Cross-Modality High-Frequency Transformer for MR Image Super-Resolution

Chaowei Fang
Xidian University
Xi'an, China

Dingwen Zhang
Northwestern Polytechnical University
Xi'an, China

Liang Wang
Xidian University
Xi'an, China

Yulun Zhang
Computer Vision Lab, ETH
Zürich, Switzerland

Lechao Cheng
Zhejiang Lab
Hangzhou, China

Junwei Han*
Northwestern Polytechnical University
Xi'an, China
jhan@nwpu.edu.cn

ABSTRACT

Improving the resolution of magnetic resonance (MR) image data is critical to computer-aided diagnosis and brain function analysis. Higher resolution helps to capture more detailed content, but typically induces to lower signal-to-noise ratio and longer scanning time. To this end, MR image super-resolution has become a widely-interested topic in recent times. Existing works establish extensive deep models with the conventional architectures based on convolutional neural networks (CNN). In this work, to further advance this research field, we make an early effort to build a Transformer-based MR image super-resolution framework, with careful designs on exploring valuable domain prior knowledge. Specifically, we consider two-fold domain priors including the high-frequency structure prior and the inter-modality context prior, and establish a novel Transformer architecture, called Cross-modality high-frequency Transformer (Cohf-T), to introduce such priors into super-resolving the low-resolution (LR) MR images. Comprehensive experiments on two datasets indicate that Cohf-T achieves new state-of-the-art performance.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Machine learning*; • **Theory of computation** → Models of computation.

KEYWORDS

magnetic resonance image processing, neural networks, super-resolution, multi-modal learning

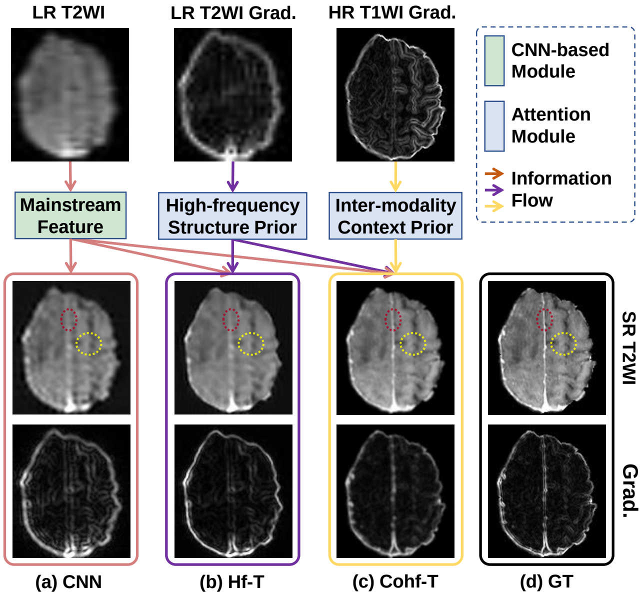


Figure 1: Aiming to super-resolve the input LR T2WI, we devise a Transformer-based framework capable of extracting both high-frequency structure prior and inter-modality context. The mainstream of our method (a) is built upon convolutional neural networks. The high-frequency structure prior and inter-modality context improves the super-resolution performance, as indicated by Hf-T (b) (short for High-frequency Transformer) and Cohf-T (c) (short for Cross-modality high-frequency Transformer), respectively. The ground-truth (GT) HR T2WI and its gradient map are shown in (d).

ACM Reference Format:
Chaowei Fang, Dingwen Zhang, Liang Wang, Yulun Zhang, Lechao Cheng, and Junwei Han*. 2018. Cross-Modality High-Frequency Transformer for MR Image Super-Resolution. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Due to the superior capacity in capturing histopathological detail of soft tissues, magnetic resonance (MR) image becomes one of the

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

most widely-used data in computer-aided diagnosis and brain function analysis. However, because of hardware and post-processing constraints, collecting MR images with higher resolution leads to lower signal-to-noise ratio or extends the scanning time [29]. For addressing this problem, one cost-effective way is to apply the super-resolution technology, which synthesizes the desired high-resolution (HR) MR image from the LR MR image.

The current main-stream technique for MR image super-resolution (SR) is based on convolutional neural networks (CNN) [28, 42]. Although encouraging results are obtained by these CNN-based approaches, the intrinsic short-range reception mechanism is disadvantageous to the exploration of the global structure and long-range context information. This makes existing CNN-based MR image SR algorithms still sub-optimal for producing satisfactory results.

This paper makes an early effort to build a Transformer-based MR image super-resolution framework. In contrast to conventional CNN models, the Transformer model constituted by self-attention blocks has advantages in modeling the long-distance dependency [35]. Such a network architecture is firstly evolved in natural language processing (NLP) and then achieves enormous success in vision tasks like image recognition [8] and semantic segmentation [32]. It is also verified that Transformer-based modelling is effective in medical image segmentation [3] and registration [17]. Here, we leverage it to advance the research on the super-resolution of T2-weighted MR image (T2WI), a typical MR data used in clinics but costing a very long scanning time.

Existing transformer designs directly model self-attention in the original image domain. However, when implementing super-resolution on MR images, we find that the structure information held in the high-frequency domain would play a paramount role for model design as the organs appearing in the MR images usually share similar anatomical structures across persons, and the relation between different parts of the organ is regular [5] (see Fig. 1). We term this as the high-frequency structure prior, and an example is provided to demonstrate the efficacy of the high-frequency structure prior in Fig. 1 (b). To this end, the transformer model designed in this work performs on the high-frequency image gradients rather than the conventionally used original image pixels.

Another important domain knowledge is that when processing LR T2WI data, the high-resolution T1-weighted images (T1WI) can be used to provide rich inter-modality context priors, as 1) the complementary morphological information [9] captured by the T1WI can help infer the structural content of the T2WI , and 2) the acquisition of HR T1WI costs much less scanning time [1]. An example is also provided to demonstrate the efficacy of such inter-modality context prior in Fig. 1 (c).

To explore the domain priors mentioned above, we propose a novel Transformer architecture called Cross-modality high-frequency Transformer (Cohf-T). A novel learning framework is set up for super-resolving LR T2WI under the guidance of gradient maps of both LR T2WI and HR T1WI. As shown in Fig. 2, our model has a main SR stream and a domain prior embedding stream, together with an input gate and an output gate. The domain prior embedding stream explores the high-frequency structure priors from the gradient map of LR T2WI and the inter-modality context from HR T1WI. Practically, both short-distance and long-distance dependencies are leveraged to explore high-frequency structure priors with the help of window attention modules. Considering there exists distribution shift between T2WI and T1WI, an adaptive instance normalization module is devised for aligning their features before performing the cross-modality attention. Additionally, a novel basic attention module that encloses both intra-head and inter-head correlations is proposed to improve the relation extraction capacity of our Transformer-based framework.

In summary, this work has three main contributions:

- We make an early effort to establish a Transformer-based framework for super-resolving T2-weighted MR images, based on the proposed Cross-modality high-frequency Transformer (Cohf-T).
- In Cohf-T, we introduce the high-frequency structure prior and inter-modality context prior by designing novel intra-modality window attention and inter-modality attention modules.
- Comprehensive experiments on two MR image super-resolution benchmarks demonstrate that the proposed Cohf-T achieves new state-of-the-art performance.

2 RELATED WORK

2.1 MR Image Super-Resolution

Improving the resolution of MR images is a long-lasting classical task in medical image analysis. Inspired by the rapid development in natural image super-resolution [7, 16, 26, 37, 41], CNN models have been the mainstream solutions for the MR image super-resolution [4, 28, 42]. According to the current studies [22, 23, 38], the low-frequency signals in MR data are relatively easy to reconstruct. In contrast, super-resolving the high-frequency signals, such as structures and textures, remains the main challenge.

Different from natural images, we can acquire MR images with multiple modalities via different imaging settings. A series of traditional algorithms [15, 31] attempt to explore prior context information from T1-weighted MR images for super-resolving T2-weighted or spectroscopy MR images. [9, 14, 39] further devise CNN models to exploit such inter-modality context information. However, directly fusing features of multiple modalities via convolutions with small kernels can not sufficiently leverage the inter-modality dependencies. To further advance this research field, we devise a novel Transformer-based super-resolution framework. The proposed framework can capture long-distance dependencies for involving the high-frequency structure prior and inter-modality context prior, which is beyond the exploration of the existing works.

2.2 Transformer for Vision Tasks

Transformer is primordially proposed for extracting long-distance relation context in NLP tasks [35]. Recently, this technique has been extensively applied to computer vision tasks, considering it can help to make up the artifact of convolution that only local features can be captured with a limited kernel size [2, 8, 18, 21, 27, 32].

Among the existing vision transformer models, self-attention [1, 20, 43] and its variations, e.g., [6, 24], are widely adopted to build the basic transformer block. For example, [6] exploits the second-order

[1] https://case.edu/med/neurology/NR/MRI%20Basics.htm

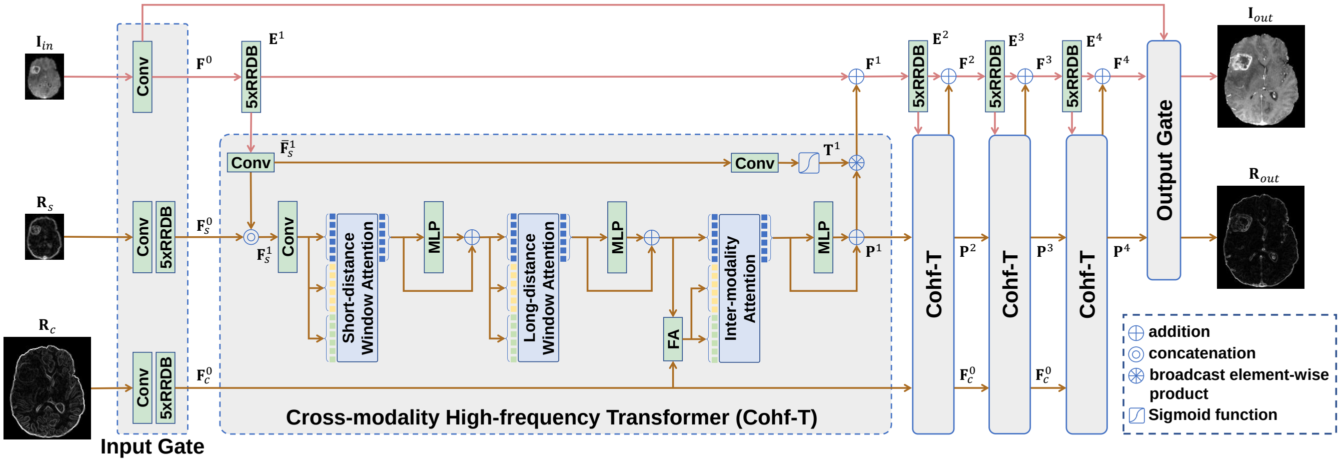


Figure 2: The pipeline of our proposed method. It consists of three main branches, a fully convolutional network for density domain super-resolution, a Transformer-based branch for restoring high-frequency signals in the gradient domain, and a guidance branch for extracting priors from the T1 modality. 'Conv' and 'RRDB' represent one 3×3 convolution layer and residual-in-residual dense block, respectively. 'MLP' stands for multi-layer perceptron.

attention based on covariance normalization for feature enhancement in image super-resolution. In [24], the computation burden of the non-local operation is reduced by removing less informative correlations and constructing a sparse attention map. Transformer modules are also applied for tackling the multi-modal vision understanding tasks. Targeted at the multispectral object detection, [30] concatenates tokens from RGB and thermal modalities and then fuse them with symmetric attention modules. [19] assigns specific modality embeddings to tokens from different modalities for solving the visible-infrared person re-identification task. Unlike these attention mechanisms, we introduce inter-head correlation in our attention modeling to further improve the feature interaction. Practically, three kinds of attention modules, namely short-distance window attention, long-distance window attention, and inter-modality attention, are incorporated in our transformer block.

3 METHOD

3.1 The Overall Learning Framework

This paper is targeted at super-resolving low-resolution (LR) T2-weighted image (T2WI) under the guidance of high-resolution (HR) T1-weighted image (T1WI). Specifically, given the main input image, namely a LR T2WI $\mathbf{I}_{in} \in \mathbb{R}^{h \times w}$, we embed the high-frequency structure prior by calculating the gradient field for $\mathbf{I}_{in}$ and define the gradient image as the structure reference input (denoted by $\mathbf{R}_s = \sqrt{(\nabla_x \mathbf{I}_{in})^2 + (\nabla_y \mathbf{I}_{in})^2 + \epsilon}$, where ϵ is a constant and is set to 10^{-6}). The gradient of the HR T1WI (denoted by $\mathbf{R}_c$) is regarded as an additional reference input for supplying inter-modality context information. h and w denote the height and width of $\mathbf{I}_{in}$, respectively. Then, a network architecture is built upon the cross-modality high-frequency transformer to predict the high-resolution T2-weighted image $\mathbf{I}_{out} \in \mathbb{R}^{rh \times rw}$. r denotes the upsampling ratio.

As shown in Fig. 2, our proposed framework is composed of two streams, including the main super-resolution stream (performing on the LR T2WI $\mathbf{I}_{in}$) and the domain prior embedding stream, together with an input gate and an output gate. The intensity levels of T1WI are not directly related to those of T2WI, e.g., the inflammation appears to be dark in T1WI but bright in T2WI. On the other hand, the two kinds of images share similar structures. Thus, we extract the domain prior information with the LR T2WI gradient image $\mathbf{R}_s$ and the HR T1WI gradient $\mathbf{R}_c$.

Input Gate. The input gate projects the input image data, including $\mathbf{I}_{in}$, $\mathbf{R}_s$, and $\mathbf{R}_c$ to their corresponding primary features $\mathbf{F}^0$, $\mathbf{F}_s^0$, and $\mathbf{F}_c^0$ through several convolutional layers and residual-in-residual dense blocks (RRDBs) [37].

Main Super-Resolution Stream. There are four stages in the main super-resolution stream. In each stage, five RRDBs are utilized for deriving more complicated latent features. For the i-th stage, the extracted feature map is denoted by, $\mathbf{E}^i = \mathcal{R}(\mathbf{F}^{i-1})$. Then, $\mathbf{E}^i$ is fed into a cross-modality high-frequency transformer (Cofh-T) block to interact with the high-frequency structure prior and inter-modality context prior, resulting in a domain prior embedding $\mathbf{P}^i = \text{Cohf-T}(\mathbf{E}^i, \mathbf{F}_s^i, \mathbf{F}_c^0)$. A more detailed description of the Cohf-T block can be referred to in the next subsection. Extra structural information is extracted from $\mathbf{E}^i$ via a convolution layer, resulting in $\bar{\mathbf{F}}_s^i = \mathcal{C}(\mathbf{E}^i)$. Then, a series of attention modules are employed to explore domain priors (namely $\mathbf{P}^i$) from $\mathbf{F}_s^i$, $\bar{\mathbf{F}}_s^i$, and $\mathbf{F}_c^0$. Finally, the prior-induced latent features of the current stage are produced via a shortcut connection: $\mathbf{F}^i = \mathbf{E}^i + \mathbf{T}^i \circ \mathbf{P}^i$, where $\mathbf{T}^i$ is a single-channel selection map inferred from $\bar{\mathbf{F}}_s^i$, and '$\circ$' is the broadcast element-wise production.

Domain Prior Embedding Stream. This network stream which is composed of four cascaded Cohf-T blocks involves the $\mathbf{F}_s^0$ and

$\mathbf{F}_c^0$ as inputs. In particular, for the i-th Cohf-T block, we first combine previous prior features and additional structure features acquired from main stream features as $\mathbf{F}_s^i = C(\mathbf{P}^{i-1}||\bar{\mathbf{F}}_s^i)^2$. Then, short-distance window attention and long-distance window attention are sequentially integrated to extract the high-frequency structure knowledge from $\mathbf{F}_s^i$. Next, together with inter-modality context prior features $\mathbf{F}_c^0$, the obtained attention features are further passed through an inter-modality attention module to generate the final domain prior embedding features $\mathbf{P}^i$.

Output Gate. Based on the latent feature $\mathbf{F}^4$ extracted from the main super-resolution stream and the domain prior embedding feature $\mathbf{P}^4$, an output gate is built up for jointly synthesizing high-resolution intensity and gradient images. The concrete architecture is shown in Fig. 3 (a).

Objective Function. We adopt the mean square error and structural similarity index measure to estimate the consistency between network predictions and the ground-truths. The following objective function is adopted for constraining the main prediction $\mathbf{I}_{out}$,

$$L_{in} = \alpha \text{MSE}(\mathbf{I}_{out}, \mathbf{I}_{gt}) - (1-\alpha)(\text{SSIM}(\mathbf{I}_{out}, \mathbf{I}_{gt})), \quad (1)$$

where α is a weighting coefficient, and $\mathbf{I}_{gt}$ represents the ground-truth HR T2WI. MSE$(\cdot)$ denotes the mean square error function, and SSIM$(\cdot)$ denotes the function for calculating the structural similarity index measure.

The similar objective function is adopted for calculating the training loss for the gradient prediction $\mathbf{R}_{out}$,

$$L_c = \alpha \text{MSE}(\mathbf{R}_{out}, \mathbf{R}_{gt}) - (1-\alpha)\text{SSIM}(\mathbf{R}_{out}, \mathbf{R}_{gt}), \quad (2)$$

where $\mathbf{R}_{gt} = \sqrt{(\nabla_x \mathbf{I}_{gt})^2 + (\nabla_y \mathbf{I}_{gt})^2 + \epsilon}$. The overall objective function is formed by combining (1) and (2), $L = L_{in} + \lambda L_c$. λ is a weighting coefficient for the gradient restoration constraint.

3.2 Main Designs in Cofh-T

3.2.1 The Basic Attention Module. Different from the classic multi-head self-attention (MHSA) that is widely used in the existing transformer models [8, 35], we propose an alternative attention module as the basic unit of our model by taking both intra-head and inter-head correlations into consideration.

Given an input feature map $\mathbf{X}_1^{BA} \in \mathbb{R}^{h_1 \times w_1 \times d}$ and a reference context feature map $\mathbf{X}_2^{BA} \in \mathbb{R}^{h_2 \times w_2 \times d}$ ($\mathbf{X}_2^{BA}$ might be equal to $\mathbf{X}_1^{BA}$), the target of the attention module is to extract the well-aligned context information from $\mathbf{X}_2^{BA}$ for enhancing the representation of $\mathbf{X}_1^{BA}$. The whole calculation process is illustrated in Fig. 3 (b).

First, the layer normalization is applied to separately standardize $\mathbf{X}_1^{BA}$ and $\mathbf{X}_2^{BA}$. A 1×1 convolution accompanied with the other layer normalization and activation function GELU is utilized to embed $\mathbf{X}_1^{BA}$ into a d-dimensional feature map. Afterwards, it is decomposed into local patches with size of $p \times p$ and then arranged into patch-wise representation $\mathbf{x}_1^{BA} \in \mathbb{R}^{N \times dp^2}$ ($N = h_1 w_1 / p^2$) via the unfolding operation which flattens local patches in the feature map into vectors. The similar process is utilized to transform $\mathbf{X}_2^{BA}$ to patch-wise representation $\mathbf{x}_2^{BA}$, in which the size and stride of the convolution are both set to $\rho \times \rho$ for registering the spatial dimensions of $\mathbf{X}_2^{BA}$ with those of $\mathbf{X}_1^{BA}$, i.e., $\rho = h_2/h_1 = w_2/w_1$. M groups of linear layers are used for generating M query representations denoted by $\cup_{m=1}^{M} \mathbf{q}^{(m)} \in \mathbb{R}^{N \times d'}$ ($d' = dp^2/M$) from $\mathbf{x}_1^{BA}$. Key representations (denoted by $\cup_{m=1}^{M} \mathbf{k}^{(m)} \in \mathbb{R}^{N \times d'}$) and value representations (denoted by $\cup_{m=1}^{M} \mathbf{v}^{(m)} \in \mathbb{R}^{N \times d'}$) are calculated by feeding $\mathbf{x}_2$ into another $2M$ linear layers.

Then, like the classic MHSA, M intra-head correlation maps are calculated by the softmax function. The value at (i, j) of the m-th correlation map $\mathbf{s}^{(m)}$ is estimated as,

$$s_{i,j}^{(m)} = \frac{\exp(\|q_i^{(m)} \circ k_j^{(m)}\| / \sqrt{d'})}{\sum_{j'=1}^{N} \exp(\|q_i^{(m)} \circ k_{j'}^{(m)}\| / \sqrt{d'})}, \quad (3)$$

where $q_i^{(m)}$ and $k_j^{(m)}$ represent the i-th and the j-th rows of $\mathbf{q}^{(m)}$ and $\mathbf{k}^{(m)}$, respectively, and $\| \cdot \|$ denotes the summation of all elements in the input tensor. These correlation maps estimate point-wise dependencies between two input feature maps from the same heads. With the correlation encoded by $\mathbf{s}^{(m)}$, we renew the value feature by, $\hat{\mathbf{v}}^{(m)} = \mathbf{s}^{(m)} \mathbf{v}^{(m)}$.

To explore the dependencies among different heads, we devise an inter-head correlation modelling algorithm. First, the inter-head correlation matrix $\mathbf{A} \in \mathbb{R}^{N \times M \times M}$ is estimated as below,

$$A_{n,i,j} = \frac{\exp(\|\hat{v}_n^{(i)} \circ \hat{v}_n^{(j)}\|)}{\sum_{j=1}^{m} \exp(\|\hat{v}_n^{(i)} \circ \hat{v}_n^{(j)}\|)}, \quad (4)$$

where $\hat{v}_n^{(i)}$ represents the n-th row of $\hat{\mathbf{v}}^{(i)}$. Then, the value features are combined with the inter-head correlation, resulting in $\cup_{m=1}^{M} \mathbf{u}^{(m)} \in \mathbb{R}^{N \times d'}$. The n-th row of $\mathbf{u}^{(m)}$ is calculated as,

$$u_n^{(m)} = \sum_{j=1}^{M} (1 + A_{n,m,j}) \hat{v}_n^{(j)}, \quad (5)$$

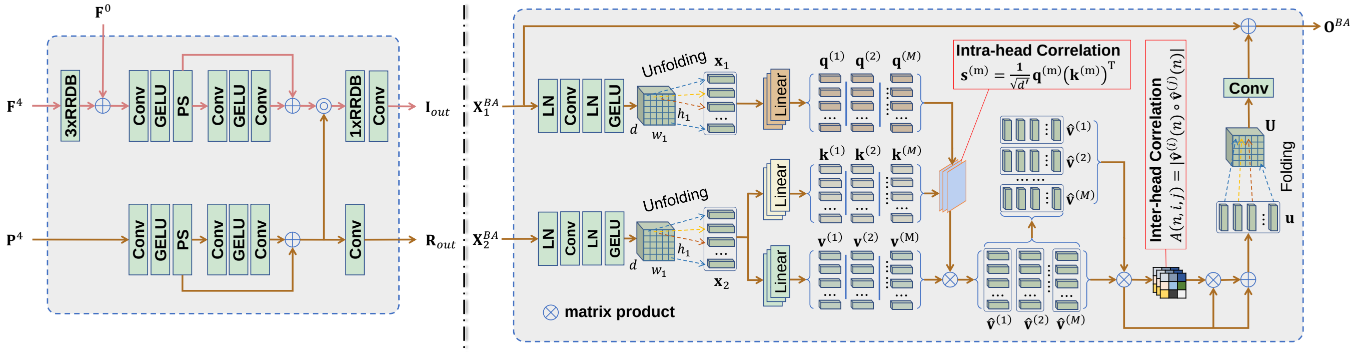
where $\mathbb{I} \in \mathbb{R}^{N \times M \times M}$ is formed by stacking N unit matrices with size of $M \times M$. Afterwards, $\mathbf{u}^{(m)}$-s are transformed into a $h_1 \times w_1 \times d$ tensor denoted by $\mathbf{U}$ with the folding operation[3]. Finally, a 3×3 convolution layer is adopted to post-process $\mathbf{U}$ which is then fused into $\mathbf{X}_1^{BA}$ via a skip connection, deriving of an enhanced variant of $\mathbf{X}_1^{BA}$ (namely $\mathbf{O}^{BA}$).

3.2.2 Intra-modality Window Attention. Short-distance and long-distance windows are adopted to model the intra-modality dependency for restoring the high-resolution gradient field. In particular, the short-distance structure helps to amend local details such as noncontinuous boundaries with surrounding structure information. At the same time, the long-distance structure can leverage the repetition of textural and structural patterns to enhance the pixel-wise feature representation.

Inspired from [21], for acquiring the short-distance structure, we uniformly decompose the input feature map $\mathbf{X} \in \mathbb{R}^{h \times w \times d}$ into compact windows with the size of $g \times g$ as in Fig. 4 (a). Regarding the input feature map itself as the reference context features, the attention module in Section 3.2.1 is applied for enhancing features of every window. Afterward, the features are processed with a residual

[2] For the first Cohf-T block, $\mathbf{F}_s^1 = C(\mathbf{P}^0||\bar{\mathbf{F}}_s^1)$ where $\mathbf{P}^0 = \mathbf{F}_s^0$.

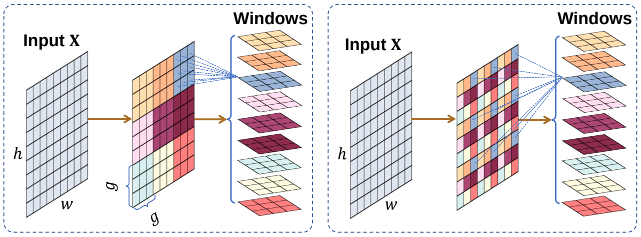
[3] The folding operation first expands the first dimension of $\mathbf{u}$ into the spatial size of $h_1/p \times w_1/p$ and then allocates the p^2d-dimensional feature vector of every point into a $p \times p \times d$ patch.



(a) Output Gate

(b) Basic Attention Module

Figure 3: Details of the output gate (a) and basic attention module (b). In (a), 'PS' denotes the pixel shuffle operation, and 'GELU' denotes the Gaussian error linear unit [12]. In (b), 'LN' stands for layer normalization; 'Unfolding' aggregates features of local patches into patch-level representations; 'Folding' is the reverse operation of 'unfolding'.



(a) Short-distance Window Sampling

(b) Long-distance Window Sampling

Figure 4: Two window sampling designs are employed for exploring short-distance (a) and long-distance (b) dependencies, respectively.

multi-layer perceptron (MLP) with a structure of 'LN-Conv1x1-LN-GELU-Conv1×1' (see Cohf-T in Fig. 2).

For extracting the long-distance dependency information, we propose to sample dilated windows from the input feature map (as shown in Fig. 4 (b)). The horizontal or vertical distance between neighboring points in each window is w/g or h/g. Then, the attention module, together with a residual MLP, is employed to enhance the features of each long-distance window. Once g is larger than w/g and h/g, cascading the short-distance and long-distance window attention modules can efficiently extract context information from the full image. The former module aggregates the context information from the surrounding $g \times g$ neighborhood, while the latter module involves the knowledge of all windows.

3.2.3 Inter-modality Attention. To involve the valuable inter-modality context information, we devise a cross-modality attention module within each Cohf-T block. Considering that there would be a large domain gap between the gradient fields of different modality data, directly injecting features of the T1 modality into those of the T2 modality would not be the optimal solution. Inspired from [13, 22], we adopt a point-wise adaptive instance normalization scheme to reduce the cross-modality representation gap as shown in Fig. 5.

Given the input feature maps $\mathbf{X}_1^{IA} \in \mathbb{R}^{h \times w \times d}$ and $\mathbf{X}_2^{IA} \in \mathbb{R}^{rh \times rw \times d}$, the goal is to transfer the content of $\mathbf{X}_2^{IA}$ to a feature space that is finely aligned to $\mathbf{X}_2^{IA}$. We estimate channel-wise mean $\boldsymbol{\mu}_2 \in \mathbb{R}^d$ and variance $\boldsymbol{\sigma}_2 \in \mathbb{R}^d$ from $\mathbf{X}_2^{IA}$, and then apply the instance normalization to standardize $\mathbf{X}_2^{IA}$, resulting in $\mathbf{X}_2'$ as below,

$$\mathbf{X}_2'[x, y, j] = \frac{\mathbf{X}_2^{IA}[x, y, j] - \boldsymbol{\mu}_2[j]}{\boldsymbol{\sigma}_2[j]}. \tag{6}$$

To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we first align the height and width of $\mathbf{X}_1^{IA}$ with those of $\mathbf{X}_2^{IA}$. Specifically, one convolution layer is first used to increase the channel number of $\mathbf{X}_1^{IA}$ to dr^2. Then, the resulted feature map is arranged to obtain $\mathbf{X}_1^h \in \mathbb{R}^{rh \times rw \times d}$ via the pixel shuffle operation. Afterward, $\mathbf{X}_2^{IA}$ and $\mathbf{X}_1^h$ are concatenated and compressed into a $rh \times rw \times d$ tensor (denoted by $\mathbf{X}^h$) via one convolution layer. Point-wise affine parameters $\boldsymbol{\beta} \in \mathbb{R}^{rh \times rw \times 1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{rh \times rw \times 1}$ are generated from $\mathbf{X}^h$ with separate convolution layers. We can also calculate the channel-wise mean ($\boldsymbol{\mu}_1$) and variance ($\boldsymbol{\sigma}_1$) vectors of $\mathbf{X}_1^{IA}$. Then, the module outputs $\mathbf{O}^{IA}$ as the updating of $\mathbf{X}_2^{IA}$:

$$\mathbf{O}^{IA}[x, y, j] = \mathbf{X}_2'[x, y, j](\boldsymbol{\sigma}_1[j] + \boldsymbol{\gamma}[x, y]) + \boldsymbol{\mu}_1[j] + \boldsymbol{\beta}[x, y]. \tag{7}$$

The adaptive normalization in [22] calculates global affine parameters are calculated in each batch normalization layer. Our devised normalization to differs from it in the adoption of point-wise affine parameters, which model local distribution deviation more finely.

The adaptive normalization helps project features of T1 modality to get close to the feature distribution of the T2 modality. The feature streams of $\mathbf{R}_c$ and $\mathbf{R}_s$ correspond to $\mathbf{X}_2^{IA}$ and $\mathbf{X}_1^{IA}$, respectively. Subsequently, the cross-modality attention is performed by using $\mathbf{X}_1^{IA}$ to generate the query features and using $\mathbf{O}^{IA}$ to create the key and value features. Then, another residual MLP is attached for further feature enhancement.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Two multi-modal MR image datasets are employed for validating super-resolution algorithms.

- **BraTS2018** is composed of 750 MR volumes [25]. They are registered to a uniform anatomical template and interpolated

Method	BraTS2018 dataset						IXI dataset					
	2×		3×		4×		2×		3×		4×	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	30.92	0.9198	28.19	0.9062	25.87	0.8388	29.71	0.9067	26.01	0.8070	24.32	0.7296
RDN [41]	34.57	0.9536	32.70	0.9324	30.92	0.9172	33.54	0.9435	31.02	0.9307	29.89	0.8998
RCAN [40]	34.82	0.9556	33.11	0.9310	30.88	0.9301	33.66	0.9422	30.79	0.9302	29.60	0.8936
SAN [6]	34.81	0.9544	32.91	0.9332	30.90	0.9281	34.31	0.9513	31.28	0.9401	29.59	0.8825
CFSR [33]	34.79	0.9566	32.66	0.9299	30.77	0.9290	34.24	0.9521	31.30	0.9320	29.78	0.8874
HAN [26]	35.01	0.9572	33.19	0.9351	31.03	0.9262	34.10	0.9480	31.36	0.9377	29.96	0.8993
SRResCGAN [34]	34.21	0.9492	33.00	0.9388	30.99	0.9048	34.02	0.9522	31.21	0.9388	30.24	0.8924
SPSR [23]	35.11	0.9585	33.32	0.9370	31.11	0.9264	34.47	0.9511	31.61	0.9382	30.47	0.8931
SMSR [36]	35.31	0.9601	33.62	0.9402	31.32	0.9342	34.77	0.9588	31.88	0.9477	30.36	0.9004
NLSN [24]	35.43	0.9655	33.99	0.9423	31.72	0.9321	34.51	0.9577	31.65	0.9432	30.64	0.8859
SwinIR [18]	35.42	0.9519	33.87	0.9425	31.84	0.9386	34.62	0.9401	31.83	0.9385	30.91	0.8915
TTSR [38]	-	-	-	-	31.33	0.9322	-	-	-	-	30.63	0.9044
MASA-SR [22]	-	-	-	-	31.44	0.9357	-	-	-	-	30.78	0.9032
MINet [9]	35.11	0.9587	33.25	0.9377	31.24	0.9355	34.34	0.9529	31.33	0.9208	30.58	0.8987
T2-Net [11]	35.32	0.9590	33.68	0.9420	31.64	0.9377	34.46	0.9532	31.60	0.9421	30.40	0.8992
MTrans [10]	35.35	0.9595	33.73	0.9411	31.73	0.9320	34.67	0.9556	31.69	0.9410	30.44	0.8993
Ours-S	35.91	0.9649	34.21	0.9501	32.10	0.9411	35.08	0.9598	32.17	0.9479	31.22	0.9081
Ours-M	36.13	0.9666	34.55	0.9516	32.30	0.9420	35.33	0.9612	32.45	0.9484	31.40	0.9101
Ours-L	**36.32**	**0.9692**	**34.73**	**0.9530**	**32.63**	**0.9431**	**35.60**	**0.9632**	**32.70**	**0.9492**	**31.71**	**0.9113**

Table 1: Comparison with existing methods on BraTS2018 and IXI datasets, under 2×, 3×, and 4× upsampling settings.

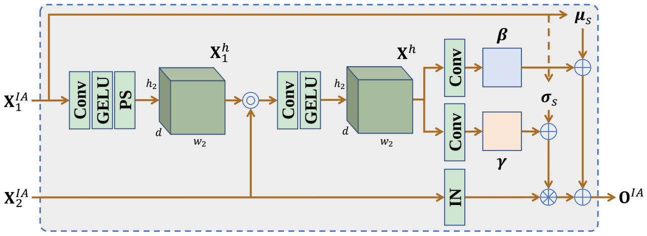


Figure 5: We devise a cross-modality feature adaptation module for aligning the feature distributions between two modalities.

into the same resolution (1mm×1mm×1mm). They are split into 484 volumes (including 75,020 images) for training, 66 volumes (including 10,230 images) for validating, and 200 volumes (including 31,000 images) for testing. The width and height of all images are both 240.

- **IXI** is composed of 576 MR volumes collected from three hospitals in London, including Hammersmith Hospital using a Philips 3T system, Guy's Hospital using a Philips 1.5T system, and Institute of Psychiatry using a GE 1.5T system. They are split into 404 volumes (including 48,480 images) for training, 42 volumes (including 5,040 images) for validating, and 130 volumes (including 15,600 images) for testing. The width and height of all images are both 256.

We follow [10] to synthesize low-resolution T2WIs. PSNR and SSIM are used for performance evaluation.

4.2 Implementation Details

Our method is implemented under PyTorch with a 32GB V100 GPU. Adam is used for network optimization, and the weight decay is set to 10^{-4}. The network is trained for 400 epochs with a mini-batch size of 4. The learning rate is initially set to 10^{-4} and decayed by half every 100 epochs. Without specification, the feature dimension d is set to 32; in the intra-modality window attention module, the input feature map is decomposed into windows with the size of 6×6 (namely $g = 6$), and the patch size p is set to 1; in the inter-modality attention module, the patch size p is set to 5; the number of attention heads M is set to 4; other parameters are set as: $\lambda = 0.5$ and $\alpha = 0.95$.

4.3 Comparison with Other Methods

We compare our method against extensive existing super-resolution methods including [6, 9–11, 18, 22–24, 26, 33, 34, 36, 38, 40, 41]. For [22, 38], the T1WI is regarded as the reference image. We implement three variants of our method with different model sizes, including: 1) For 'Ours-S', d is set to 16, two feature extraction and enhancement stages are used, and each RRDB block only contains two RDBs composed of three convolutions; 2) For 'Our-M', d is set to 16, three feature extraction and enhancement stages are used, and each RRDB block only contains three RDBs composed of three convolutions; 3) 'Our-L' is the final variant of our method with default settings. Experiments on BraTS2018 and IXI datasets are presented in Table 1.

Quantitative Comparisons. On BraTS2018 dataset, our small variant 'Our-S' performs better than the best natural image super-resolution method SwinIR [18] by 0.26dB, and the best reference-based image super-resolution method MASA-SR [22] by 0.66dB

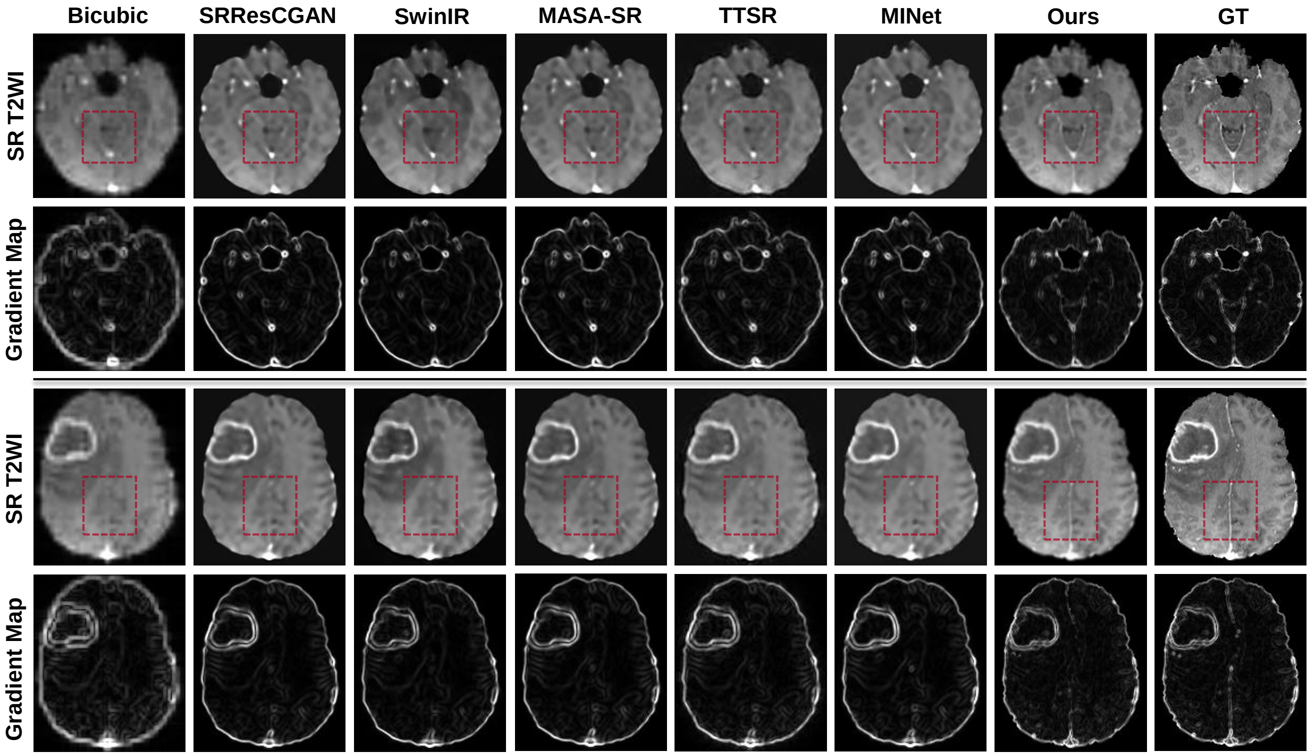


Figure 6: Qualitative comparison with other methods on 4× image super-resolution.

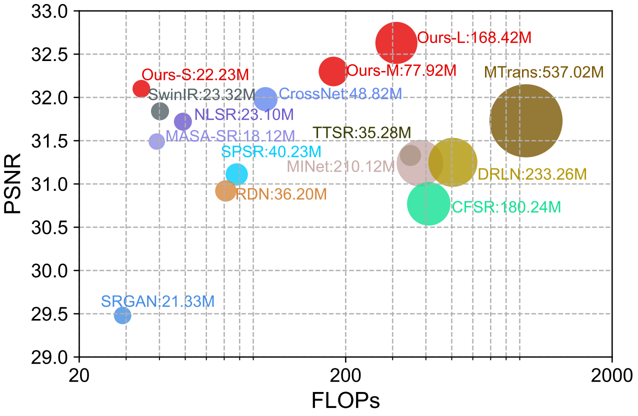


Figure 7: Visualisation of PSNR, FLOPS, and model size of image super-resolution algorithms. All metrics are evaluated in the 4× super-resolution task on BraTS2018 dataset.

under the 4× unsampling setting, while consuming fewer parameters and less memory cost (see Fig. 7). Compared to existing state-of-the-art MR image super-resolution method MTrans [10], our final variant 'Ours-L' gives rise to PSNR gains of 2.7%, 3.0%, and 2.8% under 2×, 3×, and 4× unsampling settings respectively. On IXI dataset, our method performs better than other methods as well. Particularly, it derives results with 2.7%, 3.2%, and 4.2% higher PSNR than the results of MTrans, under 2×, 3×, and 4× upsampling settings, respectively.

Qualitative Comparisons against other methods including SRResGAN [34], SwinIR [18], TTSR [38], MASA-SR and MINet are presented in Fig. 6. Compared to images super-resolved by other methods, the results of our approach have more accurate local details. The gradient maps indicate that our method can generate HR T2WI with sharper and more complete structures.

Model Complexity. The model sizes and memory consumption of different methods are illustrated in Fig. 7. Overall, our method outperforms other super-resolution methods with fewer parameters and less memory cost.

4.4 Ablation Study

Extensive ablation studies are conducted on the BraTS2018 dataset under the 4× upsampling setting to validate the efficacy of critical components in our method.

Efficacy of Attention Modules. We implement variants of our method by removing the attention modules or their inner units. The experimental results are reported in Table 2. The baseline model is formed by removing all intra-modality window attention and inter-modality attention modules. Without using the inter-modality attention (encoded as 'w/o InterM-A'), the PSNR is dramatically decreased by 0.75dB compared to the full version of our method. Qualitative comparisons are provided in Fig. 9 to illustrate the

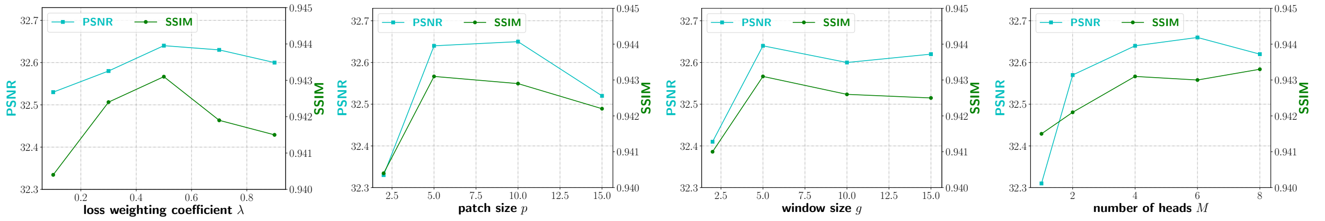


Figure 8: Performance of using different parameters. From left to right: the loss weighting factor λ; patch size p in the inter-modality attention module; window size g in intra-modality window attention modules; the number of attention heads M.

Variant	PSNR	SSIM
baseline	31.06	0.9225
w/o S-IntraM-WA or L-IntraM-WA	32.14	0.9411
w/o S-IntraM-WA	32.33	0.9417
w/o L-IntraM-WA	32.40	0.9415
w/o InterM-A	31.89	0.9377
w/o InterH-Corr	32.31	0.9410
w/o AdaptIN	32.23	0.9393
full version	32.64	0.9431

Table 2: Ablation study on attention modules. 'S/L-IntraM-WA': short-distance/long-distance intra-modality window attention; 'InterM-A': inter-modality attention; 'InterH-Corr': inter-head correlation; 'AdaptIN': adaptive instance normalization.

Domain	Arch.	#Params	PSNR	SSIM
Input	CNN-RB	171.127M	31.86	0.9377
Input	Cohf-T	168.418M	32.36	0.9401
Gradient	CNN-RB	171.127M	31.90	0.9383
Gradient	Cohf-T	168.418M	32.64	0.9431

Table 3: Performance of applying different designs for domain prior embedding. 'CNN-RB' means the CNN-based residual block is used to replace the Cohf-T.

efficacy of using inter-modality context information. As can be observed, after incorporating the gradient map of the T1WI with our devised inter-modality attention module, the structural information is recovered more completely. When both short-distance and long-distance intra-modality window attentions are abandoned (encoded as 'w/o S-IntraM-WA or L-IntraM-WA'), the PSNR is decreased to 32.14dB, which is 0.50 lower than that of the full version. Removing any separate short-distance (w/o S-IntraM-W) or long-distance (w/o L-IntraM-WA) window attention leads to performance degradation, which indicates that short-distance and long-distance dependencies have complementary effects in extracting the high-frequency structure priors. Without using the adaptive instance normalization (w/o AdaptIN) for aligning feature distributions across modalities, the PSNR is decreased by 0.41dB. The removing of the inter-head correlations (w/o InterH-Cor) in the basic attention module induces to 0.33dB PSNR reduction.

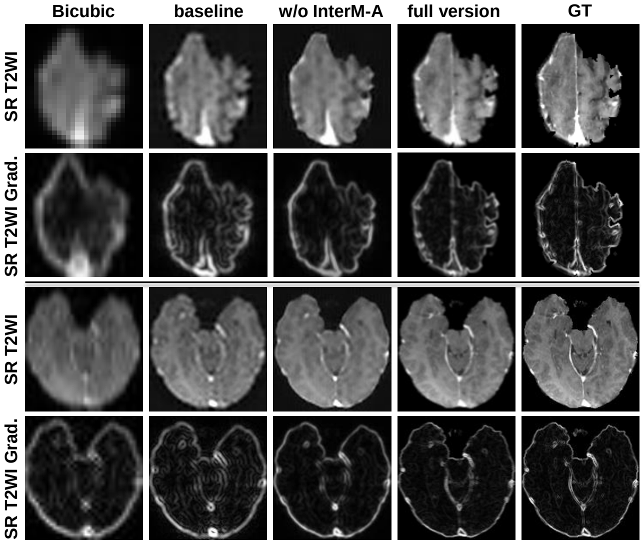


Figure 9: Qualitative comparison between different variants of our method. The baseline method is the exact CNN-based mainstream branch. 'w/o InterM-A' is the variant of our method in which only the gradient map of T2WI is utilized for extracting structure priors and the T1WI is not used. The high-frequency structure information and the cross-modality context can improve the super-resolution performance separately.

Different Designs for Domain Prior Embedding. We apply different designs for exploring high-frequency structure priors and inter-modality context priors as in Table 3. We try to replace the Cofh-T with CNN-based residual blocks to capture intra-modality and inter-modality dependencies. Though more parameters are used, CNN-based residual blocks perform worse than Cofh-T (e.g., having 0.74dB lower PSNR under the gradient domain). It is also validated that the original input domain is sub-optimal to the exploration of domain prior information. Performing domain prior embedding in the original input domain produces results with 0.28dB lower PSNR than the gradient domain when Cofh-T is used for attention modeling.

Choice of Hyper-parameters. The impact of hyper-parameters on the performance of our method is demonstrated in Fig. 8, from which we can observe that: 1) Loss weighting coefficient λ has a subtle impact on the learning framework; 2) For the inter-modality attention module, using too small patch size p would be unfavorable to constructing reliable cross-modality dependencies, while using a too large patch size would bring in noisy context knowledge; 3) For intra-modality window attention modules, larger window size g would benefit to a more thorough exploration of relational structure priors, while the performance tends to be saturated after g reaches 5; 4) Adopting more attention heads helps extract diversified types of relations. The performance increase as the M grows within four and tends to be saturated when $M > 4$.

5 CONCLUSION

In this paper, we devise a novel Transformer-based framework to tackle the MR image super-resolution task. A Cross-modality high-frequency Transformer (Cohf-T) module is proposed for exploring structure priors from the gradient domain and inter-modality context from an additional modality. We extract intra-modality and inter-modality dependencies to capture the domain priors via the Transformer module. The inter-head correlations can bring extra promotion to feature enhancement in attention modules. The distribution alignment strategy based on adaptive instance normalization is beneficial for fusing features of different modalities. Experiments on two datasets demonstrate that our method achieves state-of-the-art performance on the MR image super-resolution task.

REFERENCES

[1] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. 2011. Non-local means denoising. *Image Processing On Line* 1 (2011), 208–212.

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-trained image processing transformer. In *CVPR*.

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv:2102.04306* (2021).

[4] Yuhua Chen, Yibin Xie, Zhengwei Zhou, Feng Shi, Anthony G Christodoulou, and Debiao Li. 2018. Brain MRI super resolution using 3D deep densely connected neural networks. In *Proceedings of IEEE International Symposium on Biomedical Imaging*.

[5] Venkateswararao Cherukuri, Tiantong Guo, Steven J Schiff, and Vishal Monga. 2019. Deep MR brain image super-resolution using spatio-structural priors. *IEEE TIP* 29 (2019), 1368–1383.

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

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2015. Image super-resolution using deep convolutional networks. *IEEE TPAMI* 38, 2 (2015), 295–307.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929* (2020).

[9] Chun-Mei Feng, Huazhu Fu, Shuhao Yuan, and Yong Xu. 2021. Multi-Contrast MRI Super-Resolution via a Multi-Stage Integration Network. *arXiv:2105.08949* (2021).

[10] Chun-Mei Feng, Yunlu Yan, Geng Chen, Huazhu Fu, Yong Xu, and Ling Shao. 2021. MTrans: Multi-Modal Transformer for Accelerated MR Imaging. *arXiv:2106.14248* (2021).

[11] Chun-Mei Feng, Yunlu Yan, Huazhu Fu, Li Chen, and Yong Xu. 2021. Task Transformer Network for Joint MRI Reconstruction and Super-Resolution. *arXiv:2106.06742* (2021).

[12] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv:1606.08415* (2016).

[13] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

[14] Zohaib Iqbal, Dan Nguyen, Gilbert Hangel, Stanislav Motyka, Wolfgang Bogner, and Steve Jiang. 2019. Super-resolution 1h magnetic resonance spectroscopic imaging utilizing deep learning. *Frontiers in oncology* 9 (2019), 1010.

[15] Saurabh Jain, Diana M. Sima, Faezeh Sanaei Nezhad, Gilbert Hangel, Wolfgang Bogner, Stephen Williams, Sabine Van Huffel, Frederik Maes, and Dirk Smeets. 2017. Patch-Based Super-Resolution of MR Spectroscopic Images: Application to Multiple Sclerosis. *Frontiers in Neuroscience* 11 (2017), 13.

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

[17] Matthew CH Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. 2019. Image-and-spatial transformer networks for structure-guided image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. SwinIR: Image restoration using swin transformer. In *ICCV*.

[19] Tengfei Liang, Yi Jin, Yajun Gao, Wu Liu, Songhe Feng, Tao Wang, and Yidong Li. 2021. CMTR: Cross-modality Transformer for Visible-infrared Person Re-identification. *arXiv preprint arXiv:2110.08994* (2021).

[20] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. 2018. Non-local recurrent network for image restoration. *arXiv:1806.02919* (2018).

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030* (2021).

[22] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. 2021. MASA-SR: Matching Acceleration and Spatial Adaptation for Reference-Based Image Super-Resolution. In *CVPR*.

[23] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. 2020. Structure-preserving super resolution with gradient guidance. In *CVPR*.

[24] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. 2021. Image Super-Resolution With Non-Local Sparse Attention. In *CVPR*.

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

[26] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. 2020. Single image super-resolution via a holistic attention network. In *ECCV*.

[27] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. 2021. Conformer: Local Features Coupling Global Representations for Visual Recognition. *arXiv:2105.03889* (2021).

[28] Chi-Hieu Pham, Aurélien Ducournau, Ronan Fablet, and François Rousseau. 2017. Brain MRI super-resolution using deep 3D convolutional networks. In *Proceedings of IEEE International Symposium on Biomedical Imaging*.

[29] Esben Plenge, Dirk HJ Poot, Monique Bernsen, Gyula Kotek, Gavin Houston, Piotr Wielopolski, Louise van der Weerd, Wiro J Niessen, and Erik Meijering. 2012. Super-resolution methods in MRI: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? *Magnetic Resonance in Medicine* 68, 6 (2012), 1983–1993.

[30] Fang Qingyun, Han Dapeng, and Wang Zhaokui. 2021. Cross-Modality Fusion Transformer for Multispectral Object Detection. *arXiv preprint arXiv:2111.00273* (2021).

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

[32] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2021. Segmenter: Transformer for Semantic Segmentation. *arXiv:2105.05633* (2021).

[33] Chunwei Tian, Yong Xu, Wangmeng Zuo, Bob Zhang, Lunke Fei, and Chia-Wen Lin. 2020. Coarse-to-fine CNN for image super-resolution. *IEEE Transactions on Multimedia* 23 (2020), 1489–1502.

[34] Rao Muhammad Umer, Gian Luca Foresti, and Christian Micheloni. 2020. Deep generative adversarial residual convolutional networks for real-world super-resolution. In *CVPRW*.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

[36] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. 2021. Exploring Sparsity in Image Super-Resolution for Efficient Inference. In *CVPR*.

[37] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*.

[38] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. Learning texture transformer network for image super-resolution. In *CVPR*.

[39] Junwei Yang, Xiao-Xin Li, Feihong Liu, Dong Nie, Pietro Lio, Haikun Qi, and Dinggang Shen. 2021. Fast T2w/FLAIR MRI Acquisition by Optimal Sampling of Information Complementary to Pre-acquired T1w MRI. arXiv:2111.06400 [eess.IV]

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

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

[42] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. 2019. Channel splitting network for single MR image super-resolution. *IEEE TIP* 28, 11 (2019), 5649–5662.

[43] Feida Zhu, Chaowei Fang, and Kai-Kuang Ma. 2020. Pnen: Pyramid non-local enhanced networks. *IEEE TIP* 29 (2020), 8831–8841.