Application of Deep Learning-Based Image Compression Restoration Technology in Power System Unstructured Data Management

Junjie Zha*, Aiguo Teng, Xinwen Shan, Hao Tang, Zihan Liu
State Grid Jiangsu Electric Power Co., Ltd, Information & Telecommunication Branch, Jiangsu Nanjing, 210000 China

***Abstract*—In power-system unstructured-data management, a large volume of images from inspection drones, substation cameras, and smart meters is heavily compressed due to bandwidth and storage constraints, resulting in lower resolution that hinders defect detection and maintenance decisions. Although deep-learning super-resolution (SR) techniques have made significant advances, real-world deployments still require a balance between reconstruction accuracy and model lightweightness. To meet this need, we introduce a channel-attention-embedded Transformer SR method (CAET). The approach adaptively injects channel attention into both the Transformer's global features and the convolutional local features, harnessing their complementary strengths while dynamically enhancing critical information. Tested on five public datasets and compared with six representative algorithms, CAET achieves the best or second-best performance across all upscaling factors; at 4× enlargement, it outperforms the advanced SwinIR method by 0.09 dB in PSNR on Urban100 and by 0.30 dB on Manga109, with noticeably improved visual quality. Experiments demonstrate that CAET delivers high-precision, low-latency restoration of compressed images for the power sector while keeping model complexity low.**

Keywords*—Image compression; attention mechanism; multimodal fusion; unstructured data in the power industry; image data

I. Introduction

In today's data-intensive power systems, images captured by UAV patrols, robot inspections, smart meters and substation cameras constitute a major source of unstructured information. Because these visual records are often transmitted or stored under stringent bandwidth and storage limits, aggressive compression and down-sampling are common, which sharply degrades resolution and hinders defect detection, asset identification and other downstream analytics. High-resolution (HR) images convey richer detail than their low-resolution (LR) counterparts [1-2], so recovering spatial fidelity after compression is crucial for reliable power-equipment management. Single-image super-resolution (SISR) addresses this need by reconstructing an HR image from one LR input and has been proven in domains such as security, medical and satellite imaging [3-4].

Deep-learning methods dominate modern SISR, thanks to rapid progress in GPU computing power, achieving large leaps in reconstruction quality [17]. The seminal SRCNN network used a three-layer CNN to learn an LR–HR mapping [5]. Deeper CNNs such as VDSR with 20 layers [13], EDSR that widens and deepens residual blocks [20][8], WDSR that activates wider feature dimensions, and RCAN that embeds channel attention to model inter-channel dependencies, have steadily improved results. Yet convolution itself has limitations: fixed filters ignore image content, and locality prevents long-range modeling. Transformer architecture [30] overcomes these issues via global self-attention that exploits self-similarity [7]. Large-scale Transformer variants (e.g., IPT [4]) and remote-sensing specific designs like TransENet [18, 26], as well as SwinIR built on Swin Transformer blocks [19] [22-23] have shown strong performance.

However, state-of-the-art networks are typically heavy, conflicting with the real-time and resource-constrained requirements of power-system back-end servers and edge devices. Lightweight CNN strategies—FSRCNN that operates directly on LR inputs [6], recursive models DRCN and DRRN that cut depth via weight sharing [14, 28], and distillation-based IMDN and RFDN that progressively extract informative features while trimming parameters [11, 21]—have mitigated this to a degree. The Swin Transformer's windowed attention further lowers complexity without sacrificing quality and performs well on restoration tasks.

Building on these insights, we propose a channel-attention-embedded Transformer (CAET) for compressed-image restoration in power-grid data platforms. The method alternates Transformer attention with convolution, exploiting their complementary strengths, and uses channel attention to adaptively fuse features, enhancing learning capacity while remaining lightweight for deployment. Experimental comparisons on multiple datasets confirm that CAET attains notable quality gains with modest model size, making it suitable for large-scale unstructured image archives in the power sector. Key advantages include effective integration of convolution and Transformer feature extraction via the CAET block, adaptive weighting of multi-level features through channel attention, and superior performance versus existing SR approaches at lower computational cost.

In response to the common issues of complexity, high resource consumption, and the difficulty of balancing performance and efficiency in existing image super-resolution methods, this study conducts research aimed at improving image reconstruction quality while reducing model complexity. We propose a Channel Attention-Embedded Transformer Image Super-Resolution Network (CAET), which integrates

the local modeling capabilities of convolutional networks with the global modeling advantages of Transformers. By introducing a channel attention mechanism to enhance the discrimination of multi-level features, the method improves the ability to restore image details. While maintaining a low parameter count and computational complexity, our approach achieves superior reconstruction performance compared to existing methods across multiple public datasets, demonstrating good lightweight properties and practicality. The main innovations of this study are: 1) the design of a channel attention-embedded Transformer module that effectively combines convolutional and Transformer features; 2) the proposal of a channel attention discrimination enhancement strategy to improve the network's learning ability; and 3) the validation of the superiority of this method under lightweight conditions across various diverse datasets.

II. Image Processing Related

Due to the demand for implementing super-resolution tasks on commonly used real-world devices, lightweight super-resolution models have attracted widespread attention. FSRCNN replaced the approach of SRCNN [5], which first upsampled images before feeding them into the network, by applying the core network directly to low-resolution (LR) images, significantly reducing the computational resources required [6]. DRCN [14] and DRRN [28-31] introduced recursive neural networks, which, while reducing the number of parameters and network depth, led to a significant increase in computational cost due to repeated recursion. LapSRN (Deep Laplacian Pyramid Network for Fast and Accurate Super-Resolution) [16] adopted a progressive strategy to gradually increase image resolution, achieving more stable results for high-scale upscaling. IMDN [10-12] proposed a lightweight Information Multi-distillation Network, which effectively extracts hierarchical features using an information distillation mechanism. LatticeNet (Image Super-Resolution with Lattice Block) [4, 18, 24], inspired by the Fast Fourier Transform, designed a lattice network capable of efficiently utilizing and adjusting multi-level information. ECBSR (Edge-oriented Convolutional Block for Real-time Super-Resolution) [15, 19, 36] introduced a convolutional block incorporating edge information based on re-parameterization techniques, which enhances the model's learning capacity while reducing inference time. Despite significant progress made by lightweight super-resolution algorithms, there is still room for improvement in reconstruction quality.

III. Channel Attention Embedded Transformer

A. Algorithmic Framework

The specific design diagram of the lightweight super-resolution network based on the combination of Transformer and convolution is shown in Fig. 1. It mainly consists of four parts: shallow feature extraction stage, deep feature extraction stage, multi-level feature fusion stage, and image reconstruction stage. Among them, the deep feature extraction stage is composed of Transformer modules embedded with attention mechanisms (CAETB), which will be described in detail. The network takes the given I_{LR} (input low-resolution image) and I_{SR} (inferred super-resolution image) as the low-resolution input and the predicted high-resolution output, respectively.

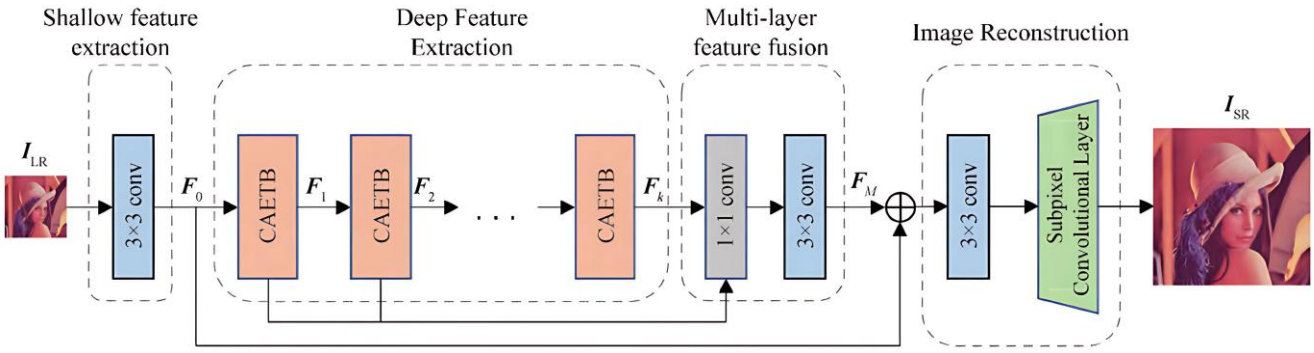

Fig. 1. Overall structure of Channel-Attention-Embedded Transformer network.

1) Shallow feature extraction stage. This stage uses a convolutional layer with a kernel size of 3 × 3 to extract shallow features F_0 from the given LR image. The process can be expressed as:

$$F_0 = H_{\mathrm{SF}}(I_{\mathrm{LR}}) \tag{1}$$

In this equation, H_{SF} represents the convolution operation. The extracted shallow feature F_0 will be further used for deep feature extraction. Meanwhile, F_0 is also directly passed to the reconstruction module to preserve the low-frequency information of the image.

2) Deep feature extraction stage. This stage takes the shallow feature F_0 as input and uses multiple CAETBs to extract deep feature information. Assuming the number of CAETBs is k, the output of the i^{th} CAETB, denoted as F_i $(1 \le i \le k)$, can be expressed as:

$$\boldsymbol{F}_i = H_{\mathrm{CAETB}}(\boldsymbol{F}_{i-1}), i = 1, 2, \cdots, k \tag{2}$$

In this equation, H_{CAETB} represents the operation of the CAETB, which is used to extract deep features of the image. The structure of CAETB is shown in Fig. 2.

3) Multi-level feature fusion stage. Hierarchical information from different stages is beneficial to the final reconstruction result. Therefore, in the multi-level feature fusion stage, the network integrates all low-level and high-level

information from the deep feature extraction stage. The fused result is denoted as $\boldsymbol{F}_M$.

$$\boldsymbol{F}_M = H_{\text{MFF}}(\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_i, \cdots, \boldsymbol{F}_k) \tag{3}$$

In this equation, H_{MFF} represents the multi-level feature fusion operation, which provides sufficient reference and guidance for the final image reconstruction.

4) Image reconstruction stage. The fused result F_{M} and the shallow feature F_0 are further input into the image reconstruction stage to recover high-resolution images adapted to different tasks. The process of obtaining the final high-resolution image I_{SR} can be expressed as follows:

$$\boldsymbol{I}_{\text{SR}} = H_{\text{REC}}(\boldsymbol{F}_0 + \boldsymbol{F}_M) \tag{4}$$

In this equation, HREC represents the operation in the reconstruction stage, which uses a 3 × 3 convolutional layer and the sub-pixel convolution layer from ESPCN [27] to upsample the features to the corresponding size of the super-resolution image.

B. Attention-Embedded Transformer Block

Convolutional layers provide more stable optimization and better feature extraction in early visual processing. Additionally, the shared weights of spatially invariant filters can enhance the network's translation equivariance [19]. Stacking convolutional layers effectively enlarges the network's receptive field. Therefore, three cascaded convolutional layers are placed at the front end of the CAETB to receive feature outputs from the previous module. To better adjust features from different levels and transformation units, the network employs a channel attention-based feature discrimination enhancement strategy, which performs channel attention-based discriminative enhancement and interactive fusion between Transformer features and convolutional features. Specifically, the input Transformer-transformed features and convolution-processed features undergo learnable channel attention feature enhancement and cross-fusion. The generation method of channel attention will be detailed in Section III(D). Taking the input feature as F_i, the process of channel attention feature discrimination enhancement can be expressed as follows.

$$\boldsymbol{F}_P = A \times H_{\text{CL}}(\boldsymbol{F}_i) + \boldsymbol{F}_i \tag{5}$$

$$\boldsymbol{F}_Q = H_{\text{CL}}(\boldsymbol{F}_i) + D \times \boldsymbol{F}_i \tag{6}$$

In this equation, A and D represent channel attention parameters, and H_{CL} denotes the cascaded convolutional layer operation. The three convolutional layers have channel numbers of 60, 45, and 60, respectively, with Leaky ReLU (LReLU) activation applied between them. Then, a convolutional layer with a kernel size of 1 × 1 is used to adjust the cascaded features back to the original number of channels. Its output, $\boldsymbol{F}_R$, can be expressed as:

$$F_R = H_C\left(Concat\left(F_P, F_Q\right)\right) \tag{7}$$

In this equation, Concat represents the feature concatenation operation, and H_C denotes the 1 × 1 convolutional layer operation. After passing through the attention embedding part, the features are input into the ST module for further feature extraction. The output of the module, F_S, can be expressed as:

$$F_s = H_{\text{Swin}}(F_R) \tag{8}$$

In this equation, H_{Swin} represents the operation of the ST module. The number of Swin Transformer layers (STL) in the CAETB module is set to 4.

The specific structure of CAETB is shown in Fig. 2. Since CAETB adopts a residual connection structure, the final output of the module, $\boldsymbol{F}_{i+1}$, can be expressed as:

$$F_{i+1} = F_s + F_i \tag{9}$$

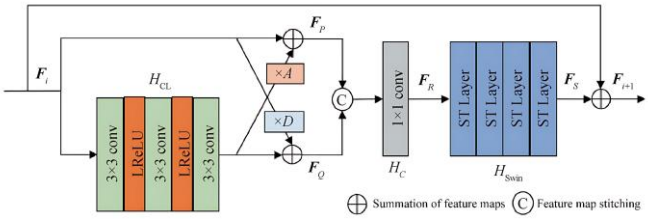

Fig. 2. Internal structure of the CAETB.

C. ST Module

The ST module is an improvement over the multi-head attention mechanism in the standard Transformer architecture. The standard Transformer performs global self-attention calculations on images, but the complexity of the global attention mechanism increases sharply as the image size grows. Therefore, when dealing with larger images in downstream vision tasks, the standard Transformer faces excessive memory requirements. To address this issue, the ST module introduces local attention mechanisms and a window-shifting mechanism. The overall structure of the Swin Transformer layer (STL) is shown in Fig. 3.

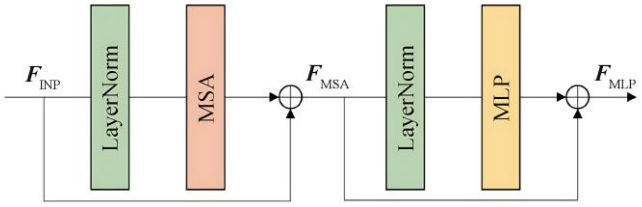

Fig. 3. Internal structure of STL.

For a given input of size $H \times W \times C$, ST first uses non-overlapping local windows of size $M \times M$ to reshape the input features into $\frac{HW}{M^2} \times M^2 \times C$, where $\frac{HW}{M^2}$ represents the total number of windows. Then, standard self-attention is computed separately for each window. For the local window feature $X \in R^{M^2 \times C}$, the query, key, and value matrices Q, K, and V are computed as follows:

$$Q = XP_Q, K = XP_K, V = XP_V \tag{10}$$

In this equation, PQ, PK, and PV are shared learnable projection matrices across different windows, and Q, K, V $\in R^{M^2 \times d}$. The attention matrix Attention (Q, K, V) is computed through the self-attention mechanism within the local window and can be expressed as follows:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}} + B\right)V \tag{11}$$

In this equation, *B* represents learnable relative position encoding, and *d* denotes the dimension of the multi-head attention. The results of multi-head self-attention (MSA) are concatenated to maintain the feature dimensions. Next, a multilayer perceptron (MLP) is used for further feature enhancement, consisting of two fully connected layers with GELU (Gaussian error linear unit) activation between them. Layer normalization (LN) is applied before both the MSA and MLP, and residual connections are used in both parts. Taking the input feature as F_{INP}, the overall process can be mathematically described as follows.

$$F_{\text{MSA}} = H_{\text{MSA}}\left(H_{\text{LN}}\left(F_{\text{INP}}\right)\right) + F_{\text{INP}} \tag{12}$$

$$F_{\text{MLP}} = H_{\text{MLP}}\left(H_{\text{LN}}\left(F_{\text{MSA}}\right)\right) + F_{\text{MSA}} \tag{13}$$

In this equation, HLN represents the layer normalization operation, HMSA denotes the multi-head attention operation, and HMLP stands for the multilayer perceptron operation.

Although the local attention mechanism based on window partitioning reduces computational complexity, fixed window partitions do not allow information exchange between windows. The ST module adopts shifted window partitioning and alternates between shifted and non-shifted windows to achieve cross-window connections. Specifically, shifted window partitioning means that features are shifted by (M/2, M/2) pixels before window partitioning, and the alternating rule means that shifted and non-shifted Swin Transformer layers (STLs) are used alternately. This approach addresses the problem of no information exchange between non-overlapping windows and significantly increases the receptive field of the ST module.

D. Channeling Attention Mechanisms

This section provides a detailed introduction to the calculation of channel attention parameters *A* and *D* mentioned. The channel attention mechanism [9] models the interdependencies between feature channels and can adaptively adjust the feature responses of different channels by assigning corresponding weights. Embedding channel attention enables adaptive enhancement and fusion of the corresponding convolutional and Transformer features within CAETB. Moreover, due to the local operation characteristics of convolution, each output value cannot represent the overall information of the entire image. Therefore, global information from channel attention is needed as guidance to select the most effective features across the whole image.

The specific operation of the channel attention mechanism is shown in Fig. 4. The input is denoted as X = [x1, x2, ..., xc], containing c channel feature maps each of size H×W. To obtain the global characteristics between feature channels, global average pooling (GAP) is used to acquire the statistical features of each channel, denoted as Z = [z1, z2, ..., zc]:

$$\boldsymbol{z}_C = H_{\text{GAP}}(\boldsymbol{x}_C) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_c(i,j) \tag{14}$$

In this equation, xc(i, j) represents the value at position (i, j) in the feature map xc, and HGAP denotes the global average pooling operation, which calculates the average value representing the global information of each feature map.

The above information is then input into the second process, called weight learning (WL). The weight learning process consists of two fully connected layers, a ReLU function, and a sigmoid function. The former learns the nonlinear interactions between channels through channel squeeze and excitation, while the latter restores the channels and normalizes the parameters to ensure the network can focus on multiple important channels simultaneously. This process can be expressed as:

$$W = H_{\text{WL}}(Z) = s\left(f_2 r\left(f_1 Z\right)\right) \tag{15}$$

In this equation, H_{WL} represents the overall weight learning process, $r(\cdot)$ and $s(\cdot)$ represent the *ReLU* and *sigmoid* functions, respectively, and f_1 and f_2 denote the two fully connected layers.

The learned weights $W = [w_1, w_2, ..., w_c]$ are multiplied by the original input features. The adjusted result is denoted as $X' = [x'_1, x'_2, ..., x'_c]$, where the adjusted feature map corresponding to the c^{th} channel can be expressed as:

$$x'_c = w_c x_c \tag{16}$$

In this equation, w_c represents the weight factor used to adjust the proportion of the original feature's weight, allowing the weight of important feature information to be increased.

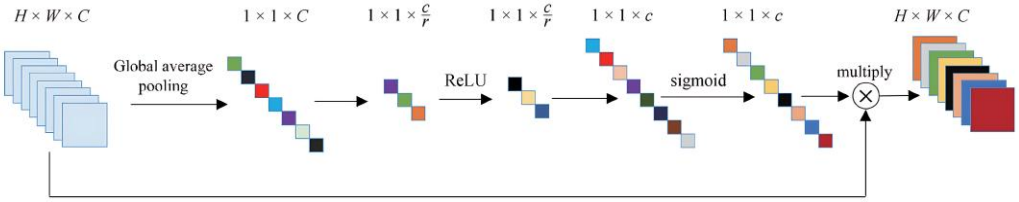

Fig. 4. Architecture of channel attention module.

E. Loss Function

For the sake of fairness in comparison, only the L_1 loss function is used to optimize the network. Given a training dataset of N image pairs $\left\{I_{LR}^{i}, I_{HR}^{i}\right\}_{i=1}^{N}$, the optimization process can be expressed as follows:

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}_{SR}^{i} - \boldsymbol{I}_{HR1}^{i} \tag{17}$$

In this equation, I_{SR}^{i} represents the high-resolution image predicted by the network for I_{LR}^{i}, I_{HR}^{i} denotes the ground-truth high-resolution image, and θ represents the learnable parameters of the network.

IV. Experimental Results and Analysis

A. Experimental Setup

For the model training part, this study uses the widely applied DIV2K dataset for image super-resolution tasks. Specifically, bicubic downsampling with three scale factors (×2, ×3, and ×4) is performed on 800 training images to obtain low-resolution input images. Data diversity is enhanced by randomly applying vertical rotations and horizontal flips to the training images. The network optimizer is *Adam*, with parameters β_1 = 0.9, β_2 = 0.999, and $\varepsilon = 10^{-8}$. The initial learning rate is set to 2×10^{-4} and is halved at the 150,000th, 300,000th, 400,000th, and 450,000th batches, with a total of 500,000 training batches. During training, each input batch consists of 32 randomly cropped low-resolution images of size 64 × 64 pixels (batch size = 32). For the super-resolution networks with ×3 and ×4 scale factors, training is performed based on the pretrained model with ×2 scale factor, and the total number of training iterations is halved.

For model evaluation, this study uses five commonly used public benchmark datasets: Set5, Set14, BSD100 (Berkeley Segmentation Dataset 100), Urban100, and Manga109. These datasets cover various types of image features and resolutions. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are used as objective evaluation metrics, with quantitative assessments performed on the luminance channel. Meanwhile, model complexity is measured by the number of parameters and the number of multiply-add operations (MAdds). The MAdds represent the cumulative number of multiplication and addition operations required by the model to process a single input image, with an output image size of 1280 × 720 pixels used as the benchmark. All experiments are conducted on the PyTorch platform, using an NVIDIA GTX 4090 GPU.

B. Analysis of CAETB Structure and Volume

To verify the effectiveness of the CAETB structure, the impact of different embedding combinations of channel attention and Transformer on the results is explored. Fig. 5 shows the case where channel attention is embedded after the Transformer, while the case where channel attention is embedded before the Transformer is the same as in CAETB.

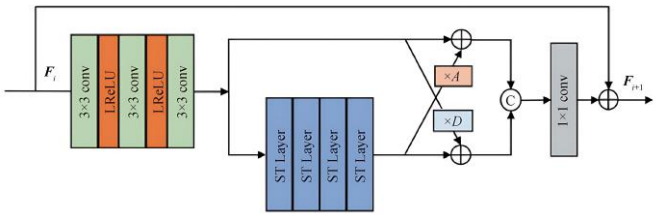

Fig. 5. Channel Attention（CA）after Transformer layers.

Table I shows the impact of different combination methods on reconstruction quality for the Urban100 and Set5 datasets at a scale factor of ×2. It can be seen that embedding channel attention before the Transformer layer achieves better reconstruction performance. During the actual training process, the model occupied approximately 7.6 GB of GPU memory when trained on an RTX 4080, with a single iteration taking about 0.85 seconds and a total training time of approximately 68 hours. In the inference stage, when the input image size is 1280×720, the average reconstruction time for a single image is 46 milliseconds, demonstrating real-time processing capability. Therefore, this model can be efficiently deployed on conventional high-performance GPU platforms, making it suitable for practical application scenarios that require a balance between performance and resources.

TABLE I. Effect of the Combination Style in CAETB on the Reconstruction Performance

Combinatorial model	Number of parameters/k	Urban 100	Set5
		PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
CA embedded in front	851	**32.79/0.9348**	**38.15/0.9618**
CA is embedded after	851	32.74/0.9340	38.12/0.9615

Note: Bold font indicates the best result in each column.

Since network depth plays an important role in improving reconstruction performance, the effect of increasing the number of CAETBs from two to five on the network's reconstruction results was investigated. Table II shows the reconstruction results with different numbers of CAETBs at a scale factor of ×2 on the Set14 dataset, along with an analysis of the required number of parameters. As shown in Table II, due to the strong nonlinear abstraction capability of deep networks, the performance improves as the number of modules increases. However, it can also be observed that the improvement in reconstruction results slows down as the number of CAETBs increases, a phenomenon known as the saturation of deep networks. To balance model complexity and performance, four CAETBs are selected to form the basic reconstruction network.

TABLE II. Effect of the Number of CAETBs on Parameter Size and Reconstruction Performance on SET14

Quantities	Number of parameters/k	PSNR/dB	SSIM
2	446	33.73	0.9195
3	649	33.83	0.9201
4	851	33.89	0.9204
5	1054	33.93	0.9206

C. Comparison with other Algorithms

To validate the effectiveness of the algorithm, this study compares it with lightweight networks such as bicubic interpolation, SRCNN [5], CARN [2], IMDN [11], LatticeNet [24-25], and SwinIR [19] under different scale factors (×2, ×3, ×4). These algorithms are representative and have superior performance.

The comparison results of PSNR and SSIM between the proposed algorithm and other methods are shown in Tables III to V. It can be seen that CAET leads in objective metrics PSNR and SSIM across all datasets and different upsampling factors. Specifically, on the Urban100 dataset, CAET outperforms the second-best method by 0.03 dB at an upsampling factor of ×2, 0.08 dB at ×3, and 0.09 dB at ×4. The most significant improvement is observed on the Manga109 dataset, where CAET surpasses the second-best by 0.13 dB at ×2, 0.33 dB at ×3, and 0.30 dB at ×4. In terms of model complexity, CAET maintains relatively low parameter counts and FLOPs, achieving better restoration performance while having lower complexity than SwinIR, which also uses a Transformer model. To further demonstrate the effectiveness of the proposed algorithm, the number of CAETBs was reduced to two, resulting in the CAET-M variant. Despite its significantly lower complexity compared to convolution-based lightweight SR algorithms like IMDN, CAET-M still achieves some performance gains. These results indicate that the proposed algorithm offers better overall performance compared to these lightweight methods.

To further analyze the impact of dataset diversity on model performance, this study conducts a statistical comparison of five test datasets from multiple dimensions, including image types (natural images, urban scenes, hand-drawn images, etc.), texture complexity, and image detail density. Set5 and Set14 primarily consist of simple natural images, while BSD100 has moderate complexity. In contrast, Urban100 and Manga109 represent urban architectural images with high structural complexity and detail-dense comic images, respectively. From the objective metrics presented in Tables III to V, it can be observed that the performance improvement of the CAET model is most significant on Urban100 and Manga109, achieving the highest PSNR gains (e.g., an increase of 0.09 dB and 0.30 dB at a 4x upscaling factor). This indicates that the model demonstrates stronger reconstruction capabilities on data rich in structure and texture details. The results suggest that the CAET model, through the combination of channel attention enhancement mechanisms and the global modeling capabilities of Transformers, can effectively extract features and recover high-frequency details even when faced with increased image content complexity. Thus, it achieves robust adaptability to diverse data. This outcome verifies that the model's structural design possesses good generalization capabilities for varied image content.

TABLE III. AVERAGE PSNR AND SSIM COMPARISON OF DIFFERENT ALGORITHMS UNDER MAGNIFICATION IS 2

Model	Number of parameters/k	Multiplier/G	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
Interpolation	-	-	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN	57	52.7	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
CARN	1592	222.8	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
LatticeNet	756	169.5	38.06/0.9610	33.70/0.9193	32.20/0.8999	32.25/0.9288	-/-
IMDN	694	158.8	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
SwinIR-light	878	243.7	38.14/0.9611	33.86/**0.9206**	32.31/0.9012	32.76/0.9340	39.12/**0.9783**
CAET (this study)	851	214.7	**38.15/0.9618**	**33.89**/0.9204	**32.34/0.901**	**32.79/0.9348**	**39.25**/0.9781
CAET	446	110.4	38.04/0.9613	33.73/0.9195	32.26/0.9007	32.39/0.9308	38.93/0.9777

Note: Bold and underlined fonts indicate the best and second-best results in each column, respectively. A dash "–" indicates the absence of corresponding data.

TABLE IV. AVERAGE PSNR AND SSIM COMPARISON OF DIFFERENT ALGORITHMS UNDER MAGNIFICATION IS 3

Model	Number of parameters/k	Multiplier/G	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
Interpolation	-	-	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN	57	52.7	32.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
CARN	1592	118.8	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.50/0.9440
LatticeNet	765	76.3	34.40/0.9272	30.32/0.8416	29.10/0.8049	28.19/0.8513	-/-
IMDN	703	71.5	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
SwinIR-light	886	19.5	34.62/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624	33.98/0.9478
CAET (this study)	859	98.4	**34.65/0.9297**	**30.61/0.8482**	**29.26/0.8099**	**28.74/0.8652**	**34.31/0.9491**
CAET	454	51.1	34.40/0.9278	30.44/0.8445	29.17/0.8076	28.40/0.8577	33.89/0.9466

Note: Bold and underlined text indicate the best and second-best results in each column, respectively. A dash "–" denotes missing or unavailable data.

TABLE V. AVERAGE PSNR AND SSIM COMPARISON OF DIFFERENT ALGORITHMS UNDER MAGNIFICATION IS 4

Model	Number of parameters/k	Multiplier/G	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
Interpolation	-	-	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN	57	52.7	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
CARN	1592	90.9	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.47/0.9084
LatticeNet	777	43.6	32.18/0.8943	28.61/0.7812	27.57/0.7355	26.14/0.7844	-/-
IMDN	715	40.9	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
SwinIR-light	897	61.7	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980	30.92/0.9151
CAET (this study)	871	55.4	**32.47/0.8997**	**28.80/0.7871**	**27.74/0.7427**	**26.56/0.8016**	**31.22/0.9168**
CAET	465	28.8	32.24/0.8963	28.70/0.7847	27.65/0.7394	26.26/0.7928	30.73/0.9118

Note: Bold and underlined fonts indicate the best and second-best results in each column, respectively. A dash "–" indicates that no corresponding data is available.

To further validate the advantages of the proposed network, representative visual results on standard benchmark datasets are analyzed. Fig. 6 shows the local enlargement results for image Urban100_img012 with an upsampling factor of ×4. It can be observed that other algorithms incorrectly restore the orientation of building textures, whereas CAET (proposed in this study) accurately preserves the texture structure of the buildings. Fig. 7 presents the local enlargement results for image B100_253027 at a ×4 scale. Compared to other methods, the proposed algorithm restores the zebra stripes with greater clarity and accuracy. Fig. 8 shows the ×4 enlargement results for image Set14_barbara. The proposed algorithm accurately reconstructs the arrangement of the books, while other methods exhibit varying degrees of distortion. These results demonstrate that the proposed method not only leads in objective metrics but also produces clearer super-resolved images than all the compared approaches.

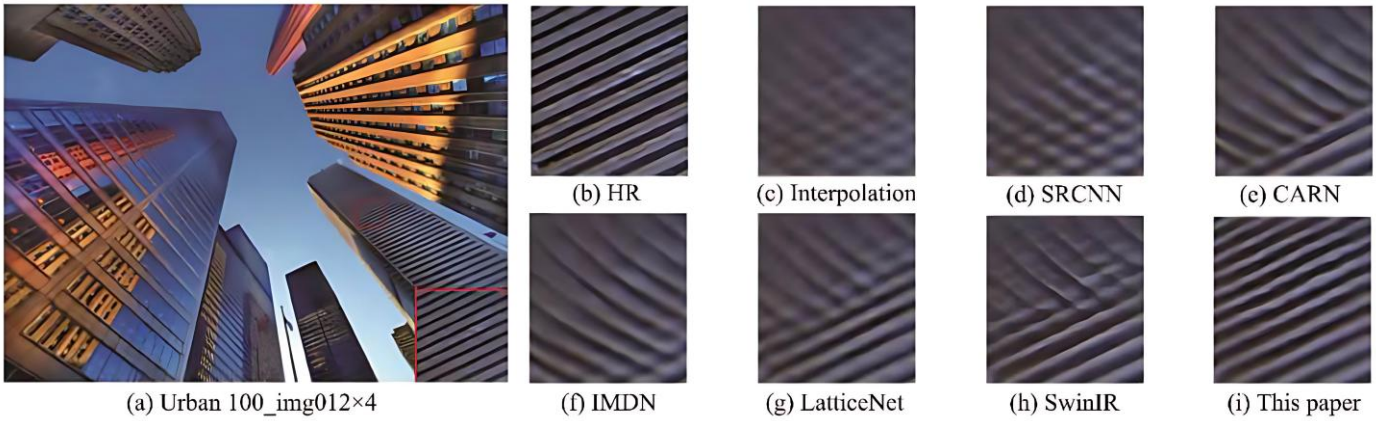

Fig. 6. Comparison of reconstructed HR images of img012 in Urban100 by different SR algorithms at the scale factor ×4 [(a) Urban100_img012×4; (b) HR; (c) bicubic interpolation; (d) SRCNN; (e) CARN; (f) IMDN; (g) LatticeNet; (h) SwinIR; i) ours].

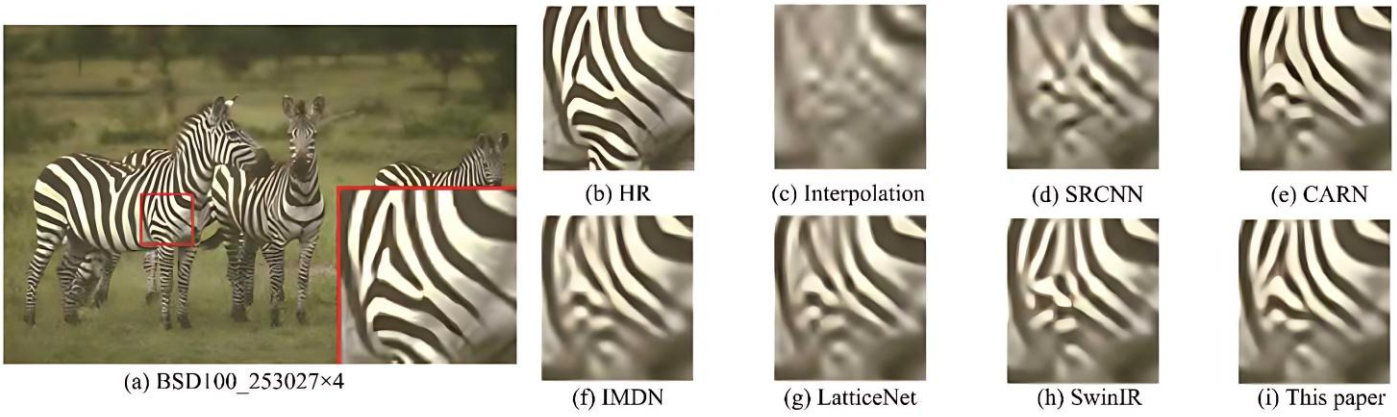

Fig. 7. Comparison of reconstructed HR images of 253027 in BSD100 by different SR algorithms with the scale factor ×4 [(a) BSD100_253027×4; (b) HR; (c) bicubic interpolation; (d) SRCNN; (e) CARN; (f) IMDN; (g) LatticeNet; (h) SwinIR; (i) ours].

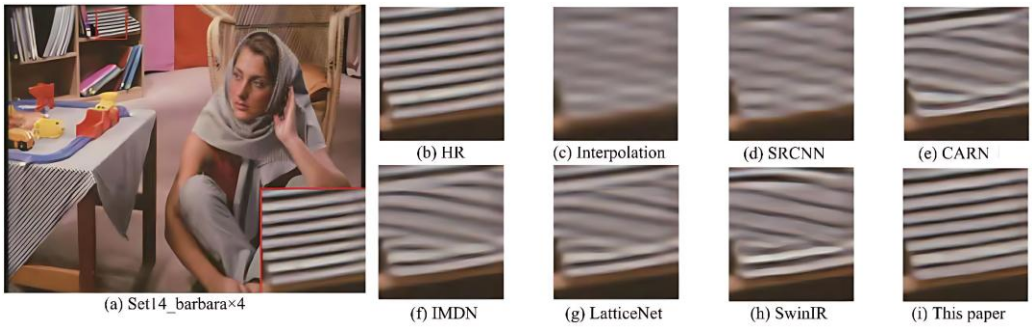

Fig. 8. Comparison of reconstructed HR images of barbara in Set14 by different SR algorithms with the scale factor ×4 [(a) Set14_barbara×4; (b) HR; (c) bicubic interpolation; (d) SRCNN; (e) CARN; (f) IMDN; (g) LatticeNet; (h) SwinIR; (i) ours].

D. Ablation Experiment

To effectively adjust features at different levels, this study adopts a channel attention-based discriminative enhancement strategy. Features from various levels are adaptively weighted using channel attention parameters. Unlike LatticeNet [24], which uses channel mean and standard deviation as modulation parameters, this approach applies channel attention weights directly to features from different levels. Table VI presents a comparison of reconstruction performance and model parameters using the two different weight generation strategies across multiple datasets with an upscaling factor of ×3. Experimental results show that, compared to the mean and standard deviation (MSD) based feature enhancement strategy proposed by Luo et al., the channel attention (CA) based discriminative enhancement method in this study enables better interaction and fusion of Transformer and convolutional features, achieving superior image reconstruction performance. To further investigate the roles of channel attention, linear weighting, and the feature aggregation module, ablation experiments were conducted to evaluate the impact of each component on overall performance. Specifically, the linear weighting structure was replaced with residual connections, and both channel attention and feature aggregation were removed to form the Base model. The models with different strategies are designated as Base A, B, C, and D.

TABLE VI. COMPARISON OF DIFFERENT WEIGHT GENERATION METHODS OF MODEL AT ×3 AMPLIFICATION FACTOR

Methodologies	Number of parameters/k	Manga109	Set5	Set14
CA	851	**34.31**	**34.65**	**30.61**
MSD	858	34.24	34.63	30.59

Note: Bold font indicates the best result in each column.

Table VII shows how these strategies affect reconstruction quality on the Manga109 dataset at a ×4 upscaling factor. The Base model performs slightly worse, achieving only 30.93 dB PSNR. Adding multi-level feature fusion in model A yields a clear 0.19 dB improvement, indicating that this fusion approach remains effective even in a Transformer-dominated network. Model B embeds channel attention in series on top of model A, while model C adds linear weighting on model A but removes learnable channel-attention parameters. Both models B and C achieve varying degrees of PSNR improvement, but combining both strategies in model D produces the most significant gain. Thus, the design strategies in this work are effective individually and even more so when combined, greatly enhancing the network's restoration capability. Compared with the Base model, model D boosts PSNR by 0.29 dB with only a 32 k increase in parameter count. For a more intuitive demonstration of each component's effect, representative images are visualized, with particular focus on how embedding channel attention improves the reconstruction of fine details. Fig. 9 shows the local magnification results of the drone aerial photo at a ×4 upscaling factor. It can be observed that, compared to model C, model D—which incorporates channel attention—can more accurately restore image details. This demonstrates that the integration of channel attention has a positive impact on image reconstruction performance. Experiments show that the model effectively improves the super-resolution performance while maintaining low complexity, and can provide a high-precision and low-latency solution for fast recovery of compressed images in the power industry. The method proposed in this study demonstrates significant advantages in practical applications. Firstly, the network architecture achieves high-fidelity restoration of image details while maintaining a low parameter count and computational complexity, resulting in a good balance between performance and efficiency. Secondly, the model exhibits stronger texture modeling and detail restoration capabilities in urban architectural images (Urban100) and comic images (Manga109), making it suitable for high-resolution demand scenarios such as video surveillance, remote sensing imagery, and anime image enhancement. Additionally, the model can run efficiently on consumer-grade GPUs with a single card, showing good deployment adaptability, which is ideal for resource-constrained edge devices or real-time image enhancement systems.

TABLE VII. COMPARISON OF THE EFFECTS OF DIFFERENT STRATEGIES ON MODEL RECONSTRUCTION PERFORMANCE ON MANGA109 AT AMPLIFICATION FACTOR ×4

Methodologies	Feature aggregation	Channel attention	Linear weighting	PSNR/dB	Number of parameters/k
Base	×	×	×	30.93	**839**
A	√	×	×	31.08	850
B	√	√	×	31.09	853
C	√	×	√	31.14	864
D (This study)	√	√	√	**31.22**	871

Note: Bold font indicates the best result in each column. "√" indicates adoption, while "×" indicates not adopted.

(a) Drone Aerial Photo (b) HR (c) Model C (d) Model D (This Paper)

Fig. 9. Comparison of the effects of different strategies on the reconstruction of ARMS in drone aerial maps at a scale of 4 [(a) Drone Aerial Photo; (b) HR; (c) model C; (d) model D (ours)].

V. CONCLUSION

In power-system unstructured-data management, images captured for patrol monitoring, equipment status awareness, and fault diagnosis are key information carriers. Because of limited storage and transmission bandwidth, these images are often heavily compressed, which degrades quality and removes fine details, ultimately impairing downstream intelligent recognition and O&M decision-making. To address this, we introduce a channel-attention-embedded Transformer super-resolution method (CAET) that balances reconstruction accuracy and model complexity for compressed-image restoration.

CAET adaptively embeds a channel-attention mechanism between Transformer global-context features and convolutional local-perception features, enabling interactive fusion. This design leverages the complementary strengths of Transformer and CNN feature extraction while dynamically enhancing critical information, markedly improving restoration quality. Across five public datasets and in comparison with six representative SR algorithms, CAET achieves best- or second-best results at all upscaling factors. Under 4× enlargement, it boosts PSNR by 0.09 dB on Urban100 and by 0.30 dB on Manga109 relative to the advanced SwinIR method, with noticeably better visual quality. Although CAET delivers strong, lightweight performance and restoration accuracy, all current experiments are conducted using bicubic-downsampled degradation. Real-world power-system images suffer from additional compression artifacts, transmission losses, and noise, so models trained on synthetic degradation still require improved generalization.

Future work will focus on enhancing the practical applicability of compressed-image restoration in power scenarios—particularly on lightweight, robust blind SR networks that can cope with multiple unknown degradations—so that image restoration technology can better support intelligent management and utilization of unstructured data in the power industry.

Although this method has demonstrated good performance on multiple public datasets, there are still two main limitations: first, the training and testing of this study are based on an idealized bicubic downsampling degradation model, which does not cover the complex degradation scenarios commonly found in real images, such as compression artifacts and noise pollution. As a result, the model may face performance degradation in practical applications. Second, there is still some computational redundancy in the integration of channel attention and Transformer features in CAET. Although it already possesses strong lightweight capabilities compared to similar methods, further optimization is needed for deployment on extremely resource-constrained devices (such as embedded systems).

REFERENCES

[1] Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution : dataset and study. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, IEEE: 2017, 1122-1131.

[2] Ahn N, Kang B, Sohn K A. Fast, accurate, and lightweight super-resolution with cascading residual network. Proceedings of the 15th European Conference on Computer Vision. Munich, 2018, 256-272.

[3] HOUQ,ZHOUD,FENGJ.Coordinate attention for efficient mobile network design. CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN: IEEE, 2021: 13708-13717.

[4] Chen H T, Wang Y H, Guo T Y, Xu C, Deng Y P, Liu Z H, Ma S W, Xu C J, Xu C, Gao W. Pre-trained image processing Transformer. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021, 12294-12305.

[5] LIU Na. Semantic segmentation algorithm of transmission lines based on CBAM attention me-chanism. Yangtze Information Communication, 2023, 36(09):60-62.

[6] LI Cuiming, WANG Hua, XU Long'er, et al. Road recognition method of photovoltaic plant based on improved DeepLabv3+. Journal of Shanghai Jiao Tong University, 2024, 58(5):776-782.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. An image is worth 16 × 16 words: Transformers for image recognition at scale. 2021.

[8] ZHOU Peng, XIONG Kai, XING Yan. Mars terrain segmentation algorithm based on improved DeepLab-v3+. Space Control Technology andApplications, 2023, 49(02):10-19.

[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 7132-7141.

[10] ZHANG Z, QIAN Z, ZHONG T, et al. Vectorized rooftop area data for 90 cities in China. Scientific Data, 2022, 9(1):66.

[11] Hui Z, Gao X B, Yang Y C, Wang X M. Lightweight image super-resolution with information multi-distillation network. Proceedings of the 27th ACM International Conference on Multimedia. 2019, 2024-2032.

[12] Jiang M J, Qian W H, Xu D, Wu H, Liu C Y. Gradual model reconstruction of Dongba painting based on residual dense structure. Journal of Image and Graphics, 2022, 27(4): 1084-1096.

[13] WANGHaifeng, XU Yilin, XU Dayi, et al. Evaluation of distributed rooftop photovoltaic hosting capacity in distribution networks based on high definition satellite map images. Guangdong Electric Power, 2023, 36(10): 105-113.

[14] Liu H, Zhang Z Z, Song R M, Shu Z Q, Wang J X, Tian H Y, Song Y X, Chen W G. Pattern Recognition Method for Detecting Partial Discharge in Oil-paper Insulation Equipment using Optical F-P Sensor Array based

on KAN-CNN Algorithm, Journal Lightwave Technology, 2025, 43(12) 6004 - 6012.

[15] MU Jingru, YU Kun, ZENG Xiangjun, TONG Haixin, et al. Leakage. fault detection in low-voltage station area with photovoltaic power supply considering multi-disturbance factors. Southern Power System Technology, 2024, 18(10): 130-141.

[16] Liu H, Zhang Z X, Tian H Y, Song Y X, Wang J X, Shu Z Q, Chen W G. Comparison of Different Coupling Types of Fiber Optic Fabry-Perot Ultrasonic Sensing for Detecting Partial Discharge Faults in Oil-Paper Insulated Equipment, IEEE Transactions on Instrumentation and Measurement. 2024, 73, 9519612.

[17] Lei P C, Liu C, Tang J G, Peng D L. Hierarchical feature fusion attention network for image super-resolution reconstruction. Journal of Image and Graphics, 2020, 25(9): 1773-1786.

[18] Lei S, Shi Z W, Mo W J. Transformer-based multistage enhancement for remote sensing image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5615611.

[19] Liang J Y, Cao J Z, Sun G L, Zhang K, Van Gool L, Timofte R. SwinIR: image restoration using Swin Transformer. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). 2021, 1833-1844.

[20] Liu H, Yang T H, Zhang Z X, Tian H Y, Song Y X, Sun Q X, Wang W, Geng YJ, Chen W G. Ultrasonic localization method based on Chan-WLS algorithm for detecting power transformer partial discharge faults by fibre optic F-P sensing array. High Voltage, 2024, 9(6),1234-1245.

[21] Liu J, Tang J, Wu G S. Residual feature distillation network for lightweight image super-resolution. Proceedings of 2020 European Conference on Computer Vision. 2020, 41-55.

[22] Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S, Guo B N. Swin Transformer: hierarchical vision Transformer using shifted windows. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021, 9992-10002.

[23] Lu Z S, Li J C, Liu H, Huang C Y, Zhang L L, Zeng T Y. Transformer for single image super-resolution. Proceedings of 2022 IEEE/ CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022, 456-465.

[24] Luo X T, Xie Y, Zhang Y L, Qu Y Y, Li C H, Fu Y. LatticeNet: towards lightweight image super-resolution with lattice block. Proceedings of the 16th European Conference on Computer Vision. 2020, 272-289.

[25] Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 2017, 76(20): 21811-21838.

[26] Wang Z D, Cun X D, Bao J M, Zhou W G, Liu G Z, Li H Q. Uformer: a general U-shaped Transformer for image restoration. Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 17683-17693.

[27] Wang Z H, Chen J, Hoi S C H. Deep learning for image super resolution: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3365-3387.

[28] Xiong C Y, Shi X D, Gao Z R, Wang G. Attention augmented multi-scale network for single image super-resolution. Applied Intelligence, 2021, 51(2): 935-951.

[29] Yu J H, Fan Y C, Yang J C, Xu N, Wang Z W, Wang X C, Huang T. Wide activation for efficient and accurate image super-resolution. 2018.

[30] Zhang X D, Zeng H, Zhang L. Edge-oriented convolution block for real-time super resolution on mobile devices. Proceedings of the 29th ACM International Conference on Multimedia. 2021, 4034-4043.

[31] Zhang Y L, Li K P, Li K, Wang L C, Zhong B N, Fu Y. Image super-resolution using very deep residual channel attention networks. Proceedings of the 15th European Conference on Computer Vision. 2018, 294-310.