Learning Attention from Attention: Efficient Self-Refinement Transformer for Face Super-Resolution

Guanxin Li[1] , **Jingang Shi**[1*] , **Yuan Zong**[2] , **Fei Wang**[1] , **Tian Wang**[3] , **Yihong Gong**[1]
[1]School of Software Engineering, Xi'an Jiaotong University
[2]Key Laboratory of Child Development and Learning Science, Southeast University
[3]Institute of Artificial Intelligence, Beihang University
{liguanxin, jingang, feynmanw, ygong}@xjtu.edu.cn,
xhzongyuan@seu.edu.cn, wangtian@buaa.edu.cn

Abstract

Recently, Transformer-based architecture has been introduced into face super-resolution task due to its advantage in capturing long-range dependencies. However, these approaches tend to integrate global information in a large searching region, which neglect to focus on the most relevant information and induce blurry effect by the irrelevant textures. Some improved methods simply constrain self-attention in a local window to suppress the useless information. But it also limits the capability of recovering high-frequency details when flat areas dominate the local searching window. To improve the above issues, we propose a novel self-refinement mechanism which could adaptively achieve texture-aware reconstruction in a coarse-to-fine procedure. Generally, the primary self-attention is first conducted to reconstruct the coarse-grained textures and detect the fine-grained regions required further compensation. Then, region selection attention is performed to refine the textures on these key regions. Since self-attention considers the channel information on tokens equally, we employ a dual-branch feature integration module to privilege the important channels in feature extraction. Furthermore, we design the wavelet fusion module which integrates shallow-layer structure and deep-layer detailed feature to recover realistic face images in frequency domain. Extensive experiments demonstrate the effectiveness on a variety of datasets. The code is released at https://github.com/Guanxin-Li/LAA-Transformer.

1 Introduction

Face Super-resolution (FSR) is a specific super-resolution problem which needs to consider the unique textures on the face, such as eyes, mouth, nose, etc. The reconstruction of these structures are critical to distinguishing identity information. In recent years, FSR technology has been widely used and has attracted much attention. With the development of Convolutional Neural Networks (CNN), many researchers have designed various CNN networks [Huang *et al.*, 2017; Zhang *et al.*, 2018a; Zhang *et al.*, 2018c; Shi and Zhao, 2019; Li *et al.*, 2021] to improve FSR performance. Moreover, face priors, such as facial landmarks and heatmaps, were also incorporated into some methods [Chen *et al.*, 2018; Bulat and Tzimiropoulos, 2018] in order to improve global face contour recovery.

The CNN-based network has limited receptive field since convolution is a local operation. Due to the symmetry of the face, we also need to establish long-term dependencies in the FSR task in order to reconstruct the complex details on facial components. Different from CNN, Transformer [Vaswani *et al.*, 2017] proposes a self-attention mechanism to establish global dependencies. In the field of computer vision, the pioneering work ViT [Dosovitskiy *et al.*, 2021] separates the image into patches of equal size and then calculates the self-attention between the patches in the whole image, yielding better results compared to CNN in the classification task. Following this, various vision Transformers [Wang *et al.*, 2021; Chu *et al.*, 2021; Yuan *et al.*, 2021; Liu *et al.*, 2021] were proposed to address different visual tasks. Subsequently, Transformers were also introduced into the SR task. Generally, recent Transformers mainly conduct self-attention in the global searching region or pre-defined local window. However, it may encounter practical difficulties when deals with FSR. For the self-attention performed on global image region, the reconstructed texture details are calculated by the combination of all input tokens, which fails to focus the attention on the most relevant ones. The integration of irrelevant textures could cause blurry artifacts on the reconstructed results. For self-attention conducted on local window, it may fail to produce high-frequency details on complex facial components (e.g., eyes) if flat textures dominate the pre-defined rectangular window.

To solve these problems, we propose an efficient self-refinement mechanism for Transformer, called Region Selection Attention (RSA), which first produces the coarse attention map for conducting self-attention and then learns fine-grained attention map from the coarse one for further refinement. In the coarse-grained self-attention, the attention map is calculated on the down-sampled scale, which is effective for reconstructing flat facial regions and has the advantage

*Corresponding author

of saving computational cost. To enhance the features of key patches in the coarse attention map, we adaptively divide several regions with the strongest attention as refined searching field to calculate fine-grained self-attention. In this way, we could further restore the detailed features to compensate the reconstructed coarse textures on the key regions. The advantages of the proposed RSA can be summarized in two aspects. First, it conducts a content-aware feature reconstruction that treats the coarse structure and detailed texture in different manners. Second, it could explore fine-grained self-attention in the receptive field with irregular shape, which is more robust compared to the traditional rectangular regions.

Another drawback in self-attention is that the operation takes the channel information of tokens equally without considering the importance. To privilege the important channels in feature extraction, we propose Feature Integration Module (FIM) which consists of alternating channel attention module and depth-wise convolution in dual-branch. It could promote a further step to achieve cross-spatial and cross-channel integration simultaneously in Transformer.

Furthermore, we design the Wavelet Fusion Module (WFM) which could modulate the global facial structure information and local detailed texture in frequency-domain. As we know, the shallow layer in the deep network contains the structural information (e.g., facial contour) while the deep layer has the advantage of extracting complex local details (e.g., eyelids). Different from previous methods which conduct simple concatenation or summation in the temporal domain, the proposed WFM achieves reconstruction by exploring the frequency property. We employ Wavelet Transformation (WT) to separate low-frequency and middle-frequency parts from shallow-layer features, while obtaining the high-frequency parts from deep-layer features. The modulation of frequency-specific feature maps is then conducted in WFM for better restoration.

Overall, our main contributions are summarized as follows:

- We propose an efficient self-refinement Transformer-based architecture for FSR task. It could adaptively conduct texture-aware reconstruction in a coarse-to-fine manner.
- The Feature Integration Module (FIM) is employed to consider cross-channel difference in Transformer, which promotes the integration of spatial-wise and channel-wise information for feature extraction.
- We design the Wavelet Fusion Module (WFM) to achieve the modulation of shallow-layer and deep-layer features through frequency decomposition and recombination.
- Our method achieves state-of-the-art quantitative metrics and visualizations. It obtains the advantages of more than 0.32dB for PSNR values on the Helen datasets.

2 Related Works

2.1 Face Super-Resolution

In recent years, deep learning has achieved great developments in the field of computer vision. The FSR task has attracted many researchers because of its wide application prospects. GLN [Tuzel *et al.*, 2016] designs the Global Upsampling Network to reconstruct the overall face and the Local Refinement Network to enhance the local details of the face. Attention-FH [Cao *et al.*, 2017] utilizes the context dependencies between facial components to recursively restore the details. RCAN [Zhang *et al.*, 2018c] presents a very deep residual channel attention network, which significantly enhances the learning capacity of CNN. FSRNet [Chen *et al.*, 2018] adds facial geometry priors such as facial landmarks, heatmaps and parsing maps to the network and achieves excellent results. SPARNet [Chen *et al.*, 2020] proposes a network built from Face Attention Units that can efficiently capture key features in very low-resolution face images. DIC [Ma *et al.*, 2020] adopts a novel iterative collaboration network to gradually obtain accurate facial landmarks and super-resolution images. IGAN [Li *et al.*, 2021] considers SR as the information-growth process and recovers HR images by exploring information differences in images of different resolutions. SRDD [Maeda, 2022] proposes a high-resolution (HR) dictionary that can be learned explicitly, which reduces the information that the network needs to process in the HR space. HGSRCNN [Tian *et al.*, 2022] adopts a heterogeneous structure to enhance the internal and external correlations of channels in parallel, which promotes the recovery of images.

2.2 Vision Transformer

The Transformer is originally used for sequence processing in natural language tasks. The proposal of ViT [Dosovitskiy *et al.*, 2021] proves that Transformer can achieve state-of-the-art performance in the image classification task. By recursively aggregating tokens of neighboring items, T2T [Yuan *et al.*, 2021] greatly decreases tokens length for practical application. [Wu *et al.*, 2021] builds a Pyramid Pooling Transformer to achieve better performance in various downstream vision tasks such as semantic segmentation and object detection. Swin Transformer [Liu *et al.*, 2021] decomposes the image into non-overlapping windows, calculates multi-head self-attention (MHSA) within the window, and introduces a shifted window mechanism to establish the cross-window connection. Recently, various vision Transformers [Wang *et al.*, 2021; Chu *et al.*, 2021] are proposed to address different visual tasks. Meanwhile, Transformers are also applied in the field of super-resolution. [Liang *et al.*, 2021] builds SwinIR based on Swin Transformer and achieves excellent results in SR tasks. VSR [Cao *et al.*, 2021] turns MHSA into a spatial-temporal convolutional self-attention to achieve state-of-the-art performance in video super-resolution tasks. [Shi *et al.*, 2022] proposes a pyramid encoder/decoder Transformer architecture to extract and restore feature textures in different spaces through a hierarchical structure.

3 Method

3.1 Overview

As shown in Figure 1(a), we describe three core stages of the proposed network: Feature Extraction Stage, Feature Transformation Stage, and Feature Recovery Stage. Let $I^L \in$

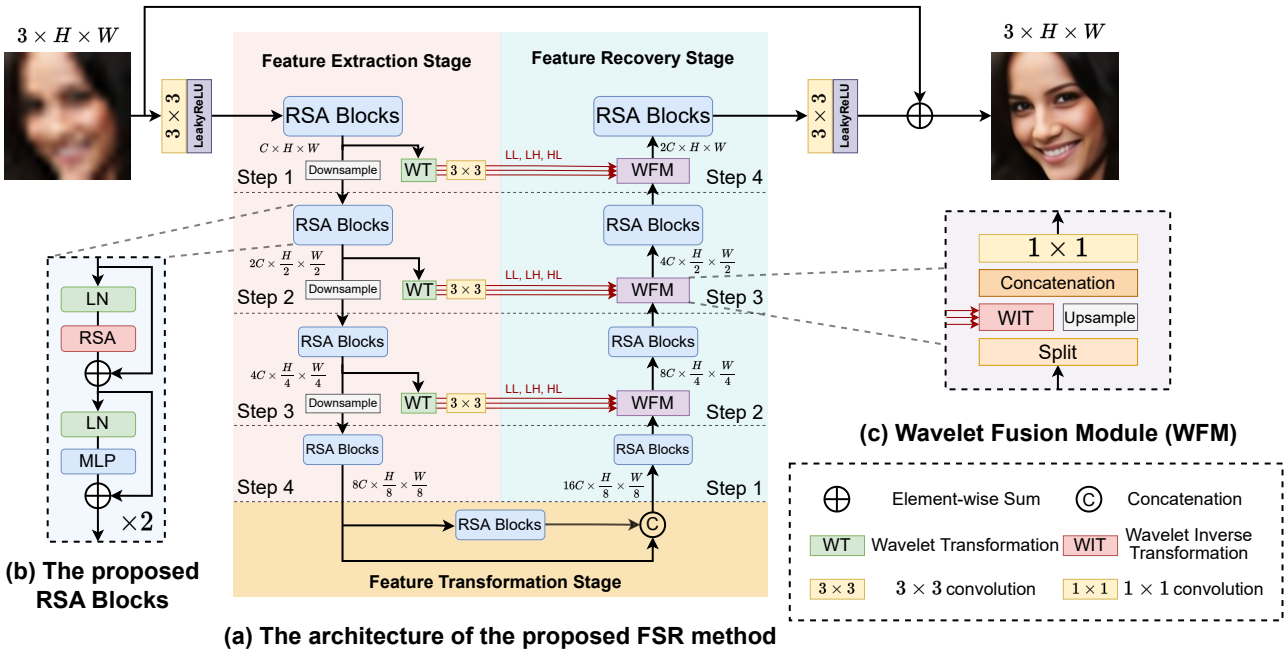

Figure 1: An illustration of the proposed face super-resolution architecture. (a) Three main stages of the network: Feature Extraction Stage, Feature Transformation Stage, and Feature Recovery Stage. (b) The Region Search Attention (RSA) Block consists of Layer Normalization (LN), Region Selection Attention (RSA), multi-layer perceptron (MLP), and residual connections. (c) The Wavelet Fusion Modules (WFM) can fuse three general features (LL, LH, and HL) from the Feature Extraction Stage and one high-frequency feature (HH) from the Feature Recovery Stage.

$\mathbb{R}^{3\times H\times W}$ be an input low-resolution RGB image, where H and W represent the height and width, respectively. We first expand the channel of input image I^L to C as:

$$F_{es}^0 = H_{EF}(I^L), \tag{1}$$

where $H_{EF}(\cdot)$ consists of a 3×3 convolution and a LeakyReLU activation. The convolutional layer expands the feature to a higher dimensional feature space, which is beneficial for recovering details in different feature channels.

Feature Extraction Stage. We extract features through a hierarchical feature pyramid structure containing 4 steps. Each step consists of 2 Region Selection Attention (RSA) Blocks, a down-sampling operation, a Wavelet Transformation (WT), and a 3×3 convolution. In particular, the 4-th step only contains 2 RSA Blocks. We get the output features F_{es}^i of i-th step as:

$$F_{es}^i = \begin{cases} Down(H_{RB}(F_{es}^{i-1})), & i = 1, 2, 3, \\ H_{RB}(F_{es}^{i-1}), & i = 4, \end{cases} \tag{2}$$

where $H_{RB}(\cdot)$ denotes 2 consecutive RSA Blocks, and $Down(\cdot)$ is the down-sampling operation. For the down-sampling operation, we apply a 4×4 convolutional layer with stride 2 to double the channel number and reduce the height and width to 1/2 of their original size, respectively. In each step, the features are also decomposed into frequency-domain features by Wavelet Transformation (WT). Our proposed WT is composed of a low-pass filter and a high-pass filter. The features sequentially pass through the combination of two filters, which can be converted into frequency-domain features: LL, LH, HL, and HH (see Sec. 3.3 for details). In the i-th step, for the output $\hat{F}_{es}^i$ of 2 RSA Blocks, we extract the low-frequency and middle-frequency features to present the global facial structure information:

$$\begin{aligned} LL^i, LH^i, HL^i &= H_{WT}(\hat{F}_{es}^i), \\ LL^i &= Conv_{3\times3}(LL^i), \\ LH^i &= Conv_{3\times3}(LH^i), \\ HL^i &= Conv_{3\times3}(HL^i), \end{aligned} \tag{3}$$

where $H_{WT}(\cdot)$ denotes WT and $Conv_{3\times3}(\cdot)$ is a 3×3 convolutional layer.

Feature Transformation Stage. At this stage, the output of the previous step F_{es}^4 is fed into 2 RSA Blocks and concatenated with itself as:

$$F_{ts} = Concat[H_{RB}(F_{es}^4), F_{es}^4], \tag{4}$$

where $Concat[\cdot]$ denotes concatenation operation.

Feature Recovery Stage. This stage also contains 4 steps to gradually restore high-resolution images. Each step consists of 2 RSA Blocks, and a Wavelet Fusion Module (WFM). The 1-st step doesn't contain the WFM. WFM can integrate the general facial features extracted from the Feature Extraction Stage and the corresponding high-frequency features in the Feature Recovery Stage through Wavelet Inverse Transformation (WIT).

We will describe the WFM in detail in Sec 3.4. More specifically, the output of j-th step can be formulated as:

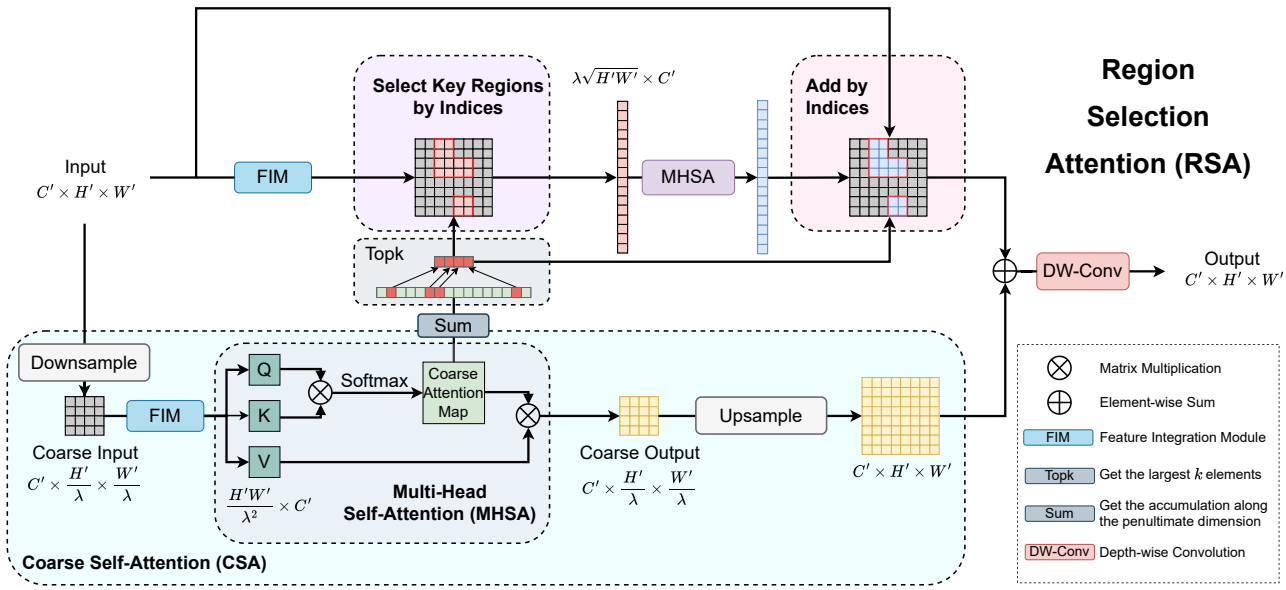

Figure 2: The architecture of our proposed Region Selection Attention (RSA). It consists of Feature Integration Module (FIM), Coarse Self-Attention (CSA), key regions selection, key regions self-attention, and a depth-wise convolution.

$$F_{rs}^{j} = \begin{cases} H_{RB}(F_{ts}), & j = 1, \\ H_{RB}(H_{WFM}(F_{rs}^{j-1}, LL^{i}, \\ \qquad LH^{i}, HL^{i})), & j = 2, 3, 4, \end{cases} \tag{5}$$

where $i = 5 - j$, and $H_{WFM}(\cdot)$ represents the WFM. Finally, the features $F_{rs}^{4} \in \mathbb{R}^{2C \times H \times W}$ are refined by a 3×3 convolution and a LeakyReLU activation to obtain the residual image $I^{R} \in \mathbb{R}^{3 \times H \times W}$. The final recovered image $\hat{I}$ can be obtained by $\hat{I} = I^{L} + I^{R}$.

3.2 Efficient Region Search Attention (RSA) Block

As shown in Figure 1(b), we replace the traditional multi-head self-attention (MHSA) with our proposed Region Selection Attention (RSA). Given the input as F_s, our proposed RSA Block can be formulated as:

$$\begin{aligned} F_s' &= RSA(LN(F_s)) + F_s, \\ \hat{F}_s &= MLP(LN(F_s')) + F_s', \end{aligned} \tag{6}$$

where $LN(\cdot)$ is Layer Normalization and $RSA(\cdot)$ is RSA, which selects multiple regions with rich high-frequency information and computes their self-attention.

RSA is a data-driven approach to adaptively implement texture-aware reconstruction in a coarse-to-fine procedure. Specifically, we get a global coarse attention map via Coarse Self-Attention (CSA). This coarse attention map can reflect the high-frequency information density of each region of the image. Under the guidance of the coarse attention map, we can find the largest k attention values in this feature map, representing the most important k regions in the original features. These regions are extracted from the original features and calculated MHSA to refine the results. Key regions with rich high-frequency information are thus reconstructed by the refinement procedure.

As illustrated in Figure 2, given the input $F_{input} \in \mathbb{R}^{C' \times H' \times W'}$, F_{input} is first downsampled into a smaller feature map F_{coarse} as:

$$F_{coarse} = Downsample(F_{input}), \tag{7}$$

where $Downsample(\cdot)$ uses a convolutional layer with kernel size of $\lambda \times \lambda$ and stride λ. Then the features are fed into the Feature Integration Module (FIM) to privilege the channel-wise importance. This can be formulated as:

$$\hat{F}_{input} = H_{FIM}(F_{input}), \hat{F}_{coarse} = H_{FIM}(F_{coarse}) \tag{8}$$

Then we transpose the coarse feature $\hat{F}_{coarse} \in \mathbb{R}^{C' \times \frac{H'}{\lambda} \times \frac{W'}{\lambda}}$ into $F_c \in \mathbb{R}^{n_c \times C'}$, where $n_c = \frac{H'}{\lambda} \times \frac{W'}{\lambda}$. We feed it into MHSA with M heads as:

$$\begin{aligned} Q_c = F_c W_q, K_c = F_c W_k, V_c = F_c W_v, \\ AM^{(h)} = Softmax(\frac{Q_c^{(h)} K_c^{(h)T}}{\sqrt{D_h}}), h = 1, \ldots, M, \\ SA^{(h)} = AM^{(h)} V_c^{(h)}, \\ Output_c = Concat(SA^{(1)}, \ldots, SA^{(M)}) W_o, \end{aligned} \tag{9}$$

where W_q, W_k, W_v, W_o are learnable parameters, and $D_h = C'/M$ is the number of dimensions for one head. $SA^{(h)}$, $AM^{(h)}$, $Q_c^{(h)}$, $K_c^{(h)}$ and $V_c^{(h)}$ represent the output of self-attention, attention map, query embedding, key embedding, and value embedding from the h-th attention head, respectively. Then we select the key regions according to the coarse attention map. It is can be defined as:

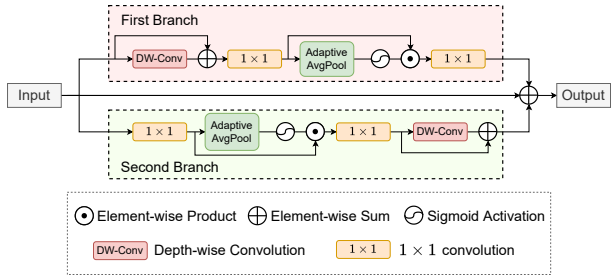

Figure 3: The structure of the Feature Integration Module (FIM).

$$\begin{aligned} AM &= Concat(AM^{(1)}, \ldots, AM^{(M)}), \\ Index &= GetIndex(Topk(Sum(AM))), \end{aligned} \tag{10}$$

where $AM \in \mathbb{R}^{M \times n_c \times n_c}$ means coarse attention map and $Sum(\cdot)$ represents accumulation along the penultimate dimension of AM, which turns region-to-region attention into region-to-global attention. $Topk(\cdot)$ means to get the largest k elements and $GetIndex(\cdot)$ denotes getting the indices of elements in feature $\hat{F}_{input}$. Here we let $k = \lfloor \frac{\sqrt{H'W'}}{\lambda} \rfloor$ to balance the computational cost and effect. Thus we get the indices of the key regions in $\hat{F}_{input}$. Next, we extract these regions to calculate the MHSA and add them back to $\hat{F}_{input}$ by indices. The output of this process can be obtained by:

$$\begin{aligned} F_{regions} &= SelectByIndex(\hat{F}_{input}, Index), \\ Output_f &= AddByIndex(F_{input}, \\ &\qquad MHSA(F_{regions}), Index), \end{aligned} \tag{11}$$

where $SelectByIndex(\cdot)$ means to select the regions in $\hat{F}_{input}$ according to $Index$ and $AddByIndex(\cdot)$ denotes adding $F_{regions}$ after MHSA to F_{input} according to $Index$. Finally, we upsample $Output_c$ and add it to $Output_f$, which is then fed into a depth-wise convolutional layer. It could be formulated as:

$$Output = DWConv(Output_f + Upsample(Output_c)), \tag{12}$$

where $Upsample(\cdot)$ denotes the up-sampling operation, which is a 4×4 transposed convolution with stride 2. $DWConv(\cdot)$ means 3×3 depth-wise convolution.

Feature Integration Module (FIM). As shown in Figure 3, our proposed FIM consists of depth-wise convolutions, channel attention module, and residual connection. Depth-wise convolution can effectively model the spatial information of features. The channel attention module allows the network to focus on channels which should be paid more attention. We change the combination of the depth-wise convolution and the channel attention for better interaction. Moreover, we utilize the residual connections to merge dual branches and the original input. This design complements the features with rich cross-spatial and cross-channel information.

In FIM, the kernel size of depth-wise convolution is 3×3, and Adaptive AvgPool means global average pooling operation.

3.3 Wavelet Transformation (WT) and Wavelet Inverse Transformation (WIT)

Our proposed Wavelet Transformation (WT) and Wavelet Inverse Transformation (WIT) consist of a convolutional block and a transposed convolution block, respectively. To decompose features into several frequency components, we adopt a high-efficiency wavelet transformation, namely the Haar wavelet.

$$L = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix}^T, H = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \end{bmatrix}^T, \tag{13}$$

where L and H represent low-pass and high-pass filters, respectively. Low-pass filters can capture general information like global contour, and facial structure. In contrast, a high-pass filter extracts local details such as texture, eyes, facial components, etc.

With the combination of two filters, we can achieve four kernels of Haar wavelet:

$$\begin{aligned} LL^T = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, LH^T = \frac{1}{2} \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \\ HL^T = \frac{1}{2} \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, HH^T = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{aligned} \tag{14}$$

The above four kernels could be utilized to decompose the feature map into frequency-domain components: LL, LH, HL, and HH. Given an arbitrary feature map F, the WT can be defined as:

$$\begin{aligned} LL &= Conv_{LL^T}(F), \\ LH &= Conv_{LH^T}(F), \\ HL &= Conv_{HL^T}(F), \\ HH &= Conv_{HH^T}(F), \end{aligned} \tag{15}$$

where $Conv_{LL^T}(\cdot)$, $Conv_{LH^T}(\cdot)$, $Conv_{HL^T}(\cdot)$, $Conv_{HH^T}(\cdot)$ represent group convolution with stride 2 and weights LL^T, LH^T, HL^T and HH^T, respectively.

We use WIT to integrate the general information (LL, LH, and HL) and the high-frequency details HH to reconstruct the face image. The WIT can be defined as:

$$\begin{aligned} \hat{F} = Deconv_{LL^T}(LL) + Deconv_{LH^T}(LH) + \\ Deconv_{HL^T}(HL) + Deconv_{HH^T}(HH), \end{aligned} \tag{16}$$

where $Deconv_{LL^T}(\cdot)$, $Deconv_{LH^T}(\cdot)$, $Deconv_{HL^T}(\cdot)$, $Deconv_{HH^T}(\cdot)$ represent four separate transposed convolutions with the weights as Eq. 14.

The WT operation achieves the modulation of shallow-layer and deep-layer feature maps from the perspective of the frequency domain. We employ skip connections to perform WIT in the Feature Recovery Stage, which can stabilize the generation of detailed information.

Figure 4: Visual comparisons for 8× FSR. Our method can produce more accurate details. In contrast, other methods generate undiscernible artifacts in complex regions. Zoom in for the best view.

3.4 Wavelet Fusion Module (WFM)

In this section, we describe our proposed Wavelet Fusion Module (WFM). As shown in Figure 1(c), for the feature F_{rs}, it is split evenly into two features F_{rs_L} and F_{rs_R} as:

$$F_{rs\_L}, F_{rs\_R} = Split(F_{rs}) \tag{17}$$

Then, F_{rs_L} and F_{rs_R} are fed into a Wavelet Inverse Transformation (WIT) and an up-sampling block, respectively, which are defined as:

$$\begin{aligned} \hat{F}_{rs\_L} &= H_{WIT}(F_{rs\_L}, LL, LH, HL), \\ \hat{F}_{rs\_R} &= Upsample(F_{rs\_R}), \end{aligned} \tag{18}$$

where $H_{WIT}(\cdot)$ and $Upsample(\cdot)$ separately represents the WIT and a 4×4 transposed convolution with stride 2. Deep-layer feature F_{rs_L} and the corresponding shallow-layer features LL, LH, HL are modulated to refine the detail in the WIT.

Then we feed the combination of two features into a convolutional layer to get the output of this module, which can be formulated as:

$$\hat{F}_{rs} = Conv_{1\times1}(Concat[\hat{F}_{rs\_L}, \hat{F}_{rs\_R}]), \tag{19}$$

where $Concat[\cdot]$ denotes concatenation operation. $Conv_{1\times1}(\cdot)$ is an 1×1 convolution, which reduces the channel number of the features to half.

4 Experiments

4.1 Datasets

The CelebA [Liu *et al.*, 2015] and the Helen [Le *et al.*, 2012] are two publicly available face image datasets. First, we use MTCNN [Zhang *et al.*, 2016] to crop the face region. After excluding images with a resolution smaller than 128×128, the image is resized to 128×128. So we obtained about 178k images from CelebA, of which 177k images were used as HR training images. Using bicubic interpolation, the HR images are downsampled to 16×16 to generate LR images. In the testing phase, we extract the remaining 1000 images from the cropped CelebA dataset and randomly extract 100 images from the cropped Helen dataset. For evaluation, we employ the following metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [Wang *et al.*, 2004] computed on the Y channel of the image YCbCr space, and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018b].

4.2 Implementation Details

Before the Feature Extraction Stage, we first extend the number of feature channels to 32. In the Feature Extraction Stage, the channel number of feature map is $2^{i-1} \times 32$ and the attention head number of RSA Block is 2^{i-1} in the i-th step. In the Feature Transformation Stage, the number of feature channels is set to 256. The number of attention heads is set to 16. In the Feature Recovery Stage at the j-th step, the channel number of feature map is $2^{5-j} \times 32$ and the attention head number of RSA Block is 2^{4-j}. When the height $\times$ width of the feature map is less than or equal to 32×32, we only

Method	Helen			CelebA		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Bicubic	24.25	0.6798	0.5236	24.10	0.6605	0.5267
RCAN [Zhang *et al.*, 2018c]	27.94	0.8196	0.1927	28.08	0.8149	0.1825
DIC [Ma *et al.*, 2020]	26.61	0.7755	0.2292	27.38	0.7911	0.1950
SPARNet [Chen *et al.*, 2020]	27.62	0.8094	0.2012	27.83	0.8067	0.1874
IGAN [Li *et al.*, 2021]	27.97	0.8213	**0.1845**	28.15	0.8171	**0.1767**
SwinIR [Liang *et al.*, 2021]	**28.04**	**0.8219**	0.1905	**28.27**	**0.8192**	0.1808
SRDD [Maeda, 2022]	27.64	0.8093	0.2184	27.88	0.8071	0.2058
HGSRCNN [Tian *et al.*, 2022]	27.93	0.8192	0.1885	28.20	0.8175	0.1793
Ours	**28.36**	**0.8318**	**0.1626**	**28.58**	**0.8297**	**0.1542**

Table 1: Quantitative comparison on Helen and CelebA test set for 8× FSR. The best and second-best performances are denoted by the *red* and *blue*.

employ CSA in the RSA with $\lambda = 1$. Otherwise, we apply the full RSA and $\lambda = 4$. The AdamW optimizer is used to train our model with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay is set to 0.02. The learning rate is initially set to 4×10^{-4} and dropped by half every 20 epochs. Meanwhile, L_1 loss is used for training. The batchsize is set to 32. Our model is implemented in PyTorch and trained for 100 epochs using 2 NVIDIA GeForce RTX 3090 GPUs.

4.3 Comparisons with State-of-the-Art Methods

We compare the proposed FSR method with several state-of-the-art methods: RCAN [Zhang *et al.*, 2018c], DIC [Ma *et al.*, 2020], SPARNet [Chen *et al.*, 2020], IGAN [Li *et al.*, 2021], SwinIR [Liang *et al.*, 2021], SRDD [Maeda, 2022], and HGSRCNN [Tian *et al.*, 2022]. Our quantitative results for Helen and CelebA dataset are shown in Table 1. The results demonstrate that our method achieves the best PSNR, SSIM, and LPIPS performance on both datasets, outperforming other state-of-the-art approaches by a significant margin. Specifically, for the PSNR value, our method outperforms the best other methods by 0.32dB and 0.31dB on the Helen and CelebA datasets, respectively. Qualitative results are shown in Figure 4. Our method can recover more realistic details in the blurry region, and the reconstructed result is closer to the real image. The pupils and nose of the face are restored more clearly. Other methods tend to produce blurring artifacts in regions with complex textures.

4.4 Ablation Study

Effectiveness of the Region Selection Attention (RSA). The core part of RSA is the region selection strategy. Relying on this design, our RSA can effectively focus on regions with long-range dependencies and rich high-frequency information. To demonstrate the effectiveness of this strategy, we remove the part that selects regions to compute key regions self-attention and only retain the Coarse Self-Attention (CSA). As shown in Table 2, our RSA improves the PSNR value by 0.12dB on the Helen test set. The above experiments demonstrate the effectiveness of the proposed RSA.

Effectiveness of the Feature Integration Module (FIM). Our proposed FIM is used to enhance cross-channel awareness and cross-spatial awareness of features. To verify the effectiveness of FIM, we use learnable position parameters, depth-wise convolution or single-branch FIM to replace full FIM. As can be seen from Table 3, compared to learnable position parameters and depth-wise convolution, our first-branch-only FIM and second-branch-only FIM improve the PSNR values by 0.04dB and 0.03dB, respectively. The full FIM achieved a significant PSNR value improvement of 0.11dB, which illustrates the superiority of our proposed innovative FIM.

Methods	RSA	CSA
PSNR	**28.36**	28.24
SSIM	**0.8318**	0.8297

Table 2: Comparison of Region Selection Attention (RSA) and Coarse Self-Attention (CSA).

Methods	PSNR	SSIM
learnable position parameters	28.25	0.8296
DW-Conv	28.25	0.8300
FIM only first branch	28.29	0.8301
FIM only second branch	28.28	0.8293
FIM	**28.36**	**0.8318**

Table 3: Comparison of learnable position parameters, the depth-wise convolution(DW-Conv), and different branches of FIM.

Methods	PSNR	SSIM
concatenation	28.31	0.8306
WFM w/o splitting	28.32	0.8310
WFM	**28.36**	**0.8318**

Table 4: Quantitative comparisons of WFM. WFM without splitting means that the entire feature participates in WIT.

Effectiveness of the Wavelet Fusion Module (WFM). WFM is used in the Feature Recovery Stage to recombine the frequency-domain features. In WFM, we retain half of the features that do not participate in the WIT. Preserved features help stabilize feature recovery. Therefore, we designed two ablation experiments: In the first one, we simply replaced WFM with a concatenation operation. In the second one, we let all the features in the WFM participate in the WIT without splitting. Table 4 demonstrates that our proposed WFM achieves 0.05dB and 0.04dB improvements, respectively. Experiments prove that the designed WFM is efficient.

5 Conclusion

In this paper, we propose a self-refinement Transformer for FSR. It could conduct the coarse-grained self-attention and further compensate for the details by fine-grained self-attention on key regions. To consider the importance of channel information in Transformer, we also employ the FIM to achieve cross-spatial and cross-channel integration simultaneously. Furthermore, WFM is designed to modulate the shallow and deep feature maps in frequency domain for restoration. Extensive experiments demonstrate the effectiveness of the proposed method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62002283, 62311530046, U2003207, 61921004, U21B2048, 62102307, 61972016, 62071380), the Key Research and Development Program of Shaanxi (No.2021GXLH-Z-021) and the Fundamental Research Funds for the Central Universities.

References

[Bulat and Tzimiropoulos, 2018] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2018.

[Cao *et al.*, 2017] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 690–698, 2017.

[Cao *et al.*, 2021] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[Chen *et al.*, 2018] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018.

[Chen *et al.*, 2020] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K Wong. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231, 2020.

[Chu *et al.*, 2021] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Huang *et al.*, 2017] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017.

[Le *et al.*, 2012] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[Li *et al.*, 2021] Zhuangzi Li, Ge Li, Thomas Li, Shan Liu, and Wei Gao. Information-growth attention network for image super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 544–552, 2021.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[Ma *et al.*, 2020] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5569–5578, 2020.

[Maeda, 2022] Shunta Maeda. Image super-resolution with deep dictionary. In *European Conference on Computer Vision*, pages 464–480. Springer, 2022.

[Shi and Zhao, 2019] Jingang Shi and Guoying Zhao. Face hallucination via coarse-to-fine recursive kernel regression structure. *IEEE Transactions on Multimedia*, 21(9):2223–2236, 2019.

[Shi *et al.*, 2022] Jingang Shi, Yusi Wang, Songlin Dong, Xiaopeng Hong, Zitong Yu, Fei Wang, Changxin Wang, and Yihong Gong. IDPT: interconnected dual pyramid transformer for face super-resolution. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1306–1312, 2022.

[Tian *et al.*, 2022] Chunwei Tian, Yanning Zhang, Wangmeng Zuo, Chia-Wen Lin, David Zhang, and Yixuan Yuan. A heterogeneous group cnn for image super-resolution. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Tuzel *et al.*, 2016] Oncel Tuzel, Yuichi Taguchi, and John R Hershey. Global-local face upsampling network. *arXiv preprint arXiv:1603.07235*, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[Wu *et al.*, 2021] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *arXiv preprint arXiv:2106.12011*, 2021.

[Yuan *et al.*, 2021] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[Zhang *et al.*, 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[Zhang *et al.*, 2018a] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *Proceedings of the European conference on computer vision (ECCV)*, pages 183–198, 2018.

[Zhang *et al.*, 2018b] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2018c] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.