Research on Image Super-Resolution Algorithm Based on Improved GAN

Peng Gai
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 929113480@qq.com

Li Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 332099732@qq.com

***Abstract*—Image super-resolution reconstruction aims to transform low-resolution blurred images into high-resolution images under the same scene. Due to its practical value and theoretical significance, this technology is widely used in computer vision and image processing. In recent years, deep learning-based super-resolution algorithms have shown strong performance. However, most existing deep learning methods suffer from a common issue: when reconstructing images at large magnification factors, the results tend to be overly smooth and lacking in texture, resulting in unrealistic visual perception. Perceptual super-resolution methods based on generative adversarial networks (GANs) can effectively alleviate the oversmoothing problem, which has drawn significant research attention. Nevertheless, GAN-based approaches still have limitations, including single-scale reconstruction, insufficient acquisition of high-frequency information, and a tendency to generate excessive textures in smooth areas, leading to images with noticeable noise, artifacts, and insufficient texture details. To address these shortcomings, this paper proposes a novel network structure that integrates generative adversarial networks with multi-stage gated aggregation and multi-scale feature fusion mechanisms. The goal is to optimize the existing problems in current image super-resolution networks and improve the quality of image reconstruction.**

Keywords-Deep Learning; Image Super Resolution; Generative Adversarial Network; Attention Mechanism;

I. INTRODUCTION

Images are formed by display devices presenting a series of colored dots that constitute the visual content we perceive. Each colored dot is termed a pixel, which displays a specific color at a given moment. The combination of numerous pixels creates a complete image. Thus, a picture is represented by multiple pixels. The resolution of an image or video refers to the number of rows and columns of pixels it contains. It also represents the ability of the human eye to discern distinct points or lines within the visual content [1]. In recent years, intelligent devices such as robotics, instrumentation, and intelligent assisted driving systems have become increasingly prevalent in daily life. These devices fundamentally operate by using various sensors to acquire information from the physical world. The collected data is then converted into digital signals for analysis, enabling corresponding decision-making. Similar to humans, these systems rely on visual perception to obtain external information. While humans perceive visual information as vivid and intuitive, intelligent devices face challenges: although visual data is readily obtainable with current technology, its quality degrades during acquisition and transmission due to noise interference from factors like weather and lighting conditions. Consequently, enhancing the quality of visual information for intelligent devices has become a critical research focus for both technological advancement and societal development.

Current Limitations in Deep Learning-Based Image Super-Resolution work is as follow:
Most existing algorithms exhibit excessive smoothing and inadequate textural details during high-magnification upscaling, resulting in visually unrealistic reconstructions. While perceptually-driven GAN methods can partially mitigate this smoothing issue by generating richer textures to enhance subjective quality, fundamental challenges remain.

GAN-based super-resolution approaches exhibit inherent limitations: When performing high-ratio upscaling, reconstructed images frequently demonstrate excessive smoothing and insufficient

textural detail, resulting in perceptually unrealistic outputs. Paradoxically, while GANs excel at detail synthesis, they may introduce extraneous details or artifacts in naturally smooth regions, compromising final image quality. Furthermore, the generative stochasticity inherent in these methods causes output inconsistencies across reconstructions. This variability not only adversely affects training stability but also prolongs convergence time.

Current super-resolution models exhibit notable shortcomings in feature extraction: Particularly in global context modeling and long-range dependency capture, their inadequate performance constrains reconstruction fidelity. Concurrently, architectural inefficiencies in certain models lead to prolonged training duration, creating implementation bottlenecks.

Most current residual learning-based super-resolution models employ single-scale convolutional kernels, which fundamentally restricts effective multi-scale feature extraction from images.

GAN-based super-resolution methods demonstrate exceptional performance in perceptual quality enhancement, owing to their superior texture synthesis capabilities. As research progresses, novel network architectures integrating residual learning, dense connections, multi-scale feature extraction, and attention mechanisms are emerging as pivotal directions for advancing image super-resolution reconstruction quality. Addressing these limitations, this paper proposes a novel GAN-integrated network architecture incorporating multi-stage gate-controlled aggregation and multi-scale feature fusion mechanisms, designed to resolve existing issues in super-resolution networks and enhance image reconstruction fidelity.

II. RELATED WORK RESEARCH

Ledig et al. [2] pioneered the integration of generative adversarial networks (GANs) into image super-resolution (SR) and proposed the SRGAN model. In their framework, low-resolution images are fed into a generator network, which learns to produce high-resolution outputs. A discriminator network then tries to tell whether a given high-resolution image is from the original real dataset or generated by the generator. Once the discriminator can no longer reliably distinguish between real and fake images [3], the generator is considered to have learned to produce perceptually high-quality results. SRGAN enhances the visual realism of SR reconstructions by jointly optimizing perceptual loss and adversarial loss. A pre-trained CNN (e.g., VGG) serves as a feature extractor, comparing the generated and original high-resolution images in a deep feature space. This encourages the generated output to be not only visually similar to the original but also semantically close in high-level features. Moreover, the discriminator in this GAN setup is tasked with distinguishing real from synthetic images, while the generator attempts to fool it into believing its outputs are real. This adversarial training prompts the generator to learn more delicate and natural image details. The generator network of SRGAN is designed based on the SRResNet architecture containing multiple residual blocks. In the optimization process, in addition to the traditional mean square error, perceptual loss and adversarial loss are added to improve the quality and authenticity of the generated images, thus being able to generate higher-quality and more realistic high-resolution images, which is better than the traditional methods that only rely on pixel-level loss. The discriminator network of SRGAN contributes to producing super-resolution results that look visually realistic. However, these results do not achieve the highest PSNR [4].

ESRGAN [5] builds upon SRGAN to further enhance reconstruction quality. The main architectural modifications include replacing residual blocks with RRDB (Residual-in-Residual Dense Blocks) and removing batch normalization [6]. For adversarial loss, a relativistic GAN is employed, which focuses on relative realism rather than absolute authenticity. Regarding perceptual loss, features are extracted before the activation function. The training strategy involves two stages: first, the network is pre-trained to optimize PSNR; second, it is fine-tuned with the GAN objective [7].

A common shortcoming of most deep-learning-based image super-resolution algorithms is that, when a large upscaling factor is applied, the reconstructed images tend to be excessively smooth and deficient in textures,

leading to visually unnatural results. Although GAN-based methods perform well in generating detailed textures, they may introduce unnecessary details or artifacts in relatively smooth areas of the image, which affects the final image quality. In addition, due to the randomness of the generation process, the super-resolution images generated each time may vary, which not only affects the training effect but also increases the time required for training.

III. METHODS

This paper mainly focuses on improving the generator component within the generative adversarial network. The design objective of the entire reconstruction model is to restore the lost high-frequency detail information in images as much as possible, thereby enhancing the quality and efficiency of image reconstruction. Meanwhile, it performs differentiated processing on smooth regions and complex texture regions in images to avoid generating redundant pseudo-textures in smooth regions. This not only reduces the number of model parameters but also improves the learning speed of the network.

The generator adopts a multi-layer network design with a parallel structure. Among them, the first layer is improved on the basis of the traditional generative adversarial network, using a multi-stage gated aggregation module to replace the traditional convolution block, so as to effectively suppress unimportant noise features and more deeply learn the high-frequency information in the image. [8] This module stacks 8 MGAM units and introduces a second-layer feature extraction network. This layer adopts a multi-scale parallel structure to realize the fusion of features of different scales, thereby obtaining rich feature representations from different levels. The main function of the second-layer network is to deeply extract the structure and detail information of low-resolution images, capture features of multiple scales in the image through receptive fields of different sizes, and fuse this information with the main generation network, thereby significantly enhancing the expressive ability and learning effect of the entire network.

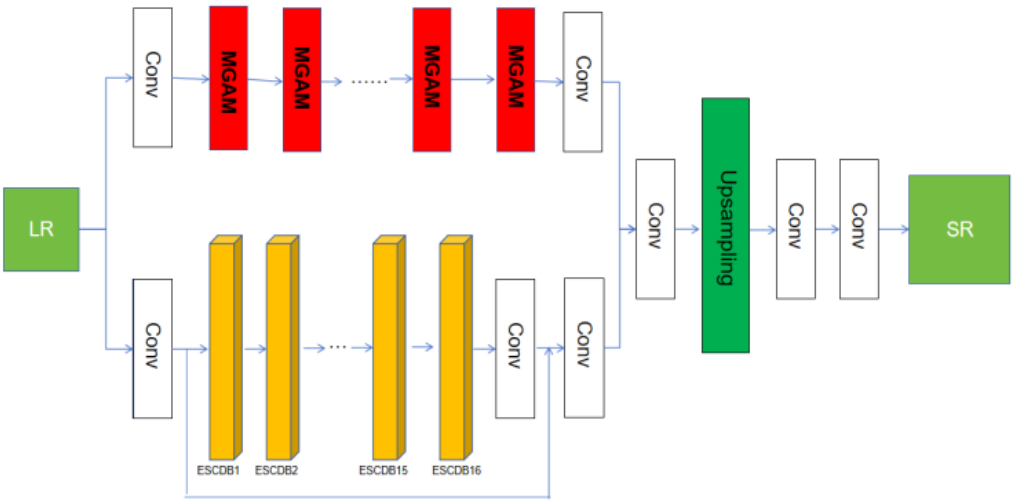

Figure 1. Generator Structural Diagram

A. MGAM Module

To effectively extract high-dimensional features of images, this paper selects MogaNet as the backbone network. [9] As shown in Figure 2, MogaNet adopts a hierarchical architecture, consisting of four stages in total. Each stage is composed of an Embedding Stem and multiple Moga Blocks (Multi-order Gated Aggregation Blocks). Each Moga Block includes a spatial aggregation module and a channel aggregation module.

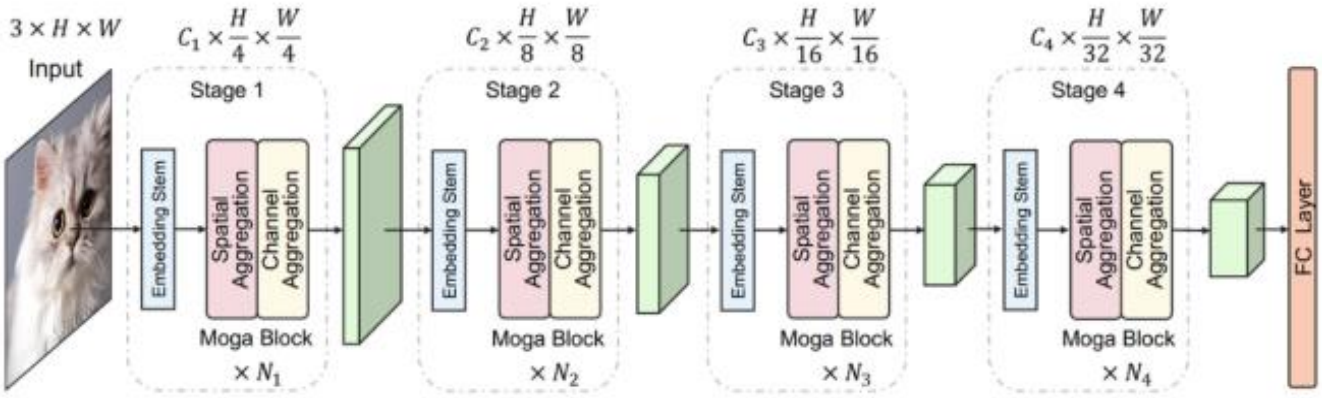

Figure 2. Efficient Multi-order Gated Aggregation Network

The embedding module in the first stage is used for initial feature extraction of the input image and reducing the resolution of the feature map. This module consists of two layers of 3x3 convolutions with a stride of 2 for the convolution operations. [10] The embedding modules in the subsequent three stages are used to further reduce the resolution of the feature map, each consisting of a single layer of 3x3 convolution with a stride of 2 as well. After the input image passes through the embedding modules of these four stages, its resolution will be reduced to 1/4, 1/8, 1/16, and 1/32 of the original in sequence.

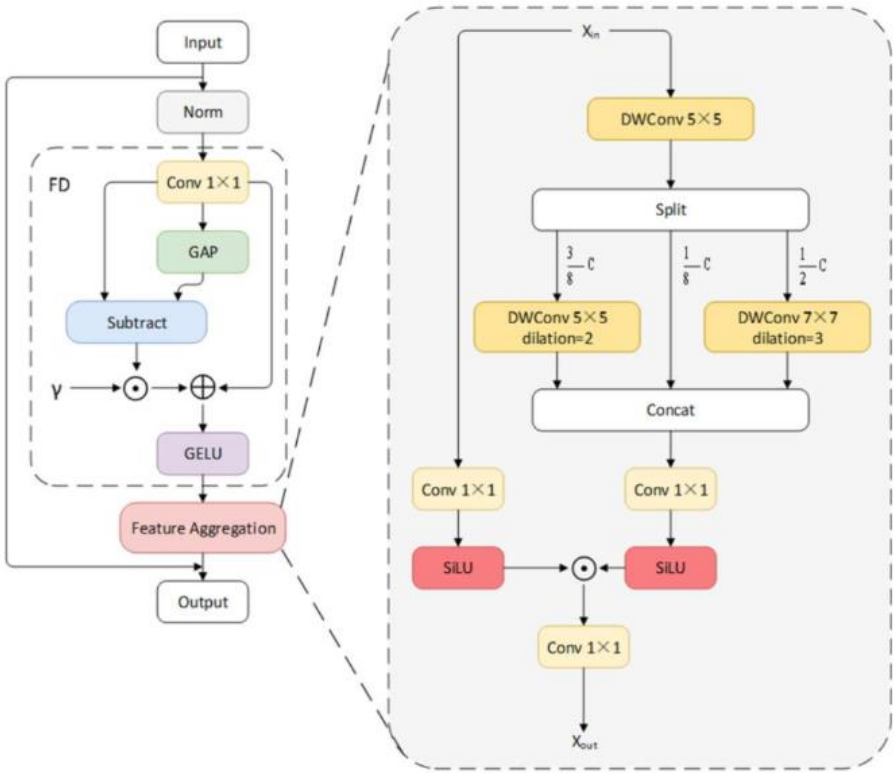

Figure 3. Spatial Aggregation

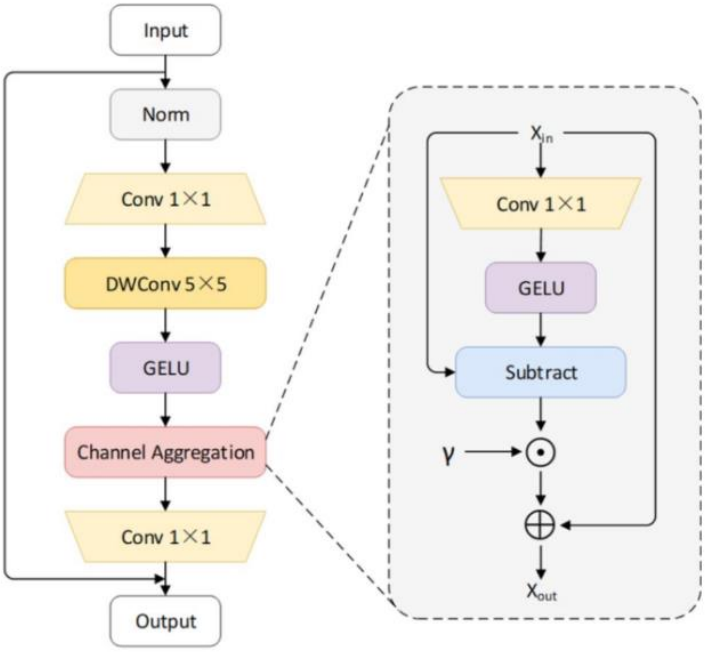

Figure 4. Channel Aggregation

As illustrated in Figure 5, the spatial aggregation module comprises two sub-modules: Feature Decomposition (FD) and Feature Aggregation (FA) [11]. The FD module captures both local and global information via a 1×1 convolution followed by global average pooling. This design enables the model to jointly consider local textures and global context. The FA module consists of a context branch and an aggregation branch. The context branch runs three depthwise convolutions in parallel to extract multi-scale contextual information. Specifically, the input feature X first passes through a 5×5 depthwise convolution. Its output is then split along the channel dimension into three parts, which retain 1/8, 3/8, and 1/2 of the original channels, respectively. The 3/8 portion is processed by a 5×5 depthwise convolution with dilation rate 2, while the 1/2 portion goes through a 7×7 depthwise convolution with dilation rate 3. The three processed parts are concatenated along the channel axis and fused by a 1×1 convolution. Meanwhile, the aggregation branch uses a 1×1 convolution to reorganize the features and enrich their representation. Finally, the outputs of the aggregation branch and the context branch are combined to produce the overall output of the feature aggregation module.

Similar to the feed-forward network design of Transformer, MogaNet also adopts a two-layer linear mapping with a certain channel expansion ratio for channel mixing, and its specific structure is shown in Figure 6. Among them, in order to aggregate channel information and enrich the diversity of features in the channel dimension, a Channel Aggregation module is added.

B. Cross-Scale Feature Fusion Module

The Cross-Scale Feature Fusion Module aims to extract multi-scale features of the input image, especially low-resolution structural information and details. [12] It captures visual information at different levels through receptive fields of different sizes and fuses these features into the original generative network, thereby significantly enhancing the learning ability and expressive power of the entire network. In the task of image super-resolution, this paper introduces the EMA (Efficient Multi-Scale Attention) mechanism, as shown in Figure 8. In Figure 9, c, h, and w represent the number of channels, height, and width of the input data respectively, and g represents the number of groups. The core idea of this mechanism is to incorporate multi-scale spatial information into the model, which can both capture local features and retain global information.

EMA realizes the extraction of information at different scales through parallel 1×1 and 3×3 convolution kernels. [13] When capturing local features, the 1×1 convolution kernel enhances the precision of feature representation through global average pooling and interactions between channels, while the 3×3 convolution kernel further expands the receptive field to capture long-range dependencies. This design enables EMA to flexibly integrate features of different spatial scales when processing images.

In the image super-resolution task, EMA processes feature channels through a grouping mechanism, effectively improving feature learning capability. High-frequency detail regions in images (such as edges and textures) usually contain more critical information, and EMA can more accurately capture and enhance these key features through attention weighting on these regions, thereby improving the clarity and detail restoration of image reconstruction. Meanwhile, by introducing the cross-spatial learning mechanism, EMA can effectively model the long-range dependencies of images in the spatial dimension (such as spanning large structures or continuous textures of the image), ensuring the spatial consistency and integrity of reconstructed features, and ultimately generating more accurate and natural super-resolution images.

To obtain multi-scale feature information from different network layers, 3×3 and 5×5 convolution kernels are used in two parallel sub-networks respectively to learn image features. [14] This ensures a relatively small number of parameters while maintaining a large receptive field. Finally, the obtained feature information is fused to obtain more abundant features. The parallel network mainly consists of convolutional layers and ESCDB (Efficient Structure Context Detail Block) modules. Figure 8 shows the network structure of ESCDB, where DCCL (Diverse Context Convolutional Layer) combines contextual information using three convolutional layers and has a large receptive field.

ESDCAB (Efficient Structure Detail Context Attention Block) is composed of one DCCL layer, one BN (Batch Normalization) layer, one PReLU layer, another DCCL layer, another BN layer, and one residual structure. DCCL consists of three parallel convolutional modules with different receptive field sizes (1×1, 3×3, and 5×5 respectively). Their outputs are concatenated together, and the weights of each module are calculated through the EMA attention mechanism module. Finally, a 1×1 convolution layer is applied to obtain the final output.

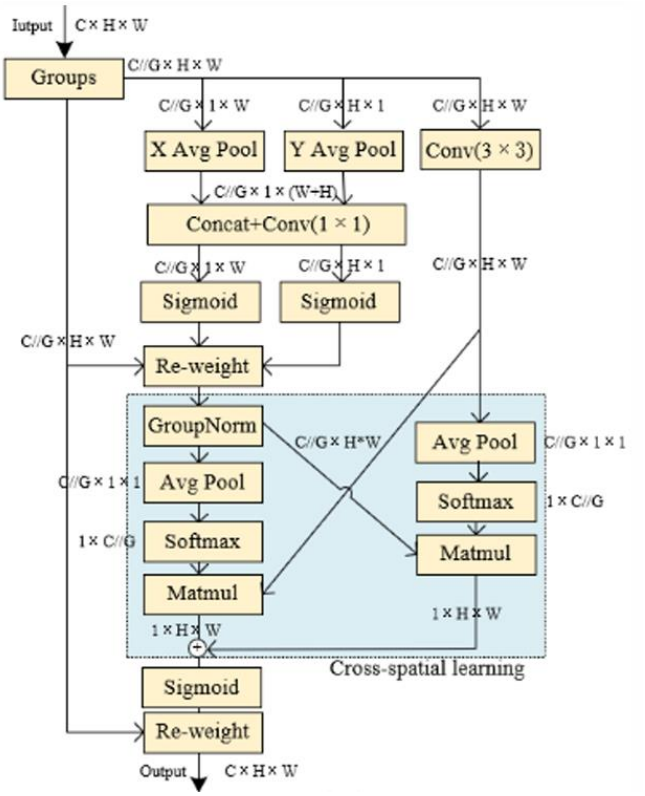

Figure 5. EMA Module Structural Diagram

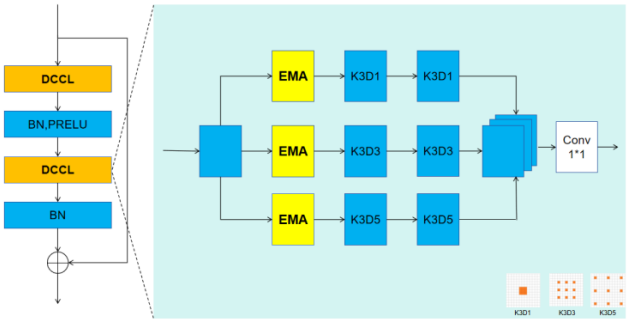

Figure 6. ESCDB Module Structural Diagram

IV. EXPERIMENTS

A. Experimental Environment

The experimental configuration employed in this paper is summarized in Table I. The system was built on a Windows 11 platform powered by an Intel

I7-13700KF CPU and an NVIDIA RTX 4090 GPU with 32GB of memory. The model was implemented using Python 3.9 and the PyTorch 2.5.1 framework.

TABLE I. EXPERIMENTAL ENVIRONMENT

Experimental environment	Configuration
CPU	Intel I7-13700KF
GPU	NVIDIA RTX 4090
Pytorch	Pytorch 2.5.1

B. Dataset

To validate the effectiveness of the proposed algorithm, the training process utilizes the DIV2K dataset, which is widely used in super-resolution reconstruction. The DIV2K dataset consists of 1000 high-quality (2K resolution) images for image restoration tasks. [15] These images are divided into three parts for training, validation, and testing. To ensure better model training and accuracy, the data for these three different tasks are allocated in an 8:1:1 ratio.

Due to the uneven distribution of samples in the training data, underfitting may occur during training. To address this issue, the experiment employs three methods for data augmentation on the images of the two datasets: (1) Scaling: Randomly downscale the images within the range of [0.5, 1.0]. (2) Rotation: Randomly rotate the images by 90°, 180°, or 270°. (3) Flipping: Horizontally or vertically flip the images. The augmented dataset provides a sufficient learning library for the reconstruction model, ensuring that the model does not bias towards learning specific types of data samples. Before training the model, the experimental data are preprocessed using image processing techniques to obtain the required low-resolution image data samples. Since the sizes of the experimental dataset images vary, each image is randomly cropped into multiple sub-images of the same size as samples. The training sub-images have a size of 64×64. Using such data samples to train the model enhances the stability of the training process. To validate the effectiveness of the proposed algorithm, the datasets Set5, Set14, BSD100, and Urban100 are used as test datasets.

C. Evaluation Metrics

Image quality assessment methods can be categorized into subjective and objective evaluation methods based on their approach. Currently, there are two mainstream objective evaluation metrics: one is the Peak Signal-to-Noise Ratio (PSNR), which assesses differences in image pixel points, and the other is the Structural Similarity (SSIM), which provides a comprehensive evaluation by taking into account image luminance, contrast, and structural factors.

PSNR is an objective benchmark for measuring image quality, used to quantify the discrepancy between a processed image and the original image. The calculation of PSNR is as follows:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[I(i,j) - K(i,j)\right]^2 \quad (1)$$

$$PSNR = 10\log\left(\frac{MAX_I^2}{MSE}\right) \quad (2)$$

Where I and K represent the reconstructed image and the original image, respectively, while *m* and *n* denote the height and width of the image. For PSNR, a higher value indicates less distortion in the reconstructed image and a closer approximation to the original image. The unit of the PSNR metric is dB.

Unlike PSNR, which calculates differences at the pixel level, SSIM measures the dissimilarity between images by computing variations in luminance, contrast, and structural information. The calculation formula is as follows:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \quad (3)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2'} \quad (4)$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \quad (5)$$

$$SSIM(x,y) = l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma} \quad (6)$$

Here, *x* and *y* represent the reconstructed image and the original image, respectively. The term l (x, y) denotes the luminance comparison, c (x, y) represents the contrast comparison, and s(x, y) captures the structural comparison. A higher

SSIM value indicates smaller structural discrepancies between the images, with an upper limit of 1, indicating that the images are identical.

D. Comparative Experiments

To validate the effectiveness of the proposed algorithm in image super-resolution reconstruction, it was compared with several classical networks, namely bicubic, SRCNN, VDSR, and ESRGAN. All methods were trained using the DIV2K dataset. Compared to images reconstructed by bicubic, those produced by the SRCNN algorithm show significant overall improvement. However, due to its shallow network depth and limited feature extraction capability, SRCNN recovers fewer fine details. VDSR, with its deeper network layers and incorporation of residual networks, achieves better reconstruction results. Nevertheless, its simplistic network architecture leads to poor internal information flow, causing the loss of some detailed textures. As for SRGAN and ESRGAN, they improve the overall visual quality of the reconstructed images and perform better than the aforementioned methods. However, they occasionally generate superfluous textural details in smooth areas. In contrast, the proposed algorithm not only achieves favorable overall reconstruction quality but also reduces the generation of extraneous noise in smooth regions.

TABLE II. COMPARATIVE EXPERIMENT RESULT

Algorithm	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM
Bicubic	27.40/0.799	26.08/0.692	25.68/0.658	23.22/0.670
SRCNN	28.72/0.852	27.27/0.710	27.48/0.688	27.01/0.701
VDSR	29.27/0.883	29.02/0.770	28.22/0.718	28.34/0.782
SRGAN	28.82/0.815	27.32/0.795	26.17/0.824	26.02/0.804
ESRGAN	27.80/0.883	27.21/0.780	27.55/0.733	27.02/0.791
PULSE	29.89/0.879	29.53/0.845	29.19/0.826	28.92/0.815
Ours	29.57/0.933	29.52/0.924	29.35/0.917	29.23/0.912

As for SRGAN and ESRGAN, they improve the overall visual quality of the reconstructed images and perform better than the aforementioned methods. However, they occasionally generate superfluous textural details in smooth areas. In contrast, the proposed algorithm not only achieves favorable overall reconstruction quality but also reduces the generation of extraneous noise in smooth regions.

Figure 7. Comparative experimental renderings

E. Ablation experiment

Ablation experiments were designed to validate the effectiveness of the Multi-stage Gated Aggregation Module (MGAM) and the Cross-Scale Feature Fusion Module (ESCDB). The model formed by removing both of these modules from the complete network serves as the Original Model (OM). The experiments compare the following configurations: the Original Model with the Multi-stage Gated Aggregation Module (OM+MGAM), the Original Model with the Cross-Scale Feature Fusion Module (OM+ESCDB), and the final proposed algorithm, which is the Original Model integrated with both the Multi-stage Gated Aggregation Module and the Cross-Scale Feature Fusion Module (OM+MGAM+ESCDB).

TABLE III. ABLATION EXPERIMENT RESULT

Algorithm	Set5 PSNR/SSIM	Set14 PSNR/SSIM
OM	27.80/0.883	27.21/0.780
OM+MGAM	28.23/0.896	27.69/0.852
OM+ESCDB	29.25/0.921	28.84/0.910
OM+MGAM+ESCDB	29.57/0.933	29.52/0.924

Figure 8. Ablation experimental renderings

The substantial SSIM improvement (e.g., 0.933 on Set5) primarily stems from the EMA-based cross-scale feature fusion module, as verified by our ablation study (Table III). OM+ESCDB alone yields a SSIM of 0.921 on Set5, while OM+MGAM achieves only 0.896. Therefore, the spatial modeling ability of EMA contributes more to

structural similarity than the multi-stage gated aggregation.

V. Copyright Forms and Reprint Orders

Although the super-resolution reconstruction algorithm proposed in this paper has achieved progressive advancements, it still exhibits notable limitations in both practical applications and theoretical simulation. While the current model outperforms traditional networks in performance metrics, the inherent complexity of its deep network architecture results in substantial computational overhead, making it difficult to meet the demands of real-world scenarios with high real-time requirements. Future research will focus on model lightweighting design, aiming to significantly improve inference efficiency while maintaining reconstruction quality through structural simplification and parameter optimization. Furthermore, the non-blind reconstruction datasets commonly used in existing algorithms fail to adequately simulate the complex degradation processes of low-resolution images in real-world settings, which undermines the practical applicability of these methods. Therefore, developing more realistic blind reconstruction datasets will be a crucial research direction for enhancing the practical value of the algorithm.

References

[1] Liu J, Zhang W, Tang Y. Residual feature aggregation network for image super-resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2359-2368

[2] Ma C, Rao Y, Cheng Y, et al. Structure-preserving super resolution with gradient guidance[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7769-7778.

[3] Bashir S M A, Wang Y, Khan M, et al. A comprehensive review of deep learning-based single image super-resolution[J]. PeerJ Computer Science,2021, 7: e621.

[4] Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution[C]//European conference on computer vision. Springer, Cham, 2014: 184-199.

[5] Dong C, Loy C C, Tang X. Accelerating the super-resolution convolutional neural network[C]//European conference on computer vision. Springer, Cham, 2016: 391-407.

[6] Li W, Zhou K, Qi L, et al. Best-buddy gans for highly detailed image super-resolution[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(2): 1412-1420.

[7] Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1646-1654.

[8] Kim J, Lee J K, Lee K M. Deeply-recursive convolutional network for image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1637-1645.

[9] Song D, Wang Y, Chen H, et al. Addersr: Towards energy efficient image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15648-15657.

[10] Ward C M, Harguess J, Crabb B, et al. Image quality assessment for determining efficacy and limitations of Super-Resolution Convolutional Neural Network (SRCNN)[C]//Applications of Digital Image Processing XL. SPIE, 2017, 10396: 19-30.

[11]

[12] Lu L, Li W, Tao X, et al. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6368-6377.

[13] Xing W, Egiazarian K. End-to-end learning for joint image demosaicing, denoising and super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3507-3516.

[14] Qiu Y, Wang R, Tao D, et al. Embedded block residual network: A recursive restoration model for single-image super-resolution[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 4180-4189.

[15] Wang W, Zhang H, Yuan Z, et al. Unsupervised real-world super-resolution: A domain adaptation perspective[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4318-4327.

[16] Pesavento M, Volino M, Hilton A. Attention-based multi-reference learning for image super-resolution[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14697-14706.