Deform-Mamba Network for MRI Super-Resolution

Zexin Ji[1,2,4], Beiji Zou[1,2], Xiaoyan Kui[1,2]*, Pierre Vera[4], and Su Ruan[3]

[1] School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Central South University, Changsha, 410083, China
[3] University of Rouen-Normandy, LITIS - QuantIF UR 4108, F-76000, Rouen, France
[4] Department of Nuclear Medicine, Henri Becquerel Cancer Center, Rouen, France
zexin.ji@csu.edu.cn

Abstract. In this paper, we propose a new architecture, called Deform-Mamba, for MR image super-resolution. Unlike conventional CNN or Transformer-based super-resolution approaches which encounter challenges related to the local respective field or heavy computational cost, our approach aims to effectively explore the local and global information of images. Specifically, we develop a Deform-Mamba encoder which is composed of two branches, modulated deform block and vision Mamba block. We also design a multi-view context module in the bottleneck layer to explore the multi-view contextual content. Thanks to the extracted features of the encoder, which include content-adaptive local and efficient global information, the vision Mamba decoder finally generates high-quality MR images. Moreover, we introduce a contrastive edge loss to promote the reconstruction of edge and contrast related content. Quantitative and qualitative experimental results indicate that our approach on IXI and fastMRI datasets achieves competitive performance.

Keywords: Magnetic Resonance Imaging · Super-Resolution · Mamba · Deformable.

1 Introduction

Magnetic resonance imaging (MRI) is one of the broadly used medical imaging techniques for disease diagnosis. The use of high-resolution MR images can allow radiologists to diagnosis accurately as it provides more anatomical and pathological information. However, high-quality MRI requires a long echo time (TE) or repetition time (TR) and needs expensive medical equipment. Obtaining high-resolution MR images from limited scanning time and acquired hardware is highly demanded in clinical. Thus, improving the resolution of MRI is particularly important. The technique of image super-resolution (SR) has drawn a lot of attention in medical image processing. Specifically, the MRI super-resolution task aims to synthesize high-resolution (HR) MR images from the low-resolution (LR) counterparts in a cost-effective way.

The two main categories of super-resolution techniques are traditional and deep learning-based techniques. Bicubic and b-spine [12] are two examples of interpolation-based traditional SR techniques. Other more complex traditional methods have also been developed, such as methods based on reconstruction [19, 11], and methods based on examples [23, 24]. These traditional methods use their limited prior knowledge to improve image resolution, which has fast processing speed but always blurs details. Benefiting from the strong feature representation ability of convolutional neural networks (CNNs), deep learning-based methods have been used for super-resolution task [5, 13, 18, 2, 7] Super-resolution convolutional neural network (SRCNN) [5] firstly applied CNNs to the SR task. Later, the further improvements are all based on the SRCNN [13, 18, 2, 15]. For example, Qiu *et al.* [18] adopted SRCNN [5] and a sub-pixel convolutional layer to obtain high-quality MR images. Furthermore, a feedback adaptive weighted dense network (FAWDN) [2] was developed for high-resolution medical image reconstruction. However, these methods only exploit the knowledge of a limited receptive field, ignoring the long-range dependencies in the image. To solve this issue, vision Transformer has emerged in the low-level tasks [14, 6, 3, 27, 20], demonstrating the ability to capture the global contexts. For example, Forigua *et al.* [8] proposed SuperFormer, which uses the swim transformer [16] for MR image SR. Fang *et al.* [6] designed a high-frequency Transformer that incorporates the high-frequency structure prior to implement image SR. Despite their remarkable success, these approaches still encounter several problems. (1) Calculating the self-attention in Transformers incurs space and time resources that increase quadratically as the number of tokens grows. (2) The interaction of the range of utilized features has not been fully explored. (3) The training strategies usually focus on pixel level differences, making it difficult to constrain the reconstruction of high-frequency and overall image distribution information.

To address the aforementioned problems, we propose a multi-scale network called Deform-Mamba for MR image super-resolution, which simultaneously incorporates the deformable locality features with computationally lightweight global complementary features into the multi-scale network. State Space Models (SSMs) [10] have been widely explored recently. It can linearly model 1D sequences and learn long-range dependencies. The SSMs have achieved good performance in continuous long sequence data analysis tasks such as natural language processing (NLP). Recently, Mamba [9] has further improved SSMs in discrete data modeling, which is a new alternative solution for CNN or Transformer. It can further model long-range dependencies effectively with linear complexity through input-dependent selection mechanisms and hardware-aware algorithms. Compared to Transformer-based methods, Mamba is suitable for both small and large datasets. Transformers usually require large datasets, risking overfitting on smaller ones. Mamba shares parameters between time steps, learning from less data and preventing overfitting. For large datasets, Transformers need substantial computational power and struggle with ultra-large datasets, while Mamba is more efficient, requiring fewer computational resources. Some studies have explored the application of Mamba in computer vision tasks. For example, Zhu

et al. [26] designed a vision Mamba backbone with bidirectional Mamba blocks, which demonstrates its effectiveness in semantic segmentation, classification, and object detection tasks. Xing *et al.* [22] proposed SegMamba for 3D medical image segmentation. Inspired by the success of Mamba, we first propose to use Mamba to realize the MR image super-resolution task. Different from the traditional method combining ordinary convolution and Transformer to learn the local and global features, we design a Deform-Mamba module as the basic unit to merge the strengths of both deformable CNNs [4] and Mamba. This module can fully explore local and global features of the image according to the image content. We combine this module with the Unet network to further learn multi-scale local and global information. Furthermore, we also develop a multi-view context module to strengthen the ability to understand image semantic content in the bottleneck layer. To further enhance the high-frequency information of super-resolved images, we introduce a contrastive edge loss (CELoss) that focuses more on the edge texture and contrast of MR images. The quantitative and qualitative experimental results demonstrate the effectiveness of our Deform-Mamba.

2 Method

2.1 Preliminaries

State Space Models (SSMs) [10], i.e., Mamba [9] that core is selective scan space state sequential model (S6), maps a 1-D function or sequence $x(t) \in \mathbb{R}$ to the output $y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^{\mathbb{N}}$. It is usually calculated by the linear ordinary differential equations (ODEs).

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{\mathbb{N} \times \mathbb{N}}$ is the state matrix. $\mathbf{B} \in \mathbb{R}^{\mathbb{N} \times 1}$ and $\mathbf{C} \in \mathbb{R}^{1 \times \mathbb{N}}$ are the projection parameters.

Mamba uses the discrete rule zero-order hold (ZOH) to discretize ODEs into discrete functions, which is more suitable for deep learning scenarios. Specifically, we use timescale parameter Δ to transform $\mathbf{A}$ and $\mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, respectively. The specific implementation is as follows:

$$\overline{\mathbf{A}} = \exp(\Delta \mathbf{A}), \overline{\mathbf{B}} = (\Delta \mathbf{A})^{-1}(\exp(\Delta \mathbf{A}) - \mathbf{I}) \cdot \Delta \mathbf{B}. \tag{2}$$

Eq. (1) will be the following form after the discretization:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, y_t = \mathbf{C}h_t. \tag{3}$$

Finally, the model generates the output via a global convolution as follows:

$$\overline{\mathbf{K}} = \left(\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}\overline{\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{M-1}\overline{\mathbf{B}} \right), \mathbf{y} = \mathbf{x} * \overline{\mathbf{K}}, \tag{4}$$

where M denotes the length of the input sequence $\mathbf{x}$. $\overline{\mathbf{K}} \in \mathbb{R}^{\mathbb{M}}$ represents a structured convolutional kernel. This model can selectively and linearly learn the long-range dependencies of sequences.

Fig. 1. Architecture of our Deform-Mamba. (A) The overall architecture mainly consists of a Deform-Mamba encoder, a multi-view context module, and a vision Mamba decoder. (B) An implementation of the Deform-Mamba module includes (B-1) modulated deform block, (B-2) vision Mamba block, and (B-3) 2D-Selective-Scan (SS2D) block. (C) Multi-view context block. (D) Contrastive edge loss (CELoss).

2.2   Proposed Architecture

We show the main network architecture in Fig.1 (A). In the first step, we use pixel shuffle to upsample low-resolution image in a learnable manner, so as not to lose image information. Our approach mainly contains patch embedding, a Deform-Mamba encoder, a multi-view context module, a vision Mamba decoder, and a final projection layer. Specifically, the image is first cropped into non-overlapping patches through patch embedding. The encoder consists of a Deform-Mamba module and patch merging that increases the number of channels. We also design the multi-view context module to further enhance feature representation in the bottleneck layer. In the decoder, each layer consists of a pure vision Mamba module and patch expanding that reduces the number of channels. The super-resolved image is obtained through the final projection layer. Furthermore, we develop a contrastive edge loss (CELoss) to regulate the generation of edges and contrast of the image.

Deform-Mamba Module. As illustrated in Fig.1 (B), we develop a Deform-Mamba module to extract local and global features that can be adaptive to image content. It mainly consists of two branches, modulated deform block and vision Mamba block. For local operation, we use the modulated deform block to adaptively modify the size of the receptive field of local operation, while for global operation, we apply a novel vision Mamba block for building long-distance

dependence between features. Then we add output features of two branches to gradually fuse the multi-scale deformable local and efficient global information.

Specifically, Fig.1 (B-1) shows the network structure of the modulated deform block. Different from traditional convolution using fixed shape sampling, the modulated deform block can dynamically learn the convolution bias according to the image content to adaptively extract features. The operation is characterized as follows:

$$Y(p) = \sum_{k=1}^{K} w_k \cdot X\left(p + p_k + \Delta p_k\right) \cdot \Delta m_k, \tag{5}$$

where k is the index. w_k denotes the k-th kernel weight. X and Y are the input and output feature map. p, p_k, $\triangle p_k$ and $\triangle m_k$ are the sampling position, k-th predefined offset, dynamic offset, and modulation scalar for the k-th location, respectively. Specifically, $\Delta p_k \in \mathbb{R}^{H \times W \times 2C}$ and $\Delta m_k \in \mathbb{R}^{H \times W \times C}$ are learned by the convoluional layer. $\Delta p_k = \text{Conv}(\mathbf{X})$, and $\Delta m_k = \sigma(\text{Conv}(\mathbf{X}))$. The σ is the sigmoid function. We then add the learned $\triangle p_k$ to p_k and adaptively change the sampling position based on the image content. In addition, $\triangle m_k$ can control the weight of the offset sampling points, which reduces the interference of irrelevant features.

Fig.1 (B-2) illustrates the architecture of the vision Mamba block. Unlike the Transformer, which employs a self-attention mechanism to calculate long-range dependencies of images. The vision Mamba block models in a sequential manner, which is more efficient than Transformer, in particular for processing high-resolution images. The input feature map first passes through layer normalization, then it diverges into dual pathways. The first pathway involves the input through a linear layer and then an activation function. In the second pathway, it is processed via a linear layer, depthwise separable convolution, and an activation function, subsequently into the 2D-Selective-Scan (SS2D) and layer normalization. Then we use multiplication to integrate both pathways. We also utilize channel attention to model the correlation between different channels. Fig.1 (B-3) shows the specific implementation of SS2D. The scan expanding first unfolds the image in four different directions to fully mine effective information. The main core module S6 of Mamba is used to connect with previous patches of the image by the hidden state space to lightly learn long-range dependencies (Eq.2-Eq.4). For more details of S6, please refer to [10]. Finally, the scan merging can merge four features of different sequences to recover the original size.

Multi-view Context Module. Inspired by [1], we also develop a multi-view context module in the bottleneck layer to improve the ability of feature extraction by exploring the multi-view information. The bottleneck layer is a crucial location for feature extraction and information aggregation. As shown in Fig.1 (C), we parallelly utilize atrous convolution with different dilation rates to increase the receptive field of the convolution kernel. These features are concatenated to fuse feature information at different scales, without increasing the number of parameters and computational complexity. Then we use the residual connection to get the output of the bottleneck layer.

Loss Function. We utilize the $\mathcal{L}_1$ loss function to measure the difference between the network prediction and the ground truth image. However, pixel-level loss functions primarily concentrate on the individual disparities among pixels, failing to comprehensively grasp the structural intricacies within an image. Therefore, we design a contrastive edge loss (CELoss) to constrain the generation of more detailed super-resolved MR images by enhancing the edge and local contrast information.

$$\mathcal{L}_{CELoss} = \sum_i \|\mathbf{E}_i \odot SR - \mathbf{E}_i \odot HR\|_2^2, \tag{6}$$

where SR is the super-resolved MR image, and HR is the ground-truth high-resolution image. E_i is the i th contrastive edge convolution kernel. As shown in Fig.1 (D), $i \in [0, 2]$. Specifically, E1 enhances horizontal and vertical edges via neighborhood differences. E2 targets diagonal edges through diagonal differential calculations, and E3 boosts local contrast by comparing the central pixel with all neighborhoods. The three kernels are designed to emphasize the edges of the image, as well as local contrast, to highlight the details of liquid and moisture regions in MR images. The final loss function ensures that the SR model not only reconstructs the image at the pixel level but also restores critical visual features about edge and contrast. It is represented as:

$$Loss = \mathcal{L}_1 + \beta \mathcal{L}_{CELoss}, \tag{7}$$

where the weight β for $\mathcal{L}_{CELoss}$ is 0.1 in our experiment.

3   Experiments

3.1   Datasets and Evaluation Metrics

To evaluate the effectiveness of our approach, we conducted experiments on the IXI [5] and fastMRI [6] with the brain and knee T2 weighted image. The slice resolutions of IXI and fastMRI are 256×256 and 320×320, respectively. We utilized 368 subjects from the IXI dataset for training and 92 for testing. For the fastMRI dataset, 227 subjects were used for training and 45 for testing. To synthesize low-resolution input images, we implement it in the frequency domain [17] to further fit the true distribution of low-resolution images. We utilized the peak signal to noise ratio (PSNR) and structural similarity index (SSIM) [21] to evaluate the quality of super-resolved images.

3.2   Experimental Details

We implemented our approach with the Pytorch toolbox and trained the network on the NVIDIA RTX A6000 GPU. We used Adam optimizer with the initial

---

Table 1. Ablation study with different components in our Deform-Mamba

Method	PSNR↑	SSIM↑
w/o Deform	32.42	0.9250
w/o MVC	32.60	0.9266
w/o CELoss	32.59	0.9264
Deform-Mamba	**32.65**	**0.9270**

Table 2. Quantitative results with different methods under fastMRI and IXI dataset

Method	fastMRI 4×		IXI 4×	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRCNN	19.74	0.3653	28.12	0.8357
VDSR	20.31	0.3839	28.34	0.8392
FMISR	24.35	0.5207	28.27	0.8349
T^{2}Net	<u>30.56</u>	<u>0.6244</u>	29.73	0.8773
HAT	29.65	0.6155	**30.73**	**0.9007**
Our	**32.11**	**0.7194**	<u>30.60</u>	<u>0.8965</u>

learning rate of 1×10^{-4} to update network parameters. The network was trained for 50 epochs with the batch size of 2. We used 4 vision Mamba blocks in each level. The channel count of each level is [96,128,384,768].

3.3 Ablation Study

We studied the importance of the designed components in our Deform-Mamba. To validate the effect of deformable local features for super-resolution, we remove the modulated deform block. It represented *w/o Deform*, which is the pure vision Mamba network. Further, we remove the multi-view context module to evaluate the effect of multi-scale contextual information in the bottleneck layer. This model is called *w/o MVC*. We also verify the effect of contrastive edge loss, which is denoted as *w/o CELoss*. The overall result of our model is *Deform-Mamba*. From the Tab.1 we can see that the results of *w/o Deform*, *w/o MVC*, and *w/o CELoss* are worse than our *Deform-Mamba*. It indicates the necessity of deformable local features and multi-view contextual information for super-resolution, and the constraint of edge and contrast is useful to generate high-frequency information.

3.4 Comparison with state-of-the-art methods

Quantitative Analysis. For quantitative evaluation, we compared the PSNR and SSIM values of SRCNN [5], VDSR [13], FMISR [25], T^{2}Net [7], HAT [3] and our approach on the IXI and fastMRI dataset under 4× upsampling factors. We include the results of the 2× upsampling in the supplementary materials. The quantitative comparison experiment results are shown in Tab.2. It can be seen from the table that the PSNR and SSIM values of our approach on the
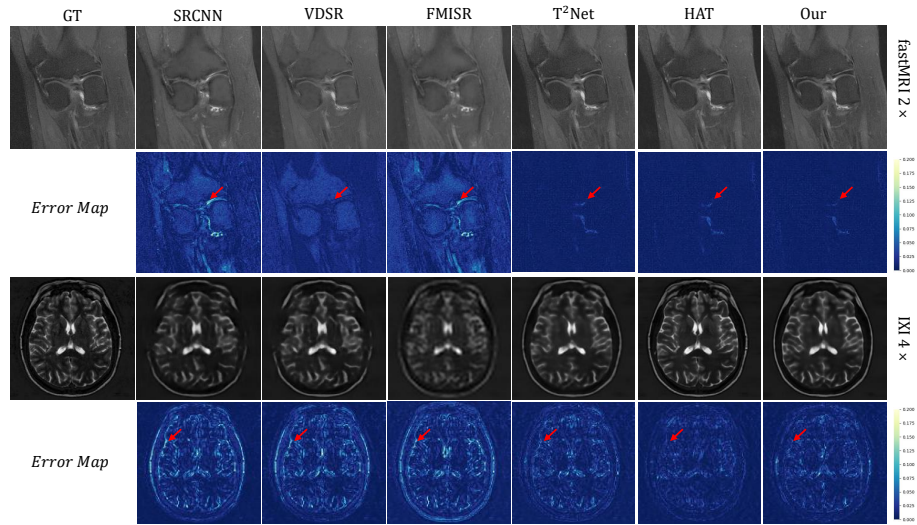
Fig. 2. Qualitative results with different methods under fastMRI and IXI dataset

fastMRI dataset obtain the best performance than those of other methods. This is because our proposed Deform-Mamba model can adaptively mine the local and global content of images. For the IXI dataset, we achieved the second-best results compared to transformer-based HAT. Concerning the two MRI datasets, IXI dataset is used for brain, and fastMRI for knee. Knee images feature simple textures and fairly clear contours, with data evenly distributed. Brain images are complex, with indistinct white and gray matter contours both physically and physiologically, occupying $2/3$ of the image center. The rest is noise-filled black background. The difference in the average values of these two zones is very large, thus slightly disrupting Mamba's attention mechanism. Furthermore, Mamba's hardware-aware algorithm processes data linearly with sequence length, significantly boosting computational speed. HAT's cost in Multi-Adds for a 64x64 input is about 5x higher than ours, our method is 2x faster in training, showing superior efficiency.

Qualitative Analysis. The super-resolved results and corresponding error maps are shown in Fig.2 under the fastMRI and IXI datasets. The error map is typically a visualization of the difference between the super-resolved image and the ground truth high-resolution image. The darker the color, the smaller the difference between the generated image and the label image, and vice versa. The figures show that the compared methods on the fastMRI dataset produce blurring artifacts since the available information in the image is not fully utilized by these approaches. It can be clearly seen in the error map that the color of our method is the darkest. For the IXI dataset, our method with fewer parameters achieved the best local visualization results.

4   Conclusion and Future Work

In this paper, we have developed a Deform-Mamba network for MR image super-resolution. Our model combines a modulated deform block and vision Mamba block as a unit in the encoder, which can activate more content-adaptive local and global features for super-resolution. The designed multi-view context block in the bottleneck layer can further enhance the fusion of multi-scale contextual information. The contrastive edge loss further reconstructs edge-enhanced and contrast-consistent high-resolution images. Quantitative and qualitative experiments demonstrate the effectiveness of our approach. In future work, our Deform-Mamba network can be added as a baseline to the diffusion model to further improve the performance of super-resolution networks. We can also apply our method to more clinical datasets with other types of imaging, such as PET and CT, to extend its clinical implications.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
2. Chen, L., Yang, X., Jeon, G., Anisetti, M., Liu, K.: A trusted medical image super-resolution method based on feedback adaptive weighted dense network. Artif. Intell. Medicine **106**, 101857 (2020)
3. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 22367–22377 (2023)
4. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: IEEE International Conference on Computer Vision, ICCV. pp. 764–773 (2017)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016)
6. Fang, C., Zhang, D., Wang, L., Zhang, Y., Cheng, L., Han, J.: Cross-modality high-frequency transformer for mr image super-resolution. In: ACM International Conference on Multimedia, ACM MM. pp. 1584–1592 (2022)

7. Feng, C., Yan, Y., Fu, H., Chen, L., Xu, Y.: Task transformer network for joint MRI reconstruction and super-resolution. In: Medical Image Computing and Computer Assisted Intervention, MICCAI. vol. 12906, pp. 307–317 (2021)
8. Forigua, C., Escobar, M., Arbelaez, P.: Superformer: Volumetric transformer architectures for mri super-resolution. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 132–141 (2022)
9. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
10. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. In: Advances in Neural Information Processing Systems, NeurIPS. pp. 572–585 (2021)
11. Irani, M., Peleg, S.: Improving resolution by image registration. CVGIP Graph. Model. Image Process. **53**(3), 231–239 (1991)
12. Keys, R.: Cubic convolution interpolation for digital image processing. IEEE Trans. Acoust. Speech Signal Process. **29**(6), 1153–1160 (1981)
13. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 1646–1654 (2016)
14. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: IEEE International Conference on Computer Vision, ICCV. pp. 1833–1844 (2021)
15. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 1132–1140 (2017)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE International Conference on Computer Vision, ICCV. pp. 10012–10022 (2021)
17. Lyu, Q., Shan, H., Steber, C., Helis, C., Whitlow, C., Chan, M., Wang, G.: Multi-contrast super-resolution MRI through a progressive network. IEEE Trans. Medical Imaging **39**(9), 2738–2749 (2020)
18. Qiu, D., Zhang, S., Liu, Y., Zhu, J., Zheng, L.: Super-resolution reconstruction of knee magnetic resonance imaging based on deep learning. Comput. Methods Programs Biomed. **187**, 105059 (2020)
19. Stark, H., Oskoui, P.: High-resolution image recovery from image-plane arrays, using convex projections. J. Opt. Soc. Am. A. **6**(11), 1715–1726 (1989)
20. Wang, W., Shen, H., Chen, J., Xing, F.: MHAN: multi-stage hybrid attention network for MRI reconstruction and super-resolution. Comput. Biol. Medicine **163**, 107181 (2023)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
22. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
23. Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T.S.: Coupled dictionary training for image super-resolution. IEEE Trans. Image Process. **21**(8), 3467–3478 (2012)
24. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
25. Zhang, S., Liang, G., Pan, S., Zheng, L.: A fast medical image super resolution method based on deep learning network. IEEE Access **7**, 12319–12327 (2019)

26. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
27. Zou, B., Ji, Z., Zhu, C., Dai, Y., Zhang, W., Kui, X.: Multi-scale deformable transformer for multi-contrast knee MRI super-resolution. Biomed. Signal Process. Control. **79**(Part), 104154 (2023)