The Fusion Oil Leakage Detection Model for Substation Oil-Filled Equipment

Zhenyu Chen[1(✉)], Lutao Wang[1], Siyu Chen[1], and Jiangbin Yu[2]

[1] Big Data Center, State Grid Corporation of China, Beijing, China
czy9907@163.com
[2] Anhui Jiyuan Software Co., Ltd., Hefei, Anhui, China

Abstract. Under the condition of long-term high-load operation, substation equipment is prone to oil leakage, which affects the operation safety of substation equipment and the stability of the power system. This paper proposes an oil leakage detection technology based on the fusion of simple linear iterative clustering (SLIC) and Transformer sub-station equipment, which is used to solve the problem of intelligent identification of oil leakage in oil-filled equipment such as transformers and transformers in substations. This paper first uses the SLIC method to segment the image to obtain superpixel of image data, and then uses the DBSCAN method based on linear iterative clustering to cluster similar superpixels. After training and learning, obtain the oil model with stable and accurate identification of oil leakage of substation oil-filled equipment. The experimental results show that the method proposed in this paper can efficiently identify oil leakage of substation equipment under the premise of ensuring stability, with an average recognition accuracy rate of 87.1%, which has high practicability and improves the detection and identification ability of oil leakage.

Keywords: Substation · Vision transformer · Simple linear iterative clustering · Oil leakage

1 Instruction

At present, most high-voltage equipment such as transformers, voltage/current transformers and capacitors in substations use insulating oil as insulating material, which can achieve insulation, cooling and arc extinguishing of high-voltage equip-ment. If the oil-filled equipment leaks oil, it will affect the safe and stable operation of the power grid and reduce the service life of the equipment. Therefore, it is necessary to study a method for detecting oil leakage of substation equipment, so as to realize the timely detection of oil leakage detection of equipment and improve the stability of power grid operation.

Substations are important nodes for stable and continuous power transmission. Traditional substation inspections rely on on-site inspections by professionals, result-ing in high inspection costs and low inspection efficiency. In addition, there are a lot of unsafe factors in the inspection process, which affects the personal safety of inspectors. With the use of monitoring and inspection robots in substation inspections, the pressure

G. Zhai et al. (Eds.): IFTC 2022, CCIS 1766, pp. 59–74, 2023.
https://doi.org/10.1007/978-981-99-0856-1_5

of manual inspections has been greatly reduced, but inspections rely on artificial intelligence recognition [1]. The combination of deep learning technology and substation inspection can greatly improve the detection efficiency [2]. Figure 1 above shows the oil leakage in several typical scenarios of the substation. The following problems can be summarized from the figure: (1) There are many equipment with oil leakage, including transformers, transformers, capacitors and other equipment. (2) The observable parts of oil leakage exist not only on the surface of the equipment, but also on the ground, with various shapes and transparency. (3) The leaking oil is transparent and similar to the shadow color of the equipment, and has no self-fixing characteristics.

Therefore, this paper proposes an oil leakage detection technology for substation equipment based on fusion SLIC, which can highlight the location of oil leakage, re-duce the influence of complex background, and improve the detection accuracy.

Fig. 1. Schematic diagram of oil leakage from substation equipment. In the picture, the background of the oil leakage part is complicated and the color is darker

At present, there are few researches on the detection of oil leakage from substation equipment. The traditional method relies on inspectors to irradiate easy-to-penetrate points such as casings and welds with flashlights, and make visual inspections through reflection, but this method has limitations in the inspection of warehouses and elevated equipment. Or regularly observe and judge through the oil level gauge, the timeliness is low.

Dong Baoguo [3] detected and segmented abnormal areas by difference method based on the color of leaking oil, and compared the color characteristics of abnormal areas in two images to obtain the result of oil leakage. However, this method relied on

pictures taken when the leaking parts did not leak in the early stage for comparison. Wang Yan fused the OTSU algorithm with the detected oil leakage area by using the difference method and the segmentation method of monitoring target image, com-pared the images before and after oil leakage in the area, and analyzed and judged the oil leakage area by using the HS color histogram method. This method still relied on the images before and after oil leakage [4], which had limitations. In order to improve the detection rate of oil spill targets and reduce the influence of shadow lighting on the detection model, Huang Wenli et al. [5] proposed an attention segmentation network based on edge fusion, which made full use of the spatial background information of oil spill forms and proposed a self-attention mechanism to improve the detection rate of oil spill. Yang Minchen and Zhang Yan et al. [6] irradiated the oil leakage position with ultraviolet flashlight based on the fluorescence characteristics of the oil leakage. In a dark environment, the oil leakage position would be purple and prominent, but this method could only be detected in the dark and had limitations in the daytime sunshine conditions. Wu et al. [2] studied the detection method based on visible image information of oil leakage, used lightweight Mobilenet-SSD deep net-work model to train oil leakage pictures, and deployed them in edge equipment to achieve intelligent positioning and detection of oil leakage. This method has high practicability. Although machine learning is extremely capable of learning image features, it has limitations in the face of challenges such as the complex background and obscure features of oil seeps.

Image segmentation [7] provides an ideal method to solve images with complex background interference and is one of the key technologies in CV field, especially color image segmentation [8], which can extract interesting or meaningful pixel sets and features in images [9]. Watershed segmentation algorithm [10], based on the similarity criterion, utilizes morphology and topological theory to traverses pixel sets and sub-merges pixels according to the threshold value. If the threshold value is greater than, a boundary will be formed to realize the classification of neighborhood pixels. This algorithm is susceptible to noise. In recent years, with the rapid development of artificial intelligence, image segmentation based on graph theory has also attracted widespread attention [11]. Image segmentation based on graph theory continuously optimizes the weight of pixel edge set after segmentation to achieve the purpose of minimum segmentation through optimization processing. GrabCut is a typical segmentation method based on graph theory [12–14]. Users input an bounding box as the seg-mentation target location to achieve the separation and segmentation of targets and complex backgrounds. However, this method has problems such as high time complexity and poor processing quality when targets and backgrounds are similar. Simple Linear Iterative Clustering (SLIC) algorithm shows advantages in generating subimages with good boundary compliance [15, 16]. SLIC is a super pixel algorithm based on K-means clustering, which has the advantages of low time complexity and better edge fitting [17]. In addition, density-based noise application spatial clustering (DBSCAN) [18, 19] performs well in grouping sub-images belonging to the same clus-ter.

Also, Vaswani et al. [20] proposed Transformer for the first time, establishing a new encoder-decoder architecture based on multi-head self-attention mechanism and feed-forward neural network. Then, Dosovitskiy et al. proposed the so-called ViT (Vision Transformer) [21], which is a complete Transformer, and has superior performance in

image classification task when it is directly applied to image patch sequence. Additionally, the training process is also able to greatly simplified due to the unique advantage of the deep learning method [22–26].

In this paper, through the research of superpixel segmentation and oil leakage detection, the oil leakage detection of oil filling equipment in the substation scene is realized. Firstly, the method uses SLIC technology to perform super-pixel segmentation on oil leakage image and obtain the super-pixel segmentation result. Then, DBSCAN technology was used to cluster the segmentation results to highlight the oil leakage area. Then the image is recognized by ViT, and good recognition results are obtained. Finally, the effectiveness and feasibility of the proposed method are verified by experiments in substation scenarios. The flow of oil leakage identification method is shown in the figure below (Fig. 2).

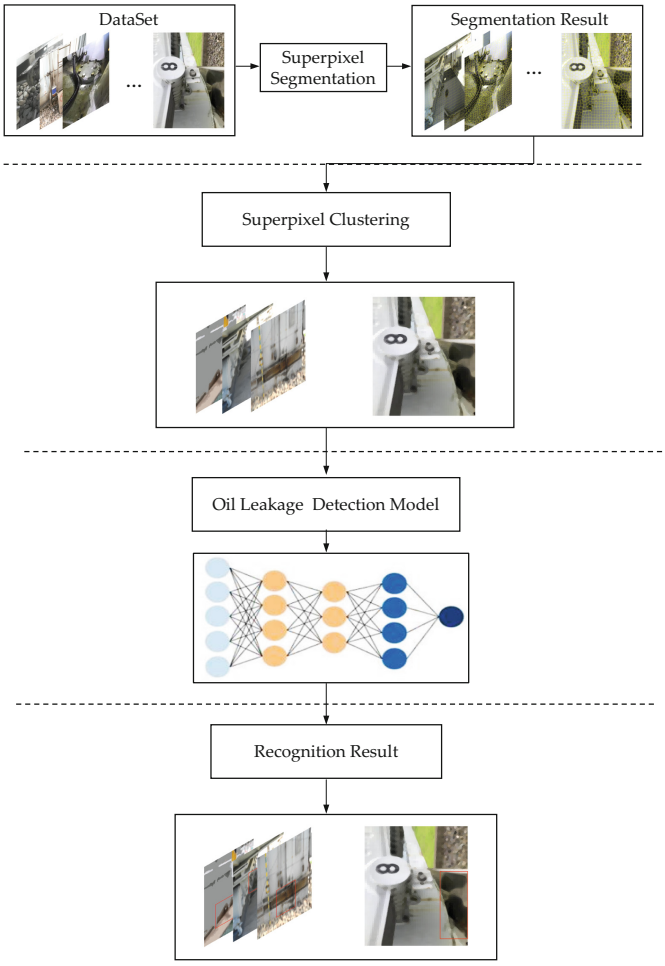

Fig. 2. Flowchart of the oil leakage detection

2 Oil Leakage Detection Based on Fusion SLIC and Transformer

2.1 Superpixel Segmentation Based on SLIC

The SLIC algorithm divides the image into superpixels, and each region has the same size and is named S. The geometric center of each region is considered as the center of the superpixel, and the coordinates of the center are updated at each iteration. Superpixels are grouped according to measurements of spatial distance d_s and d_c intensity (a measure of spatial and intensity distance).

$$d_s = \sqrt{\left(x_j - x_i\right)^2 + \left(y_j + y_i\right)^2} \tag{1}$$

$$d_c = \sqrt{\left(I_j - I_i\right)^2} \tag{2}$$

In the above formula, (x, y) represents the position of each pixel, and $\left(I_j, I_i\right)$ represents the normalized pixel intensity.

Introducing the total distance of two measurement units d_s and d_c, calculated as follows:

$$D{=}\sqrt{d_c^2 + \left(\frac{d_s}{S}\right)^2 m^2} \tag{2}$$

In the above formula, m represents the compactness coefficient. The larger the parameter m, the more compact the generated superpixel area; on the contrary, the more superpixels fit the contour of the image, but the size and shape will be irregular. Figure 3 shows the results of oil leakage data based on SLIC superpixel segmentation.

2.2 Superixel Clustering Based On DBSCAN

The main idea of DBSCAN clustering is as follows: in two-dimensional space, the neighborhood within the radius of a given object is called Eps of the object, and if the Eps of the object contains at least the minimum threshold MinPts of objects with simi-lar attributes, the object is called core object. For any sample that is in the domain of the core object, it is called density direct. DBSCAN searches the cluster by examining Eps at each point in the data set. If the Eps of point p contains more than MinPts, a new cluster with p as the core object is created. DBSCAN then iteratively collects density-reachable objects directly from these core objects, which at the same time involves merging several density-reachable clusters. The process terminates when a new point cannot be added to any cluster.

The oil leakage image can be regarded as a special spatial data set, in which each pixel has a position coordinate and corresponding color value. By finding spatial clus-ters, clusters in the oil leakage image can be found effectively. Pixels with similar col-ors and spatial connections can be grouped together to form a segmented area. The difference between spatial clustering and pixel clustering lies in that image pixels are not only distributed in spatial space, but also in other feature Spaces such as color. The pixels

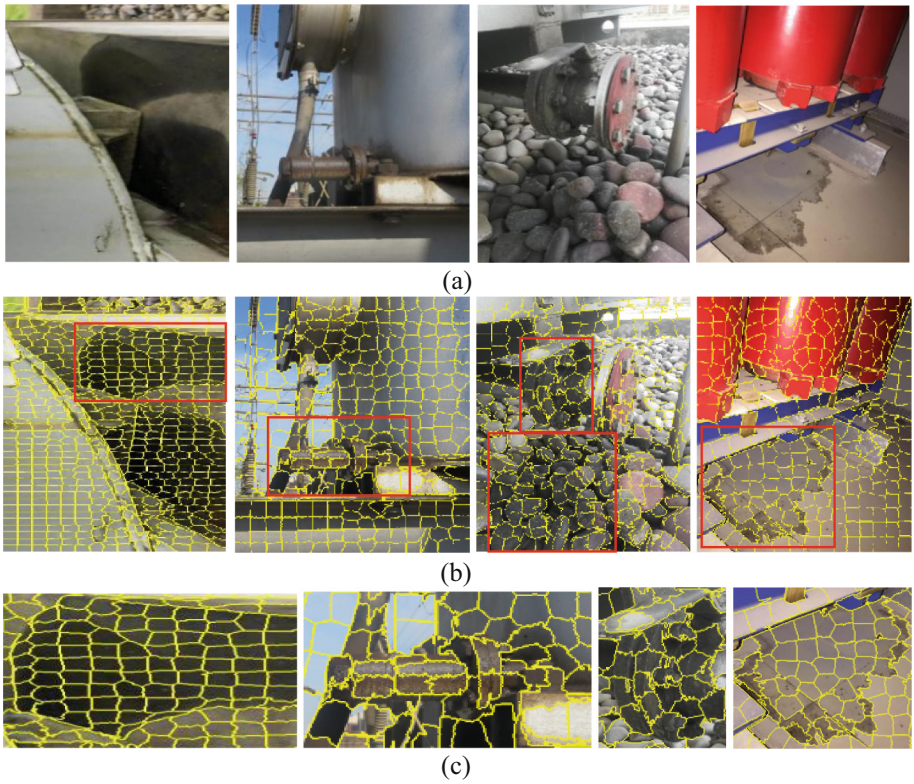

Fig. 3. Image super-segmentation results based on SLIC. (a) Pictures of oil leakage from equipment in substations; (b) Superpixel segmentation of images; (c) Local magnification of oil leakage

divided into a cluster should not only be spatially connected, but also similar in color. Table 1 shows the image clustering process of leaking oil based on DBSCAN.

Figure 4 shows the DBSCAN based superpixel clustering result. As can be seen from the figure, after DBSCAN clustering, oil stains with similar features are clustered together, eliminating other unrelated features and inhibiting complex background, which is conducive to the detection of oil leakage.

2.3 Oil Leakage Detection and Analysis Based on Transformer

In this paper, for the convenience of description, the service coding is simplified to X ($X = A, B, C\ldots$), the service node coding is simplified to i ($i = 1, 2, 3\ldots$), thus the service node identification is simplified to X_i ($X = A, B, C\ldots; i = 1,2,3\ldots$). The service overall topology diagram is shown in Fig. 1. The topology diagram involves four services, namely, A, B, C and D. The service A is provided by service node A_1. The service B can be provided by service node B_1, B_2, B_3, B_4, B_5. The service C is supplied by service node C_1, C_2 and C_3. The service D is provided by service node D_1 and D_2.

Table 1. Clustering process based on DBSCAN algorithm

Steps	Process
1	Input: Superpixel segmentation of images;
2	Get all pixels, assuming there are N;
3	Initialization parameters: ε, nb_min_points, $n = 0$;
4	For a new data point:
5	Find all reachable points using ε and nb_min_points;
6	Determine whether pixel n is an isolated point:
7	If yes, discard the noise point and go back to step 4;
9	If no, get a cluster;
10	Check if n is less than N:
11	If yes, go back to step 4 to continue execution;
12	If not, output the clustering result;
13	Output: Clustering results

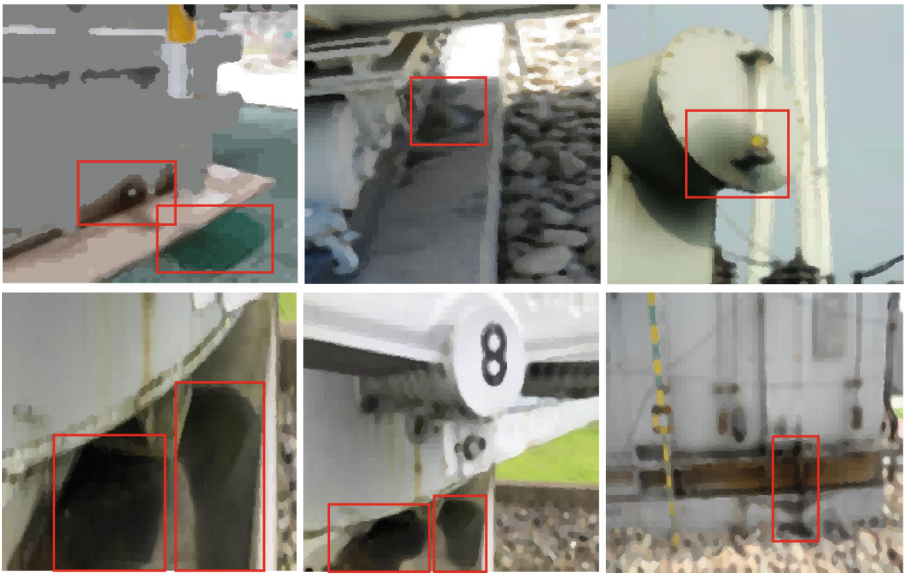

Fig. 4. Superpixel clustering results based on DBSCAN. It can be seen from the figures that after DBSCAN pro-cessing, the background is weakened, and the oil leakage part is more prominent

The traditional Transformer is mainly composed of two parts: encoding and decoding. The multi-head attention mechanism is the core of the Transformer, which enables the model to remember the key information in the picture like the human visual attention. Refer to the image sequence processing method mentioned in the paper [21]. First, the image is cut, and the image is divided into several image blocks; secondly, the image block is sent to the trainable linear projection layer, and position encoding is performed.

Before sending the image to the encoder, the extracted image features need to be positioned. Coding, the position coding adopts the sine and cosine function to generate the position code, and then adds it to the feature image of the corresponding position. The position coding adopts the random initialization method, and the position coding function is:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{mod } el}}}}\right) \tag{4}$$

$$PE(pos, 2i{+}1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{mod } el}}}}\right) \tag{5}$$

In the above formula, pos is the absolute position of pixels in the feature graph, $d_{\text{mod } el}$ is the dimension of the image, $2i$ and $2i + 1$ represent parity.

The embedded patch and position encoding are superimposed to obtain an embedded vector, which is sent to the Transformer encoding layer for processing. Transformer encoding layer consists of multi-head attention and multi-layer perceptron. As shown in Fig. 5, multi-head attention contains multiple attention mechanisms, and a single attention contains query matrix, key matrix and value matrix, which are multiplied by the embedding vector The weight matrix is obtained:

$$Q = X \times W^{Q} \tag{6}$$

$$K = X \times W^{K} \tag{7}$$

$$V = X \times W^{V} \tag{8}$$

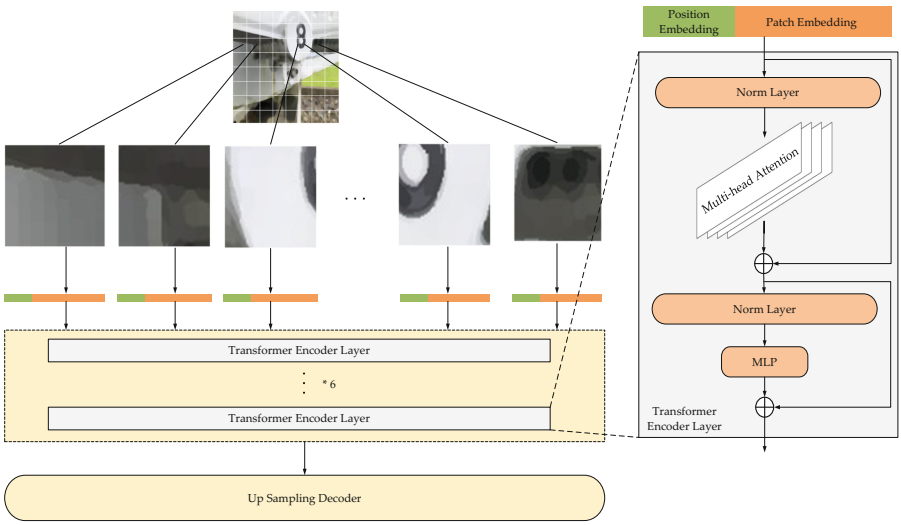

Fig. 5. Transformer Encoder structure

In the above formula, Q is query matrix, K is key matrix, V is value matrix, X is output embedding vector, W^Q, W^K and W^V corresponds to the weight matrix, respectively. The final output of the self-attention mechanism is:

$$Z = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

In the above formula, d_k is dimension of K.

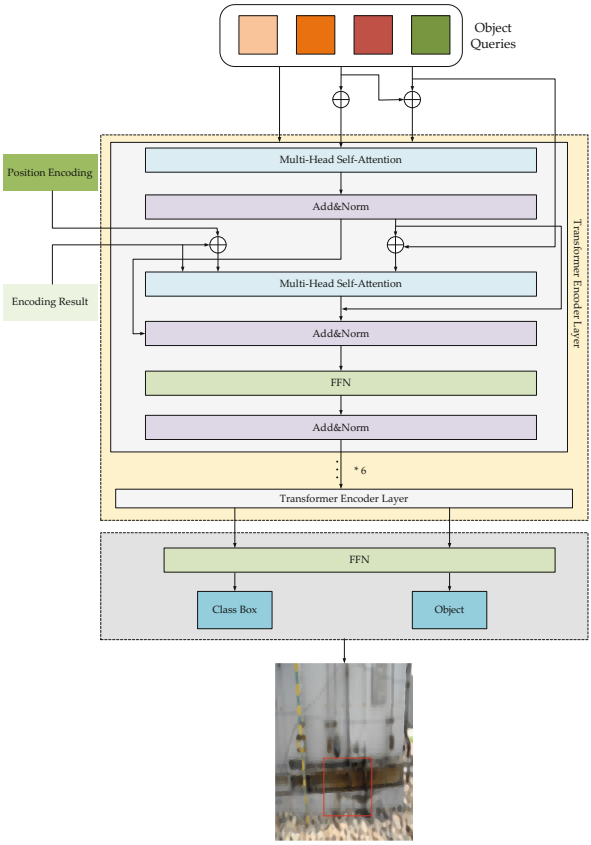

Fig. 6. Transformer Decoder structure diagram

Through the Transformer encoder, the features of the input image can be extracted. Unlike the RNN operation, there is no need for a convolutional neural network as the backbone network.

The key vector and value vector output by the encoder form a self-attention vector set, and the self-attention vector set is input to the decoding module to help the decoding module pay attention to which part of the input oil leakage image is the focus area. The decoder consists of multi-head attention and FNN layers, the location of oil leakage in

the oil leakage image is obtained by three-layer linear transformation and ReLU in FFN, and the category of the object is obtained by a single linear layer. The decoder is shown in Fig. 6.

The entire oil leakage identification model is divided into three parts. The back-bone network extracts image features, the encoder-decoder performs information fu-sion, and the feedforward network performs prediction. As shown in Fig. 7 below, the backbone network is used to learn to extract the features of the original image.

The encoder reduces the dimension of the input feature image and converts the two-dimensional feature image into one-dimensional feature image structure. Finally, the output of the top encoder is an attention vector set containing key vector and value vector. The decoder uses a small number of fixed query vectors (N) as input, and different query vectors correspond to different output vectors. The query vector is then decoded into box coordinates and class labels via FFN, resulting in N final predictions. The following figure shows the identification process of seepage oil based on ViT.

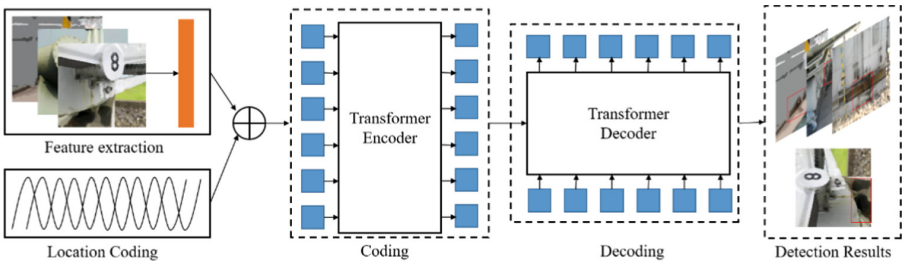

Fig. 7. Oil leakage detection results base on Transformer. The detection process includes n encoding modules and n decoding modules, 8 encoding modules and 8 decoding modules are used in this paper

3 Basis of Model Training

In this paper, the image recognition method based on the fusion SLIC method is adopted to verify the validity and accuracy of the image data of oil leakage from the substation end filling equipment. An image recognition experiment is designed to test the accuracy and difference between the proposed method and Transformer and faster-RCNN.

3.1 Software and Hardware

The test conditions of this paper are: CentOS 8, 64-bit operating system, Pytorch framework. Computer configuration: Desktop COMPUTER, NVIDIA TESLA P100, 32 GB video memory; E5-2680 V4 CPU processor, maximum main frequency 3.30 GHz, disk capacity 500 GB, Python programming language.

3.2 Path Planning

The original data set of this paper takes images for substation inspection, with a total of 4400 images of substation business scenes. In this paper, 220 images of about 5% of the 4400 images were randomly selected as the final test data, and the remaining 4180 images were used as the training data set. The original image contains two types of working conditions, oil seepage and oil leakage, which are not evenly distributed in the image. The image of the same scene is shot from multiple angles and the background is complex. The image data of oil leakage conditions of the two types of oil filling equipment used in this paper are as Fig. 8.

(a) (b)

Fig. 8. Oil leakage images. (a) Oil spill image, Oil spills are mainly distributed on the ground; (b) Oil seepage image, Oil seepage is mainly distributed on the surface of the equipment

3.3 Experimental Results

Accuracy (P), Recall (R) and Average Precision (AP) are used as evaluation indexes to evaluate the proposed method. Among them, the calculation method of AP value refers to the calculation method of Everingham et al. The calculation formula of accuracy and recall rate is as follows:

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

where TP (True Position) is a positive sample that is predicted to be a positive sample, FP (False Position) is a Negative sample that is predicted to be a positive sample, and FN (False Negative) is a positive sample that is predicted to be a Negative sample.

In this paper, the data set of oil leakage working conditions of oil filling equipment in substation scenario is classified. The data set includes 4400 samples in total, and the number of samples of each working condition is shown as Table 2.

This paper first tests the difference of image classification results between the original sample and the expanded sample. Therefore, the image recognition models of the

Table 2. Number of samples of defect categories of electric oil-filled equipment

Type	Oil seepage images	Oil spill images
number	2300	2100

proposed method, Transformer and the Faster -RCNN method are trained by using the original data set and the expanded data set. The training data set of experiment 1 was 2200 original images, the training data set of Experiment 2 was an expanded image data set containing 3200 images, and the training data set of Experiment 3 was an expanded image data set containing 4400 images. For the data of the three experiments, 70% were selected as the training set and the remaining 30% as the verification set.

As can be seen from Table 3, Table 4 and Table 5, the method presented in this paper has shown excellent performance in experiments with different data amounts of 2200, 3200 and 4400. Among them, the identification accuracy of the method presented in this paper is 3.53% higher than Transformer on average and 11.50% higher than Faster-RCNN method on average. The identification accuracy of oil leakage is 2.00% higher than ViT and 15.9% higher than Faster-RCNN method on average. The proposed method has an average identification Precision of 4.93% higher than ViT, 16.27% higher than Faster-RCNN method, 12.53% higher than ViT and 11.70% higher than Faster-RCNN method in oil leakage category. The recognition recall rate of the proposed method in oil leakage class is 6.53% higher than ViT, 15.97% higher than Faster-RCNN method, 3.53% higher than ViT and 8.13% higher than Faster-RCNN method in oil leakage class (Figs. 9, 10 and 11).

Table 3. Accuracy comparison of table

Type	Oil seepage images (2200)	Oil seepage images (3200)	Oil seepage images (4400)	Oil spill images (2200)	Oil spill images (3200)	Oil spill images (4400)
Our method	88.3%	87.7%	89.1%	87.1%	86.5%	87.3%
ViT	84.6%	84.1%	85.8%	84.7%	84.9%	85.3%
Faster-RCNN	76.1%	77.4%	77.1%	69.5%	71.2%	72.5%

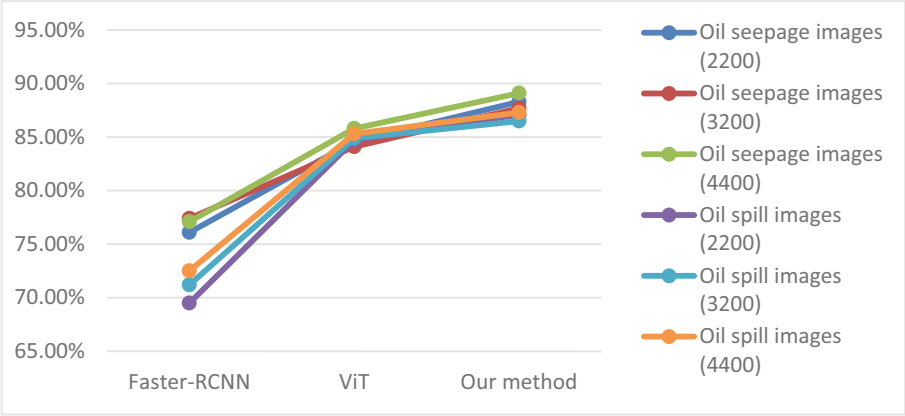

Fig. 9. The accuracy of our method is better than Fast-RCNN and ViT on different datasets

Table 4. Precision comparison of table

Type	Oil seepage images (2200)	Oil seepage images (3200)	Oil seepage images (4400)	Oil spill images (2200)	Oil spill images (3200)	Oil spill images (4400)
Our method	84.7%	86.3%	86.0%	86.6%	87.1%	87.1%
ViT	81.1%	80.4%	80.7%	74.9%	73.8%	74.5%
Faster-RCNN	69.3%	68.5%	70.4%	75.1%	75.4%	75.2%

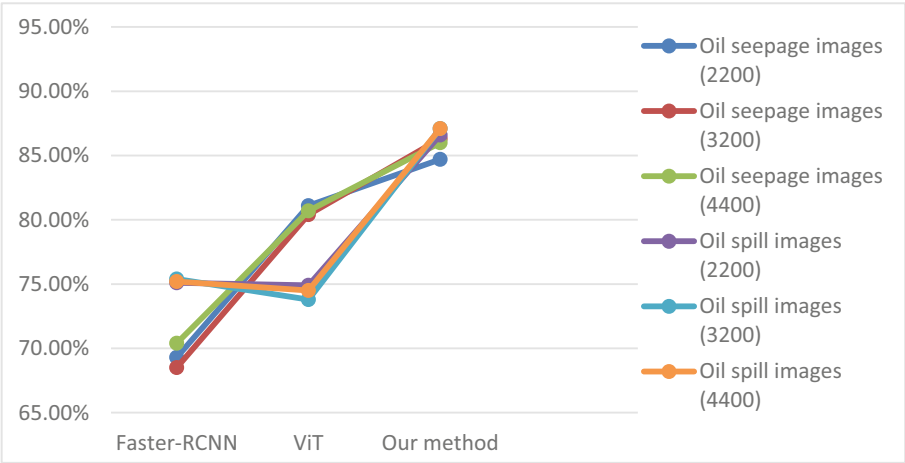

Fig. 10. The Precision of our method is better than Fast-RCNN and transformer on different datasets, while the oil spill detection based on transformer is lower than Fast-RCNN

Table 5. Recall comparison of table

Type	Oil seepage images (2200)	Oil seepage images (3200)	Oil seepage images (4400)	Oil spill images (2200)	Oil spill images (3200)	Oil spill images (4400)
Our method	82.4%	82.7%	83.4%	78.3%	78.7%	79.4%
ViT	76.7%	75.9%	76.3%	74.4%	75.8%	75.6%
Faster-RCNN	65.8%	66.7%	68.1%	69.9%	70.2%	71.9%

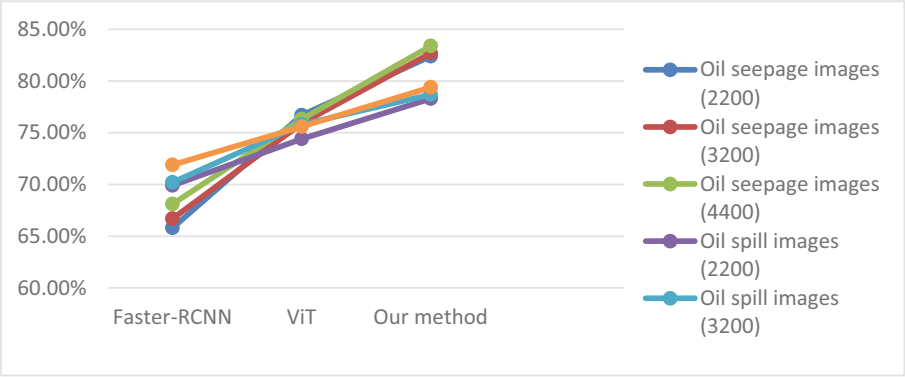

Fig. 11. On different datasets, recall comparison of our method outperforms Fast-RCNN and ViT

4 Conclusions

In this paper, aiming at the problems of difficult identification and detection of oil leakage from oil-filled equipment in daily inspection tasks of substations, the oil leakage detection technology of substation equipment based on fusion of SLIC is proposed. Firstly, the SLIC method is used to segment the image to obtain the super-pixel image data, and the oil leakage part is segmented from the background. Secondly, DBSCAN method based on linear iterative clustering was used to cluster similar superpixels to ensure accurate clustering of features of leaking oil condition images and remove the interference of background environment on leaking oil condition recognition. Finally, vision Transformer deep learning network is used to train and learn the oil leakage images collected in the substation field, and a stable and accurate oil leakage model is obtained. The oil leakage detection technology of substation equipment based on the fusion of SLIC proposed in this paper can effectively realize the accurate identification of oil leakage condition of oil-filled equipment in substation inspection task, and provide strong support for the intelligent application of power business.

Acknowledgements. This work was funded by the "Research on the key technology of intelligent annotation of power image based on image self-learning" program of the Big Data Center, State Grid Corporation of China.

References

1. Jianping, Z., Wenhai, Y., Xianhou, X., Dongfang, H.: Development and application of intelligent inspection robot in Substation. Energy and Environmental Protection **44**(01), 248–255 (2022)
2. Jianhua, W., Lihui, L., Zhe, Z., Yunpeng, L., Shaotong, P.: Oil leakage detection and recognition of substation equipment based on deep learning. Guangdong Electric Power **33**(11), 9–15 (2020)
3. Baoguo, D.: Transformer leakage oil detection based on image processing. Electric Power Construction **34**(11), 121–124 (2013)
4. Yan, W.: A Study on On-line Detection and Prevention of 35 kV Transformer Oil Leakage. Northeaat Petroleum University (2017)
5. Wenli, H., Liangjie, W., Tao, Z., et al.: A Leakage Oil Segmentation Network Based on Edge Information Fusion (2022)
6. Minchen, Y., Yan, Z., Lei, C., Jiajun, H.: Leakage oil detection method based on fluorescence characteristics of transformer oil. Electric World **59**(03), 32–34 (2018)
7. João Sousa, M., Moutinho, A., Almeida, M.: Classification of potential fire outbreaks: A fuzzy modeling approach based on thermal images. Expert Systems with Applications (2019)
8. Martin, E., Kriegel, H.P., et al. Incremental Clustering for Mining in a Data Warehousing Environment. Morgan Kaufmann Publishers Inc, pp. 323–333 (1998)
9. Yang, L., Ningning, Z.: Research on Image Segmentation Method based on SLIC. Comp. Technol. Develop. **29**(01), 75–79 (2019)
10. Parvati, K., Rao, B.S.P., Das, M.M.: Image segmentation using gray-scale morphology and marker-controlled watershed transformation. Discrete Dynamics in Nature and Society (2008)
11. Hou, Y.: Research on Image Segmentation Based on Graph Theory. Xidian University, Xi'an (2011)
12. Meng, T., Relickl, G., Veksler, O., et al.: GrabCut in one cut. IEEE international conference on computer vision. Sydney, NSW, Australia: IEEE, pp.1769–1776 (2013)
13. Meng, T., Ayed, I.B., Marin, D., et al.: Secrets of GrabCut and kernel k-means. In: IEEE international conference on computer vision. Santiago, Chile: IEEE, pp. 1555–1563 (2015)
14. Zhihua, J., Yu, N., Shibin, W., et al.: Improved GrabCut for human brain computerized tomography image segmentation. In: International conference on health information science, pp. 22–30 (2016)
15. Achanta, R., et al.: Slic superpixels. No. EPFL REPORT 149300 (2010)
16. Achanta, R., et al.: SLIC superpixels compared to state-of-the-artsuperpixel methods. Pattern Analysis and Machine Intelligence, IEEETransactions on **34**(11), 2274–2282 (2012)
17. Kanungo, T., Mount, D., Netanyahu, N., et al.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Machi. Intel-lig. **24**(7), 881–892 (2000)
18. Bi, F.M., Wang, W.K., Long, C.: DBSCAN: Density-based spatial clustering of applications with noise. Journal of Nanjing University (Natural Sciences) **48**(4), 491–498 (2012)
19. Bryant, A., Cios, K.: RNN-DBSCAN: A density-based cluste-ring algorithm using reverse nearest neighbor density estimates. IEEE Trans. Knowle. Data Eng. **30**(6), 1109–1121 (2018)
20. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in neural information processing systems, 30 (2017)
21. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR (2021)
22. Gao, X.Y., Hoi Steven, C.H., Zhang, Y.D., et al.: Sparse online learning of image similarity. ACM Trans. Intellig. Sys. Technol. **8**(5), 64:1–64:22 (2017)

23. Zhang, Y., Gao, X.Y., et al.: Learning salient features to prevent model drift for correlation tracking. Neurocomputing **418**, 1–10 (2020)
24. Zhang, Y., Gao, X.Y., Chen, Z.Y., et al.: Mining spatial-temporal similarity for visual tracking. IEEE Trans. Image Processing **29**, 8107–8119 (2020)
25. Gao, X.Y., Xie, J.Y., Chen, Z.Y., et al.: Dilated convolution-based feature refinement network for crowd localization. ACM Transactions on Multimedia Computing, Communications, and Applications (2022)
26. Tang, G.Y., Gao, X.Y., et al.: Unsupervised adversarial domain adaptation with similarity diffusion for person re-identification. Neurocomputing **442**, 337–347 (2021)