MR Image Super-Resolution with Squeeze and Excitation Reasoning Attention Network

Yulun Zhang[1], Kai Li[1], Kunpeng Li[1], Yun Fu[1,2]
[1]Department of ECE, Northeastern University, USA
[2]Khoury College of Computer Science, Northeastern University, USA
yulun100@gmail.com, {kunpengli, kaili, yunfu}@ece.neu.edu

Abstract

High-quality high-resolution (HR) magnetic resonance (MR) images afford more detailed information for reliable diagnosis and quantitative image analyses. Deep convolutional neural networks (CNNs) have shown promising ability for MR image super-resolution (SR) given low-resolution (LR) MR images. The LR MR images usually share some visual characteristics: repeating patterns, relatively simpler structures, and less informative background. Most previous CNN-based SR methods treat the spatial pixels (including the background) equally. They also fail to sense the entire space of the input, which is critical for high-quality MR image SR. To address those problems, we propose squeeze and excitation reasoning attention networks (SERAN) for accurate MR image SR. We propose to squeeze attention from global spatial information of the input and obtain global descriptors. Such global descriptors enhance the network's ability to focus on more informative regions and structures in MR images. We further build relationship among those global descriptors and propose primitive relationship reasoning attention. The global descriptors are further refined with learned attention. To fully make use of the aggregated information, we adaptively recalibrate feature responses with learned adaptive attention vectors. These attention vectors select a subset of global descriptors to complement each spatial location for accurate details and texture reconstruction. We propose squeeze and excitation attention with residual scaling, which not only stabilizes the training but also makes it flexible to other basic networks. Extensive experiments show the effectiveness of our proposed SERAN, which clearly surpasses state-of-the-art methods on benchmarks quantitatively and visually.

1. Introduction

High-resolution (HR) magnetic resonance (MR) images would provide more detailed structures and textures, which benefit accurate diagnosis and quantitative image analyses [5]. However, in real-world cases, HR MR images are obtained at the cost of longer scanning time, lower signal-to-noise ratio, and smaller spatial converge [32]. Image super-resolution (SR) reconstructs HR outputs from the given low-resolution (LR) ones. It is becoming a promising technique to upscale the spatial resolution of MR images.

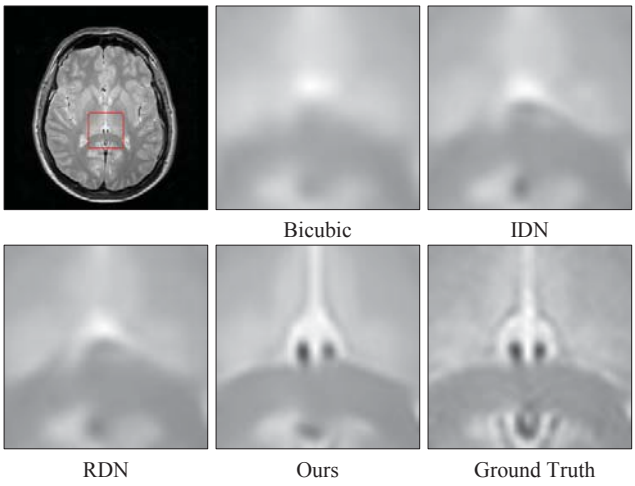

Figure 1. The visual comparison between the our method and recent image SR methods on a PD image with scaling factor $\times 4$.

Recently, deep convolutional neural network (CNN) has shown its powerful ability for high-quality image SR. Dong *et al*. firstly introduced CNN for image SR in SRCNN [7,8], which has only three convolutional layers. Kim *et al*. increased the network depth by utilizing residual learning in VDSR [18]. Hui *et al*. proposed information distillation network (IDN) [15]. Zhang *et al*. achieved better SR performance with RDN [48], which fully utilizes hierarchical features. On the other hand, some MR images oriented SR methods were also proposed. Chen *et al*. used 3D dense network and adversarial learning for MR image SR [5].

However, those deep CNN based methods either neglect the characteristics of MR images or suffer from intrinsic

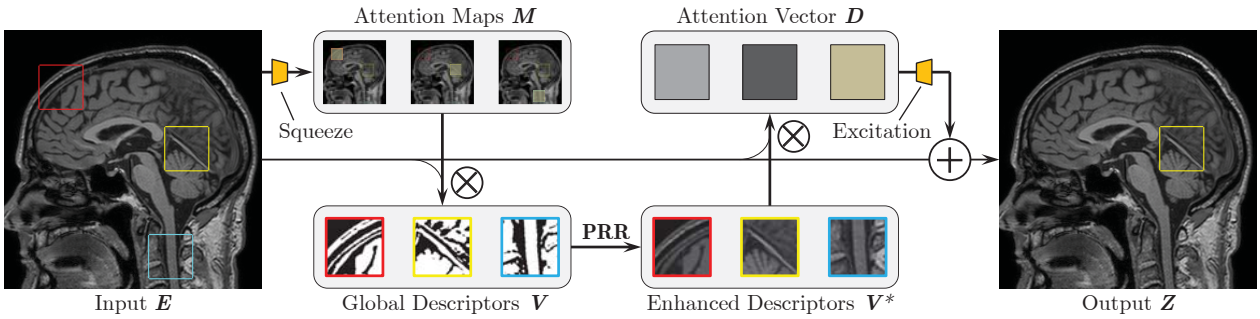

Figure 2. A brief illustration of our squeeze and excitation attention mechanism. Similar to [3], the global features are first collected via bilinear pooling and then distributed to each spatial position by considering the corresponding local feature. However, we enhance the global features through primitive relationship reasoning (PRR), as described in the text.

drawbacks. First, MR images often have repeating structural patterns and are relatively simpler than natural images. Very wide and deep networks (*e.g.*, EDSR [26]) may suffer from over-fitting problem. Second, MR images often contain large region of background, which is far less informative than the target structural regions. However, most previous CNN based methods treat all the spatial pixels of the image equally. They cannot distinguish target and background regions, hindering the representation ability. Third, most previous CNN based SR methods depend on convolutional operations, which focus on local neighborhoods and fail to capture the entire aspects of the input. However, it's important for MR images to sense the global features, which contribute to reconstruct more informative regions.

To tackle these issues and limitations, we propose squeeze and excitation attention networks (SERAN) for accurate MR image SR (see Fig. 1). As illustrated in Fig. 2, we propose to squeeze global spatial information into global descriptors with second-order attention pooling operation. It allows the model to focus on more informative regions and structures in MR images. We then build the relationship among primitives and apply graph convolutional network (GCN) to perform reasoning and obtain primitive relationship reasoning attention. We refine the global descriptors with the learned attention. To fully utilize the aggregated information, we adaptively recalibrate feature responses with learned adaptive attention vectors. These attention vectors select a subset of global descriptors to complement each spatial location for accurate details and texture reconstruction. Experimental results demonstrate the effectiveness of our SERAN, when compared with recent methods.

In summary, the main contributions of this work are:

- We propose the squeeze and excitation reasoning attention network (SERAN) for fast and accurate MR image super-resolution (SR). To our best knowledge, this is the first work investigating semantic reasoning attention to MR image SR.
- We propose to collect global visual primitives from the features. We further propose primitive relationship reasoning attention for refinement. We then adaptively allocate the global visual primitives to local feature.
- We demonstrate the effectiveness of our SERAN with extensive experiments on benchmark datasets. Our SERAN achieves significant performance gain over other state-of-the-art image SR methods.

2. Related Work

2.1. MR Image Super-resolution

As a post-processing method, image SR has been studied in lots of works related to MR image analysis, such as diffusion MRI [33, 36], structural MRI [27, 34], as well as spectroscopy MRI [16, 17] etc. In the early days, the task of image SR to MR images mainly focuses on traditional multiple frame image SR. However, when we try to reconstruct a HR image from multiple degraded LR counterparts, we often have to calibrate and fuse these LR images. Such a process is a very challenging task in itself [49]. Recently, deep learning [23] based image SR methods demonstrate superior performance on natural images [8, 18, 26, 47, 48], which promotes the application of deep learning technologies in MR image SR tasks [6, 31, 43, 49]. However, some CNN based image SR methods obtain superior results as desired, they have large model sizes and are impractical to real MRI scenarios with limited resources. In this work, our squeeze and excitation reasoning attention network is essentially a lightweight model that is more convenient for practical deployment and clinic applications.

2.2. Attention Mechanism

Attention mechanism endows neural networks with the ability to allocate resources adaptively for the informative input features, which is conducive to the full exploitation of network representational ability and the improvement of model performance [13]. Therefore, it is broadly embedded into neural networks for various machine learning

tasks in recent years, including natural language processing (NLP) [28, 39], image recognition [1, 9, 40] and image captioning [44] etc. In low-level computer vision applications like image SR, there are also some works on introducing attention mechanism to neural networks [14, 48]. However, few works have been conducted to study the role of attention on single MR image SR tasks, by considering the special characteristics of MR images: repeating patterns, relatively simpler structures, and less informative background. If the network can allocate computational resources adaptively to informative parts in MR images, it is promising to improve performance with moderate model parameters.

2.3. Semantic Reasoning

Researchers in artificial intelligence community initially investigated relational reasoning as symbolic methods [29]. By utilizing the language of mathematics and logic, they first defined the relations between abstract symbols. Then, they conducted reasoning by using abduction and deduction [12]. However, to make such systems be practical for usage, those symbols should be grounded firstly [10]. Later, extracting useful patterns with statistical learning, modern methods (e.g., path ranking algorithm [22]) turned to conduct relational reasoning on structured knowledge bases. Recently, graph-based approaches have been showing promising performance for relation reasoning. Along with the great success of convolution neural networks in computer vision areas [11], graph convolution networks (GCN) [20] was proposed for semi-supervised classification, where convolution network was used to process data in graph-structure. GCN was utilized to capture relations between objects in video recognition applications [41]. To improve semantic navigation in unseen scenes and towards novel objects, [45] took use of GCNs to encode the prior knowledge into a deep reinforcement learning framework. [4, 24] incorporate GCN into design of visual encoding and learn relationship enhanced features end-to-end towards the task of interest, such as image classification and image-text matching. Under an image captioning framework, [46] used Visual Genome dataset [21] to train a visual relationship detection model, where the detected relationship information is encoded with a GCN-based image encoder. In this work, we also introduce the reasoning advantage of graph convolutions to enhance the visual representation by considering semantic relationship among visual primitives. We incorporate the reasoning power into attention learning phase to enhance the ability of image SR models.

3. SERAN

3.1. Motivation

As we analyzed above, different from natural images, the MR images have their specific characteristics: repeating visual patterns, relatively simple structures, and less informative background. A more distinguishable mechanism is desired to handle these cases. On the other hand, attention mechanism has been demonstrated to be effective for high-level visual tasks by focusing on more informative channels [13] or spatial positions [3]. Here, we further investigate how to focus on more informative visual regions and patterns in MR images with attention.

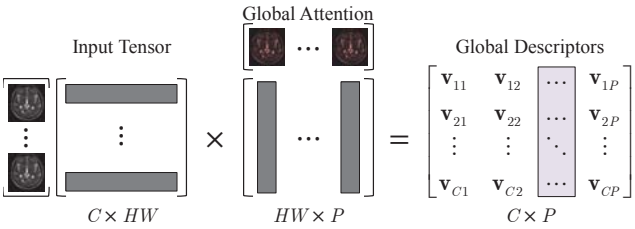

Figure 3. Global descriptors collection with global attention.

3.2. Squeeze Attention: Information Collection

We illustrate the global information collection in Fig. 3. Given an input feature $E \in \mathbb{R}^{C \times HW}$, we aim to obtain its attention guided counterpart $Z \in \mathbb{R}^{C \times HW}$. Here, C is channel number. H and W denote the height and width of the feature, respectively. For simplicity, we have reshaped the feature to 2-dimension space.

Let's rewrite the input feature as $E = [\mathbf{e}_1, \cdots, \mathbf{e}_{HW}]$. We know that the main visual primitives come from the more informative visual regions and patterns in MR features. To achieve those main visual primitives, we tend to take all input feature points into account. Mathematically, such a process can be written as

$$\mathbf{v}_i = \sum_{j=1}^{HW} \mathbf{m}_{ij} \mathbf{e}_j, \tag{1}$$

where $\mathbf{v}_i \in \mathbb{R}^{C \times 1}$ is a visual primitive and $\mathbf{m}_i \in \mathbb{R}^{1 \times HW}$ is a global attention vector. We can see each primitive $\mathbf{v}_i$ is calculated by considering all the local features weighted by a global descriptor $\mathbf{m}_i$. Supposing we target to seek P visual primitives, we have to use P global descriptors, which can be denoted as $M = [\mathbf{m}_1; \cdots; \mathbf{m}_P] \in \mathbb{R}^{P \times HW}$. Then, the global information collection can be expressed by

$$V = EM^T. \tag{2}$$

We can see that Eq. (2) not only learns a set of visual primitives $V = [\mathbf{v}_1, \cdots, \mathbf{v}_P]$, but also obtains second-order statistics. Such a second-order attention pooling operation can capture more complex long-range feature interdependencies [3]. In implementation, we have to force $\sum_{j=1}^{HW} \mathbf{m}_{ij} = 1$, which can be achieved with softmax function. Namely, $M = softmax\left(\theta\left(E; W_\theta\right)\right)$ with proper matrix reshaping, where W_θ denotes the trainable parameters for this convolutional layer.

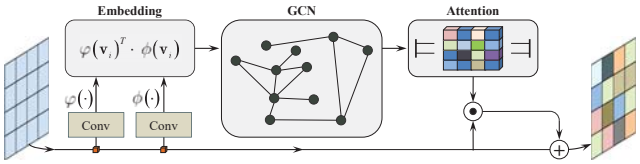

Figure 4. Illustration of our primitive relationship reasoning. GCN denotes graph convolutional network.

3.3. Primitive Relationship Reasoning Attention

After obtaining the visual primitive set V, we hope to further enhance it by considering the relationship among each visual primitive $\mathbf{v}_i$. In recent developments of deep learning, visual reasoning [2, 25, 35, 50] have been investigated to model and mine the relationship among visual components. Here, we're inspired to construct the relationship reasoning model among the visual primitives. Specifically, we firstly embed the visual primitives into two embedding spaces with weight parameters W_φ and W_ϕ. We then build the relationship by calculating the pairwise affinity via

$$R\left(\mathbf{v}_i, \mathbf{v}_j\right) = \varphi\left(\mathbf{v}_i; W_\varphi\right)^T \phi\left(\mathbf{v}_j; W_\phi\right), \tag{3}$$

where $\varphi\left(\mathbf{v}_i; W_\varphi\right)$ and $\phi\left(\mathbf{v}_j; W_\phi\right)$ are two embeddings. We use Eq. (3) to obtains the relationship between every two learned visual primitives $\mathbf{v}_i$ and $\mathbf{v}_j$, resulting in a fully-connected relationship graph.

Let's denote the graph as $G(V, R)$, where V is the set of graph nodes (i.e., visual primitives) and R is the set of graph edges (i.e., primitive relationships). Based on Eq. (3), we obtain the affinity matrix R by measuring the affinity edge of each visual primitive pair. A graph edge with large affinity score means that the corresponding visual primitive pair are highly correlated with strong semantic relationship.

Then, based on the above fully-connected graph, we conduct reasoning by utilizing graph convolutional network (GCN) [20]. For each node, its neighbors are defined by the graph relationships and can be used to compute the response of each node. Unlike some previous works [4], which incorporate the reasoning results as an enhancement of the input. Here, we achieve a reasoning attention by applying sigmoid activation function. A residual learning is then introduced to connect the original input as follows

$$V = \sigma\left(\left(\left(RV^T W_g\right) W_r\right)^T\right) \odot V + V, \tag{4}$$

where σ is sigmoid activation function. R is the $P \times P$ affinity matrix. W_g is a weight matrix of the GCN layer with size $C \times C$. For the residual structure, its weight matrix is W_r. $\odot$ denotes element-wise multiplication. We illustrate such a reasoning process in Fig. 4. With the usage of primitive relationship reasoning attention, we obtain the enhanced visual primitives, being viewed as global feature descriptors.

3.4. Excitation Attention: Feature Allocation

After collecting the global feature descriptors, we would like to distribute them to each location of the raw feature.

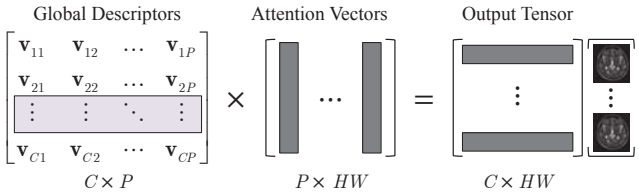

Figure 5. Adaptive feature allocation.

This would help us make better use of complex relations with the computed second-order statistics and compensate the lost information for better MR image reconstruction.

In Fig. 5, we can see that each position of raw feature has its specific need of the global descriptors. We adaptively distribute the global descriptors V based on the learned attention vector $\mathbf{d}_i$ at each location. It means that each location can adaptively select complementary visual primitives. Such a procedure can be achieved by

$$\mathbf{z}_i = \sum_{j=1}^{P} \mathbf{d}_{ij} \mathbf{v}_j, \tag{5}$$

where $\mathbf{z}_i \in \mathbb{R}^{C \times 1}$ is the i-th column of Z and $\mathbf{d}_i \in \mathbb{R}^{P \times 1}$ is a soft attention vector. For each position in Z, we have a specific soft attention vector to adaptively select complementary features from V. It will result in a soft attention matrix $D = [\mathbf{d}_1, \cdots, \mathbf{d}_{HW}]$, where $\sum_{j=1}^{P} \mathbf{d}_{ij} = 1$. We apply softmax function to achieve it via $D = softmax\left(\rho\left(E; \rho\right)\right)$ with parameter W_ρ. The adaptive distribution can finally be represented as

$$Z = EM^T D, \tag{6}$$

which acts as complementary component. We can see that the attention guidance affects the target feature Z. Specifically, Z is obtained by using adaptive complementary feature from a pool of main visual primitives.

3.5. SERAB

After obtaining the attention guided complementary feature Z, we want to encode it back to the input. A widely used practice is residual learning [11]. Here, we further adopt the residual scaling [38] with factor α to obtain the final feature output. This procedure can be written as

$$O = \alpha Z + E = \alpha EM^T D + E, \tag{7}$$

which results in a squeeze and excitation reasoning attention block (SERAB) shown in Fig. 6.

It is formulated based on two reasons. First, direct usage of residual learning (*e.g.*, $\alpha = 1$) here would make the training procedure numerically unstable. Second, the residual connection allows us to insert our SEAB to any pre-trained network, without affecting its initial behavior too much (*e.g.*, $\alpha \to 0$). With the usage of SEAB, subsequent convolutional layers could sense the entire space, even with limited receptive field size. SEAB allows the network to focus on more informative visual features and achieve better MR image SR reconstruction quality.

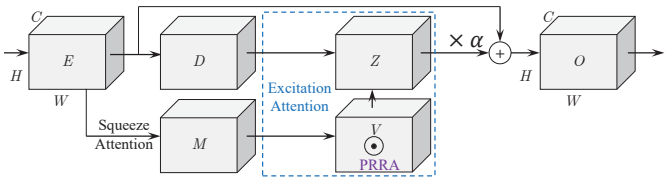

Figure 6. Squeeze and excitation attention block.

4. Experimental Results

We first briefly introduce the datasets and the details of model implementation. Then we investigate and analyze the structure of our method SERAN. Next, several recent SR methods are compared with the proposed model. For quantitative evaluation metrics, we adopt PSNR and structural similarity index metric (SSIM) [42] for SR quality. We also provide model size and running time comparisons.

4.1. Datasets

The datasets used in this paper are the same as [49] and originally derived from the IXI dataset[1]. Three types of MR images are included in the datasets (i.e., PD, T1, and T2). Each of them has 500, 70, and 6 MR volumes for model training, testing, and quick validation respectively. The size of each 3D volume is cut to 240×240×96 (height×width×depth), where 96 indicates the number of slices in the MR volume (along the imaging plane direction). Note that the datasets contain two kinds of image degradation but only the typical bicubic degradation is studied in this paper due to limited space. Due to the 2D nature of the proposed method, we get 500×96 = 48,000 2D training samples.

4.2. Implementation Details

When implementing our proposed SERAN, we use the residual network [26] (40 residual blocks) as the backbone. Specifically, we insert SEAB before the element-wise adding of the long skip connection. We set α as 0.01, allowing the network to stably and gradually learn attention guidance. We set the size of all convolutional layers as 3×3 except for that in the SEAB, where the kernel size is 1×1. Each convolutional layer has 64 filters except for the input and output layers, which have 1 channel. For convolutional layers with kernel size 3×3, we use zero-padding strategy to keep the size fixed. In the training phase, the batch size is 96. The input size of LR patches is 32×32. Our network is trained with ADAM optimizer [19] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is initialized as 10^{-4} and decreases to half every 200 epochs. We use PyTorch [30] to implement models with a Titan Xp GPU.

4.3. Ablation Study

Here, we investigate the effects of each key components of our proposed squeeze and excitation reasoning attention network (SERAN). In the SERAN, squeeze and excitation reasoning attention block (SERAB) plays an important role. So we first investigate the effectiveness of SERAB. Then, we further explore how the insert position and number of SERAB would affect the performance. We report the performance on PD, T1, and T2 with scaling factor 2×.

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

Table 1. Ablation investigation of SEAB, residual scaling (RS), and primitive relationship reasoning. We observe the PSNR (dB) values on the PD, T1, and T2 dataset with scaling factor SR×2.

Method	PD	T1	T2
Baseline	40.68	37.82	38.88
Baseline + SEAB (w/o RS)	41.31	38.41	39.80
Baseline + SEAB (with RS)	41.43	38.56	40.07
Baseline + SERAB	41.53	38.66	40.18

4.3.1 Effect of SERAB

In the SERAB, we want to demonstrate the effectiveness of our squeeze and excitation attention module. We denote this case as SEAB, namely without using primitive relationship reasoning (PRR). Then we explore how the residual scaling (RS) affect the network. Finally, we show how primitive relationship reasoning further improves network ability.

Effect of SEAB. By removing PPR, our proposed SERAB could be simplified as SEAB. We investigate the effects of SEAB with and without residual scaling (RS) by using the residual network [26] (40 residual blocks) as the baseline. As shown in Tab. 1, the introduction of SEAB (w/o RS) into the baseline would improve the network performance very obviously. Such an observation indicates that there are some redundant information in the deep features from the MR images. After collecting the global information (*i.e.*, squeeze attention), we extract more compact global descriptors for the MR deep features. Then, the global descriptors are adaptively allocated to each local feature, resulting in much stronger network representation ability. Furthermore, such a strong ability is not heavily influenced by using RS or not.

Effect of residual scaling (RS). After introducing SEAB into the baseline, we also find that RS can further helps to pursue better performance. We can see the strategy of residual scaling further boosts the performance without extra model parameters. More important, as we have described in Section 3.5, our SERAB can easily be integrated into other baselines with the residual scaling.

Effect of PRR. When we further introduce our proposed primitive relationship reasoning (PRR) into the SEAB, we achieve its enhanced version, named SERAB. As shown in Tab. 1, PRR would further achieve another performance improvement (about 0.1 dB) among each dataset. Such an improvement indicates that there still exists some redundant information in the visual primitives. Based on the fully-connected graph, we apply GCN to conduct reasoning, which would further investigate the latent relationship among visual primitives. As a result, we achieve stronger visual primitives and better network performance.

Table 2. Ablation investigation of SERAB position and number. We observe the PSNR (dB) values on the PD, T1, and T2 dataset with scaling factor SR×2.

Case	Low	Mid.	High	PD	T1	T2
1	√			41.38	38.32	39.88
2		√		41.41	38.43	39.96
3			√	41.53	38.66	40.18
4	√	√		41.55	38.67	40.20
5	√		√	41.56	38.69	40.22
6		√	√	41.59	38.72	40.25
7	√	√	√	41.62	38.76	40.29

4.3.2 Effect of SERAB position and number

After demonstrating the effectiveness of SERAB as above, we further investigate how the position and number of SERAB affect the network performance. We still use PD, T1, T2 as testsets with scaling factor $2\times$. We first name the 1-st, 20-th, and 40-th residual block (RB) in the baseline as low-, middle (mid.), and high-level RBs respectively. We insert SERAB into those three positions and record the performance in Tab. 2 for different combination cases.

Effect of SERAB position. For the cases 1, 2, and 3 in Tab. 2, we can learn that SERAB in higher-level would perform better. The main reasons might be that features in higher-level have larger receptive field size, resulting in larger sensing scope to the whole features. Also, deeper feature could be more compact and have less redundant information, which further help SERAB learn more effective visual primitives and achieve stronger representation ability.

Effect of SERAB number. We further show how the number of SERAB affects network performance. Let's visit Tab. 2, where more SERABs would obtain better performance. More higher-level SERABs performs better than those in lower-level. Compared with single high-level SERAB, the performance gains with more SERABs are not very large. But, using more SERABs would also consume more GPU memory and running time. As a result, we experimentally use one SERAB in the high-level RB, namely the last one, to report our results.

4.4. Comparison with Other Methods

We compare the proposed SERAN model with several state-of-the-art SR techniques, ranging from lightweight SRCNN [8], VDSR [18], IDN [15] to large-scale RDN [48] and CSN [49], through both quantitative and qualitative evaluations. Some quantitative results are directly cite from [49], where the compared methods use same evaluation metrics, training, and testing datasets.

4.4.1 Quantitative Comparison

Table 3 exhibits the quantitative results of the compared methods. SEARN represents the results directly obtained by our model and SERAN$^+$ indicates that geometric self-ensemble [26] technique is applied. We can se that our proposed SERAN obtains obviously higher performance than other state-of-the-art methods by a wide margin, giving the best SR performance on all types of MR images and all SR scaling factors, even without geometric self-ensemble. More importantly, our method utilizes moderate-scale model parameters and presents more accurate SR results, implying that our SERAN achieves a better trade-off between performance and model size by fully considering the particularity of MR images. Therefore, our SERAN model supports fast model inference and convenient practical deployment, indicating that it is a high-precision SR method for MR images. Moreover, our method could be a highly practical one in real-world cases.

4.4.2 Visual Comparison

Figure 7 exhibits the visual effects of the compared methods in Tab. 3, on three datasets: PD (top), T1 (middle), and T2 (bottom) images with scaling factor $\times 4$ respectively. As can be seen, the proposed SERAN model displays significantly visible superiority over other methods on all image types. For example, in the PD image, there is a black area indicated by the red arrow. This structure is almost completely lost in the results of Bicubic, SRCNN [8], and even VDSR [18]. Although it can be observed in the results of IDN [15] and RDN [48], our model presents a clearer indication and better approximation to the ground truth. Similar comparisons can also be observed from the results on the T1 and T2 images, which illustrates the superiority of our proposed model in MR image super-resolution tasks.

4.4.3 Performance on In-vivo Images

We also conduct SR experiments on in-vivo MR images in different human body positions to verify the practicability of the proposed model. Figure 8 shows the visual comparison between several SR methods, including NLM [27], SRCNN [8], VDSR [18], and RDN [48], on a real-world T2 image with SR×4. It can be seen that our model presents a clearer result with some details that are not found in the results of other methods, *e.g.* the the dark seam indicated by the red arrow. This comparison demonstrates the effectiveness of the proposed SERAN model in processing MR images with specific characteristics.

4.4.4 Model Size Analyses

In Tab. 3, we give the model sizes for each CNN based SR methods. Let's analyze the comparison about model size and performance. We can see that some large models (*e.g.*, RDN) might not be suitable for MR images, even though they perform pretty well for natural image SR. Although our SERAN is not the smallest network, it has much less parameters than that of RDN [48]. More importantly, our SERAN and SERAN$^+$ obtain the highest performance for

Table 3. PSNR (dB)/SSIM and model size comparisons between different image SR methods. We mark the highest and second higheast PSNR (dB)/SSIM values of each comparison cell in red and blue.

Method \ Image	scale	param	PD	T1	T2
Bicubic [2D]	×2	-	35.04 / 0.9664	33.80 / 0.9525	33.44 / 0.9589
NLM [27]	×2	-	37.26 / 0.9773	35.80 / 0.9685	35.58 / 0.9722
SRCNN [8]	×2	24.51K	38.96 / 0.9836	37.12 / 0.9761	37.32 / 0.9796
VDSR [18]	×2	0.67M	39.97 / 0.9861	37.67 / 0.9783	38.65 / 0.9836
IDN [15]	×2	0.73M	40.27 / 0.9869	37.79 / 0.9787	39.09 / 0.9846
FSCWRN [37]	×2	3.50M	40.72 / 0.9880	37.98 / 0.9797	39.44 / 0.9855
RDN [48]	×2	22.06M	40.31 / 0.9870	37.95 / 0.9795	38.75 / 0.9838
CSN [49]	×2	13.64M	41.28 / 0.9895	38.27 / 0.9810	39.71 / 0.9863
SERAN [Ours]	×2	3.16M	41.53 / 0.9900	38.66 / 0.9822	40.18 / 0.9872
SERAN$^+$ [Ours]	×2	3.16M	41.66 / 0.9902	38.74 / 0.9824	40.30 / 0.9874
Bicubic [2D]	×3	-	31.20 / 0.9230	30.15 / 0.8900	29.80 / 0.9093
NLM [27]	×3	-	32.81 / 0.9436	31.74 / 0.9216	31.28 / 0.9330
SRCNN [8]	×3	24.51K	33.60 / 0.9516	32.17 / 0.9276	32.20 / 0.9440
VDSR [18]	×3	0.67M	34.66 / 0.9599	32.91 / 0.9378	33.47 / 0.9559
IDN [15]	×3	0.83M	34.96 / 0.9619	33.06 / 0.9394	33.92 / 0.9591
FSCWRN [37]	×3	3.50M	35.37 / 0.9653	33.24 / 0.9423	34.27 / 0.9618
RDN [48]	×3	22.24M	35.08 / 0.9628	33.31 / 0.9430	33.91 / 0.9591
CSN [49]	×3	16.60M	35.87 / 0.9693	33.53 / 0.9464	34.64 / 0.9647
SERAN [Ours]	×3	3.34M	36.17 / 0.9713	34.08 / 0.9514	35.02 / 0.9672
SERAN$^+$ [Ours]	×3	3.34M	36.34 / 0.9721	34.18 / 0.9522	35.19 / 0.9680
Bicubic [2D]	×4	-	29.13 / 0.8799	28.28 / 0.8312	27.86 / 0.8611
NLM [27]	×4	-	30.27 / 0.9044	29.31 / 0.8655	28.85 / 0.8875
SRCNN [8]	×4	24.51K	31.10 / 0.9181	29.90 / 0.8796	29.69 / 0.9052
VDSR [18]	×4	0.67M	32.09 / 0.9311	30.57 / 0.8932	30.79 / 0.9240
IDN [15]	×4	0.95M	32.47 / 0.9354	30.74 / 0.8966	31.37 / 0.9312
FSCWRN [37]	×4	3.50M	32.91 / 0.9415	30.96 / 0.9022	31.71 / 0.9359
RDN [48]	×4	22.21M	32.73 / 0.9387	31.05 / 0.9042	31.45 / 0.9324
CSN [49]	×4	16.01M	33.40 / 0.9486	31.23 / 0.9093	32.05 / 0.9413
SERAN [Ours]	×4	3.31M	33.77 / 0.9526	31.89 / 0.9201	32.40 / 0.9455
SERAN$^+$ [Ours]	×4	3.31M	33.97 / 0.9542	32.03 / 0.9219	32.62 / 0.9472

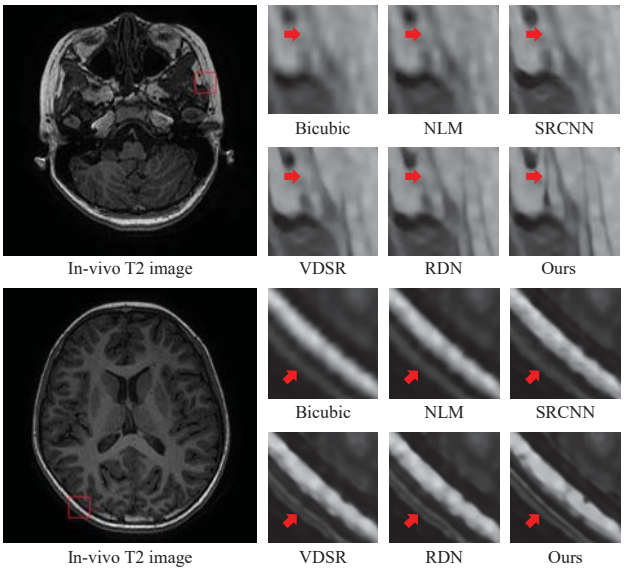

Figure 8. Real-world visual comparison (×4) between our proposed SERAN model with other typical SR methods. We show results on a in-vivo T2 data in different body positions.

each dataset and scaling factor, showing better trade-off between the SR performance and model size. These comparisons also indicate that global descriptors with squeezed attention help to offset the limited receptive field size of relative small network. Consequently, our method achieves better results with much smaller model parameters.

Table 4. Running time (seconds/3D volume) comparison of several SR models on PD images. The sizes of input volumes for scaling factors ×2, ×3, and ×4 are 120×120×96, 80×80×96, and 60×60×96, respectively.

Method	×2	×3	×4
Bicubic [2D]	0.1543	0.1578	0.1610
SRCNN [8]	0.3021	0.3211	0.3284
VDSR [18]	1.7644	2.6488	2.4131
IDN [15]	0.8123	0.4415	0.2773
RDN [48]	22.6771	11.1601	5.1250
SERAN [Ours]	1.1713	0.7475	0.7081
SERAN$^+$ [Ours]	9.9257	5.4320	4.8149

4.4.5 Running Time Comparisons

Tab. 4 exhibits the comparison of execution efficiency between several typical methods, in terms of different scaling factors. We can see that our SERAN obtains comparable running time as other leading lightweight CNN-based models. Even we use self-ensemble [26] to further improve SERAN and obtain SERAN$^+$, the running time is less than that of RDN [48]. Furthermore, when we consider the performance shown in Tab. 3 and running time together, we can see that our proposed SERAN obtains a good trade-off between performance and running time.

5. Conclusion

Attention mechanism can boost the performance of deep CNNs in low-level computer vision tasks and MR images

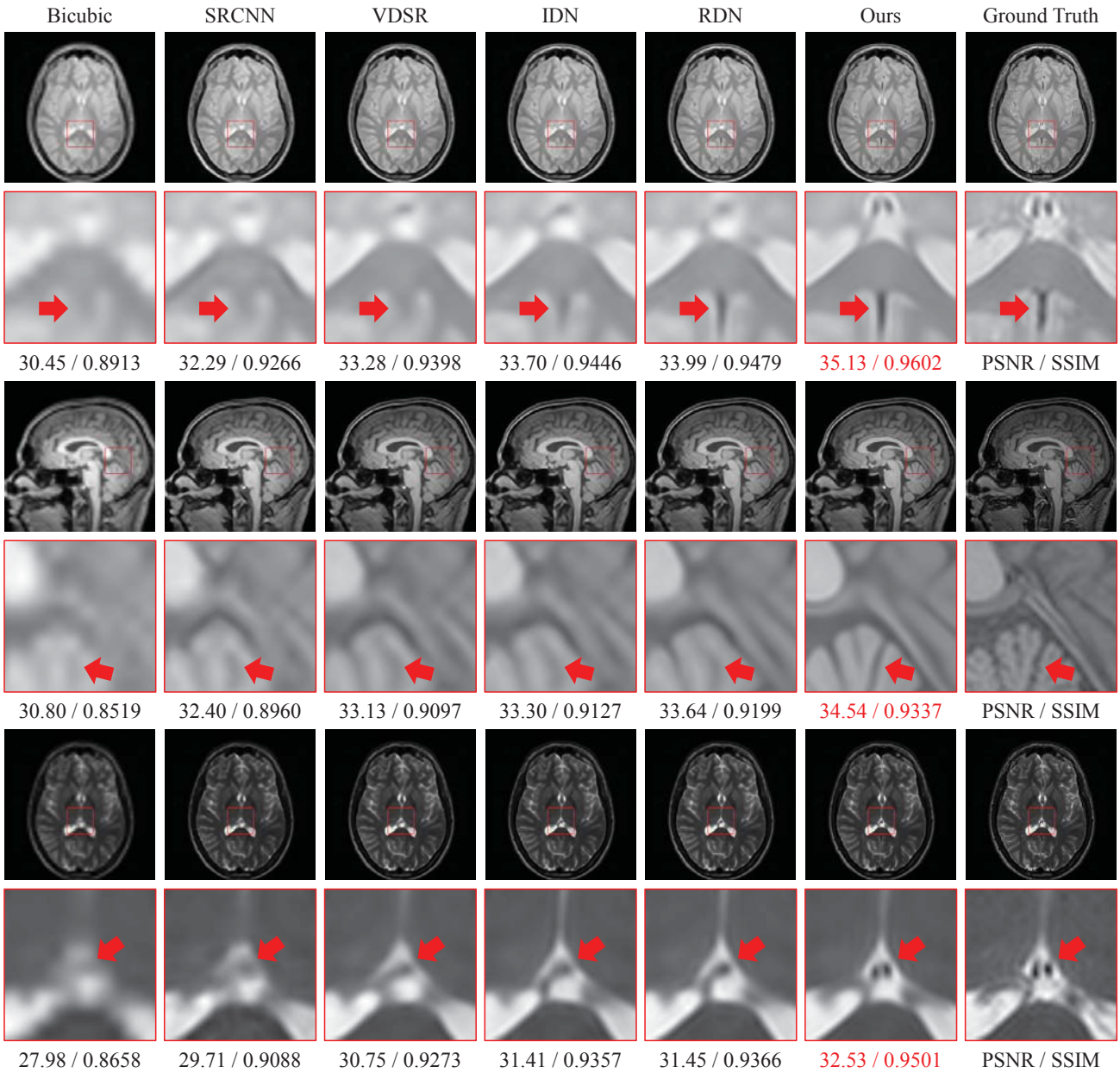

Figure 7. The visual comparison (×4) between several advanced CNN-based SISR methods on three datasets: PD (top), T1 (middle), and T2 (bottom). The highest PSNR (dB) and SSIM values for each row are marked in red.

share their specific visual characteristics, based on which we propose a SERAN model in this paper and apply it to MR image SR task. Unlike separate channel or spatial attention, we provide adaptive attention distribution for each spatial location and combine it with the learned global descriptors. Considering the repeating structure and simple distribution of MR images, our model can deal with MR image SR task more effectively and accurately. We build the basic squeeze and excitation attention block (SEAB) by using residual scaling, which helps stabilize the training. To better make use of the relationship between the learned global descriptors, namely visual primitives, we further build relationship graph among the visual primitives. Based on the relationship graph, we utilize GCN to conduct the reasoning process, resulting in primitive relationship reasoning attention. Such a learned attention can be used to further improve the representation ability of visual primitives. We demonstrate the effectiveness of each proposed modules in our SERAN. Extensive experiments quantitatively (e.g., PSNR/SSIM, model size, and running time) and qualitatively demonstrate the advantages of our proposed SERAN method over other leading CNN-based image SR methods.

Acknowledgments. The work was supported by the National Science Foundation Award ECCS-1916839.

References

[1] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015. 3

[2] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *CVPR*, 2018. 4

[3] Yunpeng Chen, Yannis Kalantidis, Jiashu Li, Shuicheng Yan, and Jiashi Feng. a^2-nets: Double attention networks. In *NeurIPS*, 2018. 2, 3

[4] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, 2019. 3, 4

[5] Yuhua Chen, Feng Shi, Anthony G Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In *MICCAI*, 2018. 1

[6] Yuhua Chen, Yibin Xie, Zhengwei Zhou, and *et al*. Brain MRI super resolution using 3D deep densely connected neural networks. In *15th IEEE International Symposium on Biomedical Imaging, ISBI 2018*, pages 739–742, 2018. 2

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2016. 1, 2, 6, 7

[9] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017. 3

[10] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 1990. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 4

[12] Jerry R Hobbs, Mark E Stickel, and Paul Martin. Interpretation as abduction. *Artificial intelligence*, 1993. 3

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 3

[14] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *CoRR*, abs/1809.11130, 2018. 3

[15] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, 2018. 1, 6, 7

[16] Zohaib Iqbal, Dan Nguyen, Gilbert Hangel, Stanislav Motyka, Wolfgang Bogner, and Steve Jiang. Super-resolution 1h magnetic resonance spectroscopic imaging utilizing deep learning. *arXiv preprint arXiv:1802.07909*, 2018. 2

[17] Saurabh Jain, Diana Maria Sima, F Sanaei Nezhad, S Williams, Sabine Van Huffel, Frederik Maes, and Dirk Smeets. Patch based super-resolution of mr spectroscopic images. In *ISBI*, 2016. 2

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 1, 2, 6, 7

[19] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[20] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3, 4

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 3

[22] Ni Lao, Tom Mitchell, and William W Cohen. Random walk inference and learning in a large scale knowledge base. In *EMNLP*, 2011. 3

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 2

[24] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. 3

[25] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. 4

[26] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2, 5, 6, 7

[27] José V. Manjón, Pierrick Coupé, Antonio Buades, Vladimir Fonov, and D. Louis Collins. Non-local MRI upsampling. *Medical Image Analysis*, 14(6):784–792, 2010. 2, 6, 7

[28] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NeurIPS*, 2014. 3

[29] Allen Newell. Physical symbol systems. *Cognitive science*, 1980. 3

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[31] Chi Hieu Pham, Aurélien Ducournau, Ronan Fablet, and François Rousseau. Brain MRI super-resolution using deep 3D convolutional networks. In *14th IEEE International Symposium on Biomedical Imaging, ISBI 2017*, pages 197–200, 2017. 2

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

[33] Dirk HJ Poot, Ben Jeurissen, Yannick Bastiaensen, Jelle Veraart, Wim Van Hecke, Paul M Parizel, and Jan Sijbers. Super-resolution for multislice diffusion tensor imaging. *Magnetic resonance in medicine*, 69(1):103–113, 2013. 2

[34] François Rousseau. Brain hallucination. In *ECCV 2008*, pages 497–508, 2008. 2

[35] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, 2017. 4

[36] Benoit Scherrer, Ali Gholipour, and Simon K Warfield. Super-resolution reconstruction to increase the spatial resolution of diffusion weighted images from orthogonal anisotropic acquisitions. *Medical image analysis*, 16(7):1465–1476, 2012. 2

[37] Jun Shi, Zheng Li, Shihui Ying, and *et al*. MR image super-resolution via wide residual networks with fixed skip connection. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1129–1140, 2019. 7

[38] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 4

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[40] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 3

[41] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 3

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[43] Zhao Xiaole, Huali Zhang, Hangfei Liu, Yun Qin, Tao Zhang, and Xueming Zou. Single mr image super-resolution via channel splitting and serial fusion network. *arXiv preprint arXiv:1901.06484*, 2019. 2

[44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 3

[45] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. *ICLR*, 2019. 3

[46] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 3

[47] 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. 2

[48] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 1, 2, 3, 6, 7

[49] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. Channel splitting network for single MR image super-resolution. *TIP*, 2019. 2, 5, 6, 7

[50] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 4