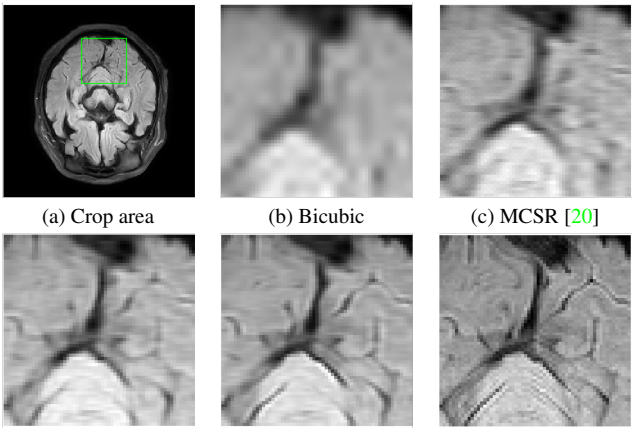
Transformer-empowered Multi-scale Contextual Matching and Aggregation for Multi-contrast MRI Super-resolution

Guangyuan Li[1], Jun Lv[1*], Yapeng Tian[2], Qi Dou[3], Chengyan Wang[4], Chenliang Xu[2], Jing Qin[5]
[1]School of Computer and Control Engineering, Yantai University, Yantai, China
[2]University of Rochester
[3]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[4]Human Phenome Institute, Fudan University, Shanghai, China
[5]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong
lgy1428275037@163.com, ljdream0710@pku.edu.cn, {yapengtian, chenliang.xu}@rochester.edu,
qidou@cuhk.edu.hk, wangcy@fudan.edu.cn, harry.qin@polyu.edu.hk

Abstract

Magnetic resonance imaging (MRI) can present multi-contrast images of the same anatomical structures, enabling multi-contrast super-resolution (SR) techniques. Compared with SR reconstruction using a single-contrast, multi-contrast SR reconstruction is promising to yield SR images with higher quality by leveraging diverse yet complementary information embedded in different imaging modalities. However, existing methods still have two shortcomings: (1) they neglect that the multi-contrast features at different scales contain different anatomical details and hence lack effective mechanisms to match and fuse these features for better reconstruction; and (2) they are still deficient in capturing long-range dependencies, which are essential for the regions with complicated anatomical structures. We propose a novel network to comprehensively address these problems by developing a set of innovative Transformer-empowered multi-scale contextual matching and aggregation techniques; we call it McMRSR. *Firstly, we tame transformers to model long-range dependencies in both reference and target images. Then, a new multi-scale contextual matching method is proposed to capture corresponding contexts from reference features at different scales. Furthermore, we introduce a multi-scale aggregation mechanism to gradually and interactively aggregate multi-scale matched features for reconstructing the target SR MR image. Extensive experiments demonstrate that our network outperforms state-of-the-art approaches and has great potential to be applied in clinical practice. Codes are available at* https://github.com/XAIMI-Lab/McMRSR.

*Corresponding author.



(a) Crop area (b) Bicubic (c) MCSR [20]
(d) MINet [8] (e) McMRSR (Ours) (f) HR

Figure 1. Compared with state-of-the-art multi-constrast MRI SR reconstruction methods: MCSR and MINet; the reconstructed MRI image by our McMRSR network contains sharper edges, more visual details, and fewer blurring artifacts.

1. Introduction

Magnetic resonance imaging (MRI) is an essential medical imaging technique in clinical application that provides clear information on tissue structure and function without causing ionizing radiation. However, due to the essential drawbacks of imaging systems [27, 36] and crepitations in some parts of the body, *e.g.*, the abdomen, it is challenging to acquire high-resolution (HR) MR images in clinics [9]. In addition, prolonged acquisition procedure may cause discomforts to patients, introduce motion artifacts, and hence affect image quality [15]. Super-resolution (SR) reconstruction is a promising way to improve the quality of MR images without upgrading hardware facilities [11].

MRI can present multi-contrast images with the same

anatomical structures at different settings, *e.g.*, T1-weighted images (T1) and T2-weighted images (T2), as well as proton density weighted images (PD) and fat-suppressed proton density weighted images (FS-PD), which can provide complementary information to each other [3, 8]. In clinical applications, as the repetition time and echo time of T1 are shorter than those of T2 and the scanning process of PD is usually shorter than that of FS-PD, T1 can be used to guide LR T2 for SR reconstruction and PD can help to reconstruct FS-PD [38]. In this regard, it is promising to leverage an HR reference image with shorter acquisition time to reconstruct the modality with longer scanning time from an LR image.

While some effort has been dedicated to multi-contrast MRI SR reconstruction [8, 20, 31, 43, 46, 47], we still face challenges in two key steps: (1) how to effectively extract the features in the reference and target images, and (2) how to transfer the features of the reference image to the features of the target image. In recent studies, Zeng *et al*. [43] employed CNN to simultaneously perform single- and multi-contrast SR reconstruction. Lyu *et al*. [20] applied a GAN-based progressive network to multi-contrast SR reconstruction. Feng *et al*. [8] used multi-stage integration network to perform multi-contrast MRI SR reconstruction. However, these methods are still incapable of sufficiently and comprehensively address the challenges in the two steps.

There are two main shortcomings. First, most existing methods harness deep convolutional layers for feature extraction. However, the convolution kernel usually has a limited receptive field and hence cannot adequately capture long-range/non-local features, which are important for MRI SR reconstruction as, for some regions with complicated anatomical structures, faithful reconstruction depends on not only local relationships but also long-range dependencies. Second, many of existing methods [8, 20] directly upsample the low-scale image into a high-scale image, and then perform the extraction and fusion of multi-contrast features. However, these methods neglect that multi-contrast features at different scales contain different anatomical details and hence can provide broad yet diverse guidance for target MRI SR reconstruction.

In order to address these two shortcomings, in this paper, we propose a novel and effective network for multi-contrast MRI SR by taming transformers to extract long-range dependencies to facilitate more comprehensive contextual matching and exploiting multi-contrast multi-scale features to guide the reconstruction at different scales with anatomical information extracted from different modalities; we call the network as *McMRSR* network. Our contributions can be summarized as follows.

1. We propose a novel network equipped with transformer-empowered multi-scale contextual matching for multi-contrast MRI SR , where Swin Transformer groups are exploited to extract deep features at different scales and from different contrasts to capture more long-range dependencies.
2. We propose multi-scale contextual matching and aggregation schemes to transfer visual contexts from reference images to target LR MR images at different scales, allowing the target LR images make full use of the guidance information to achieve SR images full of fine details.
3. Our *McMRSR* outperforms state-of-the-art approaches on three benchmark datasets: *clinical pelvic*, *clinical brain*, and *fastMRI*, demonstrating its effectiveness and great potential to be used in clinical practice.

2. Related Work

2.1. Single-Contrast MRI SR

The commonly used interpolation methods [6] are bicubic and b-spline, but they introduce edge blurring and blocking artifacts in SR images, making it impossible for clinicians to make accurate diagnosis. Traditional SR algorithms exploit redundancy in the transform domain for MRI SR reconstruction, *e.g.*, iterative deblurring algorithms [12, 33] , low rank [29] and dictionary learning [1]. However, when upsampling factor (UF) becomes large, the quality of the reconstructed SR images by these methods are not satisfactory.

Following the research in deep learning-based natural image SR methods [5, 14, 17, 23, 39, 44] and computed tomography SR method [40], some excellent MRI SR reconstruction methods emerged [4, 7, 15, 21, 22, 25, 28, 30, 32, 45]. Qui *et al*. [28] used a convolutional neural network (CNN) for knee MRI SR reconstruction. Lyu *et al*. [21] used ensemble learning for MRI SR reconstruction. Li *et al*. [15] used attention mechanism and cyclic loss in GAN for pelvic image SR reconstruction. Zhang *et al*. [45] proposed squeezed and inspired inference attention network for MR image SR, and the experimental results showed the effectiveness of the method. However, the above-mentioned algorithms all focus on reconstructing images by only using one contrast MR images.

2.2. Multi-Contrast MRI SR

The key problem of the multi-contrast MRI SR is how to get the reference image to better guide the target image in SR reconstruction. Lyu *et al*. [20] showed that the fusion of multi-contrast information in the high-level feature space yields better results than the combination in the low-level pixel space. Therefore, we consider multi-contrast feature matching and aggregation from the deep feature space to make full use of the information in the reference image. Feng *et al*. [8] used a multi-stage feature fusion mechanism for multi-contrast SR, *i.e.*, the reference features of

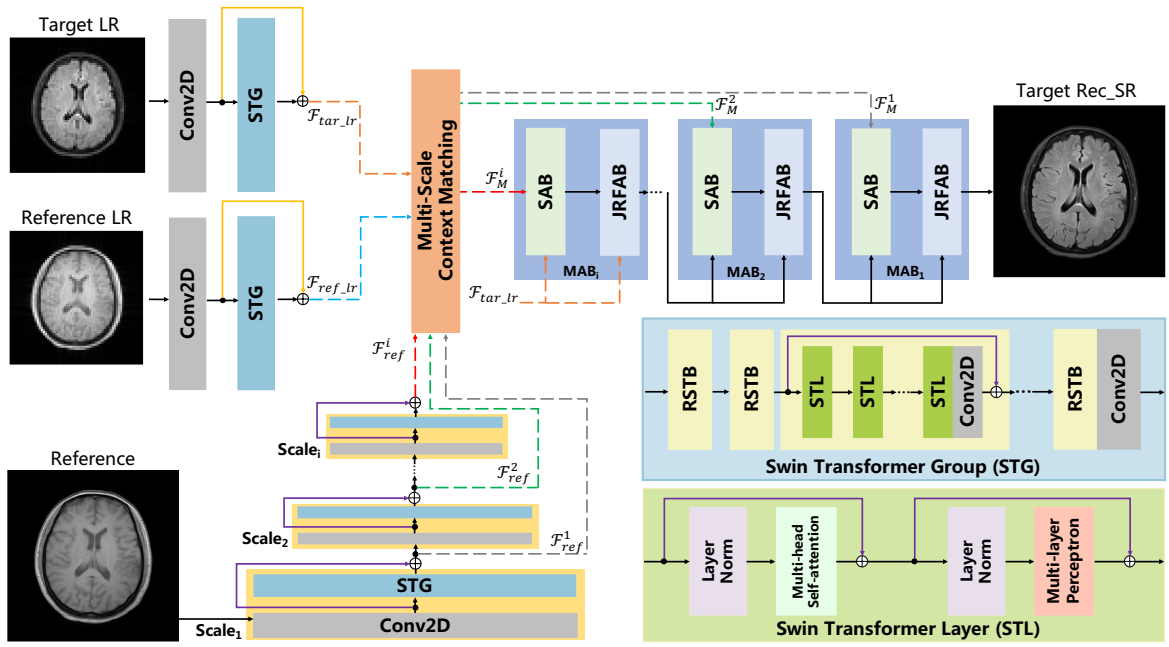


Figure 2. The overall architecture of the proposed McMRSR network. **STG**: Swin Transformer group; **RSTB**: residual Swin Transformer block; **STL**: Swin Transformer layer; **MAB**: multi-scale aggregation block; **SAB**: spatial adaptation block; **JRFAB**: joint residual feature aggregation block.

the previous stage were fused with the target features to obtain the integrated features used to guide the learning of the target features in the next stage. Inspired by [8, 19, 20], we consider fusing features from reference images of different scales in the upsampling process. Concretely, we perform multi-scale context matching and aggregation in the deep feature space and use multi-scale matched reference features to guide the recovery of target HR features.

2.3. MRI Transformer

Unlike CNNs, the transformer [37] uses a self-attentive mechanism to obtain global information between contexts and has achieved better results in dealing with visual problems [2, 18, 34]. In addition, there are several studies that have shown the effectiveness of transformer in MRI reconstruction. Feng *et al*. [11] used task transformer network to combine MRI reconstruction and SR reconstruction and proposed the use of multi-modal transformer for multi-contrast MRI reconstruction [10]. However, the general transformer is processed in the form of image patches, which results in edge pixels not learning the information of neighboring pixels outside the patches [16]. Swin Transformer [18] can be used to solve the above problem, which combines the advantages of CNN and general transformer. The method solves the problem of edge pixels in patch by shifting the window scheme to establish long-range dependencies [16]. Therefore, inspired by [16, 18], we use Swin Transformer groups consisting of multiple residual Swin Transformer blocks in McMRSR for deep feature extraction and multi-contrast features fusion.

3. Methodology

3.1. Overll Architecture

The overall architecture of the proposed McMRSR network is shown in Fig. 2. In order to obtain multi-scale feature maps for contextual matching, we carry out feature extraction through three branches, *i.e.*, target LR, reference LR and reference branches. Then, the multi-scale feature maps generated from the three branches are fed into the contextual matching module to obtain matched reference features at different scales. We then feed these matched features into the multi-scale aggregation blocks (MAB) to guide the upsampling of the target LR at multiple scales, finally obtaining the reconstructed target SR image.

3.2. Transformer-empowered Feature Extraction

As mentioned above, the long-range dependencies embedded in the feature maps are essential for efficient and robust contextual matching. In this regard, we leverage Swin Transformer group (STG) to extract the deep features of each branch; it is capable of extracting deep hierarchical representations with rich long-range dependencies [18] from both target and reference images, which facilitates the proposed network to more efficiently and precisely perform the matching. As shown in Fig. 2, the STG consists of multiple residual Swin Transformer blocks (RSTB), each employing multiple Swin Transformer layers (STL) for local

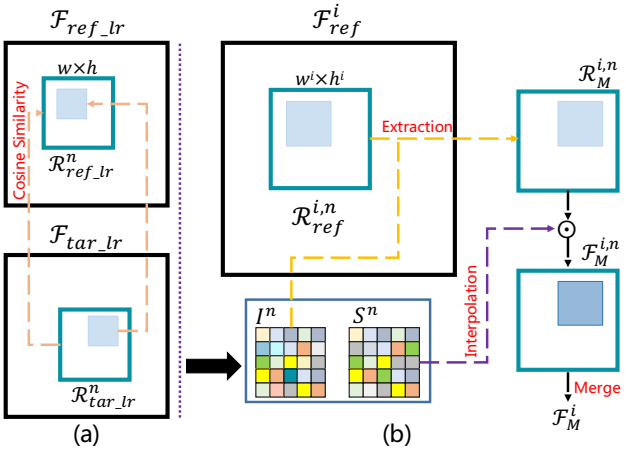


Figure 3. The process of multi-scale context matching, (a) low-scale feature context matching, (b) multi-scale feature mapping.

attention and cross-window interaction learning. The RSTB adopts residual learning to ensure the stability of feature extraction. A 3×3 convolution layer is used for feature enhancement after RSTBs and STLs. The feature extraction process of RSTB can be expressed as:

$$\mathcal{F}_{RSTB} = \mathrm{Conv}\left(\mathcal{F}_{STL}\right) + \mathcal{F}_{\mathrm{in}}\,, \tag{1}$$

where $\mathcal{F}_{STL}$ denotes the features generated from STL, Conv denotes the 3×3 Conv2D, and $\mathcal{F}_{in}$ denotes the input features of RSTB. As shown in Fig. 2, STL consists of multi-head self-attention blocks and multi-layer perceptions. More details about STL can be found in [18]. In our implementation, we set the number of RSTB and STL as 4 and 6, respectively.

We extract the multi-scale features from the reference image in a pyramid form, as shown in Fig. 2. We retain the output of each level of the pyramid and set different stride in Conv2D to ensure that the output of each layer has a different scale, named as $\mathcal{F}^i_{ref}$. Afterwards, these features are fed into the multi-scale context matching module for relevant feature mapping. The deep features obtained from LR branches are named as $\mathcal{F}_{tar\_lr}$ and $\mathcal{F}_{ref\_lr}$ respectively. Note that the features obtained in LR branches are on the same scale as the features obtained at the top of the pyramid in the reference branch.

3.3. Multi-Scale Contextual Matching

Efficient and accurate matching of features is at the core of reference image based SR reconstruction. It is crucial to leverage the details embedded in the SR reference image to make sure the reconstructed SR can contain sufficient anatomical information for clinical applications. Traditional matching schemes are incapable of achieving satisfactory results in our task because (1) directly fusing the features extracted from the multi-contrast images may bring redundant yet unnecessary features to the target images and thus reduce the quality of SR images and, (2) owing to the characteristics of medical images, long-range dependencies are quite important to offer more context-aware matching pairs for meaningful SR reconstruction but they are largely neglected in previous schemes.

In this regard, inspired by [19, 20], we perform multi-scale context matching before multi-contrast feature fusion, attempting to obtain the most relevant parts of target and reference features, *i.e.*, $\mathcal{F}_{tar\_lr}$ and $\mathcal{F}_{ref\_lr}$. Then it is mapped to the reference features at different scales, *i.e.*, $\mathcal{F}^i_{ref}$. In addition, thanks to the Transformers equipped in our network, we can implicitly harness the long-range dependencies embedded in the extracted features to enhance the matching quality. As shown in Fig. 3, our context matching can be divided into two steps: 1) context matching of low-scale features $\mathcal{F}_{tar\_lr}$ and $\mathcal{F}_{ref\_lr}$ to obtain index and similarity maps, and 2) mapping them into multi-scale features $\mathcal{F}^i_{ref}$. The details are elaborated as follows.

1) Low-scale feature context matching. To reduce the computational cost of the network, we compute the similarity maps in the target and reference features on the low-scale features. We first expand $\mathcal{F}_{tar\_lr}$ into N non-overlapping blocks to get $\mathcal{R}^n_{tar\_lr}$ $(1 \le n \le N)$; the patch size is $w \times h$ (where UF=4, w=h=13). Then, we take each $\mathcal{R}^n_{tar\_lr}$ patch center region to calculate the cosine similarity value to find the center region with the greatest similarity to $\mathcal{F}_{ref\_lr}$, and get $\mathcal{R}^n_{ref\_lr}$ patch. We crop $\mathcal{F}^i_{ref}$ with this center region to obtain multi-scale similar patches with size of $w^i \times h^i$, named as $\mathcal{R}^{i,n}_{ref}$. Thus, for each $\mathcal{R}^n_{tar\_lr}$ patch, there is a corresponding most relevant $\mathcal{R}^n_{ref\_lr}$ and $\mathcal{R}^{i,n}_{ref}$ patch. Note that as all feature maps are generated from STGs, the long-range dependencies embedded in them will implicitly affect the matching, enhancing the similarity values among patches with similar anatomical structures but located separately. Next, we perform region matching on $\mathcal{R}^n_{tar\_lr}$ and $\mathcal{R}^n_{ref\_lr}$ to get index maps $\mathcal{I}^n$ and similarity maps $\mathcal{S}^n$. For example, we first compute the similarity value between z-th region of $\mathcal{R}^n_{tar\_lr}$ and g-th region of $\mathcal{R}^n_{ref\_lr}$ to get $s^n_{z,g}$. Then, we compute the z-th elements of the index map $\mathcal{I}^n$ and similarity map $\mathcal{S}^n$, as follows:

$$\mathcal{I}^n_z = \underset{g}{\mathrm{argmax}}\ s^n_{z,g} \quad \mathcal{S}^n_z = \max_g s^n_{z,g}. \tag{2}$$

Please refer to [19] for details on how to calculate similarity values.

2) Multi-scale feature mapping. After getting the index and similarity maps, we have to map them to the $\mathcal{R}^{i,n}_{ref}$ patch at different scales to ensure that the reference features at multiple scales all contain the most similar features to the target LR, as shown in Fig. 3 (b). Specifically, according to $\mathcal{I}^n$, we extract related regions $\mathcal{R}^{i,n}_{M}$ from $\mathcal{R}^{i,n}_{ref}$ patch. Then, we multiply $\mathcal{R}^{i,n}_{M}$ with the corresponding similarity map

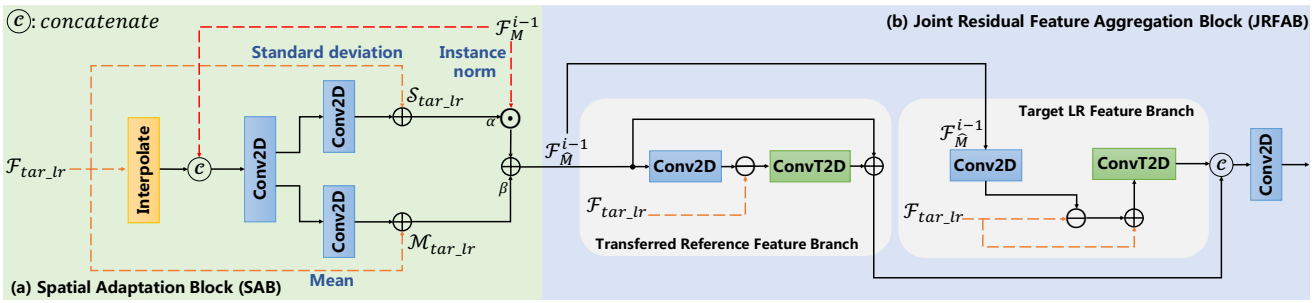


Figure 4. Multi-scale aggregation block, i,e., the fusion strategy of target LR in the upsampling process. **SAB:** spatial adaptation block, **JRFAB:** joint residual feature aggregation block. This strategy maximizes the use of the information in the matched reference features. Note that ConvT2D means ConvTranspose2D.

$\mathcal{S}^n$, and get the weighted features block $\mathcal{F}_M^{i,n}$. Note that i represents the reference features of different scale sizes. As the similarity map $\mathcal{S}^n$ is obtained on the LR scale, when $i > 1$, interpolation is required for $\mathcal{S}^n$. The above process is formulated as:

$$\mathcal{F}_M^{i,n} = multiply\left(\mathcal{R}_M^{i,n}, up\left(\mathcal{S}^n\right)\right), \tag{3}$$

where $multiply$ and up denote multiplication and bilinear interpolation. Finally, we merge N patches, and obtain multi-scale matched reference features *i.e.*, $\mathcal{F}_M^i$.

3.4. Multi-Scale Feature Aggregation

After obtaining multi-scale matched reference features, how to fuse them into the target LR features is an important yet challenging step. For the low-scale targeted LR features, fusing the matched reference features at different scales in the upsampling stage can make full use of the matched similar information and recover the details in the image to the maximum. Therefore, inspired by [19], we design MAB_i (the number corresponds to $Scale_i$) to help target LR aggregate multi-scale matched reference features, *i.e.*, $\mathcal{F}_M^i$. As shown in Fig. 2, the low-scale target LR features aggregate the features matched at the top of the pyramid, and then sequentially aggregate the reference features at different scales. This approach ensures that the matched features are fully utilized for the target LR features at each scale during upsampling. As shown in Fig. 4, this block consists of a spatial adaptation block (SAB) and a joint residual feature aggregation block (JRFAB).

Spatial Adaptation Block. We know that multi-contrast MR images have different colors and brightnesses for different contrasts, although they mirror the same anatomical structures. The previous multi-contrast MRI SR methods [8, 20] simply combine the reference and target features together and then perform the next convolution operation, which is not the optimal choice. To enhance the consistency of the matched reference features with the target LR feature distribution, inspired by [26], we use SAB to remap the distribution of matched reference features onto the distribution of traget LR features.

As shown in Fig. 4 (a), the target LR features are upsampled by 2×, and then connected with the matched reference features $\mathcal{F}_M^{i-1}$. We use stride of 1, 3×3 Conv2D to get the two parameters α and β. We then figure out the standard deviation and mean of the unsampled target LR features, and calculate $\mathcal{S}_{tar\_lr}$ and $\mathcal{M}_{tar\_lr}$ to update α and β. Next, we perform instance normalization [35] on $\mathcal{F}_M^{i-1}$, and the operation is performed with α and β to obtain the transferred reference features $\mathcal{F}_{\hat{M}}^{i-1}$ as:

$$\mathcal{F}_{\hat{M}}^{i-1} = multiply(\mathcal{F}_M^{i-1}, \alpha) + \beta. \tag{4}$$

Joint Residual Feature Aggregation Block. After SAB, we obtain the transferred reference features. In order to make the multi-scale features more fully aggregated, we consider further refining the high-frequency details in the transferred reference and target features so as to ensure that the aggregated features assimilate more anatomical details. We adopt the JRFAB to divide the aggregation process into two branches, *i.e.*, transferred reference branch and target LR branch, as shown in Fig. 4 (b). Transferred reference branch is used to enhance the high-frequency details in $\mathcal{F}_{\hat{M}}^{i-1}$, which can be formulated as:

$$\tilde{\mathcal{F}}_{\hat{M}}^{i-1} = \mathcal{F}_{\hat{M}}^{i-1} + \mathrm{ConvT}\left(\mathrm{Conv}\left(\mathcal{F}_{\hat{M}}^{i-1}\right) - \mathcal{F}_{\mathrm{tar\_lr}}\right), \tag{5}$$

where Conv denotes 3×3 Conv2D with stride of 2, and ConvT denotes 3×3 ConvTranspose2D with stride 2. Similarly, the refinement of high-frequency information in the target LR features can be expressed as:

$$\tilde{\mathcal{F}}_{\mathrm{tar\_lr}} = \mathrm{ConvT}\left(\mathcal{F}_{\mathrm{tar\_lr}} + \left(\mathcal{F}_{\mathrm{tar\_lr}} - \mathrm{Conv}\left(\mathcal{F}_{\hat{M}}^{i-1}\right)\right)\right). \tag{6}$$

Finally, we concatenate the output of the two branches and get the output of MAB_{i-1} after 3×3 Conv2D with stride 1. Note that when $\mathcal{F}_M^i$ has the same scale as $\mathcal{F}_{tar\_lr}$, upsampling of LR features is not required in SAB, and Conv2D is used in JRFAB instead of ConvT2D.

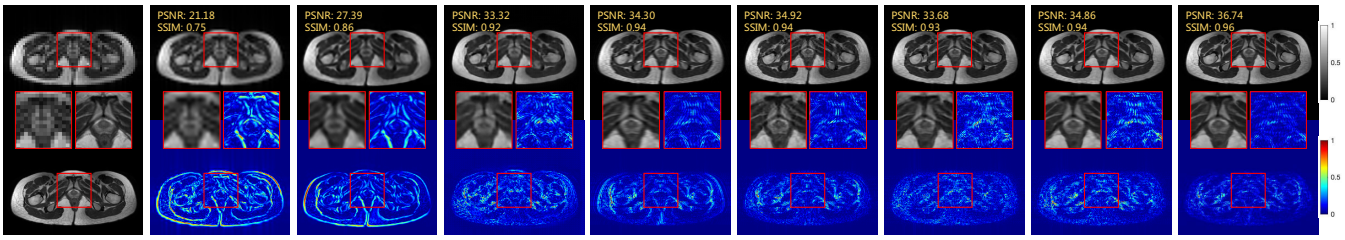


Figure 5. Qualitative results of different SR reconstruction methods on pelvic dataset with UF=4. The reconstructed images and the corresponding error maps are provided. The McMRSR recovers fine anatomical structures, as shown in the inset image.

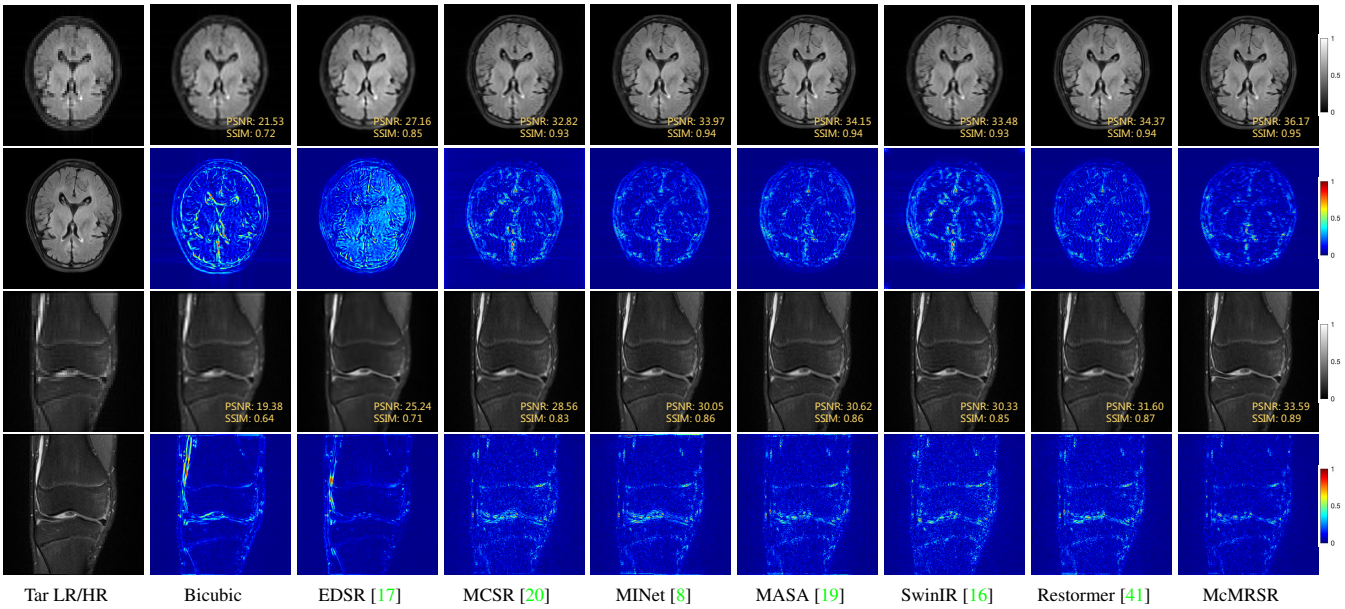


Figure 6. Qualitative results of different SR reconstruction methods on in-house brain and fastMRI knee datasets with UF=4. The reconstructed images and the corresponding error maps are provided.

3.5. Loss Functions

3.5.1 Reconstruction Loss

The L1 pixel loss is utilized as reconstruction loss to improve the overall detail of SR images [24], named as $\mathcal{L}_{rec}$:

$$\mathcal{L}_{rec} = \mathbb{E}_{(\boldsymbol{I}_{SR}, \boldsymbol{I}_{HR})} \left\| \boldsymbol{I}_{SR} - \boldsymbol{I}_{HR} \right\|_1 , \tag{7}$$

where $\boldsymbol{I}_{SR}$ denotes reconstructed MR images and $\boldsymbol{I}_{HR}$ denotes original HR MR images.

3.5.2 *k*-space Data Consistency Loss

The reconstructed SR images may lose some frequency domain information in the original HR images. We introduce the k-space data consistency [48] to militate this. Specifically, $\boldsymbol{K}_{SR}$ and $\boldsymbol{K}_{HR}$ denotes the fast Fourier transform of $\boldsymbol{I}_{SR}$ and $\boldsymbol{I}_{HR}$. Then, the sampling judgment is performed using R_{lr}. If the coefficients in $\boldsymbol{K}_{SR}$ have been sampled, they are replaced with those in $\boldsymbol{K}_{HR}$, otherwise they remain unchanged. The final fidelity of the k-space image is obtained, this process can be expressed as:

$$\boldsymbol{K}_{DC}[a,b] = \begin{cases} \boldsymbol{K}_{SR}[a,b] & \text{if } (a,b) \notin R_{lr} \\ \frac{\boldsymbol{K}_{SR}[a,b] + n\boldsymbol{K}_{HR}[a,b]}{1+n} & \text{if } (a,b) \in R_{lr} \end{cases}, \tag{8}$$

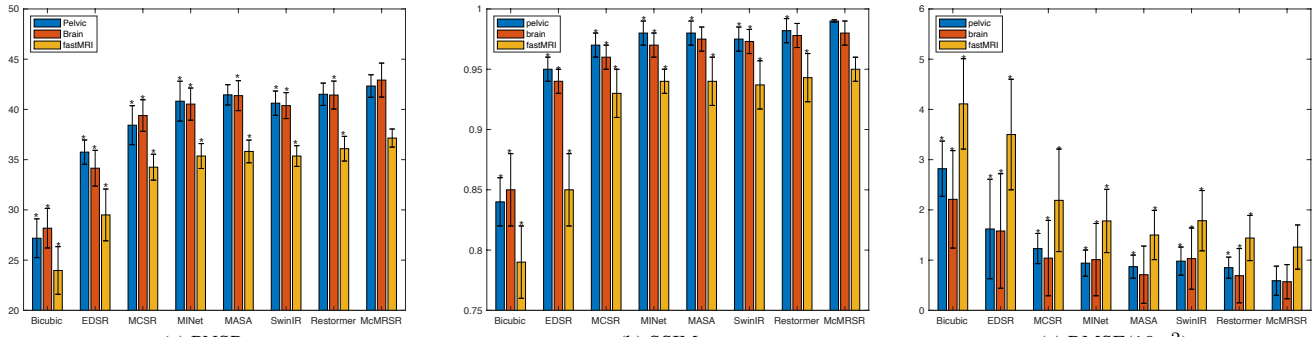
where R_{lr} is defined as the LR mask, $n \geq 0$ is the noise level (here n is set to infinity), and $[a, b]$ is the matrix indexing operation. We use mean squared error (MSE) to measure the error between $\boldsymbol{K}_{DC}$ and $\boldsymbol{K}_{HR}$ as:

$$\mathcal{L}_{dc} = \mathbb{E}_{(\boldsymbol{K}_{DC}, \boldsymbol{K}_{HR})} \left\| \boldsymbol{K}_{DC} - \boldsymbol{K}_{HR} \right\|_2 . \tag{9}$$

To the end, the full objective of the McMRSR network is defined as:

$$\mathcal{L}_{full} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{dc}\mathcal{L}_{dc}. \tag{10}$$

We set $\lambda_{rec} = 1$ and $\lambda_{dc} = 0.0001$ so that the magnitude of different loss terms can be balanced into similar scales, making their contributions reasonable.



(a) PNSR (b) SSIM (c) RMSE(10^{-2})

Figure 7. When UF=2, quantitative metrics results (mean and standard deviation) of different methods with three datasets. $*$ means significant difference between the corresponding method and McMRSR method (p <0.01).

4. Experiments

4.1. Datasets and Baselines

Three datasets are utilized to evaluate the proposed McMRSR network, including two in-house datasets of pelvic and brain and one public fastMRI [42] dataset, as shown in Tab. 1. All the complex-valued images are reshaped into the matrix size of 256×256 by cropping the k-space. We adopt a commonly used downsampling treatment, which is implemented in the frequency domain [20, 21]. Specifically, we first converted the original image of size 256×256 into the k-space. Then, only data in the central low-frequency region are kept, and all the peripheral data points are zeroed out. For the down-sampling factors 2× and 4×, the central 25% and 6.25% data points are kept. Finally, we used the inverse Fourier transform to convert the modified data into the image domain to produce the LR image.

Table 1. Three datasets used to evaluate the proposed McMRSR.

Datasets	Pelvic	Brain	fastMRI [42]
Reference	T1	T1	PD
Target	T2	T2-FLAIR	FS-PD
Train\Valid\Test	1280\320\320	513\125\125	320\80\80

We compared our McMRSR with several recent state-of-the-art methods, including a single-contrast SR method: EDSR [17], three multi-contrast SR methods: MCSR [20], MINet [8], MASA [19], and two transformer-based SR methods: SwinIR [16], Restormer [41]. Note that we concatenate the reference contrast and the target contrast as input for SwinIR and Restormer.

4.2. Implementation Details

Our proposed McMRSR is implemented in PyTorch with NVIDIA Tesla V100 GPUs (4×16GB). The Adam [13] optimizer is adopted for model training with the learning rate of 10^{-4} and epochs of 200. The performance of the SR reconstruction is evaluated by peak-to-noise-ratio (PSNR), structural similarity index (SSIM), and root mean squared error (RMSE) metrics. In addition, we use ranksum to calculate whether there is a significant difference between McMRSR and other comparison methods (p <0.01). The upsampling factors are set to 2× and 4×, respectively.

4.3. Qualitative Results

Fig. 5 provides the reconstruction results and the corresponding error maps of pelvic images when the UF=4. The predominant texture in the error map means worse reconstruction quality. As can be observed, the reconstructed SR images from the multi-contrast methods are significantly better than those from the single-contrast EDSR approach, demonstrating the effectiveness of complementary information embedded in multi-contrast images in the task of MRSR. More importantly, the SR image reconstructed by our McMRSR can better recover the uterine part and eliminate blurring edges, thanks to the proposed contextual matching and aggregation schemes.

In order to demonstrate the generalization capability and robustness of our method, we further conducted experiments on brain and fastMRI datasets, and the visual results are shown in Fig. 6. Similarly, we can see that our method is able to restore more anatomical details in both brain and knee datasets. Moreover, it can effectively handle various noises or artifacts in the images, thanks to the long-range contextual information captured by the transformers equipped in our model.

4.4. Quantitative Results

Fig. 7 reports the metrics scores with different datasets under 2× enlargement. As can be seen, our model yields the best results in terms of all metrics. We further calculate the results in terms of all metrics for each method under 4× enlargement, as shown in Tab. 2. Although it is more challenging to restore SR images under 4× enlargement than 2×, our method consistently outperforms existing methods with the best metrics scores.

Table 2. Quantitative metrics results (mean and standard deviation) on different datasets with 4× enlargement scale, in terms of PSNR, SSIM, and RMSE ($\times 10^{-2}$). **Bold** is the best results. All comparison methods have significant difference with our method (p <0.01).

Dataset	Pelvic			Brain			fastMRI [42]		
Metrics	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
Bicubic	24.38(2.30)	0.76(0.03)	6.22(2.11)	22.30(2.78)	0.74(0.02)	8.08(2.53)	19.15(2.37)	0.64(0.03)	7.23(2.16)
EDSR (CVPR2017) [17]	27.39(1.55)	0.85(0.02)	4.32(1.75)	26.01(2.30)	0.84(0.02)	6.79(2.32)	24.26(1.62)	0.69(0.02)	6.42(2.13)
MCSR (TMI2020) [20]	32.12(1.01)	0.92(0.01)	2.61(1.00)	32.09(1.95)	0.93(0.01)	4.93(1.46)	28.09(1.25)	0.82(0.03)	3.25(1.03)
MINet (MICCAI2021) [8]	34.41(0.85)	0.94(0.01)	1.82(1.27)	34.32(1.08)	0.94(0.01)	3.05(1.75)	30.58(1.38)	0.86(0.02)	2.91(0.99)
MASA (CVPR2021) [19]	34.86(1.14)	0.94(0.02)	1.59(1.46)	34.79(1.06)	0.94(0.01)	2.57(1.62)	30.97(1.14)	0.86(0.03)	2.70(0.90)
SwinIR (ICCV2021) [16]	33.92(1.09)	0.93(0.01)	2.10(1.52)	34.08(1.78)	0.93(0.02)	3.37(1.66)	30.36(1.34)	0.85(0.02)	2.98(1.09)
Restormer (arXiv) [41]	34.91(1.18)	0.94(0.02)	1.48(1.49)	34.73(1.89)	0.94(0.03)	2.54(2.07)	31.09(1.05)	0.86(0.02)	2.59(1.48)
McMRSR (Ours)	**36.23(1.07)**	**0.96(0.01)**	**1.09(0.89)**	**36.07(0.92)**	**0.95(0.01)**	**1.73(1.08)**	**33.28(0.97)**	**0.90(0.02)**	**1.82(0.85)**

Table 3. Ablation study on different variant model under fastMRI dataset with 4× enlargement scale. The best quantitative metrics results is marked in **bold**. There has a significant difference between variant models and McMRSR model (p <0.01). RMSE ($\times 10^{-2}$).

Variant	Modules				Metrics		
	Reference-Based	Multi-Scale	CM	MAB	PSNR	SSIM	RMSE
w/o reference	×	✓	✓	✓	29.05(1.24)	0.83(0.02)	3.23(0.97)
w/o multi-scale	✓	×	✓	✓	30.24(1.12)	0.85(0.03)	3.04(0.93)
w/o CM	✓	✓	×	✓	31.13(1.18)	0.86(0.02)	2.66(0.79)
w/o MAB	✓	✓	✓	×	30.56(1.03)	0.85(0.03)	2.95(0.83)
McMRSR	✓	✓	✓	✓	**33.28(0.97)**	**0.90(0.02)**	**1.82(0.85)**

4.5. Ablation Study

In this section, we demonstrate the effectiveness of the key components of McMRSR through ablation studies. The ablation studies are performed using fastMRI dataset with UF=4. In order to verify if Swin Transformer can effectively extract the deep features of images and better recover SR images, we design a single-contrast variant model using Swin Transformer, named as w/o reference. This variant model does not perform multi-scale context matching and aggregation of reference features, but only perform upsampling operation on features in target LR. To verify that context matching and aggregation at multi-scale is superior to single-scale, we design a single-scale variant model, named as w/o multi-scale. To verify the contribution of context matching and MAB in the model, we further designed variant networks without context matching (CM) for reference features, named as w/o CM and without MAB for upsampling, named as w/o MAB. The quantitative metrics results of these variant models are shown in Tab. 3.

As can be seen, the results of w/o reference are still better than the EDSR (in Tab. 2), indicating that transformers are able to extract more representative features with rich long-range dependencies for better reconstruction. More importantly, quantitative metrics scores of McMRSR are better than those of other multi-contrast variant models. This indicates the proposed multi-scale context matching and aggregation schemes are effective and capable of providing more reference features than previous approaches. Our context matching scheme ensures that the reference features at each scale contain the features most relevant to the target LR. In addition, MAB drives the target LR to maximize the use of multi-scale matched reference features during the upsampling process.

4.6. Limitation and Future Work

Here, we discuss the limitations and potential future works of this study. First, the LR-HR multi-contrast image pairs need to be co-registered in advance, which is tedious and time-consuming. In the future, we shall work on design a multi-task framework to simultaneously preform registration and SR reconstruction. Second, although our model achieves state-of-the-art performance, the reconstructed MR images still contain some artifacts, which may lead to incorrect diagnoses. In this regard, we shall further explore the fundamental limits of learning techniques for SR reconstruction and strive to design better approaches to tackle these artifacts.

5. Conclusion

We present a novel transformer-empowered multi-scale contextual matching and aggregation network for multi-contrast MRI SR reconstruction. Our model can make full use of information embedded in the reference image and reconstruct the target image with close quality to the original target HR on three representative datasets with both 2× and 4× upsampling scales. Specifically, our method provides sufficient complementary information for target LR features by harnessing contextual matching and aggregating the reference features at different scales. Experimental results show that our method is superior to the existing multi-contrast MRI SR methods and has potential to be used in clinical practice.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 61902338 and Hong Kong Research Grants Council under General Research Fund 15205919.

References

[1] Kanwal K Bhatia, Anthony N Price, Wenzhe Shi, Jo V Hajnal, and Daniel Rueckert. Super-resolution reconstruction of cardiac mri using coupled dictionary learning. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 947–950. IEEE, 2014.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[3] Wei Chen, Jun Zhao, Yaming Wen, Bin Xie, Xuanling Zhou, Lin Guo, Liu Yang, Jian Wang, Yongming Dai, and Daiquan Zhou. Accuracy of 3-t mri using susceptibility-weighted imaging to detect meniscal tears of the knee. *Knee Surgery, Sports Traumatology, Arthroscopy*, 23(1):198–204, 2015. 2

[4] Yuhua Chen, Yibin Xie, Zhengwei Zhou, Feng Shi, Anthony G Christodoulou, and Debiao Li. Brain mri super resolution using 3d deep densely connected neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 739–742. IEEE, 2018. 2

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[7] Jinglong Du, Zhongshi He, Lulu Wang, Ali Gholipour, Zexun Zhou, Dingding Chen, and Yuanyuan Jia. Super-resolution reconstruction of single anisotropic 3d mr images using residual convolutional neural network. *Neurocomputing*, 392:209–220, 2020. 2

[8] Chun-Mei Feng, Huazhu Fu, Shuhao Yuan, and Yong Xu. Multi-contrast mri super-resolution via a multi-stage integration network. In *MICCAI*, 2021. 2, 3, 5

[9] 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.

[10] Chun-Mei Feng, Yunlu Yan, Geng Chen, Huazhu Fu, Yong Xu, and Ling Shao. Accelerated multi-modal mr imaging with transformers. *arXiv e-prints*, pages arXiv–2106, 2021.

[11] Chun-Mei Feng, Yunlu Yan, Huazhu Fu, Li Chen, and Yong Xu. Task transformer network for joint mri reconstruction and super-resolution. *arXiv preprint arXiv:2106.06742*, 2021.

[12] Russell Hardie. A fast image super-resolution algorithm using an adaptive wiener filter. *IEEE Transactions on Image Processing*, 16(12):2953–2964, 2007.

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

[14] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[15] 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. 2

[16] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 6, 7, 8

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 3, 4

[19] 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. 3, 4, 5

[20] Qing Lyu, Hongming Shan, Cole Steber, Corbin Helis, Chris Whitlow, Michael Chan, and Ge Wang. Multi-contrast super-resolution mri through a progressive network. *IEEE transactions on medical imaging*, 39(9):2738–2749, 2020. 2, 3, 4, 5, 7

[21] Qing Lyu, Hongming Shan, and Ge Wang. Mri super-resolution with ensemble learning and complementary priors. *IEEE Transactions on Computational Imaging*, 6:615–624, 2020. 2, 7

[22] Steven McDonagh, Benjamin Hou, Amir Alansary, Ozan Oktay, Konstantinos Kamnitsas, Mary Rutherford, Jo V Hajnal, and Bernhard Kainz. Context-sensitive super-resolution for fast fetal magnetic resonance imaging. In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, pages 116–126. Springer, 2017. 2

[23] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 2

[24] Balamurali Murugesan, S Vijaya Raghavan, Kaushik Sarveswaran, Keerthi Ram, and Mohanasankar Sivaprakasam. Recon-glgan: a global-local context based generative adversarial network for mri reconstruction. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 3–15. Springer, 2019.

[25] Ozan Oktay, Wenjia Bai, Matthew Lee, Ricardo Guerrero, Konstantinos Kamnitsas, Jose Caballero, Antonio de Marvao, Stuart Cook, Declan O'Regan, and Daniel Rueckert. Multi-input cardiac image super-resolution using convolutional neural networks. In *International conference on med-*

ical image computing and computer-assisted intervention, pages 246–254. Springer, 2016. 2

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[27] Esben Plenge, Dirk HJ Poot, Monique Bernsen, Gyula Kotek, Gavin Houston, Piotr Wielopolski, Louise van der Weerd, Wiro J Niessen, and Erik Meijering. Super-resolution methods in mri: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? *Magnetic resonance in medicine*, 68(6):1983–1993, 2012. 1

[28] Defu Qiu, Shengxiang Zhang, Ying Liu, Jianqing Zhu, and Lixin Zheng. Super-resolution reconstruction of knee magnetic resonance imaging based on deep learning. *Computer methods and programs in biomedicine*, 187:105059, 2020. 2

[29] Feng Shi, Jian Cheng, Li Wang, Pew-Thian Yap, and Dinggang Shen. Lrtv: Mr image super-resolution with low-rank and total variation regularizations. *IEEE transactions on medical imaging*, 34(12):2459–2466, 2015.

[30] Jennifer A Steeden, Michael Quail, Alexander Gotschy, Kristian H Mortensen, Andreas Hauptmann, Simon Arridge, Rodney Jones, and Vivek Muthurangu. Rapid whole-heart cmr with single volume super-resolution. *Journal of Cardiovascular Magnetic Resonance*, 22(1):1–13, 2020. 2

[31] Bernhard Stimpel, Christopher Syben, Franziska Schirrmacher, Philip Hoelter, Arnd Dörfler, and Andreas Maier. Multi-modal super-resolution with deep guided filtering. In *Bildverarbeitung für die Medizin 2019*, pages 110–115. Springer, 2019. 2

[32] Kun Sun, Liangqiong Qu, Chunfeng Lian, Yongsheng Pan, Dan Hu, Bingqing Xia, Xinyue Li, Weimin Chai, Fuhua Yan, and Dinggang Shen. High-resolution breast mri reconstruction using a deep convolutional generative adversarial network. *Journal of Magnetic Resonance Imaging*, 52(6):1852–1858, 2020. 2

[33] Sébastien Tourbier, Xavier Bresson, Patric Hagmann, Jean-Philippe Thiran, Reto Meuli, and Meritxell Bach Cuadra. An efficient total variation algorithm for super-resolution in fetal brain mri with adaptive regularization. *NeuroImage*, 118:584–597, 2015.

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[36] Eric Van Reeth, Ivan WK Tham, Cher Heng Tan, and Chueh Loo Poh. Super-resolution in magnetic resonance imaging: a review. *Concepts in Magnetic Resonance Part A*, 40(6):306–325, 2012.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[38] Lei Xiang, Yong Chen, Weitang Chang, Yiqiang Zhan, Weili Lin, Qian Wang, and Dinggang Shen. Deep-learning-based multi-modal fusion for fast mr reconstruction. *IEEE Transactions on Biomedical Engineering*, 66(7):2105–2114, 2018.

[39] Wenzhu Xing and Karen Egiazarian. End-to-end learning for joint image demosaicing, denoising and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3507–3516, 2021. 2

[40] Haichao Yu, Ding Liu, Honghui Shi, Hanchao Yu, Zhangyang Wang, Xinchao Wang, Brent Cross, Matthew Bramler, and Thomas S Huang. Computed tomography super-resolution using convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3944–3948. IEEE, 2017. 2

[41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021. 6, 7, 8

[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.

[43] Kun Zeng, Hong Zheng, Congbo Cai, Yu Yang, Kaihua Zhang, and Zhong Chen. Simultaneous single-and multi-contrast super-resolution for brain mri images based on a convolutional neural network. *Computers in biology and medicine*, 99:133–141, 2018. 2

[44] Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2021. 2

[45] 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. 2

[46] Hong Zheng, Xiaobo Qu, Zhengjian Bai, Yunsong Liu, Di Guo, Jiyang Dong, Xi Peng, and Zhong Chen. Multi-contrast brain magnetic resonance image super-resolution using the local weight similarity. *BMC medical imaging*, 17(1):1–13, 2017. 2

[47] Hong Zheng, Kun Zeng, Di Guo, Jiaxi Ying, Yu Yang, Xi Peng, Feng Huang, Zhong Chen, and Xiaobo Qu. Multi-contrast brain mri image super-resolution with gradient-guided edge enhancement. *IEEE Access*, 6:57856–57867, 2018. 2

[48] Bo Zhou and S Kevin Zhou. Dudornet: Learning a dual-domain recurrent network for fast mri reconstruction with deep t1 prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4282, 2020.