Spatial-Frequency Mutual Learning for Face Super-Resolution

Chenyang Wang, Junjun Jiang*, Zhiwei Zhong, Xianming Liu
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{wangchy02,jiangjunjun,zhwzhong,csxm}@hit.edu.cn

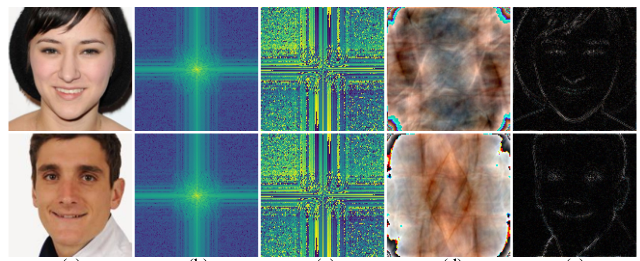
Abstract

Face super-resolution (FSR) aims to reconstruct high-resolution (HR) face images from the low-resolution (LR) ones. With the advent of deep learning, the FSR technique has achieved significant breakthroughs. However, existing FSR methods either have a fixed receptive field or fail to maintain facial structure, limiting the FSR performance. To circumvent this problem, Fourier transform is introduced, which can capture global facial structure information and achieve image-size receptive field. Relying on the Fourier transform, we devise a spatial-frequency mutual network (SFMNet) for FSR, which is the first FSR method to explore the correlations between spatial and frequency domains as far as we know. To be specific, our SFMNet is a two-branch network equipped with a spatial branch and a frequency branch. Benefiting from the property of Fourier transform, the frequency branch can achieve image-size receptive field and capture global dependency while the spatial branch can extract local dependency. Considering that these dependencies are complementary and both favorable for FSR, we further develop a frequency-spatial interaction block (FSIB) which mutually amalgamates the complementary spatial and frequency information to enhance the capability of the model. Quantitative and qualitative experimental results show that the proposed method outperforms state-of-the-art FSR methods in recovering face images. The implementation and model will be released at https://github.com/wcy-cs/SFMNet.

1. Introduction

Face super-resolution (FSR), also known as face hallucination, is a technology which can transform low-resolution (LR) face images into the corresponding high-resolution (HR) ones. Limited by low-cost cameras and imaging conditions, the obtained face images are always low-quality, resulting in a poor visual effect and deteriorating the downstream tasks, such as face recognition, face attribute analysis, face editing, *etc*. Therefore, FSR has become an emerging scientific tool and has gained more of the spotlight in the computer vision and image processing communities [20].

*Corresponding author.



(a) (b) (c) (d) (e)

Figure 1. Decomposition and reconstruction of face image in the frequency domain. (a) denote face images; (b) are their amplitude spectrum; (c) show their phase spectrum; (d) present the reconstructed images with amplitude information only; (e) are the reconstructed images with phase information only.

FSR is an ill-posed challenging problem. In contrast to general image super-resolution, FSR only focuses on the face images and is tasked with recovering pivotal facial structures. The first FSR method proposed by Baker and Kanade [1] sets off the upsurge of traditional FSR methods. These traditional methods mainly resort to PCA [6], convex optimization [23], Bayesian approach [42] and manifold learning [19] to improve the quality of face images. Nevertheless, they are still incompetent in recovering plausible face images due to their limited representation abilities. In recent years, FSR has made a dramatic leap, benefiting from the advent of deep learning [20]. Researchers develop various network frameworks to learn the transformation from LR face images to the corresponding HR ones, including single-task learning frameworks [5, 8, 17], multi-task learning frameworks [4, 9, 32, 51], *etc*., which has greatly pushed forward the frontier of FSR research.

Although existing FSR methods improve FSR performance, they still have limitations to be tackled. Face image has global facial structure which plays an important role in transforming LR face images into the corresponding HR ones. However, the actual receptive field of the convolutional neural network is limited due to the vanishing gradient problem, failing to model global dependency. To achieve large receptive field, transformer has been applied in computer vision tasks [37, 54]. The self-attention mechanism among every patch can model long-range dependency, but it usually has a high demand for both training data and computation resource. In addition, the partition strategy may also destruct the structure of the facial image. Therefore, an effective FSR method that can achieve image-size receptive field and maintain the facial structure is an urgent demand. To meet this need, frequency information is introduced. It is well-accepted that features (for each pixel or position) in frequency domain can achieve image-size receptive field and naturally have the ascendency of capturing global dependency [33], and this can well complement the local facial features extracted in the spatial domain. To obtain frequency information, Fourier transform is adopted to decompose the image into the amplitude component and the phase component, which can well characterize the facial structure information. As shown in Fig. 1, the image reconstructed with the phase component reveals clear facial structural information which is lost in the LR face images. Naturally, the phase component of the Fourier transform contains key missing information that is critical for FSR task.

Based on the above analysis, we propose a novel spatial-frequency mutual network (SFMNet) for FSR, which explores the incorporation between spatial and frequency domains. The SFMNet is a two-branch network, including a frequency branch and a spatial branch. The frequency branch is tasked with capturing global facial structure by the Fourier transform, while the spatial branch is tailored for extracting local facial features. The global information in frequency domain and the local information in spatial domain are complementary, and both of them can enhance the representation ability of the model. In light of this, we carefully design a frequency-spatial interaction block (FSIB) to mutually fuse frequency and spatial information to boost FSR performance. Based on the SFMNet, we also develop a GAN-based model with a spatial discriminator and a frequency discriminator to guide the learning of the model in both spatial and frequency domains, which can further force the SFMNet to produce more high-frequency information.

Overall, the contributions of our work are three-fold:

i) We develop a spatial-frequency mutual network for face super-resolution, to the best of our knowledge, this is the first method that explores the potential of both spatial and frequency information for face super-resolution.

ii) We carefully design a frequency-spatial interaction block to mutually fuse global frequency information and local spatial information. Thanks to its powerful modeling ability, the complementary information contained in spatial and frequency domains can be fully explored and utilized.

iii) We conduct experiments to verify the superiority of the proposed method. Experimental results on two widely used benchmark datasets (*i.e,* CelebA [30] and Helen [25]) demonstrate that our method achieves the best performance in terms of visual results and quantitative metrics.

2. Related Work

2.1. Face Super-resolution

Alongside the rise of deep learning technique, researchers develop various convolutional neural network (CNN) based frameworks for improving face image quality. Zhou *et al*. [57] develop the first CNN-based FSR method, which greatly improves the FSR performance. To capture the inter-dependency among facial parts, the pioneering work of [5] is developed by exploiting reinforcement learning. Instead of training a network in end-to-end manner, the work of [21] first coarsely recovers an intermediate result and then compensates for the missing details of the intermediate result. Recently, numerous FSR methods have shown competence in developing effective attention mechanism. For example, SPARNet [8] develops a face attention unit to bootstrap facial structure information while SISN [31] designs an inter-feature split attention to capture facial details. In contrast to the above FSR methods that recover face images in image domain, the works of [16, 17] transform the face images into the wavelet coefficient by wavelet transform to capture rich contextual information. Inspired by the generative ability of generative adversarial network (GAN), Yu *et al*. [52] build URDGN based on GAN. Relying on GAN, another GAN-based FSR method is introduced in [47] which is a collaborative suppression and replenishment framework. However, the learning of GAN-based model is difficult, limiting its effectiveness. To reduce the learning difficulty of GAN, PCA-SRGAN [11] resorts on Principal Component Analysis decomposition while SPGAN [55] constructs supervised pixel-wise loss.

Since human face is a highly structured object, many FSR methods leverage facial prior to boost the FSR performance. Super-FAN [4] designs a heatmap loss which constrains the distance between heatmaps extracted from HR and super-resolved face images. However, the constraint of heatmap loss is not applied in the inference, and the heatmap information cannot be well exploited. To address this problem, Yu *et al.* [51] estimate facial prior from LR and then incorporate the prior to assist the FSR. Considering that estimating prior from degraded LR face images is challenging, FSRNet [9] suggests to enhance LR face images first and then estimate prior from the enhanced result and utilize the estimated prior. Later on, DIC [32] performs FSR and prior estimation iteratively and incorporates the prior to boost FSR to promote the two tasks each other. However, the accuracy of prior estimation is difficult to guarantee, limiting the overall FSR result.

2.2. Fourier Transform

The Fourier transform is a widely used technique to analyze the frequency content in signals. It can be viewed as a global statistical information of signals, and thus can capture long-range dependency. Depend on the characteristic of Fourier transform, Fourier transform is used to perform computer vision tasks. Yang *et al*. [49] utilize Fourier transform to assist domain adaption for boosting cross domain semantic segmentation. Later on, the work of [46] performs domain generation from Fourier-based perspective. A computationally efficient image classification network equipped with Fourier transform is introduced in [38]. In addition, to improve perceptual quality and recover hard high-frequency details, the works of [12,22] devise Fourier-based loss functions. In low-level tasks, Mao *et al*. [33] develop a Res FFT-Conv block to capture both long- and short- range dependencies for enhancing the details while phase-aware Fourier convolutions are built to improve the light of the images in [59]. Zhou *et al*. [58] propose to recover phase and amplitude seperatively with pan as guidance. Yu *et al*. [50] build amplitude guided phase module to perform dehazing while Huang *et al*. [18] first recover amplitude and then recover phase to improve image lightness.

3. Proposed Method

3.1. Revisiting Fourier Transform

Fourier transform is an important technique in signal processing, which is also a key component in our method. In this section, we first revisit the Fourier transform before introducing our method. Given a single channel image $\boldsymbol{x}$, the Fourier transform of the image $\boldsymbol{x}$ can be expressed as:

$$\mathcal{F}(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \boldsymbol{x}(h,w) e^{-2j\pi(\frac{h}{H}u + \frac{w}{W}v)}, \tag{1}$$

where H and W are the height and weight of the image $\boldsymbol{x}$, j represents the imaginary unit, u and v are the horizontal and vertical coordinates, and $\mathcal{F}$ denotes the Fourier transform. From Eq. (1), we learn that each "pixel" in $\mathcal{F}(x)$ is the aggregation of all the pixels in the image $\boldsymbol{x}$.

In frequency domain, the two significant components of $\boldsymbol{x}$, *i.e.*, the amplitude component $\mathcal{A}(x)$ and the phase component $\mathcal{P}(x)$, can be obtained by

$$\mathcal{A}(x)(u,v) = \sqrt{R^2(x)(u,v) + I^2(x)(u,v)}, \tag{2}$$

$$\mathcal{P}(x)(u,v) = \arctan(\frac{I(x)(u,v)}{R(x)(u,v)}), \tag{3}$$

where $R(x)$ and $I(x)$ correspond to the real and imaginary parts of $\mathcal{F}(x)$. Benefiting from the Fourier transform, these two components can capture the image-size receptive field easily, which can just meet our need for efficient long-distance dependency modeling. In addition, these two components capture different characteristics of face image. In Fig. 1, we show the original face image, corresponding amplitude and phase spectrum, and images reconstructed by only amplitude component and phase component, respectively. Obviously, the face images reconstructed by the phase component have clear facial structure information that is missing in LR face images. Thus, the phase component can maintain facial structure well and can be just used as a kind of facial prior to boost the FSR performance. In light of these two points, we develop our spatial-frequency mutual network (SFMNet) for FSR, which can not only capture long-distance dependency but also exploits local dependency. Profited by the characteristic of the network, our method can achieve state-of-the-art FSR performance.

3.2. SFMNet

Considering that both long- and short-range dependencies can boost FSR performance and the Fourier transform can easily obtain an image-size receptive field, we develop a spatial-frequency mutual network (SFMNet) which is the first FSR method to explore the incorporation between the spatial and frequency domains. The proposed SFMNet is illustrated in Fig. 2, which consists of a frequency branch and a spatial branch. Equipped with Fourier transform, the frequency branch is tailored for capturing global dependency assisted by image-size receptive field. The spatial branch captures local dependency and incorporates the global frequency information to reconstruct the final super-resolution (SR) result. Since global frequency information and local spatial information are complementary and different, we carefully design a frequency-spatial interaction block (FSIB) which can generate adaptive attention maps to incorporate these complementary information mutually and effectively. For PSNR-oriented model, both pixel-level and frequency-level loss functions are adopted to guide the learning of the network. Moreover, to improve visual quality, we introduce adversarial loss in both the spatial and frequency domains based on a spatial discriminator and a frequency discriminator. Now we elaborate on our SFMNet.

3.2.1 Overview

In this subsection, we introduce the pipeline in detail. Given the LR face image $\boldsymbol{I}_{\text{LR}}$, we feed it into two convolutional layers from two branches to extract features, generating $\boldsymbol{F}^0_{\text{Fre}}$ and $\boldsymbol{F}^0_{\text{Spa}}$ corresponding to the frequency and spatial branches. Then, the extracted features are fed into L spatial-frequency mutual learning modules (SFMLM) to extract multi-scale features,

$$\boldsymbol{F}^i_{\text{Spa}}, \boldsymbol{F}^i_{\text{Fre}} = f^i_{\text{SFMLM}}(\boldsymbol{F}^{i-1}_{\text{Spa}}, \boldsymbol{F}^{i-1}_{\text{Fre}}), \tag{4}$$

where f^i_{SFMLM} is the function of the i-th SFMLM. After L SFMLMs, $\boldsymbol{F}^L_{\text{Spa}}$ and $\boldsymbol{F}^L_{\text{Fre}}$ are fed into reconstruction layers

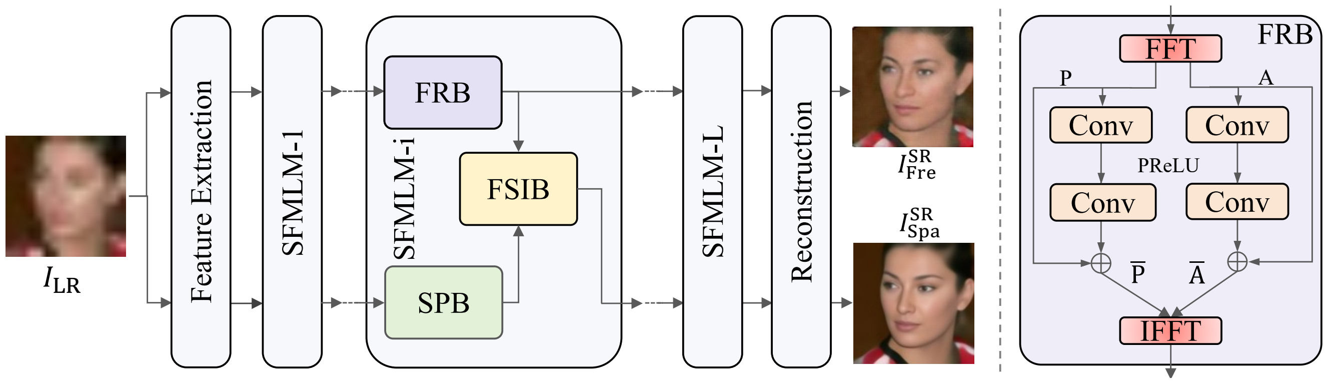


Figure 2. Overview of the proposed SFMNet in which FFT and IFFT are Fourier transform and inverse Fourier transform. SFMNet consists of a frequency branch (the top branch) and a spatial branch (the bottom branch). The former aims at capturing global facial structure and achieving image-size receptive field with Fourier transform, while the latter focuses on extracting local facial features.

(comprised of a convolutional layer) in two branches, recovering face images $\boldsymbol{I}^{\text{SR}}_{\text{Fre}}$ and $\boldsymbol{I}^{\text{SR}}_{\text{Spa}}$ as shown in Fig. 2.

To urge the model to perform FSR well, the model is supervised by pixel-level and frequency-level loss functions,

$$\mathcal{L}_{\text{Pix}} = \left\|\boldsymbol{I}^{\text{SR}}_{\text{Spa}} - \boldsymbol{I}_{\text{HR}}\right\|_1 + \left\|\boldsymbol{I}^{\text{SR}}_{\text{Fre}} - \boldsymbol{I}_{\text{HR}}\right\|_1, \tag{5}$$

$$\mathcal{L}_{\text{Fre}} = \left\|\mathcal{A}(\boldsymbol{I}^{\text{SR}}_{\text{Fre}}) - \mathcal{A}(\boldsymbol{I}_{\text{HR}})\right\|_1 + \left\|\mathcal{P}(\boldsymbol{I}^{\text{SR}}_{\text{Fre}}) - \mathcal{P}(\boldsymbol{I}_{\text{HR}})\right\|_1, \tag{6}$$

where $\mathcal{L}_{\text{Pix}}$ and $\mathcal{L}_{\text{Fre}}$ correspond to loss at the pixel-level and frequency-level, $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ are operations to extract amplitude and phase. The pixel-level loss guides the SFMNet to reconstruct high-fidelity face images, and the frequency-level loss helps it to learn frequency information. In addition, benefiting from the powerful generative ability of the generative adversarial network, we introduce adversarial losses in both spatial and frequency domains. In detail, we build a spatial discriminator and a frequency discriminator to discriminate recovered SR results and HR in the spatial and frequency domains, respectively. The two discriminators share a similar structure but take different inputs. The input of the spatial discriminator is SR or HR while that of the frequency discriminator is the concatenation of amplitude and phase of SR or HR. The specific loss functions are

$$\mathcal{L}^{\text{Adv}}_{\text{Spa}} = -log(\mathcal{SD}(\boldsymbol{I}^{\text{SR}}_{\text{Spa}})), \tag{7}$$

$$\mathcal{L}^{\text{Adv}}_{\text{Fre}} = -log(\mathcal{FD}([\mathcal{A}(\boldsymbol{I}^{\text{SR}}_{\text{Spa}}), \mathcal{P}(\boldsymbol{I}^{\text{SR}}_{\text{Spa}})])), \tag{8}$$

where $[\cdot,\cdot]$ denotes concatenation, $\mathcal{SD}$ and $\mathcal{FD}$ correspond to the spatial discriminator and the frequency discriminator respectively. Except that, perceptual loss which measures the distance between facial features is also adopted,

$$\mathcal{L}_{\text{Per}} = \left\|\Phi(\boldsymbol{I}^{\text{SR}}_{\text{Spa}}) - \Phi(\boldsymbol{I}_{\text{HR}})\right\|_1, \tag{9}$$

where Φ denotes VGG [41]. The whole loss function is

$$\mathcal{L} = \mathcal{L}_{\text{Spa}} + \gamma_1 * \mathcal{L}_{\text{Fre}} + \gamma_2 * \mathcal{L}^{\text{Adv}}_{\text{Fre}} + \gamma_3 * \mathcal{L}^{\text{Adv}}_{\text{Spa}} + \gamma_4 * \mathcal{L}_{\text{Per}}, \tag{10}$$

where $\gamma_1, \gamma_2, \gamma_3$ and γ_4 are the trade-off parameters.

3.2.2 Spatial-frequency Mutual Learning Module

Here we elaborate on the i-th spatial-frequency mutual learning module (SFMLM) in SFMNet. In detail, at the i-th SFMLM, $\boldsymbol{F}^{i-1}_{\text{Spa}}$ and $\boldsymbol{F}^{i-1}_{\text{Fre}}$ generated by the i-1-th SFMLM are fed into the spatial and frequency branches, respectively,

$$\boldsymbol{F}^{i}_{\text{Fre}} = f^{i}_{\text{FRB}}(\boldsymbol{F}^{\text{i-1}}_{\text{Fre}}), \overline{\boldsymbol{F}}^{i}_{\text{Spa}} = f^{i}_{\text{SPB}}(\boldsymbol{F}^{\text{i-1}}_{\text{Spa}}), \tag{11}$$

where f^{i}_{FRB} and f^{i}_{SPB} correspond to frequency block (FRB) and spatial block (SPB) in the frequency branch and spatial branch, respectively, $\overline{\boldsymbol{F}}^{i}_{\text{Spa}}$ and $\boldsymbol{F}^{i}_{\text{Fre}}$ are the extracted features. SPB consists of cascaded residual blocks [14]. Contrary to SPB, FRB decomposes the input into phase A and amplitude P components, and then adopts two convolutional layers to recover two components $\overline{A}$ and $\overline{P}$, respectively. Finally, the inverse Fourier transform is used to generate the output $\boldsymbol{F}^{i}_{\text{Fre}}$, as shown in Fig. 2 (right).

Thanks to the Fourier transform, the frequency branch can capture global dependency with an image-size receptive field, while the spatial branch can extract local dependency. In light of that global and local dependencies are complementary and can both facilitate FSR, we develop a frequency-spatial interaction block (FSIB) which is planted behind every SPB to harness the complementarity of them,

$$\boldsymbol{F}^{i}_{\text{Spa}} = f^{i}_{\text{FSIB}}(\overline{\boldsymbol{F}}^{i}_{\text{Spa}}, \boldsymbol{F}^{i}_{\text{Fre}}), \tag{12}$$

where f^{i}_{FSIB} represents the function of our FSIB at i-th SFMLM, and $\boldsymbol{F}^{i}_{\text{Spa}}$ is the generated feature that incorporates global and local dependencies. At this time, the final outputs of the i-th SFMLM, $\boldsymbol{F}^{i}_{\text{Spa}}$ and $\boldsymbol{F}^{i}_{\text{Fre}}$, are obtained.

3.2.3 Frequency-spatial Interaction Block

As introduced in the previous section, the frequency branch captures image-size dependency benefited by the Fourier transform while the spatial branch utilizes convolutional

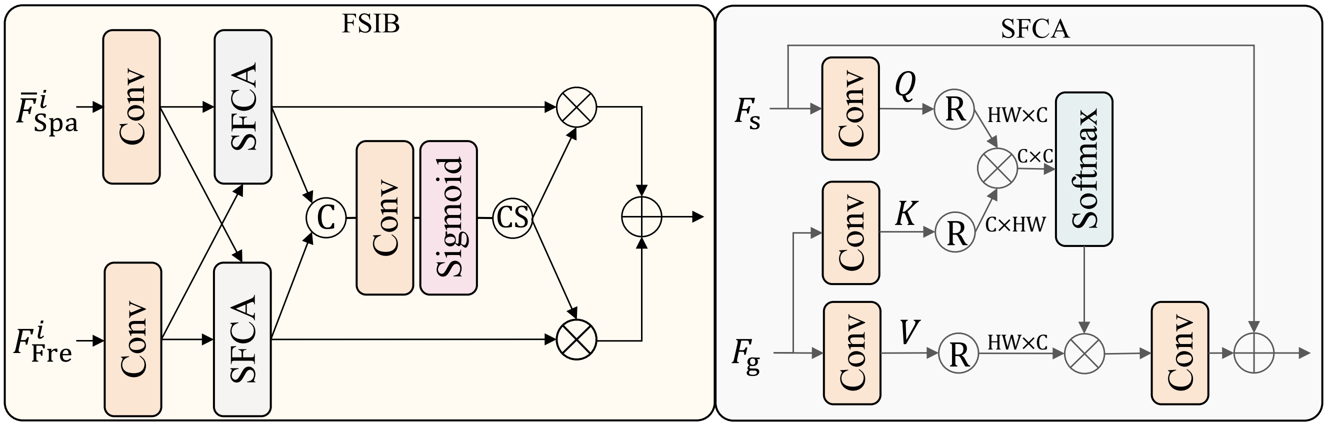


Figure 3. The architectures of the frequency-spatial interaction block (FSIB) (left) and spatial-frequency cross-attention (SFCA) (right) in which C denotes concatenation, CS is channel split operation and R is reshape operation.

layer to obtain short-range dependency. The image-size dependency and short-range dependency are complementary and both profitable to FSR task. In consideration of this, we should explore how to combine them to recover face images efficiently. To achieve this goal, we design a frequency-spatial interaction block (FSIB) which can mutually and adaptively fuse the global frequency information and local spatial information. As illustrated in Fig. 3, our FSIB first fuses two information coarsely by cross-attention and then generates different attention maps to fuse them finely.

Firstly, with $\overline{\boldsymbol{F}}^{i}_{\text{Spa}}$ and $\boldsymbol{F}^{i}_{\text{Fre}}$, FSIB applies two convolutional layers on them, obtaining $\hat{\boldsymbol{F}}_{\text{Spa}}$ and $\hat{\boldsymbol{F}}_{\text{Fre}}$ respectively. Note that we ignore the i in the equation for simplification. Then, to utilize the complementarity of global frequency information and local spatial information, we design a spatial-frequency cross-attention (SFCA) based on the self-attention mechanism.

SFCA To be specific, SFCA has two inputs, including source information $\boldsymbol{F}_{\text{s}}$ and guidance information $\boldsymbol{F}_{\text{g}}$. To fuse the two information fully, it uses $\boldsymbol{F}_{\text{s}}$ to generate query Q and uses $\boldsymbol{F}_{\text{g}}$ to obtain key K and value V by applying different convolutional layers. After that, the cross-attention between the source and guidance can be obtained by

$$\text{Attention}(K,Q,V) = f_{\text{Softmax}}(QK^{T}/\sqrt{d})V, \tag{13}$$

$$\boldsymbol{F}_{\text{Fuse}} = f_{\text{Conv}}(\text{Attention}(K,Q,V)) + \boldsymbol{F}_{\text{s}}, \tag{14}$$

where d is the hyperparameter, and $\boldsymbol{F}_{\text{Fuse}}$ is result. To capture global incorporation among channel dimension and decrease the computational cost of the SFCA, multiplication is calculated along channel dimension.

To model the incorporation between local spatial information and global frequency information, we let the frequency information $\hat{\boldsymbol{F}}^{i}_{\text{Fre}}$ and the spatial information $\hat{\boldsymbol{F}}^{i}_{\text{Spa}}$ serve as source and guidance for each other in FSIB,

$$\boldsymbol{F}^{\text{SFCA}}_{\text{Fre}} = f_{\text{SFCA}}(\hat{\boldsymbol{F}}_{\text{Fre}}, \hat{\boldsymbol{F}}_{\text{Spa}}), \boldsymbol{F}^{\text{SFCA}}_{\text{Spa}} = f_{\text{SFCA}}(\hat{\boldsymbol{F}}_{\text{Spa}}, \hat{\boldsymbol{F}}_{\text{Fre}}), \tag{15}$$

where $f_{\text{SFCA}}(\cdot,\cdot)$ is the function of SFCA, and the two parameters correspond to the source and guidance respectively, $\boldsymbol{F}^{\text{SFCA}}_{\text{Fre}}$ and $\boldsymbol{F}^{\text{SFCA}}_{\text{Spa}}$ are the fused results which combine the spatial and frequency information. With these two results, we propose to predict the pixel-wise attention with a predict network f_{PN} for fusing them mutually. In detail, we first concatenate the two results and feed the concatenated result into the predict network that consists of convolutional layers followed by sigmoid,

$$\boldsymbol{F}_{\text{Att}} = f_{\text{PN}}([\boldsymbol{F}^{\text{SFCA}}_{\text{Fre}}, \boldsymbol{F}^{\text{SFCA}}_{\text{Spa}}]), \tag{16}$$

where $\boldsymbol{F}_{\text{Att}}$ is the predicted attention and the channel of $\boldsymbol{F}_{\text{Att}}$ is twice as large as the one of the original feature. In light of the difference and complementarity between the frequency information and spatial information, we split the $\boldsymbol{F}_{\text{Att}}$ along the channel dimension to generate adaptive attentions $\boldsymbol{F}^{\text{Att}}_{\text{Spa}}$ and $\boldsymbol{F}^{\text{Att}}_{\text{Fre}}$ for reweighting them adaptively,

$$\boldsymbol{F}_{\text{SF}} = \boldsymbol{F}^{\text{Att}}_{\text{Spa}} * \hat{\boldsymbol{F}}_{\text{Spa}} + \boldsymbol{F}^{\text{Att}}_{\text{Fre}} * \hat{\boldsymbol{F}}_{\text{Fre}}, \tag{17}$$

where $\boldsymbol{F}_{\text{SF}}$ is the final fusion result.

4. Experiments

4.1. Dataset and Metrics

Two widely used face datasets are chosen in this paper, including CelebA [30] and Helen [25]. To be specific, we apply OpenFace [2,3,53] to extract 68 facial landmarks based on which face images are cropped, and then the cropped face images are resized into 128×128 pixels as ground truth. To acquire LR face images, the ground truth is downsampled into 32×32 and 16×16 corresponding to 4× and 8× FSR tasks, respectively. In the training phase, we use 168,854 face images from CelebA [30]. In the testing phase, 50 face images from Helen [25] and 1,000 face images from CelebA [30] are chosen to evaluate the performance of the model. In terms of evaluation metrics, PSNR, SSIM [44], LPIPS [56] and NIQE [35] results are used.

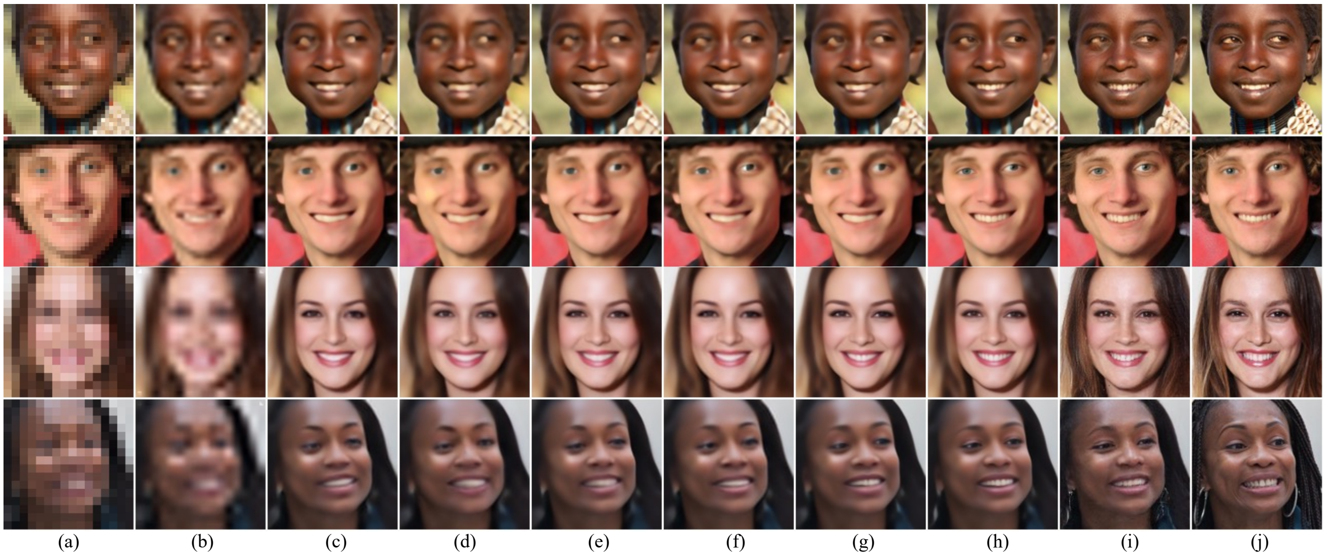


Figure 4. Visual quality comparison of state-of-the-art methods on Helen [25] dataset by the scale of $\times 4$ (the top two face images) and CelebA [30] dataset by the scale of $\times 8$ (the bottom two face images). Please zoom in to view the differences. (a): LR; (b): SRCNN [10]; (c): EDSR [29]; (d): FSRNet [9]; (e): DIC [32]; (f): SPARNet [8]; (g): SISN [31]; (h): Our SFMNet; (i): GAN-based SFMNet; (j): HR.

Table 1. Quantitative evaluation of various FSR methods on CelebA [30] and Helen [25] datasets. The **best** and the second-best results are emphasized with **bold** and underscore, respectively. Par denotes the parameter and the time is running time in the inference phase.

Dataset	CelebA [30]						Helen [25]						Par	Time
	$\times 4$			$\times 8$			$\times 4$			$\times 8$				
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓		
Bicubic	27.48	0.8166	0.1841	23.58	0.6285	0.2692	28.22	0.6628	0.1771	23.88	0.6628	0.2560	-	-
SRCNN [10]	28.04	0.8369	0.1599	23.93	0.6348	0.2559	28.77	0.8730	0.0556	24.27	0.6770	0.2430	19.6k	9.1ms
EDSR [29]	31.45	0.9095	0.0518	26.84	0.7787	0.1159	31.87	0.9286	0.0574	26.60	0.7851	0.1400	3.4M	10.0ms
FSRNet [9]	31.46	0.9084	0.0519	26.66	0.7714	0.1098	31.93	0.9283	0.0543	26.43	0.7799	0.1356	3.2M	53.0ms
DIC [32]	31.53	0.9107	0.0532	27.37	0.8022	0.0920	31.98	0.9303	0.0576	26.94	0.8026	0.1144	20.8M	84.6ms
SPARNet [8]	31.71	0.9129	0.0476	27.42	0.8036	0.0891	31.98	0.9300	0.0592	26.95	0.8029	0.1169	10.0M	45.0ms
SISN [31]	31.88	0.9157	0.0476	27.31	0.7978	0.0998	32.41	0.9351	0.0535	27.08	0.8083	0.1225	8.4M	63.8ms
SFMNet(Ours)	**32.01**	**0.9175**	0.0441	**27.56**	**0.8074**	0.0869	**32.51**	**0.9362**	0.0498	**27.22**	**0.8141**	0.1061	8.1M	51.8ms
SFMNet+GAN	30.99	0.8051	**0.0291**	26.48	0.7662	**0.0594**	31.54	0.9187	**0.0323**	26.39	0.7792	**0.0760**	8.1M	51.8ms

4.2. Implementation Details

In SFMNet, L is set to 14 and the number of residual blocks in every SRB is 2. In addition, a downsampling module (implemented by the inverse pixelshuffle [24] and convolutional layers) and an upsampling module (implemented by pixelshuffle [40] and convolutional layers) are inserted after every FRB and SRB in 1-6 SFMLMs and 9-14 SFMLMs, respectively. For training the PSNR-oriented model, γ_1 is 0.01, γ_2, γ_3 and γ_4 are set as zero. For GAN-based model, we use the pretrained PSNR-oriented model as initialization and set γ_2=0.0005, γ_3=0.001 and γ_4=0.1. We use the Adam optimizer with β_1=0.9, β_2=0.99, and ϵ=1e-8 to train our model. The learning rate is set to 1e-4. Our experiments are implemented on PyTorch [36] with NVIDIA GeForce RTX 3090.

4.3. Comparison with the state-of-the-arts

To verify the superiority of our proposed method, we compare our method with several state-of-the-art methods, including two representative convolutional neural network-based general image super-resolution methods SRCNN [10] and EDSR [29], and four FSR methods, FSRNet [9], DIC [32], SPARNet [8] and SISN [31]. In addition, Bicubic interpolation is also used as a baseline. The quantitative results are tabulated in Table 1. To be fair, all models are trained and tested with the same dataset. It can be observed that our method can achieve the best performance in both two testing datasets. SRCNN and EDSR are not designed for face images and fail to recover face images well.

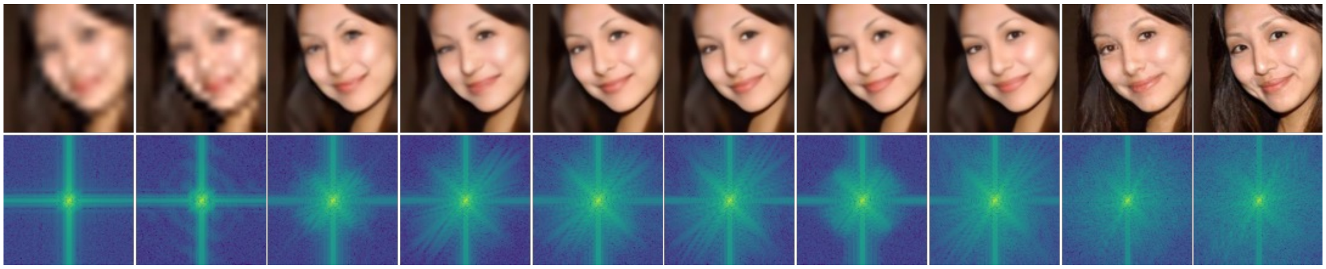


Figure 5. The visualization comparison of different FSR methods in both spatial domain (the top row) and frequency domain (the bottom row). Please zoom in to view the differences. (a): LR; (b): SRCNN [10]; (c): EDSR [29]; (d): FSRNet [9]; (e): DIC [32]; (f): SPARNet [8]; (g): SISN [31]; (h): Our SFMNet; (i): GAN-based SFMNet; (j): HR.

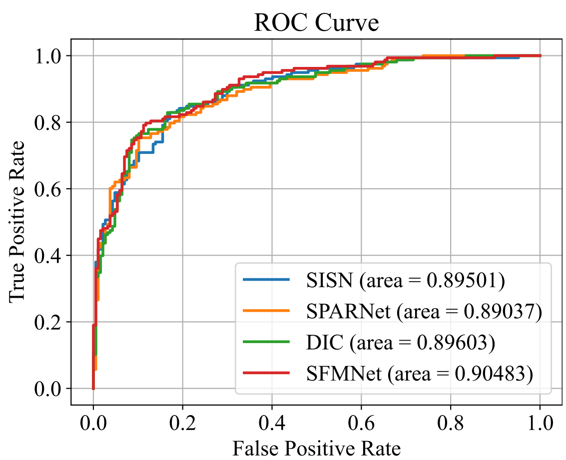


Figure 6. ROC curve on LFW [15] for face recognition task.

Figure 7. Real-world LR face images restoration comparison. (a): LR; (b): Restoreformer [45]; (c): VQFR [13]; (d): Ours.

FSRNet and DIC propose to estimate the facial prior and then utilize the prior. However, the accuracy of the estimated prior cannot be guaranteed, limiting the FSR performance. Compared to SPARNet and SISN limited in spatial space, our method can exploit the frequency information that can capture image-size receptive field and depict high-frequency details, improving FSR performance obviously. Except the PSNR and SSIM, we also present the parameter and running time of different FSR methods in Table 1 and our method achieves a good balance between performance and model complexity.

In addition, we also visualize the results hallucinated by different methods in Fig. 4. Regarding the frontal face images (*e.g.,* the third face in Fig. 4), all methods can reconstruct facial structure while our methods have the advantage of recovering facial details, especially key facial components, such as teeth and eyes. As for profile face images, our method can still have an advantage in recovering accurate and realistic details than other methods, demonstrating the robustness and stability of our method. This is because our method can utilize global face structure and implicit phase prior provided by the Fourier transform. In addition to that, we also visualize the comparison results in both the spatial domain and the frequency domain in Fig. 5. Thanks to the Fourier transform, our method can not only recover realistic facial details in the spatial domain but also reconstruct an accurate frequency spectrum in the frequency domain. To conclude, quantitative and visual quality comparisons prove the superiority of our method.

4.4. Face Recognition Results

A good FSR method can not only achieve higher PSNR and SSIM, but also improve the downstream tasks such as face recognition. Thus, we also perform face recognition as a measurement to evaluate the FSR performance of different FSR methods. To be specific, we randomly select some face images from LFW [15] dataset as reference. Then, for every reference, we select face images with the same identity and the different identities as test images. Then, we downsample the test images and use different FSR methods to recover them. Then, we adopt a pretrained face recognition model Deepface [39] to perform face recognition and judge whether the test and reference images belong to the same person. Then we plot the ROC curves in Fig. 6. As illustrated, the AUC (the area under the ROC curve) result of ours is the largest, demonstrating that our SFMNet outperforms other FSR methods in face recognition task.

Table 2. Ablation study of the proposed FSIB.

Dataset	CelebA [30]		Helen [25]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SBN	27.35	0.8010	26.98	0.8063
SFCNet	27.39	0.8033	27.01	0.8079
SFCCNet	27.40	0.8022	27.10	0.8072
SFMNet	**27.56**	**0.8082**	**27.22**	**0.8141**

Table 3. Ablation study of the frequency discriminator.

Dataset	CelebA [30]		Helen [25]	
	LPIPS↓	NIQE↓	LPIPS↓	NIQE↓
SFMNet	0.0869	10.620	0.1061	10.964
SD	0.0684	6.931	0.0847	7.475
SFD	**0.0594**	**6.690**	**0.0760**	**7.020**

4.5. Real-world Face Restoration

In addition, we also verify the performance of our model on real-world face image restoration. For real-world face image restoration, existing methods [7, 13, 26–28, 34, 43, 45, 48] explore the potential of reference prior, generative prior or vector-quantized dictionary. From them, we choose Restoreformer [45] and VQFR [13] as comparison methods. Note that we directly use the pretrained model in official public code to infer the results. As shown in Fig. 7, Restoreformer generates many artifacts while faces hallucinated by VQFR have high perceptual quality. However, the faces recovered by VQFR are slightly distorted, and the expression and the fidelity of the recovered faces is different from the original faces. Although the results of our method are not as high quality as those of VQFR, they are realistic and natural and contain key facial details. In addition, our model is only trained on 128×128 face images degraded with Bicubic while the comparison methods are trained on 512×512 face images with complex degradation. In summary, our method can be used to recover real-world LR face images.

4.6. Ablation Study

In this section, we further conduct experiments to verify the effectiveness of key components in SFMNet on ×8.

The effectiveness of the FSIB: First, we remove the frequency branch and only preserve the spatial branch in our method, and the remaining model is called SBN. Then, we recover the frequency branch and replace our FSIB with concatenation followed by convolutional layers with the similar parameters to SFMNet, named SFCNet. Finally, our carefully designed FSIB is planted into the SFCNet, generating the SFMNet. The results are reported in Table 2. The quantitative metrics of SFCNet are a little better than the ones of SBN, demonstrating that the frequency branch can provide global dependency to enhance the representation ability of the model. However, the improvement is limited due to that concatenation is too simple to perform the interaction between the frequency domain and the spatial domain. Finally, equipped with our carefully designed FSIB, our method SFMNet achieves the best performance in terms of both PSNR and SSIM. To further verify the effectiveness of the proposed SFCA, we replace the SFCA with concatenation followed by convolutional layer, generating the model SFCCNet. Compared with SFCCNet, SFCA can improve FSR performance obviously.

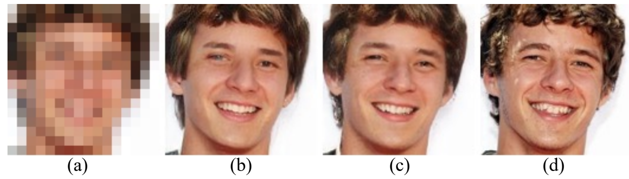


Figure 8. ×8 SR results of different discriminators. (a): LR; (b): SD; (c): SFD; (d): HR; SFD can recover realistic facial details.

The effectiveness of the frequency discriminator: We also conduct experiments to analyze the effectiveness of the frequency discriminator. Specifically, we compare the results of the model with and without the frequency discriminator in Table 3, where SD and SFD are denoted as the former and the latter model, respectively. From the aspects of quantitative metrics, the introduction of the frequency discriminator can obviously improve the LPIPS and NIQE performance of the model. The visual quality comparison of SD and SFD is shown in Fig. 8. Since frequency spectrum can capture global face structure, faces hallucinated by SFD look more realistic and visually pleasing than those of SD.

5. Conclusion

In this paper, we develop a spatial-frequency mutual network (SFMNet) for face super-resolution, which is the first work to explore the interaction between spatial domain and frequency domain in this field. The proposed SFMNet is a two-branch network, including a spatial branch and a frequency branch. The spatial branch extracts local facial features in the spatial domain. The frequency branch takes advantage of Fourier transform to finish the transformation from spatial domain to frequency domain and capture global dependency with image-size receptive field. To explore the complementarity between the global and local information, we carefully design a frequency-spatial interaction block that can fuse these dependencies mutually and boost face super-resolution performance. Finally, a frequency discriminator is developed to guide the model in frequency domain. Experimental results demonstrate that our proposed method can achieve state-of-the-art performance.

Acknowledgements: The research was supported by the National Natural Science Foundation of China (61971165, 92270116), and in part by the Fundamental Research Funds for the Central Universities (FRFCU5710050119).

References

[1] Simon Baker and Takeo Kanade. Hallucinating faces. *IEEE Int.conf.automatic Face I and Gesture Recognition*, pages 83–88, 2000. 1

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013. 5

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66, 2018. 5

[4] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2018. 1, 2

[5] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 690–698, 2017. 1, 2

[6] Ayan Chakrabarti, A. N. Rajagopalan, and Rama Chellappa. Super-resolution of face images using kernel pca-based prior. *IEEE Transactions on Multimedia*, 9(4):888–892, 2007. 1

[7] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 8

[8] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K. Wong. Learning spatial attention for face super-resolution. *IEEE Trans. Image Processing*, 30:1219–1231, 2021. 1, 2, 6, 7

[9] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. FSRNet: End-to-end learning face super-resolution with facial priors. In *Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018. 1, 2, 6, 7

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015. 6, 7

[11] Hao Dou, Chen Chen, Xiyuan Hu, Zuxing Xuan, Zhisen Hu, and Silong Peng. Pca-srgan: Incremental orthogonal projection discrimination for face super-resolution. In *Proceedings of the ACM International Conference on Multimedia*, pages 1891–1899, 2020. 2

[12] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2360–2369, 2021. 3

[13] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 2022. 7, 8

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Sun Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[15] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 7

[16] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-SRNet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017. 2

[17] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet domain generative adversarial network for multi-scale face hallucination. *International Journal of Computer Vision*, 127(6-7):763–784, 2019. 1, 2

[18] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 163–180. Springer, 2022. 3

[19] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han. Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning. *IEEE Trans. Image Process.*, 23(10):4220–4231, 2014. 1

[20] Junjun Jiang, Chenyang Wang, Xianming Liu, and Jiayi Ma. Deep learning-based face super-resolution: A survey. *ACM Computing Surveys*, 55(1):1–36, 2023. 1

[21] Junjun Jiang, Yu Yi, Jinhui Hu, Suhua Tang, and Jiayi Ma. Deep cnn denoiser and multi-layer neighbor component embedding for face hallucination. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 771–778, 2018. 2

[22] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13919–13929, 2021. 3

[23] Cheolkon Jung, Licheng Jiao, Liu Bing, and Maoguo Gong. Position-patch based face hallucination using convex optimization. *IEEE Signal Processing Letters*, 18(6):367–370, 2011. 1

[24] Junhyung Kwak and Donghee Son. Fractal residual network and solutions for real super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2114–2121, 2019. 6

[25] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Interactive facial feature localization. In *Proceedings of the European Conference on Computer Vision*, pages 679–692, 2012. 2, 5, 6, 8

[26] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration

via deep multi-scale component dictionaries. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 399–415. Springer, 2020. 8

[27] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020. 8

[28] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018. 8

[29] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1132–1940, 2017. 6, 7

[30] Ziwei Liu, Luo Ping, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2016. 2, 5, 6, 8

[31] Tao Lu, Yuanzhi Wang, Yanduo Zhang, Yu Wang, Liu Wei, Zhongyuan Wang, and Junjun Jiang. Face hallucination via split-attention in split-attention network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5501–5509, 2021. 2, 6, 7

[32] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5569–5578, June 2020. 1, 2, 6, 7

[33] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021. 2, 3

[34] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 8

[35] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 5

[36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Proceedings of the Conference on Neural Information Processing Systems Workshop*, pages 4–9, 2017. 6

[37] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 367–376, 2021. 2

[38] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 3

[39] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *Proceedings of Innovations in Intelligent Systems and Applications Conference*, pages 23–27, 2020. 7

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, June 2016. 6

[41] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. 4

[42] Marshall F Tappen and Ce Liu. A bayesian approach to alignment-based image hallucination. In *Proceedings of the European Conference on Computer Vision*, pages 236–249, 2012. 1

[43] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 8

[44] Zhou Wang and A.C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002. 5

[45] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 7, 8

[46] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 3

[47] Lingbo Yang, Chang Liu, Pan Wang, Shanshe Wang, Peiran Ren, Siwei Ma, and Wen Gao. Hifacegan: Face renovation via collaborative suppression and replenishment. In *Proceedings of the ACM International Conference on Multimedia*, pages 1551–1560, 2020. 2

[48] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 8

[49] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[50] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 181–198. Springer, 2022. 3

[51] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European Conference on Computer Vision*, pages 219–235, 2018. 1, 2

[52] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *Proceedings of the European Conference on Computer Vision*, pages 318–333, 2016. 2

[53] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2519–2528, 2017. 5

[54] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 2

[55] Menglei Zhang and Qiang Ling. Supervised pixel-wise GAN for face super-resolution. *IEEE Trans. Multimedia*, 23:1938–1950, 2021. 2

[56] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume pp, pages 586–595, June 2018. 5

[57] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Learning face hallucination in the wild. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, pages 3871–3877, 2015. 2

[58] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 274–291. Springer, 2022. 3

[59] Yunliang Zhuang, Zhuoran Zheng, and Chen Lyu. Dpfnet: A dual-branch dilated network with phase-aware fourier convolution for low-light image enhancement, 2022. 3