2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Reconstructing Reality: Robust High-Frequency Recovery for MRI via Latent Diffusion Models

Tianzhi Wang[1] and Jian Wang[2*]

[1] Department of computer and Systems Sciences, Stockholm University
terrywangsd@gmail.com
[2] Gannan Normal University
1764318284@qq.com

Abstract. Multi-contrast magnetic resonance imaging (MRI) is a widely used analytical tool for characterizing tissue contrast in neurological disorders. Although conventional MRI techniques provide rich contrast information in the diagnosis of neurological diseases, their limited spatial resolution often hinders the precise identification of subtle pathological regions. Therefore, super-resolution (SR) reconstruction of MRI images holds significant importance in the field of medical imaging. Traditional end-to-end deep neural network approaches tend to learn the average of multiple possible reconstruction outcomes, resulting in overly smoothed generated images that lack high-frequency details. In recent years, generative models have demonstrated remarkable capabilities in SR tasks by synthesizing more realistic high-frequency information, thereby substantially mitigating the aforementioned issue. However, generative models generally exhibit considerable randomness, making it challenging to ensure the stability and consistency of the results. To address this, we propose a novel MRI SR method that integrates the strengths of both generative and discriminative models. Specifically, we employ a latent diffusion model (LDT) to capture the high-frequency information in real images and utilize the low-frequency information from low-resolution (LR) images as conditional input for an autoencoder to generate high-resolution (HR) images. Quantitative experimental results demonstrate that our method outperforms existing state-of-the-art MRI SR approaches across multiple metrics while maintaining a more lightweight architecture. Furthermore, visualization results further validate the superiority of our method in reconstructing high-frequency details.

Keywords: MRI, Super-Resolution, Diffusion Model, Autoencoder, Wavelet Transform, Discriminative Model.

1 Introduction

Magnetic resonance imaging (MRI) is a non-invasive imaging technique that plays a crucial role in the diagnosis and research of neurological disorders [15]. Multi-contrast

* Tianzhi Wang and Jian Wang contributed equally. Corresponding author: 1764318284@qq.com

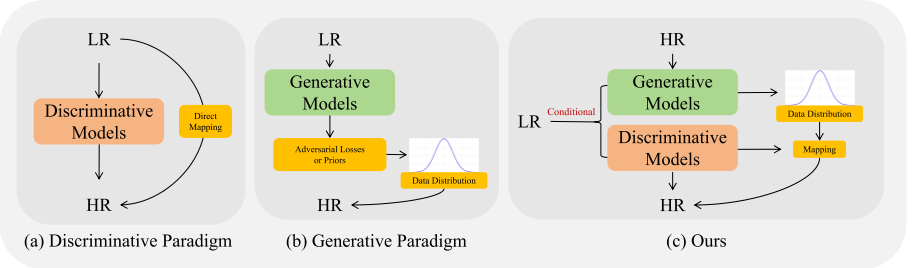


Fig. 1. Comparison of different paradigms in super-resolution. (a) Discriminative paradigms typically adopt end-to-end training with L1/L2 constraints to learn a direct mapping from low-resolution to high-resolution images. (b) Generative paradigms, represented by GANs and diffusion models, model the data distribution and generate samples accordingly. (c) Our approach combines the strengths of both generative and discriminative paradigms: the generative model captures the distribution of the real image space, while the discriminative model enforces consistency constraints, resulting in super-resolved images with both high perceptual quality and strong fidelity.

search settings [34]. However, despite its rich contrast information, the spatial resolution of conventional MRI is often insufficient to identify subtle pathological changes, such as small lesions or microstructural abnormalities. This limitation can significantly hinder early and accurate diagnosis [22].

Super-resolution (SR) techniques aim to overcome this limitation by enhancing the spatial resolution of low-resolution (LR) MRI scans, thereby improving the visibility of fine anatomical details. In recent years, deep learning-based SR methods have made significant progress in this field [30]. Most existing approaches follow an end-to-end supervised learning paradigm [38,5], where deep neural networks are trained to map LR images to their corresponding high-resolution (HR) counterparts, as shown in Fig. 1(a). Although effective to some extent, these methods often produce overly smoothed outputs due to convolutional networks' tendency to minimize pixel-level losses (e.g., L1 or L2), resulting in the averaging of multiple possible HR reconstructions [27,36]. **This leads to the loss of high-frequency details that are crucial for clinical interpretation**.

To address this issue, generative models—such as generative adversarial networks (GANs [6]) and diffusion models (DMs [7])—have recently been introduced to MRI SR tasks [41,10,14,20]. These models can generate perceptually more realistic images with sharper textures and finer structures. Among them, latent diffusion models (LDMs [28]) have emerged as a promising direction due to their ability to model complex image feature distributions in a compressed latent space. However, despite improvements in perceptual quality, generative models often suffer from randomness and output inconsistency, which is undesirable in medical imaging, where high reliability is essential [36].

In this paper, we employ wavelet transform to decompose LR and HR images into frequency-domain representations and perform downsampling. We train a frequency autoencoder to further downsample the frequency-domain inputs. Then, using the low-frequency information from LR images as conditioning, we leverage a DM to learn the

distribution of HR high-frequency components. Finally, the low-frequency information from LR images serves as a conditioning input for the autoencoder's decoder, while the high-frequency information sampled from the conditional DM is used to reconstruct high-fidelity and consistent MRI SR results.

Specifically, we first apply wavelet transform to both low-resolution (LR) and high-resolution (HR) MRI images, decomposing them into frequency-domain representations to explicitly separate low- and high-frequency components while achieving spatial downsampling. In the frequency domain, we design and train a frequency autoencoder to further compress and model frequency-domain features, obtaining a more compact yet expressive high-frequency representation. This autoencoder embeds the original frequency-domain features into a low-dimensional latent space, facilitating subsequent modeling of complex high-frequency distributions [12]. Next, we use the low-frequency information extracted from LR images as conditioning and employ a diffusion probabilistic model to learn the generative distribution of the corresponding high-frequency components in HR images. Through the conditional diffusion process, we iteratively sample high-frequency details from noise while ensuring consistency with the given low-frequency conditioning, thereby generating high-quality textures. Finally, the high-frequency information synthesized by the DM is fed as input to the decoder of the frequency autoencoder, while the low-frequency components from the LR image serve as additional conditioning. This enables the reconstruction of high-fidelity and structurally consistent HR images. As shown in Fig. 1(c), our approach effectively combines the strengths of generative models in producing high-quality details and discriminative models in maintaining structural consistency, significantly improving the quality of MRI SR reconstruction.

2 Related Work

2.1 The Paradigm of Discriminative Models

Over the past decade, the field of image SR based on convolutional neural networks (CNNs) has witnessed significant evolution [39,37,33]. This end-to-end supervised learning paradigm typically involves designing hierarchical feature extraction and upsampling modules (e.g., SRCNN [2], EDSR [16], RCAN [42]) to drive pixel-level mapping from LR to HR images, driven by L1/L2 loss functions. The field has evolved from shallow networks to deep residual structures, from local perception to global attention mechanisms, and from single-scale to multi-scale fusion. With the increasing demand for high-precision diagnosis in medical imaging, SR technology has demonstrated unique value in modalities such as MRI. For instance, Pham et al. [25] innovatively introduced a multi-scale training strategy. This approach not only addressed the issue of anisotropic resolution but also validated the feasibility of multi-modal super-resolution and clinical low-quality image enhancement. Furthermore, lyu et al. [19] utilized a two-stage progressive architecture to fuse multi-contrast information in the high-level feature space, combined with a composite loss function design, significantly improving cross-modal reconstruction performance. More recently, ji et al. [[9] broke through the local limitations of traditional CNNs by integrating deformable

convolutions with state-space models, enabling long-range dependency modeling of complex anatomical structures. These advancements collectively have propelled the translation of medical SR from theoretical methods to clinical implementation.

2.2 The Paradigm of Generative Models

MRI SR techniques based on Generative Adversarial Networks (GANs) [41,10] have evolved from early simple reconstructions relying on single-modality data to high-fidelity reconstructions that integrate multi-modal information [11] and attention mechanisms (e.g., hybrid architectures of Transformer [3,44] and UNet [31]). This evolution has significantly enhanced the recovery quality of edge details and biological tissues. In terms of network architecture, the efficiency of the generator and discriminator has been optimized through Residual Dense Blocks, local attention modules, and region-specific reconstruction strategies [29,18,10], while improving the PSNR and SSIM metrics. Moreover, the integration of GANs with other techniques such as compressed sensing and dynamic contrast-enhanced imaging (e.g., DLCS-SR [17]) has broken through the limitations of spatial and temporal resolution. This has enabled the detection of small lesions (e.g., pituitary microadenomas) and dynamic monitoring of pathology in clinical applications. In recent years, diffusion models have emerged as a powerful alternative, demonstrating stability and high fidelity in MRI SR tasks. Xie et al. [35] proposed a Measurement-conditioned Denoising Diffusion Probabilistic Model based on the measurement domain and conditioned on undersampling masks for the reconstruction of undersampled medical images. Peng et al. [24] utilized the observed signals and pre-trained diffusion models to generate diverse solutions, employing accelerated coarse-to-fine Monte Carlo sampling to approximate optimal reconstruction results. Li et al. [14] designed an efficient diffusion model for multi-contrast MRI super-resolution. This model generates high-frequency detail priors in a compact latent space to reduce the number of iterations and employs a Prior-guided Large-window Transformer to avoid distortions.

3 Method

This paper proposes a novel MRI image SR method, as illustrated in Fig. 2. The overall framework consists of four key modules: a Wavelet Transform (WT) module, a frequency encoder-decoder, a Latent Diffusion Transformer (LDT [[23]), and an Inverse Wavelet Transform (Inverse WT) module.

Specifically, both the low-resolution (LR) and corresponding high-resolution (HR) images are first decomposed into four frequency sub-bands—one low-frequency and three high-frequency components—via wavelet transform. These sub-bands exhibit explicit frequency structures, which facilitate accurate modeling of structural edges and texture details in the frequency domain. We design and train a frequency autoencoder to extract latent representations of these sub-bands and to perform conditional decoding. Subsequently, a LDT is trained to generate the high-frequency latent representations conditioned on the low-frequency latent representation. It is noteworthy that,

unlike diffusion-based MRI SR methods that mainly learn mappings and sampling between LR and HR spaces—often at the cost of prior fidelity—our model directly learns to generate real high-frequency components. During the reconstruction phase, the discriminator-constrained super-resolution-aware decoder conditions on low-frequency latent features to jointly decode both low- and high-frequency features. The final SR image is obtained through the inverse wavelet transform. Given that LR images typically retain most of the low-frequency information while losing crucial high-frequency details in the frequency domain, our method capitalizes on this property. By extracting low-frequency information through the autoencoder and leveraging the LDT to faithfully recover realistic high-frequency details, our approach achieves higher-quality reconstruction results.

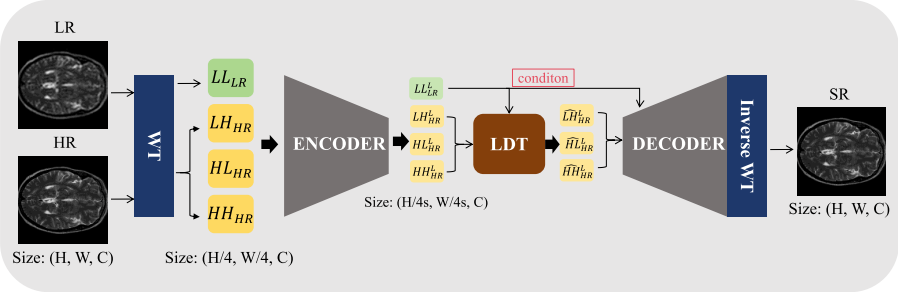


Fig. 2. Overall framework of the algorithm. WT denotes the Wavelet Transform. LDT refers to the Latent Diffusion Transformer model. s is the downsampling scale hyperparameter of the Autoencoder.

3.1 Wavelet Transform and Frequency Autoencoder

To effectively separate structural information from textural details in an MRI image, we employ the Discrete Wavelet Transform (DWT) to perform frequency-domain decomposition of the input image. DWT decomposes the image into four sub-bands: one low-frequency sub-band (LL) and three high-frequency sub-bands (LH, HL, and HH), which correspond to the image's contours, vertical edges, horizontal edges, and diagonal details, respectively. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, the wavelet decomposition can be formulated as follows:

$$\mathcal{W}(I) = \{LL, LH, HL, HH\} \tag{1}$$

where $\mathcal{W}(\cdot)$ denote the wavelet transform operator and $\mathcal{W}(I) \in H/4 \times W/4 \times C$. We apply wavelet decomposition to the low-resolution image $\mathbf{I}_{\mathrm{LR}}$ and the high-resolution image $\mathbf{I}_{\mathrm{HR}}$, resulting in the low-frequency sub-band LL_{LR} for the LR image and the high-frequency sub-bands $\{LH_{\mathrm{HR}}, HL_{\mathrm{HR}}, HH_{\mathrm{HR}}\}$ for the HR image. These high-frequency sub-bands serve as inputs to the subsequent decoder and diffusion model, facilitating multi-scale structural modeling of the image across different frequency levels. After obtaining the high-resolution frequency components, we construct Frequency Autoencoder to extract latent representations for each sub-band. The encoder

architecture is inspired by the structure of the Stable Diffusion model [26], consisting of multiple convolutional layers, residual connections, and attention mechanisms. Unlike the original Stable Diffusion model, which encodes images into a 4-channel latent representation, we modify the output to produce 3 channels to better align with the inherent structure of natural images. To reduce the overall model size and computational cost, we decrease both the depth and width of the network while preserving its multi-scale feature extraction capability, thereby enhancing the model's efficiency and suitability for lightweight deployment. The decoder mirrors the encoder in structure, with an input channel size three times that of the encoder and additional conditioning information. This condition is first converted to grayscale and then injected into the final convolutional block to provide structural guidance. The encoder, denoted as $\mathcal{E}(\cdot)$, takes the high-frequency sub-bands $\{LH_{\mathrm{HR}}, HL_{\mathrm{HR}}, HH_{\mathrm{HR}}\}$ as input and maps them into a latent space:

$$LL^{L}_{\mathrm{LR}}, LH^{L}_{\mathrm{HR}}, HL^{L}_{\mathrm{HR}}, HH^{L}_{\mathrm{HR}} = \mathcal{E}(LL_{\mathrm{LR}}), \mathcal{E}(LH_{\mathrm{HR}}), \mathcal{E}(HL_{\mathrm{HR}}), \mathcal{E}(HH_{\mathrm{HR}}) \tag{2}$$

The superscript L on the left-hand side of the equation denotes that the variables lie in the latent space. These latent features are subsequently fed into the LDT for further modeling. Compared to modeling directly in the image domain, processing high-frequency components in the frequency domain allows for more explicit separation of texture levels, thereby enabling the model to more effectively capture edge and detail information. During the decoding stage, the frequency decoder $\mathcal{D}(\cdot)$ maps the reconstructed latent features back to the frequency image space:

$$\{\widehat{LL}_{\mathrm{LR}}, \widehat{LH}_{\mathrm{HR}}, \widehat{HL}_{\mathrm{HR}}, \widehat{HH}_{\mathrm{HR}}\} = \mathcal{D}\big(\widehat{LL}^{L}_{LR}, \{\widehat{LH}^{L}_{HR}, \widehat{HL}^{L}_{HR}, \widehat{HH}^{L}_{HR}\}\big). \tag{3}$$

Finally, the reconstructed high-frequency sub-bands are combined with the low-frequency sub-band and passed into the Inverse DWT, denoted as $\mathcal{W}^{-1}$, to complete the image reconstruction:

$$\hat{\mathbf{I}}_{\mathrm{SR}} = \mathcal{W}^{-1}\big(\{\widehat{LL}_{\mathrm{LR}}, \widehat{LH}_{\mathrm{HR}}, \widehat{HL}_{\mathrm{HR}}, \widehat{HH}_{\mathrm{HR}}\}\big). \tag{4}$$

3.2 Latent Diffusion Modeling of High-Frequency Priors

In this work, the LDT model is from Peebles et al. [23], and the input and output dimensions have been changed. To model the complex distribution of high-frequency features in a compact latent space while ensuring high-fidelity details and structural consistency in the generated images, we introduce a latent conditional diffusion model that operates on the latent representations produced by a frequency-domain encoder. Diffusion models are a class of generative models that learn data distributions by simulating a gradual noising and denoising process, ultimately generating samples that resemble the training data. Our approach effectively captures the high-frequency information of real MRI images in the latent space, rather than learning a direct mapping from low-resolution (LR) images. Specifically, given the latent high-frequency subband representation of a high-resolution (HR) image, the diffusion model is used to learn the prior distribution of realistic high-frequency components. The diffusion

process constructs a Markov chain by progressively adding Gaussian noise to the latent representation over T steps. The forward diffusion process is defined as:

$$q(\mathbf{H}_t \mid \mathbf{H}_{t-1}) = \mathcal{N}\left(\mathbf{H}_t; \sqrt{1-\beta_t}\mathbf{H}_{t-1}, \beta_t \mathbf{I}\right) \tag{5}$$

where $\mathbf{H}_t = \{\widehat{LH}_{HR}^L, \widehat{HL}_{HR}^L, \widehat{HH}_{HR}^L\}_t$ represents the noised latent high-frequency representation. $\mathbf{H}_t = \mathcal{N}(\mathbf{H}_t; \sqrt{\bar{\alpha}_t}\mathbf{H}_0, (1-\bar{\alpha}_t)\mathbf{I})$ is directly sampled (noised) from the original latent variable at an arbitrary time step t, where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. Here β_t denotes the noise variance at step t, which typically increases according to a predefined linear or cosine schedule. We follow the setting used in Guided Diffusion [1]. And $\mathcal{N}$ represents a normal distribution, and $\mathbf{I}$ is the identity covariance matrix.

During the reverse generation stage, we train a neural network ϵ_θ to predict the noise ϵ,

$$\epsilon_\theta(\mathbf{H}_t, t, c): \mathbb{R}^{H\times W\times 3C} \times \mathbb{N} \times \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times 3C} \tag{6}$$

where c is the conditional information, integrated into LDT through cross-attention. In our method, c corresponds to the low-frequency sub-band LL_{LR}^L, which serves to guide the generation process towards better structural alignment and the synthesis of corresponding high-frequency components. Using the predicted noise estimation $\hat{\epsilon}$, we perform the reverse diffusion process according to the following formulation:

$$p_\theta(\mathbf{H}_{t-1} \mid \mathbf{H}_t, c) = \mathcal{N}(\mathbf{H}_{t-1}; \mu_\theta(\mathbf{H}_t, t, c), \Sigma_\theta(t)) \tag{7}$$

where Σ_θ is the learnable variance [1]. The reconstruction begins by sampling $\mathbf{H}_T$ from a standard Gaussian distribution and progressively denoises it through multiple steps to obtain $\widehat{\mathbf{H}}_0$. The final output $\widehat{\mathbf{H}}_0 = \{\widehat{LH}_{HR}^L, \widehat{HL}_{HR}^L, \widehat{HH}_{HR}^L\}$ is then used by the subsequent frequency decoder for image reconstruction.

3.3 Training Losses

We first train a frequency-domain autoencoder, and then train a latent diffusion model. The frequency-domain autoencoder is optimized using a combination of the following losses: frequency-domain loss, pixel-space loss, cycle consistency loss, and KL divergence loss. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{freq}} + \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{cycle}} + \beta\mathcal{L}_{\text{KL}} \tag{8}$$

$$\mathcal{L}_{\text{freq}} = \left\|\{\widehat{LL}_{\text{LR}}, \widehat{LH}_{\text{HR}}, \widehat{HL}_{\text{HR}}, \widehat{HH}_{\text{HR}}\} - \{LL_{\text{HR}}, LH_{\text{HR}}, HL_{\text{HR}}, HH_{\text{HR}}\}\right\|_2^2$$

$$\mathcal{L}_{\text{pixel}} = \left\|I_{\text{GT}} - \hat{I}_{\text{SR}}\right\|_2^2$$

$$\mathcal{L}_{\text{cycle}} = \left\|\mathcal{W}(\hat{I}_{\text{SR}}) - \{LL_{\text{HR}}, LH_{\text{HR}}, HL_{\text{HR}}, HH_{\text{HR}}\}\right\|_2^2$$

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q(\{LL_{\text{LR}}^L, LH_{\text{HR}}^L, HL_{\text{HR}}^L, HH_{\text{HR}}^L\} \mid \{LL_{\text{LR}}, LH_{\text{HR}}, HL_{\text{HR}}, HH_{\text{HR}}\}) \parallel \mathcal{N}(0, I))$$

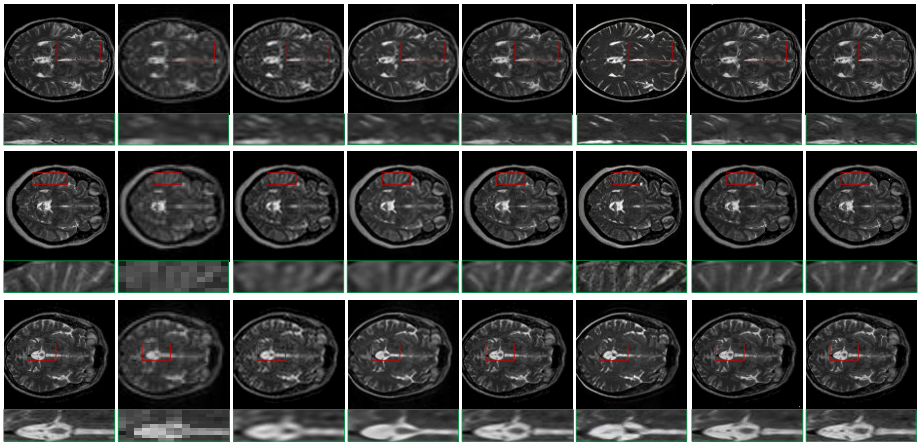
The training objective of the LDT is to minimize the error between the predicted noise and the real noise, as well as the predicted variance, using the loss function from Dhariwal et al [1]:

$$\mathcal{L}_{diff} = w_1 \cdot \| \epsilon - \epsilon_\theta(H_t, t, c) \|^2 + w_2 \cdot D_{\mathrm{KL}}(q(H_{t-1}|H_t, H_0) \parallel p_\theta(H_{t-1}|H_t)) \quad (9)$$

4 Experiments

Methods	Params	FLOPs	Time	IXI (4 ×)			FastMRI (4 ×)			Clinical (8 ×)		
				PSNR ↑	SSIM ↑	FID ↓	PSNR ↑	SSIM ↑	FID ↓	PSNR ↑	SSIM ↑	FID ↓
BICUBIC	-	-	-	23.38	0.6537	150.22	21.46	0.6124	163.61	16.58	0.5502	210.61
Discriminative Models												
MINet [4]	10.6M	-	-	28.76	0.7738	112.58	25.23	0.7304	123.01	20.82	0.6967	146.47
McMRSR [13]	12.5M	-	-	30.84	0.8413	103.41	28.73	0.7983	111.03	22.99	0.6996	132.06
MambaSR [9]	52.4M	-	-	34.64	0.9019	89.56	36.91	0.8771	99.52	29.56	0.6996	101.51
TransMRSR [8]	39.9M	-	-	35.60	0.9128	90.56	36.92	0.9036	96.52	30.31	0.8385	96.52
Generative Models												
ESRGAN [32]	16.7M	9.4G	59s	32.51	0.8423	79.06	31.73	0.8222	86.08	28.62	0.7783	92.62
DISGAN [31]	19.4M	9.7G	85s	29.65	0.7372	76.41	30.40	0.7283	83.41	29.09	0.7922	91.88
Disc-Diff [21]	86.2M	44.1G	1678s	33.87	0.8973	75.59	30.63	0.8601	83.64	30.63	0.8196	83.64
Ours	**46.3M**	**17.4G**	**47s**	**35.73**	**0.9276**	**63.21**	**35.06**	**0.9074**	**68.65**	**31.11**	**0.8442**	**77.32**

Table 1. Comparison of different SR methods across IXI (4 ×), FastMRI (4 ×), and Clinical (8 ×) datasets.



(a) GT (b) LR (c) BICUBIC (d) MambaSR (e) TransMRSR (f) DISGAN (g) DisC-Diff (h) Ours

Fig. 3. Visualization of Super-Resolution results on the IXI dataset at a scale factor of 4 for different models. Please zoom in for better perceptual quality.

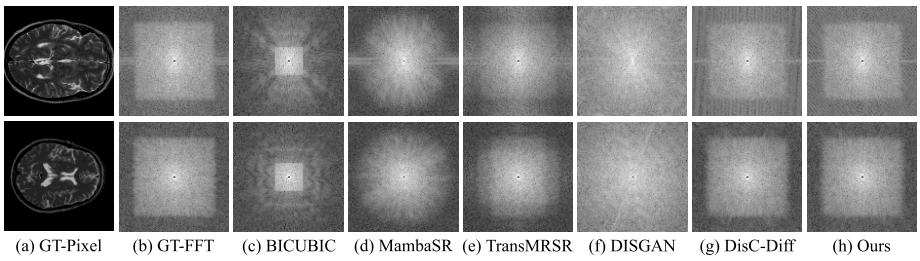


Fig. 4. Comparison of frequency domain visualization using different methods

4.1 Datasets and Evaluation Metrics

To better model high-frequency priors, we utilized two publicly available datasets and one clinical dataset, totaling 10,000 training images. The public datasets employed were IXI[3] and FastMRI [40]:

IXI: This is a widely used brain MRI dataset comprising multiple MRI sequences such as T1-, T2-, and PD-weighted images. The dataset features high image quality and diverse scanning protocols, effectively capturing image characteristics under varying acquisition conditions. The IXI dataset includes approximately 600 healthy adult brain MRI scans, each containing multiple slices. Following the protocol established by DisC-Dif [21], 578 healthy brain MRI images from the IXI dataset were partitioned as follows: 500 for training, 6 for validation, and 70 for testing[4].

Bands (%)	Ours	Disc-Diff	Trans-MRSR	Mam-baSR	McMRSR	MINet	DISGAN	ESRGAN	BICUBIC
0-10	38.26	37.92	37.76	36.57	36.21	35.84	35.11	35.22	34.56
10-20	38.04	37.65	37.54	36.24	35.98	35.72	35.27	35.18	34.42
20-30	37.99	37.53	37.32	36.02	35.36	35.35	34.99	34.23	34.26
30-40	37.57	37.22	37.06	36.21	35.86	35.46	35.29	35.16	34.28
40-50	37.12	36.94	36.62	35.77	35.17	34.57	34.16	34.21	33.94
50-60	37.08	36.98	36.24	35.31	35.27	34.33	34.11	34.17	33.88
60-70	37.56	36.46	35.96	35.03	34.91	34.86	34.37	33.95	33.72
70-80	36.54	36.34	35.54	34.65	33.91	33.77	33.18	34.72	33.05
80-90	37.91	36.81	35.21	34.07	33.21	34.57	34.17	32.82	32.74
90-100	37.73	36.50	34.80	34.11	33.92	33.06	35.03	34.75	32.85

Table 2. The recovery quality of different methods on 10 frequency bands (PSNR)

FastMRI: Created collaboratively by Facebook AI Research (FAIR) and NYU Langone Health, FastMRI is a large-scale medical imaging dataset. From the training subset, the first 3,500 images were used for training, the first 200 images for validation, and the first 200 images from the test set for testing.

[3] https://brain-development.org/ixi-dataset/
[4] https://bit.ly/3yethO4

We also leveraged a clinical cranial MRI dataset for super-resolution training and evaluation. This dataset contains multi-contrast brain MRI scans, with T2-weighted (T2W) images serving as high-quality references and diffusion-weighted imaging (DWI) as low-resolution target modalities. Axial 2D slices were extracted from 3D clinical scans, totaling 7,000 slices: 6,000 from healthy subjects and 1,000 from patients. Among these, 6,000 slices were used for training, while the remaining 1,000 slices were evenly split between validation and testing. All original scans were acquired on a Siemens 3T Prisma MRI scanner equipped with a 64-channel head coil. The 3D scans underwent standard preprocessing steps including skull stripping (FSL-BET), z-score intensity normalization, and affine registration. For supervised super-resolution training, all 3D scans were resampled into 2D axial slices and saved in PNG format. High-resolution DWI slices were synthetically downsampled using bicubic interpolation to generate paired low-resolution (LR) and high-resolution (HR) images. These LR-HR pairs were used to train conventional super-resolution models. Additionally, T2W images were provided as optional auxiliary inputs to facilitate reference-guided super-resolution frameworks. Due to dataset copyright restrictions, visualization or display of test samples is not permitted; instead, fair quantitative comparisons of different methods were conducted. We choose the commonly used pixel-level metrics PSNR and SSIM, as well as the perceptual quality metric FID, for comprehensive quantification.

4.2 Implementation Details

Our model training is conducted in two stages: (1) pretraining a frequency autoencoder that serves as the backbone for latent representations; and (2) training a LDT. All data are center-cropped to a resolution of $224 \times 224 \times 3$.

In the first stage, we employ the Adam optimizer [43] with an initial learning rate of 1×10^{-4}, following a cosine annealing schedule with a minimum learning rate of 1×10^{-6}. The batch size is set to 32 per GPU, and training is performed on four NVIDIA RTX 3090 GPUs for 200 epochs, taking approximately 14 hours. The encoder and decoder are jointly trained during this stage. All input images are normalized to the [0, 1] range prior to training. To ensure training stability, we apply gradient clipping with a maximum norm of 1.0, and incorporate a small weight decay of 1×10^{-5} to mitigate overfitting. The autoencoder uses a downsampling factor of $s = 4$, and the KL divergence term is weighted by $\beta = 1 \times 10^{-6}$.

In the second stage, the pretrained autoencoder is kept frozen, and the diffusion process is modeled over 1000 timesteps using a cosine noise schedule [1], where the objective is to predict the added Gaussian noise. We adopt the AdamW optimizer with a learning rate of 1×10^{-4}, employing linear warm-up over the first 10,000 steps, followed by cosine decay. The diffusion loss weights are set to $w_1 = 1$ and $w_2 = 1 \times 10^{-4}$. The batch size is set to 16 per GPU, and training is conducted on the same 4-GPU setup for 300 epochs, which takes approximately 20 hours. During this phase, only the parameters of the LDT are updated, while the encoder and decoder weights remain fixed.

4.3 Results

To comprehensively evaluate the effectiveness of our method, we present qualitative visual comparisons (Fig. 3) and quantitative metrics (Table 1) across three datasets at scales of ×4, ×4, and ×8, respectively.

Qualitative Results: Our method demonstrates superior visual performance over both discriminative and generative baselines in reconstructing anatomical structures and recovering high-frequency details. In particular, it excels at preserving edge sharpness and maintaining structural consistency in regions of interest (see zoomed-in areas). Discriminative models such as MambaSR [9] and TransMRSR [8] often produce overly smooth or distorted outputs, while generative approaches tend to introduce hallucinated or incorrect fine details. In contrast, our approach shows clear advantages in both high-frequency generation and low-frequency contour preservation.

Quantitative Results: Table 1 provides a detailed comparison in terms of PSNR, SSIM, and FID scores. Our method achieves the best or highly competitive performance across all datasets. Specifically, it obtains the highest PSNR and SSIM scores on both the FastMRI and clinical datasets while maintaining a low FID, indicating superior perceptual quality. Although our model is not optimal in terms of parameter count due to the use of a diffusion-based framework and autoencoder architecture, it remains competitive in inference speed. Compared with the state-of-the-art diffusion-based method Disc-Diff [21], our approach achieves a balanced trade-off between lightweight design and overall performance.

Table 3. Ablation study on the impact of different design choices of LDT. The default setting uses ϵ prediction, cosine sampling, a downsampling ratio of 1/4, and Cross-Attention for conditioning.

ID	Prediction Target	Sampling Strategy	Downsampling	Conditioning Method	PSNR ↑	FID ↓
0	ε	Cosine	1/4	Cross-Attn.	35.73	63.21
1	x_0	Cosine	1/4	Cross-Attn.	34.85	68.92
2	x_0	Linear	1/4	Cross-Attn.	35.42	66.10
3	x_0	Cosine	1/8	Cross-Attn.	34.30	70.33
4	x_0	Cosine	1/4	Cat	35.10	64.50
5	ε	Linear	1/4	Cat	33.90	71.75
6	ε	Linear	1/8	Cross-Attn.	33.10	76.20
7	x_0	Linear	1/8	Cat	34.02	69.30
8	ε	Cosine	1/8	Cat	33.55	72.60

4.4 Ablation Study

To better understand the contribution of each component in our proposed method, we conducted a series of systematic ablation studies. The default configuration adopts the ϵ prediction target, cosine noise scheduling, a 1/4 downsampling ratio (as determined by the autoencoder), and a Cross-Attention-based conditioning mechanism. In each ablation experiment, we modify only a single component to isolate its individual effect on generation quality. The results are summarized in Table 3.

Prediction Target: We compared two prediction targets: the noise vector ϵ and the original signal x_0. As shown by the comparison between ID 0 and ID 1, using ϵ as the prediction target yields better performance in both PSNR (35.73 vs. 34.85) and FID (63.21 vs. 68.92). This is consistent with findings in previous diffusion model research, indicating that predicting noise tends to be more stable and effective under the presence of sampling noise.

Sampling Strategy: We evaluated two noise scheduling strategies: cosine and linear. Cosine scheduling consistently outperforms linear scheduling. For instance, comparing ID 0 and ID 5, cosine scheduling improves PSNR by 1.83 and reduces FID by 8.54. Similar trends are observed in the comparisons between ID 1 vs. ID 2 and ID 7 vs. ID 3, demonstrating that cosine scheduling is more advantageous in maintaining high image quality.

Downsampling Ratio: We examined the impact of spatial downsampling ratios: 1/4 vs. 1/8. While a larger downsampling ratio (e.g., 1/8) reduces computational cost, it often degrades image quality. For example, reducing the ratio from 1/4 to 1/8 (ID 0 vs. ID 6) results in a PSNR drop from 35.73 to 33.10 and an FID increase from 63.21 to 76.20. This indicates that a 1/4 ratio offers a better trade-off between efficiency and quality. Moreover, considering the downsampling already introduced by wavelet transforms, the final effective ratio is 1/16.

Conditioning Method: We compared two conditioning mechanisms: Cross-Attention and Concatenation. Cross-Attention consistently outperforms Concatenation across multiple experiments. For example, changing only the conditioning method between ID 0 and ID 5 leads to a PSNR drop of 1.83 and an FID increase of 8.54, highlighting the effectiveness of attention mechanisms in modeling structured relationships between conditions.

Combined Effects: When multiple suboptimal design choices are combined, performance degradation becomes more pronounced. For example, although ID 8 employs ϵ prediction and cosine sampling, it uses a 1/8 downsampling ratio and Concatenation for conditioning. As a result, PSNR drops to 33.55 and FID increases to 72.60. This underscores the importance of a well-coordinated design among different model components.

5 Conclusion

In this work, we proposed a novel super-resolution (SR) reconstruction framework for multi-contrast magnetic resonance imaging (MRI) by integrating the complementary strengths of generative and discriminative models. Our approach leverages a latent diffusion model (LDM) to effectively synthesize realistic high-frequency details while employing an autoencoder conditioned on low-resolution (LR) inputs to guide the reconstruction process with stable structural information. Experimental evaluations across multiple benchmark datasets and metrics demonstrate that our method achieves superior quantitative performance compared to state-of-the-art SR techniques, while also offering a more efficient and lightweight architecture. Qualitative results further confirm the enhanced ability of our approach to preserve fine-grained anatomical

structures, addressing the over-smoothing issues commonly encountered in conventional deep learning-based SR methods.

References

1. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems 34, 8780–8794 (2021)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. pp. 184–199. Springer (2014)
3. Du, W., Tian, S.: Transformer and gan-based super-resolution reconstruction network for medical images. Tsinghua Science and Technology 29(1), 197–206 (2023)
4. Feng, C.M., Fu, H., Yuan, S., Xu, Y.: Multi-contrast mri super-resolution via a multi-stage integration network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 140–149. Springer (2021)
5. Feng, C.M., Yan, Y., Fu, H., Chen, L., Xu, Y.: Task transformer network for joint mri reconstruction and super-resolution. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 307–317. Springer (2021)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM 63(11), 139–144 (2020)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840–6851 (2020)
8. Huang, S., Liu, X., Tan, T., Hu, M., Wei, X., Chen, T., Sheng, B.: Transmrsr: transformer-based self-distilled generative prior for brain mri super-resolution. The Visual Computer 39(8), 3647–3659 (2023)
9. Ji, Z., Zou, B., Kui, X., Vera, P., Ruan, S.: Deform-mamba network for mri super-resolution. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 242–252. Springer (2024)
10. Jiang, M., Zhi, M., Wei, L., Yang, X., Zhang, J., Li, Y., Wang, P., Huang, J., Yang, G.: Fagan: Fused attentive generative adversarial networks for mri image super-resolution. Computerized Medical Imaging and Graphics 92, 101969 (2021)
11. Kang, L., Tang, B., Huang, J., Li, J.: 3d-mri super-resolution reconstruction using multi modality based on multi-resolution cnn. Comput. Methods Programs Biomed. 248, 108110 (2024)
12. Kara, D., Kimura, T., Liu, S., Li, J., Liu, D., Wang, T., Wang, R., Chen, Y., Hu, Y., Abdelzaher, T.: Freqmae: Frequency-aware masked autoencoder for multi-modaliot sensing. In: Proceedings of the ACM Web Conference 2024. pp. 2795–2806 (2024)
13. Li, G., Lv, J., Tian, Y., Dou, Q., Wang, C., Xu, C., Qin, J.: Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20636–20645 (2022)
14. Li, G., Rao, C., Mo, J., Zhang, Z., Xing, W., Zhao, L.: Rethinking diffusion model for multi-contrast mri super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11365–11374 (2024)

15. Li, K.R., Wu, A.G., Tang, Y., He, X.P., Yu, C.L., Wu, J.M., Hu, G.Q., Yu, L.: The key role of magnetic resonance imaging in the detection of neurodegenerative diseases-associated biomarkers: a review. Molecular Neurobiology 59(10), 5935–5954 (2022)
16. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
17. Luo, Z., Huang, H., Yu, L., Li, Y., Fan, H., Liu, S.: Deep constrained least squares for blind image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17642–17652 (2022)
18. Lv, J., Zhu, J., Yang, G.: Which gan? A comparative study of generative adversarial network-based fast MRI reconstruction. Philosophical Transactions of the Royal Society A 379(2200), 20200203 (2021)
19. Lyu, Q., Shan, H., Steber, C., Helis, C., Whitlow, C., Chan, M., Wang, G.: Multi-contrast super-resolution MRI through a progressive network. IEEE transactions on medical imaging 39(9), 2738–2749 (2020)
20. Ma, J., Jian, G., Chen, J.: Diffusion model-based MRI super-resolution synthesis. International Journal of Imaging Systems and Technology 35(2), e70021 (2025)
21. Mao, Y., Jiang, L., Chen, X., Li, C.: Disc-diff: Disentangled conditional diffusion model for multi-contrast MRI super-resolution. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 387–397. Springer (2023)
22. Nunna Jr, B., Parihar, P., Wanjari, M., Shetty, N., Bora, N., NUNNA Jr, B., PARTHAR, P., Shetty, N.D., BORA, N.: High-resolution imaging insights into shoulder joint pain: a comprehensive review of ultrasound and magnetic resonance imaging (MRI). Cures 15(11) (2023)
23. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
24. Peng, C., Guo, P., Zhou, S.K., Patel, V.M., Chellappa, R.: Towards performant and reliable undersampled MR reconstruction via diffusion model sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 623–633. Springer (2022)
25. Pham, C.H., Tor-Díez, C., Meunier, H., Bednarek, N., Fablet, R., Passat, N., Rousseau, F.: Multiscale brain MRI super-resolution using deep 3D convolutional networks. Computerized Medical Imaging and Graphics 77, 101647 (2019)
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
27. Qiao, S., Yang, J., Zhang, T., Zhao, C.: Layered input gradient for image denoising. Knowledge-Based Systems 254, 109587 (2022)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
29. Shende, P., Pawar, M., Kakde, S.: A brief review on: MRI images reconstruction using GAN. In: 2019 International Conference on Communication and Signal Processing (ICCSP). pp. 0139–0142. IEEE (2019)
30. Van Reeth, E., Tham, I.W., Tan, C.H., Poh, C.L.: Super-resolution in magnetic resonance imaging: a review. Concepts in Magnetic Resonance Part A 40(6), 306–325 (2012)
31. Wang, Q., Mahler, L., Steiglechner, J., Birk, F., Scheffler, K., Lohmann, G.: Dis-gan: Wavelet-informed discriminator guides gan to MRI super-resolution with noise cleaning. In:

Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2452–2461 (2023)
32. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
33. Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence 43(10), 3365–3387 (2020)
34. Weiskopf, N., Edwards, L.J., Helms, G., Mohammadi, S., Kirilina, E.: Quantitative magnetic resonance imaging of brain anatomy and in vivo histology. Nature Reviews Physics 3(8), 570–588 (2021)
35. Xie, Y., Li, Q.: Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–664. Springer (2022)
36. Yang, J., Dai, T., Zhu, Y., Li, N., Li, J., Xia, S.T.: Diffusion prior interpolation for flexibility real-world face super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9211–9219 (2025)
37. Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q.: Deep learning for single image super-resolution: A brief review. IEEE Transactions on Multimedia 21(12), 3106–3121 (2019)
38. Yu, M., Guo, M., Zhang, S., Zhan, Y., Zhao, M., Lukasiewicz, T., Xu, Z.: RIR-GAN: An end-to-end lightweight multi-task learning method for brain MRI super-resolution and denoising. Computers in Biology and Medicine 167, 107632 (2023)
39. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L.: Image super-resolution: The techniques, applications, and future. Signal processing 128, 389–408 (2016)
40. Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M.J., De-fazio, A., Stern, R., Johnson, P., Bruno, M., et al.: fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv preprint arXiv:1811.08839 (2018)
41. Zhang, K., Hu, H., Philbrick, K., Conte, G.M., Sobek, J.D., Rouzrokh, P., Erickson, B.J.: Soup-GAN: Super-resolution MRI using generative adversarial networks. Tomography 8(2), 905–919 (2022)
42. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
43. Zhang, Z.: Improved Adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS). pp. 1–2. IEEE (2018)
44. Zhao, X., Yang, T., Li, B., Zhang, X.: SwingAN: A dual-domain swin transformer-based generative adversarial network for MRI reconstruction. Computers in Biology and Medicine 153, 106513 (2023)