ISMU-Net: insulator image Segmentation with Mamba-based U-Net

Xunxing Liu*
College Information and Finance, Xuancheng Vocational & Technical College, Xuancheng 242000, Anhui, China

ABSTRACT

Accurate segmentation of insulators in power systems is crucial for localizing and detecting insulator defects, ensuring the safe and efficient transmission of electricity. However, current insulator segmentation models suffer from high misclassification rates and low segmentation accuracy when segmenting aerial insulator images. To address this issue and achieve precise segmentation of insulator images, we propose an Insulator image Segmentation with Mamba-based U-Net (ISMU-Net). Firstly, an enhanced feature extraction block, grounded in visual state modeling, is devised to replace the convolutional blocks of U-Net, enabling comprehensive extraction of insulator image information. Secondly, to mitigate information loss at skip connections and harness the underlying network information, we integrate the attention mechanism from SE-Net into the feature extraction block, optimizing feature fusion at skip connections. Experimental results on a collected dataset of aerial insulator images reveal that ISMU-Net achieves a Precision (Pre) of 92.7%, Recall (Rec) of 91.7%, and F-Measure (F1) of 94.8%. Moreover, ISMU-Net demonstrates strong generalization capabilities across diverse backgrounds, thereby validating its effectiveness in enhancing the accuracy and robustness of insulator segmentation.

Keywords: Attention mechanism, visual state space, local features, global features, insulator images

1. INTRODUCTION

Insulators are an indispensable component of power systems, facilitating the transmission of high-voltage electricity to various terminal devices while ensuring the safe and stable operation of the entire system. However, due to their prolonged exposure to high-voltage electric forces, insulators are prone to various defects such as explosions, cracks, and fouling. The stable operation of power systems is threatened by significant safety hazards, arising from system failures that can be caused by these defects. Therefore, insulator detection has always been a crucial research direction in the field of power systems.

The traditional approach to overhead power line inspection relies heavily on manual inspections. Given the vast number of high-voltage overhead lines located in remote, geographically complex areas, manual inspections are associated with high operational intensity and long inspection cycles. Additionally, the efficiency of manual inspections is significantly limited, and misdetections and missed detections are frequent due to factors such as the inspector's experience and perspective. In view of these challenges, a revolution in inspection methods is imperative to make power line inspections more intelligent and convenient. With the rapid technological advancements in recent years, the maturing of drone aerial photography technology has presented itself as an excellent alternative to manual inspections.

By capturing insulator images through aerial photography and further processing these images using computer-assisted algorithms, the status of the insulators can be accurately determined. Currently, transmission line insulator detection methods can be categorized into two types: conventional methodologies reliant on machine learning and advanced approaches utilizing deep learning[1-3]. Defect recognition can only be achieved under specific conditions when utilizing traditional machine learning methods, limiting their applicability in complex distribution lines, deep neural network-based methods ensure good adaptability but require significant amounts of training samples, which is contradictory to the scarcity of aerial images of distribution lines[4-6].

2. RELATED WORK

Insulators, a crucial component in transmission systems, are prone to mechanical failures due to various environmental factors such as storms, earthquakes, and heavy rains, potentially leading to malfunctions in photovoltaic power stations.

*liuxunxing2024@163.com; phone 1 380 562 1649; fax 0 563 3023305

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133952K · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049373

Therefore, detecting mechanical faults in insulators has become an important aspect in advancing intelligent power inspection.

2.1 Existing research

In recent years, numerous studies have been conducted on the mechanical fault detection of insulators. Yu et al.[7] addressed the limited availability of insulator samples by proposing a method that combines texture feature enhancement with the SINet network, achieving an accuracy of 99.82% in identifying insulator mechanical faults. Wei[8] proposed utilizing Hough ellipse detection coupled with Canny edge detection to identify mechanical faults in composite insulators. Bakshi et al.[9] introduced an improved U-Net convolutional network with dilated convolutions and full-scale skip connections to enhance the accuracy and efficiency of insulator fault detection. Hao et al.[10] presented a fault diagnosis method for insulators in aerial images, employing Otsu's thresholding for segmentation and Hough transform for ellipse detection to detect missing strings based on insulator position information. Huang et al.[11] tackled the issue of uneven illumination in natural environments by proposing an improved color difference-based image segmentation method that involves partitioning bright and dark regions, compensating for illumination variations, and utilizing an adaptive thresholding segmentation algorithm combined with geometric shape analysis.

2.2 Challenges and proposed approach

Despite the progress made using image processing techniques, existing insulator mechanical fault detection algorithms still face challenges in effectively identifying faults. While deep learning models demonstrate higher detection accuracy, they are limited by the small size of training datasets and poor model generalization. Traditional image processing algorithms, on the other hand, suffer from instability and sensitivity to uneven illumination. To address these issues, this paper proposes an improved insulator segmentation algorithm. Initially, brightness correction is applied to insulator images to normalize luminance under different lighting conditions, thereby enhancing the generalization of model training. Subsequently, an enhanced U-shaped architecture network with a visual state module and an improved attention mechanism is utilized for insulator segmentation. This approach facilitates the fusion of local and global features, further improving the accuracy of insulator segmentation and aiding in the localization and detection of defects.

3. METHODS

This section, the overall architecture of the proposed ISMU-Net is initially introduced. Subsequently, the details of the key component, the VSS block, are elaborated upon. Lastly, the loss functions utilized during the training process are explained.

3.1 Architecture overview

To better extract image features, the ISMU-Net model employs a symmetrical structure. In Figure 1, the ISMU-Net is structured with a block embedding layer, an encoder, a decoder, a final projection layer, and incorporates skip connections.

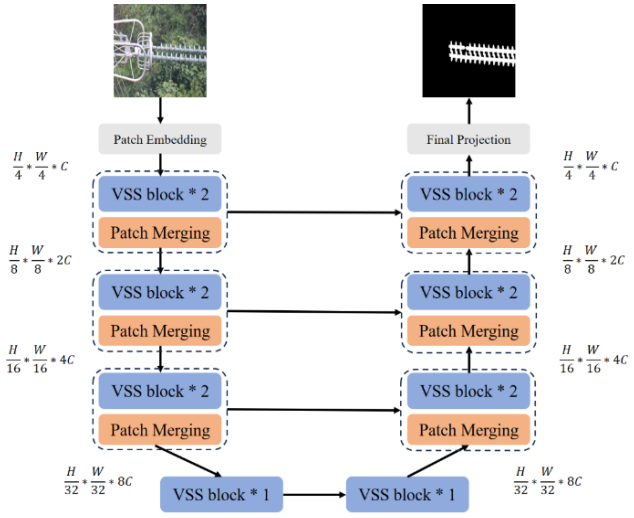

Figure 1. The overall architecture of ISMU-Net.

The input image $x \in R^{H\times W\times 3}$ is segmented into non-overlapping patches of size 4×4 by Patch Embedding, which then projects the image's dimensionality to C, with C typically set to 96. This results in an embedded representation $x' \in R^{\frac{H}{4}\times\frac{W}{4}\times 3}$. Prior to being fed into the encoder for feature extraction, x' undergoes Layer Normalization[12] for normalization. The encoder consists of four stages, and at the end of the first three stages, patch merging operations are applied to diminish the height and width of the input features while enhancing the number of channels. Each of the four stages employs [2, 2, 2, 2] VSS blocks, where the number of channels progressively increases from [C, 2C, 4C, 8C].

The decoder is likewise structured with four stages, wherein the initial three stages commence with patch expansion operations to decrease the feature channels and augment the height and width. Across the four stages, [2, 2, 2, 1] VSS blocks are utilized, with the channel count progressively decreasing from [8C, 4C, 2C, C]. Subsequent to the decoder, a final projection layer is employed to match the feature size with the segmentation target. This entails four up-sampling operations through patch expansion to reestablish the height and width of the features, followed by a projection layer to restore the original channel count.

The ISMU-Net model mitigates the bottleneck issue by incorporating two VSS blocks. Within both the encoder and decoder, skip connections are implemented in each layer to blend multi-scale features with up-sampled outcomes, thereby enriching spatial details through the integration of both shallow and deep features. A subsequent linear layer ensures dimensional consistency of the combined feature set with the up-sampled resolution. For skip connections, a straightforward addition operation is employed, eliminating the need for additional parameters.

3.2 Visual state space block (VSS)

As depicted in Figure 2, the VSS module, adapted from V-Mamaba[13], constitutes a pivotal element of ISMU-Net. Initially, the input data undergoes layer normalization and is bifurcated into two paths. In the first path, the input traverses a linear layer and is subjected to an activation function. In the second path, the input sequentially traverses a linear layer, undergoes depth-wise separable convolution, and is activated, eventually reaching the 2D-Selective-Scan (SS2D) module for feature extraction. The extracted features are subsequently normalized via layer normalization and undergo an element-wise multiplication with the output of the first path to facilitate information fusion. The fused features are then blended through a linear layer and merged with a residual connection, culminating in the output of the VSS module. This work adopts SiLU[14] as the default activation function to enhance the model's performance.

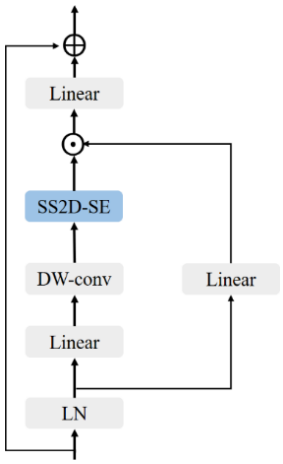

Figure 2. The detailed structure of the visual state space (VSS) block.

The SS2D module is a meticulously crafted structure consisting of three key components: a scan expansion operation, S6 blocks, and a scan merging operation. The scan expansion operation ingeniously decomposes the input image into sequences along four diverse directions, thereby comprehensively exposing the image's information from all perspectives. Subsequently, these sequences undergo feature extraction via the S6 blocks. Derived from Mamba[15], the S6 module incorporates the strengths of the S4[16] module and introduces an input-conditioned SSM parameter selection mechanism. This mechanism enables the S6 blocks to intelligently distinguish and preserve task-relevant information while filtering out irrelevant details, ensuring thorough scanning of each directional information and precise capture of various features in the image. Following the scan expansion, the scan merging operation effectively sums and integrates the generated

sequences, proficiently reconstructing the output image to match the dimensions of the input, thereby accomplishing both information integration and restoration. Through its unique components and efficient working mechanism, the SS2D module provides robust support for image processing tasks, demonstrating exceptional performance in feature extraction and processing.

3.3 2D-selective-scan and squeeze excitation block (SS2D-SE)

The integration of the selective scanning mechanism with the attention mechanism can effectively enhance the saliency of image features, addressing issues such as blurred lesion segmentation boundaries and difficult feature extraction in insulator images. Drawing on the pixel characteristics of aerial images and the distribution properties of insulator regions, this paper proposes an SS2D-SE module that guides the network to acquire feature regions spanning from shallow to deep constraints and from local to global scopes.

To enhance feature saliency and extract clearer insulator boundary information, this paper introduces the SE module based on the SS2D module, as depicted in Figure 3. The 2D-Selective-Scan (SS2D) comprises three modules: S6 block, Scan Merging and Scan Expanding. The Scan Expanding module meticulously decomposes the image into sequences along both rows and columns, subsequently scanning in four diverse directions: from top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. This methodology ensures that every pixel incorporates information from all other pixels in various orientations. Subsequent to the S6 module's processing, the Scan Merging module reconfigures each sequence into a unified image and combines all sequences into a novel sequence. The weights extracted by the SE module are weighted with the features processed by the SS2D module, as illustrated in Figure 3.

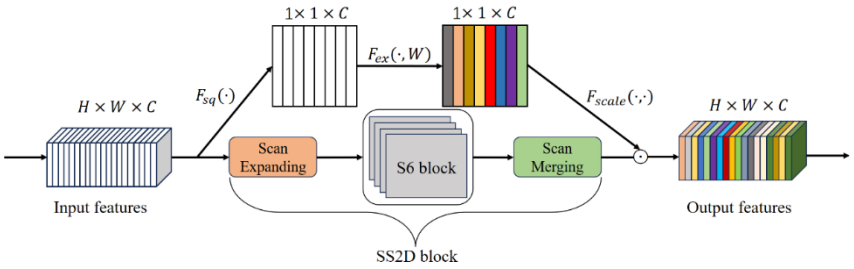

Figure 3. Illustration of the 2D-selective-scan and squeeze excitation block on image.

The amalgamation of S6 and CSM, designated as the S6 block, functions as the pivotal component in the construction of the Visual State Space (VSS) block, constituting the foundational unit of V-Mamba. The S6 block embodies the linear complexity inherent in the selective scanning mechanism while preserving a comprehensive global receptive field. The superimposition of the SE module on top of this aids in recalibrating channel feature responses and learning global information by suppressing ineffective features and emphasizing effective ones. Initially, a feature map with channel number, height, and width represented by C, H and W, is taken as input, as shown in equation (1):

$$u_c = v_c * X = \sum_{n=1}^{C'} v_c^n * x^n \tag{1}$$

Herein, u_c denotes the output of the (c)-th convolutional kernel, v_c represents the (c)-th convolutional kernel itself, and x^n signifies the (n)-th input covered by the current convolutional kernel.

In this context, the notation u_c signifies the output generated by the (c)-th convolutional kernel, v_c represents the (c)-th convolutional kernel itself, and x^n denotes the (n)-th input that is encompassed by the current convolutional kernel.

Additionally, for the purpose of condensing global spatial information into channel descriptors, we employ a global average pooling operation F_{sq} that transforms the feature map into a $1 \times 1 \times C$ dimensional feature map $Z = [z_1, z_2, \cdots, z_c]$. The statistical information z_c for the (c)-th channel is calculated as shown in equation (2):

$$z_c = F_{sq}(u_c) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{i=1}^{W} u_c(i, j) \tag{2}$$

As shown in equation (3), the excitation function is realized through the employment of a concatenation of fully connected (FC) layers, a Rectified Linear Unit (ReLU) activation, another FC layer, and a Sigmoid activation function.

$$R = F_{ex}(Z, W) = \sigma(W_2 \cdot \delta(W_1 \cdot Z)) \tag{3}$$

In this context, $R = [r_1, r_2, \cdots, r_c]$ represents the learned activation values for each channel, where r_c denotes the weight value for the (c)-th channel. The notation F_{ex} denotes the excitation operation, whereas δ stands for the Rectified Linear Unit (ReLU) activation function, and σ signifies the Sigmoid activation function. W stands for the weight parameters learned during the training process.

Finally, the final output is obtained by scaling the channel weights, as follows:

$$\tilde{x}_c = F_{scale}(u_c, r) = r_c \cdot u_c \tag{4}$$

where F_{scale} represents the weighted scaling operation, and $\tilde{x}_c$ denotes the output value for the (c)-th channel.

4. EXPERIMENTAL SETUP

4.1 Dataset and experimental environment

In this study, the experimental image dataset comprises solely of aerial images captured by drones, specifically targeting power system insulator images, all with a resolution of 512∗512 pixels, as illustrated in Figure 4. This dataset encompasses various grid insulator images captured under diverse environmental conditions, including those featuring only the insulator, the insulator with connecting components, the insulator partially occluded by connecting components, and aerial images of insulators against varying backgrounds.

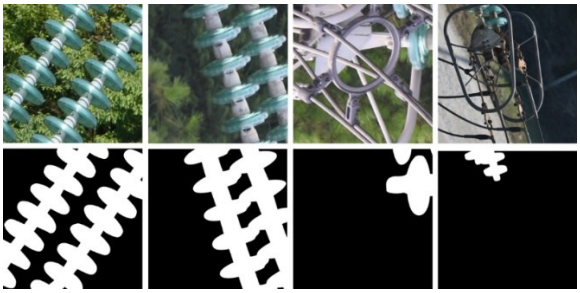

Figure 4. Aerial images of insulators.

The dataset comprises a total of 5,320 insulator images, with 4,256 images in the training set (including 850 images for the validation set) and 1,064 images in the test set, as detailed in Table 1. The hardware environment for the experiments includes an Intel I7 12700KF processor, an Nvidia RTX 3080 TI GPU, and 16 GB of RAM. The software environment is based on the PyTorch 1.7 deep learning framework running on Ubuntu 20.04, with Python 3.8 as the programming language.During training, Adam was chosen as the network optimizer, with a batch size of 8, an image size of 512×512, an initial learning rate of 0.01, and a total of 120 training iterations.

Table 1. Data setting for training, validation and test sets.

Dataset	Training	Validation	Test	Sum
Insulator images	4256	850	1064	5320

4.2 Evaluation metrics

To evaluate the performance of the model with precision, we selected Precision (Pre), Recall (Rec), and F-Measure (F1)[17,18] as the metrics. Specifically, Pre represents the precision index, while Rec represents the recall index, both of which measure the accuracy of the model's segmentation. Their formulas are expressed as follows:

$$Pre = \frac{TP}{TP+FP} \tag{5}$$

$$Rec = \frac{TP}{TP+FN} \tag{6}$$

The quantities TP, FP, and FN, respectively, signify the count of false positives, true negatives, and false positives. To achieve better performance, both Pre and Rec should be as high as possible. However, due to the small proportion of insulators in the images and the presence of interference, there is often a trade-off between Pre and Rec. Therefore, we introduce the comprehensive evaluation metric F1 to better assess the model's performance:

$$F_1 = \frac{(1+\beta^2)\times Pre\times Rec}{\beta^2\times Pre+Rec} \tag{7}$$

Here, β^2 is a hyperparameter that balances the influence of *Pre* and *Rec* on $F1$. When $\beta^2 > 1$, *Rec* has a greater impact on $F1$, and when $\beta^2 < 1$, *Pre* has a greater influence. Since insulators occupy a small proportion in the images and there are many interferences, to achieve more accurate insulator segmentation, the evaluation metric $F1$ should be more focused on *Pre*. Therefore, we set $\beta^2 = 0.4$.

5. RESULTS AND DISCUSSION

A comparative experiment was undertaken to ascertain the efficacy of the proposed ISMU-Net model as an enhancement to U-Net, by contrasting it against a range of advanced models. Specifically, U-Net, R2U-Net, and TransU-Net were selected and evaluated on the same aerial insulator dataset, using identical experimental conditions and tuning strategies. The visualization of the experimental results is shown in Figure 5.

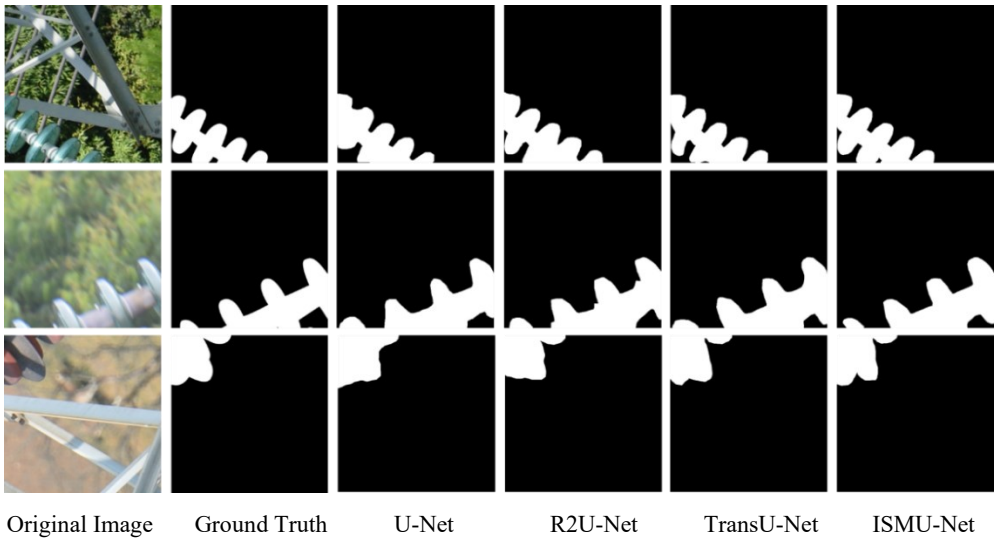

Figure 5. Segmentation results of different models.

While visual comparisons can be intuitive, they are susceptible to subjective biases. As evident from Table 2, the proposed model surpasses the other models in Accuracy, IoU, and Dice Score, thereby validating its rationality and effectiveness in capturing intricate details and displaying robust performance. Due to the complexity of background images and the lack of extensive preprocessing and post-processing, the U-Net model, which introduces deconvolution and feature layer connections, mitigates the loss of detail information to some extent, improving the results. The R2U-Net model, leveraging recursive neural networks, achieves superior feature representation and improved metrics. Meanwhile, the TransU-Net model incorporates a transformer structure in the encoder, enabling better feature extraction. However, as the encoding-decoding process deepens, it may lose some global features. Nonetheless, its segmentation performance still surpasses other current models. Compared to TransU-Net, the proposed model exhibits significant improvements in all metrics, indicating its ability to segment more cells in low-contrast regions and subtle edges, with better adaptability to brightness, noise, and other interferences, resulting in superior segmentation outcomes.

Table 2. Quantitative analysis of different segmentation models.

Model	***Pre***	***Rec***	F_1
U-Net	0.8014	0.7585	0.8462
R2U-Net	0.8219	0.8034	0.8739
TransU-Net	0.9062	0.8441	0.9136
Ours	0.9271	0.9172	0.9483

6. CONCLUSION

Addressing the challenges of insufficient utilization of image feature information and insufficient segmentation accuracy, which are caused by the complex background of aerial insulator images and the scarcity of standard datasets, this paper proposes a model based on the U-Net architecture. The core component of it comprises a 2D-Selective-Scan along with a Squeeze Excitation block. By employing a visual state space module to replace traditional convolutional blocks, the model achieves precise localization and segmentation of insulators with varying sizes, shapes, and complex backgrounds. This approach significantly reduces the number of model parameters and enhances real-time performance. Furthermore, the integration of an attention mechanism in this module effectively utilizes both local and global feature information from the current layer, along with the rich semantic information embedded in the bottom data of the model, to mitigate information loss during the encoding-decoding process.

Experimental results on an aerial insulator image dataset using ISMU-Net demonstrate that the proposed model can segment insulator regions of different sizes more comprehensively compared to other advanced models. It also exhibits strong generalization performance under various backgrounds. This model not only enables insulator segmentation under different backgrounds but also satisfies the requirements for real-time segmentation and detection during aerial photography.

ACKNOWLEDGMENTS

This research is supported by the Anhui Provincial Department of Education, with the project name: Anhui University Natural Science Key Research Project (Project No. 2022AH052783) and Anhui Higher Education Quality Engineering Project (Project No. 2022cjrh052) and Anhui Higher Education Quality Engineering Project (Project No. 2021jxtd320).

REFERENCES

[1] Jenssen, R. and Roverso, D., “Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning,” International Journal of Electrical Power & Energy Systems 99, 107-120 (2018).
[2] Zheng, X., et al., “Component identification and defect detection in transmission lines based on deep learning,” Journal of Intelligent and Fuzzy Systems 40, 3147-3158 (2021).
[3] Kumar, V., Rajesh, P., Jeyanthy, A. and Kesavamoorthy, R., “Optimization-assisted CNN model for fault classification and site location in transmission lines,” International Journal of Image and Graphics 24, 1 (2024).
[4] Huang, H., et al., “Research on recognition and location method of insulator in infrared image based on deep learning,” Journal of Physics: Conference Series 2087, 012090 (2021).
[5] Zheng, Z., Ni, C. and Zeng, G., “Pedestrian Detection Algorithm Based on Improved YOLOv4,” 2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (2023).
[6] Zai, H., Long, H. and Feng, L., “Target tracking method of transmission line insulator based on multi feature fusion and adaptive scale filter,” 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE) (2020).
[7] Yu, J., Liu, K., He, M. and Qin, L., “Insulator defect detection: A detection method of target search and cascade recognition,” Energy Reports 7, 750-759 (2021).
[8] Wei, Z., “Composite insulator defect identification and quantitative method based on random Hough transform ellipse detection,” Journal of Physics: Conference Series 2170.1 (2022).
[9] Bakshi, U., Sarkar, M., Paul, S., et al., “Assessment of virulence potential of uncharacterized Enterococcus faecalis strains using pan genomic approach-Identification of pathogen–specific and habitat-specific genes,” Scientific Reports 6(1), 38648 (2016).
[10] Jiang, H. R., et al., “Recognition and fault diagnosis of insulator string in aerial images,” Journal of Mechanical & Electrical Engineering, 32, 274-278 (2015).
[11] Huang, X., et al., “Composite insulator images segmentation technology based on improved color difference,” High Voltage Engineering, 44, 2493-2500 (2018).
[12] Ba, Jimmy, Jamie Ryan Kiros and Geoffrey E. Hinton. “Layer Normalization.” arXiv:1607.06450, (2016).
[13] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q. and Liu, Y., “Vmamba: Visual state space model,” arXiv:2401.10166, (2024).

[14] Stefan, E., Uchibe, E. and Doya, K., “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning,” Neural Networks, (2018).
[15] Gu, A. and Dao, T., “Mamba: Linear-time sequence modeling with selective state spaces,” arXiv arXiv:2312.00752, (2023).
[16] Gu, A., Goel, K. and Ré, C., “Efficiently modeling long sequences with structured state spaces,” arXiv arXiv:2111.00396, (2021).
[17] Brank, J., Mladenic, D., Grobelnik, M., et al., [F-Measure], Springer, Berlin, (2010).
[18] Wei, X. and Zhenmin, T., “Pavement crack detection based on image saliency,” Journal of Image and Graphics, 18, 01, 69-77 (2013).