applied sciences

MDPI

Article

Why Not Both? An Attention-Guided Transformer with Pixel-Related Deconvolution Network for Face Super-Resolution

Zhe Zhang * and Chun Qi *

School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
* Correspondence: zzpong_xjtu@outlook.com (Z.Z.); qichun@mail.xjtu.edu.cn (C.Q.)

Abstract: Transformer-based encoder-decoder networks for face super-resolution (FSR) have achieved promising success in delivering stunningly clear and detailed facial images by capturing local and global dependencies. However, these methods have certain limitations. Specifically, the deconvolution in upsampling layers neglects the relationship between adjacent pixels, which is crucial in facial structure reconstruction. Additionally, raw feature maps are fed to the transformer blocks directly without mining their potential feature information, resulting in suboptimal face images. To circumvent these problems, we propose an attention-guided transformer with pixel-related deconvolution network for FSR. Firstly, we devise a novel Attention-Guided Transformer Module (AGTM), which is composed of an Attention-Guiding Block (AGB) and a Channel-wise Multi-head Transformer Block (CMTB). AGTM at the top of the encoder-decoder network (AGTM-T) promotes both local facial details and global facial structures, while AGTM at the bottleneck side (AGTM-B) optimizes the encoded features. Secondly, a Pixel-Related Deconvolution (PRD) layer is specially designed to establish direct relationships among adjacent pixels in the upsampling process. Lastly, we develop a Multi-scale Feature Fusion Module (MFFM) to fuse multi-scale features for better network flexibility and reconstruction results. Quantitative and qualitative experimental results on various datasets demonstrate that the proposed method outperforms other state-of-the-art FSR methods.

Keywords: face super-resolution; transformer; feature map enhancement; attention mechanism; deconvolutional layer

Citation: Zhang, Z.; Qi, C. Why Not Both? An Attention-Guided Transformer with Pixel-Related Deconvolution Network for Face Super-Resolution. *Appl. Sci.* **2024**, *14*, 3793. https://doi.org/10.3390/app14093793

Academic Editor: Andrea Prati

Received: 25 March 2024
Revised: 26 April 2024
Accepted: 28 April 2024
Published: 29 April 2024

1. Introduction

Face super-resolution (FSR), also referred to as face hallucination [1], is a technology that enhances the quality of low-resolution (LR) face images by transforming them into high-resolution (HR) ones. Typically, face images suffer lower spatial resolution due to limited imaging conditions and low-cost imaging equipment. This degradation affects the performance of most practical downstream applications, such as face recognition and face analysis. As a result, FSR has become a popular and essential scientific tool in the fields of computer vision and image processing [2].

Different from general image super-resolution, FSR is a technique that focuses on recovering crucial facial structures. Although these structures only occupy a small portion of the face, they are essential in distinguishing different faces and improving image quality. Baker and Kanade [1] proposed the first FSR method, which triggered the upsurge of traditional FSR methods. Afterward, various traditional techniques for FSR have been developed over time, which can mainly resort to the interpolation approach [3], Principal Component Analysis (PCA) [4], convex optimization [5], Bayesian approach [6], kernel regression [7], and manifold learning [8]. Nevertheless, traditional methods are limited in producing plausible facial images due to their shallow structure and representation abilities. Recently, FSR has made significant progress due to the advent of deep learning techniques [2]. Relying on the powerful deep convolution structures, various convolution neural networks (CNNs)-based FSR methods [9–13] have been developed to predict the

fine-grained facial details. However, due to the vanishing gradient problem, the actual receptive field of most CNN-based models is limited. This makes it challenging to model global dependency, resulting in blurry effects in the reconstructed face images. Devoting to capturing both local and global dependencies, transformer-based methods [14,15] have gained significant attention nowadays.

The efficacy of transformer-based methods in improving FSR performance is noteworthy. However, they still exhibit certain limitations that require attention. Transformer-based encoder-decoder networks comprise two major parts: an up/downsample module that connects adjacent-scale feature information and a transformer module that explores and enhances the corresponding-level features. We will discuss the limitations of these components separately below:

(1) Due to its relatively small network size, the up/downsample module has not received sufficient attention in FSR methods. However, it plays a more important role than the one applied in the general image super-resolution methods. This is because face images are highly structured with eyes, nose, and mouth in a specific location, which is also why some FSR methods require additional marking on the dataset. Nonetheless, as illustrated in Figure 1b, the inner feature maps generated by pixel deconvolutional or shuffle layers have no direct relationship since they are produced by independent convolutional kernels, which can result in significant differences between the values of adjacent pixels. Therefore, an up/downsample module that can build direct relationships among adjacent pixels is in high demand.

(2) The transformer module has significantly improved FSR performance. However, raw feature maps are fed to the transformer blocks directly without examining their potential feature information, limiting their performance. As shown in Figure 1g, raw features processed by the transformer block without guiding are not always detail-rich or even buried in gray, which restricts the following transformer blocks to only selecting a limited number of feature maps based on the self-attention heatmap (Figure 1e). On the contrary, applying a guiding block to guide the transformer block about essential facial components results in more correlated feature maps (Figure 1f). Such an approach is particularly beneficial for tackling the "one-to-many" FSR problem [16] and finally yielding superior outcomes (Figure 1h).

(3) Most previous research [15,17,18] favors improving the transformer module for the transformer-based encoder-decoder FSR approaches. However, matching a compatible strong up/downsample module and transformer module is crucial; otherwise, some of the system potential could be wasted on either side.

In this work, we aim to address all the limitations mentioned above and propose a novel attention-guided transformer with pixel-related deconvolution network for face super-resolution. The proposed method utilizes a multi-scale connected encoder-decoder architecture as the backbone. In encoder-decoder branches, we carefully design an Attention-Guided Transformer Module (AGTM), which is composed of an Attention Guiding Block (AGB) and a Channel-wise Multi-head Transformer Block (CMTB). AGB aims to guide the transformer block in learning about essential facial components. Different from previous transformer-based methods [15,17] which utilized the same transformer structure for different feature layers, AGB is further divided into two subdivision modules to adapt different levels of features: AGTM at the top of the encoder-decoder network (AGTM-T) which promotes both local facial details and global facial structures, while AGTM at the bottleneck side (AGTM-B) which optimizes the encoded low-level features. Noting the problem that the usual spatial-wise transformers are limited to position-specific windows and their partition strategy may potentially alter the structure of the facial image [19], the Channel-wise Multi-head Transformer Block (CMTB) is introduced to achieve an image-size receptive field by utilizing feature map channels. The AGB and CMTB are complementary and can simultaneously enhance local facial details and global facial structures. Furthermore, considering that face images are highly structured, we design a Pixel-Related Deconvolution (PRD) layer to establish direct relationships among adjacent pixels in the upsampling process for

better face structure preservation. Moreover, different from the pyramid network [13,20] that progressively reconstructs high-resolution face images, we have also developed a Multi-scale Feature Fusion Module (MFFM) to wisely fuse multi-scale features for better network flexibility and reconstruction results.

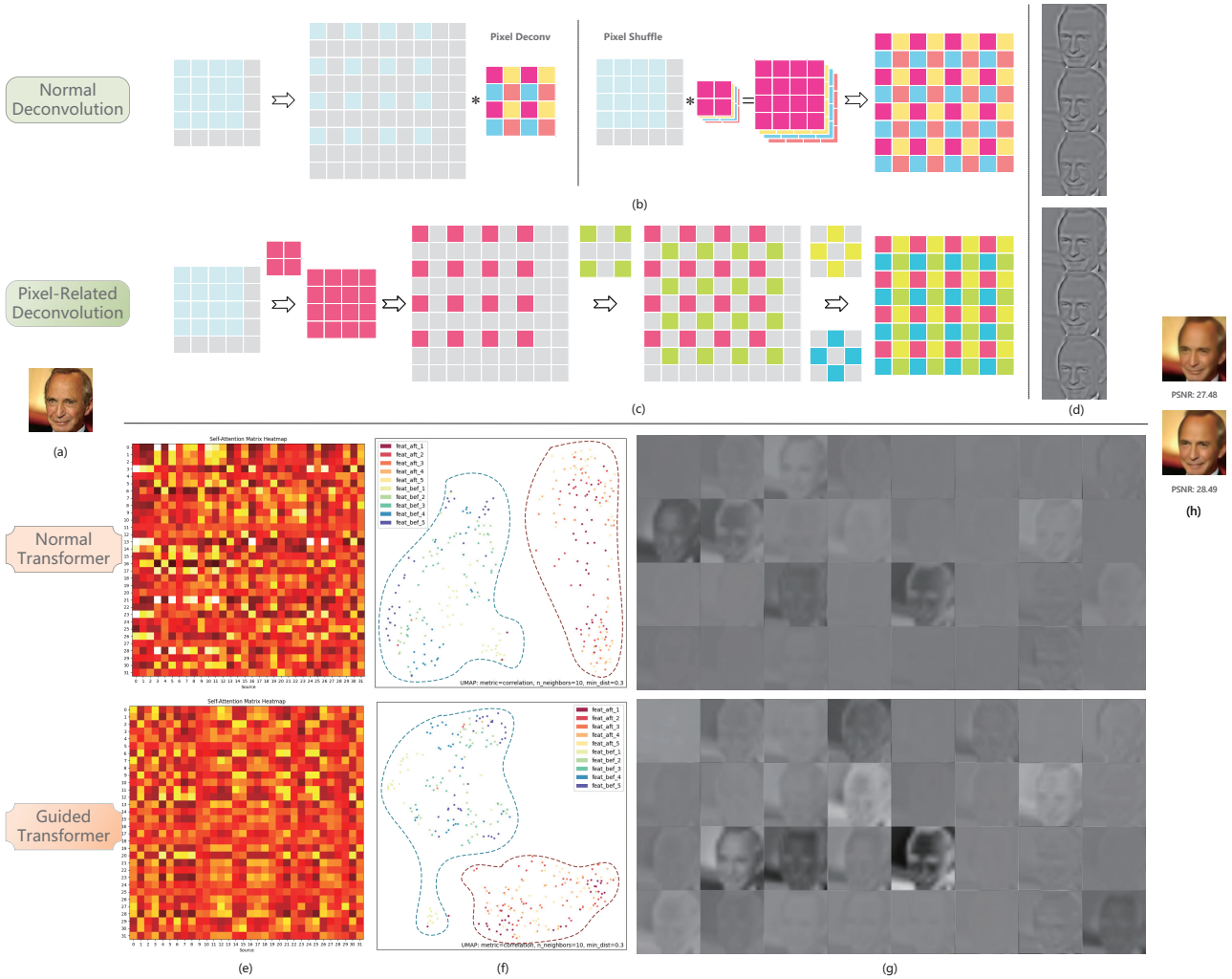

Figure 1. Visual analysis of the pixel-related deconvolution and the guiding block for transformer-based FSR methods: (**a**) is the input face image; (**b**) is the conventional pixel deconvolutional and pixel shuffle upsampling layer that neglects the relationship between adjacent pixels; (**c**) is the proposed pixel-related deconvolution that establishes direct relationships among adjacent pixels; (**d**) is the inner feature map outputs with corresponding upsampling methods; (**e**–**g**) are self-attention heatmaps, correlation maps [21] between input and output feature maps, and inner feature maps without and with guiding blocks, respectively (please note that five different images are tested in the correlation map instead of one for fair comparison); (**h**) is the output images (the top one is trained without pixel-related deconvolution and guiding blocks, while the bottom one is trained with them). Moreover, subfigure (**f**) represents the correlations between input and output feature maps. The more the output feature maps squeeze together, the more the input and output feature maps are suited and correlated, and the better it benefits the "one-to-many" FSR problem to obtain fine-grained FSR results. Please refer to Section 2.2 for more detailed information and related works.

To sum up, this work has four main contributions:

- We devise an attention-guided transformer with pixel-related deconvolution network for face super-resolution. To the best of our knowledge, neither the guiding block that mines potential inner feature map information nor the pixel-related deconvolution that

establishes direct relationships among adjacent pixels have been discussed before in the transformer-based FSR field. Results conducted on two frequently used benchmark datasets (i.e., CelebA [22] and Helen [23]) demonstrate that the proposed method surpasses other state-of-the-art methods both quantitatively and qualitatively.

- We carefully design an Attention-Guided Transformer Module (AGTM) to extract fine-grained features by enhancing the inner feature map relationship. Thanks to its powerful modeling ability, the proposed method can proficiently explore and utilize both local facial details and global facial structures.
- We develop a Pixel-Related Deconvolution (PRD) layer to establish direct relationships among adjacent pixels in the upsampling process for better face structure preservation and further strengthen the overall transformer-based FSR performance.
- We propose an elaborately designed Multi-scale Feature Fusion Module (MFFM) to fuse multi-scale features for better network flexibility and reconstruction results. The module is essential for the proposed method to acquire a wide range of features, which in turn improves the quality of the restoration performance.

2. Related Work

2.1. Face Super-Resolution

With the advent of deep learning techniques [2,24,25], deep convolution neural networks (CNNs) have been making remarkable advancements in enhancing the quality of face images. One of the first CNN-based FSR methods was proposed by Zhou et al. [26], demonstrating significant improvements in FSR performance compared to traditional FSR methods. To delve deeper into the facial information, Cao et al. [27] employed reinforcement learning to uncover the interdependent relationships among various facial components. Zhang et al. [9] introduced super-identity loss to help the network generate more accurately identified super-resolution face information. Unlike the FSR methods mentioned above that directly recover face images, Huang et al. [28] introduced the wavelet transform to project face images into wavelet spaces for capturing rich contextual information. Wang et al. [19] applied the Fourier transform to obtain an image-size receptive field to capture the global facial structures.

Motivated by the remarkable achievements of generative adversarial networks (GANs) [29], Yang et al. [30] developed a collaborative suppression and replenishment framework based on GANs. Yu [31] claimed that feature maps with additional facial attribute information could significantly reduce the ambiguity in FSR and combined these supplement residual images with GANs. Noticing that GAN-based methods are computationally intensive, PCA-SRGAN [32] leveraged Principal Component Analysis decomposition, while SPGAN [33] adopted a supervised pixel-wise loss approach to facilitate the GAN training process.

Noting that human faces are highly structured, several FSR methods utilized facial priors, including face landmarks and face parsing maps, to improve the reconstruction performance. Chen et al. [34] introduced facial parsing maps to guide the end-to-end FSR convolution network. Bulat et al. [35] combined a well-designed heatmap loss with GANs to ensure the face structure consistency between high-resolution (HR) and super-resolved (SR) face images. Hu et al. [36] proposed 3D facial priors for capturing sharp facial structures in face images with large pose variations. Due to the difficulty of directly estimating the prior from degraded LR facial images, DIC [37] developed an iterative process where FSR and prior estimation were performed repeatedly to enhance FSR performance. While the FSR models with facial priors have achieved promising outcomes, there is a problem that has not received sufficient attention. Specifically, the deconvolution or shuffle layers in the upsampling process neglect the relationship between adjacent pixels, which can break the highly structured face image prior, limiting the FSR performance.

In recent years, the attention mechanism has emerged as a prominent approach in computer vision tasks [38–41]. For instance, Chen et al. [10] introduced a face attention unit to capture facial structures, while Lu et al. [11] designed an external-internal split

attention group to reconstruct more detailed facial images. Moreover, transformers, which have already demonstrated their effectiveness in various fields, are also widely applied in computer vision tasks such as image recognition [42,43] and restoration [15,17,44]. The self-attention mechanism, which is the core of transformers, is promised to capture both long- and short-range correlations between words or pixels [45]. However, feeding feature maps directly into transformers without guidance will result in the loss of some fine-grained details, limiting facial structure reconstruction performance. Therefore, it is essential to design an effective guiding block to identify the crucial facial components for the transformers.

2.2. Feature Maps and Feature Spaces

Feature maps and feature spaces share some similarities: CNN-based FSR methods utilize convolution layers to project LR images into inner "feature maps" and then into HR ones. Meanwhile, traditional manifold learning-based FSR methods project LR images into "feature spaces" and then into HR ones, assuming that both LR and HR spaces share the same local geometry [46]. Many traditional FSR methods have been introduced based on the manifold learning assumption [47–49], aiming to enhance the LR and HR space relationship. However, the manifold learning technique has gained less attention with the rise of CNNs because complicated CNN structures are challenging to deliberate. To bridge the gap between CNNs and manifold learning, several deep learning-based FSR methods have been proposed. Yang et al. [50] introduced a manifold localized deep external compensation (MALDEC) network that provides accurate localization and mapping to the HR manifold by referring to the big data online. Menon et al. [51] searched the HR manifold space to find images that match the original LR image and then applied a generative model for image reconstruction by feeding it the downscaling loss. Chen et al. [52] introduced an LR and HR space homogenization projection to formulate FSR in a multi-stage framework. Guo et al. [16] developed a closed-loop dual regression network (DRN) with an additional constraint, claiming that limiting mapping spaces would be advantageous for image super-resolution. The methods mentioned above tried to combine the CNN-based method with the manifold learning technique. However, they overlook the vital role of inner feature maps, which restricts the image super-resolution performance. Therefore, determining how to effectively handling inner feature maps, such as a guiding block that increases feature map correlations, is vital for a high-quality image reconstruction process.

3. Proposed Method

Considering the vital role of the guiding blocks in identifying the essential facial components and aiming to establish direct relationships among adjacent pixels for better face structure preservation, we develop a novel attention-guided transformer with a pixel-related deconvolution network for face image super-resolution. This is the first study for the transformer-based FSR field to not only mine potential inner feature map information but also establish direct relationships among adjacent pixels in reconstructing highly structured face images.

To better elaborate on the proposed method, we divide it into four subsections. In the first section, we provide an overview of the architecture of the proposed method. Then, we delve into the main component of the proposed method, the Attention-Guided Transformer Module (AGTM), which consists of an Attention Guiding Block (AGB) and a Channel-wise Multi-head Transformer Block (CMTB). The AGB and CMTB are complementary and can simultaneously enhance local facial details and global facial structures. Afterward, we introduce the Multi-scale Feature Fusion Module (MFFM), which integrates features from all layers to improve network flexibility and restoration performance. Finally, we introduce the Pixel-Related Up/Downsample Module (PRUM/PRDM) that establishes direct relationships among adjacent pixels for better face structure preservation and further strengthens the overall image reconstruction performance.

3.1. Overview

The proposed method, illustrated in Figure 2, is a symmetrical hierarchical network consisting of three stages: encoding, bottleneck, and decoding. The encoding stage aims to extract and enhance both local facial details and global facial structures. Meanwhile, the bottleneck stage is intended to optimize the encoded low-level features. Finally, the decoding stage is introduced to facilitate multi-scale feature fusion and image reconstruction. To simplify the description, we use the notations $\boldsymbol{I}_{LR}$, $\boldsymbol{I}_{SR}$, and $\boldsymbol{I}_{HR}$ to represent the low-resolution (LR) images, the super-resolved (SR) images, and the ground-truth high-resolution (HR) images, respectively.

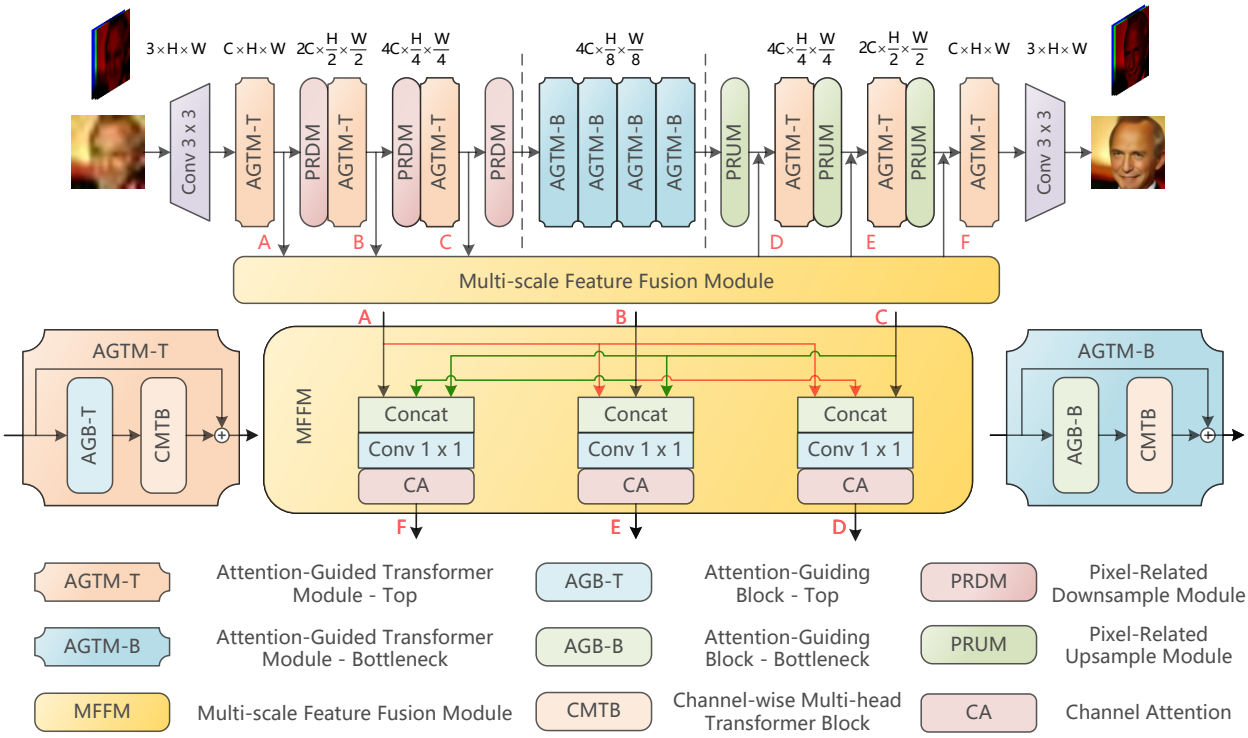

Figure 2. The structure of the proposed attention-guided transformer with pixel-related deconvolution network. It is a symmetrical hierarchical network containing three stages: encoding, bottleneck, and decoding. The goal of the encoding stage is to extract and enhance both local facial details and global facial structures. Meanwhile, the bottleneck stage is intended to optimize the encoded low-level features. Finally, the decoding stage is introduced to facilitate multi-scale feature fusion and image reconstruction. Here, the red and green arrows in MFFM represent downsampling and upsampling, respectively.

(1) Encoding Stage: The goal of the encoding stage is to extract and enhance both local facial details and global facial structures. To begin with, the face images traverse a 3×3 convolution layer to extract their low-level features. Since the output channel number should exceed the input ones while an excessive number of output channels would significantly increase the computational complexity [53], we suggest using 32 output channels for optimal performance. Afterward, the extracted shallow features are passed through three encoding stages. Each stage comprises an Attention-Guided Transformer Module—Top (AGTM-T) and a Pixel-Related Downsample Module (PRDM). The AGTM-T comprises an Attention Guiding Block—Top (AGB-T) and a Channel-wise Multi-head Transformer Block (CMTB). All the blocks and modules mentioned above will be discussed in the following subsection. Moreover, it is worth noting that the channel of the input feature maps doubles, and the size of the input feature maps halves after each encoding stage.

(2) Bottleneck Stage: In the bottleneck stage, there are a large number of encoded feature maps, but each one is relatively small in size compared to those in the encoding

stage. To better use these features in the decoding stage, we introduce the Attention-Guided Transformer Module—Bottleneck (AGTM-B). Unlike AGTM-T in the encoding stage, the guiding blocks in AGTM-B aim to further enhance the low-level encoded features. By applying AGTM-Bs, the model can continuously strengthen different facial features and focus on a broader range of facial structures.

(3) Decoding Stage: The decoding stage of the proposed method aims to reconstruct high-quality face images by utilizing previously extracted and refined multi-scale features. In this stage, low-level features are initially fed into the Pixel-Related Upsample Module (PRUM). The module halves the feature map channel and doubles the feature map size, which is the opposite of the PRDM in the encoding stage. After this, the upsampled features are combined with features from other scales using the Multi-scale Feature Fusion Module (MFFM) to enhance network flexibility and achieve better restoration performance. The well-combined features are then fed to AGTM-T for further refinement of image details. Finally, a convolutional layer with a size of 3×3 is utilized to convert the learned feature maps to the output face image $\boldsymbol{I}_{Out}$. The final SR face image output $\boldsymbol{I}_{SR}$ is obtained by adding the LR face image $\boldsymbol{I}_{LR}$, which has been upsampled to the same size as the HR image through bicubic interpolation, to the output face image $\boldsymbol{I}_{Out}$.

Moreover, to optimize the performance of FSR, the proposed model is supervised by minimizing the following pixel-level loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{I}_{SR}^{i} - \boldsymbol{I}_{HR}^{i}\right\|_{1} \tag{1}$$

where N denotes the number of training images. $\boldsymbol{I}_{SR}^{i}$ and $\boldsymbol{I}_{HR}^{i}$ are the i-th SR and ground-truth HR face image in the training dataset, respectively.

3.2. Attention-Guided Transformer Module (AGTM)

As the pivotal component of the proposed method, AGTM comprises two blocks: the Attention Guiding Block (AGB) and the Channel-wise Multi-head Transformer Block (CMTB). To ensure the feature extraction and enhancement quality on multi-scale features, the AGB has been bifurcated into two separate types: the Attention Guiding Block—Top (AGB-T) in the encoding/decoding stage and the Attention Guiding Block—Bottleneck (AGB-B) in the bottleneck stage. AGTM at the top of the encoder-decoder network (AGTM-T) promotes both local facial details and global facial structures, while AGTM at the bottleneck side (AGTM-B) optimizes the encoded low-level features. Furthermore, noticing that the usual spatial-wise transformers are limited to position-specific windows and their partition strategy may potentially alter the structure of facial images, the Channel-wise Multi-head Transformer Block (CMTB) is introduced here to achieve an image-size receptive field by utilizing feature map channels. The AGB and CMTB are complementary and can facilitate the simultaneous promotion of both local facial details and global facial structures.

3.2.1. Attention Guiding Block—Top (AGB-T)

AGB-T, which aims to locate and guide both local and global facial structures for the following transformer module, is illustrated in Figure 3a. It can be roughly divided into three parts: the Feature Distillation Network (FDN), the Hourglass Block, and the Channel Attention (CA) network. FDN aims to distill feature information from multiple levels of respective fields within the input feature maps. Firstly, a 3×3 convolutional layer is applied to the input feature maps to halve its number and select its internal principal components; then another 3×3 convolutional layer is used to restore the major information of the input feature maps by doubling the channel number. Then again, the original and the processed input feature maps are concatenated and sent to a full connection layer followed by a 3×3 convolutional layer to fully utilize the hierarchical features. After that, a CA network is applied to highlight the critical feature map channels, followed by a 3×3 convolutional layer to refine the distilled feature maps. Finally, a residual learning mechanism is applied

to avoid the gradient vanishing problem. Following the distillation process of the FDN, the Hourglass Block [54], which has demonstrated its efficacy in generating spatial attention maps [13], is employed to capture landmark features of the human face, such as the eyes, nose, and mouth. Once the feature information is appropriately processed, the CA network [38] is utilized to select and emphasize feature map channels that contain a higher number of features. Thanks to the well-designed structure that wisely distills internal principal features and mutualizes spatial and channel attention, the proposed AGB can successfully guide the following transformer block to capture the essential part of the face images for better reconstruction results.

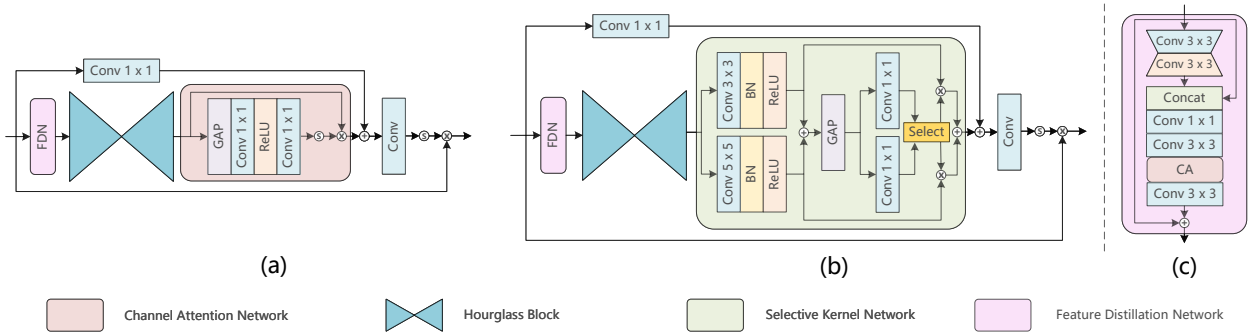

Figure 3. Architectures of the Attention Guiding Block (AGB): (**a**) is the Attention Guiding Block—Top (AGB-T); (**b**) is the Attention Guiding Block—Bottleneck (AGB-B); (**c**) is the Feature Distillation Network (FDN). Here, Ⓢ denotes the sigmoid function.

After all the above, the final attention map for CMTB is generated by applying a 3×3 convolutional layer followed by a sigmoid function. Then, input feature maps are element-wise multiplied with the attention map and fed to the following transformer block with better extracted spatial features and promoted channel information. Moreover, a residual connection with a full connection layer is also applied between the input and the CA network output to stabilize the training process.

3.2.2. Attention Guiding Block—Bottleneck (AGB-B)

Different from the above AGB-T, AGB-B in the bottleneck stage is designed to target and guide the low-level encoded features. The channel number of feature maps in the bottleneck stage is relatively large, but the size of each feature is relatively small compared to those in the encoding stage. Therefore, it is crucial to implement a dynamic selection mechanism that adaptively enables each neuron to adjust its receptive field size. Here, we introduce the selective kernel (SK) network [55] to the AGB-B, which is shown in Figure 3b. In the SK network, the input feature maps first pass through two convolution layers with different respective fields, followed by a batch normalization layer and a ReLU layer. The upper and lower outputs here are noted as $\mathbf{U}$ and $\mathbf{V}$, respectively. Then, these output feature maps are elementwise summed and traverse a global average pool (GAP) to generate channel-wise statistics with different respective fields. After that, the inner feature maps are sent through two full connection layers to enable the guidance for the adaptive selections. Lastly, a soft attention layer is applied across different channels to extract information from different respective fields selectively. Here, we use the notations $F(\mathbf{U})$ and $G(\mathbf{V}) \in \mathbb{R}^{C\times 1}$ to represent the upper and lower input of the Select layer in Figure 3b, where $F(\cdot)$ and $G(\cdot)$ denote the previous inner feature map process, and C denote the number of channels of the inner feature map, the output weight is:

$$\mathbf{w}_c^{upper} = \frac{e^{F_c(\mathbf{U})}}{e^{F_c(\mathbf{U})} + e^{G_c(\mathbf{V})}}, \quad \mathbf{w}_c^{lower} = \frac{e^{G_c(\mathbf{V})}}{e^{F_c(\mathbf{U})} + e^{G_c(\mathbf{V})}} \tag{2}$$

where c in $\mathbf{w}_c^{upper}$ denotes the c-th element of the $\mathbf{w}^{upper}$, likewise $\mathbf{w}_c^{lower}$, $F_c(\mathbf{U})$ and $G_c(\mathbf{V})$. The final attention maps of the SK network are obtained through attention weights on inner feature maps from various respective fields:

$$\mathbf{A}_c = \mathbf{w}_c^{upper} \times \mathbf{U}_c + \mathbf{w}_c^{lower} \times \mathbf{V}_c, \quad \mathbf{w}_c^{upper} + \mathbf{w}_c^{lower} = 1 \tag{3}$$

where $\mathbf{A} = [\ \mathbf{A}_1, \ \mathbf{A}_2, \ \ldots, \ \mathbf{A}_C\ ]$ denotes the output attention maps, $\mathbf{A}_c \in \mathbb{R}^{H \times W}$. H and W denote the height and width of the feature maps, respectively.

3.2.3. Channel-Wise Multi-Head Transformer Block (CMTB)

Following the pre-processing of the inner feature maps with guiding blocks, there still remains a demand for effectively aggregating previous feature data across various channels to facilitate high-quality face image restoration. However, the usual spatial-wise transformers are limited to position-specific windows, and their partition strategy may potentially alter the structure of the facial image [19]. To address this limitation, we introduce the Channel-wise Multi-head Transformer Block (CMTB)—a novel approach capable of achieving image-size receptive fields based on channels rather than position-specific windows. Furthermore, CMTB is more computation-friendly, rendering it a suitable match for the previous guiding blocks. As depicted in Figure 4, CMTB comprises two key components: the Channel-wise Multi-head Self-attention Network (CMSN) and the Gated-Dconv Feed-Forward Network (GDFN). While CMSN serves as the primary component, GDFN aims to encode information from spatially neighboring pixel positions to enable effective learning of local image structures.

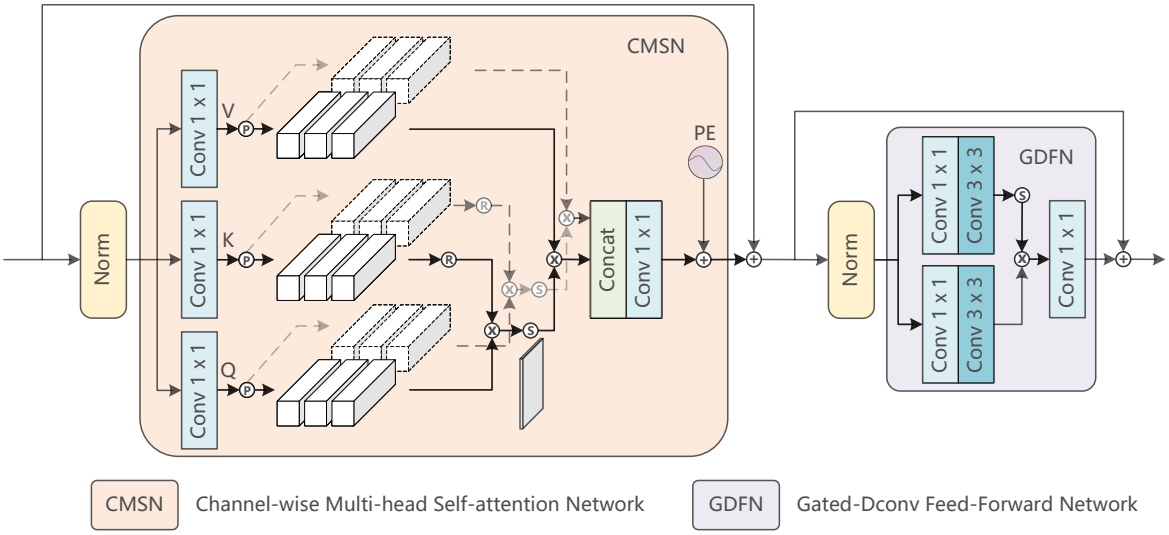

Figure 4. Architectures of the Channel-wise Multi-head Transformer Block (CMTB). Here, Ⓢ, Ⓡ, and Ⓟ denotes the sigmoid function, reshape, and split, respectively. PE denotes the position embedding generator.

CMTB has been proposed to achieve image-size receptive fields based on different channels of feature maps rather than position-specific windows. Suppose feature maps $\mathbf{X}_{in} \in \mathbb{R}^{H \times W \times C}$ as the input of the CMSN, which is reshaped into tokens $\mathbf{X} \in \mathbb{R}^{HW \times C}$ based on channels. Here, H, W, and C denote the height, width, and channel numbers of the feature maps, respectively. Then $\mathbf{X}$ is linearly projected to obtain three different matrices: *query* $\mathbf{Q} \in \mathbb{R}^{HW \times C}$, *key* $\mathbf{K} \in \mathbb{R}^{HW \times C}$, and *value* $\mathbf{V} \in \mathbb{R}^{HW \times C}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}} \tag{4}$$

where $\mathbf{W}^{\mathbf{Q}}$, $\mathbf{W}^{\mathbf{K}}$, and $\mathbf{W}^{\mathbf{V}} \in \mathbb{R}^{C \times C}$ are learnable parameters; *biases* are omitted here for simplification. Afterwards, $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are split into N *heads* along the channel dimension:

$\mathbf{Q} = [\ \mathbf{Q}_1, \ \ldots, \ \mathbf{Q}_N\ ], \mathbf{K} = [\ \mathbf{K}_1, \ \ldots, \ \mathbf{K}_N\ ], \mathbf{V} = [\ \mathbf{V}_1, \ \ldots, \ \mathbf{V}_N\ ]$, where the dimension of each head is $d = C/N$. Therefore, the self-attention matrix for $head_i$ is:

$$\mathbf{A}_i = softmax(\sigma_i \mathbf{K}_i^T \mathbf{Q}_i), \quad head_i = \mathbf{V}_i \mathbf{A}_i \tag{5}$$

where $\mathbf{K}_i^T$ denotes the transposed matrix of $\mathbf{K}_i$. By implementing the reshape strategy, the size of the generated attention maps will be $d \times d$ instead of $HW \times HW$, which greatly reduces the computational complexity. Moreover, a learnable parameter $\sigma_i \in \mathbb{R}^1$ is introduced to further improve the flexibility of the network. Subsequently, N *heads* outputs are concatenated and fed to a full connection layer. The resulting attention matrix is then added with the embedding values from the position embedding generator:

$$CMSN(\mathbf{X}) = (\overset{N}{\underset{i=1}{Concat}}(head_i))\mathbf{W} + f_p(\mathbf{V}) \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{C \times C}$ are learnable parameters. $f_p(\cdot)$ represents the position embedding generator, which is designed to encode the position information from various channel dimensions. It contains a 3×3 depth-wise convolution layer with a stride of 1 followed by a GELU layer [56] and another 3×3 depth-wise convolution layer with a stride of 1. Finally, the output feature maps $\mathbf{X}_{out} \in \mathbb{R}^{H \times W \times C}$ can be calculated by reshaping the result of Equation (6).

Additionally, we introduce GDFN [57] to effectively learn local image structures by encoding information from spatially neighboring pixel positions. Given feature maps $\mathbf{X}_{in} \in \mathbb{R}^{H \times W \times C}$ as the input of the GDFN, the output $\mathbf{X}_{out} \in \mathbb{R}^{H \times W \times C}$ can be obtained by:

$$\hat{\mathbf{X}} = \mathbf{H}_{dconv}^{3\times3}(\mathbf{H}_{fc}(\mathbf{X}_{in})), \quad \mathbf{X}_{out} = \mathbf{H}_{fc}(\hat{\mathbf{X}} \cdot \sigma(\hat{\mathbf{X}})) \tag{7}$$

where $\mathbf{H}_{dconv}^{3\times3}(\cdot)$ and $\mathbf{H}_{fc}(\cdot)$ denote the 3×3 depth-wise convolution layer and the full connection layer, respectively. $\sigma(\cdot)$ represents the GELU non-linearity.

The AGB and CMTB are two components that work in tandem to enhance facial features and strengthen the inner feature map relationship. The AGB is responsible for extracting and guiding the key features from the inner feature maps, while the CMTB aggregates and refines the previously extracted feature information. By leveraging both these blocks, the AGTM can concurrently enhance local facial details and global facial structures, making it a promising solution for face image reconstruction tasks.

3.3. Multi-Scale Feature Fusion Module (MFFM)

The importance of multi-scale feature information in the image reconstruction process has been proven by the successive pyramid super-resolution networks [13,20]. However, the pyramid methods mentioned above reconstruct high-resolution images only from adjacent layers, limiting the FSR performance. To further utilize the multi-scale feature information and enable the network with better feature representation capabilities, we introduce the Multi-scale Feature Fusion Module (MFFM), which is shown on the bottom side of Figure 1.

The first step of the MFFM is to unify the size of multi-scale feature maps. Noticing that the magnification scale of adjacent layers in the proposed method is always 2, we introduce a 3×3 convolution layer with a stride of 2 and a 6×6 transposed convolution layer with a stride and padding of 2 for $/2$ down-scale and $\times 2$ up-scale, processes respectively. Please note that the $/M$ down-scale and $\times N$ up-scale mentioned here represent the M times downsampling and N times upsampling, respectively. Furthermore, for a larger magnification scale like $/4$ or $\times 4$, double 3×3 convolution or 6×6 transposed convolution layers are applied, etc. Considering that the MFFM is trained for residual compensation for multi-scale feature maps, a single convolution or transposed convolution layer is applied here to modify the feature map size instead of the PRUM/PRDM for simplicity. After resizing the multi-scale feature maps to a uniform size, they are concatenated to undergo a

full connection layer and then fed to the CA network to highlight the essential channels. Finally, the well-handled multi-scale features are integrated with the target feature map layer from the encoding stage.

Pixel-Related Up/Downsample Module (PRUM/PRDM)

The Pixel-Related Up/Downsample Module (PRUM/PRDM) aims to establish direct relationships among adjacent pixels for better face structure preservation and further strengthen the overall image reconstruction performance, whose structure is illustrated in Figure 5.

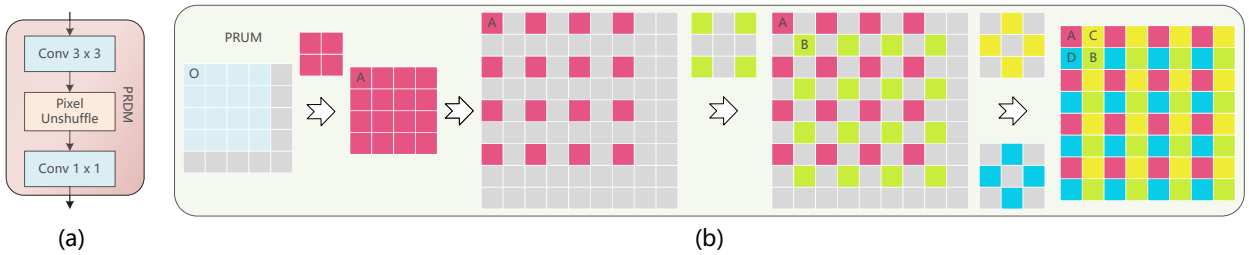

Figure 5. Architectures of the Pixel-Related Up/Downsample Module (PRUM/PRDM). (**a**) and (**b**) are the Pixel-Related Downsample Module and Pixel-Related Upsample Module, respectively.

The PRDM consists of three layers. Firstly, a 3×3 convolutional layer with a padding of 1 and a stride of 2 is applied to simulate the image degradation process. The kernel size here is set to 3×3 due to the fact that multiple 3×3 kernels from the multi-scale encoding-decoding network can simulate large kernel sizes while reducing computational complexity. After that, a Pixel Unshuffle layer instead of a convolutional layer is utilized to preserve as much original pixel information as possible. Finally, a full connection layer is applied to extract the vital image features. After the PRDM, the feature map channel doubles while the feature map size halves.

To establish direct relationships among adjacent pixels in the upsampling process, we introduce the PRUM, whose central part is the Pixel-Related Deconvolution (PRD) layer. The structure of the module is illustrated in Figure 5b. As we all know, pixel-Os obtained from */2* down-scale are not always their original values due to the complicated degradation process. Therefore, we first applied a 2×2 convolutional layer to obtain the real pixel-As. Then, the diagonal pixel-Bs are calculated by their corresponding As with another 3×3 convolutional layer with a padding of 1. After the above processing, the Cs and Ds are surrounded by known pixels and can be obtained by two separate 3×3 convolutional layers. Based on the relative pixel position, the calculated pixel-As, Bs, Cs, and Ds are highly related and structured, which assists the PRUM in establishing direct relationships among adjacent pixels. After the PRUM, the feature map channel halves while the feature map size doubles, which is the exact opposite of the PRDM in the encoding stage.

4. Experiments

4.1. Dataset and Metrics

The proposed model is trained on the CelebA dataset [22], and its performance is evaluated on both CelebA and Helen datasets, as well as on real face images. During the data preprocessing phase, we crop the images to a size of 128×128 based on their center point and treat them as the ground truth. After that, we obtain 16×16 LR face images from the ground truth using a */8* down-scale bicubic operation. It is worth noting that no additional facial landmarking is required on the datasets to train the model. We trained the model on 18,000 face images from the CelebA dataset and evaluated its performance on 1000 faces from the same dataset, along with 50 faces from the Helen dataset. Additionally, we directly applied the same model trained on CelebA to the Helen datasets and real face images to evaluate the flexibility of the model.

To evaluate the quality of the SR results, three image quality assessment metrics are introduced: Peak Signal-to-Noise Ratio (PSNR) [58], Structural Similarity (SSIM) [59], and Learned Perceptual Image Patch Similarity (LPIPS) [60].

4.2. Implementation Details

All experiments are conducted using PyTorch [61] on an NVIDIA GeForce RTX 4090 24 GB graphics card. The proposed model is optimized using Adam with $\beta 1 = 0.9$, $\beta 2 = 0.99$, and a learning rate of 2×10^{-4}.

4.3. Ablation Studies

To assess the effectiveness of individual model modules, we conducted a series of ablation studies on the CelebA test sets for ×8 SR.

(1) Study on AGTM-T: AGTM-T is a module that combines an AGB-T and a CMTB to extract and promote both local facial details and global facial structures. This module marks the first attempt to explore the potential of inner feature map information with a guiding block to reconstruct plausible face images in the transformer-based FSR area. To test its effectiveness, we design three test models by removing different module parts, of which the results are shown in Table 1.

Table 1. Ablation study of the components in the proposed AGTM-T.

AGB-T	CMTB	PSNR↑	SSIM↑	LPIPS↓
×	×	27.45	0.7870	0.2026
×	✓	27.68	0.7925	0.1804
✓	×	27.71	0.7931	0.1795
✓	✓	**27.77**	**0.7941**	**0.1766**

The symbols ✓and × indicate whether or not a corresponding block is included, respectively. The **best** results are emphasized with **bold**. ↑ and ↓ indicate whether higher or lower evaluation matrix values correspond to better image quality, respectively.

From the table, we can observe that:

(a) The performance decreases dramatically when the AGTM-T module is completely removed. The proposed model structure will be considerably shallower without the AGTM-T, which further results in the difficulty in refining input features. Moreover, the multi-scale feature fusion operation module (MFFM), which is complemented with the AGTM-T, will also be greatly affected.

(b) The AGTM-T with a single component performs better compared with the no-component one mentioned above. This demonstrates that both AGB-T and CMTB benefit the learning ability of the proposed model. However, the AGTM-T with AGB-T only lost the guiding target, while the AGTM-T with CMTB only cannot focus on crucial feature parts, limiting its performance.

(c) The carefully designed components AGB-T and CMTB ensure that the AGTM-T achieves the best performance in all evaluation matrices. This proves that the AGB-T and CMTB are complementary and can simultaneously enhance local facial details and global facial structures.

(2) Study on AGTM-B: AGTM-B, which contains an AGB-B and a CMTB, aims to enhance the low-level encoded features. Similar experiments as the above section are conducted and the results are shown in Table 2. We have arrived at comparable observations and conclusions in the preceding AGTM-T section. However, we do notice that the performance of the model without AGTM-B is better than that without AGTM-T. This is because AGTM-T and MFFM are more complementary when compared to the relationship between AGTM-B and MFFM. The removal of AGTM-T will result in a further decline in the performance of the proposed method.

Furthermore, we also conduct an evaluation of the model on the number of AGTM-Bs, whose results are shown in Table 3. It can be observed that the performance of the model

is poor without any AGTM-B, suggesting that AGTM-B plays a crucial role in the model. Meanwhile, we also notice that the performance improves when the number of AGTM-Bs increases within a specific range. However, when the number of AGTM-Bs exceeds 4, the change rate of the evaluation matrix slows down, and the performance even decreases slightly. Therefore, to maintain a good balance between model size and performance, we set the number of AGTM-Bs to 4.

Table 2. Ablation study of the components in the proposed AGTM-B.

AGB-B	CMTB	PSNR↑	SSIM↑	LPIPS↓
×	×	27.63	0.7911	0.1896
×	✓	27.73	0.7935	0.1788
✓	×	27.74	0.7938	0.1779
✓	✓	**27.77**	**0.7941**	**0.1766**

The symbols ✓and × indicate whether or not a corresponding block is included, respectively. The **best** results are emphasized with **bold**. ↑ and ↓ indicate whether higher or lower evaluation matrix values correspond to better image quality, respectively.

Table 3. Performance comparisons of different AGTM-B numbers in the proposed method.

AGTM-B Numbers	PSNR↑	SSIM↑	LPIPS↓
0	27.63	0.7911	0.1896
2	27.71	0.7933	0.1797
4	**27.77**	**0.7941**	**0.1766**
6	27.76	0.7938	0.1775

The **best** results are emphasized with **bold**. ↑ and ↓ indicate whether higher or lower evaluation matrix values correspond to better image quality, respectively.

(3) Study on MFFM: MFFM is specially designed to integrate features from all layers to improve network flexibility and restoration performance. In this part, we create three different multi-scale feature fusion models to demonstrate the effectiveness of the MFFM, whose results are shown in Table 4.

Table 4. Performance comparisons of different approaches of the multi-scale feature fusion process.

Approaches	PSNR↑	SSIM↑	LPIPS↓
Not Applied	27.67	0.7917	0.1846
Only Add	27.72	0.7926	0.1809
Only Concat	27.73	0.7930	0.1800
Our MFFM	**27.77**	**0.7941**	**0.1766**

The **best** results are emphasized with **bold**. ↑ and ↓ indicate whether higher or lower evaluation matrix values correspond to better image quality, respectively.

It can observed from the table that: (a) The experiment demonstrates the importance of incorporating multi-scale features in the image reconstruction process since the model without multi-scale feature fusion performs the worst. (b) Using an addition or concatenation layer to fuse multi-scale features has proven beneficial. However, it is imperative to note that these techniques are inadequate for the complex multi-scale feature fusion process. (c) The model with the carefully designed MFFM achieves the best performance regarding PSNR, SSIM, and LPIPS. This proves that a suitable feature fusion strategy like MFFM can benefit the image reconstruction process.

(4) Study on PRUM/PRDM: The PRUM/PRDM aims to establish direct relationships among adjacent pixels for better face structure preservation and further strengthen the overall image reconstruction performance. It is the first attempt to establish direct relationships among adjacent pixels in reconstructing highly structured face images in the transformer-based FSR area. In this part, we compare its performance with the usual deconvolution layers, of which results are shown in Table 5.

Table 5. Performance comparisons of different up/downsample approaches.

Approaches	PSNR↑	SSIM↑	LPIPS↓
Pixel Deconv	27.68	0.7924	0.1836
Pixel Shuffle	27.69	0.7930	0.1816
Ours	**27.77**	**0.7941**	**0.1766**

The **best** results are emphasized with **bold**. ↑ and ↓ indicate whether higher or lower evaluation matrix values correspond to better image quality, respectively.

It can be observed that the pixel deconvolutional and pixel shuffle layers can obtain barely satisfactory reconstruction results. This is because the inner feature maps produced by these layers have no direct pixel relationship. At the same time, our proposed PRUM/PRDM achieves better reconstruction results due to its ability to preserve face structure based on relative pixel positions.

4.4. Comparison with the State-of-the-Arts

To demonstrate the effectiveness of our proposed method, we conduct a comparison with several state-of-the-art methods. These include two GAN-based methods (SRResNet [29] and RCAN [38]), three attention-based methods (SPARNet [10], SISN [11], and IGAN [39]), and two transformer-based methods (SwinIR [17] and Uformer [15]). We evaluate these methods on the CelebA and Helen datasets, along with the real face images. In addition, we apply bicubic interpolation as the baseline for comparison. All models are trained on the same CelebA dataset to ensure a fair comparison. The quantitative results are tabulated in Table 6.

Table 6. Quantitative comparisons for ×8 SR on the CelebA and Helen test sets.

Methods	*CelebA*			*Helen*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Bicubic	23.44	0.6180	0.5900	23.79	0.6739	0.5254
SRResNet [29]	26.08	0.7502	0.2131	25.47	0.7828	0.2308
IGAN [39]	26.99	0.7801	0.2201	26.37	0.7996	0.2245
RCAN [38]	26.99	0.7796	0.2249	26.39	0.7965	0.2359
SISN [11]	26.85	0.7738	0.2337	26.33	0.7974	0.2322
SPARNet [10]	26.95	0.7794	0.2211	26.38	0.7953	0.2314
SwinIR [17]	27.15	0.7850	0.2162	26.48	0.7917	0.2413
Uformer [15]	27.33	0.7884	0.2040	26.67	0.8009	0.2063
Ours	**27.77**	**0.7941**	**0.1766**	**27.16**	**0.8117**	**0.1890**

The **best** results are emphasized with **bold**. ↑ and ↓ indicate whether higher or lower evaluation matrix values correspond to better image quality, respectively.

(1) Comparison on CelebA dataset: Quantitative comparisons of the proposed method with other existing methods on the CelebA dataset are presented in Table 6. As per the table, our proposed method outperforms other competitive methods in terms of PSNR, SSIM, and LPIPS, which implies that our method has the advantage of recovering realistic face details. We have also provided some test images from the CelebA dataset for visual comparisons, which are shown in Figure 6. Benefiting from the guiding blocks pointing to the key features and the pixel-related upsample layer preserving face structures, the proposed method can generate more precise nose contours and eye details while avoiding creating unpleasant artifacts compared with other state-of-the-art methods.

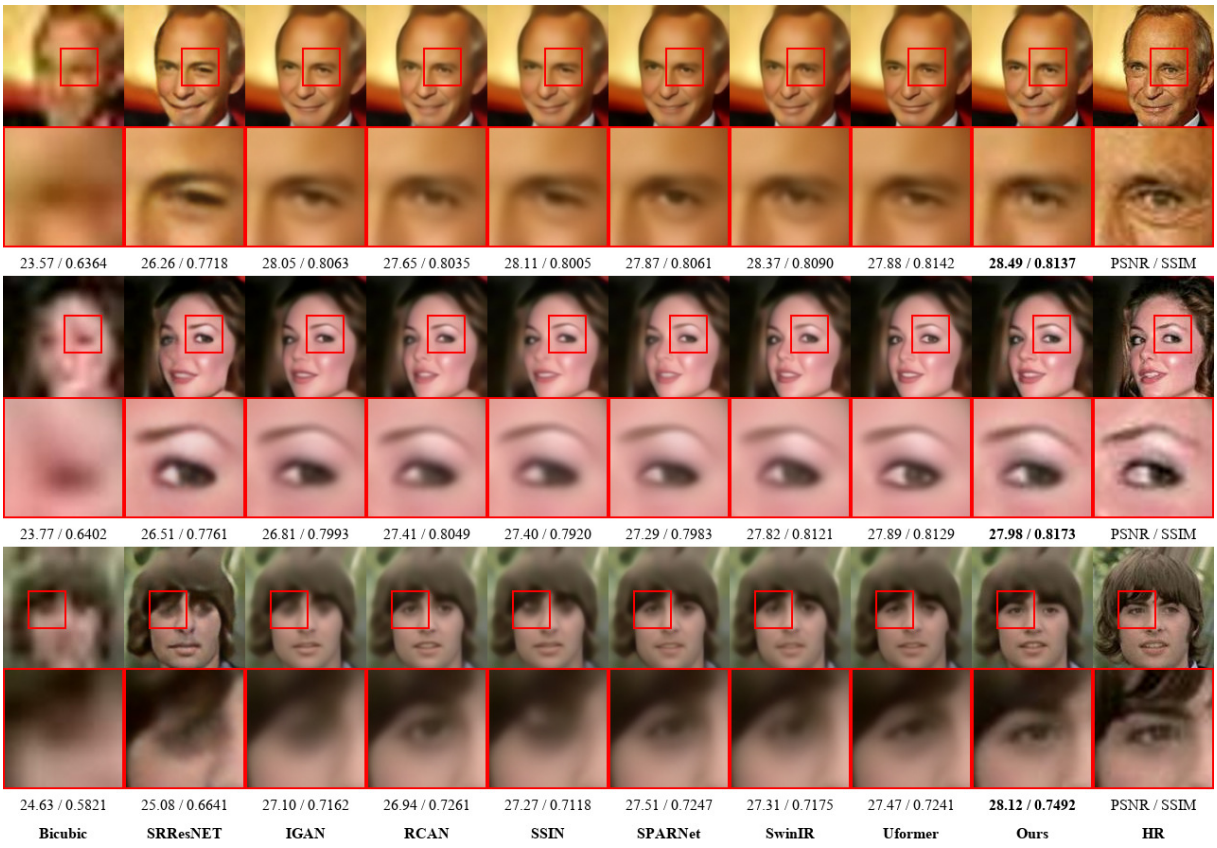

Figure 6. Visual comparisons for ×8 SR on the CelebA test set. Please zoom in for better comparison.

(2) Comparison on Helen dataset: Aiming to prove the flexibility of the proposed model, we assess its performance on the Helen dataset using the same model trained on CelebA. We present a quantitative and visual comparison of the proposed method with others on the Helen dataset in Table 6 and Figure 7, respectively. According to the results, the proposed method still shows superiority in restoring facial images both quantitatively and qualitatively. This proves the robustness and stability of the proposed method. However, it is worth noting that all methods experience a decrease in performance when the training and testing images are not from the same dataset. Therefore, investigating the styles among various datasets will be a promising way to enhance the generality of FSR methods in the future.

(3) Comparison on real face images: Restoring face images from real-world environments is a challenging task due to the complexity of the captured images. Although the CelebA dataset is a good source for simulating face images, it cannot replicate all the complexities of real-life scenarios. In order to test the effectiveness of the proposed method in restoring real-world face images, we conduct experiments on low-quality face images collected from the classic TV series "Friends". It was shot in the 90s with low-tech imaging equipment and suffered severe low-resolution issues, making it perfect for testing. The experiment results are illustrated in Figure 8. Benefiting from the guiding blocks pointing to the key features and the pixel-related upsample layer preserving face structures, our method reconstructs more detailed facial images with appealing facial structures compared with other state-of-the-art methods.

4.5. Noise Stress Test

Due to the fact that noises from image sensors are all randomly valued and located, the Gaussian noise best fits the image degradation model. However, some other noises from specific situations could also challenge the performance of the proposed FSR method. Therefore, we stress-test our model in this subsection on seven different noises: Gaussian, Poisson, Rayleigh, gamma, exponential, uniform, and salt-pepper.

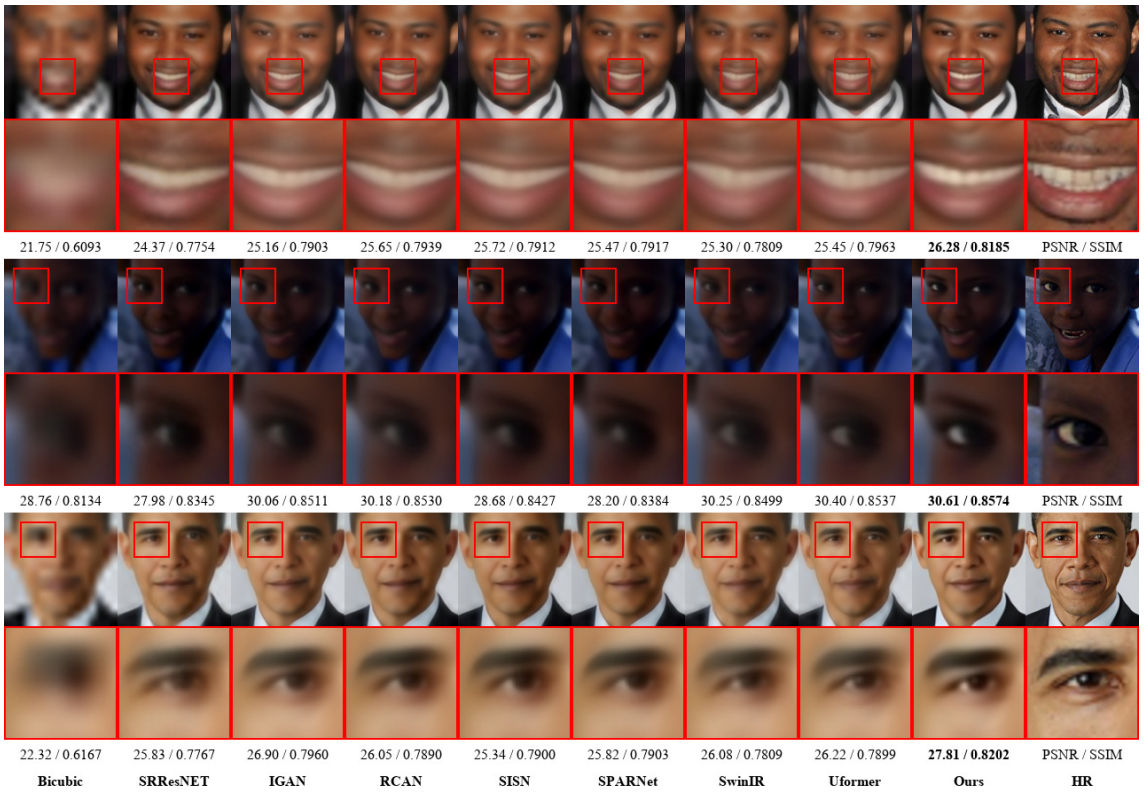

Figure 7. Visual comparisons for ×8 SR on the Helen test set. Please zoom in for better comparison.

Figure 8. Visual comparisons for ×8 SR on real face images. Please zoom in for better comparison.

Experiments in this subsection are conducted on the same 1000 test face images from the CelebA dataset as above. The noise images are multiplied by 0.3 and then added to the original HR images to simulate the noise-degrading process, except for the salt-pepper noise that directly operated on the original HR images. All the noise generation models can be obtained from Numpy [62]. To make a fair comparison, we manually alternate the parameters of these noise models by adjusting the PSNR of their HR outputs to 26.0–26.5 dB. Here are the detailed parameters of the noise models: The Gaussian noise has a mean value of 0 and a standard deviation of 70. The Poisson noise has a lambda value of 50, while the Rayleigh noise has a scale of 40. The gamma noise has a shape of 7 and a scale of 7. The exponential noise has a scale value of 43. The uniform noise has a low value of 20 and a high value of 80. Lastly, the probability of the salt-pepper noise is set to 0.01. The noise images and their HR outputs are shown in Figure 9.

These HR images, which have been impaired by noise, are */8* downscaled and subsequently directed through the proposed method. Additionally, we introduce the bicubic interpolation results as the baseline while choosing the best comparative method, Uformer, for better comparison. Table 7 and Figure 10 show the quality evaluation matrices and visual comparisons, respectively.

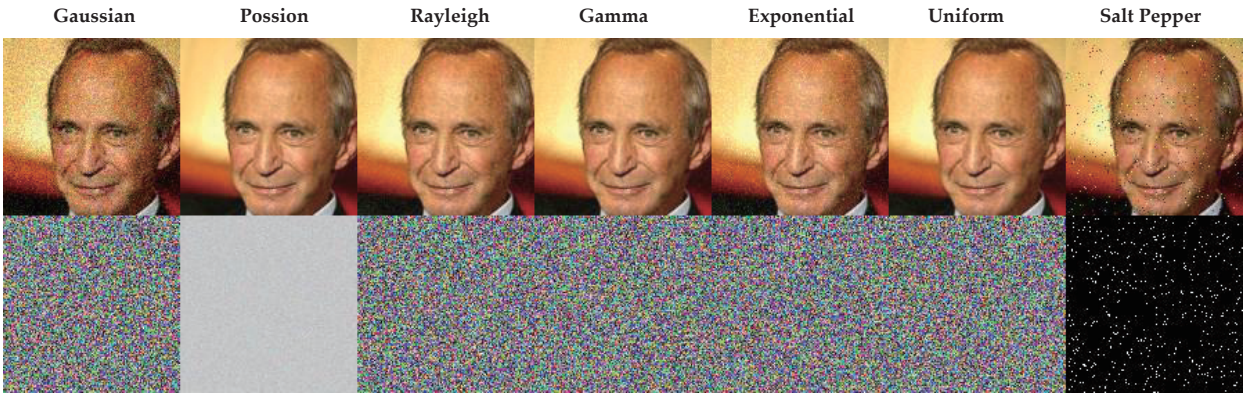

Figure 9. Visual noise images and their HR outputs. Please zoom in for better comparison.

Table 7. Performance comparisons of different ×8 SR on face images with noises (PSNR↑/SSIM↑).

Methods	Gaussian	Poisson	Rayleigh	Gamma	Exponential	Uniform	Salt Pepper
Bicubic	23.96/0.6486	21.95/0.6400	21.95/0.6398	22.02/0.6404	22.43/0.6417	21.95/0.6400	23.98/0.6489
Uformer [15]	23.77/0.6303	21.97/0.6260	21.98/0.6253	22.05/0.6262	22.44/0.6274	21.98/0.6258	23.81/0.6320
Ours	**27.38/0.7837**	24.08/0.7817	24.09/0.7808	24.19/0.7818	24.84/0.7818	24.10/0.7817	**27.36/0.7823**
Noise HR	26.49/0.6175	**26.34/0.9818**	**26.10/0.9322**	**26.43/0.9602**	**26.51/0.8309**	**26.29/0.9638**	26.40/0.6959

The **best** results are emphasized with **bold**. ↑ and ↓ indicate whether higher or lower evaluation matrix values correspond to better image quality, respectively.

From Table 7 and Figure 10, we can observe that:

(a) There are varying degrees of reduction in reconstructing face images using different methods. The Uformer, which proves its effectiveness in denoising, requests a particular denoising dataset for training, or else it will be unable to construct reasonable images. On the contrary, the proposed method, which introduces the guiding blocks and PRUM to mine and preserve face structures, has successfully reconstructed face images with noises.

(b) Noises like Gaussian and salt-pepper influence face structures (i.e., SSIM) much more than others. However, they can easily be recovered using the proposed method. This is because the natural face images taken from image sensors always contain Gaussian noises, which makes the proposed method familiar with this kind of noise. The salt-pepper noise influences only a few pixels, while the */8* times degradation further weakens its impact. Therefore, it can be overcome by the face structure preserving modules of the proposed method.

(c) The Poisson, Rayleigh, gamma, exponential, and uniform noises greatly affect the performance of all FSR methods, which proves their ability to blur images. More attention needs to be paid to overcome the influence of these types of noises.

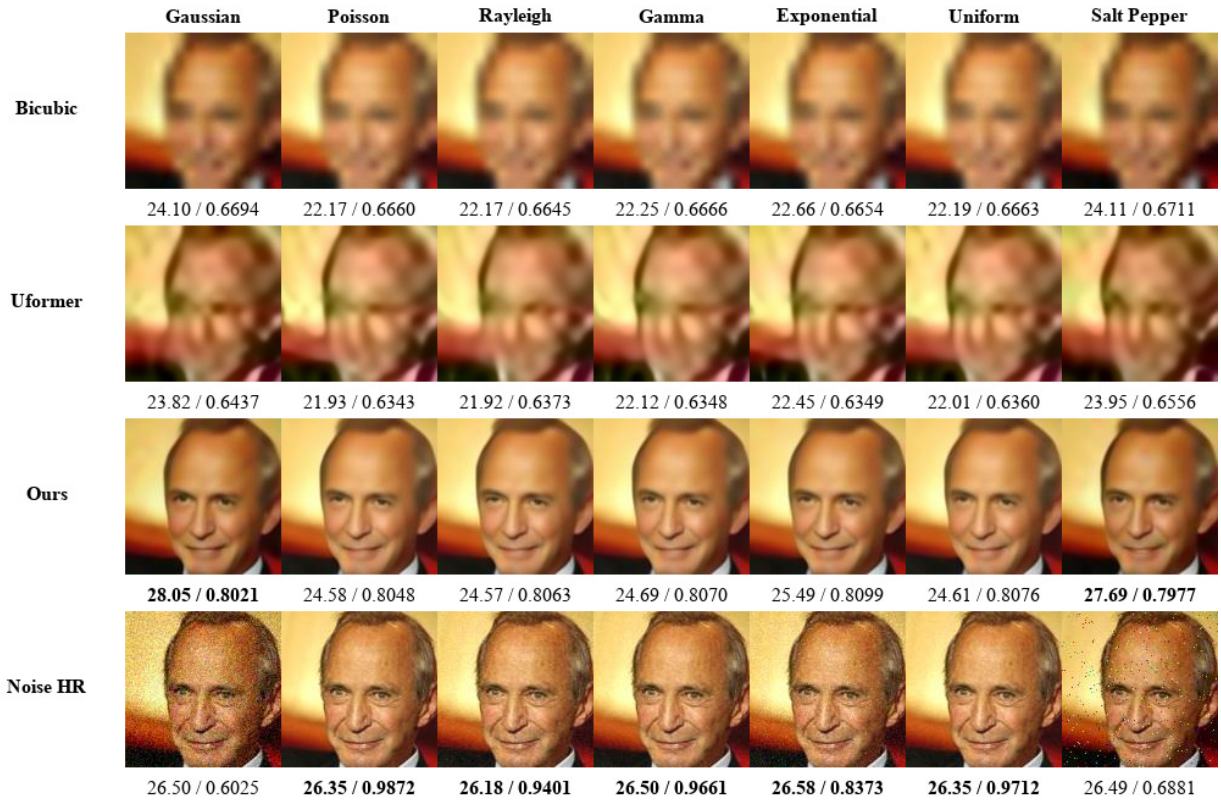

Figure 10. Visual comparisons for ×8 SR on face images with noises. Please zoom in for better comparison.

4.6. Face Recognition Results

To further prove that the proposed method can recover crucial facial structures that are essential in distinguishing different faces, we also perform face recognition as a measurement. Specifically, we chose the commonly used LFW [63] dataset as the face recognition database. Then, several images are randomly picked, downsampled, and super-resoluted as the reference images with different FSR methods. After that, we select face images with the same and other identities as test images for every reference. Finally, we adopt a pre-trained face recognition model, Deepface [64], to perform face recognition. Moreover, we also measure Uformer along with the proposed method for better comparison and the bicubic interpolation as the baseline. The Receiver Operator Characteristic (ROC) curve can be seen in Figure 11.

From Figure 11, we can observe that:

(a) The performance of Deepface [64] on the original HR images is excellent, which proves the significant improvement in the face recognition field based on the deep-learning network.

(b) SR images with bicubic interpolation are difficult for Deepface [64] to verify. This is reasonable due to the poor SR performance of the bicubic interpolation, which can be seen from the sections mentioned above.

(c) Both the SR images reconstructed by the Uformer and the proposed method have obtained satisfactory face recognition performance. Moreover, the proposed method has a larger AUC result, demonstrating its better performance in face recognition tasks and further proving its ability to recover crucial facial structures that are essential in distinguishing different faces.

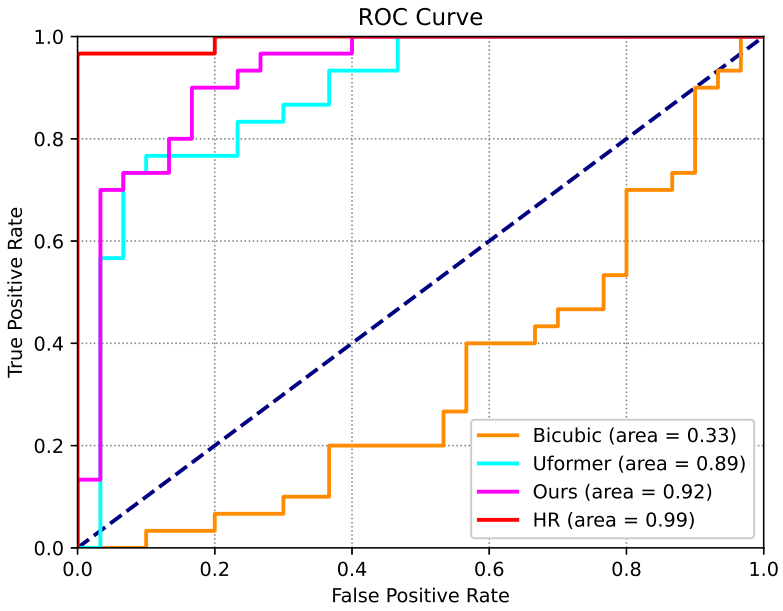

Figure 11. ROC curve on LFW [63] for face recognition task. The "area" in the legend of the figure represents the area under the ROC curve (AUC).

4.7. Model Complexity Analysis

In previous experiments, the proposed method has demonstrated its superior ability in both quantitative and qualitative FSR performance. In this section, we compare its model performance, size, and execution time with other state-of-the-art methods, whose results are shown in Figure 12. According to the figure, our method achieves the best quantitative results while maintaining comparable model size and execution time. Hence, the proposed approach strikes a better balance between model performance, size, and execution time compared to other state-of-the-art methods.

Figure 12. Model complexity scattergram for ×8 SR on the CelebA test set.

5. Conclusions

This work proposes a novel attention-guided transformer with pixel-related deconvolution network for face super-resolution. This is the first study for the transformer-based FSR field to not only mine potential inner feature map information but also establish direct

relationships among adjacent pixels in reconstructing highly structured face images. The proposed method utilizes a multi-scale connected encoder-decoder architecture as the backbone. Specifically, we design an Attention-Guided Transformer Module (AGTM), which is composed of an Attention Guiding Block (AGB) and a Channel-wise Multi-head Transformer Block (CMTB). AGTM at the top of the encoder-decoder network (AGTM-T) promotes both local facial details and global facial structures, while AGTM at the bottleneck side (AGTM-B) optimizes the encoded low-level features. The channel-wise CMTB overcomes the problem that the usual spatial-wise transformers are limited to position-specific windows and exploits feature map channels to achieve an image-size receptive field. Furthermore, considering that face images are highly structured, we design a Pixel-Related Deconvolution (PRD) layer to establish direct relationships among adjacent pixels in the upsampling process for better face structure preservation. Moreover, we have also developed a Multi-scale Feature Fusion Module (MFFM) to fuse multi-scale features for better network flexibility and reconstruction results. Quantitative and qualitative experimental results on both simulated and real-world datasets demonstrate that the proposed method can achieve state-of-the-art performance.

Author Contributions: Conceptualization, Z.Z.; data curation, Z.Z.; methodology, Z.Z.; software, Z.Z.; supervision, C.Q.; validation, Z.Z.; visualization, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z. and C.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Nos. 61572395 and 61675161).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both CelebA [22] and Helen [23] datasets are available online.

Acknowledgments: We thank the editors and reviewers for taking the time and effort to review the manuscript. We sincerely appreciate all valuable comments and suggestions that have helped us improve the quality of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baker, S.; Kanade, T. Hallucinating faces. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 28–30 March 2000; pp. 83–88.
2. Jiang, J.J.; Wang, C.Y.; Liu, X.M.; Ma, J.Y. Deep learning-based face super-resolution: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–36. [CrossRef]
3. Zhang, L.; Wu, X. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Trans. Image Process.* **2006**, *15*, 2226–2238. [CrossRef] [PubMed]
4. Chakrabarti, A.; Rajagopalan, A.N.; Chellappa, R. Super-resolution of face images using kernel pca-based prior. *IEEE Trans. Multimed.* **2007**, *9*, 888–892. [CrossRef]
5. Jung, C.K.; Jiao, L.C.; Liu, B.; Gong, M.G. Position-patch based face hallucination using convex optimization. *IEEE Signal Process. Lett.* **2011**, *18*, 367–370. [CrossRef]
6. Tappen, M.F.; Liu, C. A bayesian approach to alignment-based image hallucination. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 236–249.
7. Zhang, K.B.; Gao, X.B.; Tao, D.C.; Li, X.L. Single Image Super-Resolution With Non-Local Means and Steering Kernel Regression. *IEEE Trans. Image Process.* **2012**, *21*, 4544–4556. [CrossRef] [PubMed]
8. Jiang, J.J.; Hu, R.M.; Wang, Z.Y.; Han Z. Face Super-Resolution via Multilayer Locality-Constrained Iterative Neighbor Embedding and Intermediate Dictionary Learning. *IEEE Trans. Image Process.* **2014**, *23*, 4220–4231. [CrossRef] [PubMed]
9. Zhang, K.; Zhang, Z.; Cheng, C.W.; Hsu, W. H.; Qiao, Y.; Liu, W.; Zhang T. Super-identity convolutional neural network for face hallucination. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 183–198.
10. Chen, C.; Gong, D.; Wang, H.; Li, Z.; Wong, K.-Y.K. Learning spatial attention for face super-resolution. *IEEE Trans. Image Process* **2020**, *30*, 1219–1231. [CrossRef]

11. Lu, T.; Wang, Y.; Zhang, Y.; Wang, Y.; Wei, L.; Wang, Z.; Jiang, J. Face hallucination via split-attention in split-attention network. In Proceedings of the ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 5501–5509.
12. Yang, T.; Ren, P.; Xie, X.; Zhang, L. Gan prior embedded network for blind face restoration in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 672–681.
13. Yang, D.; Wei, Y.; Hu, C.; Yu, X.; Sun, C.; Wu, S.; Zhang, J. Multi-Scale Feature Fusion and Structure-Preserving Network for Face Super-Resolution. *Appl. Sci.* **2023**, *13*, 8928. [CrossRef]
14. Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; Luo, P. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 512–521.
15. Wang, Z.D.; Cun, X.D.; Bao, J.M.; Zhou, W.G.; Liu, J.Z.; Li, H.Q. Uformer: A General U-Shaped Transformer for Image Restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17683–17693.
16. Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; Tan, M. Closed-loop matters: Dual regression networks for single image superresolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 5407–5416.
17. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Montreal, BC, Canada, 19–25 June 2021; pp. 1833–1844.
18. Gao, G.; Xu, Z.; Li, J.; Yang, J.; Zeng, T.; Qi, G.J. CTCNet: A CNN-Transformer Cooperation Network for Face Image Super-Resolution. *IEEE Trans. Image Process* **2023**, *32*, 1978–1991. [CrossRef]
19. Wang, C.; Jiang, J.; Zhong, Z.; Liu, X. Spatial-Frequency Mutual Learning for Face Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–24 June 2023; pp. 22356–22366.
20. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 5835–5843.
21. Leland, M.; John, H.; James, M. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, arXiv:1802.03426.
22. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
23. Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; Huang, T.S. Interactive facial feature localization. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 679–692.
24. Wang, Z.; Chen, J.; Hoi, S.C. Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3365–3387. [CrossRef]
25. Li, J.; Pei, Z.; Zeng, T. From beginner to master: A survey for deep learning-based single-image super-resolution. *arXiv* **2021**, arXiv:2109.14335.
26. Zhou, E.J.; Fan, H.Q.; Cao, Z.M.; Jiang, Y.N.; Yin, Q. Learning face hallucination in the wild. In Proceedings of the Association for the Advancement of Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 3871–3877.
27. Cao, Q.X.; Lin, L.; Shi, Y.K.; Liang, X.D.; Li, G.B. Attention-aware face hallucination via deep reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 690–698.
28. Huang, H.B.; He, R.; Sun, Z.N.; Tan, T.N. Wavelet domain generative adversarial network for multiscale face hallucination. *Int. J. Comput. Vis.* **2019**, *127*, 763–784. [CrossRef]
29. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.H.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 105–114.
30. Yang, L.B.; Liu, C.; Wang, P.; Wang, S.S.; Ren, P.R.; Ma, S.W.; Gao, W. Hifacegan: Face renovation via collaborative suppression and replenishment. In Proceedings of the ACM International Conference on Multimedia, Dublin, Ireland, 8–11 June 2020; pp. 1551–1560.
31. Yu, X.; Fernando, B.; Hartley, R.; Porikli, F. Semantic Face Hallucination: Super-Resolving Very Low-Resolution Face Images with Supplementary Attributes. *IEEE Trans. PAMI* **2020**, *42*, 2926–2943. [CrossRef]
32. Dou, H.; Chen, C.; Hu, X.Y.; Xuan, Z.X.; Hu, Z.S.; Peng, S.L. Pca-srgan: Incremental orthogonal projection discrimination for face super-resolution. In Proceedings of the ACM International Conference on Multimedia, Dublin, Ireland, 8–11 June 2020; pp. 1891–1899.
33. Zhang, M.L.; Ling, Q. Supervised pixel-wise GAN for face super-resolution. *IEEE Trans. Multimed.* **2021**, *23*, 1938–1950. [CrossRef]
34. Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; Yang, J. Fsrnet: End-to-end learning face super-resolution with facial priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2492–2501.
35. Bulat, A.; Tzimiropoulos, G. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 109–117.

36. Hu, X.; Ren, W.; LaMaster, J.; Cao, X.; Li, X.; Li, Z.; Menze, B.; Liu, W. Face super-resolution guided by 3d facial priors. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 763–780.
37. Ma, C.; Jiang, Z.; Rao, Y.; Lu, J.; Zhou, J. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5569–5578.
38. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 286–301.
39. Li, Z.Z.; Li, G.; Li, T.; Liu, S.; Gao, W. Information-Growth Attention Network for Image Super-Resolution. In Proceedings of the ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 544–552.
40. Kalarot, R.; Li, T.; Porikli, F. Component Attention Guided Face Super-Resolution Network: CAGFace. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 359–369.
41. Li, C.; Xiao, N. A Face Structure Attention Network for Face Super-Resolution. In Proceedings of the International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 75–81.
42. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 347–357.
43. Xiong, L.; Zhang, J.; Zheng, X.; Wang, Y. Context Transformer and Adaptive Method with Visual Transformer for Robust Facial Expression Recognition. *Appl. Sci.* **2024**, *14*, 1535. [CrossRef]
44. Shi, A.; Ding, H. Underwater Image Super-Resolution via Dual-aware Integrated Network. *Appl. Sci.* **2023**, *13*, 12985. [CrossRef]
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
46. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323–2326. [CrossRef]
47. Zhang, Z.; Qi, C.; Asif, M.R. Investigation on Projection Space Pairs in Neighbor Embedding Algorithms. In Proceedings of the IEEE International Conference on Signal Processing, Beijing, China, 12–16 August 2018; pp. 125–128.
48. Hao, Y.H.; Qi, C. Face Hallucination Based on Modified Neighbor Embedding and Global Smoothness Constraint. *IEEE Signal Process. Lett.* **2014**, *21*, 1187–1191. [CrossRef]
49. Tu, Q.; Li, J.W.; Javaria, I. Locality constraint neighbor embedding via KPCA and optimized reference patch for face hallucination. In Proceedings of the IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016; pp. 424–428.
50. Yang, W.; Xia, S.; Liu, J.; Guo, Z. Reference-Guided Deep Super-Resolution via Manifold Localized External Compensation. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 1270–1283. [CrossRef]
51. Menon, S.; Damian, A.; Hu, S.; Ravi, N.; Rudin, C. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2434–2442.
52. Chen, L.; Pan, J.; Jiang, J.; Zhang, J.; Han, Z.; Bao, L. Multi-Stage Degradation Homogenization for Super-Resolution of Face Images With Extreme Degradations. *IEEE Trans. Image Process.* **2021**, *30*, 5600–5612. [CrossRef]
53. Howard, J.; Gugger, S. Deep Learning from Scratch. In *Deep Learning for Coders with fastai and PyTorch*; Faucher, C., Hassell, J., Potter, M., Eds.; O'Reilly Media: Sebastopol, CA, USA, 2020; pp. 493–515.
54. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
55. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
56. Hendrycks D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
57. Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5718–5729.
58. Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
59. Sheikh, H. R.; Bovik, A. C. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [CrossRef] [PubMed]
60. Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
61. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.M.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4–9.
62. NumPy. Available online: https://numpy.org/ (accessed on 20 April 2024).

63. Huang, G. B.; Mattar, M.; Berg, T.; Learned-Miller, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*; Technical Report 07-49; University of Massachusetts: Amherst, MA, USA, 2007.
64. Serengil, S. I.; Ozpinar, A. Lightface: A hybrid deep face recognition framework. In Proceedings of the Innovations in Intelligent Systems and Applications Conference, Istanbul, Turkey, 15–17 October 2020; pp. 23–27.