3D CROSS-SCALE FEATURE TRANSFORMER NETWORK FOR BRAIN MR IMAGE SUPER-RESOLUTION

Wanqi Zhang[1], *Lulu Wang*[1], *Wei Chen*[1], *Yuanyuan Jia*[2], *Zhongshi He*[1,*], *Jinglong Du*[2,*]

[1] College of Computer Science, Chongqing University, Chongqing, China
[2] College of Medical Informatics, Chongqing Medical University, Chongqing, China

ABSTRACT

High-resolution (HR) magnetic resonance (MR) images could provide reliable visual information for clinical diagnosis. Recently, super-resolution (SR) methods based on convolutional neural networks (CNNs) have shown great potential in obtaining HR MR images. However, most existing CNN-based SR methods neglect the internal priors of the MR image, which hides the performance of SR. In this work, we propose a 3D cross-scale feature transformer network (CFTN) to utilize the cross-scale priors within MR features. Specifically, we stack multiple 3D residual channel attention blocks (RCABs) as the backbone. Meanwhile, we design a plug-in mutual-projection feature enhancement module (MFEM) to extract the target-scale features with HR cues, which is able to capture the global cross-scale self-similarity within features and can be flexibly inserted into any position of the backbone. Furthermore, we propose a spatial attention fusion module (SAFM) to adaptively adjust and fuse the target-scale features and upsampled features that are respectively extracted by the MFEM and the backbone. Experimental results show that our CFTN achieves a new state-of-the-art MR image SR performance.

***Index Terms*—** Magnetic resonance image, Cross-scale self-similarity, Super-resolution, Attention mechanism

1. INTRODUCTION

Magnetic resonance (MR) images with high-resolution (HR) could provide clear tissue structures and detailed anatomical information for clinical diagnosis. However, the resolution of MR images is often limited by the scanning time and hardware constraints, etc. Reducing the scanning thickness can improve the resolution of MR images but will dramatically increase expenses and surfer from motion artifacts. In contrast, super-resolution (SR) is a popular technology that reconstructs the HR image from the given low-resolution (LR) counterpart without changing the scanning protocol.

Some early efforts have provided feasible ideas for achieving MR image SR. Manjón et al. [1] combined the non-local self-similarity of the MR image to break the local limitation of interpolation. Later, to fully mine the internal details of the MR image, Rousseau et al. [2] and Plenge et al. [3] explored the non-local similarity of different scale images and realized the multi-modal and multi-slice MR image SR by directly searching for HR details from LR images. Although the above methods incorporated internal priors of the image, they could not achieve the desired effects with limited representation capability.

Recently, the convolutional neural network (CNN) provides the MR image SR with a promising methodology. The milestone work was SRCNN [4], which resolved natural image SR using a three-layer CNN. Afterward, Pham et al. [5] utilized the 3D SRCNN to improve the resolution of brain MR images. Subsequently, various architectures were proposed for MR image SR, which improved the representation ability of the model from different aspects. Specifically, Pham et al. [6] and Du et al. [7] constructed different residual networks for brain MR image SR, making the SR network go deeper. Wang et al.[8] and Du et al. [9] reused MR features through dense connections to promote the feature propagation of deep networks. Furthermore, inspired by the excellent performance of attention mechanism in natural image SR [10–12], Wu et al. [13] integrated self-attention into a deep network to perceive the global information in MR features. Wang et al. [14] used parallel channel and spatial attention to enable the model to adaptively treat MR features of different values. These CNN-based SR methods have made impressive progress, but only exploring external image resources, without mining the internal priors, still falls short in reconstructing fine texture for the specific MR image and large-scale SR.

In this work, we propose a novel 3D cross-scale feature transformer network (CFTN) which could make full use of the internal correlation statistical priors of cross-scale recurring patches in specific MR features. Specifically, we propose a flexible plug-in mutual-projection feature enhancement module (MFEM), which uses the mutual-projection to combine the global cross-scale self-similarity within features and map LR features to the most likely HR counterparts. Moreover, we propose a spatial attention fusion module (SAFM) to achieve accurate features fusion by suppressing the useless cross-scale information and refining meaningful regions.

* Corresponding authors. This work was supported by the Intelligent Medical Project of Chongqing Medical University (ZHYXQNRC202101), the Graduate Research and Innovation Foundation of Chongqing, China (Grant No.CYS21060).

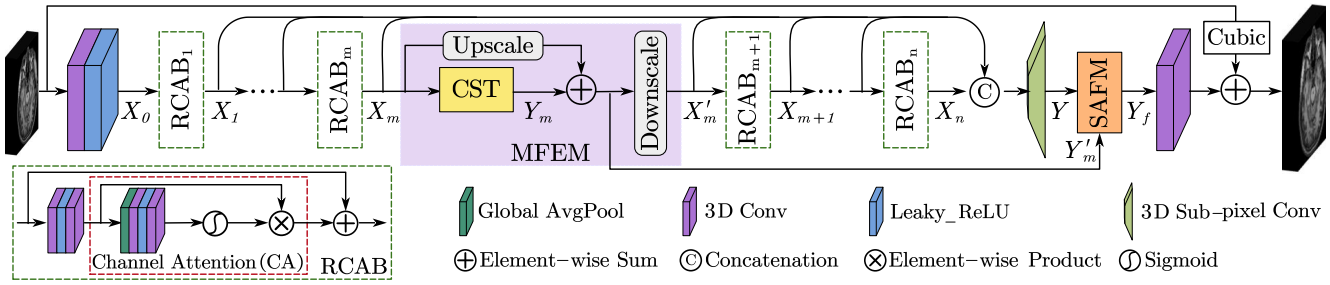

Fig. 1: The overall architecture of the proposed cross-scale feature transformer network (CFTN).

2. METHOD

2.1. Overview

The proposed cross-scale feature transformer network (CFTN) aims at predicting an HR MR image $\hat{I}_{HR}$ that is as close as possible to the ground truth I_{HR} from the given LR MR image I_{LR}. As shown in Fig.1, we extend the residual channel attention block (RCAB) in [11] to the 3D version and use it as the fundamental unit of our backbone network. Similar to [14, 15], we first extract the initial features X_0 from I_{LR} through a convolution layer and a leaky rectified linear unit (Leaky_ReLU). Then, we stack n RCABs to learn the mapping from X_0 to deep features X_n. It is worth noting that the proposed mutual-projection feature enhancement module (MFEM) can be placed between any adjacent RCABs as a network plug-in, which outputs target-scale features Y'_m with HR details and corresponding downsampling features X'_m. X'_m is used as the input of RCAB$_{m+1}$ to make the network explore more cross-scale information. The output features of each RCAB and X'_m are concatenated and upsampled to the target size. The upsampled features Y can be expressed as:

$$Y = F_{up}\left(\left[X_1, \cdots, X_m, X'_m, X_{m+1}, \cdots, X_n\right]\right) \quad (1)$$

where $F_{up}(\cdot)$ is the 3D sub-pixel convolution upsampling function, and $[\cdot]$ represents the feature concatenation operation. Later, the upsampled features Y and the HR features Y'_m extracted by MFEM are merged through the proposed spatial attention fusion module (SAFM), producing accurate HR fusion features Y_f. Finally, the desired $\hat{I}_{HR}$ can be defined as:

$$\hat{I}_{HR} = I^{\uparrow}_{LR} + F_{conv}\left(Y_f\right) \quad (2)$$

where $I^{\uparrow}_{LR}$ represents the interpolated LR MR image, $F_{conv}(\cdot)$ denotes the convolution operation.

The L_1 loss with the regular term is used to train the network, which can be formulated as:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{k=1}^{N} \|F\left(I^k_{LR}; \theta\right) - I^k_{HR}\|_1 + \lambda\|\theta\|_2 \quad (3)$$

where $F(\cdot)$ is the mapping function from I_{LR} to I_{HR} learned by our proposed CFTN with the parameter θ, and N represents the number of the training pairs. The hyper parameter λ of the regularization term is set to $1e-6$.

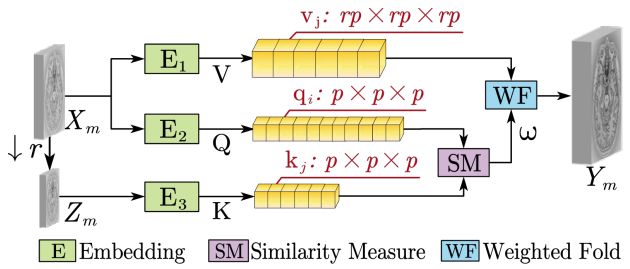

Fig. 2: The architecture of our cross-scale transformer (CST).

2.2. Mutual-projection Feature Enhancement Module

In [16, 17], researchers have proved that features in low-level tasks (e.g., image denoising and SR) contain richer details and textures at different scales. In this paper, we design the MFEM to mine the useful details by capturing the cross-scale self-similarity priors within the entire MR features. As shown in Fig.1, MFEM uses mutual-projection to learn and combine effective cross-scale residuals. For the input features X_m of MFEM, the final outputs Y'_m and X'_m are computed by:

$$Y'_m = F_{upscale}\left(X_m\right) + Y_m \quad (4)$$

$$X'_m = F_{downscale}\left(Y'_m\right) \quad (5)$$

where Y_m is the target-scale features with HR cues obtained from the cross-scale transformer (CST). $F_{upscale}(\cdot)$ and $F_{downscale}(\cdot)$ represent the upscaling function and downscaling function using transposed convolution and stride convolution, respectively. MFEM exchanges useful information between Y'_m and X_m of different scales via up/down scaling, which effectively enhances the same-scale features X_m by utilizing the cross-scale self-similarity priors in Y'_m.

Cross-scale Transformer (CST). CST aims to explore the global self-similarity priors of the MR features at different scales, as effective internal complements to the external information obtained from the training dataset. As shown in Fig.2, suppose the size of X_m is $H \times W \times L$, we first downsample X_m to $Z_m \in \mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times\frac{L}{r}}$ with the cubic operation according to the ratio r, where r is equal to the desired SR scaling factor. Then, we utilize three linear embeddings: $E_1(X_m)$, $E_2(X_m)$ and $E_3(Z_m)$ that realized by the $1\times1\times1$ convolution to obtain embedding features V, Q, and K. To capture the global cross-scale dependencies between Q and K, we use patch size as p and stride as g to unfold both Q and K into

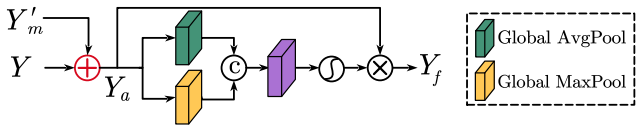

Fig. 3: The structure of the proposed spatial attention fusion module (SAFM).

patches, expressed as q_i and k_j, where $i \in \left[1, \frac{H}{g} \times \frac{W}{g} \times \frac{L}{g}\right]$ and $j \in \left[1, \frac{H}{rg} \times \frac{W}{rg} \times \frac{L}{rg}\right]$. For each patch q_i and k_j, we calculate their similarity $s_{i,j}$ by the inner product:

$$s_{i,j} = \langle q_i,\ k_j \rangle \tag{6}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product calculation. The $s_{i,j}$ is further normalized to produce the transformer weight $\omega_{i,j}$ between q_i and k_j, which is achieved by softmax operation:

$$\omega_{i,j} = \frac{\exp\left(s_{i,j}\right)}{\sum_j \exp\left(s_{i,j}\right)} \tag{7}$$

where $\omega_{i,j}$ is located at the coordinate (i, j) in the cross-scale transformer weight map ω. The larger the value of $\omega_{i,j}$, the higher the cross-scale correlation between q_i and k_j.

To enhance the feature of HR details, we assign $\omega_{i,j}$ to the corresponding HR patch v_j that extracted from V. Note that the size of v_j is r times that of q_i and k_j, which is obtained by unfolding V with the patch size of rp and the stride of rg. Formally, the weighting operation can be expressed as:

$$y_j = \sum_j \omega_{i,j} \otimes v_j \tag{8}$$

where $\otimes$ is the element-wise product, and y_j represents the HR patch obtained by weighted aggregation. Through transferring the cross-scale self-similarity to the HR receptive field, we could obtain the most likely HR correspondence from the LR patches themselves. After that, we fold y_j to get the output features $Y_m \in \mathbb{R}^{rH \times rW \times rL}$ of CST, which contains rich HR details that come from the cross-scale patches. Operationally, the above weighted aggregation and folding operations can be effectively implemented as a transposed convolution, in which v_j is the kernel and ω is the input.

2.3. Spatial Attention Fusion Module

Considering that the target-scale features Y'_m output by MFEM may contain some useless repetitive information, to integrate more meaningful parts of Y'_m into the backbone network, we utilize spatial attention to adaptively adjust and fuse features. Fig.3 illustrates our proposed SAFM. After adding Y'_m and the upsampled features Y to Y_a, we implement global average pooling and max pooling on Y_a along the channel axis to aggregate spatial features, which generates two different spatial information statistics: $Y_a^{avg} \in \mathbb{R}^{1 \times H \times W \times L}$ and $Y_a^{\max} \in \mathbb{R}^{1 \times H \times W \times L}$. Then, a convolution layer and a sigmoid function are applied to learn the weights of different positions, and the learned spatial weights are multiplied with Y_a to emphasize or suppress the spatial regions. Specifically, the final fusion features Y_f of SAFM can be given as:

$$Y_f = Y_a \otimes \sigma\left(F_{conv}\left([Y_a^{avg}, Y_a^{\max}]\right)\right) \tag{9}$$

where $\sigma(\cdot)$ represents the sigmoid function. By utilizing the inter-spatial relationship of features, SAFM makes the network pay more attention to the important informative regions, attenuates some useless features, and ensures more accurate and effective SR reconstruction.

Table 1: The effects of MFEM and SAFM

Baseline	✓	✓	✓
MFEM	✗	✓	✓
SAFM	✗	✗	✓
PSNR (dB) / SSIM	39.49 / 0.9836	39.63 / 0.9842	39.70 / 0.9847

3. EXPERIMENTS

3.1. Implementation Details

Following [14], we use 37 T1-weighted MR images (from KKI06 to KKI42) from the Kirby21 dataset [18] as the training samples. We evaluate the SR performance with the peak signal-to-noise ratio (PSNR) and structural similarity image index (SSIM) on two datasets: Kirby21 and BRATS [19]. As in [1, 5, 6, 8, 9, 14], the LR MR images are produced using Gaussian kernel (σ=1.0) and cubic downsampling.

We set the number of RCAB as 10, and insert one MFEM after RCAB_5. The convolution layers in CFTN have 48 channels with the filter size of $3 \times 3 \times 3$ except for those in CA and CST. In CST, we use 24 filters with the size of $1 \times 1 \times 1$ to carry out the convolution operation for linear embedding, and set p=3 and g=2 to unfold the feature maps. The setting of CA in RCAB are consistent with [11]. During training, the LR MR images are cropped to patches with the size $26 \times 26 \times 26$ and stride 13, and the mini-batch size is set to 16. The learning rate is fixed to $1e-4$, and the Adam optimizer with default setting is used for the parameter optimization. We train and test CFTN using PyTorch on an NVIDIA 3090 GPU.

3.2. Ablation Study

In this part, we evaluate the contribution of each component in CFTN and find the best structural configurations of the network. Each model is trained with the scale factor $\times 2$ for 100 epochs, and the results are calculated on the Kirby21 dataset.

The effects of the proposed MFEM and SAFM are shown in Table 1, where the baseline is the network without MFEM and SAFM. When the MFEM is inserted into the middle of the baseline, the PSNR increases by 0.14 dB, and the SSIM improves from 0.9836 to 0.9842. This proves the effectiveness of MFEM and indicates that the adaptive attention to cross-scale self-similarity within MR features could improve SR performance. Moreover, the network with both MFEM

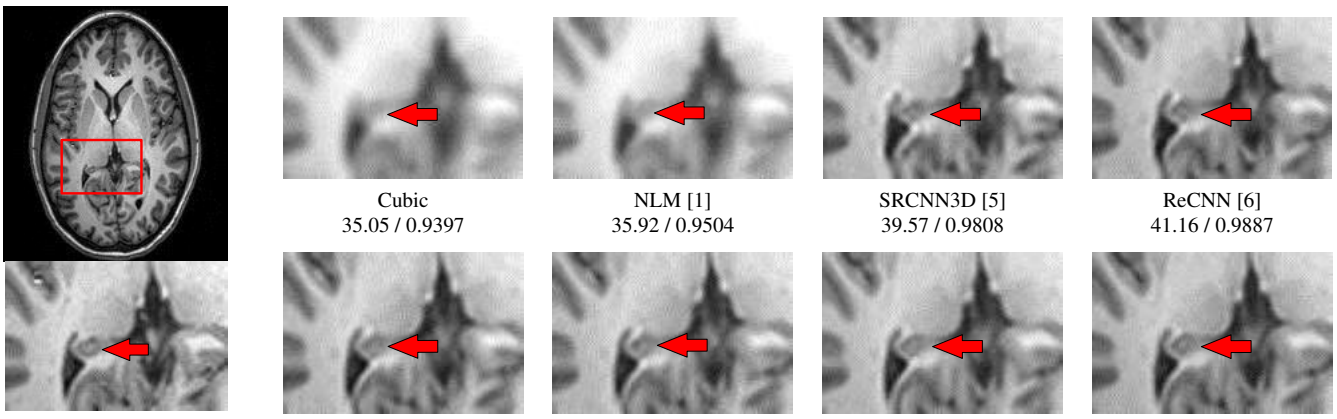

Fig. 4: Visual comparisons of a MR case (KKI03 in the Kirby21 dataset) that produced by different SR methods with scale $\times 2$.

Table 2: The effects of the position and number of MFEM

Pre		✓			✓	✓		✓
Mid			✓		✓		✓	✓
Post				✓		✓	✓	✓
PSNR	39.49	39.68	39.70	39.67	39.72	39.71	39.72	39.74
SSIM	0.9836	0.9844	0.9847	0.9844	0.9848	0.9847	0.9847	0.9849
Params	9.06M	13.35M	13.35M	13.35M	17.63M	17.63M	17.63M	21.93M

and SAFM further improves SR results in PSNR and SSIM metrics, which shows that SAFM could fuse the features more accurately. Overall, both MFEM and SAFM are essential for improving the performance of the SR network.

To explore the impact of the position and number of MFEM on the performance of CFTN, we insert MFEM into three typical positions: after the first RCAB, after the 5-th RCAB, and before the last RCAB. These three positions represent pre-processing, mid-processing, and post-processing. As can be seen from the first 4 columns of Table 2, inserting MFEM at any position brings obvious PSNR and SSIM value improvements, and the best performance gain is obtained by using it as the mid-processing. From the last 4 columns, we find that using more MFEMs slightly improves the results, but adding one MFEM increases the parameters by around 4.3M. Thus, we only insert one MFEM in the middle of the CFTN to achieve a trade-off between performance and complexity.

3.3. Comparisons with State-of-the-Art Methods

To evaluate the effectiveness of our proposed SR method, we compare CFTN with representative methods: Cubic, NLM [1], SRCNN3D [5], ReCNN [6], DDSR [9], EDDSR [8], FASR-L_1 [14] on the Kirby21 dataset (from KKI01 to KKI05) and the BRATS dataset (five randomly selected T1 contrast-enhanced MR images), which were collected from healthy volunteers and brain tumor patients, respectively.

Table 3 shows the quantitative comparisons for $\times 2$ and $\times 3$ MR image SR. Compared with other methods, our CFTN performs the best on two datasets with two scale factors. Specifically, for the Kirby21 dataset, the PSNR and SSIM gains of our model over FASR-L_1 are 0.26 dB and 0.0006 on scale $\times 2$, and CFTN also obtains the best PSNR of 36.03 dB, SSIM of 0.9612 with scale $\times 3$. For the BRATS dataset, CFTN also produces the best results in both PSNR and SSIM values over scale $\times 2$ and $\times 3$. These quantitative experimental results demonstrate the superiority of our CFTN.

Table 3: Quantitative evaluations of CFTN with start-of-the-art MR image SR methods on two datasets. The best and second best results are indicated in **bold** and underlined.

Method/Scale	Kirby21				BRATS			
	$\times 2$		$\times 3$		$\times 2$		$\times 3$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Cubic	33.35	0.9217	31.87	0.8939	35.46	0.9660	34.43	0.9640
NLM [1]	34.19	0.9349	33.01	0.9173	36.12	0.9712	35.40	0.9645
SRCNN3D [5]	37.51	0.9735	34.03	0.9384	38.67	0.9847	36.09	0.9694
ReCNN [6]	38.80	0.9797	35.20	0.9530	39.14	0.9887	36.55	0.9729
DDSR [9]	39.00	0.9811	34.99	0.9500	39.47	0.9878	36.49	0.9675
EDDSR [8]	39.24	0.9822	35.95	0.9599	39.60	0.9890	36.82	0.9728
FASR-L_1 [14]	39.44	0.9841	35.71	0.9577	39.97	0.9908	37.02	0.9750
CFTN(Ours)	**39.70**	**0.9847**	**36.03**	**0.9612**	**40.13**	**0.9910**	**37.13**	**0.9760**

Fig.4 visualizes an example slice of the Kriby21 case (KKI03) reconstructed by different SR methods. As can be seen, compared with other methods, the MR image reconstructed by CFTN has more accurate details and sharper edges (indicated by the red arrow), which produces the most similar visual effect to the ground truth. These visual comparisons show that CFTN can recover more HR MR information.

4. CONCLUSION

In this paper, we propose a 3D cross-scale feature transformer network named CFTN for MR image SR. CFTN adaptively integrates the global cross-scale self-similarity priors of MR features into the deep network. Through the mutual-projection feature enhancement module, CFTN models the cross-scale correlation within the entire MR features, which effectively mines the internal potential HR details. Moreover, CFTN could attenuate the useless cross-scale features and enhance important informative regions through the spatial attention fusion module. Experimental results show the superiority of our proposed CFTN against the state-of-the-art MR image SR methods.

References

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

[2] François Rousseau, Alzheimer's Disease Neuroimaging Initiative, et al. A non-local approach for image super-resolution using intermodality priors. *Medical Image Analysis*, 14(4):594–605, 2010.

[3] Esben Plenge, Dirk HJ Poot, Wiro J Niessen, and Erik Meijering. Super-resolution reconstruction using cross-scale self-similarity in multi-slice MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 123–130. Springer, 2013.

[4] 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 (ECCV)*, pages 184–199. Springer, 2014.

[5] Chi-Hieu Pham, Aurélien Ducournau, Ronan Fablet, and François Rousseau. Brain MRI super-resolution using deep 3D convolutional networks. In *IEEE 14th International Symposium on Biomedical Imaging (ISBI)*, pages 197–200. IEEE, 2017.

[6] Chi-Hieu Pham, Carlos Tor-Díez, Hélène Meunier, Nathalie Bednarek, Ronan Fablet, Nicolas Passat, and François Rousseau. Multiscale brain MRI super-resolution using deep 3D convolutional networks. *Computerized Medical Imaging and Graphics*, 77:101647, 2019.

[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.

[8] Lulu Wang, Jinglong Du, Ali Gholipour, Huazheng Zhu, Zhongshi He, and Yuanyuan Jia. 3D dense convolutional neural network for fast and accurate single MR image super-resolution. *Computerized Medical Imaging and Graphics*, 93:101973, 2021.

[9] Jinglong Du, Lulu Wang, Ali Gholipour, Zhongshi He, and Yuanyuan Jia. Accelerated super-resolution MR image reconstruction via a 3D densely connected deep convolutional neural network. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 349–355. IEEE, 2018.

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[11] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[12] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074, 2019.

[13] Yan Wu, Yajun Ma, Jing Liu, Jiang Du, and Lei Xing. Self-attention convolutional neural network for improved MR image reconstruction. *Information Sciences*, 490:317–328, 2019.

[14] Lulu Wang, Jinglong Du, Huazheng Zhu, Zhongshi He, and Yuanyuan Jia. Brain MR image super-resolution using 3D feature attention network. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1151–1155. IEEE, 2020.

[15] Zongcai Du, Jie Liu, Jie Tang, and Gangshan Wu. Anchor-based plain net for mobile image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2494–2502, 2021.

[16] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[17] Yiqun Mei, Yuchen Fan, Yulun Zhang, Jiahui Yu, Yuqian Zhou, Ding Liu, Yun Fu, Thomas S Huang, and Humphrey Shi. Pyramid attention networks for image restoration. *arXiv preprint arXiv:2004.13824*, 2020.

[18] Bennett A Landman, Alan J Huang, Aliya Gifford, Deepti S Vikram, Issel Anne L Lim, Jonathan AD Farrell, John A Bogovic, Jun Hua, Min Chen, Samson Jarso, et al. Multi-parametric neuroimaging reproducibility: a 3-T resource study. *Neuroimage*, 54(4):2854–2866, 2011.

[19] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.