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

Chaowei Fang
Xidian University
chaoweifang@outlook.com

Dingwen Zhang∗
Northwestern Polytechnical University
zhangdingwen2006yyy@gmail.com

Liang Wang
Xidian University
2451207324@qq.com

Yulun Zhang
Computer Vision Lab, ETH Zürich
yulun100@gmail.com

Lechao Cheng
Zhejiang Lab
chenglc@zhejianglab.com

Junwei Han
Northwestern Polytechnical University
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. Experiments on two datasets indicate that Cohf-T achieves new state-of-the-art performance.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence; Machine learning.

KEYWORDS
magnetic resonance image, super-resolution, multi-modal learning

ACM Reference Format:
Chaowei Fang, Dingwen Zhang∗, Liang Wang, Yulun Zhang, Lechao Cheng, and Junwei Han. 2022. Cross-Modality High-Frequency Transformer for MR Image Super-Resolution. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547804

∗ Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM ’22, October 10–14, 2022, Lisboa, Portugal © 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-9203-7/22/10 $15.00 https://doi.org/10.1145/3503161.3547804

1 INTRODUCTION
Due to the superior capacity in capturing histopathological detail of soft tissues, magnetic resonance (MR) image becomes one of the 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 [31]. 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) [30, 46]. 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 [37]. 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 [34]. It is also verified that Transformer-based modelling is effective in medical image segmentation [3] and registration [19]. 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 [11, 42, 43] 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. 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 image super-resolution [7, 9, 10, 18, 28, 39, 45], CNN models have been the mainstream solutions for the MR image super-resolution [4, 30, 46]. According to the current studies [24, 25, 40], 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 [17, 33] attempt to explore prior context information from T1-weighted MR images for super-resolving T2-weighted or spectroscopy MR images. [11, 16, 41] 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 [37]. 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, 20, 23, 29, 34].

Among the existing vision transformer models, self-attention [1, 22, 47] and its variations, e.g., [6, 26], are widely adopted to build the basic transformer block. For example, [6] exploits the second-order attention based on covariance normalization for feature enhancement in image super-resolution. In [26], 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, [32] concatenates tokens from RGB and thermal modalities and then fuse them with symmetric attention modules. [21] 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 $I_{in} \in \mathbb{R}^{h \times w}$, we embed the high-frequency structure prior by calculating the gradient field for $I_{in}$ and define the gradient image as the structure reference input (denoted by $R_s = \sqrt{(\nabla_x I_{in})^2 + (\nabla_y I_{in})^2} + \epsilon$, where $\epsilon$ is a constant and is set to $10^{-5}$). The gradient of the HR T1WI (denoted by $R_c$) is regarded as an additional reference input for supplying inter-modality context information. $h$ and $w$ denote the height and width of $I_{in}$, respectively. Then, a network architecture is built upon the cross-modality high-frequency transformer to predict the high-resolution T2-weighted image $I_{out} \in \mathbb{R}^{rh \times rw}$, $r$ 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 $I_{in}$) 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 $R_s$ and the HR T1WI gradient $R_c$.

**Input Gate.** The input gate projects the input image data, including $I_{in}$, $R_s$, and $R_c$ to their corresponding primary features $F^0_s$, $F^0_c$, and $F^0_i$ through several convolutional layers and residual-in-residual dense blocks (RRDBs) [39].

**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 $E_i = R(F^{i-1})$. Then, $E_i$ is fed into a cross-modality high-frequency transformer (Cohf-T) block to interact with the high-frequency structure prior and inter-modality context prior, resulting in a domain prior embedding $P^i = \text{Cohf-T}(E_i, F_s^i, F_c^i)$. A more detailed description of the Cohf-T block can be referred to in the next subsection. Extra structural information is extracted from $E_i$ via a convolution layer, resulting in $F_i^c = \text{C}(E_i)$. Then, a series of attention modules are employed to explore domain priors (namely $P^i$) from $F_i^s$, $F_i^c$, and $F_i^i$. Finally, the prior-induced latent features of the current stage are produced via a shortcut connection: $F^i = E^i + T^i \circ P^i$, where $T^i$ is a single-channel selection map inferred from $F_i^s$, and ‘$\circ$’ is the broadcast element-wise production.

**Domain Prior Embedding Stream.** This network stream consists of four cascaded Cohf-T blocks that share the same architecture as the main super-resolution stream, with $P^0_i$ 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 $F_i^s = C(P^{i-1} \circ F_i^c)$. Then, short-distance window attention and long-distance window attention are sequentially integrated to extract the high-frequency structure knowledge from $F_i^s$. Next, together with inter-modality context prior features $P^i_c$, the obtained attention features are further passed through an inter-modality attention module to generate the final domain prior embedding features $P^i$.

**Output Gate.** Based on the latent feature $F^i$ extracted from the main super-resolution stream and the domain prior embedding feature $P^i$, 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...
function is adopted for constraining the main prediction $I_{out}$,

$$L_{in} = \alpha \text{MSE}(I_{out}, I_{gt}) - (1 - \alpha)(\text{SSIM}(I_{out}, I_{gt})),$$  

where $\alpha$ is a weighting coefficient, and $I_{gt}$ 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 $R_{out}$,

$$L_c = \alpha \text{MSE}(R_{out}, R_{gt}) - (1 - \alpha)\text{SSIM}(R_{out}, R_{gt}),$$  

where $R_{gt} = \sqrt{(\nabla_xI_{gt})^2 + (\nabla_yI_{gt})^2 + \epsilon}$. The overall objective function is formed by combining (1) and (2), $L = L_{in} + \lambda L_c$. $\lambda$ 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, 37], 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 $X_{1A}^B \in \mathbb{R}^{h_1 \times w_1 \times d}$ and a reference context feature map $X_{2A} \in \mathbb{R}^{h_2 \times w_2 \times d}$, the target of the attention module is to extract the well-aligned context information from $X_{2A}$ for enhancing the representation of $X_{1A}^B$. The whole calculation process is illustrated in Fig. 3 (b).

First, the layer normalization is applied to separately standardize $X_{1A}^B$ and $X_{2A}$. A $1 \times 1$ convolution accompanied with the layer normalization and activation function GELU is utilized to embed $X_{1A}^B$ into a $d$-dimensional feature map. Afterwards, it is decomposed into local patches with size of $p \times p$, then and arranged into patch-wise representation $X_{1A}^B \in \mathbb{R}^{N \times d'}$ ($N = h_1w_1/p^2$) via the unfolding operation which flattens local patches in the feature map into vectors. The similar process is utilized to transform $X_{2A}$ to patch-wise representation $X_{2A}^B$, in which the size and stride of the convolution are both set to $p \times p$ for registering the spatial dimensions of $X_{2A}^B$ with those of $X_{1A}^B$, i.e., $p = h_2/h_1 = w_2/w_1$. $M$ groups of linear layers are used for generating $M$ query representations denoted by $\cup_{m=1}^{M} q^{(m)} \in \mathbb{R}^{N \times d'}$ ($d' = dp^2/M$) from $X_{1A}^B$.

Key representations (denoted by $\cup_{m=1}^{M} k^{(m)} \in \mathbb{R}^{N \times d'}$) and value representations (denoted by $\cup_{m=1}^{M} v^{(m)} \in \mathbb{R}^{N \times d'}$) are calculated by feeding $X_{2A}$ 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 $s^{(m)}$ is estimated as,

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

where $q_i^{(m)}$ and $k_j^{(m)}$ represent the $i$-th and the $j$-th rows of $q^{(m)}$ and $k^{(m)}$, 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 $s^{(m)}$, we renew the value feature by $\hat{v}^{(m)} = s^{(m)} v^{(m)}$.

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

$$A_{n,i,j} = \frac{\exp(\|\hat{v}_i^{(n)} \circ \hat{v}_j^{(n)}\|)}{\sum_{j'=1}^{M}\exp(\|\hat{v}_i^{(n)} \circ \hat{v}_{j'}^{(n)}\|)},$$  

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

$$u_{n}^{(m)} = \sum_{j=1}^{M}(1 + A_{n,m,j})\hat{v}_{i}^{(j)},$$  

where $A \in \mathbb{R}^{N \times M \times M}$ is formed by stacking $N$ unit matrices with size of $M \times M$. Afterwards, $u^{(m)}$s are transformed into a $h_1 \times w_1 \times d'$ tensor denoted by $U$ with the folding operation\(^3\). Finally, a $3 \times 3$ convolution layer is adopted to post-process $U$ which is then fused into $X_{2A}$ via a skip connection, deriving of an enhanced variant of $X_{2A}$ (namely $O_{BA}$).

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

---

**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 [14]. In (b), ‘LN’ stands for layer normalization; ‘Unfolding’ aggregates features of local patches into patch-level representations; ‘Folding’ is the reverse operation of ‘unfolding’.\(^4\)
At the same time, the long-distance structure can leverage the repe-

mation to standardize $X$ and variance $X$ is finely aligned to

domain gap between the gradient fields of different modality data,

To involve the valuable inter-modality

3.2.3 Inter-modality Attention. Short-distance and long-
distance windows are adopted to model the intra-modality depen-
dencies, respectively.

3.2.2 Intra-modality Window Attention. Short-distance and long-
distance windows are adopted to model the intramodality depen-
dencies, respectively.

Exploring short-distance (a) and long-distance (b) dependen-
ties, respectively.

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

Given the input feature maps $X_{1A} \in \mathbb{R}^{h \times w \times d}$ and $X_{2A} \in \mathbb{R}^{h \times w \times d}$, the goal is to transfer the content of $X_{2A}$ to a feature space
that is finely aligned to $X_{1A}$. We estimate channel-wise mean $\mu_2 \in \mathbb{R}^d$ and variance $\sigma_2 \in \mathbb{R}^d$ from $X_{1A}$, and then apply the instance
normalization to standardize $X_{2A}$, resulting in $X_2$ as below,

$$X_2[x, y, j] = \frac{X_{2A}[x, y, j] - \mu_2[j]}{\sigma_2[j]}.$$  (6)

To estimate $\beta$ and $\gamma$, we first align the height and width of $X_{1A}$ with
those of $X_{2A}$. Specifically, one convolution layer is first used to in-
crease the channel number of $X_{1A}$ to $dr^2$. Then, the resulted feature
map is arranged to obtain $X_{1h} \in \mathbb{R}^{h \times r \times w \times d}$ via the pixel shuffle op-
eration. Afterward, $X_{1A}$ and $X_{1h}$ are concatenated and compressed into
a $rh \times rw \times d$ tensor (denoted by $X^{h}$) via one convolution layer.

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 [27]. They are
registered to a uniform anatomical template and interpolated
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.
The dataset is composed of 750 MR volumes [27]. They are
split into 484 volumes (including 48,480 images) for training, 66
volumes (including 7,500 images) for validating, and 130 volumes
(including 15,600 images) for testing. The width and height of
all images are both 256.
| Method         | PSNR | SSIM |
|---------------|------|------|
| Bicubic       | 30.92| 0.9198 |
| RDN [45]      | 34.57| 0.9536 |
| RCAN [44]     | 34.82| 0.9556 |
| SAN [6]       | 34.81| 0.9544 |
| CFSR [35]     | 34.79| 0.9566 |
| HAN [28]      | 35.01| 0.9572 |
| SRResCGAN [36]| 34.21| 0.9492 |
| SPSR [25]     | 35.11| 0.9585 |
| SMSR [38]     | 35.31| 0.9601 |
| NLSN [26]     | 35.43| 0.9655 |
| SwinIR [20]   | 35.42| 0.9519 |
| TTSR [40]     | -    | -    |
| MASA-SR [24]  | -    | -    |
| MINet [11]    | 35.32| 0.9590 |
| T2-Net [13]   | 35.35| 0.9595 |
| MTrans [12]   | 36.85| 0.9739 |
| Ours-S        | 36.10| 0.9681 |
| Ours-M        | 36.29| 0.9698 |
| Ours-L        | 36.85| 0.9739 |

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

We follow [12] 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 $p$ is set to 1; in the inter-modality attention module, $p$ is set to 5; $M$ 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, 11–13, 20, 24–26, 28, 35, 36, 38, 40, 44, 45]. For [24, 40], 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 ‘Ours-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) ‘Ours-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 [20] by 0.41dB, and the best reference-based image super-resolution method MASA-SR [24] by 0.81dB under the 4x upsampling 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 [12], our final variant ‘Ours-L’ gives rise to PSNR gains of 4.2%, 3.3%, and 4.8% under 2x, 3x, and 4x upsampling settings respectively. On IXI dataset, our method performs better than other methods as well. Particularly, it derives results with 3.1%, 3.5%, and 4.5% higher PSNR than the results of MTrans, under 2x, 3x, and 4x upsampling settings, respectively.

Qualitative Comparisons against other methods including SR-ResGAN [36], SwinIR [20], TTSR [40], 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 4x 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 1.09dB compared to the full version of our method. Qualitative comparisons are provided in Fig. 9 to illustrate the 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 32.20dB, which is 1.06dB lower than that of the full version. Removing any separate short-distance (w/o S-IntraM-WA or L-IntraM-WA) 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 0.49dB. The removing of the inter-head correlations (w/o InterH-Corr) in the basic attention module induces to 0.57dB PSNR reduction.

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 Coffi-T with CNN-based residual blocks to capture intra-modality
The impact of hyper-parameters on the performance of our method is demonstrated in Fig. 8, from which we can observe that: 1) Our method achieves the best performance when the loss weighting coefficient $\lambda$ is set to around 0.5; 2) For the inter-modality attention module, using too small patch size $p$ would be unfavorable to constructing reliable cross-modality dependencies, while the performance tends to be saturated after $p$ reaches 5; 3) For intra-modality window attention modules, larger window size $g$ would benefit to a more thorough exploration of relational structure priors. However, the performance saturates when $g$ becomes larger than 5; 4) Adopting more attention heads helps extract diversified types of relations. The performance increase as the $M$ grows within 4 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.

**ACKNOWLEDGMENTS**

This work was supported in part by the National Natural Science Foundation of China (No. 62003256, 62106235, 61876140, 62027813, U1801265, and U21B2048), in part by Open Research Projects of Zhejiang Lab (No. 2019KD001D01010), and in part by the Exploratory Research Project under Grant 2022FG0AN01.
