2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428112

PYRAMID ORTHOGONAL ATTENTION NETWORK BASED ON DUAL SELF-SIMILARITY FOR ACCURATE MR IMAGE SUPER-RESOLUTION

Xiaowan Hu[1, 2], *Haoqian Wang*[1, 2] *Yuanhao Cai*[1], *Xiaole Zhao*[3], *Yulun Zhang*[4]

[1] The Shenzhen International Graduate School, Tsinghua University, China
[2] The Shenzhen Institute of Future Media Technology, Shenzhen 518071, China
[3] Southwest Jiaotong University, China [4] Northeastern University, US.
{hu-xw19, wanghaoqian, cyh20}@tsinghua.edu.cn, zxlation@foxmail.com, yulun100@gmail.com

ABSTRACT

For magnetic resonance (MR) images sharing visual characteristics, the internal structure repetitions of different scales are considerable image-specific priors. Following the traditional algorithms, we try to combine external dataset-driven learning with the internal self-similarity for MR image super-resolution (SR). We propose a pyramid orthogonal attention network (POAN) based on dual self-similarity. On the one hand, by combining the point-similarity and the pyramid-similarity, sufficient spatial autocorrelation is explored to alleviate less training data limitation. On the other hand, the non-reduction channel attention mechanism maximizes inter-channel dependence. It increases the probability of the high-frequency region (e.g., structural textures and edges) being activated while suppresses low-frequency regions (e.g., background) adaptively. Out proposed POAN reconstructs the MR image under the guidance of pyramid orthogonal attention. Extensive experiments demonstrate that our method obtains the best results compared with state-of-the-art MR image SR methods quantitatively and visually.

Index Terms— Magnetic resonance (MR) images, super-resolution, self-similarity, pyramid orthogonal attention

1. INTRODUCTION

Compared with other medical images, high-quality MR images has good imaging resolution for small structures and detailed textures. So they are more suitable for clinical diagnosis and quantitative image analysis [1]. However, due to hardware limitations, the higher the resolution of the output MR image, the longer the scanning time required and the more expensive imaging physical equipment. In recent years, single-image super-resolution (SR) technologies have achieved excellent performance on natural images. However, its role in medical image super-resolution has not been explored extensively. For balancing cost and imaging quality, it is essential to use the advanced SR method to restore high-resolution (HR) MR images from low-quality images accurately [2].

The pattern repeatability prior in images has been proved to be important guidance in image restoration. There are many good attempts in traditional methods to utilize the prior knowledge of image [3,4]. Subsequently, self-similarity (SS) as an important natural property in images extensively explored on image SR tasks [5–7]. Relying on its powerful learning and expression capabilities, the convolutional neural network (CNN) shows great superiority over traditional algorithms in extracting image SS. NLRN [8], RNAN [9], and SAN [10] incorporate point-similarity features. Mei et al. proposed the pyramid attention [11] and further explored the cross-scale non-local SS in the deep network [12]. Due to the imaging characteristics of human tissues, MR images have various repeatable structure redundancy. So extracting rich SS information can improve SR performance and reduce the excessive model dependence on external datasets.

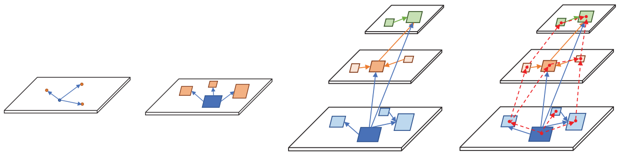
(a) Point SS (b) Patch SS (c) Pyramid SS (d) Dual SS

Fig. 1. Different self-similarity (SS) structures comparison.

Different self-similarity structures search different matching targets. We compare different SS extraction structures in Fig 1. The non-local attention at a fixed scale limits the search area. As shown in Fig 1(a), the point SS calculates pair-wise pixel correlation and captures the long-range dependence of the entire image. The patch SS in Fig 1(b) computes pair-wise patch correlation on a single scale. The pyramid SS shown in Fig 1(c) expands the patch search space from a single feature map to a multi-scale feature pyramid for better matches. However, the multi-scale pyramid attention focuses more on large-scale SS with regular shapes, which is very unfavorable for MR images that contain rich details and changeable micro-structures. Existing methods that separate the point SS and patch SS will lose the feature specificity of MR images. Therefore, we explore an efficient way to combine point-similarity and multi-scale patch-similarity simultaneously. As shown in Fig 1(d), the dual SS computes pair-wise pixel correlation and pair-wise patch correlation on the

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

pyramid scale. The point-similarity helps reconstruct subtle structural textures, and the single-scale patch-similarity avoids errors caused by pixel-by-pixel matching. The multi-scale dual SS fusion mechanism tends to find the best matches of cross-scale correlation patches faster.

Extensive background area in MR images will interfere with SS feature extraction in salient regions. The existing spatial attention mechanism treats the high-frequency textures and low-frequency backgrounds equally, which will bring feature redundancy. It is difficult for the network to pay more attention to sophisticated textures. Although the channel attention mechanism can distinguish salient features adaptively, most models use channel reduction operations to reduce complexity [13, 14]. We find the sudden channel narrowing in the middle layer will destroy the dense channel correlation and prediction accuracy. So we design the non-reduction channel attention block (NCAB) to preserve the non-linear correlations and evaluate spatial SS importance. Besides, MR images lack color information and contain rich local texture details with variable gradients. Compared with the first-order global pooling, the high-order statistics are more capable of discovering high-frequency features [10]. Therefore, we replace the global pooling with covariance pooling. The probability of being activated can be adjusted adaptively according to the feature complexity for a specific position.

To address existing problems and limitations of MR images SR, we incorporate the pyramid orthogonal attention block (POAB) based on dual self-similarity to the deep network in this paper. Based on the characteristics of MR images, we build a complete spatial statistical correlation and channel self-attention mapping to guide the HR restoration. The spatial attention performs full-scale SS matching from local to global in the degraded MR image. Moreover, we design the NCAB for efficient high-level interactive attention and adaptive region activation. The main contributions of our pyramid orthogonal attention network (POAN) are as follows:

- We design the pyramid dual self-similarity block (PDSB) containing the non-local point-similarity and the multi-scale pyramid-similarity simultaneously. The fused SS mechanism covers the global full-scale receptive field and models the correlation between features of different sizes in the entire input image.
- The channel reduction operation is removed to maximize the continuous cross-channel interaction. Furthermore, we use covariance pooling instead of global pooling for capturing high-order gradient differences of gray value. An orthogonal attention structure is proposed by combining the spatial attention and channel attention with a multi-level residual structure.
- An extensive ablation study verifies the effectiveness and efficiency of each component in the proposed POAN. Compared with recently leading methods, our novel method achieves state-of-the-art performance in quantitative and qualitative experiments.

2. PROPOSED METHOD

2.1. Pyramid Dual Self-Similarity

Point-Similarity: As shown in Fig. 1(a), for an entire image, we search for points similar to the current pixel to get the point-similarity matrix. By traversing the correlation of pixels at two different coordinates through specific transformation functions, the feature response that includes non-local self-similarity attention is obtained. Formally, given the image feature map X, the point-similarity is defined as:

$$V_{i,j} = \sum_{m,n} \frac{\exp\left(\Phi\left(X_{i,j}, X_{m,n}\right)\right)}{\sum_{g,h}\exp\left(\Phi\left(X_{i,j}, X_{g,h}\right)\right)} \Psi\left(X_{m,n}\right), \tag{1}$$

where (i,j), (m,n), and (g,h) are pairs of coordinates of X. We define $\{\Psi\left(X_{m,n}\right) = W_{\Psi}X_{m,n}\}$ as the feature transformation function, and $\Phi(\cdot,\cdot)$ is the correlation function for feature maps to measure similarity, which is denoted as:

$$\Phi\left(X_{i,j}, X_{m,n}\right) = e^{g\left(X_{i,j}\right)^{T} f\left(X_{m,n}\right)}, \tag{2}$$

where $g\left(X_{i,j}\right) = W_g X_{i,j}$ and $f\left(X_{m,n}\right) = W_f X_{m,n}$ are transformation functions that generate the new representation of $X_{i,j}$ and $X_{m,n}$. The output response $V_{i,j}$ obtains point-similarity information from feature maps by calculating all pixels.

Patch-Similarity: In addition to similar pixels, some similar image patches also tend to be repeated at multiple locations on the current scale. The patch closest to the query patch is found by matching the query patch with other image patches. Compared with point-similarity, extracting patch-similarity with an appropriate neighborhood size can effectively reduce the amount of calculation and capture a wider range of correlations. As shown in Fig. 1(b), we define the image patch of $k \times k$ as the minimum search unit. Moreover, similar matching pairs are searched in the global scope of the entire image. The patch-similarity item can be adapted as:

$$V_{ki,kj} = \sum_{m,n} \frac{\exp\left(\Phi\left(X_{ki,kj}, X_{km,kn}\right)\right)}{\sum_{g,h}\exp\left(\Phi\left(X_{ki,kj}, X_{kg,kh}\right)\right)} \Psi\left(X_{km,kn}\right), \tag{3}$$

where $V_{ki,kj}$ is the patch-similarity matrix of each feature patch of size $k \times k$ located at (ki, kj). We look for other similar patches for each position within the entire feature map at the current scale. And the patch-correlation information is obtained directly through global weighted normalization.

Pyramid-Similarity: Since multi-scale recursion will not lose features, the structural information will still be well preserved after scaling down. Therefore, by scaling the original input image, the patch-similarity can be extended to multiple scales. We built the image-space pyramid of the original LR image to get the cross-scale patch-similarity. As shown in Fig. 1(c), to make full use of the image prior, the single-scale attention can be extended to the pyramid attention, which can calculate the correlation between multiple scales. In such a unit, the multi-level block matching correspondence is captured on the entire feature pyramid.

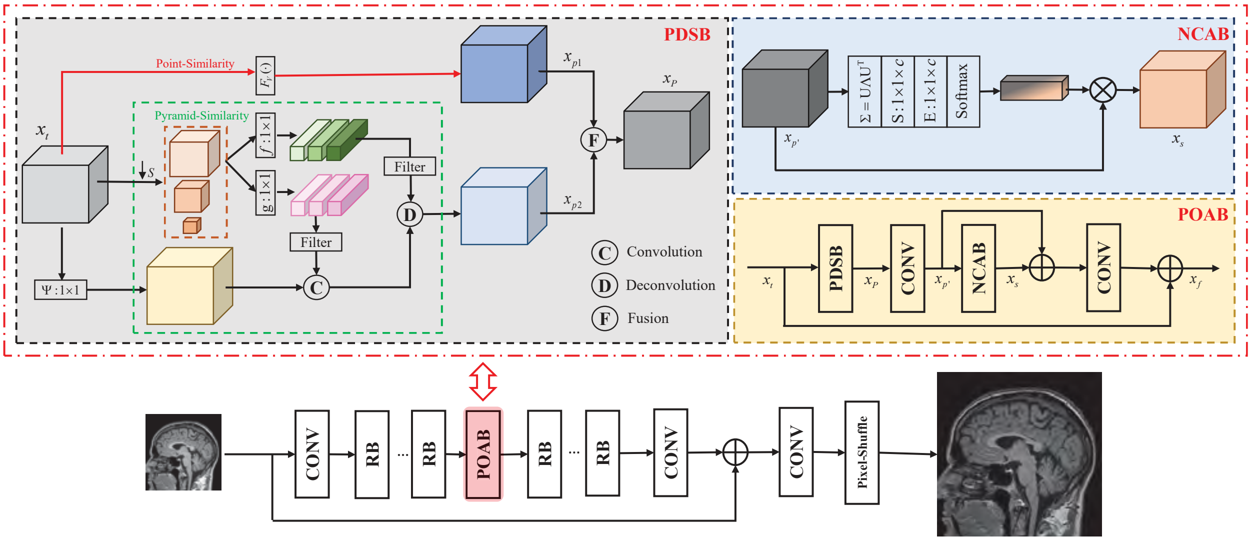

Fig. 2. The architecture of the proposed POAN and the internal implementation details of different network components. PDSB fuses dual self-similarity from the two branches of point-similarity and pyramid-similarity. NCAB captures high-order attention in the direction of the channel. POAB is composed of POSB and NCAB through multi-level residual structure.

In particular, a series of scales are given as $S = \{s_0, s_1, s_2, \ldots, s_n\}$, where n represents the number of pyramid levels. By traversing each value in the set S, the regional feature descriptors with different pyramid levels of the input X are obtained. Therefore, in every specific level, the pyramid-similarity matrix $V_{Si,Sj}$ can be expressed as:

$$\begin{aligned} V_{s_e i, s_e j} &= \sum_{m,n} \frac{\exp\left(\Phi\left(X_{s_e i, s_e j}, X_{s_e m, s_e n}\right)\right)}{\sum_{g,h} \exp\left(\Phi\left(X_{s_e i, s_e j}, X_{s_e g, s_e h}\right)\right)} \Psi\left(X_{s_e m, s_e n}\right), \\ e &= 0,1,2,\ldots,n, \\ V_{Si,Sj} &= \left[V_{s_0 i, s_0 j}, V_{s_1, s_1, j}, V_{s_2 i, s_2 j}, \ldots, V_{s_n i, s_n j}\right], \end{aligned} \tag{4}$$

where e represents different levels of the pyramid, and $V_{s_e i, s_e j}$ corresponds to the self-similarity matrix calculated on the corresponding scale. The $[\cdots]$ represent the concatenation operation between channels. Compared with a single scale self-similarity, the pyramid-similarity summarizes regional descriptors of various sizes. When concatenating patch-similarities of different scales together, the response contains richer and more authentic information intuitively.

Pyramid Dual Self-Similarity: Pyramid-similarity includes the autocorrelation of multi-scale image patches. The point-similarity matrix captures the long-range dependence of each pixel in a non-local range. In this paper, as shown in Fig. 1(d), we combine the similarities of the above two types to construct a pyramid dual self-similarity block (PDSB). We define the function of point SS as $F_{V_{i,j}}$ and the function of pyramid SS as $F_{V_{Si,Sj}}$, then the fusion feature is expressed as:

$$\begin{aligned} x_{p1} &= F_{V_{i,j}}\left(x_t\right), \\ x_{p2} &= F_{V_{Si,Sj}}\left(x_t\right), \\ x_P &= F_{c'}\left(\left[x_{p1}, x_{p2}\right]\right), \end{aligned} \tag{5}$$

where $F_{c'}(\cdot)$ means the convolution layer with 1×1 kernel. x_{p1} represents the output feature of the point-similarity operation, and x_{p2} represents the output feature of the pyramid-similarity operation. We combine x_{p2} and x_{p2} through channel concatenation and convolution layer to obtain x_P as the fused pyramid dual self-similarity feature. The internal operations in the proposed PDSB are shown in Fig. 2. This block is mainly implemented by basic convolution and deconvolution operations. It should be noted that the size of the input feature x_t and the output feature x_P are the same.

2.2. Pyramid Orthogonal Attention

Non-Reduction Channel Attention: As shown in Fig. 2, we use the non-reduction channel attention mechanism to obtain distinguishing feature representation. The previous method obtains the first-order statistical information through the global pooling layer and channel reduction, which destroy the serial correlation between channels. So we use the global covariance pooling to extract the high-order statistical distribution of the image. The number of remapping convolution kernels is set to be equal to the input feature maps.

The subsequent operation is similar to the usual channel attention mechanism. Particularly, we get the final attention vector through the non-reduction squeeze-and-excitation operation. We define the input of NCAB is $x_{p'} \in R^{h\times w\times c}$. The process can be expressed as follows:

$$\begin{aligned} \Sigma &= x_{p'} \bar{\mathrm{I}} x_{p'}^T = \mathrm{U}\Lambda\mathrm{U}^\mathrm{T}, \\ \hat{Y} &= \Sigma^\alpha = \mathrm{U}\Lambda^\alpha\mathrm{U}^\mathrm{T}, \\ z &= F_{GCP}(\hat{Y}), \\ x_s &= \mathrm{Softmax}\left(F_S\delta\left(F_E z\right)\right) \otimes x_{p'}, \end{aligned} \tag{6}$$

where Σ is the corresponding covariance matrix. We set $\bar{\mathrm{I}} = \frac{1}{h\times w}\left(\mathrm{I} - \frac{1}{h\times w}1\right)$, where I means identity matrix. F_{GCP} is the global covariance pooling function, which averages the power of the eigenvalues to obtain the normalized vector z. The δ stands for the ReLU function. F_S and F_E represent the squeeze and excitation operations with the same number of channel filters. The final channel attention vector is obtained through the remapping of the Softmax function, and then multiplied by the input feature $x_{p'}$ of NCAB to obtain the final high-order attention feature output x_s.

Pyramid Orthogonal Attention Block: PDSB outputs the self-similarity attention feature in the spatial domain while NCAB outputs the high-order non-reduction attention feature in the channel domain. We concatenate and merge the outputs in these two directions to obtain pyramid orthogonal attention feature based on dual self-similarity. As shown in Fig. 2, the multi-level residual mechanism in POAB promotes information flow and stabilizes the training process through skip connections between layers. The internal implementation of POAB can be expressed by the following formula:

$$\begin{aligned} x_P &= F_{PSDB}\left(x_t\right), \\ x_{p'} &= F_{CONV}\left(x_P\right), \\ x_s &= F_{NCAB}\left(x_{p'}\right), \\ x_f &= F_{CONV}\left(x_s + x_{p'}\right) + x_t, \end{aligned} \tag{7}$$

where F_{PSDB} and F_{NCAB} represent the corresponding functions of PDSB and NCAB respectively. F_{CONV} is defined as a convolutional layer with a 1×1 convolution kernel, which is used to adjust the number of channels of the output. x_f is the output after the orthogonal attention. We use the residual structure in POAB to create an identity mapping that is easier to learn by adding cross-layer elements directly.

2.3. Pyramid Orthogonal Attention Network

The architecture of the proposed POAN is shown in Fig. 2. We use a ResNet [15] with multiple residual blocks (RB) stacked as the backbone. The proposed POAB is directly embedded in the network, which can explain the effectiveness of our method more intuitively. For reducing calculations and redundancy, we follow the enhanced deep residual network [16] and remove some redundant layers. We use pixel-shuffle layer based on sub-pixel convolution and multi-channel recombination to perform upsampling of different scales in image reconstruction. Besides, considering the disappearance of gradients and shallow features in network training, we introduce a global skip connection in the main branch.

2.4. Training Objective

Some excellent loss functions have appeared in the natural images SR task recently. However, excessive smoothing and generating adversarial learning will bring the risk of texture and structure distortion to MR images. So we choose the L_1 loss as the optimization goal. Given a set including N-paired LR and HR training images $\{I_i^{LR}, I_i^{HR}\}_i^N$. We define the corresponding function between the LR and SR image as F_{POAN} and minimize the loss between the HR and SR image:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\| I_i^{HR} - F_{POAN}\left(I_i^{LR}; \theta\right) \right\|_1, \tag{8}$$

where θ denotes the parameters of our model, I_i^{HR} is the ground truth corresponding to I_i^{LR}.

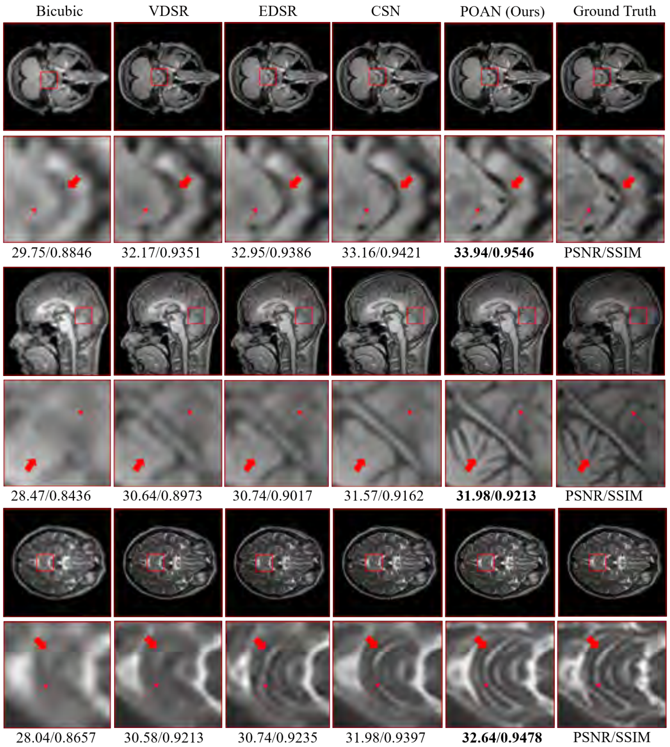

Fig. 3. The visual comparison on PD (top) , T1 (middle), and T2 (bottom) dataset with scaling factor SR×4.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

We use the IXI dataset[1] for training. It comprises three types of MR images (PD, T1, and T2), and the numbers are 578, 581, and 578, respectively. Each image is a 3D volume with a size of 240×240×96 (height×width×chnnel) and bicubic downsample with 3 scaling factors (×2, ×3 and ×4). The peak signal-to-noise ratio (PSNR) and structural similarity index metric (SSIM) [17] are used for quantitative evaluation.

3.2. Implementation Details

There are 40 RBs in the POAN, and a POAB is inserted after the 20th block. In PDSB, we set $S = \{0.8, 0.6, 0.4\}$ to construct a 3-level SS feature pyramid. The 3 × 3 patches and 1 × 1 points are used for the dual SS extraction. The convolution and deconvolution filters in PDSB have an equal size. 96 LR 64×64 patches are integrated as a training batch, and the number of all feature maps is set to 64. Our model is trained by ADAM optimizer with $\{\beta_1 = 0.9, \beta_2 = 0.999,\ epsilon = 10^{-8}\}$. The learning rate is initialized as 1×10^{-4} and then reduce to half after every 200 epochs. We implement all models with the Pytorch framework and train them on NVIDIA GeForce GTX 1080 Ti GPU for 1000 epochs.

3.3. Comparisons with Other Methods

Some advanced SR methods on MR images are compared quantitatively and qualitatively, including NLM [18], SRCNN [19], VDSR [20], RDN [21], EDSR [16], and CSN [22].

[1] http://brain-development.org/ixi-dataset/

Table 1. Quantitative comparison between different SR methods. The maximal PSNR (dB) and SSIM values of each comparison cell are marked in *red*, and the second ones are marked in *blue* (PSNR / SSIM).

Data	Scale	Bicubic	NLM	SRCNN	VDSR	RDN	EDSR	CSN	POAN (Ours)	POAN+ (Ours)
PD	×2	35.04/0.9664	37.26/0.9773	38.96/0.9861	39.97/0.9861	40.31/0.9870	39.87/0.9857	41.28/0.9895	41.59/0.9901	41.92/0.9906
	×3	31.20/0.9230	32.81/0.9436	33.60/0.9516	34.66/0.9599	35.08/0.9628	34.39/0.9678	35.87/0.9693	36.46/0.9731	36.65/0.9739
	×4	29.13/0.8799	30.27/0.9044	31.10/0.9181	32.09/0.9311	32.73/0.9387	31.80/0.9284	33.40/0.9486	33.94/0.9546	34.18/0.9563
T1	×2	33.80/0.9525	35.80/0.9685	37.12/0.9761	37.67/0.9783	37.95/0.9795	37.56/0.9774	38.27/0.9810	38.79/0.9827	38.86/0.9829
	×3	30.15/0.8900	31.74/0.9216	32.17/0.9276	32.91/0.9378	33.31/0.9430	32.76/0.9347	33.53/0.9464	34.22/0.9530	34.34/0.9538
	×4	28.28/0.8312	29.31/0.8655	29.90/0.8796	30.57/0.8932	31.05/0.9042	30.46/0.8902	31.23/0.9093	32.06/0.9231	32.21/0.9250
T2	×2	33.44/0.9589	35.58/0.9722	37.32/0.9796	38.65/0.9836	38.75/0.9838	38.28/0.9824	39.71/0.9863	40.43/0.9877	40.56/0.9879
	×3	29.80/0.9093	31.28/0.9330	32.20/0.9440	33.47/0.9559	33.91/0.9591	33.15/0.9528	34.64/0.9647	35.25/0.9686	35.43/0.9694
	×4	27.86/0.8611	28.85/0.8875	29.69/0.9052	30.79/0.9240	31.45/0.9324	30.52/0.9198	32.05/0.9413	32.60/0.9477	32.83/0.9495

Table 2. Performance comparison of different SS combination on PD (×2) in 400 epochs.

Method	Point SS	Patch SS	Pyramid SS	PSNR (dB)
A				36.45
B	✓			37.72
C		✓		37.67
D			✓	38.31
E	✓	✓		38.94
F		✓	✓	39.35
G	✓		✓	40.21
H (Ours)	✓	✓	✓	**41.53**

Table 3. Performance comparison of different attention domain combinations on T1 (4×) in 400 epochs.

Method	Channel Domain	Spatial Domain	PSNR (dB)
A			26.71
B	✓		28.94
C		✓	30.37
D (Ours)	✓	✓	**31.62**

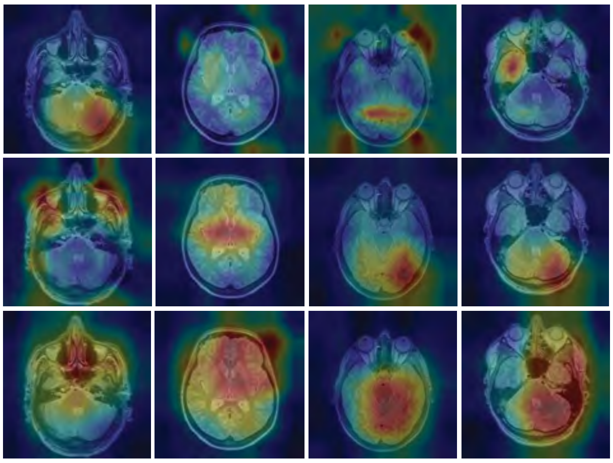

Fig. 4. The visualization of CAMs before attention layer (top), after the CAB (middle), and after the NCAB (bottom).

Quantitative Results: As shown in Tab. 1, all the quantitative results for ×2, ×3, and ×4 on PD, T1, and T2 datasets are reported. We use geometric self-ensemble like the method of EDSR [16], which is defined as POAN+. Compared with other advanced methods, it can be seen that the proposed POAN shows superiority and outperforms other state-of-the-art methods by a large margin on all scaling factors.

Visual Comparisons: As shown in Fig. 3, we compare SR results on the PD (top), T1 (middle), and T2 (bottom) images with scale×4 visually. It can seem that most of the compared methods cannot reconstruct the detailed textures accurately and even suffer structural distortions. For example, in the PD image, most of the methods listed cannot recover the sharp triangular shape and bright white spots well, and the result of CSN appears wrong arc structure distortion. Our POAN achieves clearer and more reliable results with more refined details. A similar comparison can also be seen from T2 images. Our method can still extract more sophisticated features from severely degraded LR images with limited information and produce the SR MR image closest to the ground truth.

3.4. Ablation Study

Self-Similarity Combinations: We compared the performance of networks focusing on different types of SS priors. The specific quantitative results are listed in Tab. 2. We take the structure that does not extract any similarity features as the basic baseline. According to the performance of the first three methods in Tab. 2, we can observe that pyramid ss in method D achieves the highest PSNR gain (1.9 dB) compared with method B and C. This observation proves that cross-scale pyramid attention captures more feature correlations. Then we combined different SS structures in pairs. The results show that method E adding pixel-wise features attention based on single scale patch SS can improve the highest PSNR from 37.67 dB to 38.94 dB. It indicates the effectiveness of the pixel-wise non-local attention. If we replace the simple patch SS with pyramid SS in method G, the SR performance will be improved slightly (40.21 dB). The last row of Tab. 2 indicates that the proposed PDSB achieves the highest PSNR and the closest result to the ground truth.

Pyramid Orthogonal Attention: We analyze the impact of different attention domain combinations on network performance. As shown in Tab. 3, Method A is non-attention. We insert channel domain attention NCAB or the spatial domain attention PSDB into the block, respectively, defined as methods B and C. The results in the first three rows prove each attention domain's effectiveness, as each of them brings improvement over the backbone. The network with POAB is defined as method D, which is equipped with attention from two orthogonal directions. We found that the network gets the best performance and improves the highest PSNR from 30.37 dB to 31.62 dB, achieving significant improvements.

Attention Visualization Comparison: We visualize the class activation map (CAM) before and after the first-order channel attention block (CAB) or the NCAB. The CAMs displays which areas are critical by highlighting in different colors, where red means the most important. Meanwhile, orange, green, and blue represent decreasing importance. As shown in

Fig. 4, the CAMs without any attention (top) and the second and third rows represent the feature visualization after CAB and NCAB remapping, respectively. The activation regions in non-attention CAMs are evenly distributed and pay too much attention to the background wrongly. This phenomenon is alleviated after adding the CAB (middle). Although the network with CAB began to focus on areas with richer textures, a large number of useful features were still be suppressed. The CAMs from NCAB (bottom) show excellent superiority. Larger high-frequency regions are activated, and more suppression is applied to the ineffective background regions. Compared with the first-order information using channel reduction, the experiment results prove that the non-reduction channel recalibration and high-order covariance statistics capture the continuous correlation between channels.

4. CONCLUSION

Based on the visual characteristics of MR images, we design a pyramid orthogonal attention network (POAN) to guide image SR in this paper. It is the first attempt to combine autocorrelation prior of the MR image itself in traditional methods and the deep CNNs. Using pyramid orthogonal attention containing both spatial and channel domains, we explore the complete self-similarity on full scales in LR MR images. The non-reduction attention mechanism with high-order statistics captures distinguishing features adaptively. Using points and pyramid-scale patches highly similar to the current region, the micro-structures and sharp textures in MR images can be restored more accurately. The proposed POAN improves the performance of the MR images SR task and alleviates the deep network's excessive dependence on external datasets efficiently. Self-similar visual characteristics also exist in other types of medical images (e.g., X-ray and CT), so our method can be further explored in other image restoration tasks.

Acknowledgments. This work is partially supported by the NSFC fund (61831014), the Guangdong Provincial Science and Technology Project (2017B010110005), the Shenzhen Science and Technology Project under Grant (ZDYBH201900000002,JCYJ20180226181021364)

5. REFERENCES

[1] Eyal Carmi, Siuyan Liu, Noga Alon, Amos Fiat, and Daniel Fiat, "Resolution enhancement in mri," *MRI*, 2006.

[2] 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*.

[3] Maria Zontak and Michal Irani, "Internal statistics of a single natural image," in *CVPR*, 2011.

[4] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *TIP*, 2007.

[5] Gilad Freedman and Raanan Fattal, "Image and video upscaling from local self-examples," *ACM TOG*, 2011.

[6] Or Lotan and Michal Irani, "Needle-match: Reliable patch matching under high uncertainty," in *CVPR*, 2016.

[7] Maria Zontak, Inbar Mosseri, and Michal Irani, "Separating signal from noise using patch recurrence across scales," in *CVPR*, 2013.

[8] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang, "Non-local recurrent network for image restoration," in *NeurIPS*, 2018.

[9] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu, "Residual non-local attention networks for image restoration," in *ICLR*, 2019.

[10] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.

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

[12] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *CVPR*, 2020.

[13] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu, "Squeeze-and-excitation networks," *TPAMI*, 2017.

[14] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017.

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004.

[18] José V Manjón, Pierrick Coupé, Antonio Buades, Vladimir Fonov, D Louis Collins, and Montserrat Robles, "Non-local mri upsampling," *Medical image analysis*, 2010.

[19] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *TPAMI*, 2015.

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016.

[21] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *CVPR*, 2018.

[22] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou, "Channel splitting network for single mr image super-resolution," *TIP*, 2019.