# Accurate Multi-contrast MRI Super-Resolution via a Dual Cross-Attention Transformer Network

Shoujin Huang[1], Jingyu Li[1], Lifeng Mei[1], Tan Zhang[1], Ziran Chen[1], Yu Dong[2], Linzheng Dong[2], Shaojun Liu[1(✉)], and Mengye Lyu[1(✉)]

[1] Shenzhen Technology University, Shenzhen, China
liusj14@tsinghua.org.cn, lvmengye@sztu.edu.cn
[2] Shenzhen Samii Medical Center, Shenzhen, China

**Abstract.** Magnetic Resonance Imaging (MRI) is a critical imaging tool in clinical diagnosis, but obtaining high-resolution MRI images can be challenging due to hardware and scan time limitations. Recent studies have shown that using reference images from multi-contrast MRI data could improve super-resolution quality. However, the commonly employed strategies, e.g., channel concatenation or hard-attention based texture transfer, may not be optimal given the visual differences between multi-contrast MRI images. To address these limitations, we propose a new Dual Cross-Attention Multi-contrast Super Resolution (DCAMSR) framework. This approach introduces a dual cross-attention transformer architecture, where the features of the reference image and the upsampled input image are extracted and promoted with both spatial and channel attention in multiple resolutions. Unlike existing hard-attention based methods where only the most correlated features are sought via the highly down-sampled reference images, the proposed architecture is more powerful to capture and fuse the shareable information between the multi-contrast images. Extensive experiments are conducted on fastMRI knee data at high field and more challenging brain data at low field, demonstrating that DCAMSR can substantially outperform the state-of-the-art single-image and multi-contrast MRI super-resolution methods, and even remains robust in a self-referenced manner. The code for DCAMSR is avaliable at https://github.com/Solor-pikachu/DCAMSR.

**Keywords:** Magnetic resonance imaging · Super-resolution · Multi-contrast

## 1 Introduction

Magnetic Resonance Imaging (MRI) has revolutionized medical diagnosis by providing a non-invasive imaging tool with multiple contrast options [1,2]. However,

S. Huang and J. Li contribute equally to this work.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43999-5_30.


H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 313–322, 2023.
https://doi.org/10.1007/978-3-031-43999-5_30

generating high-resolution MRI images can pose difficulties due to hardware limitations and lengthy scanning times [3,4]. To tackle this challenge, super-resolution techniques have been developed to improve the spatial resolution of MRI images [5]. However, while several neural network-based super-resolution methods (e.g., EDSR [6], SwinIR [7], and ELAN [8]) have emerged from the computer vision field, they primarily utilize single-contrast data, ignoring the valuable complementary multi-contrast information that is easily accessible in MRI.

Recent studies have shown that multi-contrast data routinely acquired in MRI examinations can be used to develop more powerful super-resolution methods tailored for MRI by using fully sampled images of one contrast as a reference (Ref) to guide the recovery of high-resolution (HR) images of another contrast from low-resolution (LR) inputs [9]. In this direction, MINet [10] and SANet [11] have been proposed and demonstrated superior performance over previous single-image super-resolution approaches. However, these methods rely on relatively simple techniques, such as channel concatenation or spatial addition between LR and Ref images, or using channel concatenation followed by self-attention to identify similar textures between LR and Ref images. These approaches may overlook the complex relationship between LR and Ref images and lead to inaccurate super-resolution.

Recent advances in super-resolution techniques have led to the development of hard-attention-based texture transfer methods (such as TTSR [12], MASA [13], and McMRSR [14]) using the texture transformer architecture [12]. However, these methods may still underuse the rich information in multi-contrast MRI data. As illustrated in Fig. 1(a), these methods focus on spatial attention and only seek the most relevant patch for each query. They also repetitively use low-resolution attention maps from down-sampled Ref images ($\text{Ref}_{\downarrow\uparrow}$), which may not be sufficient to capture the complex relationship between LR and Ref images, potentially resulting in suboptimal feature transfer. These limitations can be especially problematic for noisy low-field MRI data, where down-sampling the Ref images (as the key in the transformer) can cause additional image blurring and information loss.

As shown in Fig. 1(b), our proposed approach is inspired by the transformer-based cross-attention approach [15], which provides a spatial cross-attention mechanism using full-powered transformer architecture without Ref image down-sampling, as well as the UNETR++ architecture [16], which incorporates channel attention particularly suitable for multi-contrast MRI images that are anatomically aligned. Building upon these developments, the proposed Dual Cross-Attention Multi-contrast Super Resolution (DCAMSR) method can flexibly search the reference images for shareable information with multi-scale attention maps and well capture the information both locally and globally via spatial and channel attention. Our contributions are summarized as follows: 1) We present a novel MRI super-resolution framework different from existing hard-attention-based methods, leading to efficient learning of shareable multi-contrast information for more accurate MRI super-resolution. 2) We introduce a dual cross-attention transformer to jointly explore spatial and channel information,

substantially improving the feature extraction and fusion processes. 3) Our proposed method robustly outperforms the current state-of-the-art single-image as well as multi-contrast MRI super-resolution methods, as demonstrated by extensive experiments on the high-field fastMRI [17] and more challenging low-field M4Raw [18] MRI datasets.

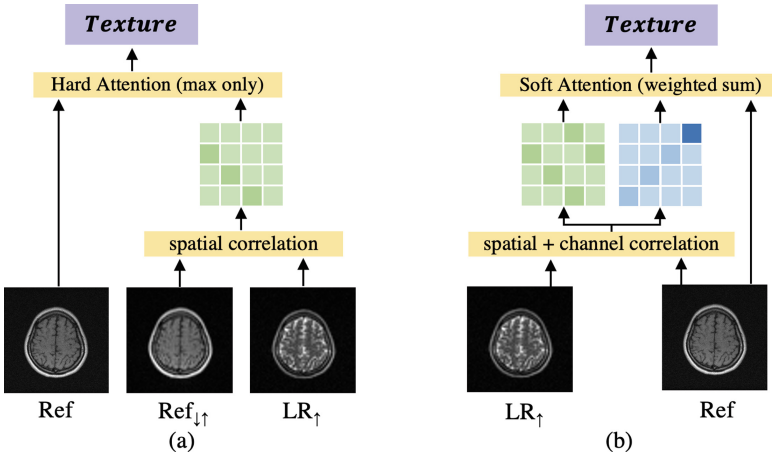


**Fig. 1.** (a) Illustration of Texture Transformer. (b) Illustration of the proposed Dual Cross-Attention Transformer.

## 2 Methodology

**Overall Architecture.** Our goal is to develop a neural network that can restore an HR image from an LR image and a Ref image. Our approach consists of several modules, including an encoder, a dual cross-attention transformer (DCAT) and a decoder, as shown in Fig. 2. Firstly, the LR is interpolated to match the resolution of HR. Secondly, we use the encoder to extract multi-scale features from both the up-sampled LR and Ref, resulting in features $F_{LR}$ and $F_{Ref}$. Thirdly, the DCAT, which contains of dual cross-attention (DCA), Layer Normalization (LN) and feed-forward network (FFN), is used to search for texture features from $F_{LR}$ and $F_{Ref}$. Fourthly, the texture features are aggregated with $F_{LR}$ through the Fusion module at each scale. Finally, a simple convolution is employed to generate SR from the fused feature.

**Encoder.** To extract features from the up-sampled $LR$, we employ an encoder consisting of four stages. The first stage uses the combination of a depth-wise convolution and a residual block. In stages 2–4, we utilize a down-sampling layer and a residual block to extract multi-scale features. In this way, the multi-scale features for the $LR_{\uparrow}$ are extracted as $F_{LR}^{H\times W}$, $F_{LR}^{\frac{H}{2}\times\frac{W}{2}}$, $F_{LR}^{\frac{H}{4}\times\frac{W}{4}}$ and $F_{LR}^{\frac{H}{8}\times\frac{W}{8}}$,

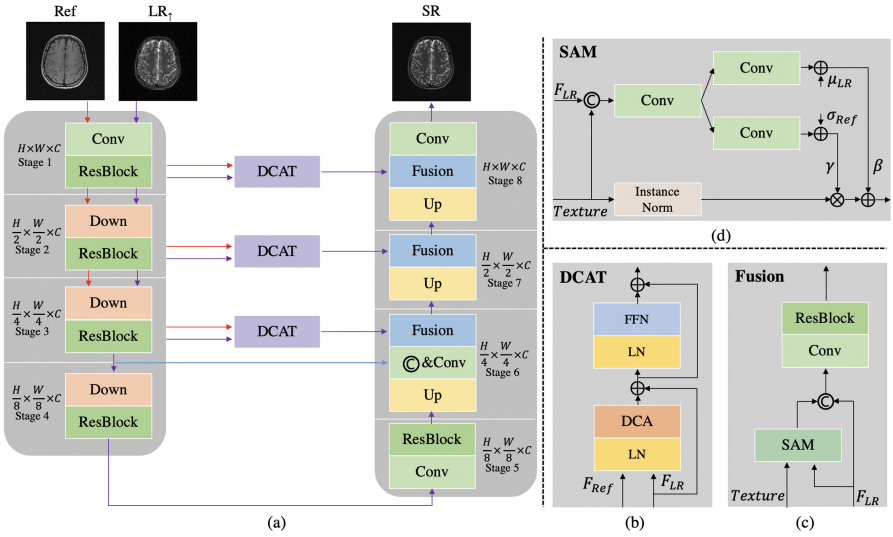


**Fig. 2.** (a) Network architecture of the proposed Dual Cross-attention Multi-contrast Super Resolution (DCAMSR). (b) Details of Dual Cross-Attention Transformer (DCAT). (c) Details of Fusion block. (d) Details of Spatial Adaptation Module (SAM).

respectively. Similarly, the multi-scale features for $Ref$ are extracted via the same encoder in stages 1–3 and denoted as $F_{Ref}^{H\times W}$, $F_{Ref}^{\frac{H}{2}\times\frac{W}{2}}$ and $F_{Ref}^{\frac{H}{4}\times\frac{W}{4}}$, respectively.

**Dual Cross-Attention Transformer (DCAT).** The DCAT consists of a DCA module, 2 LNs, and a FFN comprising several 1×1 convolutions.

The core of DCAT is dual cross-attention mechanism, which is diagrammed in Fig. 3. Firstly, we project $F_{LR}$ and $F_{Ref}$ to $q$, $k$ and $v$. For the two cross-attention branches, the linear layer weights for $q$ and $k$ are shared, while those for $v$ are different:

$$q_{share} = W^{q}_{share}(F_{LR}), k_{share} = W^{k}_{share}(F_{Ref}), \tag{1}$$

$$v_{spatial} = W^{v}_{spatial}(F_{Ref}), v_{channel} = W^{v}_{channel}(F_{Ref}), \tag{2}$$

where $q_{share}$,$k_{share}$,$v_{spatial}$ and $v_{channel}$ are the parameter weights for shared queries, shared keys, spatial value layer, and channel value layer, respectively. In spatial cross-attention, we further project $k_{share}$ and $v_{spatial}$ to $k_{project}$ and $v_{project}$ through linear layers, to reduce the computational complexity. The spatial and channel attentions are calculated as:

$$X_{spatial} = softmax(\frac{q_{share} \cdot k^{T}_{share}}{\sqrt{d}}) \cdot v_{project}, \tag{3}$$

$$X_{channel} = softmax(\frac{q^{T}_{share} \cdot k_{share}}{\sqrt{d}}) \cdot v^{T}_{channel}. \tag{4}$$

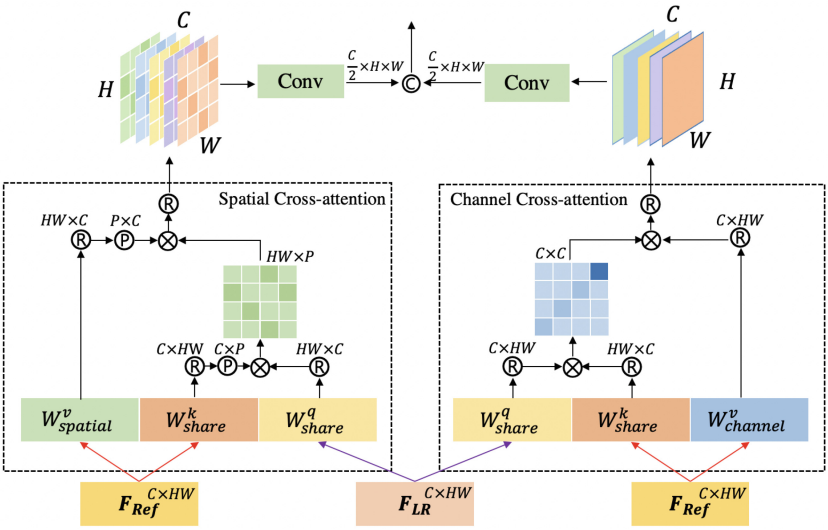


**Fig. 3.** Details of Dual Cross-Attention (DCA).

Finally, $X_{spatial}$ and $X_{channel}$ are reduced to half channel via $1 \times 1$ convolutions, and then concatenate to obtain the final feature:

$$X = Concat(Conv(X_{spatial}), Conv(X_{channel})). \tag{5}$$

For the whole DCAT, the normalized features $LN(F_{LR})$ and $LN(F_{Ref})$ are fed to the DCA and added back to $F_{LR}$. The obtained feature is then processed by the FFN in a residual manner to generate the texture feature. Specifically, the DCAT is summarized as:

$$X = F_{LR} + DCA(LN(F_{LR}), LN(F_{Ref})), \tag{6}$$

$$Texture = X + FFN(LN(X)). \tag{7}$$

Feeding the multi-scale features of $LR_{\uparrow}$ and $Ref$ to DCAT, we can generate the texture features in multi-scales, denoted as $Texture^{H\times W}$, $Texture^{\frac{H}{2}\times\frac{W}{2}}$, and $Texture^{\frac{H}{4}\times\frac{W}{4}}$.

**Decoder.** In the decoder, we start from the feature $F_{LR}^{\frac{H}{8}\times\frac{W}{8}}$ and process it with a convolution and a residual block. Then it is up-sampled and concatenated with $F_{LR}^{\frac{H}{4}\times\frac{W}{4}}$, and then feed to a convolution to further incorporate the both information. Next, the incorporated feature is fed to the Fusion module along with $Texture^{\frac{H}{4}\times\frac{W}{4}}$, to produce the fused feature at $\frac{H}{4} \times \frac{W}{4}$ scale, denoted as $Fused^{\frac{H}{4}\times\frac{W}{4}}$. $Fused^{\frac{H}{4}\times\frac{W}{4}}$ is then up-sampled and feed to Fusion along with $Texture^{\frac{H}{2}\times\frac{W}{2}}$, generating $Fused^{\frac{H}{2}\times\frac{W}{2}}$. Similarly, $Fused^{\frac{H}{2}\times\frac{W}{2}}$ is up-sampled

and feed to Fusion along with $Texture^{H\times W}$, generating $Fused^{H\times W}$. Finally, $Fused^{H\times W}$ is processed with a $1 \times 1$ convolution to generate $SR$.

In the Fusion module, following [13], the texture feature $Texture$ and input feature $F_{LR}$ are first fed to Spatial Adaptation Module (SAM), a learnable structure ensuring the distributions of $Texture$ consistent with $F_{LR}$, as shown in Fig. 2(d). The corrected texture feature is then concatenated with the input feature $F_{LR}$ and further incorporated via a convolution and a residual block, as shown in Fig. 2(c).

**Loss Function.** For simplicity and without loss of generality, $L_1$ loss between the restored $SR$ and ground-truth is employed as the overall reconstruction loss.

## 3 Experiments

**Datasets and Baselines.** We evaluated our approach on two datasets: 1) fastMRI, one of the largest open-access MRI datasets. Following the settings of SANet [10,11], 227 and 24 pairs of PD and FS-PDWI volumes are selected for training and validation, respectively. 2) M4Raw, a publicly available dataset including multi-channel k-space and single-repetition images from 183 participants, where each individual haves multiple volumes for T1-weighted, T2-weighted and FLAIR contrasts [18]. 128 individuals/6912 slices are selected for training and 30 individuals/1620 slices are reserved for validation. Specifically, T1-weighted images are used as reference images to guide T2-weighted images. To generate the LR images, we first converted the original image to k-space and cropped the central low-frequency region. For down-sampling factors of 2× and 4×, we kept the central 25% and 6.25% values in k-space, respectively, and then transformed them back into the image domain using an inverse Fourier transform. The proposed method is compared with SwinIR [7], ELAN [8], SANet (the journal version of MINet) [11], TTSR [12], and MASA [13].

**Implementation Details.** All the experiments were conducted using Adam optimizer for 50 epochs with a batch size of 4 on 8 Nvidia P40 GPUs. The initial learning rate for SANet was set to $4 \times 10^{-5}$ according to [11], and $2 \times 10^{-4}$ for the other methods. The learning rate was decayed by a factor of 0.1 for the last 10 epochs. The performance was evaluated for enlargement factors of 2× and 4× in terms of PNSR and SSIM.

**Quantitative Results.** The quantitative results are summarized in Table 1. The proposed method achieves the best performance across all datasets for both single image super-resolution (SISR) and multi-contrast super-resultion (MCSR). Specifically, our LR-guided DCAMSR version surpasses state-of-the-art methods such as ELAN and SwinIR in SISR, and even outperforms SANet (a MCSR method). Among the MCSR methods, neither SANet, TTSR or MASA achieves better results than the proposed method. In particular, the PSNR for

**Table 1.** Quantitative results on two datasets with different enlargement scales, in terms of PSNR and SSIM. SISR means single image super resolution, MCSR means multi-contrast super resolution. The best results are marked in for multi-contrast super resolution, and in blue for single image super resolution. Note that TTSR and MASA are not applicable to 2× enlargement based on their official implementation.

| Dataset | | fastMRI | | | | M4Raw | | | |
|---|---|---|---|---|---|---|---|---|---|
| Scale | | 2× | | 4× | | 2× | | 4× | |
| Metrics | | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| SISR | ELAN | 32.00 | 0.715 | 30.45 | 0.619 | 31.71 | 0.770 | 28.70 | 0.680 |
| | SwinIR | 32.04 | 0.717 | 30.58 | 0.624 | 32.08 | 0.775 | 29.42 | 0.701 |
| | DCAMSR | 32.07 | 0.717 | 30.71 | 0.627 | 32.19 | 0.777 | 29.74 | 0.709 |
| MCSR | SANet | 32.00 | 0.716 | 30.40 | 0.622 | 32.06 | 0.775 | 29.48 | 0.704 |
| | TTSR | NA | NA | 30.67 | 0.628 | NA | NA | 29.84 | 0.712 |
| | MASA | NA | NA | 30.78 | 0.628 | NA | NA | 29.52 | 0.704 |
| | DCAMSR | 32.20 | 0.721 | 30.97 | 0.637 | 32.31 | 0.779 | 30.48 | 0.728 |

MASA is even 0.18 dB lower than our SISR version of DCAMSR at 4× enlargement on M4Raw dataset. We attribute this performance margin to the difficulty of texture transformers in extracting similar texture features between Ref and $\text{Ref}_{\downarrow\uparrow}$. Despite the increased difficulty of super-resolution at 4× enlargement, our model still outperforms other methods, demonstrating the powerful texture transfer ability of the proposed DCA mechanism.

**Qualitative Evaluation.** Visual comparison is shown in Fig. 4, where the upsampled *LR*, the ground-truth *HR*, the restored *SR* and the error map for each method are visualized for 4× enlargement on both datasets. The error map depicts the degree of restoration error, where the more prominent texture indicating the poorer restoration quality. As can be seen, the proposed method produces the least errors compared with other methods.

**Ablation Study.** We conducted ablation experiments on the M4Raw dataset and the results are shown in Table 2. Three variations are tested: *w/o reference*,

**Table 2.** Ablation study on the M4Raw dataset with 4× enlargement.

| Variant | Modules | | | Metrics | | |
|---|---|---|---|---|---|---|
| | reference | multi-scale attention | channel attention | PSNR↑ | SSIM↑ | NMSE↓ |
| *w/o reference* | ✗ | ✓ | ✓ | 29.74 | 0.709 | 0.035 |
| *w/o multi-scale attention* | ✓ | ✗ | ✓ | 30.40 | 0.725 | 0.031 |
| *w/o channel attention* | ✓ | ✓ | ✗ | 29.79 | 0.712 | 0.035 |
| DCAMSR | ✓ | ✓ | ✓ | 30.48 | 0.728 | 0.029 |

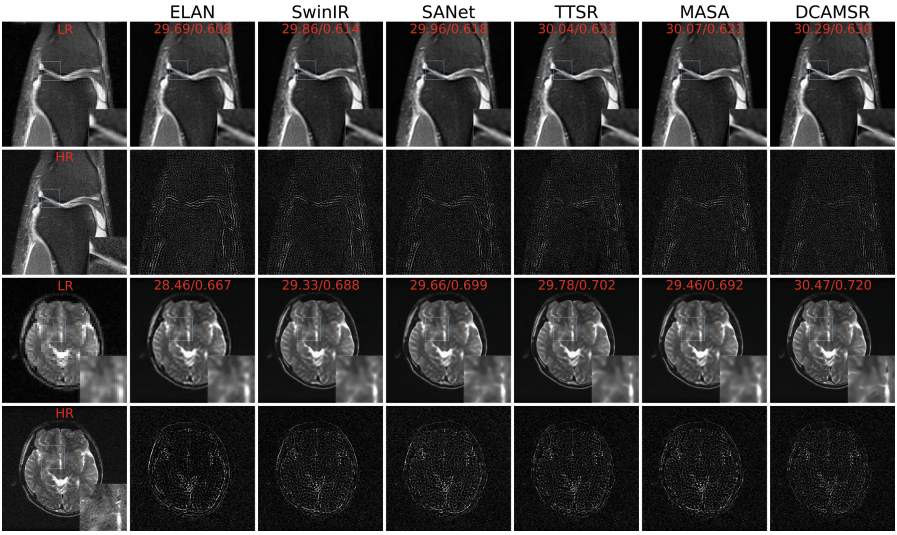


**Fig. 4.** Visual comparison of reconstruction results and error maps for 4× enlargement on both datasets. The upper two rows are fastMRI and the lower two rows are M4Raw.

where $LR_{\uparrow}$ is used as the reference instead of *Ref*; *w/o multi-scale attention*, where only the lowest-scale attention is employed and interpolated to other scales; and *w/o channel attention*, where only spatial attention is calculated. The improvement from *w/o reference* to DCAMSR demonstrates the effectiveness of MCSR compared with SISR. The performance degradation of *w/o multi-scale attention* demonstrates that the lowest-scale attention is not robust. The improvement from *w/o channel attention* to DCAMSR shows the effectiveness of the channel attention. Moreover, our encoder and decoder have comparable parameter size to MASA but we achieved higher scores, as shown in Table 1, demonstrating that the spatial search ability of DCAMSR is superior to the original texture transformer.

**Discussion.** Our reported results on M4Raw contain instances of slight inter-scan motion [18], demonstrating certain resilience of our approach to image misalignment, but more robust solutions deserve further studies. Future work may also extend our approach to 3D data.

## 4 Conclusion

In this study, we propose a Dual Cross-Attention Multi-contrast Super Resolution (DCAMSR) framework for improving the spatial resolution of MRI images. As demonstrated by extensive experiments, the proposed method outperforms

existing state-of-the-art techniques under various conditions, proving a powerful and flexible solution that can benefit a wide range of medical applications.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China under Grant 62101348, the Shenzhen Higher Education Stable Support Program under Grant 20220716111838002, and the Natural Science Foundation of Top Talent of Shenzhen Technology University under Grants 20200208, GDRC202117, and GDRC202134.

## References

1. Plenge, E., et al.: Super-resolution methods in MRI: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? Magn. Reson. Med. **68**(6), 1983–1993 (2012)
2. Van Reeth, E., Tham, I.W., Tan, C.H., Poh, C.L.: Super-resolution in magnetic resonance imaging: a review. Concepts Magn. Reson. Part A **40**(6), 306–325 (2012)
3. Feng, C.M., Wang, K., Lu, S., Xu, Y., Li, X.: Brain MRI super-resolution using coupled-projection residual network. Neurocomputing **456**, 190–199 (2021)
4. Li, G., Lv, J., Tong, X., Wang, C., Yang, G.: High-resolution pelvic MRI reconstruction using a generative adversarial network with attention and cyclic loss. IEEE Access **9**, 105951–105964 (2021)
5. Chen, Y., Xie, Y., Zhou, Z., Shi, F., Christodoulou, A.G., Li, D.: Brain MRI super resolution using 3D deep densely connected neural networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 739–742. IEEE (2018)
6. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
7. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: image restoration using Swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021)
8. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part XVII. LNCS, vol. 13677, pp. 649–667. Springer, Cham (2022)
9. Lyu, Q., et al.: Multi-contrast super-resolution MRI through a progressive network. IEEE Trans. Med. Imaging **39**(9), 2738–2749 (2020)
10. Feng, C.M., Fu, H., Yuan, S., Xu, Y.: Multi-contrast MRI super-resolution via a multi-stage integration network. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) MICCAI, Part VI. LNCS, vol. 13677, pp. 140–149. Springer, Cham (2021)
11. Feng, C.M., Yan, Y., Yu, K., Xu, Y., Shao, L., Fu, H.: Exploring separable attention for multi-contrast MR image super-resolution. arXiv preprint arXiv:2109.01664 (2021)
12. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5791–5800 (2020)

13. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: 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, pp. 6368–6377 (2021)
14. Li, G., et al.: Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20636–20645 (2022)
15. Jaegle, A., et al.: Perceiver IO: a general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
16. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: delving into efficient and accurate 3d medical image segmentation. arXiv preprint arXiv:2212.04497 (2022)
17. Zbontar, J., et al.: fastMRI: an open dataset and benchmarks for accelerated MRI. arXiv preprint arXiv:1811.08839 (2018)
18. Lyu, M., et al.: M4raw: a multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. Sci. Data **10**(1), 264 (2023)