

Article

Frequency Domain Mamba-Based Diffusion Model for Power Image Super-Resolution

Quan Quan ^{1,2}, Yunlong Sun ^{1,2}, Jian Xiao ^{1,2}, Kaixiang Xie ³, Jianwen Hu ³, Wen Wang ³, Zhuo Chen ^{1,2}
and Yuanjun Zuo ^{1,2,*}

¹ State Grid Hunan Electric Power Corporation Limited Research Institute, Changsha 410007, China; quanquan20b@163.com (Q.Q.); yl-sun@outlook.com (Y.S.); xiaojianhn@163.com (J.X.); chenzhuo19930401@163.com (Z.C.)

² Hunan Province Engineering Technology Research Center of Electric Power Multimodal Perception and Edge Intelligence, Changsha 410000, China

³ School of Artificial Intelligence, Changsha University of Science and Technology, Changsha 410114, China; xkx203818@163.com (K.X.); hujianwen1@163.com (J.H.); wew@csust.edu.cn (W.W.)

* Correspondence: hjzuoyuanjun@163.com

Abstract

To address the issues of low resolution and missing edge details in current power imagery, this paper proposes a diffusion-based super-resolution method. By applying the iterative denoising process of diffusion models, the proposed method learns feature distributions from training images, guiding the recovery of edge details. Unlike conventional diffusion methods, which directly use low-resolution images to guide noise towards high-resolution images, we propose a frequency-domain Mamba module. This module separates the image into high-frequency and low-frequency components through wavelet transform. The low-frequency components are processed using a Mamba network to extract features, while multi-scale convolution is employed to restore high-frequency details. The initial prediction generated by the frequency-domain Mamba module guides the diffusion model to restore high-resolution images. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art approaches on power images.

Keywords: power image; image super-resolution; diffusion models; frequency domain Mamba

1. Introduction

With the advancement of smart grid technology, real-time monitoring and fault diagnosis of electrical equipment can effectively enhance grid operational reliability. The comprehensive understanding of operational data, fault types, and equipment status across the entire grid has become a critical component in building the power Internet of Things [1,2]. Power scene imagery plays an indispensable role in monitoring the status of electrical equipment and detecting faults. Recent studies on intelligent fault diagnosis have shown that the quality of acquired data directly affects the reliability of equipment condition assessment and fault identification, underscoring the importance of obtaining high-quality power scene images [3,4]. However, existing image acquisition techniques still suffer from drawbacks such as low signal-to-noise ratios and insufficient high-frequency details. Due to the constraints in manufacturing costs and technological

Received: 12 May 2026

Revised: 12 June 2026

Accepted: 29 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

limitations, improving sensor resolution through hardware upgrades is challenging [5]. Therefore, software-based approaches to enhance image resolution are of great significance.

In recent years, deep learning has achieved breakthroughs across numerous fields such as image super-resolution (SR). Image super-resolution methods aim to reconstruct high-resolution (HR) images from low-resolution (LR) images, which is an ill-posed problem. In recent years, image super-resolution has attracted considerable attention. Dong et al. [6] first proposed the Super-Resolution Convolutional Neural Network (SRCNN), which constructs a three-layer Convolutional Neural Network (CNN) to learn an end-to-end mapping between LR images and their HR counterparts, achieving promising SR results. Subsequently, they introduced the Fast Super-Resolution Convolutional Neural Network (FSRCNN) [7], which employs post-upsampling and modifies feature dimensions to reduce the number of parameters and improve efficiency. Kim et al. [8] proposed the Very Deep Super-Resolution network (VDSR), which deepens the network via residual connections to expand the receptive field and extract richer features. Shi et al. [9] introduced the Efficient Sub-pixel CNN (ESPCN) using sub-pixel convolution instead of deconvolution for upsampling, thereby reducing computational cost.

Compared to traditional CNN architectures, the self-attention mechanism in Transformers captures global context, which is crucial for reconstructing clear images. Chen et al. [10] pioneered the application of standard Vision Transformers to low-level vision tasks such as image super-resolution. The approach involves pre-training the model on a large dataset followed by fine-tuning for specific tasks. Liang et al. [11] proposed SwinIR, a robust image restoration model based on Swin Transformers. By introducing local window attention mechanisms, they reduced computational complexity from quadratic to linear scale relative to image dimensions. Lu et al. [12] designed a lightweight hybrid Transformer architecture by combining shallow feature extracted by CNNs with deep feature extracted by Transformers, substantially reducing the number of parameters and computational burden. Compared to the quadratic complexity introduced by Transformers, the complexity of the Mamba model is linear, allowing it to handle long sequences data. The Mamba model retains global receptive field while being more efficient. Shi et al. [13] pioneered the systematic application of visual state space models to image restoration tasks. The model adopts an encoder–decoder architecture. The encoder comprises multiple stacked VMamba blocks for feature extraction, and the decoder is adapted to different image restoration tasks. Guo et al. [14] introduced a simple yet effective baseline termed MambaIR, which introduces both local enhancement and channel attention to improve the vanilla Mamba. Liu et al. [15] proposed a novel scanning strategy and designed a new Visual State Space Module (VSSM) for image deblurring.

Super-resolution models based on CNN, Transformers, and Mamba often produce overly smooth images with blurred edges and reduced perceptual quality. To overcome this limitation, Li et al. [16] introduced SrDiff, the first diffusion-based model for image super-resolution. SrDiff demonstrates the potential of diffusion techniques for image super-resolution. Saharia et al. [17] proposed SR3, which employs a Denoising Diffusion Probabilistic Models (DDPM) for image super-resolution and yields perceptually superior results. Shang et al. [18] present ResDiff, a novel Diffusion Probabilistic Model based on a residual structure for Single Image Super-Resolution (SISR). The combination of the CNN and Diffusion Probabilistic Model accelerated the generation process and yielded improved sample quality. Yue et al. [19] introduced ResShift, an efficient diffusion-based SR model that shifts residuals between HR and LR images, achieving high-quality restoration in only 15 sampling steps.

While existing super-resolution approaches based on convolutional neural networks, Transformer architectures, and Mamba models have demonstrated promising reconstruction performance, they often suffer from a loss of high-fidelity details in the reconstructed images. Current image super-resolution methods based on diffusion models use low-resolution images to guide noise into higher resolution, which limits the performance of image super-resolution. Furthermore, the lack of effective guidance may cause diffusion models to generate artifacts and hallucinate details, thereby compromising the fidelity and faithfulness of the reconstructed images. To address this drawback, we propose a frequency-domain Mamba module based on ResShift architecture. This module applies wavelet transforms to separate the high- and low-frequency components of the image. The Mamba module processes the low-frequency data to capture global information, while a multi-scale convolutional module restores fine details from the high-frequency information. After combining the high- and low-frequency information, a two-dimensional inverse wavelet transform is applied to generate an enhanced image. This enhanced image is then used as the input for the denoising network, guiding the model to produce the corresponding high-resolution image. This approach effectively restores both global features and local details, enabling diffusion models to recover more detailed fine-scale information and improve overall image quality. In summary, the main contributions of this work are given as follows:

1. By introducing a diffusion model for super-resolution reconstruction of power images, high-frequency details are generated through progressive denoising. This process learns the feature distribution from training power images to refine and enhance image quality.
2. To improve feature extraction, a wavelet transform module is used to separate high-frequency and low-frequency components. Different methods are applied to process different components, effectively integrating both local and global features from the frequency domain. This approach fully utilizes the frequency domain information, significantly enhancing the quality of super-resolution.
3. The Mamba module is used to process low-frequency information, with the extracted initial prediction guiding noise restoration to a high-resolution image, thus preventing the generation of textures not present in the original image.

The rest of this article is divided into four sections. In Section 2, the relevant work is described in detail. In Section 3, the proposed method is described in detail. Section 4 is devoted to experiments and results. In Section 5, a conclusion is given.

2. Theoretical Background and Related Work

In this section, the related work is presented in three parts: denoising diffusion probabilistic models, state space models, and wavelet-based approaches.

2.1. Denoising Diffusion Probabilistic Models

Ho et al. [20] proposed Denoising Diffusion Probabilistic Models (DDPM). Diffusion models are generative models that define a sequence of diffusion steps via Markov chains. These steps progressively add random noise to data, and the model learns to reverse this process to reconstruct data samples from noise. The key idea involves simulating a forward diffusion process from the true data distribution to noisy distribution, and a reverse diffusion process from noise back to data.

2.1.1. Forward Noise Addition Process

The forward noise addition process follows a Markov chain, which gradually adds Gaussian noise to high-resolution images by T iterations. The forward noise addition process can be defined as:

$$X_t = \sqrt{1 - \beta_t} * X_{t-1} + \sqrt{\beta_t} * Z_t \quad Z_t \sim N(0, I) \quad (1)$$

where X_t denotes the noisy image of the t -th iteration, t represents the time series with a range of values $[1, T]$, β_t is the forward process variance, and Z_t corresponds to the Gaussian noise.

The forward noise addition process can be defined as:

$$X_t = \sqrt{\alpha_t} * X_0 + \sqrt{1 - \alpha_t} * Z \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{k=1}^t \alpha_k$; this formula directly transforms the original image X_0 into the noisy one X_t in a single step, reducing the time of converting the original image into noise.

2.1.2. Reverse Inference Process

Inverse inference in diffusion models reverses the forward diffusion process by following the joint distribution of the Markov chain, reconstructing images from noise samples drawn from a standard Gaussian distribution. The noise sample at the previous step is predicted by the following Gaussian distribution:

$$q(X_{t-1} | X_t) \sim N \left[\frac{1}{\sqrt{\alpha_t}} \left(X_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} * \tilde{Z}, \frac{1 - \bar{\alpha}_{t-1}}{1 - \alpha_t} * \beta_t \right) \right] \quad (3)$$

Therefore, the inverse denoising process satisfies the following formula:

$$X_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(X_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} * \tilde{Z} \right) + \sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \alpha_t}} * \beta_t * Z \quad (4)$$

where $\tilde{Z} = UNet(X_t, t)$, $Z \sim N(0, I)$. This formula enables the gradual restoration of a clear image from a noisy one.

To accelerate the inference process, several efficient diffusion-based super-resolution frameworks have recently been proposed. For instance, a simple yet effective one-step SR generation method, termed SinSR [21], has been introduced. This approach is capable of generating super-resolved images within a single sampling step, thereby significantly reducing inference time. However, such acceleration is often accompanied by a certain degree of performance degradation. Wang et al. proposed SAMDiffSR [22], which exploits the fine-grained structural information provided by the Spectral Angle Mapper (SAM) during the noise sampling process to enhance image quality. Notably, this strategy improves reconstruction performance without introducing additional computational overhead during inference. In addition, Huang et al. developed a wavelet-based diffusion model, referred to as WaveDM [23]. Instead of directly modeling image distributions in the spatial domain, WaveDM learns the distribution of clean images in the wavelet domain based on the wavelet spectra of degraded images obtained through wavelet decomposition. By transferring the diffusion process to the wavelet domain, WaveDM substantially reduces the computational cost of each sampling step, resulting in more efficient super-resolution reconstruction.

2.2. State Space Models

State Space Models [24] provide a mathematical framework designed to effectively model temporal data. Mamba [25], an SSM-based model, has linear computational complexity and has recently garnered significant attention. Its core concept involves employing a selective mechanism to perform context-dependent reasoning. The selective state space model dynamically adjusts parameters based on input, enabling the model to selectively transmit or forget information according to current data. This facilitates better handling of discrete and information-dense data and can be expressed as a linear ordinary differential equation:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) \end{aligned} \quad (5)$$

where $h(t)$ is the hidden state vector at time t , and A , B and C are weight parameters.

The above equations can be discretized using a zero-order holder, transforming the continuous parameters A and B into discrete parameters $\bar{A}$ and $\bar{B}$, which may be defined as:

$$\begin{aligned} \bar{A} &= \exp(\Delta A) \\ \bar{B} &= (\Delta A)^{-1} (\exp(\Delta A) - I) \cdot \Delta B \end{aligned} \quad (6)$$

The discretized form of the equation can be expressed as:

$$\begin{aligned} h_k &= \bar{A}h_{k-1} + \bar{B}x_k \\ y_k &= Ch_k \end{aligned} \quad (7)$$

Moreover, the computation of selective state-space models can be extended to the following convolutional form:

$$\begin{aligned} \bar{K} &= (C\bar{B}, C\bar{A}\bar{B}, \dots, C\bar{A}^{L-1}\bar{B}) \\ y &= x * \bar{K} \end{aligned} \quad (8)$$

where L denotes the length of the input sequence, $*$ represents the convolution operation, and $\bar{K}$ indicates the convolution kernel.

In light of Mamba, Liu et al. developed SRMamba-T [26], a hybrid framework that combines Mamba and Transformer architectures to balance computational efficiency and reconstruction capability. Subsequently, Huang et al. proposed IRSRMamba [27], a novel framework that integrates Mamba with wavelet transforms, where customized structure and texture preserving mechanisms are employed to mitigate the spatial inconsistency introduced by patch-wise processing. Moreover, Li et al. presented HSRMamba [28], which preserves the computational advantages of VisualMamba while incorporating a stripe-based scanning scheme and wavelet decomposition. Such a design effectively alleviates the pattern conflict between high-frequency spatial information and low-frequency spectral information, leading to improved super-resolution performance.

2.3. Wavelet-Based Approaches

Wavelet decomposition [29] is a classical method widely employed across diverse computer vision tasks, capable of separating low-frequency information and high-frequency details from the original image. It has extensive application in fields such as image compression [30]. The Haar wavelet [31] is widely adopted in practical applications. It involves two types of operations: the Discrete Wavelet Transform (DWT) and the Discrete Inverse Wavelet Transform (IWT). The low-pass and high-pass filters are set as $L = 1/\sqrt{2} \begin{bmatrix} 1 & 1 \end{bmatrix}$ and $H = 1/\sqrt{2} \begin{bmatrix} -1 & 1 \end{bmatrix}$, respectively. Through the Haar wavelet transform, an input image of size $X \in H \times W$ can be decomposed into four subbands: X_{ll} , X_{lh} , X_{hl} and X_{hh} , each inherently possessing a size of $H/2 \times W/2$. As these filters are pairwise orthogonal, they form an invertible matrix of size 4×4 , enabling the precise

reconstruction of the original signal from its frequency components via inverse wavelet transform (IWT).

Wavelets offer a novel perspective for image super-resolution, providing significant advantages. This is because the spatial size of the image can be reduced to one-fourth of its original dimensions; this facilitates separate attention to high-frequency and low-frequency information, allowing for distinct processing approaches [32]. A novel recursive structure-preserving mechanism has been proposed, which employs multi-scale wavelet transform (WT) as an image prior and recursively processes low-frequency and high-frequency subnetworks at each decomposition level [33].

The Haar wavelet [33] is widely adopted in practical applications. It involves two types of operations: the Discrete Wavelet Transform (DWT) and the Discrete Inverse Wavelet Transform (IWT). The low-pass and high-pass filters are set as $L = 1/\sqrt{2} \begin{bmatrix} 1 & 1 \end{bmatrix}$ and $H = 1/\sqrt{2} \begin{bmatrix} -1 & 1 \end{bmatrix}$, respectively. Through the Haar wavelet transform, an input image of size $X \in H \times W$ can be decomposed into four subbands: X_{ll} , X_{lh} , X_{hl} and X_{hh} , each inherently possessing a size of $H/2 \times W/2$. As these filters are pairwise orthogonal, they form an invertible matrix of size 4×4 , enabling the precise reconstruction of the original signal from its frequency components via inverse wavelet transform (IWT).

3. Proposed Method

In this section, the proposed network is presented in three parts: the overall network, Mamba module, and training and sampling of the proposed method.

3.1. Overall Network

The proposed diffusion model processes low-resolution power images by three branches illustrated in Figure 1. Three branches perform noise addition, image enhancement, and residual connection on the input image. The first branch constitutes the forward diffusion and noise addition process. The diffusion model progressively incorporates Gaussian white noise into the input power scene image until the input image is entirely transformed into Gaussian white noise. The second branch incorporates a frequency-domain Mamba module comprising a two-dimensional wavelet transform module, the

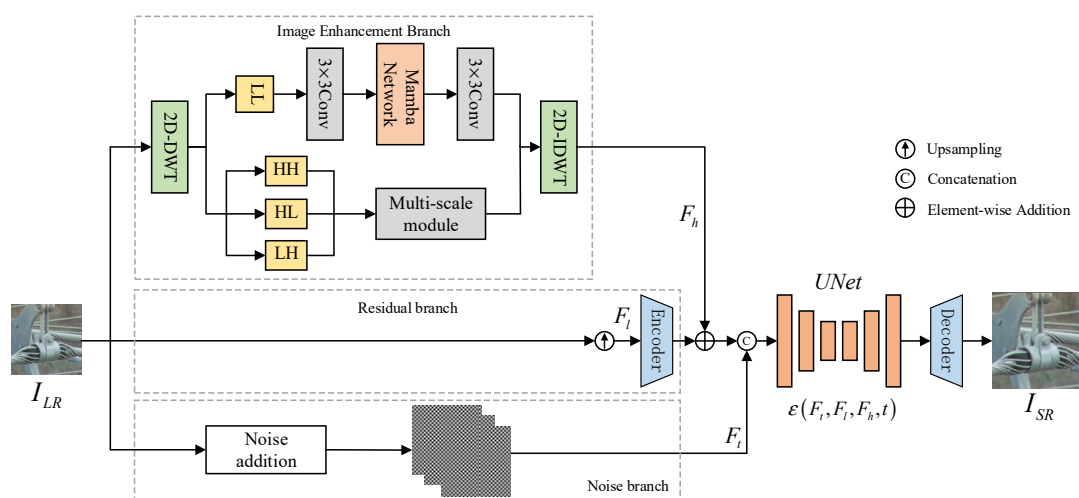


Figure 1. Overall network structure of the proposed method.

Mamba module, a multi-scale convolution module, and a two-dimensional inverse wavelet transform module. The third branch incorporates low-resolution images via

residual connections, employing bicubic interpolation to enlarge input image. The diffusion-step embedding is embedded within the noisy images. The low-resolution input is first upscaled using bicubic interpolation. This preliminary image is then processed by a VQGAN encoder [34], which reduces its spatial dimensions by a factor of four to extract compact feature representations. These features are combined with a noise signal and an initial prediction, forming the input for the subsequent U-Net denoising network [35]. Through the reverse inference process of the diffusion model, image noise is progressively removed. After decoding, the super-resolved image is obtained.

Let $I_{LR} \in \mathbb{R}^{H \times W \times 3}$ and $I_{SR} \in \mathbb{R}^{H' \times W' \times 3}$ denote the input image and the output image, respectively, where H and W are the height and width of input image, and H' and W' are the height and width of output image. We first perform a two-dimensional wavelet transform on the I_{LR} .

$$F_{LL, LH, HL, HH} = f_{2D-DWT}(I_{LR}) \quad (9)$$

Specifically, F_{LL} preserves the overall structure of the image, while $F_{LH, HL, HH}$ capture variations in edges and textures across three spatial directions. Next, a 2-D convolution with the kernel size of 3×3 is used to initially extract features from F_{LL} , $F'_{LL} = GELU[f_{3 \times 3 \text{ conv}}(F_{LL})]$. Subsequently, deep features are extracted using the Mamba module and processed through convolution with a kernel size of 3×3 , $F''_{LL} = f_{3 \times 3 \text{ conv}}[f_{MB}(F'_{LL})]$. High-frequency information is further extracted by a multi-scale convolutional module, $F'_{LH, HL, HH} = f_{1,2,3 \text{ conv}}(F_{LH, HL, HH})$. After merging the processed low-frequency and high-frequency features, a two-dimensional wavelet inverse transform is performed to obtain an enhanced image.

$$F_h = f_{2D-IWT}(F''_{LL}, F'_{LH, HL, HH}) \quad (10)$$

The upsampled low-resolution images are fed into the encoder, $F_i = \text{encoder}[\text{upsample}(I_{LR})]$. The low-resolution image I_{LR} is transformed into a noisy image F_i via a Markov chain. The enhanced image F_h is then added to the encoded image F_i and concatenated with the noisy image F_i . The diffusion-step embedding t is incorporated, and the result is fed into the U-Net network. After denoising via a U-Net network, the decoder is subsequently processed to obtain the super-resolved image I_{SR} .

$$I_{SR} = \text{Decoder}[\mathcal{E}(F_h, F_i, F_i, t)] \quad (11)$$

where $\mathcal{E}$ denotes the denoising process within the U-Net network.

3.2. Mamba Module

The Mamba module consists of layer normalization, the Mamba unit, a feedforward network (FFN), and residual connections, as shown in Figure 2. Layer normalization is applied to normalize the input features across the feature dimension, helping stabilize the training of intermediate layers. The Mamba unit effectively captures dependencies between distant pixels in the image, scanning and restoring the low-frequency information. The FFN, which includes two convolutions with the kernel size of 1×1 and activation functions, integrates and refines the spatial features extracted by the Mamba module, enabling feature transformation and dimensionality recovery. Residual connections add the input of a layer to its output to form the final output. The residual connection supports better data flow, eases network training, and enhances performance in deep architectures by mitigating issues like vanishing and exploding gradients.

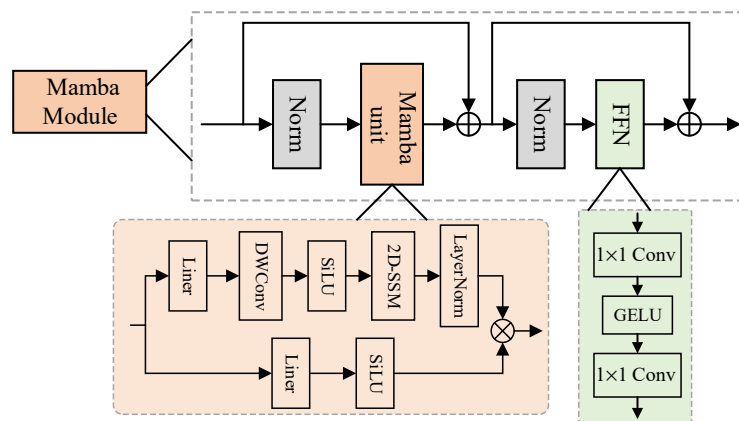


Figure 2. Architecture of the proposed Mamba module.

Let F'_{LL} denote the input features of the Mamba module. These features are fed into the normalization layer and the Mamba unit, where initial features are added via skip connections, $F_1 = f_{Mamba} [f_{Norm} (F'_{LL})] + F'_{LL}$. The previous output is fed into another normalization layer and a feedforward network, which includes two convolution layers with a kernel size of 1×1 and GELU activation. The skip connections are used to prevent feature loss, $F_2 = f_{FFN} [f_{Norm} (F_1)] + F_1$. The final output features are expressed as follows:

$$F''_{LL} = f_{FFN} [f_{Norm} (f_{Mamba} (f_{Norm} (F'_{LL})) + F'_{LL})] + f_{Mamba} (f_{Norm} (F'_{LL})) + F'_{LL} \quad (12)$$

3.3. Training and Sampling of the Proposed Method

During training, the network first converts low-resolution images into noisy versions F_t . The low-resolution images are then processed by the frequency-domain Mamba network to generate enhanced images F_h . The low-resolution images are also upsampled by bicubic interpolation to create enlarged versions F_l . These three versions—original, enhanced, and enlarged—are combined, along with a temporal step t , and encoded through the VQGAN encoder before being fed into the U-Net network. The denoising network U-Net is trained using the loss between the predicted output and the original image.

During sampling, the network selects a standard Gaussian noise tensor and the low-resolution image, encoded by the VQGAN encoder. The encoded low-resolution features and the initial prediction from the frequency-domain Mamba network are added. The denoising network U-Net is then applied, followed by decoding through the VQGAN encoder. This process restores the noisy image to a clear image within 15 sampling steps.

4. Experiments

In this section, extensive experiments are conducted to evaluate the performance of the proposed method. Next, the ablation experiments are constructed to fully demonstrate the effectiveness of the proposed method.

4.1. Datasets

The experiment uses 2000 1k-resolution power scene images, including critical power system components such as insulators, bolts, transmission towers, Transformers, and transmission lines. During training, these images are cropped to a size of 256×256 with a stride of 30, resulting in a training set of 56,000 images. These images are treated as high-resolution images and used as labels. The low-resolution images were obtained by

downsampling the corresponding high resolution images with a scaling factor of four using bicubic interpolation, yielding images of size 64×64 . The test set is constructed using the same approach, consisting of 1080 images. The proposed method utilizes pre-trained ResShift weights and integrates a frequency-domain Mamba network to perform $4\times$ super-resolution on the power scene images.

4.2. Evaluation Metrics and Implementation Details

The evaluation metrics include PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index) [36], CLIP-IQA (CLIP Image Quality Assessment) [37], and MUSIQ (Multi-Scale Image Quality) [38]. PSNR quantifies reconstruction fidelity by measuring the global pixel-wise error between the reconstructed and reference high-resolution images. SSIM evaluates perceptual similarity in terms of luminance, contrast, and structural information, making it more sensitive to local structural distortions. CLIP-IQA utilizes a pretrained CLIP model to assess perceptual quality through semantic alignment between image features and quality-related textual prompts, exhibiting strong sensitivity to texture realism and semantic plausibility. MUSIQ adopts a multi-scale Transformer to capture both global and local image characteristics, enabling robust quality assessment across different resolutions and achieving high consistency with human perceptual judgments.

In the proposed method, the multi-scale convolution module adopts a multi-branch parallel convolution architecture to extract features at different receptive fields using dilation rates of $\{1, 2, 3\}$. The hidden state dimension of the Mamba model is set to 32. The network is optimized using Adam [39] with the default setting, and the learning rate is set at 0.0005. The batch size is 32, and the number of iterations is 5×10^4 . The proposed method is implemented on the PyTorch 2.8.1 framework. The GPU employed in the experiments is an NVIDIA RTX5090D.

4.3. Super-Resolution Comparison Experiments

A $\times 4$ super-resolution experiment is conducted. The proposed method is compared with three diffusion-based approaches (SR3 [17], ResDiff [18], and ResShift [19]) and a Transformer-based method SwinIR [11]. The results of evaluation metrics are shown in Table 1, and the bold represents the best results.

Table 1. Evaluation metrics for super-resolution results in power image dataset.

Methods	PSNR	SSIM	CLIP-IQA	MUSIQ
SR3	31.002	0.9372	0.195	32.555
ResDiff	33.643	0.9171	0.248	25.816
ResShift	36.369	0.9751	0.351	48.100
SwinIR	36.496	0.9763	0.326	48.230
Ours	36.574	0.9769	0.360	48.412

Table 1 indicates that the proposed method outperforms other methods. Specifically, our method achieves the following results: a PSNR of 36.574 dB, an SSIM of 0.9769, a CLIP-IQA of 0.360, and a MUSIQ of 48.412. The first two metrics indicate that the reconstructed images produced by the proposed method exhibit higher similarity to the ground-truth images, while the latter two metrics demonstrate its superior perceptual quality.

The super-resolved images for cable and bolt samples are shown in Figure 3. Local regions are magnified for clearer observation and comparison in Figure 4. The proposed method achieves the most effective restoration of detailed information for components such as cables and bolts.

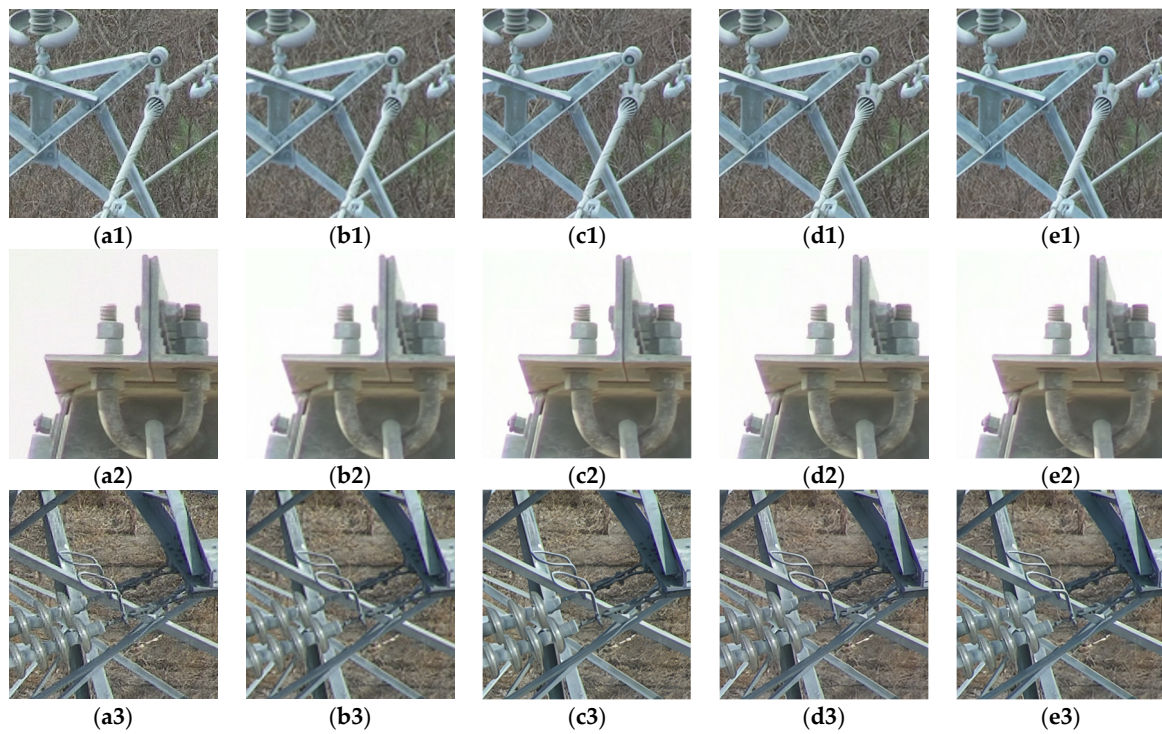
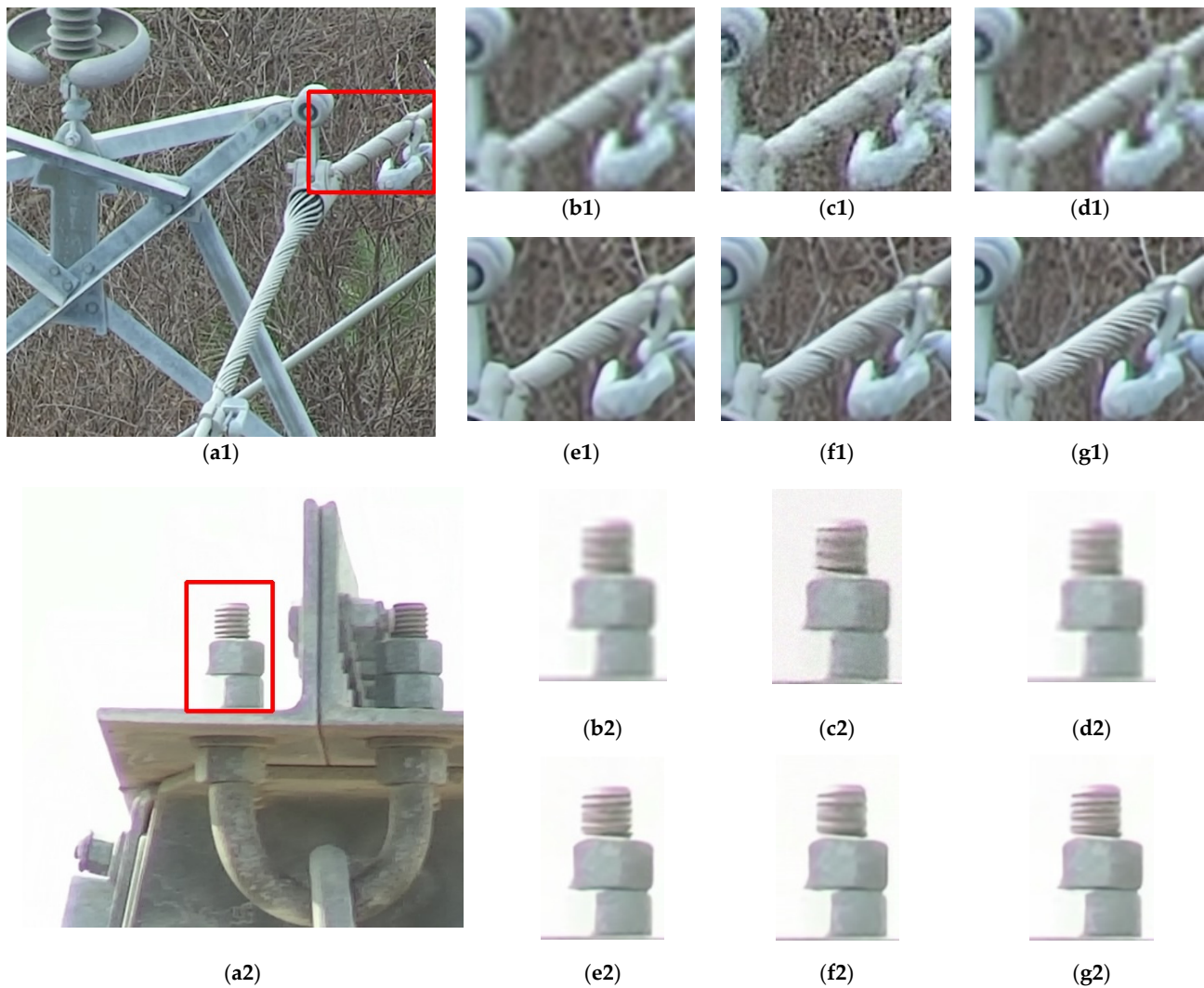


Figure 3. Architecture SR results in the cable, bolt, and insulator (scale factor 4). (a1–a3) SR3. (b1–b3) ResDiff. (c1–c3) ResShift. (d1–d3) SwinIR. (e1–e3) Ours.



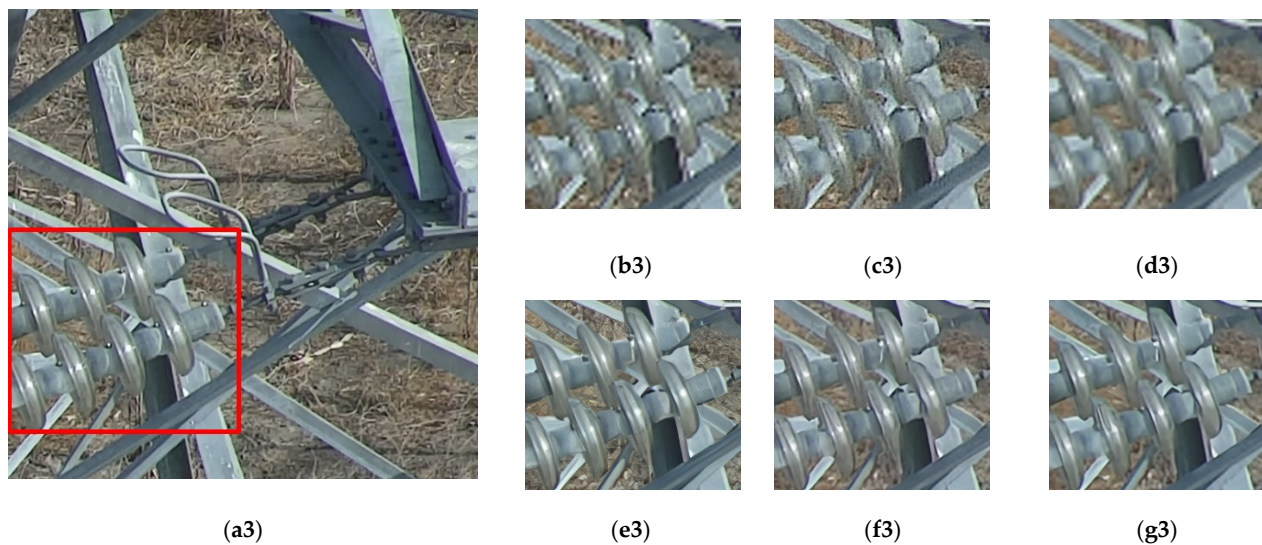


Figure 4. Magnified image of SR results in the cable, bolt and insulator dataset. (a1–a3) Ground truth. (b1–b3) Bicubic. (c1–c3) SR3. (d1–d3) ResDiff. (e1–e3) ResShift. (f1–f3) SwinIR. (g1–g3) Ours.

4.4. Ablation Study

To evaluate the effectiveness of the frequency domain separation and Mamba feature extraction modules in the proposed diffusion super-resolution network, an ablation study is conducted. The pre-trained weights from ResShift are used as the baseline. Three approaches are tested: fine-tuning the pre-trained ResShift weights with the training set, combining the Mamba frequency domain network with noise images via channel-wise concatenation, and combining the Mamba frequency domain network with noise images via additive fusion. The evaluation metrics are shown in Table 2.

Table 2. Objective evaluation metrics for ablation experiments.

Method	PSNR	SSIM
Pre-trained	30.651	0.9379
Pre-trained + Fine-tuning	36.369	0.9751
Mamba + Concat	36.034	0.9730
Mamba + Add	36.574	0.9769

As shown in Table 2, the method of integrating the Mamba frequency-domain network with the noisy image through direct addition effectively enhances model performance in super-resolution. The second row shows that training on pre-trained ResShift weights increased the PSNR and SSIM values by 5.7 dB and 0.0372, respectively, reducing the gap between super-resolved and original images. The third row demonstrates that combining the Mamba frequency domain network with noise images via channel concatenation yields inferior results compared to pre-trained ResShift weights. Compared with the fine-tuned model, lower PSNR and SSIM values are observed at this stage. The underlying reason is that feature concatenation changes the number of input channels, making the modified network architecture incompatible with a portion of the pretrained parameters and thereby reducing the benefits of pretrained initialization. The fourth row shows that combining the Mamba frequency domain network with the noisy image via direct addition improves PSNR and SSIM values by 0.54 dB and 0.0039, respectively, indicating better recovery of global and local features. Training with pre-trained weights and incorporating the Mamba frequency domain module result in PSNR and SSIM values 5.923 dB and 0.039 higher than direct testing on pre-trained weights.

These ablation experiments validate the effectiveness of the Mamba frequency domain module for image super-resolution. Direct addition is chosen to combine the Mamba module with the noisy image.

To validate the effectiveness of the Mamba module, an ablation study was conducted by removing the Mamba module while keeping all other components of the network unchanged. The quantitative results are summarized in Table 3. As shown in Table 3, removing the Mamba module results in noticeable performance degradation, with the PSNR decreasing from 36.574 dB to 36.159 dB and the SSIM dropping from 0.9769 to 0.9743. These results demonstrate that the Mamba module plays a crucial role in the reconstruction process. The performance improvement can be attributed to the ability of Mamba to effectively model long-range dependencies and capture global contextual information, thereby facilitating more accurate feature representation and image reconstruction. Moreover, the incorporation of the Mamba module introduces only a marginal increase in model complexity. Specifically, the number of parameters increases by merely 0.001 M (from 1.178 M to 1.179 M). Although the FLOPs increase from 0.051 G to 0.118 G, the computational overhead remains moderate relative to the achieved performance gains. Therefore, the proposed Mamba module provides an effective balance between reconstruction accuracy and computational efficiency.

Table 3. Ablation study on the effectiveness of the Mamba module.

Method	PSNR	SSIM	Flops (G)	Param (M)
Mamba Removed	36.159	0.9743	0.051	1.178
Full Model	36.574	0.9769	0.118	1.179

5. Conclusions

This paper proposes a diffusion-based super-resolution method for power imagery, incorporating a frequency-domain Mamba approach. By integrating diffusion models into the super-resolution process, the method applies a stepwise denoising generation technique to learn feature distributions from power images. This enables the generation of high-resolution images with authentic textures and fine details that align with real-world patterns. To mitigate the risk of introducing artifacts or unrealistic content, we propose a frequency-domain Mamba module. This module generates an initial prediction that guides the diffusion model in reconstructing high-resolution images. It decomposes the image into high- and low-frequency components, applying distinct feature extraction strategies for each frequency domain to ensure effective fusion of local and global features. The effectiveness of the proposed method is validated on a test dataset, showing strong performance in both objective evaluation metrics and visual quality. However, the iterative sampling process of diffusion models incurs considerable computational overhead. In future work, we will explore more lightweight network architectures to improve computational efficiency while maintaining reconstruction performance.

Author Contributions: Conceptualization, Q.Q., Y.S., J.X., K.X., J.H. and Y.Z.; Methodology, Q.Q., Y.S., J.X., K.X., J.H. and Z.C.; Software, Q.Q.; Validation, Y.Z.; Formal analysis, Q.Q., J.X. and Z.C.; Investigation, Y.S., J.H., W.W. and Y.Z.; Data curation, K.X. and W.W.; Writing—original draft, Q.Q., Y.S. and Y.Z.; Writing—review & editing, Q.Q., Y.S., J.X., K.X., J.H., W.W., Z.C. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Project of State Grid Hunan Electric Power Company Limited, grant number 5216A5250009.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors Quan Quan, Yunlong Sun, Jian Xiao, Zhuo Chen and Yuanjun Zuo were employed by the State Grid Hunan Electric Power Corporation Limited Research Institute. The authors declare that this study received funding from Science and Technology Project of State Grid Hunan Electric Power Company Limited. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Xia, C.; Ren, M.; Wang, B.; Dong, M.; Xu, G.; Xie, J.; Zhang, C. Infrared thermography-based diagnostics on power equipment: State-of-the-art. *High Volt.* **2021**, *6*, 387–407.
2. Yang, L.; Fan, J.; Liu, Y.; Li, E.; Peng, J.; Liang, Z. A review on state-of-the-art power line inspection techniques. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9350–9365. <https://doi.org/10.1109/TIM.2020.3031194>.
3. Khan, A.; Al Farid, F.; Junaid, A.; Siddique, M.F.; Iqbal, A.; Siddique, M.S.; Uddin, J.; Karim, H.A.; Husnain, G. Early-warning industrial fault detection based on physics-guided residual learning and calibrated CRNNs. *Sci. Rep.* **2026**, *16*, 17488.
4. Siddique, M.F.; Ahmad, Z.; Kim, J.M. Pipeline leak diagnosis based on leak-augmented scalograms and deep learning. *Eng. Appl. Comput. Fluid Mech.* **2023**, *17*, 2225577.
5. Furse, C.M.; Kafal, M.; Razzaghi, R.; Shin, Y.-J. Fault diagnosis for electrical systems and power networks: A review. *IEEE Sens. J.* **2021**, *21*, 888–906. <https://doi.org/10.1109/JSEN.2020.2987321>.
6. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307.
7. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2016; pp. 391–407.
8. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
9. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
10. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
11. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, Montreal, QC, Canada, 10–17 October 2021; pp. 1833–1844.
12. Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; Zeng, T. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, New Orleans, LA, USA, 18–24 June 2022; pp. 456–465.
13. Shi, Y.; Xia, B.; Jin, X.; Wang, X.; Zhao, T.; Xia, X.; Xiao, X.; Yang, W. Vmambair: Visual state space model for image restoration. *arXiv* **2024**, arXiv:2403.11423.
14. Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; Xia, S.-T. Mambair: A simple baseline for image restoration with state-space model. *arXiv* **2024**, arXiv:2402.15648.
15. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
16. Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **2022**, *479*, 47–59.
17. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4713–4726.
18. Shang, S.; Shan, Z.; Liu, G.; Zhang, J. ResDiff: Combining CNN and diffusion model for image super-resolution. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 8975–8983.
19. Yue, Z.; Wang, J.; Loy, C.C. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *Proceedings of the International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 10–15 December 2024; pp. 13294–13307.

20. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2020; pp. 6840–6851.
21. Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A.C.; Wen, B. Sinsr: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 16–22 June 2024.
22. Wang, C.; Hao, Z.; Tang, Y.; Guo, J.; Yang, Y.; Han, K.; Wang, Y. Sam-diffsr: Structure-modulated diffusion model for image super-resolution. *arXiv* **2024**, arXiv:2402.17133.
23. Huang, Y.; Huang, J.; Liu, J.; Yan, M.; Dong, Y.; Lv, J.; Chen, C.; Chen, S.; Lyu, J. Wavedm: Wavelet-based diffusion models for image restoration. *IEEE Trans. Multimed.* **2024**, *26*, 7058–7073.
24. Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; Van Gool, L. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022; pp. 17481–17490.
25. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
26. Liu, C.; Zhang, D.; Lu, G.; Yin, W.; Wang, J.; Luo, G. Srmamba-t: Exploring the hybrid mamba-transformer network for single image super-resolution. *Neurocomputing* **2025**, *624*, 129488.
27. Huang, Y.; Miyazaki, T.; Liu, X.; Omachi, S. Irsrmamba: Infrared image super-resolution via mamba-based wavelet transform feature modulation model. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5005416.
28. LLi, B.; Wang, X.; Xu, H. HSRMamba: Efficient Wavelet Stripe State Space Model for Hyperspectral Image Super-Resolution. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2025.
29. Zhang, D. “Wavelet transform,” in *Fundamentals of Image Data Mining (Texts in Computer Science)*; Springer: Cham, Switzerland, 2019; pp. 35–44. https://doi.org/10.1007/978-3-030-17989-2_3.
30. Taubman, D.S.; Marcellin, M.W. *JPEG2000: Image Compression Fundamentals, Standards and Practice*; Springer: Berlin/Heidelberg, Germany, 2001.
31. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693.
32. Tan, J.; Pei, S.; Qin, W.; Fu, B.; Li, X.; Huang, L. Wavelet-based mamba with Fourier adjustment for low-light image enhancement. In *Proceedings of the Asian Conference on Computer Vision*, Hanoi, Vietnam, 8–12 December 2024; pp. 3449–3464.
33. Hsu, W.-Y.; Jian, P.-W. Wavelet pyramid recurrent structure-preserving attention network for single image super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 15772–15786.
34. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 20–25 June 2021; pp. 12868–12878.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the MICCAI*, Munich, Germany, 5–9 October 2015.
36. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612.
37. Wang, J.; Chan, K.C.; Loy, C.C. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, 7–14 February 2023.
38. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 10–17 October 2021; pp. 5128–5137.
39. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.