applied sciences

MDPI

Article

Feature Maps Need More Attention: A Spatial-Channel Mutual Attention-Guided Transformer 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: Recently, transformer-based face super-resolution (FSR) approaches have achieved promising success in restoring degraded facial details due to their high capability for capturing both local and global dependencies. However, while existing methods focus on introducing sophisticated structures, they neglect the potential feature map information, limiting FSR performance. To circumvent this problem, we carefully design a pair of guiding blocks to dig for possible feature map information to enhance features before feeding them to transformer blocks. Relying on the guiding blocks, we propose a spatial-channel mutual attention-guided transformer network for FSR, for which the backbone architecture is a multi-scale connected encoder–decoder. Specifically, we devise a novel Spatial-Channel Mutual Attention-guided Transformer Module (SCATM), which is composed of a Spatial-Channel Mutual Attention Guiding Block (SCAGB) and a Channel-wise Multi-head Transformer Block (CMTB). SCATM on the top layer (SCATM-T) aims to promote both local facial details and global facial structures, while SCATM on the bottom layer (SCATM-B) seeks to optimize the encoded features. Considering that different scale features are complementary, we further develop a Multi-scale Feature Fusion Module (MFFM), which fuses features from different scales for better restoration performance. Quantitative and qualitative experimental results on various datasets indicate that the proposed method outperforms other state-of-the-art FSR methods.

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

Citation: Zhang, Z.; Qi, C. Feature Maps Need More Attention: A Spatial-Channel Mutual Attention-Guided Transformer Network for Face Super-Resolution. *Appl. Sci.* **2024**, *14*, 4066. https://doi.org/10.3390/app14104066

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 1 March 2024
Revised: 1 May 2024
Accepted: 9 May 2024
Published: 10 May 2024

1. Introduction

Face super-resolution (FSR), also known as face hallucination [1], is a technology for enhancing low-resolution (LR) face images into high-resolution (HR) ones. Due to low-cost imaging equipment and limited imaging conditions, face images often have a lower spatial resolution, which severely degrades the performance of most practical downstream applications such as face analysis and face recognition. Therefore, FSR has become increasingly popular in the computer vision and image processing fields, making it an important scientific tool [2].

Unlike general image super-resolution, FSR focuses on recovering pivotal facial structures. Although facial structures only occupy a small portion of the face, they are crucial in distinguishing different faces. The first FSR method, proposed by Baker and Kanade [1], triggered the upsurge of traditional FSR methods. Over time, various traditional techniques for FSR have been developed, including the interpolation approach [3], PCA [4], convex optimization [5], Bayesian approach [6], kernel regression [7], and manifold learning [8]. Traditional methods are limited by their shallow structure and representation abilities, making them incompetent in producing plausible facial images.

Alongside the rise of deep learning techniques [2,9,10], deep convolution neural networks (CNNs) [11–15] have achieved remarkable advancements in improving face image quality. The first CNN-based FSR method was introduced by Zhou et al. [11], greatly improving FSR performance compared with traditional FSR methods. To further explore facial

information, Cao et al. [12] utilized reinforcement learning to capture the interdependency among facial parts. Zhang et al. [13] introduced super-identity loss to assist the network in generating more accurately identified super-resolution face images. In contrast to the FSR methods above, which recover face images directly, Huang et al. [14] used the wavelet transform to project face images into wavelet spaces to capture rich contextual information, while Wang et al. [15] applied the Fourier transform to obtain an image-size receptive field for capturing global facial structure.

Inspired by the great success of generative adversarial networks (GANs) in the image processing field [16–18], Yang et al. [19] introduced a collaborative suppression and replenishment framework based on GANs. Gao et al. [20] proposed a conditional generative model based on the diffusion model, which replaces the U-Net in super-resolution to capture complex details and fine textures. Addressing the fact that GAN-based methods require greater computational resources, PCA-SRGAN [21] uses Principal Component Analysis decomposition, while SPGAN [22] employs supervised pixel-wise loss to ease the GAN training process.

Due to the highly structured nature of the human faces, many FSR methods leverage facial priors, e.g., face landmarks and parsing maps, to enhance the reconstruction performance. Chen et al. [23] utilized facial parsing maps to guide the end-to-end FSR convolution network. Meanwhile, Bulat et al. [24] combined GANs with a well-designed heatmap loss to constrain the face structure between HR and super-resolved (SR) face images. To better capture sharp facial structures in face images with large pose variations, Hu et al. [25] introduced 3D facial priors instead of the commonly used 2D ones. Considering the challenge of estimating priors from degraded LR face images directly, DIC [26] used an iterative process in which FSR and prior estimation were performed repeatedly to enhance FSR performance.

Recently, the attention mechanism has emerged as a new trend in computer vision tasks [27–31]. Chen et al. [32] devised a face attention unit to capture facial structure information, while Lu et al. [33] designed an external–internal split attention group to reconstruct clear facial images. Furthermore, the performance of transformers has already been proven and widely applied in computer vision tasks, such as image recognition [34,35] and image restoration [28,36,37]. The core of the transformers is a self-attention mechanism that can capture both long- and short-range correlations between words/pixels [38]. Therefore, transformer-based methods have gained dramatic attention for their high ability to capture both local and global dependencies.

While transformer-based methods have led to significant improvements in FSR performance, they have a number of limitations that need to be addressed. First, there is a lack of sufficient discussion regarding the inner feature maps. As shown in Figure 1b, without guidance, the features of the inner feature maps may not always be detail-rich, and may even be buried in gray. This restricts the transformer block to selecting only a limited number of feature maps based on the self-attention matrix (Figure 1c), while leaving others untouched. On the contrary, from Figure 1d it can be seen that utilizing a guiding block to enhance essential facial components results in more correlated output feature maps, benefiting the "one-to-many" FSR problem [39] and yielding superior outcomes (Figure 1e). Therefore, it is crucial to have a guiding block that can identify the essential facial component for inner feature maps in transformer-based methods. Second, most previous transformer-based methods [28,37] have utilized the same transformer structure for different feature layers. However, the transformer structure on high-level features cannot be thoroughly applied to low-level features [40], and vice versa, otherwise resulting in unsatisfactory FSR performance. Lastly, the usual spatial-wise transformers are limited to position-specific windows, meaning that the partition strategy may potentially alter the structure of the facial image [15].

Based on the above analysis, we propose a spatial-channel mutual attention-guided transformer network for face super-resolution, which explores the potential of inner feature maps for reconstructing plausible face images. The proposed method is a multi-scale

connected encoder–decoder network. In the encoder–decoder branches, a Spatial-Channel Mutual Attention-guided Transformer Module (SCATM) is carefully designed to extract more detailed features by enhancing the relationship between inner feature maps. It is composed of a Spatial-Channel Mutual Attention Guiding Block (SCAGB) and a Channel-wise Multi-head Transformer Block (CMTB). Among them, SCAGB aims to guide the transformer block in identifying the essential facial component. SCAGB on the top layer (SCAGB-T) aims to guide and promote both local facial details and global facial structures, while SCAGB on the bottom layer (SCAGB-B) seeks to identify the crucial encoded features. Unlike the usual spatial-wise transformersm which are limited to position-specific windows, the CMTB utilizes feature map channels to achieve image-size receptive fields. The combination of SCAGB and CMTB is complementary, and can simultaneously promote both local facial details and global facial structures. Meanwhile, unlike pyramid super-resolution networks [41,42], which progressively reconstruct high-resolution images, we further develop a Multi-scale Feature Fusion Module (MFFM) which fuses features from all layers, making for network flexibility and better restoration performance.

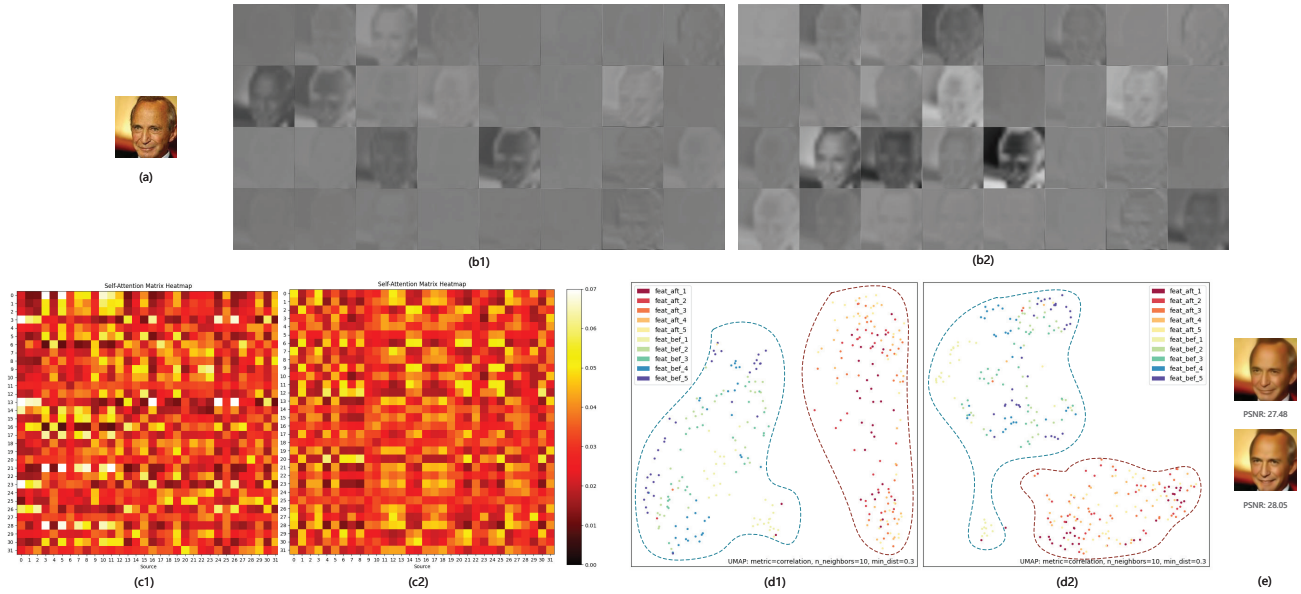

Figure 1. Visual analysis of the inner feature maps trained with/without guiding blocks for transformer-based FSR methods: (**a**) the input face image; (**b1**) the inner feature maps without guiding blocks; (**b2**) the inner feature maps with guiding blocks; (**c1**) the self-attention heatmap without guiding blocks; (**c2**) the self-attention heatmap with guiding blocks; (**d1**) the correlation [43] between input and output feature maps without guiding blocks (please note that for fair comparison five different images are tested here instead of one); (**d2**) the correlation between input and output feature maps with guiding blocks; (**e**) the output images (the top one trained without guiding blocks and bottom one trained with guiding blocks).

In summary, the main contributions of this work are four-fold:

- We devise a spatial-channel mutual attention-guided transformer network for face super-resolution. To the best of our knowledge, this is the first paper to explore the potential of inner feature maps in reconstructing plausible face images in the transformer-based FSR area.
- We carefully design a Spatial-Channel Mutual Attention-guided Transformer Module (SCATM) to extract more detailed features by enhancing the relationship between inner feature maps. Thanks to its powerful modeling ability, both local facial details and global facial structures can be fully explored and utilized.

- We propose an elaborately designed Multi-scale Feature Fusion Module (MFFM) to fuse multi-scale features during the reconstruction process. This module is crucial in enabling our method to acquire a wide range of features, which in turn enhances the quality of the restoration performance.
- We conduct experiments to confirm the effectiveness of the proposed method. The results of our experiments, conducted on two frequently used benchmark datasets (CelebA [44] and Helen [45]) demonstrate that our method surpasses others in terms of both visual outcomes and quantitative measurements.

2. Proposed Method

Considering the vital role of inner feature maps in identifying the essential facial component for better FSR performance, we develop a spatial-channel mutual attention-guided transformer network for FSR, which is the first study to explore the potential of inner feature maps in reconstructing plausible face images in the transformer-based FSR area.

To better elaborate the proposed method, we divide the method into four subsections. In the first subsection, we briefly revisit the difference between feature maps and feature spaces, which is the key foundation of the proposed method. Then, we take an overall view of the architecture of the proposed method. Next, we focus on the main part of the proposed method, the Spatial-Channel Mutual Attention-guided Transformer Module (SCATM), which contains a Spatial-Channel Mutual Attention Guiding Block (SCAGB) and a Channel-wise Multi-head Transformer Block (CMTB). The combination of SCAGB and CMTB is complementary, and can simultaneously promote both local facial details and global facial structures. Finally, we introduce the Multi-scale Feature Fusion Module (MFFM), which fuses features from all layers for network flexibility and better restoration performance.

2.1. Revisiting Feature Maps and Feature Spaces

Feature maps and feature spaces are similar in several ways: CNN-based FSR methods utilize complex convolution layers to project LR images into inner "feature maps" and then into HR ones; on the other hand, manifold learning-based FSR methods, which belong to the traditional FSR method, project LR images into "feature spaces" and then into HR ones, assuming that LR and HR spaces share the same local geometry [46]. Many traditional FSR methods have been developed based on manifold learning [47–49], focusing on further enhancing the LR and HR space relationship. However, with the rise of CNNs, manifold learning has gained less attention due to the complicated nature of CNN structures, which are difficult to deliberate on. Yang et al. [50] proposed a manifold localized deep external compensation (MALDEC) network that references online big data to provide an accurate localization and mapping to the HR manifold for image super-resolution. Menon et al. [51] traversed the HR manifold spaces to search for images that suit the original LR image, then feed the downscaling loss to a generative model for image reconstruction. Chen et al. [52] proposed a homogenization projection in LR space and HR space to formulate FSR in a multi-stage framework. Guo et al. [39] claimed that limiting mapping spaces would benefit image super-resolution, and introduced a closed-loop dual regression network (DRN) containing an additional constraint. The above methods try to merge traditional manifold learning with CNN-based methods; however, they overlook the key role of inner feature maps (Figure 1), limiting their image super-resolution performance. Therefore, how to effectively handle inner feature maps is vitally important for a high-quality image reconstruction process.

2.2. Overview

The proposed method, illustrated in Figure 2, is a multi-scale connected symmetrical hierarchical network containing three stages: encoding, bottleneck, and decoding. The encoding stage aims to extract and promote both local facial details and global facial structures, while the bottleneck stage is designed to optimize the encoded low-level features. Finally,

the decoding stage is introduced for multi-scale feature fusion and image reconstruction. For convenience of description, we denote $\boldsymbol{I}_{\mathrm{LR}}$, $\boldsymbol{I}_{\mathrm{SR}}$, and $\boldsymbol{I}_{\mathrm{HR}}$ as the low-resolution (LR) images, super-resolved (SR) images, and ground-truth high-resolution (HR) images, respectively.

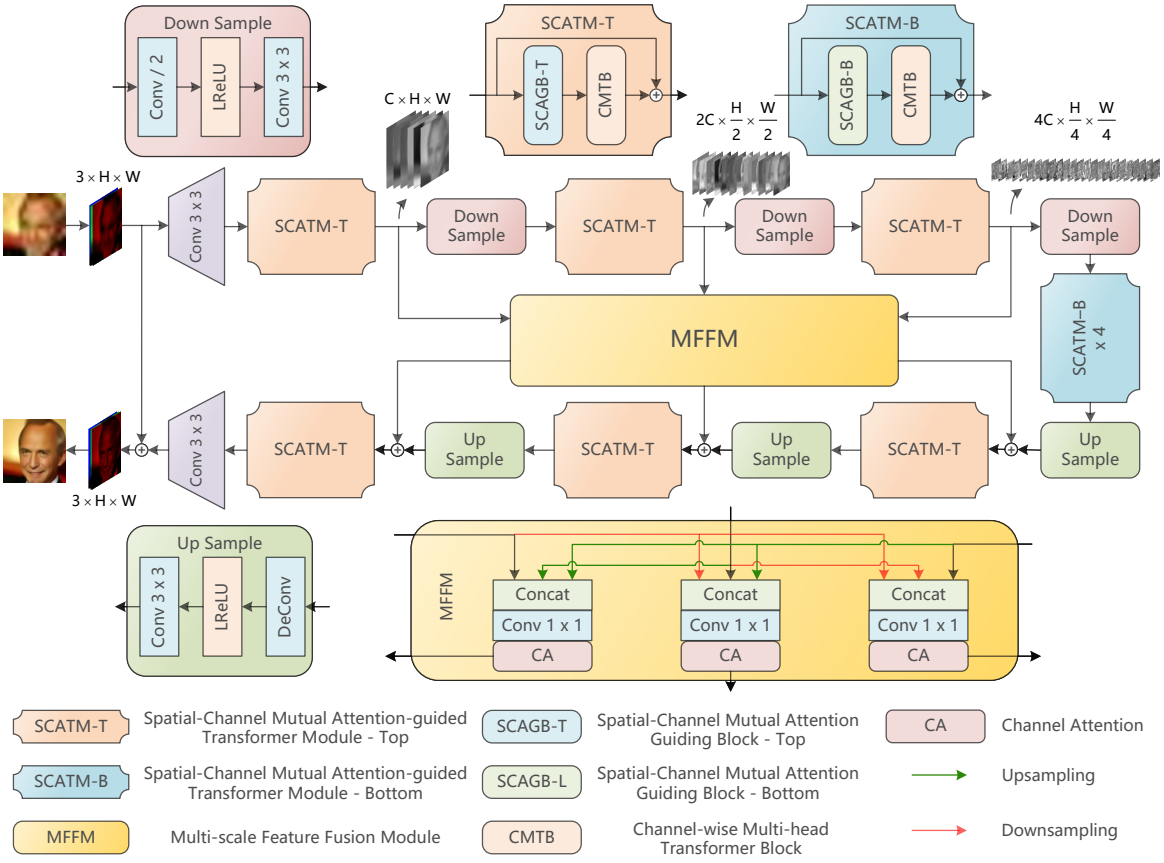

Figure 2. The structure of the proposed spatial-channel mutual attention-guided transformer network. It is a symmetrical hierarchical network containing three stages: encoding, bottleneck, and decoding. The encoding stage aims to extract and promote both local facial details and global facial structures, while the bottleneck stage is designed to optimize the encoded low-level features. Finally, the decoding stage is introduced for multi-scale feature fusion and image reconstruction.

(1) Encoding Stage: The encoding stage aims to extract and promote both local facial details and global facial structures. First, a 3×3 convolution layer is applied to extract shallow features. Here, we suggest using 32 output channels for optimal performance. It is essential to keep in mind that the number of output channels should exceed the number of input channels, while an excessive number of output channels may lead to a significant increase in computation complexity and a decrease in feature information storage effectiveness [53]. Afterwards, the extracted shallow features are passed through three encoding stages. Each stage contains a carefully designed Spatial-Channel Mutual Attention-guided Transformer Module-Top (SCATM-T) and a downsampling block. The SCATM-T contains a Spatial-Channel Mutual Attention Guiding Block-Top (SCAGB-T) and a Channel-wise Multi-head Transformer Block (CMTB), which are discussed in the following subsection. The downsampling block consists of three layers: a 3×3 convolutional layer with a stride of 2, followed by a LeakyReLU activation function layer, then another 3×3 convolutional layer with a stride of 1. Please note that the first convolutional layer doubles the feature channels while simultaneously halving the feature map size.

(2) Bottleneck Stage: The bottleneck stage contains a large number of encoded feature maps, though each map is relatively small compared to those in the encoding stage. To better utilize these feature map features in the decoding stage, we introduce the Spatial-

Channel Mutual Attention-guided Transformer Module-Bottom (SCATM-B). Unlike the SCATM-T in the encoding stage, the guiding blocks in SCATM-B aim to further enhance the low-level encoded features. With the help of SCATM-Bs, the model can focus on a greater variety of facial structures and continuously strengthen different facial features.

(3) Decoding Stage: The decoding stage aims to reconstruct high-quality face images based on the previously extracted and refined multi-layer features. In the decoding stage, the low-level features are first fed to the upsampling block, which contains a 6×6 transposed convolutional layer with a stride of 2 followed by a LeakyReLU activation function layer and a 3×3 convolution layer with a stride of 1. The feature channels are halved and the feature map size is doubled in the first transposed convolutional layer, which is the opposite of the downsampling process in the encoding stage. Afterward, the upsampled features are combined with features from other scales by the Multi-scale Feature Fusion Module (MFFM), extending the network flexibility and resulting in better restoration performance. Then, the well-combined features are fed to the SCATM-T for further image detail enhancement. Lastly, a 3×3 convolutional layer is utilized to convert the learned feature maps into the output face image $\boldsymbol{I}_{\text{Out}}$. The final SR face image output is $\boldsymbol{I}_{\text{SR}} = \boldsymbol{I}_{\text{Out}} + \boldsymbol{I}_{\text{LR}}$. Please note that the LR face images in the paper have already been upsampled to the same size as the HR ones by bicubic interpolation.

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

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{I}_{\text{SR}}^{i} - \boldsymbol{I}_{\text{HR}}^{i} \right\|_1 \tag{1}$$

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

2.3. Spatial-Channel Mutual Attention-Guided Transformer Module (SCATM)

The SCATM is the central part of the proposed method, which contains a Spatial-Channel Mutual Attention Guiding Block (SCAGB) and a Channel-wise Multi-head Transformer Block (CMTB). Based on the need for multi-scale feature maps to apply different suitable blocks for better feature extraction and enhancement, SCAGB is further divided into two distinct blocks: the Spatial-Channel Mutual Attention Guiding Block-Top (SCAGB-T) in the encoding–decoding stage, and the Spatial-Channel Mutual Attention Guiding Block-Bottom (SCAGB-B) in the bottleneck stage. SCATM-T aims to promote both local facial details and global facial structures, while SCATM-B seeks to optimize the encoded low-level features. Moreover, unlike the usual spatial-wise transformers, which are limited to position-specific windows, the CMTB utilizes feature map channels to achieve image-size receptive fields. The combination of SCAGB and CMTB is complementary, and can simultaneously promote both local facial details and global facial structures.

2.3.1. Spatial-Channel Mutual Attention Guiding Block-Top (SCAGB-T)

The SCAGB-T aims to locate and guide both local facial details and global facial structures for the following CMTB, the detailed architecture of which is shown in Figure 3a. First, the Hourglass Block [54], which has already proven its effectiveness in generating spatial attention maps [41], is utilized to capture facial landmark features such as the eyes, nose, and mouth. Moreover, aiming to guide the weights of different feature map channels, the channel attention (CA) network [29] is applied to select and pay more attention to feature map channels that are rich in features. Thanks to a carefully designed structure that mutualizes spatial and channel attention wisely, the SCAGB can guide the following CMTB to capture the essential part of the face images of all channels for better reconstruction results with more details. Afterwards, a 3×3 convolutional layer followed by a sigmoid function is applied to generate the spatial-channel mutual attention map. Then, the input feature maps are multiplied element-wise by the attention map and fed to the subsequent transformer block with better extracted spatial features and promoted channel information.

Aiming to eliminate the gradient vanishing problem, the residual connection with a full connection layer is applied between the input feature maps and the output of the channel attention network.

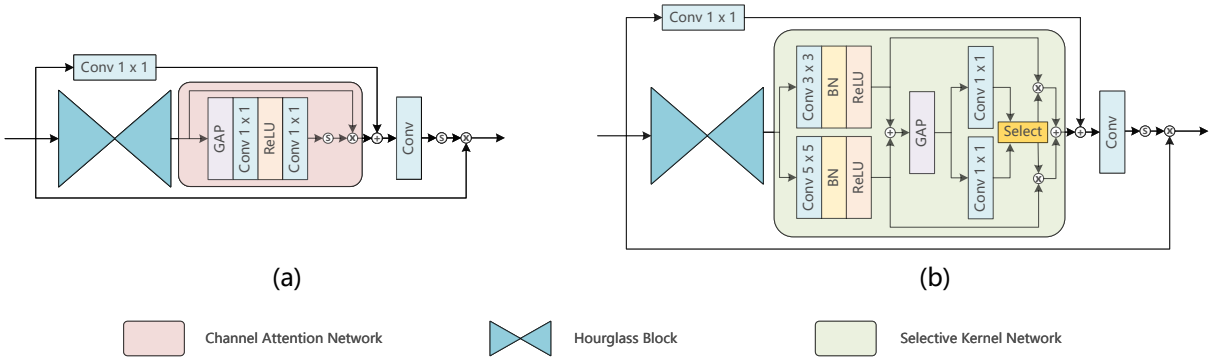

Figure 3. Architectures of the Spatial-Channel Mutual Attention Guiding Block (SCAGB): (**a**) the Spatial-Channel Mutual Attention Guiding Block-Top (SCAGB-T) and (**b**) the Spatial-Channel Mutual Attention Guiding Block-Bottom (SCAGB-B). Here, Ⓢ denotes the sigmoid function.

2.3.2. Spatial-Channel Mutual Attention Guiding Block-Bottom (SCAGB-B)

Unlike the above SCAGB-T, the SCAGB-B in the bottleneck stage aims to guide the enhancement of the low-level encoded features. The bottleneck stage contains a large number of encoded feature maps; however, each map is relatively small compared to the ones in the encoding stage. Therefore, it is essential to introduce a dynamic selection mechanism that allows each neuron to adjust its receptive field size adaptively. The dynamic selection mechanism introduced here, which is the major difference between SCAGB-T and SCAGB-B, is the selective kernel (SK) network [55] shown in Figure 3b. Given a sequence of feature maps, the SK network firstly conducts two convolution layers with different respective fields, followed by a batch normalization layer and a ReLU layer. The calculated upper and lower inner feature maps are denoted as $\mathbf{U}$ and $\mathbf{V}$, respectively. Then, the feature maps are integrated via an element-wise summation and sent through the global average pool (GAP) layer to generate channel-wise statistics with different respective fields. Afterwards, two full connection layers are applied to the inner feature maps to enable the guidance for the adaptive selections. Finally, a soft attention layer across different channels is applied to adaptively select the information from different respective fields. Assuming that the upper and lower input of the Select layer in Figure 3b is $F(\mathbf{U})$ and that $G(\mathbf{V}) \in \mathbb{R}^{C\times 1}$, where $F(\cdot)$ and $G(\cdot)$ denote the previous inner feature map process and C denotes 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 the attention weights on the 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 $\mathbf{A} = [\ \mathbf{A}_1, \quad \mathbf{A}_2, \quad \ldots, \quad \mathbf{A}_C\ ]$ denotes the output attention maps, $\mathbf{A}_c \in \mathbb{R}^{H\times W}$, and H and W denote the height and width of the feature maps, respectively.

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

After preprocessing the inner feature maps with the guiding blocks, there is still a high demand to effectively aggregate the previous feature information from various channels for high-quality face image restoration. Moreover, the usual spatial-wise transformers

are limited to position-specific windows, where the partition strategy may potentially alter the structure of the facial image [15]. Therefore, we introduce the Channel-wise Multi-head Transformer Block (CMTB). It can achieve image-size receptive fields based on channels instead of position-specific windows and is more computation-friendly, making it a good match for the previous guiding blocks. As depicted in Figure 4, CMTB consists of a Channel-wise Multi-head Self-attention Network (CMSN) and a Gated-Dconv Feed-Forward Network (GDFN). The CMSN is the primary component of the CMTB, while the GDFN aims to encode information from spatially neighboring pixel positions to effectively learn local image structures.

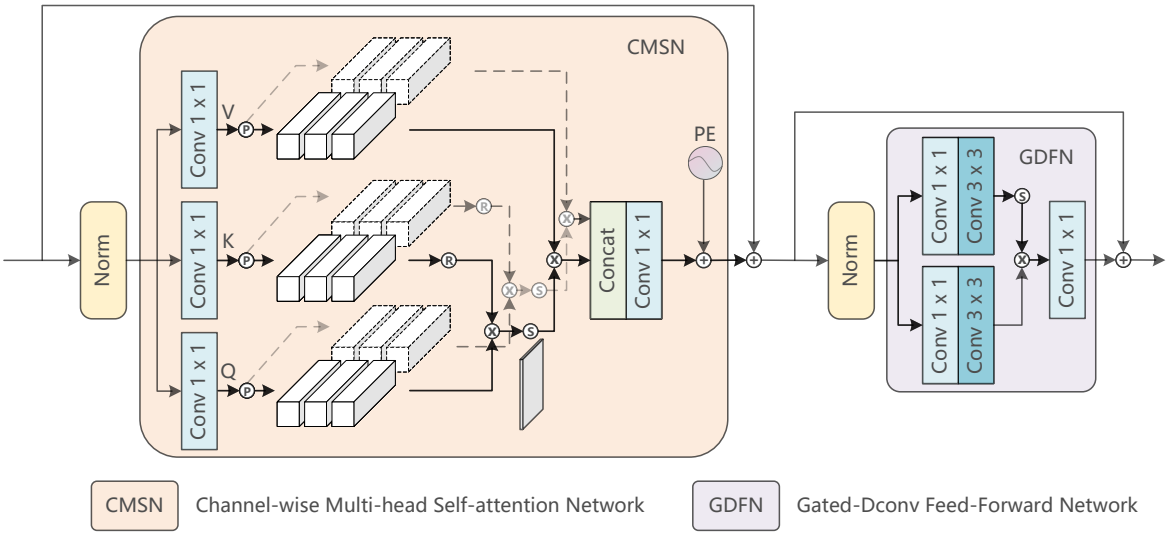

Figure 4. Architecture of the Channel-wise Multi-head Transformer Block (CMTB). Here, Ⓢ, Ⓡ, and Ⓟ denotes the sigmoid function, reshape, and split, respectively, while PE denotes the position embedding generator.

The CMTB can achieve image-size receptive fields based on feature map channels instead of position-specific windows. Given feature maps $\mathbf{X}_{in} \in \mathbb{R}^{H\times W\times C}$ as the input of the CMSN, which are reshaped into tokens $\mathbf{X} \in \mathbb{R}^{HW\times C}$ based on channels (where H, W, and C, respectively denote the height, width, and channel numbers of the feature maps), then $\mathbf{X}$ is linearly projected to achieve 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 this reshape strategy, the size of the generated attention map will be $d \times d$ instead of $HW \times HW$, which greatly cuts down the computational complexity. Moreover, a learnable parameter $\sigma_i \in \mathbb{R}^1$ is introduced to further extend the flexibility of the network. Subsequently, the outputs of N *heads* are fed to the concatenate layer followed by a full connection layer, and the resulting attention matrix is then added with the values from the position embedding generator:

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

where $\mathbf{W} \in \mathbb{R}^{C\times C}$ are learnable parameters and $f_p(\cdot)$ represents the position embedding generator, which aims to encode the position information of different 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 obtained by reshaping the result of Equation (6).

Moreover, we introduce GDFN [57] to encode information from spatially neighboring pixel positions in order to effectively learn local image structures. 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

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

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

The SCAGB distills and guides the key features from inner feature maps, while the CMTB further aggregates and refines the previous feature information. These two blocks complement each other and gradually enhance the relationship between the inner feature maps. Benefiting from the SCAGB and CMTB, the SCATM can simultaneously promote both local facial details and global facial structures, offering a promising solution to the challenging face image reconstruction task.

2.4. Multi-Scale Feature Fusion Module (MFFM)

Successive pyramid super-resolution networks [41,42] have already proven the importance of multi-scale feature information in the image reconstructing process; however, the pyramid methods mentioned above progressively reconstruct high-resolution images from neighboring layers, ignoring the others. Aiming to further utilize the multi-scale feature information and to provide the network with better feature representation capabilities, the Multi-scale Feature Fusion Module (MFFM), the details of which shown on the bottom side of Figure 1, is introduced here. First, the multi-scale feature map sizes are unified to the same size as the target feature map layer. Considering that the magnification scale between different layers is always $2^n (n \in \mathbb{Z})$, a 3×3 convolution layer with a stride of 2 and a 6×6 transposed convolution layer with a stride and padding of 2 are introduced for $/2$ downscaling and $\times 2$ upscaling processes, respectively. Moreover, for a larger magnification scale such as $/4$ or $\times 4$, double 3×3 convolution or 6×6 transposed convolution layers, etc., would be applied. After unifying all required feature map sizes, the multi-scale feature maps are concatenated to pass through a full connection layer, followed by the channel attention network, the details of which can be viewed in Figure 3a. Finally, the well-handled multi-scale feature information is added to the target feature map layer in the encoding stage.

3. Experiments

3.1. Dataset and Metrics

In our experiments, the model was trained on the CelebA [44] dataset and evaluated on the CelebA and Helen [45] datasets along with the real face images. In the data preprocessing phase, images were simply cropped to a size of 128×128 based on their center point and then treated as the ground truth. Afterwards, 16×16 LR face images were obtained from the ground truth images with a $/8$ downscaling bicubic operation. It should be noted that the model does not require any additional facial landmarking on the dataset for training. Then, we trained the model on 18,000 face images from the CelebA dataset and evaluated its performance on 1000 face images from the same dataset along with 50 face images from the Helen dataset. Moreover, to prove the flexibility of the model, we directly evaluated the model on the Helen dataset and real face images with the model trained on CelebA. It should be noted that the proposed deep learning method cannot be trained with a single image. Thus, sufficient face images must be provided should anyone want to train it on their own dataset.

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

3.2. Implementation Details

All experiments were conducted in PyTorch [61] with an GeForce RTX 4090 24 GB graphic card made by NVIDIA from Santa Clara, CA, USA. The proposed model was optimized by Adam with $\beta 1 = 0.9$, $\beta 2 = 0.99$, and a learning rate of 2×10^{-4}.

3.3. Ablation Studies

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

(1) Study on SCATM-T: The SCATM-T, which consists of an SCAGB-T and a CMTB, aims to extract and promote both local facial details and global facial structures. It is the first attempt to explore the potential of inner feature maps in reconstructing plausible face images in the transformer-based FSR area. To verify the effectiveness of SCATM-T and its components, we designed three test models by removing different module parts. The test results are shown in Table 1.

Table 1. Ablation study of the components in the proposed SCATM-T. Please note that the **best** results are emphasized with **bold** in the experiment part for better visualization.

SCAGB-T	CMTB	PSNR↑	SSIM↑	LPIPS↓
×	×	27.30	0.7833	0.2047
×	✓	27.55	0.7897	0.1831
✓	×	27.54	0.7891	0.1839
✓	✓	**27.63**	**0.7905**	**0.1797**

From the table, it can be observed that:

(a) The performance of SCATM-T without any components (i.e., removing SCATM-T) decreased dramatically. This is because the proposed model structure is much shallower without the SCATM-T, making it challenging to refine features. Moreover, without the processed fine-detailed feature maps from the SCATM-T, the Multi-scale Feature Fusion operation Module (MFFM) is greatly affected.

(b) The quantitative metrics of SCATM-T with one single component inside are better than the no-component one above, demonstrating that both SCAGB-T and CMTB can enhance the representation ability of the model. However, SCATM-T with only SCAGB-T cannot guide anything, while SCATM-T with only CMTB cannot focus on feature map parts that are rich in features, limiting its performance.

(c) Equipped with both of the carefully designed components SCAGB-T and CMTB, SCATM-T achieves the best performance in terms of all evaluation matrices, which proves that the combination of SCAGB-T and CMTB is complementary and can simultaneously promote both local facial details and global facial structures.

(2) Study on SCATM-B: The SCATM-B, which contains an SCAGB-B and a CMTB, aims to enhance the low-level encoded features. Due to their comparable inner structures, we conducted similar experiments to the ones reported in the "Study on SCATM-T" part, the results of which are shown in Table 2. Similar observations and conclusions to those made in the previous section can be derived. However, it is worth noting that the performance of the model without SCATM-B is better than that without SCATM-T. This is because the SCATM-T and the MFFM are more complementary compared with the relationship between the SCATM-B and the MFFM; with one removed, the performance of others will also be greatly impacted.

Moreover, we evaluated the influence on model performance of the number of SCATM-Bs, the results of which are shown in Table 3. From the table, it can be observed that the

model performs poorly without any SCATM-Bs, which proves that the SCATM-B plays a vital role in the model. Meanwhile, the performance of the model rises with the increase in the number of SCATM-Bs in a particular range. However, as the number of SCATM-Bs increases and exceeds four, the evaluation matrix change rate slows down and even decreases slightly. Additionally, the model size increases, which leads to an upsurge in the computational complexity of the model. Therefore, the number of SCATM-Bs is set to four for a good balance between model size and performance.

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

SCAGB-B	CMTB	PSNR↑	SSIM↑	LPIPS↓
×	×	27.50	0.7877	0.1897
×	✓	27.60	0.7902	0.1811
✓	×	27.59	0.7899	0.1817
✓	✓	**27.63**	**0.7905**	**0.1797**

Table 3. Performance comparison of different SCATM-B numbers in the proposed method.

SCATM-B Numbers	PSNR↑	SSIM↑	LPIPS↓
0	27.50	0.7877	0.1897
2	27.57	0.7900	0.1823
4	**27.63**	**0.7905**	**0.1797**
6	27.62	0.7903	0.1804

(3) Study on MFFM: MFFM is specially designed to further utilize the multi-scale feature information for better FSR performance. In this part, we design three different test models to demonstrate the effectiveness of the MFFM, the results of which are shown in Table 4.

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

Approaches	PSNR↑	SSIM↑	LPIPS↓
Not Applied	27.54	0.7885	0.1873
Only Add	27.58	0.7892	0.1838
Only Concat	27.59	0.7895	0.1832
Our MFFM	**27.63**	**0.7905**	**0.1797**

From the table, it can be observed that: (a) The model without the MFFM performs the worst, which proves the importance of multi-scale features in the image reconstruction process. (b) An add operation or a concatenate layer to fuse multi-scale features does benefit the performance of the FSR model; however, they are too simple to take responsibility for the multi-scale feature fusion part. (c) With the carefully designed MFFM, the model achieves the best performance in terms of PSNR, SSIM, and LPIPS, which proves that a suitable feature fusion strategy such as MFFM can benefit the image reconstruction process.

3.4. Comparison with the State-of-the-Art

To verify the superiority of the proposed method, in this section we compare our method with several state-of-the-art methods, including two GAN-based methods (SRResNet [16] and RCAN [29]), three attention-based methods (SPARNet [32], SISN [33], and IGAN [30]), and two pioneering transformer-based methods (SwinIR [37] and Uformer [28]) on the CelebA and Helen datasets along with real face images. Moreover, bicubic interpolation is applied as the baseline. All models were trained using the same CelebA dataset for fair comparison. Quantitative results are tabulated in Table 5.

(1) Comparison on CelebA dataset: Quantitative comparisons with other methods on the CelebA dataset are shown in Table 5. According to the table, the proposed method outperforms other competitive methods in terms of PSNR, SSIM, and LPIPS, which proves

that the proposed method has the advantage of recovering accurate and realistic face details. Moreover, we provide some test images from the CelebA dataset for visual comparisons, shown in Figure 5. From the first two samples in Figure 5, it can be seen that our method can better restore nose contours and eye details compared to other state-of-the-art methods, while avoiding unpleasant artifacts. This is due to the combination of SCAGB and CMT, which complement each other and can simultaneously enhance both local facial details and global facial structures. Furthermore, despite all the methods mentioned above achieving satisfactory evaluation metrics on the last sample, the visual results are not very compelling due to the inability to reconstruct eyeglasses. Therefore, adding face identifications in the FSR process that can help to identify lost details such as eyeglasses in the LR image represents a good opportunity for future work.

Table 5. 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 [16]	26.08	0.7502	0.2131	25.47	0.7828	0.2308
IGAN [30]	26.99	0.7801	0.2201	26.37	0.7996	0.2245
RCAN [29]	26.99	0.7796	0.2249	26.39	0.7965	0.2359
SISN [33]	26.85	0.7738	0.2337	26.33	0.7974	0.2322
SPARNet [32]	26.95	0.7794	0.2211	26.38	0.7953	0.2314
SwinIR [37]	27.15	0.7850	0.2162	26.48	0.7917	0.2413
Uformer [28]	27.33	0.7884	0.2040	26.67	0.8009	0.2063
Ours	**27.63**	**0.7905**	**0.1797**	**26.97**	**0.8069**	**0.1945**

Figure 5. Visual comparisons for ×8 SR on the CelebA test set. Please zoom in for better comparison.

(2) Comparison on Helen dataset: In this part, we directly evaluate the model on Helen datasets with the model trained on CelebA to prove the flexibility of the model.

Quantitative and visual comparisons with other methods on the Helen dataset are shown in Table 5 and Figure 6, respectively. According to the results, our proposed method has the advantage of recovering facial images both quantitatively and qualitatively, which demonstrates the robustness and stability of our method. However, it is worth noting that the methods mentioned above showed varying degrees of performance decrease on the Helen test set as compared to the CelebA test set. Therefore, digging into the inner differences among various datasets is a good choice for enhancing the generality of FSR methods, especially for real face image restoration.

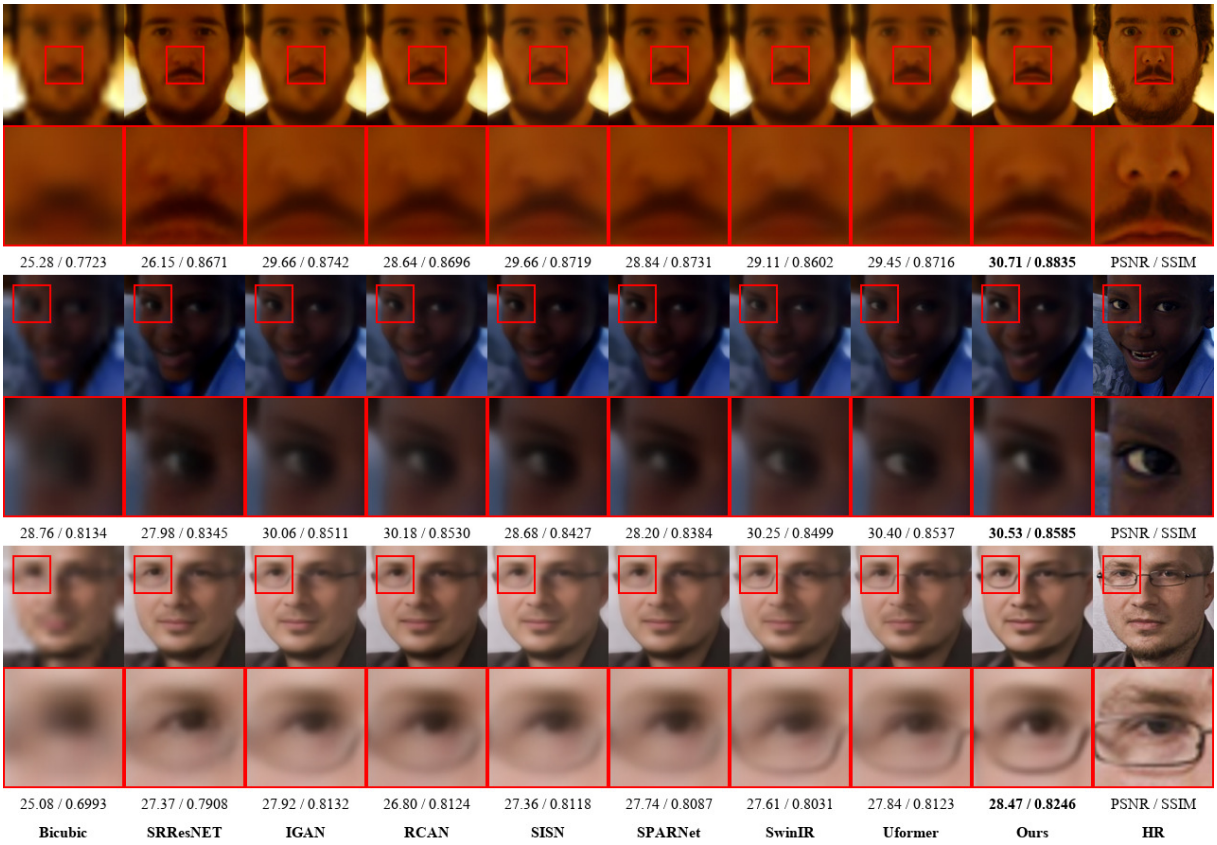

Figure 6. Visual comparisons for ×8 SR on the Helen test set. Please zoom in for better comparison.

(3) Comparison on real face images: Due to the fact that real face images are captured from a variety of complex environments that the aforementioned CelebA dataset cannot simulate, restoring face images from the real world is always a considerable challenge. Aiming to validate the effectiveness of our proposed method in real-world face images, we conducted experiments on low-quality real-world face images collected from the classic "I Love My Family" TV series. This is a popular sitcom made in the 1990s that suffers severe issues with low resolution due to the use of outdated imaging equipment, making it perfect for testing. Experiments were conducted with the aim of reconstruct more detailed facial images with appealing facial structures, the reconstructed results of which are shown in Figure 7. Benefiting from the guiding blocks and transformers in the proposed method, which complement each other and can simultaneously enhance both local facial details and global facial structures, our method achieves appealing performance with reasonable results compared to other state-of-the-art methods.

Figure 7. Visual comparison for ×8 SR on real face images. Please zoom in for better comparison.

3.5. JPEG Artifacts Analysis

To further prove the robustness of the proposed method on FSR tasks, we introduced JPEG artifacts to blur face images and tested these images without training another network based on JPEG artifacts. The images with JPEG artifacts were generated using the "JPEG Artifact Generator v1.0" [62] with the "Base JPEG compression" set to 0.2. We also selected two representative methods, Uformer and IGAN, as the comparative methods and the bicubic method as the baseline. The results are shown in Figure 8.

As the figure illustrates, all models, including ours, suffer performance degradation due to the lack of training with JPEG artifacts. However, our proposed method leveraging guiding blocks to preserve face features still manages to outperform the others. This is particularly evident in the fourth-row images.

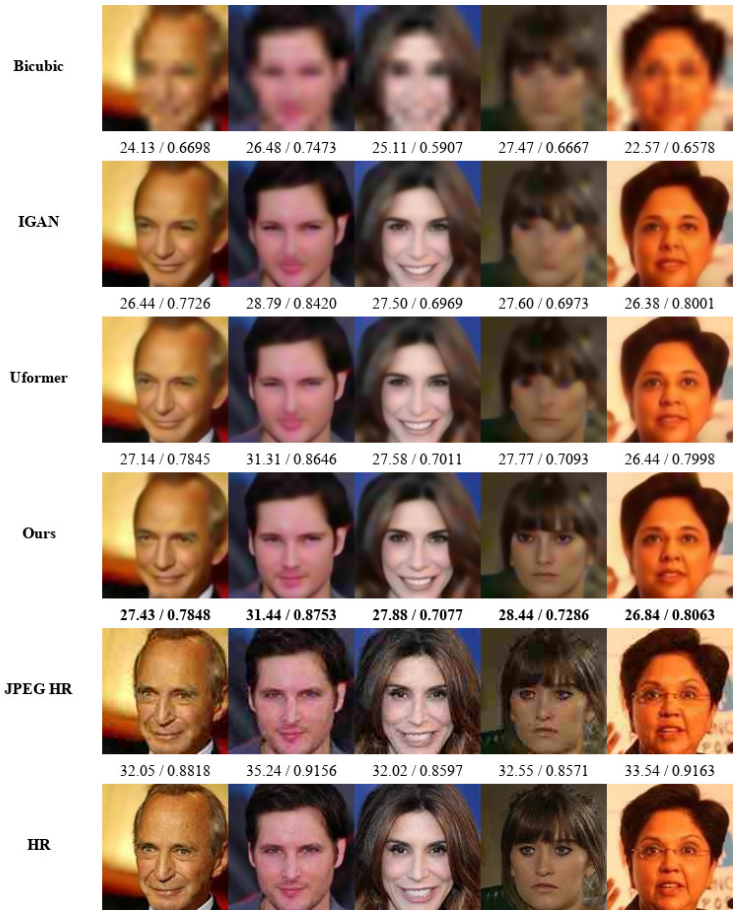

Figure 8. Visual comparison for ×8 SR on face images with JPEG artifacts; "JPEG HR" represents the original HR images with JPEG artifacts. Please zoom in for better comparison.

3.6. Model Complexity Analysis

The proposed method has proven its superior ability in quantitative and qualitative FSR performance based on the previous experiments. In this section, we compare the model performance, size, and execution time of our approach with other state-of-the-art methods. The results are shown in Figure 9. From the figure, it can be observed that our method achieves the best quantitative results while maintaining comparable execution time and model size, which makes it a possible choice for FSR tasks.

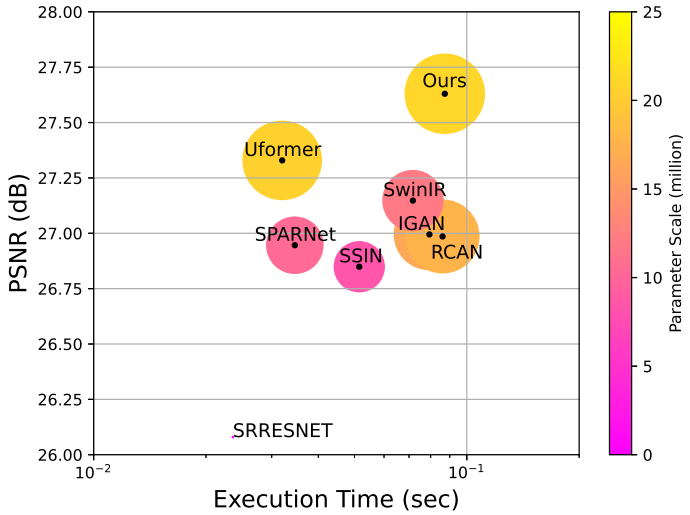

Figure 9. Model complexity scattergram for ×8 SR on the CelebA test set.

4. Conclusions

In contrast to existing transformer-based approaches that focus on introducing sophisticated structures, we propose a spatial-channel mutual attention-guided transformer network for face super-resolution. This is the first study to explore the potential of inner feature maps for reconstructing plausible face images in the transformer-based FSR area. The proposed method is a multi-scale connected encoder–decoder network. The primary component of the network is a Spatial-Channel Mutual Attention-guided Transformer Module (SCATM), which is composed of a Spatial-Channel Mutual Attention Guiding Block (SCAGB) and a Channel-wise Multi-head Transformer Block (CMTB). The SCATM on the top layer (SCATM-T) aims to promote both local facial details and global facial structures, while the SCATM on the bottom layer (SCATM-B) seeks to optimize the encoded low-level features. Unlike the usual spatial-wise transformers, which are limited to position-specific windows, the CMTB utilizes feature map channels to achieve image-size receptive fields. Moreover, we develop a Multi-scale Feature Fusion Module (MFFM), which fuses features from different scales for better restoration performance. Extensive experiments 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 No. 61572395 and 61675161).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both the CelebA [44] and Helen [45] datasets are available online.

Conflicts of Interest: The authors declare no conflicts 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]
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]
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]
9. 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] [PubMed]
10. 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.
11. 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.
12. 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.

13. 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.
14. 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]
15. 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.
16. 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.
17. 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.
18. Zhang, Y.; Yu, X.; Lu, X.; Liu, P. Pro-uigan: Progressive face hallucination from occluded thumbnails. *IEEE Trans. Image Process.* **2022**, *31*, 3236–3250. [CrossRef]
19. 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.
20. Gao, J.; Tang, N.; Zhang, D. A Multi-Scale Deep Back-Projection Backbone for Face Super-Resolution with Diffusion Models. *Appl. Sci.* **2023**, *13*, 8110. [CrossRef]
21. 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.
22. Zhang, M.L.; Ling, Q. Supervised pixel-wise GAN for face super-resolution. *IEEE Trans. Multimed.* **2021**, *23*, 1938–1950. [CrossRef]
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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]
33. 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.
34. 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.
35. 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]
36. Shi, A.; Ding, H. Underwater Image Super-Resolution via Dual-aware Integrated Network. *Appl. Sci.* **2023**, *13*, 12985. [CrossRef]
37. 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.
38. 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.

39. 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.
40. 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]
41. 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]
42. 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.
43. Leland, M.; John, H.; James, M. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, arXiv:1802.03426.
44. 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.
45. 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.
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] [PubMed]
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 Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4–9.
62. JPEG Artifact Generator. Available online: https://impliedchaos.github.io/artifactor.html (accessed on 1 May 2024).