applied sciences

MDPI

Article

Multi-Scale Feature Fusion and Structure-Preserving Network for Face Super-Resolution

Dingkang Yang [1], Yehua Wei [1], Chunwei Hu [1], Xin Yu [1], Cheng Sun [2], Sheng Wu [3] and Jin Zhang [1,3,*]

[1] College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China; ydkang99@163.com (D.Y.)
[2] School of Mathematics and Statistics, Hunan Normal University, Changsha 410081, China
[3] School of Computer and Communication Engineering, Changsha University of Science & Technology, Changsha 410114, China
* Correspondence: mail_zhangjin@163.com

Abstract: Deep convolutional neural networks have demonstrated significant performance improvements in face super-resolution tasks. However, many deep learning-based approaches tend to overlook the inherent structural information and feature correlation across different scales in face images, making the accurate recovery of face structure in low-resolution cases challenging. To address this, this paper proposes a method that fuses multi-scale features while preserving the facial structure. It introduces a novel multi-scale residual block (MSRB) to reconstruct key facial parts and structures from spatial and channel dimensions, and utilizes pyramid attention (PA) to exploit non-local self-similarity, improving the details of the reconstructed face. Feature Enhancement Modules (FEM) are employed in the upscale stage to refine and enhance current features using multi-scale features from previous stages. The experimental results on CelebA, Helen and LFW datasets provide evidence that our method achieves superior quantitative metrics compared to the baseline, the Peak Signal-to-Noise Ratio (PSNR) outperforms the baseline by 0.282 dB, 0.343 dB, and 0.336 dB. Furthermore, our method demonstrates improved visual performance on two additional no-reference datasets, Widerface and Webface.

Keywords: face super-resolution; structure-preservation; attention mechanism; feature fusion

Citation: 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. https://doi.org/10.3390/app13158928

Academic Editors: Mukesh Prasad, Pang-jo Chun, Xian Tao and Ali Braytee

Received: 7 July 2023
Revised: 28 July 2023
Accepted: 28 July 2023
Published: 3 August 2023

1. Introduction

Face Super-Resolution (FSR) is a subfield of image super-resolution that focuses on restoring Low-Resolution (LR) face images to High-Resolution (HR) counterparts using algorithms. It can increase the resolution of an LR face image of low quality and recover the details. Its purpose is to address the limitations posed by image acquisition systems or environmental conditions in many real-world scenarios [1]. By doing so, FSR aims to enhance the quality and improve the visibility of key facial features that are often degraded in LR images. This technique finds significant applications in various domains, including face vision tasks [2], public security, and other relevant fields [3,4].

Existing FSR methods can be categorized into two main groups: traditional methods and deep learning-based methods. Traditional methods can be further classified into interpolation-based methods, reconstruction-based methods, and learning-based methods. Notably, Baker et al. [5] were among the first to propose FSR methods that employed manual image priors and Gaussian image pyramids for face reconstruction. Since then, numerous significant advancements have been made in this field. During the initial stages of FSR research, researchers primarily focused on designing shallow learning-based methods that leveraged techniques such as local linear embedding [6], eigentransformation [7], and principal component analysis [8]. However, these methods demonstrated limited representation capabilities, making it challenging to generate high-quality HR face images.

In recent years, deep learning-based approaches leveraging Convolutional Neural Networks (CNNs) have made remarkable advancements in various computer vision tasks, including FSR. The limited information available in LR images makes the task of recovering high-resolution images unstable and non-unique. As a result, there are infinite possible high-resolution images corresponding to a single low-resolution image, making super-resolution challenging. To obtain stable and meaningful results, it is essential to employ appropriate optimization algorithms. Researchers have explored various methods, including deep learning networks and image priors, to address the ill-posed nature of super-resolution problems and enhance the performance and stability of super-resolution techniques. Neural networks, with their ability to extract deep features from images, are particularly meaningful for super-resolution tasks with insufficient feature information.

Zhang et al. [9] utilize an extremely deep residual network to learn feature representations of images. The increased depth enhances the network's capability for deep feature extraction, leading to the learning of more powerful and informative feature representations, thus achieving more accurate and detailed super-resolution reconstruction. On the other hand, Lai et al. [10] employ a Laplacian pyramid to decompose LR images into multiple scales. Deep convolutional neural networks are then applied at each scale to extract features, resulting in improved SR performance.

To exploit neural networks for recovering finer facial structures, researchers have proposed several studies to incorporate additional face prior information [11], such as facial parsing maps [12], facial heat maps [13], and facial landmarks [14,15], to guide the network during face reconstruction. Although the inclusion of prior information enhances network performance, it presents significant limitations. Firstly, acquiring and annotating the prior information requires additional effort, and obtaining reliable priors from low-resolution face images is challenging [16]. Secondly, inaccurate face prior information can lead to erroneous reconstruction outcomes. To address these challenges, alternative methods without prior knowledge have been proposed [16–19]. These approaches leverage the powerful representation and learning capabilities of neural networks and achieve excellent performance by enhancing facial feature representation or utilizing attention mechanisms to guide the network in recovering facial structures. By circumventing the reliance on explicit prior information, these methods offer a more robust and effective solution for FSR.

Although deep learning-based methods have significantly improved FSR tasks, they often overlook the intrinsic features of face images. Firstly, the face possesses fixed key parts and shapes, such as symmetrical features on both sides and consistent texture structures in the hair. However, these inherent features are underutilized, leading to a limited representation capability of the network. As a result, facial deformations occur in the reconstructed faces, making it challenging to recover accurate details. Additionally, LR face images exhibit non-local self-similarity, where similar feature patterns occur at different parts and scales. Unfortunately, existing CNN-based methods tend to neglect the exploitation of this self-similarity, hindering the recovery of fine facial details.

In light of the aforementioned issues, inspired by SPARNet [16], this paper presents an improved face super-resolution network that fuses multi-scale features while preserving structural information, without additional facial prior to supervising the reconstruction process. The primary contributions are as follows:

- We propose a novel multi-scale residual structure that effectively extracts features and integrates feature information from two branches: key face components and intrinsic image structure. This approach aims to restore facial images with improved structural clarity.
- To address feature loss resulting from network depth and maximize the utilization of information at different scales, we incorporate pyramid attention and feature enhancement module into the network architecture. These components effectively explore the correlations among features at various scales, compensating for the loss of information and aiding in the reconstruction of finer details.

- The proposed method is evaluated on five publicly available datasets and compared with other state-of-the-art methods, and the results show that the proposed method outperforms other methods in both qualitative and quantitative results.

2. Related Work

2.1. Face Super-Resolution

Deep neural networks have revolutionized FSR tasks by achieving remarkable advancements. Face super-resolution methods can be broadly categorized into two groups based on the utilization of facial priori information.

The first category comprises methods that leverage facial prior information. Yu et al. [13] introduced a two-branch multitasking network that utilizes underlying and intermediate feature information to enforce constraints on the facial components of LR faces, leading to improved preservation of the complete facial structure. Chen et al. [12] proposed FSRNet, which achieves impressive results in SR of very low-resolution faces by incorporating facial landmark heat maps into the feature map to resolve facial features, it aims to concentrate on the localization of facial signs, but it does not adequately consider the region around the sign's facial attributes. Kim et al. [14] devised a face attention loss based on facial landmark heat maps and employed a progressive training method for face reconstruction, while the process of extracting heat maps of facial landmarks greatly increases the training process. In order to restore fundamental facial features without distortion, Ma et al. [15] proposed an iterative collaboration approach that employs facial priors generated by their face keypoint recovery network to assist in FSR. However, the multi-stage iterative process also amplifies errors due to incorrect priors. To obtain SR images at arbitrary scales, Grm et al. [11] incorporated identity prior into the reconstruction process and employed multiple models for progressive cascade reconstruction of faces, but the cascading structure also makes the network larger In order to fully capture the potential of prior information and multi-scale information, Wang el al. [20] proposed a two-stage network, a ParsingNet is used to extract parsing map, which is then combined with LR image as input to the reconstruction network, crucial facial details and contours are restored by integrating multi-stage features. While utilizing additional facial prior information enhances network performance, it necessitates extra data annotation efforts, and the challenge of facial structure deformation arising from inaccurate face prior information persists.

The second category encompasses methods that do not rely on facial a priori information. In order to obtain sharper facial details, Tuzel et al. [18] introduced a method featuring a two-branch sub-network, where one branch focuses on global constraints to reconstruct face images while the other branch enhances local facial details. Chen et al. [16] proposed a novel spatial residual attention network that employs facial attention units to prioritize the recovery of important facial structures, thereby enhancing the network's representation capability, while it lacks inter-channel correlation interaction. To fully leverage the facial attribute information. Xin et al. [19] presented a facial attribute capsule network that transforms extracted facial feature maps into facial attribute capsules to obtain a complete facial structure, leveraging semantic and probabilistic information to generate corresponding high-resolution faces. To restore facial fine details and textures, Dastmalchi et al. [17] incorporated wavelet prediction into their network to predict missing wavelet details of facial images, resulting in finer face reconstructions. However, the incorporation of wavelet prediction also increases the training overhead of the network. These methods demonstrate the effectiveness of alternative strategies that exploit network architectures and attention mechanisms to recover facial details without relying on explicit facial prior information.

2.2. Attention Mechanism

Attention mechanisms have gained significant popularity in both high-level and low-level computer vision tasks, such as image recognition, target detection, image classification, image super-resolution, and image defogging. These mechanisms empower neural networks to dynamically adjust weights by leveraging attention graphs, enhancing network

performance through the emphasis on crucial features and suppression of less informative ones. Hu et al. [21] introduced Squeeze-and-Excitation Network (SE-Net), which employs a channel attention mechanism. By calculating the adaptive weights for each channel using a fully connected layer after converting them to single values through Global Average Pooling (GAP), SE-Net achieves improved feature representation and network efficiency. However, SE-Net overlooks the importance of spatial information. To address this limitation, Convolution Block Attention Module (CBAM) [22] enriches the attention graph by effectively combining spatial and channel attention, incorporating both GAP and global maximum pooling to enhance feature diversity. The utilization of attention mechanisms has demonstrated its value as a versatile tool in various vision tasks, enabling adaptive resource allocation within networks and promoting enhanced performance in visual understanding and reconstruction tasks.

In the realm of image super-resolution, Zhang et al. [9] introduced Residual Channel Attention Network (RCAN), which was the first to combine channel attention with image super-resolution tasks. Zhao et al. [23] incorporated a pixel attention mechanism into their network, aiming to enhance the network's reconstruction performance. Lu et al. [24] proposed an attention split face super-resolution network that encompasses an internal and external attention resolution network, resulting in improved texture details in the SR faces. Zeng et al. [25] propose a self-attention learning network for three-stage FSR, which fully explores the interdependence of both low and high spaces to enhance the features. These works have contributed to the advancement of attention-based techniques in the field of image super-resolution.

3. Methods

This section presents the improved model and its architecture, which comprises three key improvements built upon the original model (SPARNet). Firstly, an additional structure extraction branch is incorporated within the residual blocks to capture the intrinsic structural information of the face image. Secondly, pyramid attention is introduced during the feature extraction stage to exploit the inter-scale correlations among the multi-scale features. Lastly, a feature enhancement module is introduced in the upscaling stage to mitigate feature loss and enhance the overall quality of the reconstructed image. These improvements aim to enhance the model's capacity in capturing facial structure, utilizing multi-scale information, and preserving fine details throughout the super-resolution process.

3.1. Network Structure

As illustrated in Figure 1, the proposed residual network for fusing multi-scale features while preserving structure comprises three key components: a downscale module, a depth feature extraction module, and an upscale module. Notably, both the downscale module and upscale module are constructed using the Multi-Scale Residual Block (MSRB), which plays a crucial role in extracting and integrating features at different scales.

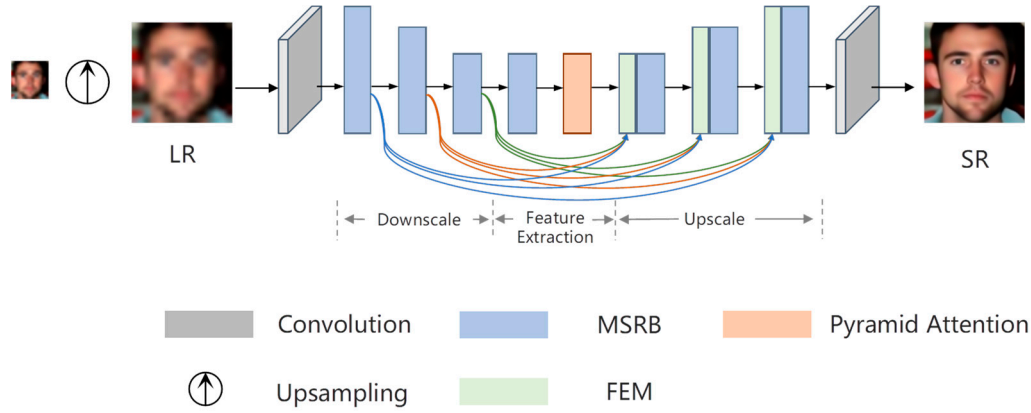


Figure 1. The architecture of the proposed network.

Since it is difficult to extract effective facial features directly from LR face images, a preprocessing step is employed to upsample the LR face image I_{LR} using Bicubic interpolation, bringing it to the same dimension as the HR image. Shallow features are then extracted from the upsampled image using a 3×3 convolutional layer, as depicted in Equation (1). Here, H_{sp} denotes the feature extraction process, H_{up} represents the upscale operation, and the resulting shallow features F_0 are subsequently fed as input into the downscale module.

$$F_0 = H_{sp}(H_{up}(I_{LR})) \tag{1}$$

In the downscale module, the face image undergoes gradual encoding and downsampling after the extraction of shallow features. This process involves applying three consecutive MSRBs to obtain the feature map F_1, which shares the same spatial dimensions as I_{LR}. To mitigate the issue of gradient attenuation with increasing network depth, the skip connections are incorporated within each Feature Enhancement Module (FEM) to fuse features from different scales. This fusion mechanism enables the feature map to retain a more comprehensive representation of facial information. The process can be described as follows:

$$F_1 = H_{down-i}(F_0) \tag{2}$$

The downscaled face image is passed through the depth feature extraction module, as indicated in Equation (3). To extract multi-scale features from F_1, both a MSRB and a Pyramid Attention (PA) module [26] are utilized, resulting in the generation of the deep facial feature F_2.

$$F_2 = H_{deep}(F_1) \tag{3}$$

Finally, the deep facial feature F_2 is passed into the upscale module to match the spatial dimension of HR. It is then fused with the output feature F_i from the ith MSRB in the downscale module using an additional feature enhancement module (FEM). This fusion process generates the final output feature F_{out}, which is subsequently subjected to a 3×3 convolutional layer to adjust the channel dimension to 3. The resulting SR image I_{SR} is then obtained. The operation process is as follows:

$$I_{SR} = H_{sp}(H_{up-i}(F_2) + F_i) \tag{4}$$

3.2. Multi-Scale Residual Block

We utilize a multi-scale residual block structure, as illustrated in Figure 2. The input feature F_{in} undergoes a pre-activation layer and two consecutive 3×3 convolutions, resulting in the output feature F_1. Subsequently, F_1 is fed into two separate branches: the Hourglass Block (HB) [27] and the Efficient Structure Extraction Module (ESEM), facilitating multi-scale feature extraction in both spatial and channel dimensions. The spatial dimension emphasizes key facial parts, such as the nose and eyes, which are rich in feature information. On the other hand, the channel dimension focuses on extracting structural and edge features from the face image. To mitigate artifacts in the reconstructed faces, additional 3×3 convolution layers with a stride of 2 are introduced in the downscale stage for downsampling and fusion of shallow features. Similarly, an additional 3×3 convolution layer is used in the upscale stage for upsampling and reconstruction of deep features. The resulting multi-scale features F_{scale} are multiplied with the convolutional extracted features and then added to the original features F_{in} to obtain the final features F_{out}.

The Hourglass structure was originally introduced in human pose estimation networks [27]. Its exceptional multi-scale feature extraction capability has yielded impressive results in various studies [12,15,16]. Recognizing the importance of independent facial components, such as the eyes and nose, in facial tasks, and considering the hourglass structure's ability to extract richer facial feature information and preserve high-frequency details, we adopt hourglass blocks for feature extraction of key facial parts. The facial features F_{in} are extracted in depth through a series of successive 3×3 convolutional layers, with each layer focusing on a different facial structure. To enhance convergence and

prevent gradient explosion, normalization layers and LeakyReLU activation functions are incorporated. Additionally, hopping connections are employed to facilitate the fusion of feature information from different layers, mitigating information loss during layer-by-layer feature transfer. Finally, the output features F_{out} are obtained for further processing.

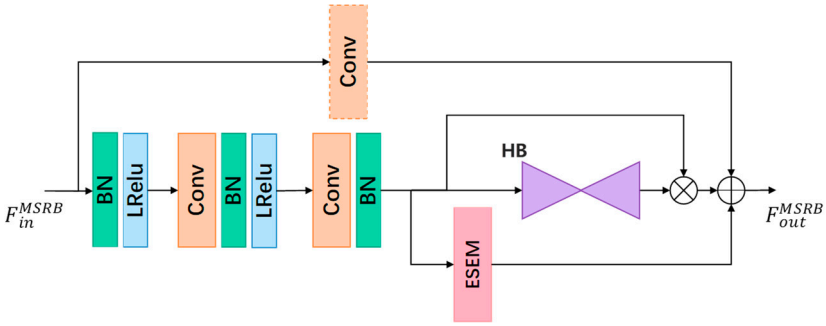


Figure 2. The architecture of MSRB.

3.3. Efficient Structure Extraction Module

Based on the mechanism of the human visual system, the human eyes are highly sensitive to edge information in images [28]. Consequently, edge information holds significant importance in visual perception [29], and the visual quality of an image is closely related to its edge information [30]. In the context of image super-resolution, the inherent structural information of low-resolution face images, which can be considered as a form of edge information, plays a crucial role. Building upon the inspiration from [31], we have developed an efficient structural information extraction module that extracts edge features and structural information from a multi-scale perspective. This approach allows us to fully leverage the structural feature correlations present in images and effectively recover more precise and distinct structural details in facial images.

Given the high sensitivity of edges and structures to scale variations, the input features F_{in} are divided into four segments with varying channel counts using three Cross-Convolution Blocks (CCBs) to extract features at different scales. The architecture of the CCBs is illustrated in Figure 3a. To progressively weigh the features in the channel dimension, we employ Cross Convolution [31] and Efficient Channel Attention (ECA) [32]. This weighting strategy enables the network to prioritize the restoration of facial structure and enhance the preservation of contour details. Subsequently, the features extracted by the CCB across different channels are fused, and the hierarchical features of the channels are consolidated through two convolutional layers with PRelu activation functions and ECA. This process yields the final output feature F_{out}, which can be expressed as follows:

$$F_{out} = H_{fuse}\left(\left[HC_3^0, HC_2^1, HC_1^2, HC_0^3\right]\right) \tag{5}$$

where HC_j^k denotes the kth group after the jth CCB and H_{fuse} represents the fusion operation.

In contrast to conventional convolutions, cross-convolution employs two distinct vertical filters $k_{1\times m}$ and $k_{m\times 1}$, for asymmetric filtering of the features. By leveraging gradient information in both the vertical and horizontal directions, these filters emphasize the structural characteristics of edge contours. The edge information is subsequently fused and reinforced to obtain the final output feature F_{out}^{Cross}. The process can be expressed as follows, where b represents the bias term:

$$F_{out}^{Cross} = Conv(k_{1\times m}, F_{in}^{CCB}) + Conv\left(k_{m\times 1}, F_{in}^{CCB}\right) + b \tag{6}$$

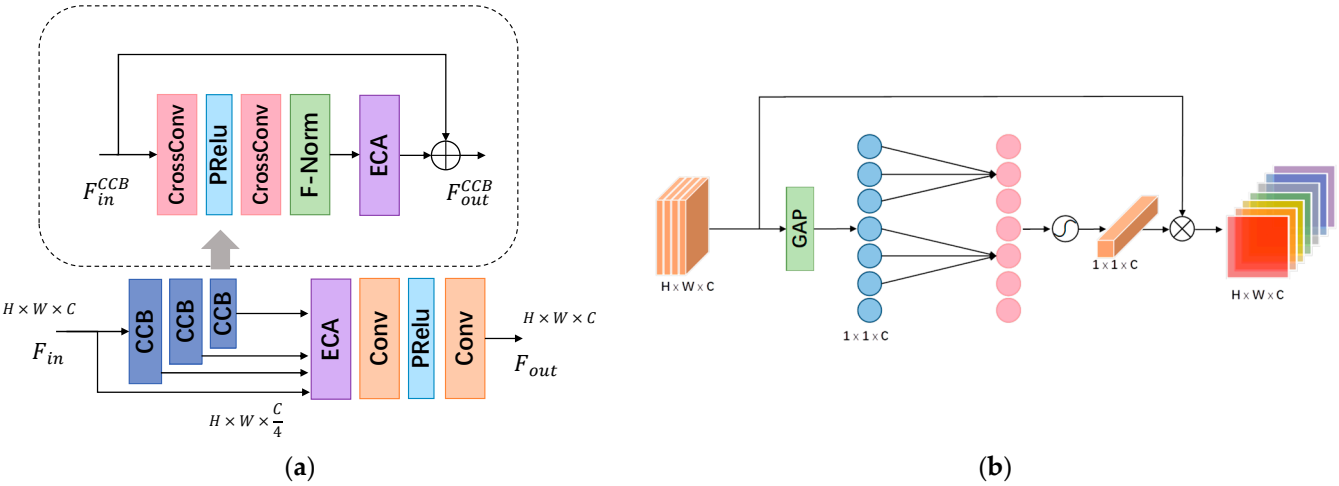


Figure 3. (**a**) The architecture of efficient structure extraction module. (**b**) Illustration of efficient channel attention.

The structure of ECA, as depicted in Figure 3b, offers a more efficient and lightweight alternative to SE-Net [21]. It overcomes the challenge of diminished learning caused by dimensionality reduction. ECA starts by performing channel-wise averaging of the features. The weight for each channel is then computed by jointly considering the aggregated features of that channel and its neighboring channels. This is accomplished using a one-dimensional convolution with a kernel size of k, which facilitates inter-channel feature interactions. The calculation can be expressed as follows:

$$w = Sigmoid(Conv_{1\times k}(y)) \tag{7}$$

For a given channel dimension C, the size k of the convolution kernel can be calculated by Equation (8).

$$k = \left| \frac{log_2(C)+1}{2} \right|_{odd} \tag{8}$$

where $|t|_{odd}$ indicates the nearest odd number. Finally, the information of each channel is fused to obtain the output features, which are used to enhance the information exchange between channels at different levels by capturing the local cross-channel information.

3.4. Pyramid Attention

The presence of pattern repeatability in images has been established as a crucial factor in image restoration. Self-similarity, as a form of repeatability, refers to the occurrence of small but similar patterns at different locations and scales within an image. It serves as valuable prior information for image restoration algorithms [26]. However, most existing deep neural network-based face super-resolution methods employ attention mechanisms that solely focus on the same scale, neglecting the full potential of self-similarity in face images. Given the symmetric nature of faces, facial images exhibit various repeatable and similar structures. Therefore, capturing rich self-similarity information can significantly enhance the performance of super-resolution reconstruction and reduce the model's reliance on external datasets. Based on this, we introduce a pyramid attention that captures feature correspondences at a distance from a multi-scale feature pyramid. This mechanism enhances the interaction between nearby and distant features, resulting in the recovery of finer facial details.

The pyramid attention structure, as depicted in Figure 4, incorporates a scale-independent attention module called Scale Agnostic Attention (SAA) to capture the intrinsic correlation of the multi-scale feature maps. To input the feature F_{in}^{PA} into the pyramid attention, a five-layer feature pyramid with varying scales is constructed using Bicubic interpolation. The pixel feature at layer i is represented as x_i, and its corresponding mapping

across different scales is denoted as K. Additionally, the pixel feature at the subsequent level is downsampled to x_i and referred to as z^j. This downsampling operation is beneficial for reducing image noise and effectively preserving the original structural information even after scaling down. Hence, this operation contributes to noise reduction in the image while maintaining the integrity of the original structural details.

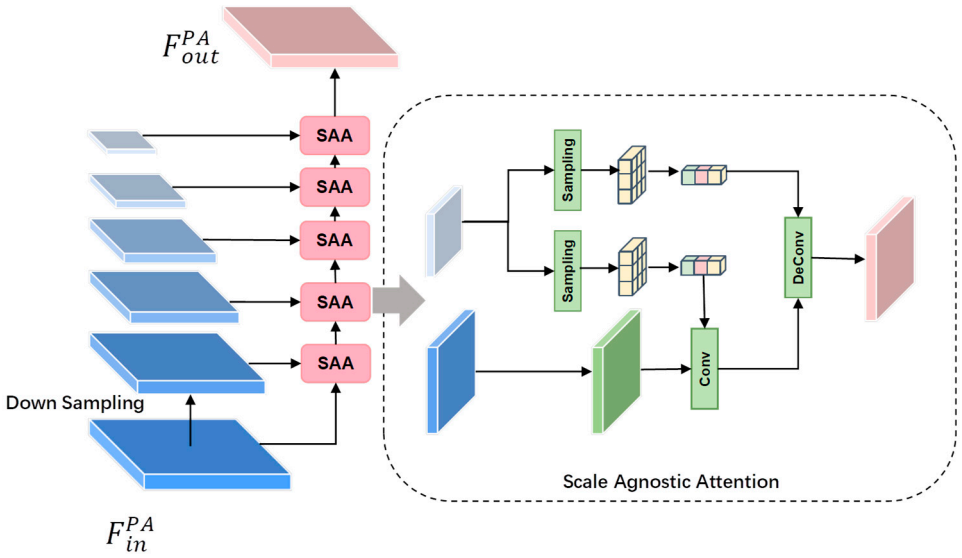


Figure 4. The architecture of Pyramid Attention.

For each scale feature, two image blocks are extracted: block f for reconstruction and block g for matching, corresponding to the block feature W_f and block feature W_g, respectively. Regarding the block features W_g, channel concatenation is applied as weights, and convolutional matching is performed with the input feature x_i. This process yields the self-similarity feature map, which is obtained through a Softmax operation. On the other hand, the block feature W_f, channel concatenation is directly applied as transposed convolution kernel weights. The correlation f_{out}^i between the two scales is obtained after performing the deconvolution operation with the self-similarity feature map. Finally, the pyramid attention is obtained by summing up the contributions from all positions of each scale, as expressed by the following Equation:

$$f_{out}^i = \frac{1}{\sigma(x,K)} \sum_{z \in K} \sum_{j \in Z} \varnothing\left(x_{\delta(r)}^i, z_{\delta(r)}^j\right) \theta\left(z^j\right) \tag{9}$$

where $\varnothing$ represents the Gaussian embedding function for calculating the similarity between two-pixel features, θ denotes the linear feature transformation function, σ represents a scalar function for normalizing the features, and $\delta_{(r)}$ denotes nearest neighbor similarity constraint, where two-pixel features are considered highly correlated when and only when they are also highly similar to their corresponding neighbors, which helps to make the network focus more on feature-related regions while suppressing irrelevant regions and is used to improve the robustness of the feature matching process.

3.5. Feature Enhancement Module

High-resolution image features are known to possess more precise spatial information, while low-resolution image features contain richer contextual information [20]. Given the complementary nature of high and low resolutions, it is crucial to leverage features at different scales effectively. To achieve this, we introduce a Feature Enhancement Module that refines the features during the upscale stage using the scale-specific features generated in the downscale stage. The FEM is built with reference to [20]. We strategically place

the FEM before each MSRB to enhance feature details within each stage and facilitate information exchange between features at different scales.

The structure of FEM is depicted in Figure 5. Its input comprises four components: the current feature F and the previous features F_1, F_2 and F_3 obtained during the downscale phase. These parts are individually fused in the Refine Block to compensate for missing information in F. Subsequently, the output results of the three parts are weighted using ECA to incorporate channel dimension information. Finally, the fused refinement feature F_r is obtained by adding the weighted outputs of F and the channel attention. For instance, when fusing F and F_3, the refinement block initially aligns F and F_3 to the same dimension through nearest-neighbor interpolation or average pooling. It then calculates the feature difference F_d between them and updates the current feature F using F_d to compensate for feature loss due to network deepening. This process can be expressed through Equations (10) and (11).

$$F_d = F - F_3 \tag{10}$$

$$F_r = H_p(F_d) + F \tag{11}$$

where H_p represents the projection function constructed by two consecutive convolution layers.

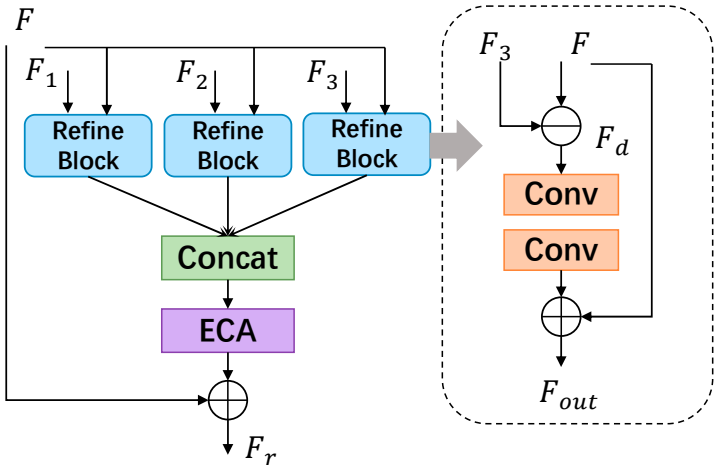


Figure 5. The architecture of Feature Enhancement Module.

3.6. Loss Function

To minimize the discrepancy between the reconstructed image and the original image, this study employs the $L1$ loss to optimize the network parameters during training. The $L1$ loss is less sensitive to outliers, thereby promoting the preservation of high-frequency features and facilitating a smoother training process to prevent gradient explosion. Given an image pair consisting of LR and HR, the calculation process of the $L1$ loss is as follows:

$$L_{Pixel} = \frac{1}{N}\sum_{k=1}^{N} \left\| H_{SR}(I_{LR}^{K} - I_{HR}^{K}) \right\| \tag{12}$$

where k denotes the kth image pair trained, n denotes the image pixel size, H_{SR} is the proposed network, and I_{LR} and I_{LR} represent the LR and HR face images, respectively. The optimized network undergoes continuous training to minimize the difference between the SR image and the original image, aiming to achieve a high level of resemblance between them.

4. Experiment and Results

4.1. Experiment Settings

We conducted extensive experiments on five datasets: CelebA [33], Helen [34], LFW [35], Widerface [36], and WebFace provided by [37]. Among them, Widerface and Webface are real-

world datasets that exhibit unknown complex degradation processes. CelebA serves as the training set, consisting of 202,599 face images belonging to 10,177 individuals with 40 attribute classes. Similar to [16], we performed preprocessing on the face images. Specifically, we selected 158,026 face images from CelebA, ensuring a balanced distribution across different ages and genders, as the training set. To maximize the utilization of training data, we applied data augmentation techniques such as random scaling, mirroring, horizontal flipping, and random rotation (90°, 180°, or 270°) to enhance the training samples. For the testing phase, we used CelebA, Helen, LFW, Widerface, and WebFace datasets, ensuring that the face images in the training and test sets are mutually exclusive. Following DIC [15], we employed MTCNN [38] for face detection, and the face regions were cropped from the center without pre-alignment to obtain HR images of size 128×128 through Bicubic interpolation. Subsequently, these HR images were downsampled to obtain LR images of size 16×16.

All experiments were conducted in a virtual environment using Python 3.7, CUDA 11.1, and PyTorch 1.10 on a RTX 3070 GPU. The batch size was set to 16, with the momentum coefficient β_1 set to 0.9 and the squared momentum coefficient β_2 set to 0.99. We employed the Adam optimizer with a fixed learning rate of 10^{-5} and a scale factor of $\times 8$.

4.2. Evaluation Metrics

We utilize two evaluation metrics to assess the quality of SR images: PSNR and Structural Similarity (SSIM). Additionally, for evaluating the naturalness of restored face images on two real-world test sets, we employ the widely used Natural Image Quality Evaluator (NIQE). PSNR measures the difference between the SR image and the HR image by calculating the pixel mean square error between the two images. A higher PSNR value, expressed in dB, indicates less distortion in the reconstructed image. The PSNR is calculated as follows:

$$PSNR = 10log_{10}\left(\frac{(2^8-1)^2}{f_{MSE}(y,y')}\right) \tag{13}$$

$$f_{MSE} = \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left(y_{i,j} - y'_{i,j}\right) \tag{14}$$

where y indicates the HR image, y' indicates the SR image, and H and W are the corresponding height and width of the image, respectively.

SSIM evaluates the similarity between the SR image and the HR image based on various factors such as brightness, contrast, and structure. It provides a value within the range of [0, 1], where a higher value indicates a higher similarity and better reconstruction effect between the two images. The calculation process of SSIM can be expressed as follows:

$$SSIM = \frac{(2\mu_x\mu_y+c_1)(\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{15}$$

where c_1 and c_2 are constants, μ_x and μ_y are the means of HR images and SR images, μ_x^2 and μ_x^2 are the standard deviations of HR images and SR images, respectively, and σ_{xy} is the covariance.

NIQE is a no-reference metric that measures the difference between two multivariate Gaussian models: one for natural images and the other for evaluated images without ground truth. It utilizes quality-aware features from the natural scene statistic model. A smaller NIQE value indicates higher visual quality. The calculation process is as follows:

$$D(v_1, v_2, \varepsilon_1, \varepsilon_2) = \sqrt{((v_1 - v_2)^T\left(\frac{\varepsilon_1+\varepsilon_2}{2}\right)^{-1}(v_1 - v_2))} \tag{16}$$

where v_1, v_2, ε_1 and ε_2 represent the mean vectors and covariance matrices of the natural images and the evaluated images.

4.3. Ablation Study

This section presents ablation experiments conducted to analyze and verify the effectiveness of each component in our model. The components used in the experiments include ESEM, PA, and FEM. The experiments are performed on the CelebA [33] training set, and the Helen test set. All parameters are kept consistent with the original network.

A total of eight sets of experiments were designed by combining the three modules. Model 1 does not incorporate any module, models 2 to 7 incorporate only one or two modules, and model 8 incorporates all modules. The final results are compared quantitatively to evaluate the impact of each component on the performance.

After training the networks used for ablation separately, the objective evaluation results are presented in Table 1. Comparing model 1, which does not have any module added, with the other models, it is evident that the addition of modules positively impacts the network performance. Among the added modules, the inclusion of ESEM yields the most significant improvement. The models incorporating two modules demonstrate better reconstruction results compared to those with only one module, with model 5, incorporating both ESEM and PA, exhibiting particularly notable effects. The PSNR value of Model 8, after incorporating all three modules, reaches 27.744 dB, which is 0.355 dB higher than that of Model 1. This confirms the effectiveness of the proposed method. It can be concluded that the addition of ESEM enhances the interaction ability of internal features across channels at each network level and improves the network's capability to extract structural information from the image, resulting in clearer contours and details in the reconstructed face. The incorporation of PA enhances the network's capability to mine the intrinsic correlations within multi-scale feature maps. Furthermore, the inclusion of the FEM enables the reconstruction process to retain rich spatial and contextual information across different scale features, thereby enhancing the quality of the SR image.

Table 1. Results of ablation experiments on Helen.

Models	ESEM	PA	FEM	PSNR	SSIM
1				27.389	0.817
2			√	27.533	0.822
3		√		27.425	0.819
4	√			27.585	0.823
5	√	√		27.612	0.825
6	√		√	27.684	0.828
7		√	√	27.579	0.823
8	√	√	√	27.744	0.830

To assess the impact of each module on the model's convergence, we present the comparison results of the training curves in Figure 6. The horizontal axis represents the number of iterations, while the vertical axis indicates the PSNR values. Notably, the training process of the model exhibits a smooth trajectory, characterized by minimal oscillation. Upon reaching approximately 30K iterations, the model essentially converges. Furthermore, it is evident that the model incorporating all three modules outperforms the other configurations, as indicated by significantly higher PSNR values.

4.4. Comparison with Other Methods

To validate the effectiveness of the model, this study compares it with several existing FSR methods, including the traditional interpolation method Bicubic, as well as DIC [15], KDFSRNet [39], SISN [24], WIPA [17], and SPARNet [16]. Since SISN and KDFSRNet utilize the same CelebA training set as ours, we utilize the pre-trained weights for testing on our test set. For all other models, we train and test them using the official open-source code and the datasets used in this study. The results are subjected to both qualitative and quantitative analysis to provide a comprehensive evaluation.

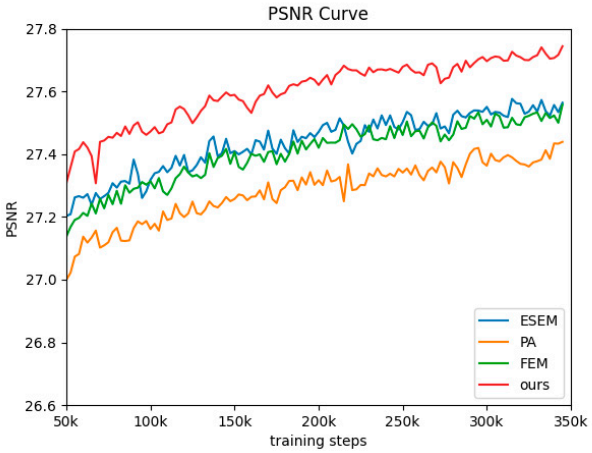


Figure 6. Curves of the training procedure.

Table 2 lists the PSNR and SSIM values of our method and other compared methods tested on the datasets with the scale factor of ×8.

Table 2. Quantitative comparison with other methods on CelebA, Helen and LFW for scale factor ×8.

Methods	CelebA		Helen		LFW	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Bicubic	23.572	0.637	24.138	0.681	24.893	0.693
DIC [15]	27.155	0.789	26.790	0.797	28.478	0.815
KDFSRNet [39]	27.245	0.793	26.515	0.788	-	-
SISN [24]	26.146	0.750	26.271	0.776	27.744	0.791
WIPA [17]	27.025	0.786	26.945	0.806	28.545	0.818
SPARNet [16]	27.167	0.789	27.401	0.818	28.829	0.825
Ours	27.449	0.800	27.744	0.830	29.165	0.838

Based on the experimental results presented in Table 2, it is evident that our model outperforms other methods in terms of achieving optimal PSNR and SSIM values on the CelebA, Helen, and LFW test sets. Specifically, the highest PSNR value of 27.744 dB is attained on the Helen test set, surpassing SPARNet [16], DIC [15], and KDFSRNet [39] by 0.343 dB, 0.954 dB, and 1.1229 dB, respectively. The corresponding SSIM value of 0.830 is also notably higher by 0.012, 0.033, and 0.042, respectively. On the LFW test set, our model achieves PSNR and SSIM values of 29.165 dB and 0.838, respectively, which are 0.687 dB and 0.336 dB better than [15,16], while exhibiting a higher SSIM value by 0.023 and 0.013, respectively. This is because SPARNet emphasizes spatial information while neglecting the interaction between feature channels, leading to a reduction in the quality of the SR images due to the underutilization of channel information. On the other hand, the DIC method lacks sufficient effective LR prior information, resulting in errors between the SR images and the HR images. KDFSRNet directly explores prior knowledge during the training phase, propagating from the teacher network to the student network, which introduces some dataset dependency. These experimental results highlight the superior performance of ours compared to the other methods in producing high-quality SR results.

In order to gain a more intuitive understanding of the face reconstruction results, Figure 7 presents the SR results of the test set images using a scale factor of 8. Upon observation, it becomes apparent that while other compared methods also achieve satisfactory face reconstructions, our model produces results with a clearer face outline and more comprehensive facial structure. This improvement can be attributed to the preservation of crucial face structure information during the reconstruction process. In contrast, the face images obtained using the Bicubic interpolation method lack significant detail and appear excessively blurred due to a simplistic zooming approach. The DIC [15] method,

which incorporates face-prior knowledge to facilitate network learning, exhibits noticeable deformations in the eye position and blurred edges. The utilization of incorrect priors in LR face images leads to substantial errors in the reconstruction, which are further amplified by the multi-stage iterative process employed in it. Although WIPA [17], based on generative adversarial networks, demonstrates more realistic facial textures, it still exhibits artifacts that impair the visual effect and poorly recover essential eye positions.

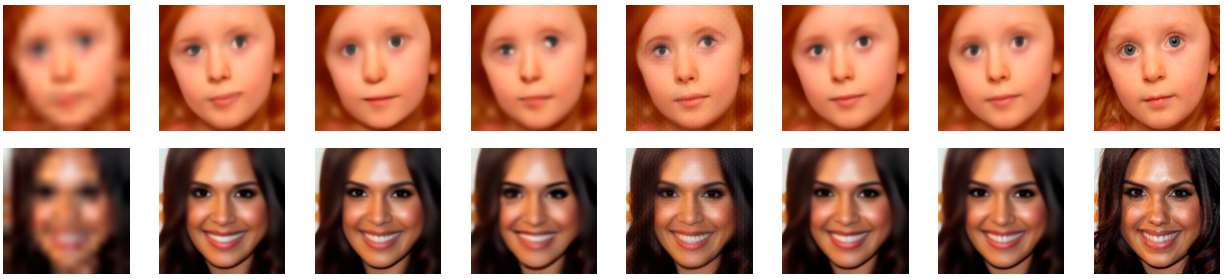


Figure 7. Visual quality comparison with other methods. (**a**) Bicubic; (**b**) DIC; (**c**) KDFSRNet; (**d**) SISN; (**e**) WIPA; (**f**) SPARNet; (**g**) Ours; (**h**) HR.

Figure 8 presents a comparison of local details for different methods. While KDFSRNet [39] and SPARNet [16] excel in overall image reconstruction, they struggle to fully recover certain structural details, resulting in an overly smooth appearance. A specific example is the interdental region highlighted in Figure 8. On the other hand, SISN [24] incorporates attention separation between channels but fails to prioritize individual facial key parts, leading to a lack of clarity in the reconstructed face images. In contrast, our model exhibits richer facial textures (such as teeth, eyes, and hair) and well-defined edge contours, while minimizing distortion. These results demonstrate the effectiveness of our model in both face structure recovery and detail reconstruction.

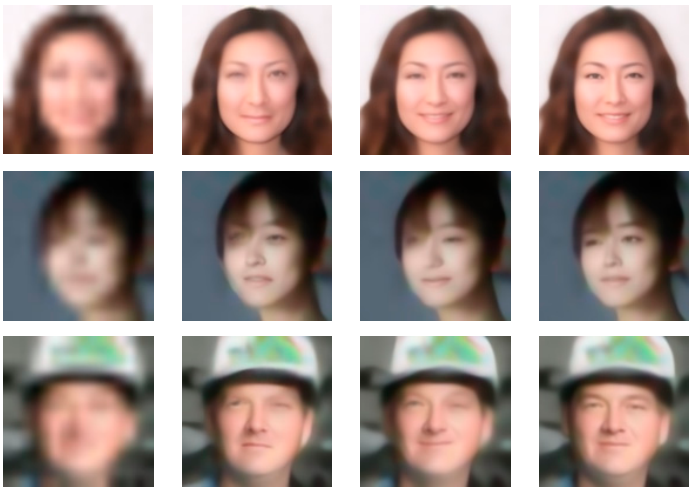
Figure 8. Visual quality comparison of details with other methods. (**a**) Bicubic; (**b**) DIC; (**c**) KDFSRNet; (**d**) SISN; (**e**) WIPA; (**f**) SPARNet; (**g**) Ours; (**h**) HR.

The aforementioned experiments were conducted on datasets that represent an unrestricted environment, which may differ significantly from real-world low-resolution face images. To bridge this gap, we applied the proposed method and CNN-based methods to real-world faces with unknown complex degradation, sourced from WiderFace [36] and WebFace [37]. In order to assess the naturalness of the recovered face images, we employed the NIQE since high-resolution reference images were not available. A lower value indicates a better reconstruction. The results of this evaluation are presented in Table 3, alongside comparisons with existing methods.

Table 3. Quantitative comparison of NIQE with other methods on WiderFace and WebFace.

Methods	WiderFace	WebFace
Bicubic	13.6051	13.7006
DIC [15]	12.1322	12.1569
SPARNet [16]	12.1075	12.1780
Ours	11.7509	11.8835

Figure 9 presents a comparative analysis of the reconstruction results obtained by various methods on real-world datasets. It is evident that all methods experience degradation in performance when confronted with unknown complex degradation, resulting in smoothing and blurring of the reconstructed details. In contrast, our model demonstrates clear visual superiority and delivers more detailed facial reconstruction outcomes. It effectively captures and reproduces high-frequency textures and intricate details, thereby enhancing the recovery of key facial components.



(**a**) (**b**) (**c**) (**d**)

Figure 9. Visual quality comparison on the real-world face images. (**a**) Bicubic; (**b**) DIC; (**c**) SPARNet; (**d**) Ours.

5. Conclusions

This paper introduces a novel network for face super-resolution that incorporates multi-scale features while preserving facial structure. The proposed approach leads to improved preservation of facial structure and edge details in the SR images, consequently enhancing the overall quality of the reconstructed images. However, it is essential to acknowledge its current limitation. The focus of the study is currently restricted to face super-resolution at the scale of $\times 8$, and the network involves a higher number of parameters compared to some lightweight alternatives. As part of our future research direction, we plan to explore FSR at arbitrary scales. This will allow us to cater to a wider range of practical applications with varying resolution requirements. Additionally, we will investigate and develop lightweight network architectures to strike a balance between computational efficiency and high-quality SR results.

Author Contributions: Conceptualization, D.Y.; methodology, D.Y.; software, D.Y.; validation, D.Y., S.W. and C.H.; formal analysis, D.Y. and J.Z.; investigation, X.Y. and C.S.; writing—original draft preparation, D.Y.; writing—review and editing, D.Y. and C.H.; visualization, D.Y. and C.S.; supervision, Y.W.; project administration, D.Y.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Hunan Province (2021JJ30456, 2021JJ30374), the Open Research Project of the State Key Laboratory of Industrial Control Technology (No. ICT2022B60), the National Defense Science and Technology Key Laboratory Fund Project (2021-KJWPDL-17), the National Natural Science Foundation of China (61972055), and the Research Foundation of Education Bureau of Hunan Province, China (20C0030).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 8 September 2022), reference number [33].

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

References

1. Jiang, J.; Wang, C.; Liu, X.; Ma, J. Deep Learning-based Face Super-resolution: A Survey. *ACM Comput. Surv. CSUR* **2021**, *55*, 13. [CrossRef]
2. Wang, G.Q.; Li, J.Y.; Xie, J.; Xu, J.; Yang, B. EfficientSRFace: An Efficient Network with Super-Resolution Enhancement for Accurate Face Detection. *arXiv* **2023**, arXiv:2306.02277.
3. Lau, C.P.; Castillo, C.D.; Chellappa, R. Atfacegan: Single face semantic aware image restoration and recognition from atmospheric turbulence. *IEEE Trans. Biom. Behav. Identity Sci.* **2021**, *3*, 240–251. [CrossRef]
4. Zheng, X.; Guo, Y.; Huang, H.; Li, Y.; He, R. A survey of deep facial attribute analysis. *Int. J. Comput. Vis.* **2020**, *128*, 2002–2034. [CrossRef]
5. Baker, S.; Kanade, T. Hallucinating faces. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 28–30 March 2000; pp. 83–88.
6. Chang, H.; Yeung, D.-Y.; Xiong, Y. Super-resolution through neighbor embedding. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; Volume I.
7. Wang, X.; Tang, X. Hallucinating face by eigentransformation. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2005**, *35*, 425–434. [CrossRef]
8. Chakrabarti, A.; Rajagopalan, A.; Chellappa, R. Super-resolution of face images using kernel PCA-based prior. *IEEE Trans. Multimed.* **2007**, *9*, 888–892. [CrossRef]
9. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
10. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5835–5843.
11. Grm, K.; Scheirer, W.J.; Štruc, V. Face hallucination using cascaded super-resolution and identity priors. *IEEE Trans. Image Process.* **2019**, *29*, 2150–2165. [CrossRef] [PubMed]
12. 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.
13. Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; Hartley, R. Face super-resolution guided by facial component heatmaps. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 217–233.
14. Kim, D.; Kim, M.; Kwon, G.; Kim, D.-S. Progressive face super-resolution via attention to facial landmark. *arXiv* **2019**, arXiv:1908.08239.
15. 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/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5569–5578.
16. 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] [PubMed]
17. Dastmalchi, H.; Aghaeinia, H. Super-resolution of very low-resolution face images with a wavelet integrated, identity preserving, adversarial network. *Signal Process. Image Commun.* **2022**, *107*, 116755. [CrossRef]
18. Tuzel, O.; Taguchi, Y.; Hershey, J.R. Global-local face upsampling network. *arXiv* **2016**, arXiv:1603.07235.
19. Xin, J.; Wang, N.; Jiang, X.; Li, J.; Gao, X.; Li, Z. Facial attribute capsules for noise face super resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12476–12483.
20. Wang, C.; Jiang, J.; Zhong, Z.; Zhai, D.; Liu, X. Super-Resolving Face Image by Facial Parsing Information. *IEEE Trans. Biom. Behav. Identity Sci.* **2023**. *early access*. [CrossRef]
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

22. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.-S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
23. Zhao, H.; Kong, X.; He, J.; Qiao, Y.; Dong, C. Efficient image super-resolution using pixel attention. In *Computer Vision–ECCV 2020 Workshops, Proceedings of the ECCV European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Part III 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 56–72.
24. 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 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 5501–5509.
25. Zeng, K.; Wang, Z.; Lu, T.; Chen, J.; Wang, J.; Xiong, Z. Self-attention learning network for face super-resolution. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **2023**, *160*, 164–174. [CrossRef] [PubMed]
26. Mei, Y.; Fan, Y.; Zhang, Y.; Yu, J.; Zhou, Y.; Liu, D.; Fu, Y.; Huang, T.S.; Shi, H. Pyramid attention networks for image restoration. *arXiv* **2020**, arXiv:2004.13824.
27. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Part VIII 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 483–499.
28. Ran, X.; Farvardin, N. A perceptually motivated three-component image model-Part I: Description of the model. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **1995**, *4*, 401–415. [CrossRef] [PubMed]
29. Wang, H.; Hu, X.; Zhao, X.; Zhang, Y. Wide Weighted Attention Multi-Scale Network for Accurate MR Image Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 962–975. [CrossRef]
30. Mandal, S.; Sao, A.K. Edge preserving single image super resolution in sparse environment. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 967–971.
31. Liu, Y.; Jia, Q.; Fan, X.; Wang, S.; Ma, S.; Gao, W. Cross-SRN: Structure-Preserving Super-Resolution Network With Cross Convolution. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 4927–4939. [CrossRef]
32. Wang, Q.; Wu, B.; Zhu, P.F.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2019; pp. 11531–11539.
33. 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.
34. Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; Huang, T.S. Interactive facial feature localization. In *Computer Vision–ECCV 2012, Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012*; Part III 12; Springer: Berlin/Heidelberg, Germany, 2012; pp. 679–692.
35. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Marseille, France, 1–18 September 2008.
36. Yang, S.; Luo, P.; Loy, C.-C.; Tang, X. Wider face: A face detection benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5525–5533.
37. Hou, H.; Xu, J.; Hou, Y.; Hu, X.; Wei, B.; Shen, D. Semi-cycled generative adversarial networks for real-world face super-resolution. *IEEE Trans. Image Process.* **2023**, *32*, 1184–1199. [CrossRef] [PubMed]
38. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
39. Wang, C.; Jiang, J.; Zhong, Z.; Liu, X. Propagating Facial Prior Knowledge for Multitask Learning in Face Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7317–7331. [CrossRef]