Research on Visual Fault Diagnosis Technology for Power Transformers Based on Diffusion Model Sample Augmentation

1st Wenxuan Ye
College of Automation & Artificial Intelligence
College of Advanced Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
1222056625@njupt.edu.cn

2nd Xiangsen Wei
College of Automation & Artificial Intelligence
College of Advanced Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
weixiangsen97@163.com

***Abstract*—The issue of insufficient sample size is a key factor restricting the effectiveness of deep learning techniques in the field of visual fault diagnosis for power transformers. To address this, we propose a visual fault diagnosis technology for power transformers based on diffusion model sample augmentation. This method first fine-tunes the diffusion model using LoRA to generate visual fault samples for transformers, thereby expanding the dataset. Subsequently, the diagnostic model is improved to better adapt to the complexity of transformer fault samples, aiming to enhance diagnostic accuracy. The enhanced diagnostic performance of the model is evaluated using an assessment system composed of accuracy, F1 score, and a confusion matrix. The experimental results demonstrate that our proposed method has good diagnostic effectiveness for visible faults in transformers.**

***Index Terms*—Diffusion Model, LoRA, Sample Augmentation, power transformer, fault diagnosis**

I. Introduction

Power transformers, as one of the key devices in the operation of electric grids, are used on a large scale. Their safe and stable operation is of great significance for ensuring the supply of high-quality electricity. Therefore, the use of new technologies such as artificial intelligence for fault diagnosis of power transformers has broad application prospects [1]–[3]. However, due to the lack of transformer fault cases and abnormal operating condition data, visual fault samples for power transformers are extremely scarce, limiting the application of many advanced technologies in transformer fault diagnosis. Currently, traditional classification models based on artificial neural networks [4], random forests [5], and others have achieved certain results in transformer fault diagnosis. However, These models primarily focus on minimizing loss values or enhancing class distinctions. When constrained by small sample sizes, they are prone to overfitting, which can significantly compromise diagnostic reliability and present substantial safety threats to power systems.

Moreover, the diagnosis of transformer fault currently relies heavily on Dissolved Gas Analysis (DGA) [6]. This technology effectively identifies transformer faults by analyzing the gases released during the fault process, playing a crucial role in fault diagnosis. However, DGA technology may have delayed response times in situations where faults need time to produce sufficient gases, potentially missing early diagnosis and repair opportunities.

In addition to the widely used Dissolved Gas Analysis (DGA) technique, image-based methods have also demonstrated potential for transformer fault diagnosis. For instance, Jiang et al. [7] applied infrared images of transformer bushings to diagnose overheating faults, achieving commendable experimental results. Similarly, Li et al. [8] used infrared images as a dataset for transformer fault diagnosis, receiving positive feedback. However, many visible faults do not cause temperature changes, limiting the effectiveness of diagnoses using only infrared images. Therefore, we combine visible and infrared fault image samples to expand the range of detectable faults. Yet, both types of samples are quite scarce.

In this paper, we propose a transformer fault diagnosis method based on diffusion model sample augmentation. Diffusion models, as the latest generative models, have been proven by Dhariwal et al. [9] to surpass traditional GANs in terms of image generation quality. Furthermore, considering the high training and inference costs of original diffusion models like DDIM [10], the Latent Diffusion Model proposed by Rombach et al. [11] significantly reduce training parameters. On this basis, Stability AI made further innovations by adopting a pretrained CLIP Text Encoder as the domain-specific encoder for the Latent Diffusion model and developing the stable diffusion model through larger scale data training. Stable diffusion model inherits the efficiency of latent diffusion model, making high-quality image generation more feasible and economical. However, fully training the stable diffusion model still requires a significant amount of training data and computational resources, which is clearly impractical for the scenario of power transformer fault diagnosis. Therefore, we consider employing the LoRA [12] fine-tuning method on the pretrained stable diffusion model to better adapt it to our specific task. Although the LoRA fine-tuning method was initially used for language models, experiments have shown

2024 China Automation Congress (CAC) | 979-8-3503-6860-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CAC63892.2024.10865784

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on November 19,2025 at 07:22:42 UTC from IEEE Xplore. Restrictions apply.

that it is equally applicable to the fine-tuning of diffusion models, facilitating their application in downstream tasks.

Using the aforementioned method, we generated numerous visual fault samples for transformers, addressing the issue of scarcity in transformer fault samples. Additionally, we proposed an advanced fault diagnosis model, this model not only focuses on the global long-range dependencies of fault images but also pays attention to their critical local information, thereby better adapting to the complexity of transformer fault images and improving the precision of fault diagnosis. This paper explores how advanced image processing techniques can be applied effectively in specific engineering contexts to address real-world challenges.

II. Methodology

In this chapter, we will delve into the two core components of our study: the technology of using diffusion models for the augmentation of transformer fault image samples, and the fault diagnosis model designed for diagnosing faults in power transformers. First, we will introduce the basic principles of Stable Diffusion and its application in augmenting transformer fault samples. Subsequently, we will elaborate on the design concept and working principle of the fault diagnosis model.

A. Stable Diffusion Model

Diffusion models are a form of probabilistic model that learn the data distribution $p(x)$ by progressively denoising a variable from a normal distribution. These models consist of two processes: the forward process, also known as the diffusion process, and the reverse process. Diffusion models can be interpreted as a sequence of equally weighted denoising autoencoders $\epsilon_\theta(x_t, t)$. They are trained to predict the denoised version of their input x_t , where x_t is a noisy version of the input x. The specific process is illustrated in Fig. 1.

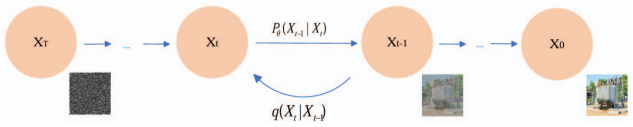

Fig. 1. Schematic diagram of diffusion model principles

The forward process is a noise-adding process. Given an initial image $x_0 \sim q(x)$, we progressively add noise to obtain $x_1, x_2, \ldots, x_T$. The final image x_T is entirely noise, conforming to a normal distribution. Moreover, during the forward process, x_T is only related to its immediate predecessor x_{t-1} , thus it can be considered as a Markov process:

$$q(x_t \mid x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{1}$$

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \tag{2}$$

In the process, β_t is the hyperparameter for the variance of the Gaussian distribution, which increases gradually. Based on the above formula, x_t can be obtained through reparameterized sampling, establishing the relationship between x_t and x_0 :

$$q(x_t \mid x_0) = N\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t) I\right) \tag{3}$$

In this context, $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$.

The backward process is a step-by-step denoising process that can gradually restore an image from the random noise x_T , Since it's impossible to directly infer $q(x_{t-1} \mid x_t)$, the U-net model is used to predict the reverse distribution p_θ .

$$p_\theta(X_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \tag{4}$$

$$p_\theta(x_{t-1} \mid x_t) = N\left(x_{t-1}; \mu(x_t, t), \sum_\theta(x_t, t)\right) \tag{5}$$

Given that x_0 is known, the Bayesian formula yields $q(x_{t-1} \mid x_t)$ as follows:

$$q(x_{t-1} \mid x_t, x_0) = N\left(x_{t-1}; \tilde{\mu}(x_t, x_0), \widetilde{\beta}_t I\right) \tag{6}$$

Its variance and mean are:

$$\widetilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{7}$$

$$\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}z_t\right) \tag{8}$$

Where z_t is the random normal distribution at time t originating from reparameterization. Ultimately, the training function can be obtained as follows:

$$L = E_{x_0, \bar{z}_t}\left[\left\|\bar{z}_t - z_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{z}_t, t\right)\right\|\right] \tag{9}$$

Diffusion models can produce high-quality images, but they consume a lot of memory and are slow when generating complex images. The stable diffusion model can address these issues.

The stable diffusion model first requires a well-trained Auto-Encoder. The encoder ε compresses the image, performing diffusion operations in a latent space, and then the decoder D restores the image back to its original pixel space. This process of compressing images with the encoder ε is known as perceptual compression [11]. Perceptual compression can ignore the high-frequency information in images, retaining only the essential and fundamental features, markedly lowering the computational load during training and sampling stages. In addition, the stable diffusion model introduces a conditional mechanism for conditional image generation. This is mainly achieved through a U-net model with cross-attention layers, denoted as $\epsilon_\theta(z_t, t, y)$. Furthermore, the stable diffusion model introduces a domain-specific encoder (CLIP Text Encoder) to map y (images, text, etc.) to an intermediate representation $\tau_\theta(y)$, which is then integrated into the Cross-Attention layer, allowing the control of the image synthesis process through y. Ultimately, the training function of the stable diffusion model is as follows:

$$L_{SDM} = E_{\varepsilon_x, y, \epsilon \sim N(0,1), t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \tag{10}$$

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on November 19,2025 at 07:22:42 UTC from IEEE Xplore. Restrictions apply.

B. Utilizing LoRA for Fine-Tuning the Stable Diffusion Model

Although the stable diffusion model has significantly reduced the computational difficulty during the training and sampling phases, training a stable diffusion model to its full capacity still requires a large amount of data. For power transformers, the availability of visual fault images is extremely limited, thus failing to meet the requirements for fully training stable diffusion model. However, pretrained models often have inherent dimensions [12]. By using LoRA fine-tuning on the inherent dimensions of stable diffusion model, it requires only a small number of training samples to achieve good training results.

In the stable diffusion model, the LoRA technique is applied to both the Cross-Attention layer and the domain-specific encoder. Therefore, by training the LoRA model, it is possible to integrate both visual and semantic features of electric power transformers into the stable diffusion model, thereby generating visual fault image samples of electric power transformers.

Specifically, we initially collected 100 high-quality images for each type of fault and conducted detailed annotations on them, as shown in Fig. 2. We paid special attention to aligning the visual and semantic features of the images, which included key information such as the type of fault, color, and background. Based on these efforts, we created a series of transformer fault image samples with corresponding text labels, providing a solid data foundation for training the LoRA model.

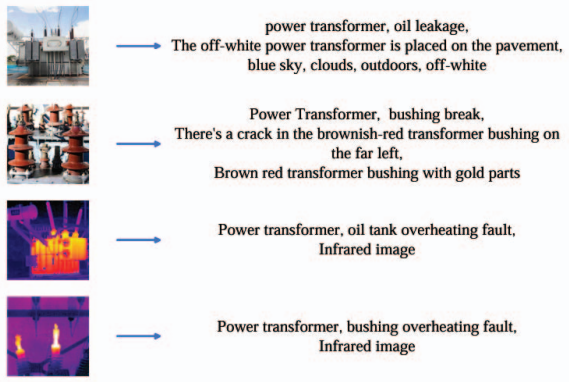

Fig. 2. Display of the LoRA model training dataset.

Then, we froze the original pre-trained model parameters and added additional trainable layers to the Cross-Attention layer and CLIP Text Encoder for training. For the four types of faults, we individually trained the corresponding LoRA models.

C. Synthesizing Power Transformer Visual Fault Images

We combined LoRA models with the pretrained stable diffusion model to generate visual fault samples of electric power transformers. Here, we employed two methods of generation:

The first method used text-conditioning to generate corresponding fault images, as illustrated in Fig. 3. Initially, we randomly generated a fixed-dimension latent noise vector. Then, We inputted textual information, including fault type and background, to map them into corresponding textual vectors $\tau_\theta(y)$ using the CLIP text encoder,which was fine-tuned with LoRA. Subsequently, the latent noise vector x_t and textual vectors $\tau_\theta(y)$ were jointly inputted into the U-net, which had been fine-tuned with LoRA. Through the Cross attention mechanism within it, the textual information was integrated with the image generation process, guiding the denoising process to generate images consistent with the textual vectors of transformer faults. Utilizing the DDIM sampling method, we executed 100 iterations within the U-net, progressively refining the denoising to capture the specific characteristics of power transformer fault information in the latent vector representation. Finally, this refined latent vector was decoded, resulting in the generation of high-quality transformer fault images.

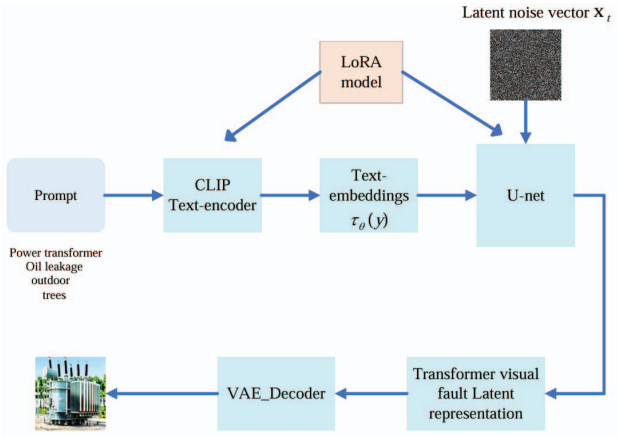

Fig. 3. By inputting textual features such as power transformer, oil leakage, outdoor, and trees, we can generate images of power transformer oil leakage fault in the corresponding environment.

The second method involved using normal transformer images to generate corresponding fault images. The working principle of this method was depicted in Fig. 4. Initially, a normal transformer image was mapped from the pixel space to the latent space to obtain a latent representation. Then, noises were added to this latent representation, resulting in a noise-affected latent representation, denoted as z_{noise}. Meanwhile, the fault information was also input and mapped to the corresponding textual vector $\tau_\theta(y)$. Subsequently, $\tau_\theta(y)$ and z_{noise} were fed into the fine-tuned U-net model. Using the DDIM sampling method, the U-net underwent 100 iterations to continuously denoise, ultimately obtaining the latent space vector of the transformer fault image. This latent space vector was then input into the decoder for decoding, yielding transformer fault image.

D. Power Transformer Fault Diagnosis Model

The difficulty in visual fault diagnosis of power transformers lies in the complex construction of power transformers and the intricate background environments, leading to high complexity

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on November 19,2025 at 07:22:42 UTC from IEEE Xplore. Restrictions apply.

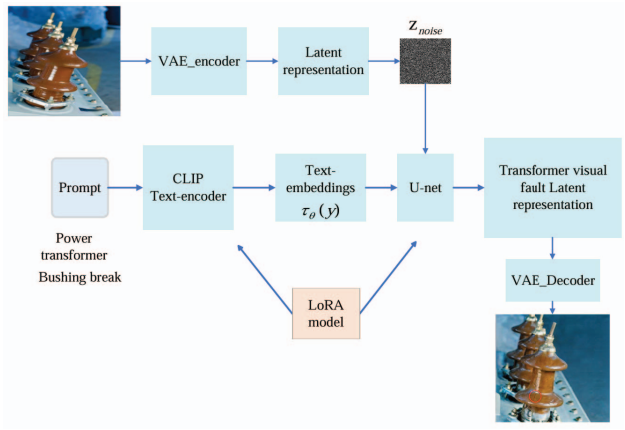

Fig. 4. We input a normal transformer bushing image and generate its image in the fault state.

in fault images. Moreover, since overheating faults are primarily reflected through low-resolution, low-contrast infrared images, the processing and interpretation of infrared images add extra difficulty to fault diagnosis. To address these issues, we employed sample augmentation techniques to generate a large number of samples, providing the diagnostic model with ample image information. In addition, we had optimized the diagnostic model to make it more suitable for handling such complex tasks. We utilized the ConvNeXt-T network as the foundational network for the diagnostic model. The ConvNeXt network [13], introduced in 2022, references the Swin Transformer [14] in the stacking of Blocks. It boasts a faster inference speed and higher accuracy compared to the Swin Transformer.

Based on the architecture of the ConvNeXt-T network, we made specific improvements to better adapt to the complexity of transformer fault samples. Particularly in dealing with a mix of visible light and infrared images, we optimized the network to enhance its ability to classify faults in different types of image data.

Specifically, instead of using the traditional approach of directly adding the input to the features after convolution, the input is processed in parallel through a Multi-Encoder Feature Enhancement (MEFE) module and an Efficient MLP module. The MEFE module includes a series of convolutional layers and an encoding layer designed to capture and encode input features, focusing on aggregating the local corner areas of the input image. This design ensures that the model retains these critical local information during the recognition process, preventing the loss of important details [15]. The Efficient MLP module primarily consists of two residual structures: a depthwise convolution-based residual block and a ReLU-MLP module with ReLU activation function. Both modules include a channel scaling layer and a DropPath layer to enhance the model's generalization and robustness, and the ReLU activation function is used to improve the network's representational capability. The primary function of the Efficient MLP module is to capture the global long-range dependencies of the input features. Finally, the features processed by the MEFE and Efficient MLP modules are concatenated and then added to the transformed features after convolution. The improved Block is shown in Fig. 5.

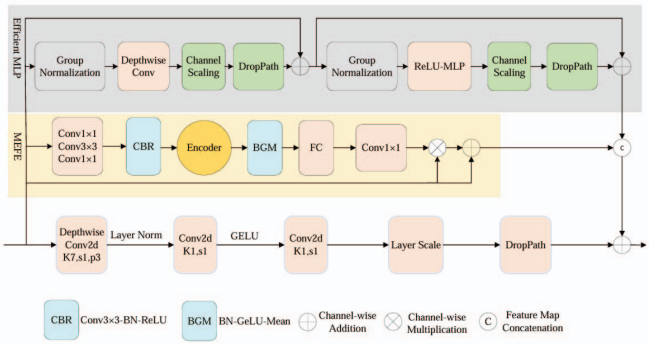

Fig. 5. Block with MEFE and Efficient MLP modules.

By introducing the new MEFE-MLP Block, the design of the ConvNeXt network has been optimized, enabling the network to more accurately differentiate between transformer faults, thereby enhancing the accuracy of fault diagnosis. Fig. 6 shows the structure of the ConvNeXt network after integrating the MEFE-MLP Block.

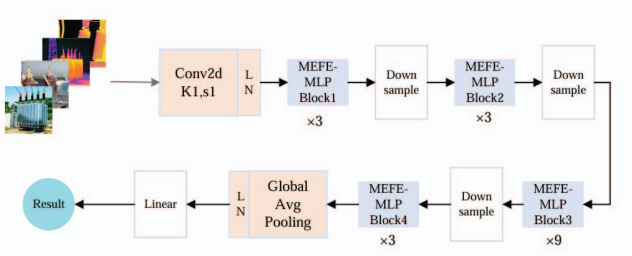

Fig. 6. Improved power transformer fault diagnosis network.

III. Experiments And Results Analysis

In this section, we first demonstrated various visually faulty transformer samples that have been generated. Then, we conducted experiments using these samples on various models to validate the effectiveness of the generated samples and the Progressiveness of the diagnostic model we proposed. All experiments were conducted on a computer using Python 3.7 and a single RTX4090.

A. Displaying Generated Images

Using the aforementioned method, we generated a large number of transformer fault image samples, as shown in Fig. 7. We then merged these generated samples with the real samples, creating an augmented electric power transformer fault image dataset. This dataset is of crucial importance for the study of electric power transformer fault diagnosis. Especially for research using deep learning methods for transformer fault diagnosis, this dataset provides valuable training data.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on November 19,2025 at 07:22:42 UTC from IEEE Xplore. Restrictions apply.

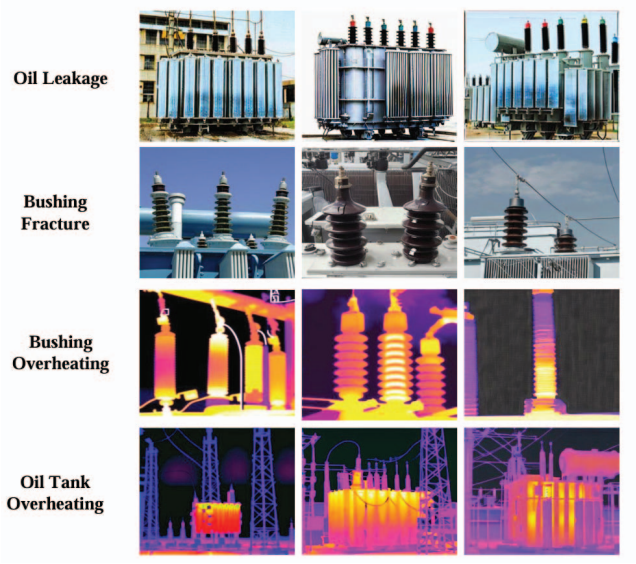

Fig. 7. Display of the generated visual fault images of power transformers.

B. Analysis of Transformer Fault Diagnosis Results

In the initial phase of our study, we used a total of 100 authentic fault images for each of the four fault categories. By employing a random sampling method, 90% of these images were allocated to the training set for model training, while the remaining 10% constituted the test set for model performance evaluation. Subsequently, we carried out the following experiments to validate the contribution of the generated images to the improved accuracy of transformer fault diagnosis, as well as to confirm the enhanced performance of the proposed model: we initially used 250 generated images per fault category as the training set, while employing all authentic images as the test set. Subsequently, in each experiment, we added an additional 250 generated samples to the training set for each fault category until the total reached 1000 samples per category. Throughout all experiments, we utilized the complete set of real fault samples as the test set to assess the diagnostic performance of the model. Upon training completion, we meticulously documented and compared the different models' accuracy across different scales of training sets, as demonstrated in Table I . As can be seen from Table I , When using only real images for fault diagnosis, the limited sample size prevents the model from learning sufficiently, resulting in lower diagnostic accuracy for various models. In contrast, when using generated samples as the training set and gradually increasing the sample size, the diagnostic accuracy of various models on real samples has improved. Moreover, when using 4000 generated images as the training set, the diagnostic model proposed in this paper has demonstrated higher accuracy compared to other models, reaching up to 96%.

Simultaneously, we recorded the confusion matrix of MEFE-MLP ConvNeXt when utilizing 4000 generated images as the training set, as illustrated in Fig. 8. The Data shows the model has an overall high diagnostic accuracy across the four categories, but there are still some misclassifications between categories.

TABLE I
COMPARISON OF ACCURACY TESTING OF DIFFERENT DIAGNOSTIC MODELS ON VARIOUS TRAINING

Algorithm	Training set	Test set	Accuracy	F1 score
Swin Transformer	360	40	87.75%	0.8778
	1000	400	89.00%	0.8903
	2000	400	91.25%	0.9122
	3000	400	91.50%	0.9154
	4000	400	92.75%	0.9273
MobileViT [16]	360	40	87.50%	0.8752
	1000	400	89.00%	0.8901
	2000	400	91.00%	0.9105
	3000	400	90.50%	0.9054
	4000	400	91.50%	0.9153
ShuffleNet [17]	360	40	90.00%	0.9004
	1000	400	92.25%	0.9221
	2000	400	92.75%	0.9276
	3000	400	93.25%	0.9321
	4000	400	94.00%	0.9400
EfficientNet [18]	360	40	90.00%	0.8998
	1000	400	89.50%	0.8951
	2000	400	91.50%	0.9156
	3000	400	92.75%	0.9271
	4000	400	95.00%	0.9505
DenseNet [19]	360	40	92.50%	0.9249
	1000	400	92.75%	0.9271
	2000	400	93.50%	0.9356
	3000	400	93.50%	0.9351
	4000	400	93.75%	0.9375
ConvNeXt	360	40	92.50%	0.9248
	1000	400	92.75%	0.9279
	2000	400	93.50%	0.9352
	3000	400	94.25%	0.9423
	4000	400	94.50%	0.9449
MEFE-MLP ConvNeXt	360	40	92.25%	0.9227
	1000	400	93.00%	0.9302
	2000	400	94.25%	0.9423
	3000	400	95.00%	0.9504
	4000	400	96.00%	0.9600

And, as shown in Fig. 9, when the model proposed in this paper uses 4000 generated samples as the training set and all real samples as the test set, except for the diagnostic accuracy for bushing fracture being slightly lower than that of the EfficientNet, the diagnostic accuracy for other fault types is the highest among all models.

IV. CONCLUSION

This paper proposes a visual fault diagnosis technique for power transformers based on sample augmentation using the stable diffusion model. The main research contents are as follows:

- Proposing a method for sample augmentation of various types of visual fault images in power transformers by using LoRA to fine-tune the stable diffusion model.
- Developing a unique dataset of visual fault images for power transformers, which is validated for its effectiveness and reliability through experiments and expert experience.
- By improving the diagnostic network, we enhance the accuracy of fault diagnosis, providing innovative ideas

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on November 19,2025 at 07:22:42 UTC from IEEE Xplore. Restrictions apply.

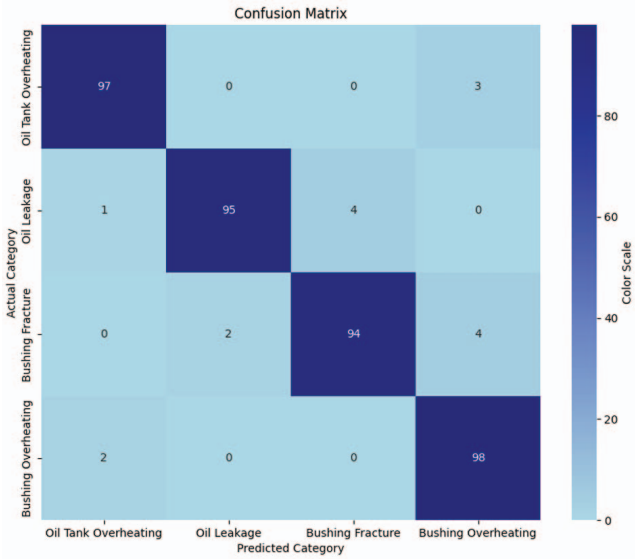

Fig. 8. Confusion matrix of the MEFE-MLP ConvNeXt model trained On 4000 generated images and tested on 400 real images.

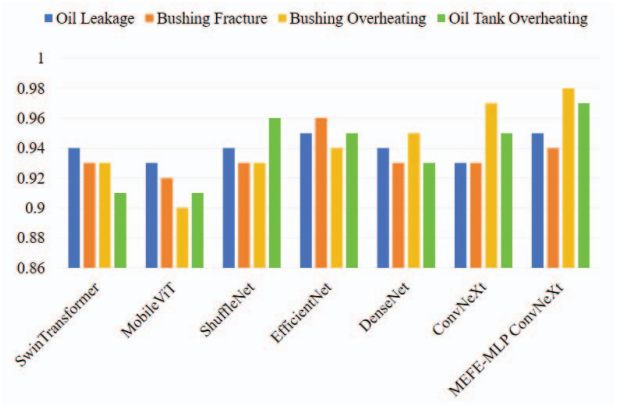

Fig. 9. When using 4000 generated samples as the training set, the diagnostic results of various models for four types of faults.

for the further development of visual fault diagnosis techniques for power transformers.

In the future, we aim to further enhance the quality and realism of the generated images to more accurately simulate various complex fault scenarios. At the same time, we plan to explore more advanced fault diagnosis algorithms to further improve diagnostic accuracy.

REFERENCES

[1] J. Wang, F. Zhang, H. Liu, J. Ding, and C. Gao, "Interruptible load scheduling model based on an improved chicken swarm optimization algorithm," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 232–240, 2020.

[2] L. Ruijin, W. Youyuan, L. Hang, L. Hongbo, and M. Zhipeng, "Research status of condition assessment method for power equipment," *High Voltage Engineering*, vol. 44, no. 11, pp. 3454–3464, 2018.

[3] Z. Zhang, C. Dou, D. Yue, Y. Xue, X. Xie, C. Deng, and B. Zhang, "Voltage sensitivity-related hybrid coordinated power control for voltage regulation in adns," *IEEE Transactions on Smart Grid*, vol. 15, pp. 1388–1398, 2024.

[4] Z. Xing and Y. He, "Multimodal mutual neural network for health assessment of power transformer," *IEEE Systems Journal*, vol. 17, no. 2, pp. 2664–2673, 2023.

[5] A. M. Shah and B. R. Bhalja, "Fault discrimination scheme for power transformer using random forest technique," *IET Generation, Transmission & Distribution*, vol. 10, no. 6, pp. 1431–1439, 2016.

[6] M. S. Ali, A. Omar, A. S. A. Jaafar, S. H. Mohamed *et al.*, "Conventional methods of dissolved gas analysis using oil-immersed power transformer for fault diagnosis: A review," *Electric Power Systems Research*, vol. 216, p. 109064, 2023.

[7] J. Jiang, Y. Bie, J. Li, X. Yang, G. Ma, Y. Lu, and C. Zhang, "Fault diagnosis of the bushing infrared images based on mask r-cnn and improved pcnn joint algorithm," *High voltage*, vol. 6, no. 1, pp. 116–124, 2021.

[8] X. Li, "Design of infrared anomaly detection for power equipment based on yolov3," in *2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2)*. IEEE, 2019, pp. 2291–2294.

[9] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[10] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[15] Y. Quan, D. Zhang, L. Zhang, and J. Tang, "Centralized feature pyramid for object detection," *IEEE Transactions on Image Processing*, 2023.

[16] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[17] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[18] B. Koonce and B. Koonce, "Efficientnet," *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pp. 109–123, 2021.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on November 19,2025 at 07:22:42 UTC from IEEE Xplore. Restrictions apply.