IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 4, August, 2023, pp. 1401-1412
Journal homepage: http://iieta.org/journals/ts

Semantic Segmentation Optimization in Power Systems: Enhancing Human-Like Switching Operations

Jin Hua[1*], Yue Zhao[1], Huijun Zhang[2], Haiming Zhao[2], Lei Wang[2]

[1] Electronic Information Engineering, Xi'an Technological University, Xi'an 710021, China
[2] China Coal Shaanxi Yulin Energy Chemical Co., LTD., Yulin 719100, China

Corresponding Author Email: huajin@xatu.edu.cn

https://doi.org/10.18280/ts.400409

Received: 5 March 2023
Revised: 18 June 2023
Accepted: 26 June 2023
Available online: 31 August 2023

Keywords:
semantic segmentation, attention mechanism, multi-scale fusion, visual bag-of-words, SLAM, digital transformation

ABSTRACT

In addressing the digital and intelligent transformation challenges within the traditional desulfurization process of coal secondary utilization, a perception learning model rooted in semantic segmentation networks has been developed. This model, when integrated with real-world operational environments, is shown to confront issues arising from symmetry and multi-scale features inherent to thermal power distribution room contexts. A combination of multi-scale feature fusion and attention mechanisms has facilitated the precise detection of human-like operation knobs on switch operation panels. To cope with extended dynamic scenes, a method relying on the visual bag-of-words has been adopted, wherein local image features are extracted and matched against a visual dictionary, resulting in a refined visual representation. The subsequent selection of consecutive symmetrically similar scene keyframes and the elimination of superfluous data have been observed to augment the efficacy of loop-closure detection. Such enhancements have culminated in improved accuracy in the SLAM (Simultaneous Localization and Mapping) of mobile robots, enabling their autonomous navigation to designated switching operation task locations. Experimental findings underscore the superiority of this optimized model over traditional semantic segmentation networks, with its ability to pinpoint operation knobs on electrical control cabinet panels in distribution rooms. Moreover, before initiating grasping actions under the Eye-in-hand architecture, visual servo grasping maneuvers can be executed, irrespective of the target's appearance angle within the field of view. This optimization offers an insightful foundation for potential integrations into patrol operation mobile robots, marking a feasible and effective stride forward.

1. INTRODUCTION

In coal-fired enterprises, circulating fluidized bed boilers, which utilize circulating fluidized bed technology for combustion and thermal energy conversion, have been identified as a unique type of boiler. Challenges are presented in the realm of coal secondary utilization, particularly within the thermal power supply link which necessitates superior operation and malfunction handling. It has been observed that the control of these boilers is intimately linked to their thermal power supply capacities. Such control is realized through switching operations executed in the electrical distribution room. A pivotal transition towards the automation and enhancement of these operations lies in the robots' ability to semantically segment the instruments on the electrical control cabinet, thereby enriching their environmental perception.

One frequent operation observed within a boiler's electrical distribution room is the shifting of a 10 kV switch from a cold standby state to hot standby. This involves the rotation of the "local/remote" knob on the electrical distribution cabinet, an operation acknowledged as vital to the stability and reliability of the power system. Limitations exist in current robots; they are primarily conditioned to execute operations based on pre-established guidelines and exhibit a deficiency in executing interactive operations responsive to real-time conditions in the distribution room. As illustrated in Figure 1, which presents the schematic diagram of the switchgear's intelligent operation display device, the numeral '1' demarcates the “local/remote” knob.

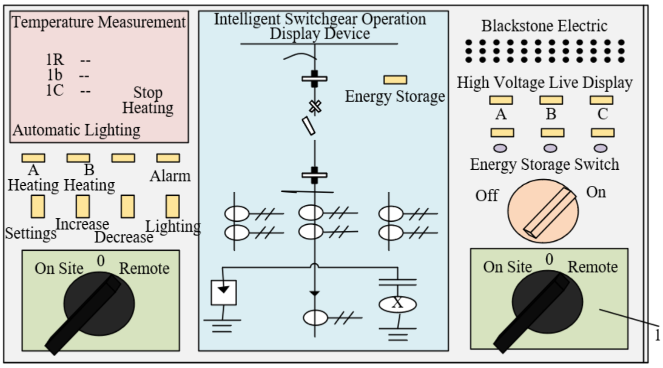

Figure 1. Schematic diagram of switchgear intelligent operation display device

It has been documented that during operations, robots employ the visual bag-of-words technique to capture keyframes, subsequently securing images of the knobs on the electrical control cabinet's panel. Once procured, these image feature datasets are subsequently fed into the semantic segmentation network, which, when integrated with attention mechanisms and multi-scale feature fusion modules, is geared towards the extraction of knob data. By harnessing this deep feature information, accurate segmentation of the knob becomes feasible, paving the way for autonomous robot interaction with the knob.

The underlying pursuit of this research underscores the design and realization of a robot system adept in precise semantic segmentation, thus enhancing its interactive perception of the distribution room environment. Such advancements aim to capacitate the robot to accomplish the switching operation task predicated on real-time conditions. Such endeavors promise to raise the automation standard of circulating fluidized bed boiler control in thermal power plants, thereby bolstering system safety and reliability.

In the evolving industrial milieu, which underscores resource optimization and sustainable progression, coal enterprises have been motivated to augment the secondary utilization rate of coal, consolidate coal systems, and envisage a transition wherein robots supplant human intervention in environments recognized for their hazards, such as distribution rooms. Concurrently, concerted efforts by researchers in the sphere of intelligent robot design have been noted, merging technologies spanning intelligent sensors, machine learning, and artificial intelligence [1, 2]. Emphasis on semantic segmentation emerges as a linchpin in this discourse, revered for its capacity in scene comprehension. As a technology that segments images at the pixel level based on semantic understanding, it is positioned as a vanguard in the vision-based environmental perception arena [3]. Predominance of deep learning-oriented semantic segmentation approaches in this domain has been documented [4-8].

Li et al. [9] elucidated a method wherein a spectral index-driven Fully Convolutional Network (FCN) was devised for the extraction of aquatic regions from multispectral remote sensing images. These multispectral images were first subjected to preprocessing measures, encompassing image enhancement and denoising, with the objective of enhancing the data quality. Based on the distinct characteristics of water bodies, specific spectral indices were selected as input data. An FCN-based deep learning model, designed to utilize these spectral indices, was then constructed. Training of this model was facilitated using the backpropagation algorithm, enabling precise water area extraction. Post-extraction, the results underwent further refinement through noise removal and morphological operations. Nonetheless, a limitation in FCNs was noted, particularly in their ability to segment edge details and preserve intricate information.

Zhu et al. [10] discussed the utilization of the Pyramid Scene Parsing Network (PSPNet) for coronary angiography image segmentation. These images, once input into the PSPNet, underwent prediction procedures, leading to the formulation of a probability distribution of class labels for every pixel. A subsequent thresholding method transformed this probability map into a dichotomous image, yielding the segmented coronary artery region. The post-processing stage involved refining the segmented image through noise elimination and coronary artery geometric feature extraction. It was emphasized that the PSPNet, grounded in the principle of pyramid pooling, adeptly captures semantic data across various image scales. However, the pyramid pooling process was found to amplify computational demands, thereby curtailing processing speed.

In the study of Wang and Liu [11], the Deeplab v3+ neural network was adopted for the discernment and segmentation of gastric cancer pathology slices. Preliminary steps involved preprocessing techniques such as image enhancement, normalization, and cropping, all aimed at augmenting image clarity and minimizing input dimensions. The processed slices were subsequently input into the Deeplab v3+ neural network for both training and inference. During the training regimen, supervised learning was executed using annotated slices, optimizing the loss function for network parameter adjustments. In the inference stage, the trained model was applied to the classification and segmentation of unlabeled slices on a pixel basis, delivering segmentation outcomes for gastric cancer tissues and their normal counterparts. A dominant challenge recognized was the intricate and voluminous nature of medical imagery, which necessitated vast computational resources during both training and inference stages. Additionally, the complex features of gastric cancer pathology slices were occasionally linked to segmentation inaccuracies.

A technique anchored on the Mask R-CNN framework was elucidated, tailored specifically for the quantitative shape assessment of granular materials [12]. Preliminary preprocessing, encompassing denoising and image enhancement, was applied to images of granular materials, aiming to elevate data fidelity and reduce noise artifacts. The processed images were subsequently integrated into the Mask R-CNN framework for both training and inference. During the training regimen, labeled granular images were utilized, facilitating supervised learning and optimization of the network's loss function. In the inference phase, particle localization and segmentation were performed on unlabeled granular images, with additional computations conducted to ascertain particle shape parameters for quantitative assessment. Nonetheless, significant computational costs were incurred by the method, coupled with a pronounced demand for labeled datasets. Furthermore, potential inaccuracies in the recognition and segmentation of granular material particles were identified.

While the previously mentioned semantic segmentation models demonstrated admirable efficacy in certain domains, intrinsic limitations became evident. Specifically, FCNs were noted to grapple with the segmentation of edge details and the retention of granular information. PSPNet's prowess was impeded by its computational latency. The Deeplab series occasionally introduced image artifacts and was associated with relatively heightened computational complexity. Mask R-CNN, although sophisticated, exhibited extensive computational demands and was sometimes found deficient in executing pixel-level segmentation tasks efficiently. Given these constraints, focus was redirected towards U-Net, as described by Chen et al. [13]. Introduced by the University of Olm in 2015, this model's distinctive architecture garnered significant acclaim within the realm of semantic segmentation. U-Net's inherent design advantageously harnesses multi-scale feature information from both foundational and advanced layers, enabling a nuanced feature representation. Such prowess facilitates the precise localization of edge intricacies. Additionally, U-Net's reduced parameter requirements and computational overhead ensure efficiency during training and inference, while its modular design fosters adaptability to task-specific needs. Highlighting these merits, this investigation probes into U-Net's capability for the real-time identification and recognition of knobs on electrical cabinet control panels, bearing implications for the automation and intelligent operations in coal enterprises.

A U-Net based approach for the semantic segmentation of digital mammographic X-rays was presented [14]. The model allowed for pixel-level delineation of these X-rays, automating the detection and localization of breast cancer markers. During its training phase, labeled mammographic X-rays facilitated supervised learning, with model parameter tuning conducted through loss function optimization. The inference stage permitted the segmentation of unlabeled mammographic X-rays, thereby facilitating the subsequent classification of breast cancer regions. Notwithstanding its capabilities, this approach exhibited an innate reliance on a vast corpus of annotated data, exhibited constraints in detecting rare or non-typical breast cancer manifestations, and demonstrated specificity to digital mammographic X-rays, suggesting limited adaptability to other imaging modalities.

Jiang and Li [15] introduced a medical image segmentation approach named TransCU-Net. By integrating the U-Net architecture and the cross-attention Transformer technology, the model was devised. The U-Net structure was initially employed for medical image feature extraction and encoding. The cross-attention Transformer module was subsequently incorporated, fostering long-range dependencies between the encoder and the decoder, thus capturing global context. Such an integration was observed to refine the segmentation accuracy by facilitating the comprehension of the global contextual information within images. Nevertheless, an increase in the computational complexity of the network was noted due to the integration of the Transformer module, which may lead to higher time and resource consumption during both training and inference. It was also observed that this approach is particularly sensitive to data preprocessing and hyperparameter selection, emphasizing the need for meticulous adjustment and optimization.

Fernández and Mehrkanoon [16] proposed a method termed Broad-U-Net, aimed at multi-scale feature learning in on-site prediction tasks. Built upon the U-Net framework, multi-scale feature maps were utilized for feature learning and extraction. Within the encoder, convolution kernels of varied sizes and pooling operations were introduced, allowing the capture of image features at multiple scales. These features were then incrementally upsampled and fused in the decoder, restoring the original spatial resolution of the image. Enhanced prediction accuracy and robustness were observed, suggesting that this multi-scale learning architecture effectively captures multi-scale information in on-site prediction tasks.

A U-Net-based model for crack detection, integrating visualization interpretation techniques, was discussed [17]. The U-Net structure, recognized for its effectiveness in image segmentation tasks, was utilized for crack detection. Its integration with visualization interpretation techniques was found to provide insight into the network's decisions. By generating visual explanations of network predictions, a clearer understanding of the network's crack detection results was achieved. However, certain factors, such as image quality, noise, and interference, were identified as potential disruptors to the reliability of the visual explanations. The method's general applicability was found to be limited, suggesting that it is tailored primarily for crack detection, requiring further investigation for broader image analysis tasks.

Kumar et al. [18] detailed a novel approach to glaucoma detection, where segmentation based on U-Net++ was combined with optimization via ResNet and GRU. The U-Net++ structure, an advanced iteration of the traditional U-Net with augmented depth and additional path connections, was harnessed for the segmentation process. It was adept at pinpointing glaucoma-specific regions in retinal images. However, the complexity of merging U-Net++ with ResNet and GRU was reported to potentially heighten the model's computational complexity, potentially elongating both training and inference times. The model's performance was indicated to be influenced by the inherent quality and noise present in retinal images, which might result in reduced accuracy for intricate or subpar quality images.

A method coupling selective kernel convolution U-Net with a fully connected conditional random field was presented, specifically tailored for semantic segmentation in mechanical assembly [19]. Within the domain of mechanical assembly, semantic segmentation was characterized as the meticulous separation and labeling of various components and parts in images. Through selective kernel convolution, an attention mechanism, the weighted processing of different image regions was facilitated, potentially enhancing segmentation accuracy. In parallel, the introduction of a fully connected conditional random field, through its ability to infer and optimize both local and global constraints, was found to further refine segmentation outcomes. However, a notable rise in the model's computational complexity was identified due to the intricacy of the fully connected conditional random field, indicating potential extensions in training and inference times. Challenges in achieving optimum segmentation accuracy for complex mechanical assembly images were also reported.

For the task of segmenting switch knobs in distribution cabinets within dynamic environments, the necessity of input pre-processing was underscored. Given the visually analogous and symmetrical characteristics of the environment, a keyframe selection method rooted in the visual bag-of-words model was introduced by Zhang et al. [20]. The visual bag-of-words model was found to succinctly represent image content by aligning visual features with a visual dictionary, facilitating the efficient recognition and segmentation of target scenes.

Qi et al. [21] discussed a method wherein microscopy-based sensors were combined with a visual bag-of-words model for the close-range discernment of soil particle sizes. After imaging soil samples using the specified sensors, visual bag-of-words models in computer vision were employed to transform these images into fixed-length feature vectors. However, it was observed that these microscopy-based sensors required specialist personnel for operation and potential challenges stemming from sample preparation and environmental conditions might influence the imaging process.

An enhanced gradient direction histogram method, anchored in the visual bag-of-words model, was introduced by Abouzahir et al. [22] targeting proficient weed detection. Here, the HOG algorithm was initially deployed to extract the local gradient information from images. Following this process, clustering algorithms were applied to the derived gradient features, leading to the generation of a set of visual words. For each image, the frequency of each visual word's occurrence was computed to form a representative feature vector, which was subsequently proposed for use in a weed detection classifier. Yet, this approach was noted to have certain limitations when attempting to detect weeds in extensive agricultural fields, as the density and distribution of weeds could influence accurate detection.

In the study of Asiyabi et al. [23], a method based on the visual bag-of-words model, utilizing polarimetric SAR data for urban land cover classification, was detailed. The polarimetric SAR data was initially segmented, segmenting the image into distinct regions. Within each region, a myriad of features, possibly including polarimetric parameters and texture specifics, were extracted, leading to the creation of the visual bag-of-words model. The frequency of visual word occurrences within each region was then computed to establish a feature vector, which was subsequently utilized by a classifier to determine each region's classification, resulting in the final land cover delineation. In specific contexts, such as electrical distribution rooms, the visual bag-of-words model was recognized as particularly beneficial for switch knob segmentation tasks.

Given this literature backdrop, the essence of this study lies in the profound exploration of semantic segmentation technology, pivotal for scene comprehension. Emphasis is on advancing the real-time detection and recognition capacities for knobs on electrical control cabinet panels. Such efforts are anticipated to greatly support the automation and intelligentization of switch operations in coal-burning enterprises and to proffer solutions aiming to boost resource utilization efficiency, further propelling sustainable growth. The composition of this study is segmented into four primary sections: the introduction, methodology, experiments & results, and conclusion. Within the introduction, insights into the prevailing state of coal reutilization technology and the integration of intelligent robots in coal-burning enterprises are provided. The significance and potential applications of semantic segmentation, a cornerstone technique in environmental perception for industrial production, are emphasized. The methodology delineates the architecture and cardinal technologies of the perception learning model predicated on the semantic segmentation network. This encompasses multi-scale feature integration, attention mechanisms, and key frame selection methodologies based on the visual bag-of-words. The section on experiments & results illuminates the digital transformation of traditional desulfurization processes during coal reutilization, emphasizing the precision and performance of the optimized network model in actual operational environments. The requirements for robot accuracy in switch operations, spanning both local and global spaces, are also underscored.

2. SEMANTIC PERCEPTION TECHNOLOGY

2.1 U-Net Network

The U-Net Network is recognized as a widely applied semantic segmentation model composed of an encoder-decoder. Convolutional layer parameters within the network are trained using provided input images, enabling the acquisition of a well-trained model that can subsequently be employed for predictions. The structure of the U-Net Network is depicted in Figure 2.

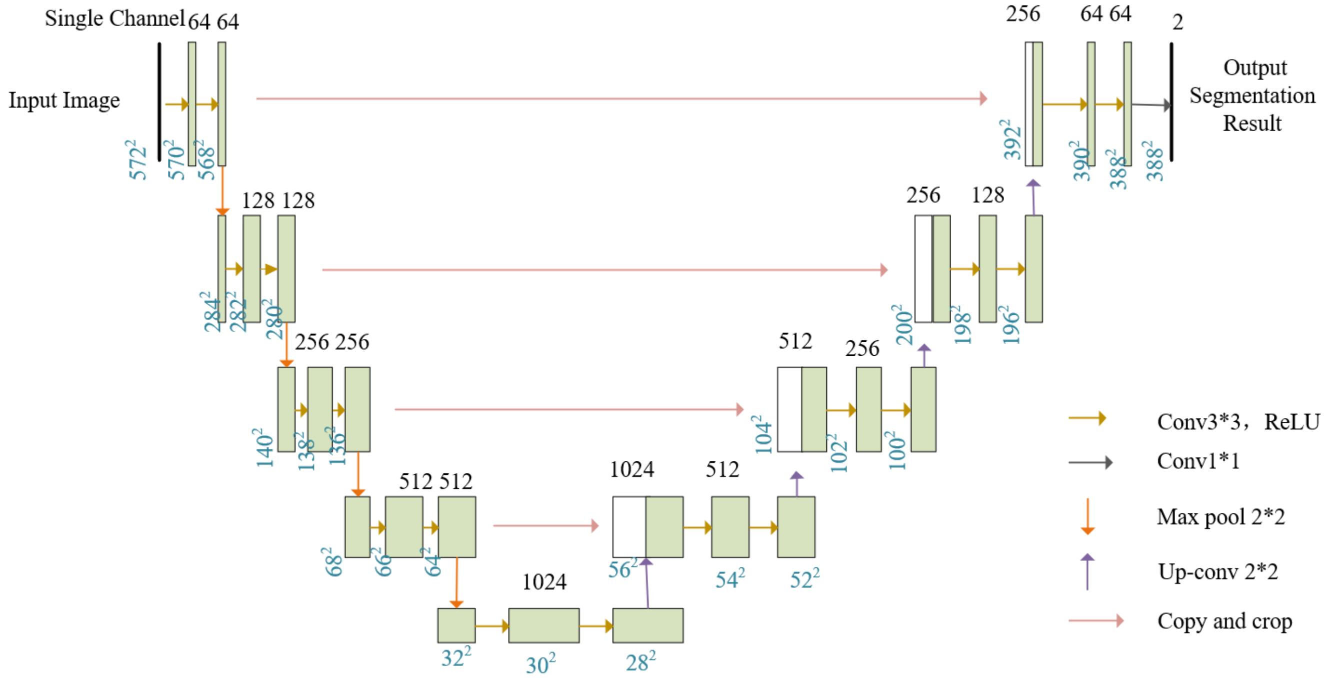

Figure 2. The overall structure of the semantic segmentation network

The architectural design of this model employs a 5-layer U-shaped encoder-decoder structure. It can be divided into two primary sections: a contracting path and an expansive path. The contracting path, consisting of a series of convolutional and pooling layers, is utilized for the extraction of image features. In contrast, the expansive path, comprising a series of convolutional and upsampling layers, serves to transform feature vectors into pixel values for prediction output.

2.2 Semantic segmentation network based on attention mechanism

Within the electrical distribution room environment, not all environmental information is required. For the "local-remote" knobs on the electrical control cabinet panel that demand operation, other control cabinets and instruments of the same cabinet can be deemed secondary and thus disregarded. Drawing inspiration from human visual attention, where attention is focused on the target object, the concept of attention mechanisms was introduced [24]. In this study, a semantic segmentation model based on the attention mechanism was adopted. This mechanism emphasizes monitoring the electrical control cabinet and the information on the instruments of the control panel within the distribution room, while neglecting other equipment and irrelevant information. Such focus serves to amplify target feature information and enhance segmentation accuracy. The structure of the attention mechanism is shown in Figure 3.

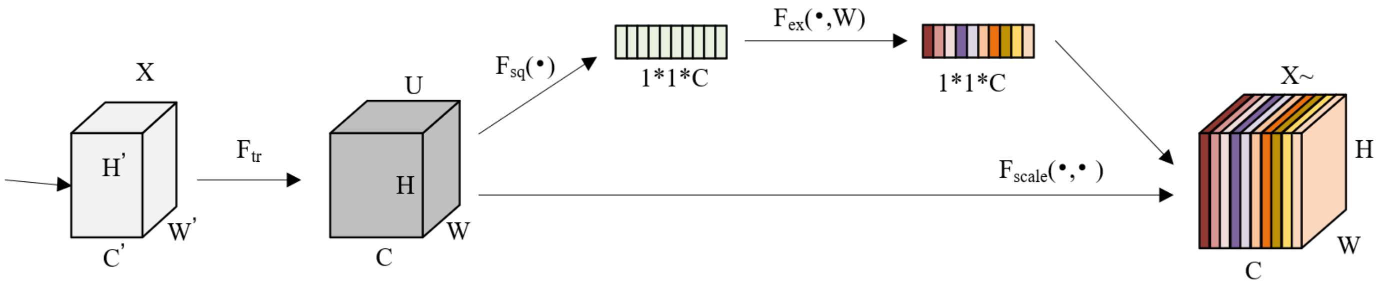

Figure 3. Structure of attentional mechanisms

As illustrated above, the attention mechanism module employs a weight matrix, assigning different weights to distinct positions within the image to capture more vital feature information. This module predominantly amplifies the network's focus on essential features by leveraging channel correlations. It encompasses two operations: Squeeze and Excitation. Initially, for each output channel, global average pooling is performed on the input X, resulting in a scalar. This process yields C scalars, where C represents the number of channels. Subsequently, these scalars undergo a fully connected (FC) layer, ReLU activation function, another FC layer, and then a Sigmoid function, producing C scalars ranging between 0 and 1, indicative of the channel weights. Finally, by multiplying each original output channel with its corresponding weight, weighted feature maps are derived.

(1) Squeeze Operation

For the provided feature map, a global average pooling operation is conducted over the channel dimension, resulting in a feature map of size 1×1×C (where C represents the number of channels). The global features of each channel are compressed from a low dimension, and a dimensionality reduction maps them to a lower dimension, effectively projecting the input feature map onto a smaller vector through a fully connected layer, as demonstrated by the following formula:

$$z_c = \frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{i,\,j,\,c} \qquad (1)$$

where, H and W are the height and width of the feature map respectively, and F represents the input feature map.

(2) Excitation Operation

Two fully connected layers are employed to learn the weights of each channel, subsequently producing a channel attention vector which is then scaled using the sigmoid function. The specific formula is given as:

$$s = f_{ReLU}(FC_2(f_{ReLU}(FC_1(z)))) \qquad (2)$$

where, FC_1 and FC_2 represent the two fully connected layers, and f_{ReLU} stands for the activation function.

Lastly, the channel attention vector is multiplied with the input feature, yielding the output feature adjusted by the attention mechanism. The corresponding formula is:

$$F^{'} = a\,\mathrm{e}\;\, F \qquad (3)$$

where, a symbolizes the channel attention vector, ⊙ denotes the Hadamard product operation, F signifies the input feature map, and F' represents the output feature map.

Through this attention mechanism module, features are weighted based on the importance of each channel, enhancing focus on critical features. This method harnesses channel correlations to aid the network in better capturing significant feature information. With the introduction of the attention mechanism, importance for each channel can be autonomously learned, and the feature map can be weighted in accordance with channel correlations. Consequently, more focus is directed towards pivotal feature channels, while less crucial channels are overlooked. Such a mechanism facilitates a more concentrated allocation of computational resources and attention, enabling more precise segmentation. Upon training and testing the network model, a set of test results, as depicted in Figure 4, was obtained.

In the aforementioned figure, (a), (b), (c), and (d) represent the original image, segmentation results without attention, segmentation results with attention, and original image segmentation results, respectively. It was observed that the attention mechanism aids the network in zeroing in on key areas, capturing essential semantic information more accurately. When segmenting, the attention mechanism allows the network to be more focused on the semantic region of the "local-remote" knob. By assigning varying weights to each channel, the attention mechanism amplifies the network's attention to the features of the "local-remote" knob, while reducing focus on non-essential features. Hence, pivotal targets or areas can be segmented more precisely, elevating the accuracy of segmentation.

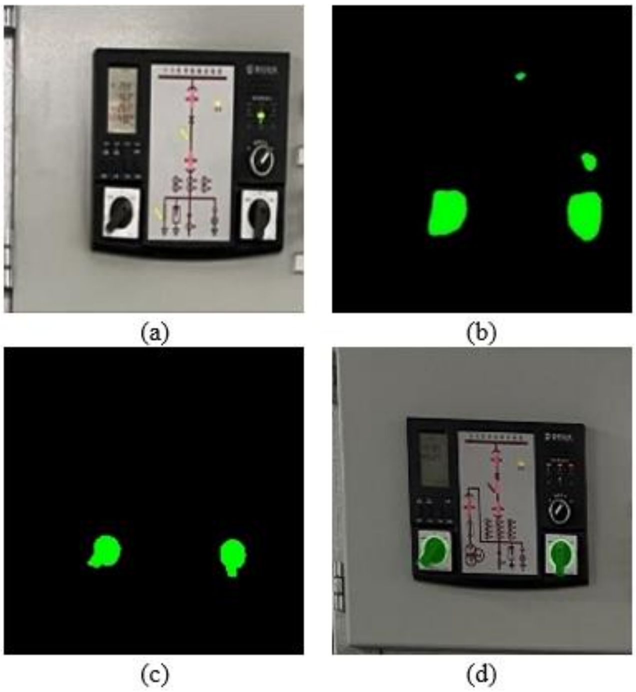

Figure 4. Segmentation diagrams with and without the attention mechanism

2.3 Semantic segmentation network based on multi-scale feature fusion

To further enhance the segmentation accuracy of the "local-remote" knob on the electrical control cabinet using the attention mechanism model and to address the issue of missing contours, a multi-scale feature fusion module was introduced. The structure of this module is depicted in Figure 5.

Multi-scale feature fusion is achieved by aligning local features with global features through the use of edge loss, progressively carrying out image alignment and upsampling operations. Initially, local features 3, obtained from the image using the attention mechanism, are upsampled to produce local features 2. Subsequent upsampling of these features yields result image III. Edge loss is then applied to supervise the training of result image III, correcting the offset of target category pixels. Following this, the corrected local features 2 and global features 2 are subjected to convolution and upsampling operations, generating higher resolution local features 1. These features, upon upsampling, yield result image II, which undergoes supervised training. Finally, convolution and upsampling operations are performed on local features 1 and global features 1, producing the output result image I. This output is then subjected to supervised training using cross-entropy loss.

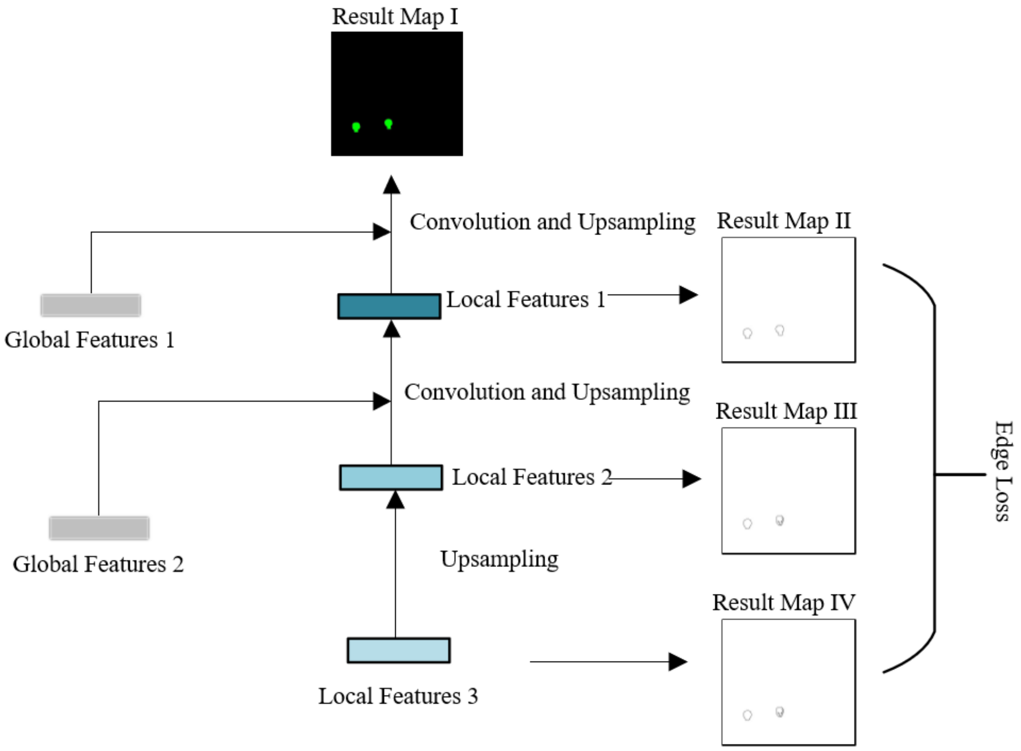

Figure 5. Structure diagram of multi-scale feature fusion

The formula for cross-entropy loss is expressed as:

$$l_{cls} = -\sum_{i=1}^{c} p_i * \log(q_i) \tag{4}$$

where, c is the number of target categories, q_i represents the probability of each pixel being predicted as the ith target category, and $pi \in \{0,1\}$ determines whether each pixel belongs to the ith target category.

The formula for edge loss is:

$$L_{m\arg in} = \sum_H \sum_W \sum_C (\bar{y} - y)^2 \tag{5}$$

In this context, H, W, and C are denoted as the height, width, and number of channels of the image feature respectively, $\bar{y} \in \{0,1\}$ represents whether the pixel point is an edge point of the target category, and y signifies the prediction of a particular pixel point as an edge point.

By introducing the multi-scale feature fusion module, the perceptual capability of the model can be enhanced by leveraging features of varying scales. Specifically, feature maps with different receptive fields can be acquired from different levels of the network and then merged together, providing a more comprehensive and enriched feature representation. Within the multi-scale feature fusion module, commonly utilized methods include upsampling and downsampling, after which feature maps from different levels are combined either by element-wise addition or concatenation. Such operations can enable the model to better capture details and contour information across various scales and resolutions, thereby enhancing segmentation accuracy and mitigating the contour omission issue. Upon testing the network incorporated with the multi-scale feature fusion structure, results as depicted in Figure 6 are obtained. These results display the segmentation effects post multi-scale feature fusion, indicating that this structure can refine segmentation accuracy and more adeptly capture the details and contour information of the target.

In Figure 6, (a), (b), (c), and (d) respectively represent the segmentation result images for distant small-scale original, distant small-scale segmentation, oblique scale result, and oblique scale segmentation. Based on the segmentation chart employing multi-scale feature fusion, when capturing the electrical control cabinet from varying distances and angles, the "local-remote" knob can appear in different scales. After the integration of the multi-scale feature fusion module, it is observed that segmentation outcomes under this methodology can delineate the "local-remote" knob more precisely. By adjusting the model incorporated with the attention mechanism and introducing the multi-scale feature fusion module, segmentation accuracy can be further enhanced, thereby addressing the contour omission issue prevalent in the segmentation of the "local-remote" knob on the electrical control cabinet. Employing this method allows for the harnessing of feature information from different scales and angles, capturing the knob region's details and contours more comprehensively, and boosting segmentation precision and stability.

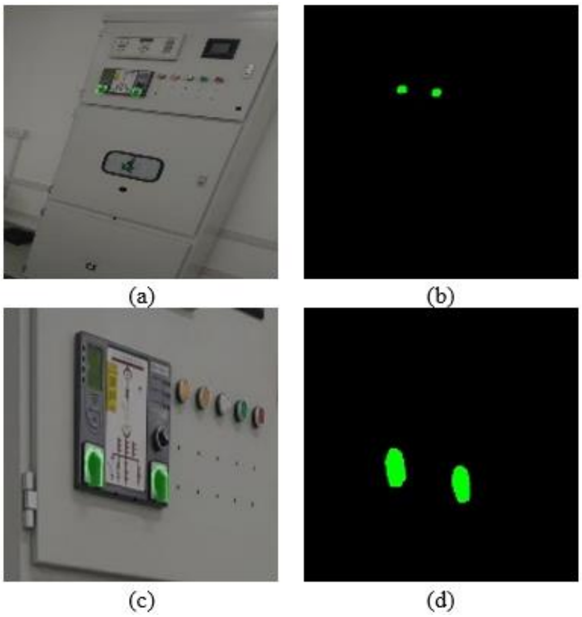

Figure 6. Segmentation comparison chart for multi-scale feature fusion

2.4 Visual bag-of-words

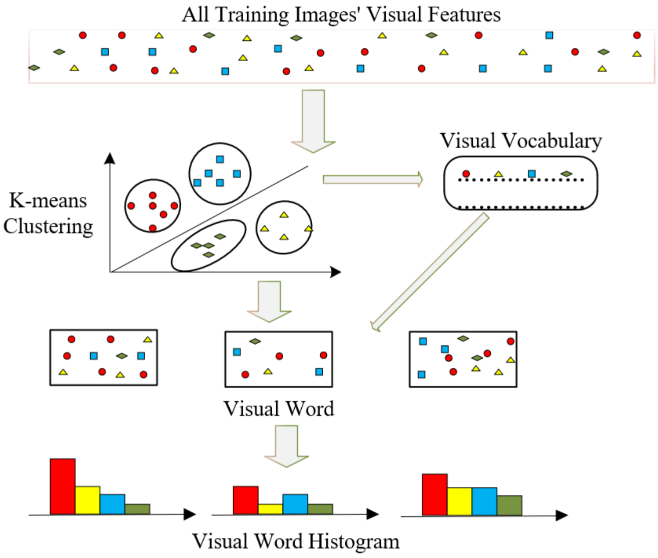

Figure 7. Image representation process diagram based on visual bag-of-words model

In actual distribution room environments, robots are required to operate in dynamic surroundings over extended periods. While testing the model, using video streams, as opposed to merely static images, better simulates real-world scenarios. However, utilizing every frame from a video stream might lead to information redundancy and an increase in computational load. Hence, a keyframe selection method based on the visual bag-of-words model is adopted to select pivotal frames from the distribution room video stream, thereby reducing redundant information and computational overhead. The image representation process grounded in the visual bag-of-words model is illustrated in Figure 7.

The process of image feature extraction and the construction of the visual bag-of-words model includes several steps: extraction of visual features from the image, clustering to obtain visual words, establishment of a vocabulary, and feature extraction from the test image to form a representation vector. A weight is assigned to each visual word, typically represented as term frequency-inverse document frequency (TF-IDF) – denoting the word's distinctiveness within the entire dictionary. Given the definition of term frequency-inverse document probability, the weight η_x of word x is:

$$\eta_x = TF_x * IDF_x = \frac{m}{M}\log\frac{N}{N_x} \tag{6}$$

where, *TF* represents term frequency, indicating the frequency of the word appearing in the total feature library; *IDF* signifies inverse document frequency, where a higher *IDF* denotes fewer images containing that word, implying that the word possesses robust class differentiation capabilities; *M* is the word count of an image frame with word *x* appearing *m* times; *N* is the total feature count in the dictionary; N_x corresponds to the feature count of a word.

For any given image I, representation through words is:

$$I = \{(x_1,\eta_1),(x_2,\eta_2),...,(x_n,\eta_n)\} = f_I \tag{7}$$

where, (x_i,η_i), i=1,2,...,n is the i-th word in image I and its corresponding weight respectively.

Thus, the similarity computation formula between I_1 and I_2 is:

$$f_{I1} - f_{I2} = 2\sum_{i=1}^{n} | f_{I1i} | + | f_{I2i} | - | f_{I1i} - f_{I2i} | \tag{8}$$

The keyframe selection method based on the visual bag-of-words model transforms frames within the video stream into feature vectors, and clustering algorithms are then employed on these vectors. Representative keyframes are subsequently chosen from each cluster as the final selection results, preserving essential information from the video stream while minimizing repetition and redundancy. As evidenced in Figure 8, such methods yield the following test outcomes.

Figure 8. Keyframe extraction test chart based on visual bag of words

From the Figure 8, it can be discerned that keyframes extracted through the visual bag-of-words model vividly showcase instrument information on the electrical control cabinet panel. This provides a more accurate image for subsequent semantic segmentation of the "local-remote" knob, enhancing the precision of segmentation.

3. EXPERIMENTS AND RESULTS ANALYSIS

3.1 Experimental setup and parameter configuration

Experiments were conducted on a 64-bit Windows system computer. An Intel(R) Core(TM) i7-13700KF CPU was incorporated alongside an NVIDIA GeForce RTX 4080 GPU. CUDA version 12.0 was utilized. Necessary dependencies for the experiments were configured through PyCharm and Anaconda, creating the experimental environment. Initialization training of the model was carried out using a dataset from the distribution room. During model training, specific parameters were set: a batch size of 1, an epoch count of 200, and an initial learning rate of 1e-5. The choice and setting of the learning rate are considered critical steps in model training. An appropriate learning rate can be chosen based on actual conditions. A common approach is to train using an initial learning rate, followed by dynamic adjustments or decay depending on model performance and loss during training. Batch size determines the number of samples input in each training step. An appropriate batch size balances training speed with memory usage. Iteration number, which denotes the number of times the model traverses the entire training set, is selected based on the training set size and model convergence rate.

To assess network model performance, indicators employed in the semantic segmentation model included Mean Intersection over Union (MIoU), recall, precision, and F1-score. Performance of the holistic system constructed was evaluated based on the semantic segmentation model's performance.

The formulas for IoU and MIoU are as follows:

$$IoU = \frac{TP}{TP+FP+FN} \tag{9}$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FP+FN} \tag{10}$$

where, k is the number of categories. TP represents true positives, FN denotes false negatives, and FP signifies false positives.

The recall formula is:

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

The precision formula is:

$$precision = \frac{TP}{TP+FP} \tag{12}$$

The F1-score, which consolidates recall and precision, is used for comprehensive performance assessment. The formula is:

$$F1-score = \frac{2*precision*recall}{precision+recall} \tag{13}$$

An *F1-score* value range is from 0 to 1, with values closer to 1 indicating superior model performance.

3.2 Experimental platform

To address the need for robots to accurately perceive their surroundings within distribution room environments, a four-wheel drive circuit breaker operation robot prototype was designed. This prototype consists of a robot chassis, a vision module, an inertial navigation module, a power module, a communication module, and a master control module. The vision module is responsible for capturing visual information from the environment. Power requirements for the robot are catered to by the power module. Data interchange with external systems is facilitated by the communication module, while the master control module oversees the robot's behavior and decision-making.

A perception learning model based on semantic segmentation networks was proposed, aiming at digital and intelligent transformation of traditional desulfurization processes in secondary coal utilization. To validate the effectiveness of this method, a circuit breaker operation robot prototype was independently developed. The appearance of this robot prototype is presented in Figure 9, wherein the proposed method underwent verification and testing. By integrating the perception learning model with actual operational conditions, satisfactory results were achieved. These findings highlight the independent development of the prototype and validate the feasibility of the introduced method.

In the figure, 1 refers to the electric control cabinet, 2 to the "local/remote" knob, 3 to the mechanical gripper, 4 to the robotic arm, 5 to the robot chassis, 6 to the encoder, and 7 to the camera.

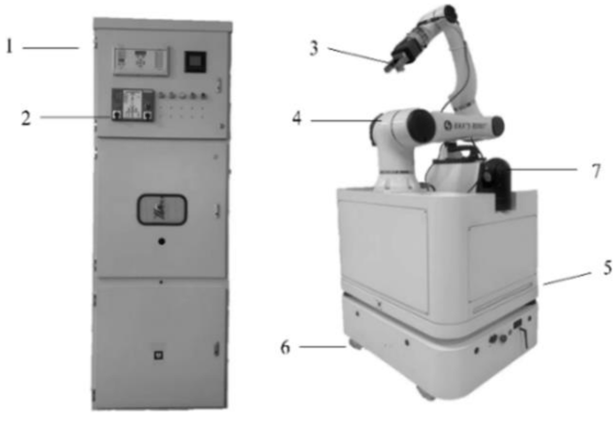

Figure 9. Keyframe extraction test chart based on visual bag of words

When the robot transitions the distribution room's state from 10kV switch cold standby to hot standby, images of the control panel on the distribution room's electric control cabinet are captured using the vision module. Images of the control panel are acquired by the camera and relayed to the master control module for segmentation processing via the communication module. Segmentation of the "local/remote" knob is performed on the image by the master control module to determine its location. Based on this location, the robotic arm is moved to the corresponding area, and the gripper is commanded to open, grasping the "local/remote" knob. Subsequently, the gripper closes and rotates the knob, positioning it to the "remote" setting. Through the application of the vision module, the robot's ability to accurately identify the knob's location on the control panel and carry out the necessary operations for state transition is demonstrated.

3.3 Experimental design

To enable the semantic segmentation model proposed in this study to operate on a computer and utilize a robot to collect environmental information from the distribution room, predicting the position of the "local/remote" knob, it was deemed necessary to train and test the model on a distribution room-related dataset and undergo optimization processes. The overall procedure is depicted in Figure 10.

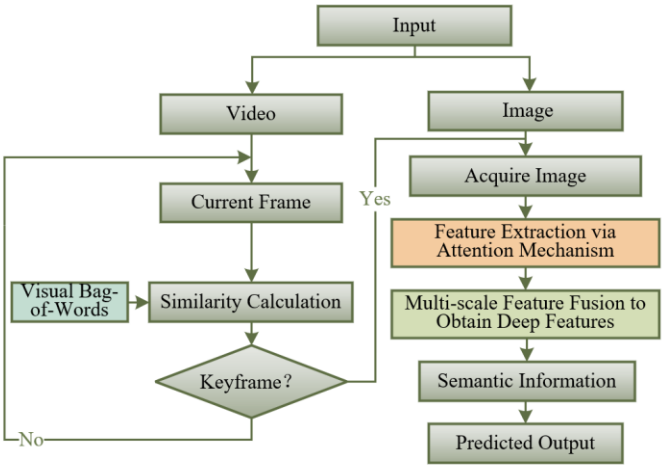

Figure 10. Model flowchart of scene semantics

Initially, the attention mechanism was employed to reinforce key information. Subsequently, a multi-scale feature fusion method was utilized to align global and local features. The fused feature information from the first step was then mapped to the reinforced key information from the second step. Semantic segmentation was conducted to identify target and background areas. By incorporating a keyframe selection method based on the visual bag of words model, multiple consecutive similar scenes were discerned, facilitating the cognitive understanding of the distribution room environment.

3.4 Analysis of experimental results

The model was tested on a computer using the PyCharm programming software, and the curves for loss rate and accuracy for both the training and validation sets were plotted, as illustrated in Figures 11 and 12.

The curve reflecting the loss rate is an indication of the changing trend of the loss function value during the training process, while the accuracy curve portrays the classification accuracy of the model on both training and validation sets. By observing these curves, it was discerned that the loss rate curve gradually decreased and stabilized, indicating that the model progressively learned to extract pertinent features from input images and classify them accurately, thereby minimizing the discrepancy between predictions and actual labels. This observation underscores the capability of the proposed method to effectively assimilate semantic information from the distribution room environment. A gradual increase in the accuracy curve during the training process signifies the model's escalating classification accuracy on both the training and validation sets. Such results affirm the method's efficacy in distinguishing the "local/remote" knob from other background entities, culminating in accurate target identification. It is inferred that the proposed method, by optimizing the loss function, enables the model to glean accurate semantic information from the data and incrementally enhances target object classification accuracy during the training process. This underpins the superiority of the proposed method in the distribution room environmental perception task. As training advanced, the model's classification prowess consistently improved. By introducing attention mechanisms and multi-scale feature fusion techniques, the model capitalized on contextual information within images more effectively, further enhancing classification accuracy.

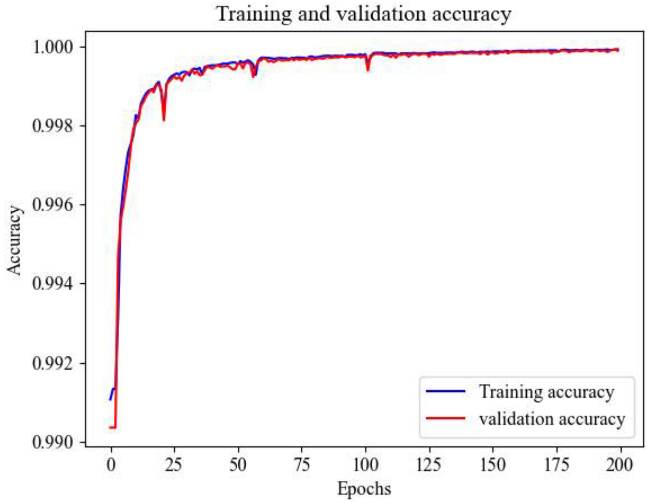

Figure 11. Comparison of the curves of the accuracy rate for the training and validation sets

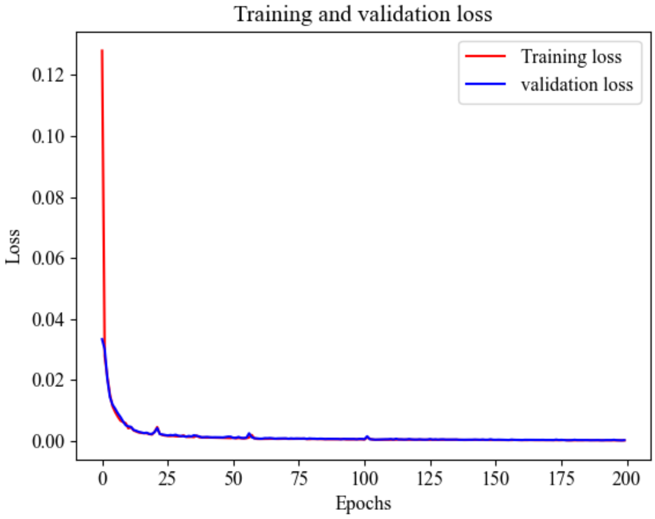

Figure 12. Comparison of the loss rate curves for the training and validation sets

In conclusion, through the loss and accuracy curves depicted in the figures, the superior results achieved by the proposed method in the distribution room environmental perception task were evident. The model, throughout its training, gradually acquired precise semantic information and successfully identified the "local/remote" knob, demonstrating commendable classification accuracy.

In specific evaluation metrics, MIoU measures the overlap between the model's segmentation results and the actual labels. In this experiment, the Mean Intersection over Union (MIoU) for the "local/remote" knob reached 0.92, denoting the model's adeptness at differentiating the knob from the background, effectively capturing the target object. This ascertains the outstanding visual results achieved by the proposed method in the semantic segmentation task within the distribution room environment. Moreover, recall, precision, and F1 scores are critical indicators in evaluating model performance. Their metrics are presented in Table 1.

Table 1. Evaluation indicators

Indicators	Recall	Precision	F1-Score
Back	83%	94%	88%

Based on the data presented in the table, recall, which gauges the capability of the model to correctly identify target objects, was found to be 83%. This implies that the model can proficiently recognize the presence of the "local/remote" knob. Precision, reflecting the model's ability to correctly classify target objects, was recorded at 94%, suggesting that the background is infrequently misidentified as the knob by the model. The F1 score, which comprehensively considers both recall and precision, stood at 88%. This demonstrates that while the model adeptly identifies target objects, it also maintains a degree of control over misclassifications. These findings affirm that the proposed method can accurately perceive target objects in the distribution room environment, offering dependable environmental cognition for precise robotic controls in practical applications. Through the incorporation of attention mechanisms and multi-scale feature fusion techniques, the model was shown to harness the contextual information within images more effectively, enhancing the accuracy of semantic segmentation. Furthermore, by integrating the visual bag-of-words technique, the perception speed was accelerated, facilitating its potential for real-time applications.

In summary, through the evaluation metrics of the experimental results and their practical implications, the superior performance of the proposed method in the distribution room environmental perception task was evident. The "local/remote" knob was accurately segmented, and commendable performance in both target identification and background exclusion was displayed, fortifying the performance of the overarching system.

In this study, environmental perception technology based on semantic segmentation was utilized, complemented with the visual bag-of-words approach for key frame selection in dynamic scenarios. By minimizing redundant information, accurate perception of the distribution room environment was achieved. Figure 13 showcases the test outcomes of this technique.

In the aforementioned figure, a clear observation can be made of the model adeptly segmenting the "local/remote" knob on the electrical control cabinet. This indicates the successful identification of the target object within that region, distinguishing the "local/remote" knob from its background.

The experimental results provided evidence for the efficacy of the semantic segmentation model, incorporating attention mechanisms and multi-scale feature fusion, in robotic perception of the distribution room environment. Additionally, to optimize model performance, the visual bag-of-words technique was employed to boost perception speed. The results asserted that the method could swiftly and precisely segment the "local/remote" knob in the distribution room, further substantiating its effectiveness.

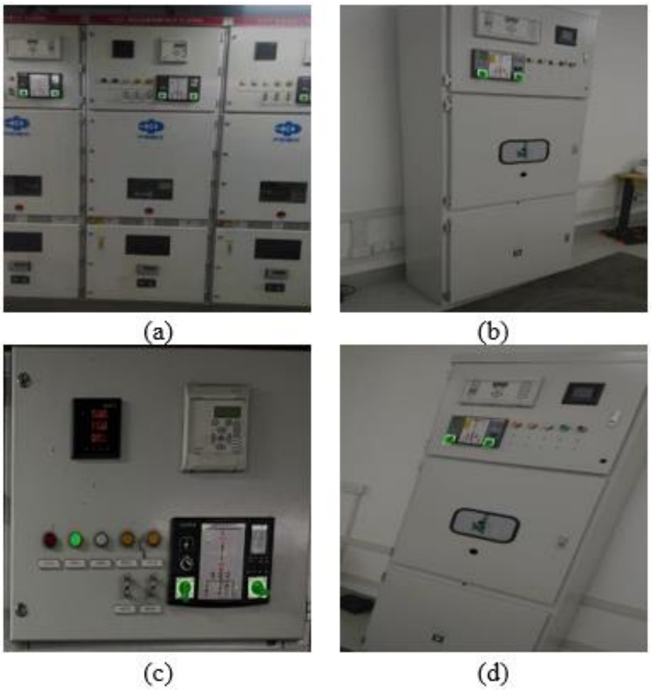

Figure 13. Comparison of test results of the proposed model

4. CONCLUSION

In this study, a novel robotic semantic perception technique grounded in semantic segmentation networks was presented. This technique, when applied to the "local/remote" knob in distribution room environments, underscored its salience for electrical system stability and reliability. Through integration into real-world operational scenarios, digitization and intelligent transformation of traditional desulfurization processes within secondary coal utilization were observed. Challenges posed by symmetry and multi-scale features in the boiler distribution room's thermal supply environment were addressed. The use of multi-scale feature fusion and attention mechanisms facilitated the accurate detection of anthropomorphic operational switch knobs. Furthermore, to ameliorate the localization precision of mobile robots in extended dynamic scenarios, a key frame selection method rooted in the visual bag-of-words was incorporated. Key frames, chosen from successive symmetrical similar scenes, effectively eliminated redundant data, optimizing loop closure detection efficiency.

Experimental outcomes revealed that operational knobs on the control panel of the distribution room's electric control cabinet were accurately identified by the optimized network model. Additionally, it was observed that robots could adeptly undertake precision-necessitated switch operation tasks across both local and global autonomous navigation spheres.

The presented technique affirmed the viability and efficiency of semantic perception technology within distribution room environments for robotic applications. The resultant enhancements in robotic perception and understanding tasks advocate for a more reliable electrical system and a reduction in potential fault risks. The potential for this technology to aid distribution room inspection and maintenance procedures signals its broad applicability.

However, it must be acknowledged that in these experiments, semantic segmentation was solely conducted for the "local/remote" knob, prompting inquiries regarding the scalability of the method to diverse devices and contexts. Further research and empirical validation are imperative to address these queries. Prospective research endeavors could prioritize refining the semantic perception model, elevating segmentation accuracy and processing speed, and expanding its utility to encompass a broader range of distribution room equipment. Delving deeper into advanced deep learning methodologies and pioneering intelligent algorithms might further elevate robotic perception and operational capacities within electrical systems, thus heralding a transformative era for the electrical sector.

ACKNOWLEDGEMENTS

This paper was supported by the Natural Science Special Research Program of Shaanxi Provincial Department of Education (Grant No.: 21JK0677) and the Science and Technology Innovation Project Program of China Coal Energy Group Corporation, a central enterprise.

REFERENCES

[1] Tripathi, A.K., Aruna, M., Prasad, S., Pavan, J., Kant, R., Choubey, C.K. (2023). New approach for monitoring the underground coal mines atmosphere using IoT technology. Instrumentation Mesure Métrologie, 22(1): 29-34. https://doi.org/10.18280/i2m.220104

[2] Felleman, D.J., Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex (New York, NY: 1991), 1(1): 1-47. https://doi.org/10.1093/cercor/1.1.1-a

[3] Harris, Z.S. (1954). Distributional structure. Word, 10(2-3): 146-162. https://doi.org/10.1080/00437956.1954.11659520

[4] Sivic, Z. (2003). Video Google: A text retrieval approach to object matching in videos. In Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, pp. 1470-1477. https://doi.org/10.1109/ICCV.2003.1238663

[5] Warda, L., Kaladzavi, G., Samdalle, A., Kolyang (2022). Integration of Ontology Transformation into Hidden Markov Model. Information Dynamics and Applications, 1(1): 2-13. https://doi.org/10.56578/ida010102

[6] Song, X.N., Liu, H.C., Wang, L.J., Wang, S., Cao, Y.Y., Xu, D.L., Zhang, S.F. (2022). A semantic segmentation method for road environment images based on hybrid convolutional auto-encoder. Traitement du Signal, 39(4): 1235-1245. https://doi.org/10.18280/ts.390416

[7] Padamata, R.B., Atluri, S.K. (2023). Tomato crop disease classification using semantic segmentation algorithm in deep learning. Revue d'Intelligence Artificielle, 37(2): 415-423. https://doi.org/10.18280/ria.370218

[8] Amaria, S., Guidedi, K., Lazarre, W., Kolyang (2022). A survey on multimedia ontologies for a semantic annotation of cinematographic resources for the web of data. Acadlore Transactions on AI and Machine Learning, 1(1): 2-10. https://doi.org/10.56578/ataiml010102

[9] Li, Z., Zhang, X., Xiao, P. (2022). Spectral index-driven FCN model training for water extraction from multispectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 192: 344-360. https://doi.org/10.1016/j.isprsjprs.2022.08.019

[10] Zhu, X., Cheng, Z., Wang, S., Chen, X., Lu, G. (2021). Coronary angiography image segmentation based on PSPNet. Computer Methods and Programs in Biomedicine, 200: 105897. https://doi.org/10.1016/j.cmpb.2020.105897

[11] Wang, J., Liu, X. (2021). Medical image recognition and segmentation of pathological slices of gastric cancer based on Deeplab v3+ neural network. Computer Methods and Programs in Biomedicine, 207: 106210. https://doi.org/10.1016/j.cmpb.2021.106210

[12] Yang, D., Wang, X., Zhang, H., Yin, Z.Y., Su, D., Xu, J. (2021). A Mask R-CNN based particle identification for quantitative shape evaluation of granular materials. Powder Technology, 392: 296-305. https://doi.org/10.1016/j.powtec.2021.07.005

[13] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[14] Soulami, K.B., Kaabouch, N., Saidi, M.N., Tamtaoui, A. (2021). Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation. Biomedical Signal Processing and Control, 66: 102481. https://doi.org/10.1016/j.bspc.2021.102481

[15] Jiang, S., Li, J. (2022). TransCUNet: UNet cross fused transformer for medical image segmentation. Computers in Biology and Medicine, 150: 106207. https://doi.org/10.1016/j.compbiomed.2022.106207

[16] Fernández, J.G., Mehrkanoon, S. (2021). Broad-UNet: Multi-scale feature learning for nowcasting tasks. Neural Networks, 144: 419-427. https://doi.org/10.1016/j.neunet.2021.08.036

[17] Liu, F., Wang, L. (2022). UNet-based model for crack detection integrating visual explanations. Construction and Building Materials, 322: 126265. https://doi.org/10.1016/j.conbuildmat.2021.126265

[18] Kumar, V.V.N.S., Reddy, G.H., GiriPrasad, M.N. (2023). A novel glaucoma detection model using Unet++-based segmentation and ResNet with GRU-based optimized deep learning. Biomedical Signal Processing and Control, 86: 105069. https://doi.org/10.1016/j.bspc.2023.105069

[19] Chen, C., Zhang, C., Wang, J., Li, D., Li, Y., Hong, J. (2023). Semantic segmentation of mechanical assembly using selective kernel convolution UNet with fully connected conditional random field. Measurement, 209: 112499. https://doi.org/10.1016/j.measurement.2023.112499

[20] Zhang, L., Shen, J., Zhu, B. (2021). A research on an improved Unet-based concrete crack detection algorithm. Structural Health Monitoring, 20(4): 1864-1879. https://doi.org/10.1177/1475921720940068

[21] Qi, L., Adamchuk, V., Huang, H.H., Leclerc, M., Jiang, Y., Biswas, A. (2019). Proximal sensing of soil particle sizes using a microscope-based sensor and bag of visual words model. Geoderma, 351: 144-152. https://doi.org/10.1016/j.geoderma.2019.05.020

[22] Abouzahir, S., Sadik, M., Sabir, E. (2021). Bag-of-visual-words-augmented histogram of oriented gradients for efficient weed detection. Biosystems Engineering, 202: 179-194.

https://doi.org/10.1016/j.biosystemseng.2020.11.005

[23] Asiyabi, R.M., Sahebi, M.R., Ghorbanian, A. (2022). Segment-based bag of visual words model for urban land cover mapping using polarimetric SAR data. Advances in Space Research, 70(12): 3784-3797. https://doi.org/10.1016/j.asr.2021.10.042

[24] Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint, arXiv:1409.0473.