipment and improving the intelligent level of substation.

ACKNOWLEDGMENT

The work is partially supported by “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2023C03135), Ningbo Science and Technology Project (2021Z029), and National Natural Science Foundation of China (11621101). The authors are grateful to Dr. Julian Evans of Zhejiang University for valuable discussions.

REFERENCES

1. Wang, T., W. Liu, J. Zhao, et al., "A rough set-based bio-inspired fault diagnosis method for electrical substations," *International Journal of Electrical Power and Energy Systems*, Vol. 119, 105961, 2020.
2. Li, Y. J., H. T. Li, S. Q. Song, et al., "Research on temperature detection of internal conductor in GIS basing on infrared thermal imaging," *Electric Power Engineering Technology*, 142–146, 2019.
3. Chen, F., J. G. Yao, Z. S. Li, et al., "The method to extract shed surface image of a single insulator from infrared image of a insulator string," *Power System Technology*, 220–224, 2010.
4. Tang, Q. J., J. Y. Liu, Y. Wang, et al., "Infrared image edge recognition and defect quantitative determination based on the algorithm of fuzzy C-means clustering and canny operator," *Infrared and Laser Engineering*, 281–285, 2016.
5. Cui, J. Y., Y. D. Cao, and W. J. Wang, "Application of an improved algorithm based on watershed combined with Krawtchouk invariant moment in inspection image processing of substations," *Proceedings of the CSEE*, Vol. 35, No. 6, 1329–1335, 2015.
6. Gong, D., T. Ma, J. Evans, and S. He, "Deep neural networks for image super-resolution in optical microscopy by using modified hybrid task cascade U-Net," *Progress In Electromagnetics Research*, Vol. 171, 185–199, 2021.
7. Zhang, X., W. Lin, M. Xiao, and H. Ji, "Multimodal 2.5D convolutional neural network for diagnosis of Alzheimer's disease with magnetic resonance imaging and positron emission tomography," *Progress In Electromagnetics Research*, Vol. 171, 21–34, 2021.
8. Hou, C. P., H. G. Zhang, W. Zhang, et al., "Identification method for spontaneous explosion defects of transmissionline insulators," *Electrical Power System Automatic*, Vol. 31, No. 6, 1–6, 2019.
9. Sampedro, C., J. Rodriguez-Vazquez, A. Rodriguez-Ramos, et al., "Deep learning-based system for automatic recognition and diagnosis of electrical insulator strings," *IEEE Access*, Vol. 7, 101283–101308, 2019.
10. Wang, B., M. Dong, M. Ren, et al., "Automatic fault diagnosis of infrared insulator images based on image instance segmentation and temperature analysis," *IEEE Transactions on Instrumentation and Measurement*, Vol. 69, No. 8, 5345–5355, 2020.
11. Vaswani, A., N. Shazeer, N. Parmar, et al., "Attention is all you need," *Proceedings of the International Conference on Neural Information Processing Systems*, 6000–6010, 2017.
12. Han, K., Y. Wang, H. Chen, et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 1, 87–110, 2023.
13. Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," [EB/OL], 2020, https://arxiv.org /abs/2010.11929.
14. Zheng, S., J. Lu, H. Zhao, et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6881–6890, 2021.
15. Xie, E., W. Wang, Z. Yu, et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Proceedings of the International Conference on Neural Information Processing Systems*, 2021.
16. Wang, X. J., Z. N. Zheng, Y. C. Fang, et al., "Defect diagnosis method for composite insulators based on U-net segmentation," Fujian Province: CN114037694A, Feb. 11, 2022.
17. Long, J., E. Shelhamer, T. Darrell, et al., "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440, 2015.
18. Ronneberger, O., P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241, 2015.
19. Badrinarayanan, V., A. Kendall, and R. Cioplla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 12, 2481–2495, 2017.

20. Chen, L. C., Y. Zhu, G. Papandreou, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Proceedings of the European Conference on Computer Vision*, 801–818, 2018.