Rethinking Diffusion Model for Multi-Contrast MRI Super-Resolution

Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing*, Lei Zhao*
College of Computer Science and Technology, Zhejiang University, China
{cslgy, raochen, csmjc, cszzj, wxing, cszhl}@zju.edu.cn

Abstract

Recently, diffusion models (DM) have been applied in magnetic resonance imaging (MRI) super-resolution (SR) reconstruction, exhibiting impressive performance, especially with regard to detailed reconstruction. However, the current DM-based SR reconstruction methods still face the following issues: (1) They require a large number of iterations to reconstruct the final image, which is inefficient and consumes a significant amount of computational resources. (2) The results reconstructed by these methods are often misaligned with the real high-resolution images, leading to remarkable distortion in the reconstructed MR images. To address the aforementioned issues, we propose an efficient diffusion model for multi-contrast MRI SR, named as DiffMSR. Specifically, we apply DM in a highly compact low-dimensional latent space to generate prior knowledge with high-frequency detail information. The highly compact latent space ensures that DM requires only a few simple iterations to produce accurate prior knowledge. In addition, we design the Prior-Guide Large Window Transformer (PLWformer) as the decoder for DM, which can extend the receptive field while fully utilizing the prior knowledge generated by DM to ensure that the reconstructed MR image remains undistorted. Extensive experiments on public and clinical datasets demonstrate that our DiffMSR[1] outperforms state-of-the-art methods.

1. Introduction

Magnetic resonance imaging (MRI) is a widely utilized in clinical imaging technology as it can provide clear information on tissue structure and function while being non-invasive and radiation-free. However, obtaining high-resolution (HR) magnetic resonance (MR) images is challenging due to acquisition limitations [5, 16]. Super-resolution (SR) technology can handle this challenge by reconstructing low-resolution (LR) images into their corresponding high-resolution (HR) version. Traditional methods typically utilized model-based [26, 30] and learning-based [40, 43] SR reconstruction approaches. Nonetheless, these methods have insufficient reconstruction capabilities under high upsampling factors due to the complicated anatomical structures in MR images[20].

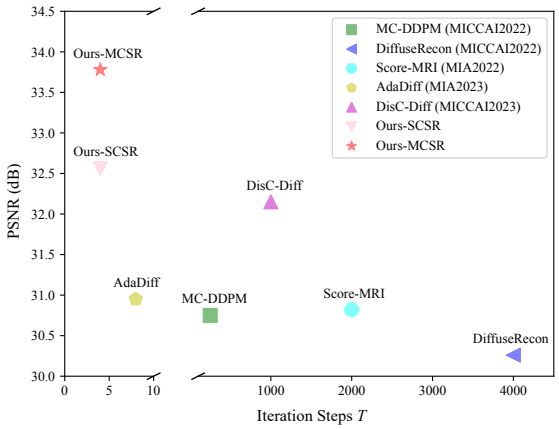

Figure 1. Comparison with DM-based MRI reconstruction methods on FastMRI dataset. Note that the experimental settings used by these methods are the same as those in Sec. 4. As can be seen, our method has the best reconstruction metric and it only requires 4 iteration steps. Note that DiffuseRecon [32], MC-DDPM [39], Score-MRI [2], and AdaDiff [9] are employed for single-contrast SR (SCSR) reconstruction. DisC-Diff [27] is specifically designed for multi-contrast SR (MCSR) reconstruction.

Following the advent of deep learning, convolutional neural networks (CNNs) have found extensive application in MRI SR reconstruction tasks [4, 13, 25, 35, 44]. Besides, Transformer-based approaches [6, 11, 17, 18, 20, 23, 24] are introduced as an alternative to CNNs for modeling long-range dependencies in MR images. While these approaches enhance the performance of SR reconstruction, they tend to yield images that lack detail [34], and for MR images with complicated anatomical structures, they fail to reconstruct some high-frequency details satisfactorily. Apart from CNN- and Transformer-based methods, deep generative models like generative adversarial networks (GANs) [8] provide different strategies for creating intri-

* Corresponding author
[1] Code: https://github.com/GuangYuanKK/DiffMSR

cate details. Very recently, diffusion models (DMs) [7, 10] have shown impressive performance in MR image synthesis [12, 28, 29, 41], reconstruction tasks [9, 27, 32, 37, 39]. DMs generate high-fidelity images through a stochastic iterative denoising process employing a pure white Gaussian noise. In contrast to GANs, DMs yield a more accurate target distribution without facing optimization instability or mode collapse issues.

However, DM-based approaches encounter several issues when applied to MRI reconstruction. On the one hand, DMs are required to perform a large number of iteration steps to generate samples, aiming to simulate the accurate details of the data. However, unlike MRI synthesis tasks that generate every pixel from scratch, MRI SR reconstruction tasks only require adding details on the given LR MR images. Therefore, in theory, the SR reconstruction tasks only require a small number of iterations to generate satisfactory results, which will greatly save computational resources [1, 38]. On the other hand, DMs tend to introduce artifacts in the generated MR images that are not present in the original HR images. Furthermore, some details of complicated anatomical structures may also be misaligned with the real target, resulting in image distortion.

These challenges lead us to rethink DM for MRI super-resolution reconstruction from a new perspective. First, we consider compressing the latent features in the DM into a low-dimensional latent space to reduce computational complexity and the number of iteration steps. Second, given the advantage of the Transformer in capturing long-range dependencies, we integrated DM and Transformers to solve the distortion problem of DM. Third, MRI can acquire images with different contrasts but the same anatomical structure by adjusting scanning parameters. Numerous studies [4, 11, 17–20, 23, 25, 27] have demonstrated that utilizing HR T1 contrast images with shorter scan times can offer valuable supplementary information for LR T2 contrast images with longer scan times. Therefore, we integrate HR T1 contrast images as reference images into the network.

Based on the above motivations, we propose an efficient diffusion model for multi-contrast MRI SR reconstruction, which it integrates the Prior-Guide Large Window Transformer (PLWformer) as the decoder. We call it *DiffMSR*. PLWformer can benefit from large window self-attention while having less computational burden, effectively enhancing the performance of DM in terms of generating accurate details. Following previous practice [1, 3, 38], we divide the training process into two stages to achieve prior extraction and training of the DM. In the first stage, we employ the Prior Extraction (PE) module to compress the original HR MR image into highly compact latent features as prior knowledge, which is utilized to guide the PLWformer. To fully utilize the prior knowledge, we design Prior-Guide Large Window Multi-Head Self-Attention (PLMSA) and Prior-Guide Feed-Forward Network (PG-FFN) in PLWformer to utilize the compressed prior knowledge while expanding the attention receptive field. We jointly optimize the PE and PLWformer to effectively obtain reliable prior knowledge. In the second stage, we employ the PE module trained in the first stage to train the DM, allowing it to generate latent prior features in the latent space from Gaussian noise to guide the PLWformer for accurate reconstruction. Since the latent prior features are low-dimensional, the DM can estimate accurate details with simple iterations.

The main contributions are summarized as follows:

(1) We propose an efficient diffusion model for multi-contrast MRI SR, named as DiffMSR. Our method leverages diffusion models to generate effective prior knowledge, which is integrated into the SR process to reconstruct more satisfactory MR images with accurate details.

(2) We design the PLWformer with the aim of fully leveraging the prior knowledge generated by DM and having less computational burden while expanding the receptive field, which ensures that the reconstructed MR image is undistorted.

(3) Extensive experiments conducted on public and clinical datasets demonstrate the superior performance of the DiffMSR in comparison to state-of-the-art methods.

2. Related Works

2.1. Multi-Contrast MRI SR

Multi-contrast MRI super-resolution (MCSR) involves using multiple contrasts of MR images in the SR process. Specifically, MCSR utilizes T1 contrast images with shorter repetition time and echo time as reference images to provide valuable high-frequency information during SR for T2 contrast images with longer scan times. For instance, Li *et al*. [17] were the first to introduce Transformer in MCSR tasks and proposed the Transformer-empowered multi-scale contextual matching and aggregation network. Lyu *et al*. [23] introduced a texture-preserving branch and a contrastive constraint in MCSR. Lei *et al*. [15] proposed a decomposition-based variational network for MCSR task. Li *et al*. [20] designed the reference-aware implicit attention to extend MCSR to arbitrary scale upsampling. Mao *et al*. [27] proposed a disentangled conditional diffusion model for MCSR. The majority of the aforementioned MCSR methods primarily employ Transformer-based models, while few methods employ diffusion models to perform MCSR tasks.

2.2. MRI Diffusion Model

Diffusion models (DM) [10], as a type of probabilistic generative model, can construct the required data samples from Gaussian noise through a stochastic iterative denoising pro-

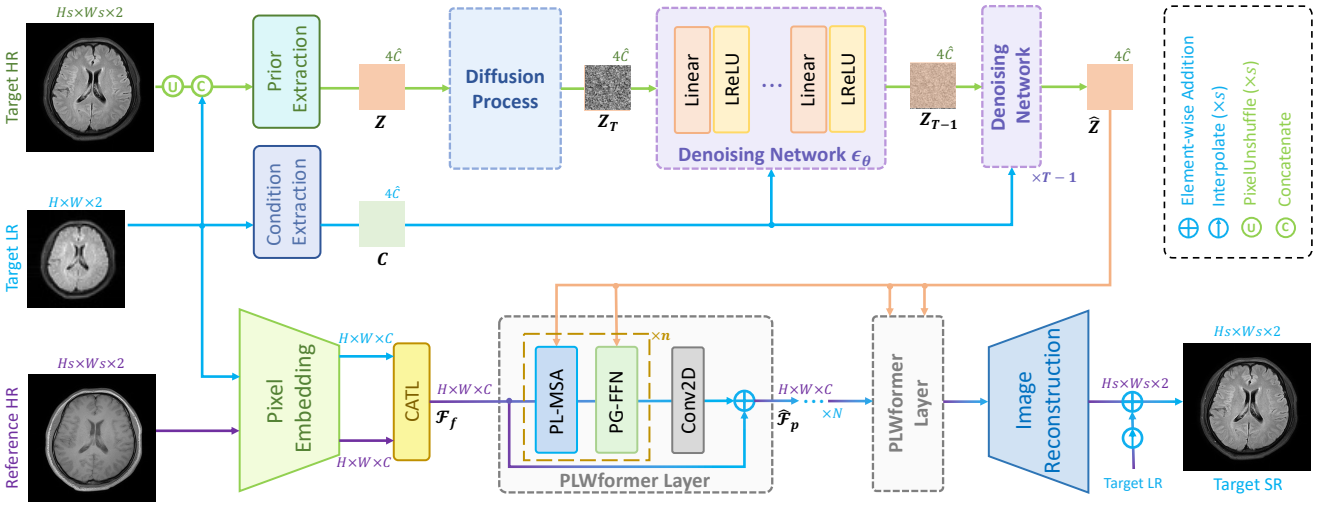

Figure 2. The overall architecture of our proposed DiffMSR, which mainly divided into two parts: (1) Diffusion Model; (2) Prior-guide Large Window Transformer (PLWformer) including N PLWformer layers and a image reconstruction module. CATL: Cross-Attention Transformer Layer.

cess. DM has exhibited impressive performance in MRI reconstruction tasks [2, 9, 27, 32, 39]. For instance, Xie *et al*. [39] proposed a measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. Peng *et al*. [32] introduced a novel diffusion model-based MR reconstruction method. Chung *et al*. [2] designed a score-based diffusion models for accelerated MRI. Mao *et al*. [27] designed a disentangled conditional diffusion model for multi-contrast brain MRI SR. Nonetheless, the above DM-based methods require a significant number of iteration steps to generate samples, which costs many computing resources. Additionally, they tend to introduce artifacts into the generated MR SR images that are not present in the original HR images. To address the above issues, we propose to combine DM and Transformer to reconstruct artifact-free and non-distorted MR images.

2.3. MRI Transformer

The Transformer architecture is widely applied in MCSR [17, 18, 20, 23] as it can model long-range dependencies effectively, allowing for the reconstruction of complicated anatomical structures in MR images. However, due to computational burdens, these methods typically limit the window size to 8×8. Expanding the window size can effectively increase the receptive field, capture longer-range dependencies, and enhance the quality of reconstructed images. Inspired by [45], to maintain a smaller computational burden while expanding the window size, we design a prior-guide large window transformer (PLWformer). PLWformer can utilize the prior knowledge generated by DM while possessing a larger receptive field to better reconstruct details.

3. Methodology

3.1. Overall Architecture

The overall architecture of the proposed DiffMSR is shown in Figure 2. As can be seen, DiffMSR mainly comprises two components: the diffusion model (DM) and the prior-guide large window transformer (PLWformer). We design DiffMSR with the goal of integrating DM and Transformer to overcome the shortcomings of DM and achieve better reconstruction results. Specifically, DM is employed to generate highly compact latent features as prior knowledge, which includes detailed information from real MR images, guiding the PLWformer for reconstruction to produce artifact-free and non-distorted MR images. Meanwhile, PLWformer utilizes large window attention to model longer-range dependencies in MR images.

We follow the previous practice [1, 3, 38] and split the training process into two stages, as shown in Figure 3. In the first stage, we utilize the Prior Extraction (PE) module to compress the target HR image $I_{HR} \in \mathbb{R}^{Hs \times Ws \times 2}$ into a highly compact latent space, obtaining prior knowledge $Z \in \mathbb{R}^{4\hat{C}}$, and employ it to guide the PLWformer for high-quality reconstruction. Here, H, W, and s represent the height, width, and upsampling scale. $\hat{C}$ denotes the latent feature channel number. We jointly optimize the PE and PLWformer to obtain more reliable prior knowledge Z. In the second stage, we first employ the PE module trained in the first stage to generate the target sample Z. Then, we train DM to learn to generate prior knowledge $\hat{Z} \in \mathbb{R}^{4\hat{C}}$ and subsequently utilize $\hat{Z}$ to enhance the reconstruction capabilities of the PLWformer. Additionally, we employ Pixel Embedding and Cross-Attention Transformer Layer

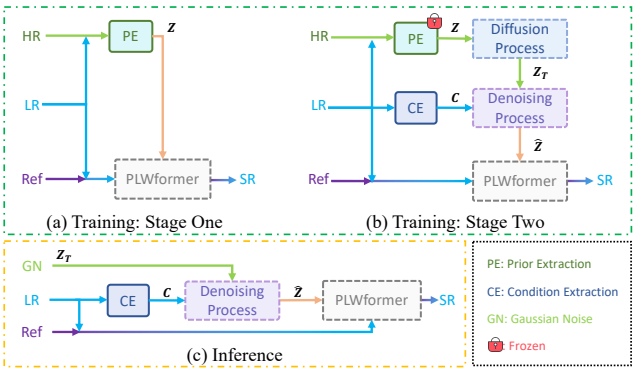

Figure 3. The process of training stage and inference stage.

(CATL) [18, 22] to fuse target LR image $I_{LR} \in \mathbb{R}^{H\times W\times 2}$ and reference HR image $I_{Ref} \in \mathbb{R}^{Hs\times Ws\times 2}$, obtaining $\mathcal{F}_f$ as input for the PLWformer.

3.2. Training: Stage One

The purpose of the first stage is to jointly train Prior Extraction (PE) and PLWformer to obtain more reliable prior knowledge Z, as shown in Figure 3(a). Hence, we primarily utilize two modules: PE and PLWformer. PE compresses HR images to obtain a compact representation Z, which serves as prior knowledge. PLWformer employs Z as a guide to model long-range dependencies using large window attention without increasing computational burden, thereby enhancing the accuracy of reconstructed details.

Prior Extraction. The structure of PE is primarily composed of residual blocks and linear layers stacked together. For more detailed architecture, please see *Supplementary Material*. Given the target LR image I_{LR} and its corresponding HR image I_{HR}, we first perform a PixelUnshuffle operation on I_{HR}, then concatenate them along the channel dimension and feed them into the PE to generate the prior knowledge Z:

$$Z = PE(Concat(PixelUnshuffle(I_{HR}), I_{LR})). \quad (1)$$

Since Z is a highly compact feature, this effectively reduces the computational burden for the subsequent DM.

PLWformer. The vast majority of Transformer-based MCSR methods [17, 18, 20, 23] employ multi-head self-attention (MSA) with a window size of 8×8. Appropriately expanding the window size of MSA in Transformer can effectively enhance reconstruction performance, but this will result in higher computational burden[45]. Inspired by [45], we introduce the Prior-Guide Large Window Transformer (PLWformer) to enjoy the benefits of the large window while utilizing prior knowledge. PLWformer consists of two components, the Prior-Guide Large Window Multi-Head Self-Attention (PL-MSA) and the Prior-Guide Feed Forward Network (PG-FFN), as shown in Figure 2. Next, we will introduce how PLWformer employs prior knowledge Z as guidance and how to expand the window size without increasing the computational burden.

After obtaining the prior knowledge Z through PE compression, it is input into PL-MSA and PG-FFN as dynamic modulation parameters to guide the reconstruction, as shown in Figure 4:

$$\begin{aligned} \mathcal{F}'_f &= \mathbb{L}(Z) \odot \mathbb{N}(\mathcal{F}_f) + \mathbb{L}(Z), \\ \mathcal{F}'_p &= \mathbb{L}(Z) \odot \mathbb{N}(\mathcal{F}_p) + \mathbb{L}(Z), \end{aligned} \quad (2)$$

where $\mathcal{F}'_f \in \mathbb{R}^{H\times W\times C}$, $\mathcal{F}'_p \in \mathbb{R}^{H\times W\times C}$, C means the channel dimension of the feature map, $\odot$ indicates element-wise multiplication, $\mathbb{N}$ denotes layer normalization, $\mathbb{L}$ represents linear layer.

In PL-MSA, $\mathcal{F}'_f$ is first splitted into N non-overlapping square windows $\mathcal{F}^{s\prime}_f \in \mathbb{R}^{NL^2\times C}$, where L is the length of each window. Then, linear layers $\mathbb{L}_Q$, $\mathbb{L}_K$, and $\mathbb{L}_V$ are used to embed $\mathcal{F}^{s\prime}_f$ to obtain $Q \in \mathbb{R}^{NL^2\times C}$, $K \in \mathbb{R}^{NL^2\times C/k^2}$, and $V \in \mathbb{R}^{NL^2\times C/k^2}$, where k is the token reduction factor. Next, permutation is applied to K and V, obtaining in permuted tokens $K_p \in \mathbb{R}^{NL^2/k^2\times C}$ and $V_p \in \mathbb{R}^{NL^2/k^2\times C}$. This way, a factor of k (k=2) reduces the window size for K and V, but their channel dimension is still unchanged to ensure the expressiveness of the attention map generated by each attention head [36]. Finally, self-attention (SA) computation is performed:

$$SA(Q, K, V) = softmax(QK'_p/\sqrt{d_k} + B)V_p, \quad (3)$$

where B is the aligned relative position embedding, K'_p is the transpose matrix of K_p. We reduce the channel dimensions of the K and V matrices and employ permutation operations to transfer some spatial information to the channel dimensions, which effectively reduces the computational burden in the case of large windows.

In PL-FFN, we first employ 1×1 convolution to aggregate information from different channels. Then, 3×3 depthwise convolution is utilized to aggregate information from spatially adjacent pixels, and a gating mechanism is employed to enhance information encoding:

$$\hat{\mathcal{F}}_p = Conv(GELU(\Phi(\mathcal{F}'_p)) \odot \Phi(\mathcal{F}'_p)) + \mathcal{F}_p, \quad (4)$$

where Φ means 1×1 convolution and 3×3 depthwise convolution.

Optimization function. To generate more reliable prior knowledge, we jointly optimize PE and PLWformer in the first training stage using an image-domain reconstruction loss $\mathcal{L}_{img}$ and frequency-domain data consistency loss $\mathcal{L}_{dc}$:

$$\mathcal{L}^1_{stage} = \lambda_1\mathcal{L}_{img} + \lambda_2\mathcal{L}_{dc}, \quad (5)$$

where $\mathcal{L}_{img} = \|I_{SR} - I_{HR}\|_1$, $\mathcal{L}_{dc} = \|K_{DC} - K_{HR}\|_2$. I_{SR} and I_{HR} represent the reconstructed T2 image and

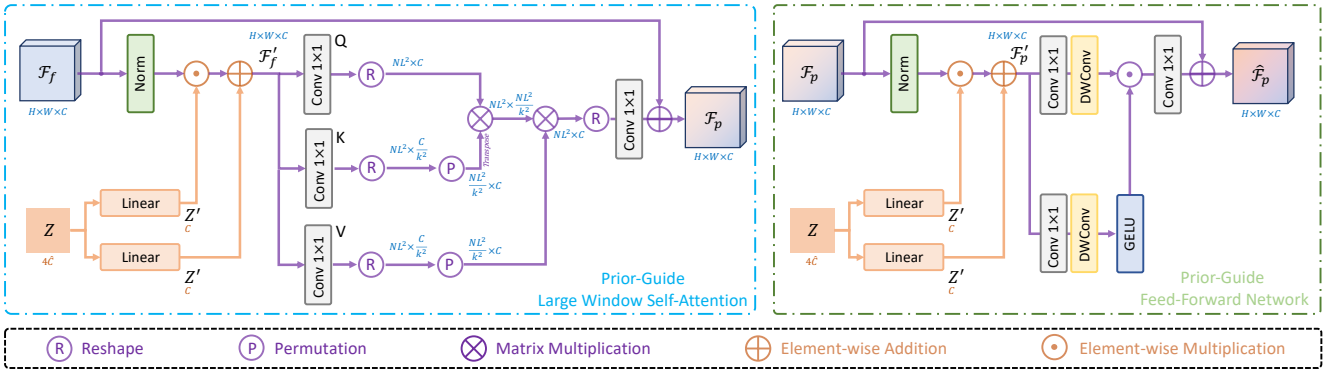

Figure 4. The architecture of prior-guide large window self-attention and prior-guide feed-forward network.

original HR T2 image, respectively. K_{DC} and K_{HR} are the frequency domain data after fidelity and the original frequency domain data, respectively. We set λ_1=1 and λ_2 =0.001 to balance the contributions of the two losses.

3.3. Training: Stage Two

The purpose of the second stage is to train DM to learn how to generate prior knowledge consistent with the distribution of real MR images for guiding and enhancing the reconstruction process of the PLWformer, as shown in Figure 3(b). DM includes the forward diffusion process and the reverse denoising process.

Diffusion Process. We employ the PE trained in the first stage to capture the prior knowledge Z. After that, we apply the diffusion process on Z to sample $Z_T \in \mathbb{R}^{4\hat{C}}$, which can be described as:

$$q\left(\mathbf{Z}_T \mid \mathbf{Z}\right)=\mathcal{N}\left(\mathbf{Z}_T ; \sqrt{\bar{\alpha}_T}\mathbf{Z}, \left(1-\bar{\alpha}_T\right)\mathbf{I}\right), \tag{6}$$

where T is the total number of iterations, $\mathcal{N}$ denotes the Gaussian distribution, α and $\bar{\alpha}_T$ are defined as: $\alpha = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, where $t = (1, \ldots, T)$, $\beta_{1:T} \in (0, 1)$ are hyperparameters that control the variance of the noise.

Denoising Process. The reverse process is a Markov chain running backwards from Z_T to $\hat{Z}$. Taking the reverse step from Z_t to Z_{t-1} as an example:

$$q\left(\mathbf{Z}_{t-1} \mid \mathbf{Z}_t, \mathbf{Z}_0\right)=\mathcal{N}(\mathbf{Z}_{t-1} ; \boldsymbol{\mu}_t\left(\mathbf{Z}_t, \mathbf{Z}_0\right), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{I}), \tag{7}$$

$$\boldsymbol{\mu}_t\left(\mathbf{Z}_t, \mathbf{Z}_0\right)=\frac{1}{\sqrt{\alpha_t}}(\mathbf{Z}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}), \tag{8}$$

where $\boldsymbol{\epsilon}$ represents the noise in Z_t and a denoising network $\boldsymbol{\epsilon}_\theta$ is employed to estimate the noise $\boldsymbol{\epsilon}$ for each step. Inspired by [33], we design a Condition Extraction (CE) module, as shown in Figure 2. The structure of this module is consistent with PE, except that it only takes the target LR image I_{LR} as input and outputs the conditional latent feature $C \in \mathbb{R}^{4\hat{C}}$. Therefore, the denoising network predicts noise conditioned on Z_t and C:

$$\boldsymbol{Z}_{t-1}=\frac{1}{\sqrt{\alpha_t}}(\boldsymbol{y}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{Z}_t, \mathbf{C}, t\right))+\sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t, \tag{9}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$. After T iterations of sampling, the DM can generate the predicted prior knowledge $\hat{Z}$, and then it can be utilized to guide the PLWformer, as shown in Figure 2. Since the prior knowledge Z is highly compact, DM in the second stage can employ fewer iterations to obtain considerably better estimations than traditional DMs [10, 27, 33].

Optimization function. We employ $\mathcal{L}^2_{stage}$ to joint train CE, denoising network, and PLWformer:

$$\mathcal{L}^2_{stage} = \lambda_1\mathcal{L}_{img} + \lambda_2\mathcal{L}_{dc} + \mathcal{L}_{diff}, \tag{10}$$

where $\mathcal{L}_{diff} = \frac{1}{4C'}\sum_{i=1}^{4C'}|\hat{\mathbf{Z}}(i) - \mathbf{Z}(i)|$.

3.4. Inference

In the inference, CE is first used to compress the target LR image I_{LR} into a conditional latent C. Then, we randomly sample Gaussian noise Z_T. The denoising network employs Z_T and C to generate prior knowledge $\hat{Z}$ after T iterations (T=4). $\hat{Z}$ is then utilized to guide PLWformer to reconstruct the final SR image, as shown in Figure 3(c). Here, the PLWformer takes I_{LR} and the reference HR image I_{ref} as input and employs CAT to fuse I_{LR} and I_{ref} to make full use of the valuable supplementary information in I_{ref}.

4. Experiments

4.1. Datasets and Baselines

Public Dataset. The public dataset employed is the FastMRI Knee [42], where the reference contrast is PD, and the target contrast is FS-PD. We selected 1600 slices with a training, validation, and test set split ratio of 7:1:2.

Clinical datasets. Clinical datasets include Brain (with T1 reference contrast and T2-FLAIR target contrast) and Pelvic (with T1 reference contrast and T2 target contrast). Specifically, the brain dataset consists of 637 slices from healthy subjects and 305 slices from tumor subjects. Among them, 512 slices of healthy subjects are utilized for training, 125 slices of healthy subjects are employed for validation and testing, and 305 slices of tumor subjects are used for additional testing. The pelvic dataset comprises 1600 slices, with a training, validation, and test split ratio of 7:1:2. The raw clinical datasets are generated by scanning with a 3T Philips Ingenia MRI Scanner. The scanning parameters for the brain are TE (T1): 2.3ms, TE (T2-FLAIR): 120ms. The scanning parameters for the pelvic are TE (T1): shortest, TE (T2): 130ms. The sum-of-squares (SOS) method is used for coil combination.

Baselines. We compared our DiffMSR with several recent state-of-the-art methods, including MCSR [25], MINet [4], MASA [21], WavTrans [18], McMRSR [17], MC-VarNet [15] and DisC-Diff [27]. Note that we focus on 4× upscaling reconstruction in our experiments.

4.2. Implementation Details

We implement our proposed approach in PyTorch [31] with a single NVIDIA RTX4090 GPU. For PLWformer, we set the number of Transformer blocks as [6, 6, 6, 6], the attention heads as [4, 4, 4, 4], the number of channels C as 64, the window size as 16×16. For the diffusion model, the channel dimension $\hat{C}$ is 64, the linear layers in the denoising network are set to 5, and the total time-step T is 4. The Adam [14] optimizer is adopted for network training with iterations of 500K. We set the batch size as 4 and the learning rate is 2e-4 and decayed by factor 0.5 at [250K, 400K, 450K, 475K]. Furthermore, the complex data $[H \times W]$ is divided into two channels, *real* and *imag*, *i.e.*, $[H \times W \times 2]$.

4.3. Qualitative results

Figure 5 provides the qualitative comparison of the various methods on the four datasets at a scale of 4×. The top, second, third, and bottom rows are the SR results under the FastMRI, clinical brain, clinical tumor and clinical pelvic datasets, respectively. The red boxes indicate the zoom-in region of complicated anatomical structures along with their corresponding error maps. Note that the brighter textures in the error maps, the lower the quality of the reconstructed images. As can be seen, compared to methods based on Transformers and CNNs, diffusion-based methods like DisC-Diff and DiffMSR (Ours) are capable of reconstructing high-realistic images with promising reconstruction metric scores (PSNR and SSIM). Nevertheless, while DisC-Diff can reconstruct high-precision MR images, it does not preserve the structure present in the original HR images, introducing some additional information that can affect medical diagnosis. In contrast, our method combines DM and PLWformer, which can preserve the original image's structure while restoring high-frequency information.

4.4. Quantitative results

In Table 1, we provide a comprehensive quantitative analysis, comparing our DiffMSR with other state-of-the-art MCSR methods on four datasets with a 4× upscaling factor. As we can see, our DiffMSR performs best among all comparison methods in terms of PSNR and SSIM metrics in all MRI datasets. Specifically, we notice that the performance of CNN-based methods is relatively poor as CNNs struggle to capture long-range dependencies. Although Transformer-based methods address some of the issues with CNNs, the ability of Transformers to reconstruct high-frequency details is limited, restricting further improvements in metrics. The DisC-Diff method based on DM achieves high metric values, as DM can generate some high-frequency details, but it also introduces unnecessary information. Our proposed DiffMSR combines the strengths of Transformer and DM, preserving the original image structure while maximizing the reconstruction of complicated anatomical structures, obtaining the best performance.

In addition, we provide the model parameters, FLOPs, and inference speed of each model in Table 1. As can be seen, our proposed method has mid-range model parameters, the smallest FLOPs, and the fastest inference speed. This is due to the two strategies we introduce to reduce computational overhead and speed up inference. (1) In the PLWformer, we employ permutation operations to decrease the computational cost of self-attention. (2) We apply the diffusion model within a highly compact latent space, requiring only a simple denoising network and a few iterations to generate prior knowledge.

4.5. Ablation Study

In this section, we explore the effectiveness of each key component of our proposed DiffMSR. All variants are retrained the same way as before and tested on the FastMRI dataset with an upsampling scale of 4×.

Effect of Reference Image. To investigate the contribution of the reference image, we design a variant *w/o* reference, which only utilizes the target LR image for reconstruction without the reference image, as shown in Table 2. As can be seen, without employing the reference image, the reconstruction performance is significantly decreased, which demonstrates that the reference image can provide some valuable supplementary information.

Effect of Prior. To validate the role of prior knowledge, we design a variant that only utilizes the first stage and excludes the prior extraction and guidance module, denoted as *w/o* prior, as shown in Table 2. As can be seen, without the

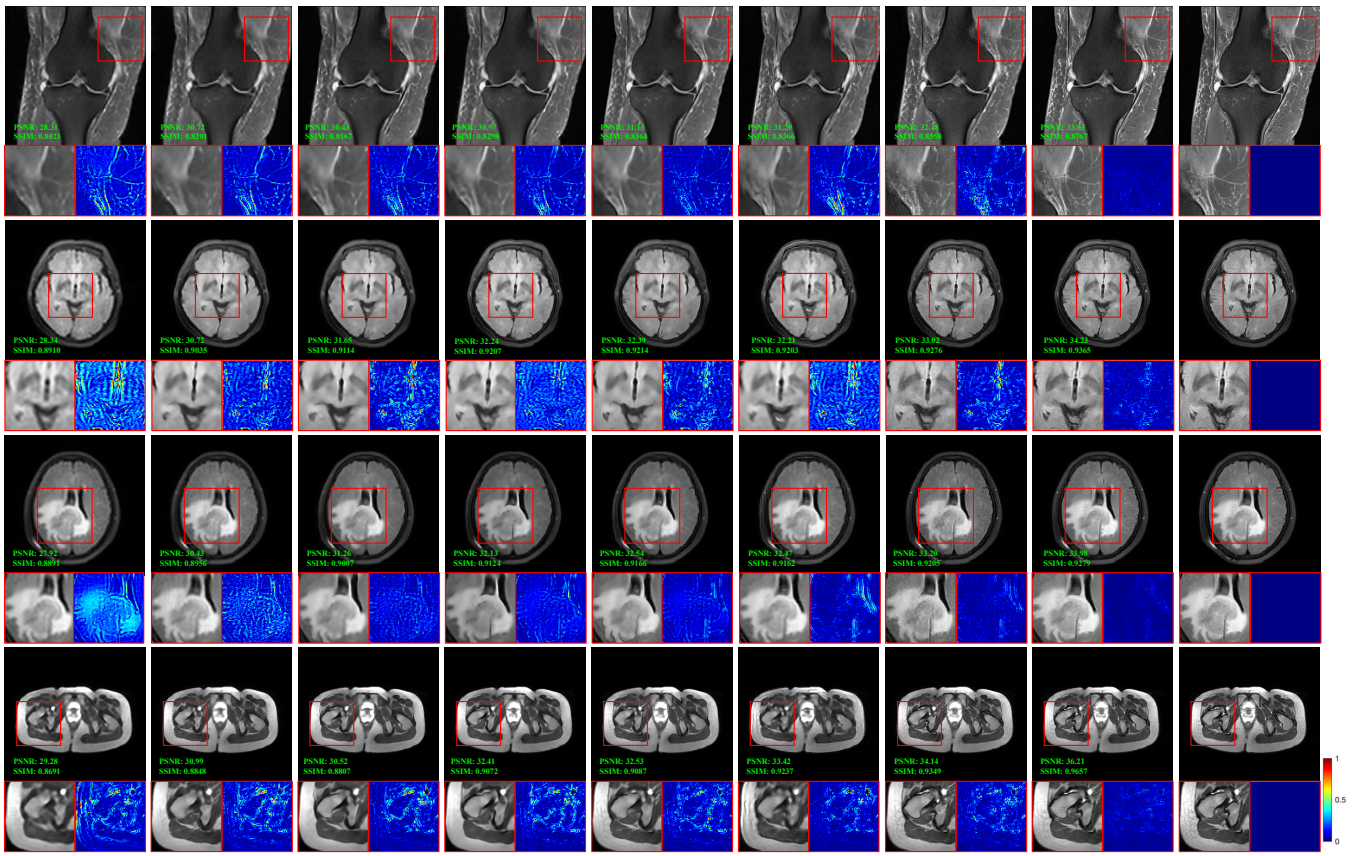

Figure 5. Qualitative comparison of SOTA MCSR methods on four datasets with an upsampling scale of 4×. The top, second, third, and bottom rows are the SR results under the FastMRI, clinical brain, clinical tumor and clinical pelvic datasets. Please zoom-in for details.

Methods	Param	FLOPs	Speed	FastMRI [42]		Clinical Brain		Clinical Tumor		Clinical Pelvic	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MCSR [25]	3.5*M*	90.399*G*	103*ms*	28.41	0.8039	28.55	0.8930	28.24	0.8909	29.59	0.8713
MINet [4]	11.9*M*	866.933*G*	115*ms*	30.19	0.8101	30.93	0.9049	30.57	0.8962	30.75	0.8861
MASA [21]	4.0*M*	180.134*G*	110*ms*	30.51	0.8192	31.17	0.9076	30.81	0.8981	30.94	0.8879
WavTrans [18]	2.1*M*	162.889*G*	173*ms*	30.95	0.8343	32.42	0.9215	31.99	0.9107	32.15	0.9054
McMRSR [17]	3.5*M*	269.860*G*	168*ms*	30.96	0.8345	32.48	0.9217	32.07	0.9121	32.19	0.9063
MC-VarNet [15]	5.7*M*	139.862*G*	97*ms*	31.65	0.8412	32.74	0.9231	32.16	0.9153	33.51	0.9265
DisC-Diff [27]	86.1*M*	461.002*G*	1455*ms*	32.24	0.8530	32.85	0.9285	32.64	0.9187	34.26	0.9388
DiffMSR (Ours)	6.8*M*	58.617*G*	39*ms*	33.78	0.8765	34.12	0.9362	33.85	0.9274	35.47	0.9607

Table 1. Quantitative comparison with state-of-the-art methods on four datasets, with performance measured in terms of PSNR (dB) ↑ and SSIM ↑. The most outstanding results are indicated in red (the best) and blue (the second-best). Param refers to the model parameters. Speed means the inference speed.

high-frequency detail information provided by prior knowledge, the reconstruction performance has significantly declined, indicating the limited ability of the Transformer to reconstruct high-frequency details.

Effect of Joint-Training. To investigate the effect of the joint training strategy in the second stage, we only optimize the diffusion model and CE in stage two, denoted as *w/o* joint. Specifically, we employ $\mathcal{L}_{diff}$ in stage two to solely train the diffusion model. After training, the diffusion model is directly combined with the PLWformer trained in the first stage for evaluation. As shown in Table 2, the results indicate that the performance of this separated training is lower than that of joint training, demonstrating the effectiveness of employing joint training in stage two.

Variants	Components						Metrics	
	Reference	Diffusion	Joint-training	DC Loss	Condition	PLWformer	PSNR ↑	SSIM ↑
w/o reference	✗	✓	✓	✓	✓	✓	32.56	0.8562
w/o prior	✓	✗	✗	✓	✗	✓	31.51	0.8428
w/o joint	✓	✓	✗	✓	✓	✓	32.35	0.8541
w/o DC	✓	✓	✓	✗	✓	✓	32.92	0.8604
w/o CE	✓	✓	✓	✓	✗	✓	33.21	0.8638
w/o PLWformer	✓	✓	✓	✓	✓	✗	33.30	0.8650
Full model	✓	✓	✓	✓	✓	✓	**33.78**	**0.8765**

Table 2. Ablation study on various variants under FastMRI with an upsampling scale of 4×. The best quantitative result is marked in **bold**.

Ratio k	$k=1$	$k=2$	$k=4$
PSNR	33.81	33.78	33.70
FLOPs	68.691G	58.617G	56.224G

Table 3. Ablation study on various ratios of k in FastMRI with an upsampling scale of 4×. The best and second-best results are indicated in red and blue, respectively.

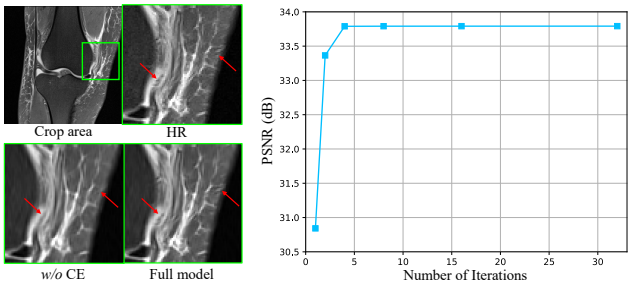

Figure 6. Left: Visualization comparison between *w/o* CE and the full model. Right: Ablation study of the number of iteration steps in the diffusion model.

Effect of Data Consistency Loss. To evaluate the contribution of data consistency (DC) loss, we conduct an ablation study by removing $\mathcal{L}_{dc}$ in the optimization function, named as *w/o* DC, as shown in Table 2. As can be seen, the reconstruction performance of *w/o* DC has significantly decreased, with a reduction of 0.86 in PSNR, indicating that the DC loss can effectively supplement frequency domain information for MR images, thereby improving the reconstruction results in the image domain.

Effect of Condition. To evaluate the effect of the condition extraction module, we design a variant by removing CE (named as *w/o* CE), which means that condition C is not employed in the denoising process, as shown in Table 2. As can be seen, without condition C, the reconstruction performance decreases, which demonstrates that condition C extracted by the CE can provide supplementary information for the denoising network. Besides, Figure 6 provides a qualitative comparison *w/o* CE. As can be seen, without condition C, the reconstructed image will lose some complicated anatomical structures.

Effect of Iterations. To explore the impact of the number of iterations on the diffusion model, we set up six variants utilizing different numbers of iteration steps T=[1, 2, 4, 8, 16, 32]. We plot the PSNR for different iterations in Figure 6. We notice that when T=1, the DM fails to generate reasonable prior knowledge, thus limiting the reconstruction performance. When T=4, the growth rate of PSNR becomes very flat, indicating that generating valuable prior knowledge only requires a small number of iterations as the simple distribution of the highly compact latent space.

Effect of Ratio *k*. We conduct an ablation study to investigate the impact of k. Specifically, we set k to 1, 2, and 4, respectively and the results are shown in Table 3. As can be seen, compared to k=1, the model with k=2 shows a significant reduction in FLOPs by about 10G, while the PSNR only slightly decreases by 0.03dB. However, when k=4, the model's performance is deceased as the spatial dimensions of K and V are greatly reduced, resulting in partial information loss. Therefore, to balance the FLOPs and performance, we set k to 2.

5. Conclusion

We propose an efficient diffusion model for multi-contrast MRI SR, which combines DM and Transformer, requiring only four iterations to reconstruct high-quality images. Besides, we introduce the PLWformer, which can expand the window size of attention without increasing the computational burden and can utilize the prior knowledge generated by DM to reconstruct MR images with high-frequency information. Extensive experiments demonstrate that our DiffMSR outperforms existing SOTA methods.

Acknowledgements. This work was supported in part by Zhejiang Province Program (2022C01222, 2023C03199, 2023C03201), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), Ningbo Science and Technology Plan Project (022Z167, 2023Z137), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

[1] Zheng Chen, Yulun Zhang, Liu Ding, Xia Bin, Jinjin Gu, Linghe Kong, and Xin Yuan. Hierarchical integration diffusion model for realistic image deblurring. In *NeurIPS*, 2023. 2, 3

[2] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80: 102479, 2022. 1, 3

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3

[4] Chun-Mei Feng, H. Fu, Shuhao Yuan, and Yong Xu. Multi-contrast mri super-resolution via a multi-stage integration network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. 1, 2, 6, 7

[5] Chun-Mei Feng, Kai Wang, Shijian Lu, Yong Xu, and Xuelong Li. Brain mri super-resolution using coupled-projection residual network. *Neurocomputing*, 456:190–199, 2021. 1

[6] Cristhian Forigua, Maria Escobar, and Pablo Arbelaez. Superformer: Volumetric transformer architectures for mri super-resolution. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 132–141. Springer, 2022. 1

[7] Mingming Gong, Shaoan Xie, Wei Wei, Matthias Grundmann, Tingbo Hou, et al. Semi-implicit denoising diffusion models (siddms). In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[9] Alper Güngör, Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Hasan A Bedel, Gokberk Elmas, Muzaffer Ozbey, and Tolga Çukur. Adaptive diffusion priors for accelerated mri reconstruction. *Medical Image Analysis*, page 102872, 2023. 1, 2, 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[11] Shoujin Huang, Jingyu Li, Lifeng Mei, Tan Zhang, Ziran Chen, Yu Dong, Linzheng Dong, Shaojun Liu, and Mengye Lyu. Accurate multi-contrast mri super-resolution via a dual cross-attention transformer network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 313–322. Springer, 2023. 1, 2

[12] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–408. Springer, 2023. 2

[13] Li Kang, Guojuan Liu, Jianjun Huang, and Jianping Li. Super-resolution method for mr images based on multi-resolution cnn. *Biomedical Signal Processing and Control*, 72:103372, 2022. 1

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Pengcheng Lei, Faming Fang, Guixu Zhang, and Tieyong Zeng. Decomposition-based variational network for multi-contrast mri super-resolution and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21296–21306, 2023. 2, 6, 7

[16] Guangyuan Li, Jun Lv, Xiangrong Tong, Chengyan Wang, and Guang Yang. High-resolution pelvic mri reconstruction using a generative adversarial network with attention and cyclic loss. *IEEE Access*, 9:105951–105964, 2021. 1

[17] Guangyuan Li, Jun Lv, Yapeng Tian, Qingyu Dou, Cheng-Yan Wang, Chenliang Xu, and Jing Qin. Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20604–20613, 2022. 1, 2, 3, 4, 6, 7

[18] Guangyuan Li, Jun Lyu, Chengyan Wang, Qi Dou, and Jin Qin. Wavtrans: Synergizing wavelet and cross-attention transformer for multi-contrast mri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022. 1, 3, 4, 6, 7

[19] Guangyuan Li, Wei Xing, Lei Zhao, Zehua Lan, Zhanjie Zhang, Jiakai Sun, Haolin Yin, Huaizhong Lin, and Zhijie Lin. Dudoinet: Dual-domain implicit network for multi-modality mr image arbitrary-scale super-resolution. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 7335–7344, 2023.

[20] Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Zhanjie Zhang, Jiafu Chen, Zhijie Lin, Huaizhong Lin, and Wei Xing. Rethinking multi-contrast mri super-resolution: Rectangle-window cross-attention transformer and arbitrary-scale upsampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21230–21240, 2023. 1, 2, 3, 4

[21] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2021. 6, 7

[22] Jun Lyu, Bin Sui, Chengyan Wang, Yapeng Tian, Qi Dou, and Jing Qin. Dudocaf: Dual-domain cross-attention fusion with recurrent transformer for fast multi-contrast mr imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 474–484. Springer, 2022. 4

[23] Jun Lyu, Guangyuan Li, Chengyan Wang, Qing Cai, Qi Dou, David Zhang, and Jing Qin. Multicontrast mri super-resolution via transformer-empowered multiscale contextual matching and aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 2, 3, 4

[24] Jun Lyu, Guangyuan Li, Chengyan Wang, Chen Qin, Shuo Wang, Qi Dou, and Jing Qin. Region-focused multi-view transformer-based generative adversarial network for cardiac cine mri reconstruction. *Medical Image Analysis*, 85: 102760, 2023. 1

[25] Qing Lyu, Hongming Shan, Cole R. Steber, Corbin A. Helis, Chris Whitlow, Michael Chan, and Ge Wang. Multi-contrast super-resolution mri through a progressive network. *IEEE Transactions on Medical Imaging*, 39:2738–2749, 2019. 1, 2, 6, 7

[26] José V Manjón, Pierrick Coupé, Antonio Buades, Vladimir Fonov, D Louis Collins, and Montserrat Robles. Non-local mri upsampling. *Medical image analysis*, 14(6):784–792, 2010. 1

[27] Ye Mao, Lan Jiang, Xi Chen, and Chao Li. Disc-diff: Disentangled conditional diffusion model for multi-contrast mri super-resolution. Springer, 2023. 1, 2, 3, 5, 6, 7

[28] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13 (1):12098, 2023. 2

[29] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 2

[30] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003. 1

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[32] Cheng Peng, Pengfei Guo, S Kevin Zhou, Vishal M Patel, and Rama Chellappa. Towards performant and reliable undersampled mr reconstruction via diffusion model sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2022. 1, 2, 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[34] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 1

[35] Chun-Yuan Shin, Yi-Ping Chao, Li-Wei Kuo, Yi-Peng Eve Chang, and Jun-Cheng Weng. Improving the brain image resolution of generalized q-sampling mri revealed by a three-dimensional cnn-based method. *Frontiers in Neuroinformatics*, 17:956600, 2023. 1

[36] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021. 4

[37] Jueqi Wang, Jacob Levman, Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, M Jorge Cardoso, and Razvan Marinescu. Inversesr: 3d brain mri super-resolution using a latent diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 438–447. Springer, 2023. 2

[38] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. *ICCV*, 2023. 2, 3

[39] Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer, 2022. 1, 2, 3

[40] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 1

[41] Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, pages 388–400. Springer, 2023. 2

[42] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018. 5, 7

[43] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 1

[44] Yulun Zhang, Kai Li, Kunpeng Li, and Yun Fu. Mr image super-resolution with squeeze and excitation reasoning attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13425–13434, 2021. 1

[45] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 4