Research on Medical Image Super-Resolution Reconstruction Using a Transformer-RRDB Hybrid Generator

Yingjie Song
School of Electrical and Information Engineering, Heilongjiang Institute of Technology, Jixi, Heilongjiang, China

Abstract: The limited resolution of magnetic resonance imaging (MRI) is one of the important factors affecting the accuracy of clinical diagnosis. This paper proposes a medical image super-resolution reconstruction method based on a Transformer-RRDB hybrid generator, targeting the 4×super-resolution task of brain T2-weighted MRI images. The method employs a cascade structure of residual dense blocks and Swin Transformer, where RRDB is used to extract local texture features and the window-based self-attention mechanism models the global spatial correlations of brain anatomical structures. In the design of the loss function, we combine the pixel-wise L1 loss, perceptual loss, frequency-domain amplitude-phase joint constraint, and DDO implicit discriminative regularization with a fixed scoring proxy, forming multi-dimensional supervisory signals. The training adopts a two-stage transfer learning strategy of pre-training on DF2K natural images followed by fine-tuning on IXIT2 medical images. Experimental results show that the proposed method achieves PSNR 28.42 dB, SSIM 0.845, LPIPS 0.107, and gEn 0.092 on the IXIT2 test set, outperforming comparative methods such as Bicubic, SRCNN, ESRGAN, and SwinIR in all metrics, which validates the effectiveness of the hybrid architecture and frequency-domain constraints in medical image super-resolution.

Keywords: Medical Image Super-Resolution; Transformer; RRDB; Frequency-Domain Loss; Transfer Learning

1. Introduction

Magnetic resonance imaging (MRI) plays an important role in clinical diagnosis due to its excellent soft-tissue contrast and absence of ionizing radiation. However, its spatial resolution is limited by factors such as scan time, hardware cost, and patient tolerance; therefore, improving resolution via computational methods without modifying hardware has significant clinical value.

Single image super-resolution (SISR) aims to recover a high-resolution image from its low-resolution input. SRCNN first demonstrated the feasibility of end-to-end learning; SRGAN introduced generative adversarial networks and perceptual loss; ESRGAN further improved reconstruction quality through residual dense blocks (RRDB) and relativistic discriminators; and SwinIR exploited window-based self-attention to model global contextual information. Directly transferring these methods to medical images still faces three major challenges. First, MRI textures differ significantly from natural images, limiting the generalization ability of pre-trained models [1]. Second, the convolutional operations in RRDB are insufficient for modeling global correlations of anatomical structures across different regions [2]. Third, pixel-wise losses tend to cause over-smoothing, which weakens the sharpness of tissue boundaries [3].

To address these issues, this paper proposes a Transformer-RRDB hybrid generator, which employs RRDB to extract local textures and Swin Transformer to model global spatial correlations. We introduce a frequency-domain amplitude-phase joint constraint to enhance high-frequency structural supervision, and adopt DDO implicit discriminative regularization with a fixed scoring proxy in place of a conventional explicit discriminator to improve training stability. The training follows a two-stage transfer strategy of pre-training on DF2K followed by fine-tuning on IXIT2.

The main contributions of this paper are as follows: (1) proposing a hybrid generator architecture that cascades RRDB and Swin Transformer; (2) introducing the frequency-domain amplitude-phase joint constraint into medical super-resolution; (3)

adopting DDO implicit discriminative regularization with a fixed scoring proxy; and (4) verifying the effectiveness of two-stage transfer learning in MRI super-resolution, and quantitatively evaluating the contribution of each module through ablation and comparative experiments.

2. Related Work

2.1 CNN- and GAN-Based Image Super-Resolution

SRCNN applied three convolutional layers to image super-resolution [4]. SRGAN introduced perceptual loss and adversarial loss to improve visual quality, with a generator designed using residual blocks [5]. ESRGAN systematically improved upon this by replacing residual blocks with residual dense blocks (RRDB) [6], which enhance feature reuse through dense connections and residual scaling; removing batch normalization layers; and introducing a relativistic discriminator to improve training stability. ESRGAN achieved state-of-the-art performance on natural images, yet its purely convolutional architecture suffers from limited receptive fields. In the medical domain, ESRGAN has been applied to MRI and CT super-resolution [7], but GANs often suffer from training instability in small-sample scenarios—when the discriminator is too strong, gradients vanish, and when too weak, supervision becomes insufficient [8].

2.2 Transformer in Image Restoration

SwinIR applied Swin Transformer to tasks such as super-resolution, denoising, and deblurring [9]. Its core lies in the window-based multi-head self-attention and shifted-window mechanism: self-attention is restricted to non-overlapping local windows, reducing complexity to scale with window area rather than the square of the image pixel count; adjacent Transformer blocks alternately use regular and shifted windows to enable cross-window interaction. SwinIR outperformed pure CNN methods on several benchmarks, yet pure Transformer architectures are less efficient in recovering local textures compared to CNNs.

2.3 Frequency-Domain Constraints and Implicit Discrimination

Traditional methods mainly rely on pixel-domain losses. In recent years, frequency-domain constraints have gained increasing attention. The 2D Fourier transform decomposes an image into amplitude and phase spectra, where the amplitude spectrum reflects the distribution of frequency energy and the phase spectrum encodes spatial structure. Studies have shown that joint amplitude-phase constraints are more effective in preserving sharp anatomical boundaries than purely pixel-wise losses [10]. On the discriminator side, DDO proposed an optimization paradigm that does not require an explicit discriminator, instead constructing implicit discriminative signals using a frozen reference scoring model. In this work, we adopt the engineering implementation of the DDO paradigm by employing a fixed scoring proxy network rather than a full pre-trained diffusion model, thereby circumventing the training instability issues inherent in conventional GANs.

2.4 Two-Stage Transfer Learning

Transfer learning has been widely used in medical image analysis. Pre-trained models on natural images can be adapted to medical tasks through fine-tuning, alleviating the problems of high annotation costs and limited sample sizes. Models such as ESRGAN, after being trained on large-scale datasets like DF2K, have already learned general image priors. Transferring such priors to medical super-resolution is a key strategy in this paper.

3. Method

3.1 Overall Framework

The workflow of the proposed method is as follows. Given a low-resolution input image ILR, the generator produces a super-resolved image ISR, with the high-resolution reference image IHR serving as the supervision during training. The generator pipeline proceeds as: ILR is first passed through a 3×3 convolutional layer for shallow feature extraction, then sequentially through a front RRDB block, the Swin Transformer module, and a rear RRDB block for deep feature extraction. The deep features are residually fused with the shallow features, followed by PixelShuffle upsampling and a 3×3 output convolution to yield ISR. The total loss comprises four terms: pixel-wise L1 loss, perceptual loss, frequency-domain loss, and DDO loss. Among these, the weights for the frequency-domain and DDO terms are gradually increased during training, while the weights for

the pixel and perceptual losses are kept fixed.

3.2 Generator Network Architecture

(1)Shallow feature extraction and global residual connection: The first layer is a 3×3 convolution (stride 1, padding 1) that maps the input into a 64-dimensional feature space F0 . F0 is retained as a global residual branch and fused with the deep features after the deep feature processing is completed, allowing basic luminance, edges, and low-frequency structures to be directly propagated to the output.

(2) Residual dense blocks and RRDB: Each residual dense block (RDB) consists of five convolutional layers. The first four layers each output 32 growth features, which are concatenated along the channel dimension; the fifth layer compresses the concatenated features back to 64 channels. The input channel count for the i-th layer is 64+i×32 (i=0,1,2,3). LeakyReLU (with a negative slope of 0.2) is applied after each convolution. The block-level output is given by:

$$RDB(x)=x+0.2\times Conv_5([x,x_1,x_2,x_3,x_4]) \quad (1)$$

The RRDB is composed of three RDBs cascaded in series, and also employs residual scaling:

$$RRDB(x)=x+0.2\times RDB_3(RDB_2(RDB_1(x))) \quad (2)$$

The generator contains 23 RRDB modules by default, divided into 11 in the front section and 12 in the rear section. The dense connections reuse local features from different hierarchical levels, while the residual scaling coefficient suppresses gradient oscillations in the deep network.

(3) Swin Transformer middle stage: The output of the front RRDB section is fed into a sequence of Swin Transformer blocks. By default, the sequence contains 4 Swin Transformer blocks with a window size of 8 and 4 attention heads in the multi-head self-attention. Each block sequentially performs layer normalization, multi-head self-attention (with learnable relative position bias), output projection, layer normalization, and a two-layer MLP with GELU activation. Adjacent Transformer blocks alternately use regular windows and cyclically shifted windows to enable cross-window information interaction. The complexity of window attention scales with the window area, making it suitable for large-sized feature maps. If the input size is not an integer multiple of the window size, reflection padding is first applied for alignment, and the output is cropped back to the original size after computation. The Transformer module is designed to compensate for the limited receptive field of RRDB by modeling long-range dependencies of anatomical structures such as the ventricles and the corpus callosum.

(4) Residual fusion and upsampling: The features processed by the Swin Transformer and the rear RRDB section are passed through a 3×3 convolution and then added to the shallow features:

$$F_{fused} = F_0 + Conv_{3\times3}(F_{after}) \quad (3)$$

Each upsampling stage consists of a 3×3 convolutional layer (which outputs four times the number of main channels), a PixelShuffle operation with an upsampling factor of 2, and a LeakyReLU activation. For 4× super-resolution, two such stages are employed. PixelShuffle increases the resolution via channel rearrangement, effectively avoiding checkerboard artifacts. Finally, two 3×3 convolutional layers are used to produce the output image.

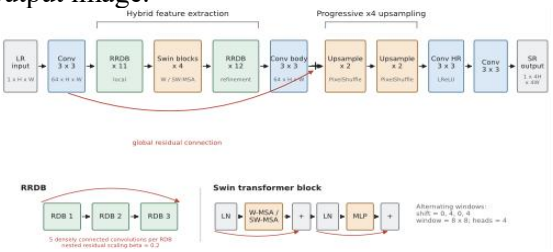

Figure 1. Schematic Diagram of the Generator Network Architecture (a) Hybrid Feature Extraction Backbone; (b) Internal Structure of the RRDB; (c) Structure of the Swin Transformer Block; (d) Progressive Upsampling Module.

Figure 1 illustrates the complete network structure of the generator. As shown in Figure 1(a), the low-resolution input is sequentially fed through a shallow convolution, 11 RRDB modules, 4 Swin Transformer modules, and 12 RRDB modules. The shallow features are fused via a global residual connection, followed by two progressive upsampling stages to achieve 4× magnification. Figure 1(b) presents the internal structure of the RRDB — each RRDB consists of three Residual Dense Blocks (RDBs) in series, and each RDB contains five densely connected convolutional layers with a residual scaling factor of 0.2. Figure 1(c) details the composition of the Swin Transformer block: layer normalization, Window-based Multi-head Self-Attention (W-MSA) or Shifted Window-based Multi-head Self-Attention (SW-MSA), layer normalization, and a two-layer

MLP. Adjacent blocks alternate between window partitioning strategies with a window size of 8×8 and shift amounts of 0 and 4. Figure 1(d) shows the detailed flow of the upsampling module, where each stage consists of convolution, PixelShuffle, and LeakyReLU; the two stages successively increase the spatial resolution by a factor of 2.

3.3 Loss Function Design

(1) Pixel-wise L1 loss: computes the mean absolute error between the super-resolved image and the high-resolution reference image on a per-pixel basis.

$$L_{pixel}=\text{mean}(|I_{SR}-I_{HR}|) \quad (4)$$

Compared with the L2 loss, it is more robust to outlier pixels and constrains the overall luminance, contrast, and low-frequency structures.

(2)Perceptual loss: By default, the FixedPerceptualLoss is adopted, which computes the L1 distance using fixed Sobel horizontal/vertical filters, a Laplacian filter, and multi-scale average-pooled features. If local VGG-19 weights are provided, it can be switched to frozen convolutional features before the 35th layer of VGG-19.

(3) Frequency-domain loss: For each channel, a normalized 2D FFT is performed, yielding FSR=F(ISR) and FHR=F(IHR). For the amplitude term, a log-amplitude L1 distance is adopted:

$$L_{amp}=\text{mean}(|\log(1+|F_{SR}|)-\log(1+|F_{HR}|)|) \quad (5)$$

For the phase term, a period-safe cosine difference is adopted:

$$\mathcal{L}_{phase}=\text{mean}(1-\cos(\angle F_{SR}-\angle F_{HR})) \quad (6)$$

The frequency-domain loss is the weighted average of the two terms, each with a weight of 0.5:

$$\mathcal{L}_{fourier}=0.5\mathcal{L}_{amp}+0.5\mathcal{L}_{phase} \quad (7)$$

The amplitude spectrum reflects the distribution of texture energy, while the phase spectrum is associated with spatial structure and edge localization. The joint constraint is therefore more suitable for preserving the sharpness of anatomical boundaries.

(4) DDO implicit discriminative regularization: A fixed scoring proxy network (rather than a full pre-trained diffusion model) is adopted in place of a conventional trainable discriminator. Its parameters are frozen, serving solely as an implicit discriminative signal source that provides gradient feedback to the generator:

$$\mathcal{L}_{DDO}=-(S(I_{SR})-S(I_{HR}).\text{detach}()) \quad (8)$$

This eliminates the need to train and maintain a separate discriminator, thereby avoiding the training instability issues inherent in conventional GANs.

(5) Progressive weighting strategy: The weights of the frequency-domain and DDO losses are increased linearly in a progressive manner, while the weights of the pixel-wise and perceptual losses are kept fixed. At epoch e, the progress is defined as t=(e−1)/(E−1), frequency-domain weight:

$$\lambda_f(e)=\lambda_f^{start}+t\cdot(\lambda_f^{end}-\lambda_f^{start}) \quad (9)$$

DDO weight:

$$\lambda_d(e)=\lambda_d^{start}+t\cdot(\lambda_d^{end}-\lambda_d^{start}) \quad (10)$$

During the DF2K stage, the frequency-domain weight increases from 0.01 to 0.1, and the DDO weight from 0.001 to 0.01; during the IXIT2 stage, the upper limits are raised to 0.15 for the frequency-domain loss and 0.02 for the DDO loss. This allows the model to first learn low-frequency structures and then progressively strengthen high-frequency and perceptual constraints.

3.4 Two-Stage Transfer Learning Strategy

(1) DF2K pre-training: The DF2K dataset combines DIV2K (2,000 images) and Flickr2K (2,650 images), split into training and validation sets at a ratio of 9:1. The Adam optimizer (betas=0.9, 0.99) is employed with an initial learning rate of 1e-4, cosine annealing to 1e-7, and a gradient clipping threshold of 1.0. The objective is to learn general degradation inversion, edge recovery, and texture reconstruction capabilities. Upon completion, the generator weights with the optimal OQS_DF2K on the validation set are saved.

(2) IXIT2 fine-tuning: Using the T2-weighted modality from the IXI dataset, the model is initialized with the optimal DF2K weights. The learning rate is set to 1e-5 to mitigate catastrophic forgetting. The upper bounds of the frequency-domain and DDO weights are increased to strengthen the supervision specific to medical image structures. Fine-tuning is performed for 150 epochs without early stopping, and the final checkpoint at epoch 150 is used as the model weights. PSNR, SSIM, LPIPS, and gEn metrics are computed on the validation set at each epoch.

(3) Training stability safeguards: The DDO reference model is frozen; mixed-precision training is adopted; gradient clipping is set to 1.0;

the frequency-domain and DDO weights are progressively increased; Adam is used with cosine annealing; and all random seeds are fixed to ensure reproducibility.

4. Experimental Setup

4.1 Dataset and Preprocessing

The DF2K dataset is a fusion of DIV2K (2,000 images) and Flickr2K (2,650 images). 4× low-resolution images are generated via bicubic downsampling, and the dataset is split into training and validation sets at a ratio of 9:1. On the pre-training validation set, the optimal PSNR is 26.94 dB (with an OQS of 0.670), and the optimal OQS is 0.671.

The IXIT2 dataset is a publicly available brain MRI dataset. High-resolution images are automatically paired with their 4× downsampled counterparts by filename, and the training/validation sets are divided according to directories. All images are converted to single-channel grayscale and normalized to [0, 1]. During training, random patches are cropped from the LR images (with a size of patch_size / scale) and synchronously cropped from the corresponding HR patches.

4.2 Data Augmentation Strategy

Paired augmentation is adopted, where geometric transformations are applied simultaneously to both LR and HR images: horizontal flipping (p=0.5), vertical flipping (p=0.5), random rotation within [-10°, 10°], and contrast adjustment (p=0.2, range 0.95–1.05).

4.3 Training Hyperparameters and Network Architecture

DF2K pre-training stage: epochs=250, batch size=8, lr=1e-4 with cosine annealing to 1e-7, frequency-domain weight from 0.01 to 0.1, DDO weight from 0.001 to 0.01.

IXIT2 fine-tuning stage: epochs=150, batch size=8, lr=1e-5, frequency-domain upper bound=0.15, DDO upper bound=0.02, no early stopping.

Network architecture hyperparameters: main channels=64, RRDB blocks=23, RDB growth rate=32, Swin modules=4, window size=8, attention heads=4. Optimizer: Adam (betas=0.9, 0.99), weight decay=0, gradient clipping=1.0. Perceptual loss weight=0.1, pixel loss weight=1.0.

4.4 Evaluation Metrics

PSNR, SSIM, LPIPS (AlexNet), and gEn (mean absolute error of Sobel gradient magnitudes) are adopted as evaluation metrics.

4.5 Comparison Models and Training Configurations

The comparison methods include Bicubic, SRCNN, ESRGAN, and SwinIR, all of which are trained from scratch on IXIT2 for 100 epochs (lr=1e-5) until the validation loss reaches a plateau. Our method is pre-trained on DF2K for 250 epochs, followed by fine-tuning on IXIT2 for 50 epochs (in ablation studies) or 150 epochs (for the full model).

5. Experimental Results and Analysis

5.1 DF2K Pre-training Results

The DF2K pre-training stage is conducted for a total of 250 epochs. The optimal PSNR checkpoint achieves 26.94 dB (with a corresponding OQS of 0.670), the optimal OQS_DF2K value is 0.671, and the validation SSIM is 0.777. During training, the pixel loss decreases from 0.065 to 0.028, and the perceptual loss from 0.143 to 0.072. The frequency-domain and DDO losses steadily decline with the progressive weighting strategy, without exhibiting gradient oscillations or non-convergence.

5.2 IXIT2 Fine-tuning Results

After initialization from the optimal DF2K pre-trained model, fine-tuning is conducted on IXIT2 for 150 epochs. At the optimal OQS_IXI checkpoint (the 111th epoch), the validation set achieves a PSNR of 25.51 dB, SSIM of 0.809, LPIPS of 0.0873, and gEn of 0.1695, with an optimal OQS_IXI value of −1.634. During the fine-tuning process, all metric curves remain smooth without severe oscillations or signs of overfitting, which validates the effectiveness of the progressive weighting strategy and gradient clipping in maintaining training stability.

5.3 Comparative Experimental Results

The quantitative results of all methods on the IXIT2 test set are presented in Table 1.

SRCNN achieves 26.66 dB, which is slightly superior to Bicubic, but its LPIPS and gEn remain relatively high. ESRGAN exhibits abnormal performance (19.74 dB). The analysis suggests that conventional GANs fail to

converge stably within 100 epochs on the limited IXIT2 data, as the discriminator becomes overly strong too early, causing unstable gradients for the generator. In contrast, our method adopts DDO with a fixed scoring proxy in place of an explicit discriminator, completely circumventing this issue. In ablation setting E, it achieves 28.42 dB, confirming the stability advantage of the DDO scheme. SwinIR obtains 25.62 dB, indicating that 100 epochs are insufficient to fully exploit its global modeling potential; by contrast, our method reaches 28.42 dB with only 50 epochs of fine-tuning, highlighting the crucial role of transfer learning in medical small-sample tasks—the number of medical images required to train a complex architecture from random initialization far exceeds the amount that can be practically acquired.

To qualitatively evaluate the visual performance of each method, Figure 2 presents a comparison of the reconstruction results produced by different approaches on the IXI T2 test set. From the locally magnified regions, it is evident that the Bicubic interpolation results are generally blurry, with unclear boundaries of brain tissues. SRCNN achieves moderate improvements in edge preservation, yet the reconstructed details remain oversmoothed. ESRGAN outputs exhibit noticeable texture artifacts, which are consistent with its abnormally low PSNR values reported in the quantitative results. SwinIR produces relatively sharper reconstructions, but still shows insufficient sharpness in certain anatomical structures. In contrast, the proposed method achieves results that are visually closest to the high-resolution references (HR), particularly in terms of ventricular boundaries and gray-white matter contrast. The reconstructed images are visually the most natural, with sharp edges and no significant artifacts, aligning well with the optimal quantitative performance of the proposed method across all metrics reported in Table 1.

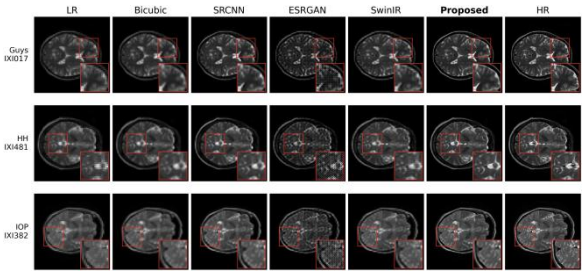

Figure 2. Qualitative Comparison of Super-Resolution Results Produced by Different Methods on the IXI T2 Test Set

Table 1. Quantitative Comparison Results of Different Methods on the IXIT2 Test Set

Method	Training strategy	PSNR(dB)	SSIM	LPIPS	gEn
Bicubic	—	25.84	0.782	0.290	0.131
SRCNN	Trained from scratch for 100 epochs	26.66	0.783	0.235	0.115
ESRGAN	Trained from scratch for 100 epochs	19.74	0.609	0.232	0.170
SwinIR	Trained from scratch for 100 epochs	25.62	0.756	0.175	0.124
Ours (ablation configuration E)	Pre-training + fine-tuning for 50 epochs	28.42	0.845	0.107	0.092

5.4 Ablation Experimental Results

All five ablation configurations are initialized from the same DF2K pre-trained model and fine-tuned on IXIT2 for 50 epochs. The results are presented in Table 2.

Table 2. Quantitative Results of Ablation Experiments

Configuration	Swin	Frequency	DDO	PSNR(dB)	SSIM	LPIPS	gEn
A (baseline)	✗	✗	✗	28.21	0.839	0.0997	0.0935
B (A+Swin)	✓	✗	✗	28.31	0.843	0.0969	0.0923
C (A+Frequency)	✗	✓	✗	28.31	0.842	0.1065	0.0930
D (A+Swin+Frequency)	✓	✓	✗	28.43	0.846	0.1082	0.0915
E (Full model)	✓	✓	✓	28.42	0.845	0.1073	0.0916

Contribution of SwinTransformer (B vs. A): All four metrics show positive improvements, with PSNR increased by 0.10 dB, validating the effectiveness of window-based self-attention in modeling global anatomical structures in brain MRI.

Contribution of frequency-domain loss (C vs. A): PSNR improves by 0.10 dB, with gains in SSIM and gEn; however, LPIPS increases from 0.0997 to 0.1065, which may be attributed to the amplification effect of the frequency-domain loss on noise in low-SNR regions of MRI.

Synergistic effect of Swin and frequency-domain loss (D vs. B, D vs. C): D outperforms both single-module configurations in terms of PSNR and SSIM, indicating a synergy between global modeling and high-frequency enhancement.

Contribution of DDO (E vs. D): LPIPS improves from 0.1082 to 0.1073, while PSNR and SSIM remain largely unchanged. The modest gain may

be due to the short 50-epoch fine-tuning period, which does not allow the progressive weights to be fully exploited. Nevertheless, the value of DDO in training stability remains clear—as evidenced by the collapse of ESRGAN in the comparative experiments, which serves as a counterexample.

5.5 Analysis of Transfer Learning Effectiveness

A substantial gap of 5.75 dB exists between training from scratch (22.67 dB) and pre-training on DF2K followed by fine-tuning (28.42 dB), which strongly demonstrates the effectiveness of transfer learning. For the hybrid architecture with 17.05M parameters, DF2K pre-training provides an initialization point close to the optimum, enabling the subsequent fine-tuning to converge within only a few epochs.

5.6 Discussion of Limitations

The frequency-domain loss has a slightly negative impact on perceptual quality; this could potentially be mitigated in future work through frequency-domain masking or adaptive weighting. The current DDO implementation employs a fixed scoring proxy rather than a full pre-trained diffusion model, which limits its representational capacity. The ablation experiments are conducted with only 50 epochs of fine-tuning, so the long-term effects of certain modules remain to be verified. In addition, the experiments are performed solely on the IXIT2 modality, and the generalization capability of the proposed method to other modalities or datasets awaits further exploration.

6. Conclusion

This paper proposes a super-resolution method based on a Transformer-RRDB hybrid generator, targeting the 4×reconstruction task of brain T2-weighted MRI images. In terms of generator architecture, a cascade structure combining RRDB and SwinTransformer is adopted, endowing the model with both local detail recovery and global structure preservation capabilities. Regarding the loss function, a frequency-domain amplitude-phase joint constraint and DDO implicit discriminative regularization with a fixed scoring proxy are introduced, forming multi-dimensional supervisory signals. For the training strategy, a two-stage transfer learning approach is employed, consisting of pre-training on DF2K natural images followed by fine-tuning on IXIT2 medical images.

Experimental results on the IXIT2 dataset demonstrate that the proposed method (ablation configuration E) achieves a PSNR of 28.42 dB, SSIM of 0.845, LPIPS of 0.107, and gEn of 0.092, outperforming Bicubic, SRCNN, ESRGAN, and SwinIR across all metrics. Ablation experiments validate the individual effectiveness and synergistic effects of the SwinTransformer, frequency-domain loss, and DDO regularization. The collapse of ESRGAN in the comparative experiments further highlights the advantage of our DDO scheme with a fixed scoring proxy in terms of training stability. The analysis of transfer learning effectiveness indicates that DF2K pre-training makes a decisive contribution to IXIT2 fine-tuning, yielding a 5.75 dB improvement in PSNR.

Future work will explore the generalization capability of the proposed architecture on other MRI modalities, as well as upgrading the DDO reference model to a full pre-trained diffusion model to further enhance perceptual quality.

References

[1] Huang W F, Liao X Y, Chen H, et al. Deep local-to-global feature learning for medical image super-resolution. Computerized Medical Imaging and Graphics, 2024, 115:102374.

[2] Li G Y, Lv J, et al. Multicontrast MRI Super-Resolution via Transformer-Empowered Multiscale Contextual Matching and Aggregation. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(9):12004-12014.

[3] Liu Q H, Zhang W K, Zhang Y T, et al. DGEDDGAN: A dual-domain generator and edge-enhanced dual discriminator generative adversarial network for MRI reconstruction. Magnetic Resonance Imaging, 2025, 119:110381.

[4] Dong C, Loy C C, He K M, Tang X O. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295-307.

[5] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network//Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition (CVPR). IEEE, 2017: 4681-4690.

[6] Wang X, Yu K, Wu S, et al. ESRGAN: Enhanced super-resolution generative adversarial networks//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. Springer, 2018: 63-79.

[7] Nandal P, Pahal S, Khanna A, et al. Super-resolution of medical images using Real ESRGAN. IEEE Access, 2024, 12: 176155-176170.

[8] Sun J W, Li Z Y, Li P C, et al. Improving the diagnostic performance of computed tomography angiography for intracranial large arterial stenosis by a novel super-resolution algorithm based on multi-scale residual denoising generative adversarial network. Clinical Imaging, 2023, 96: 1-8.

[9] Liang J Y, Cao J Z, Sun G L, et al. SwinIR: Image Restoration Using Swin Transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, 2021: 1833-1844.

[10] Li J L, Wang G Z, Zhao H W. Magnetic resonance image super-resolution reconstruction based on frequency-domain constraints and cross-fusion feature. Chinese Journal of Medical Physics, 2023, 40(1): 31-38.

