

MSFSR: A Multi-Stage Face Super-Resolution with Accurate Facial Representation via Enhanced Facial Boundaries

Yunchen Zhang, Yi Wu, Liang Chen*

Fujian Provincial Engineering Technology Research Center of Photoelectric Sensing Application,
Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education,
Fujian Provincial Key Laboratory of Photonics Technology,
Fujian Normal University

cydiachen@cydiachen.tech, wuyi@fjnu.edu.cn, cl.0827@126.com

Abstract

The majority of Face Super-Resolution (FSR) approaches apply specific facial priors as guidance in super-resolving the given low-resolution (LR) into high-resolution (HR) images. To improve the FSR performance, various kinds of facial representations were explored in the past decades. Nevertheless, there remains a challenge in estimating high-quality facial representations for LR images. To address this problem, we propose novel facial representation - enhanced facial boundaries. By semantically connecting the facial landmark points, enhanced facial boundaries retain rich semantic information and are robust to different spatial resolution scales. Based on the enhanced facial boundaries, we design a novel Multi-Stage FSR (MSFSR) approach, which applies the multi-stage strategy to recover high-quality face images progressively. The enhanced facial boundaries and the coarse-to-fine supervision facilitate the facial boundaries estimation process in producing high quality facial representation. The one-time projection of the FSR task is decomposed into multiple simpler sub-processes. In these ways, the MSFSR estimates a more robust facial representation and achieves better performance. Experimental results indicate the superiority of our approach to the state-of-the-art approaches in both qualitative and quantitative measurements.

1. Introduction

Face images, different from other real-world images, have distinct distribution in their highly structural shapes and rich contextual information. However, face images are sometimes in low-resolution (LR) mode as they are captured under low-quality scenarios with inferior camera sen-

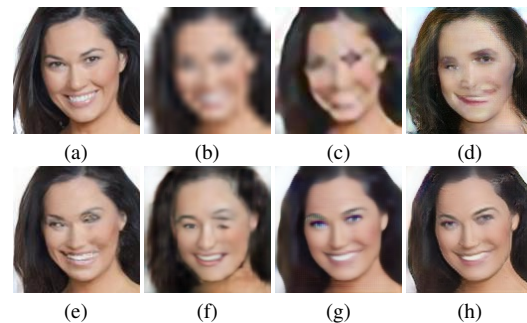


Figure 1. Visual results of different SR approaches on upscale factor of $8\times$. (a) Ground Truth. (b) Bicubic. (c) EDSR [1]. (d) URDGN [2]. (e) PFSR [3]. (f) FSRNet [4]. (g) Ours. (h) Ours-GAN.

sors and far-shooting distance from the interested faces [5], making it difficult for high-level face applications, e.g., face recognition, face manipulation and face alignment.

To address this problem, Face Super-Resolution (FSR), a.k.a. Face hallucination, is the super-resolution (SR) technique that reconstructs only face images. The FSR technique aims at reconstructing visually appealing high-resolution (HR) images from LR face images. As a classical domain-specific SR technique, the FSR can effectively reduce the blurry and mismatched texture (the first line in Figure 1) in the reconstructed HR face image due to its consideration on the particular geometry structure of face images.

It is now decades since the seminal work made by Baker and Kanade [6] that first proposed FSR. Afterwards, various kinds of FSR techniques have developed. Most classical models focus on discovering appropriate mappings between LR inputs and corresponding HR face images. Recently, a variety of convolutional neural networks (CNNs) [7, 8, 9] and generative adversarial networks (GANs) [10, 11] have

*Corresponding author.

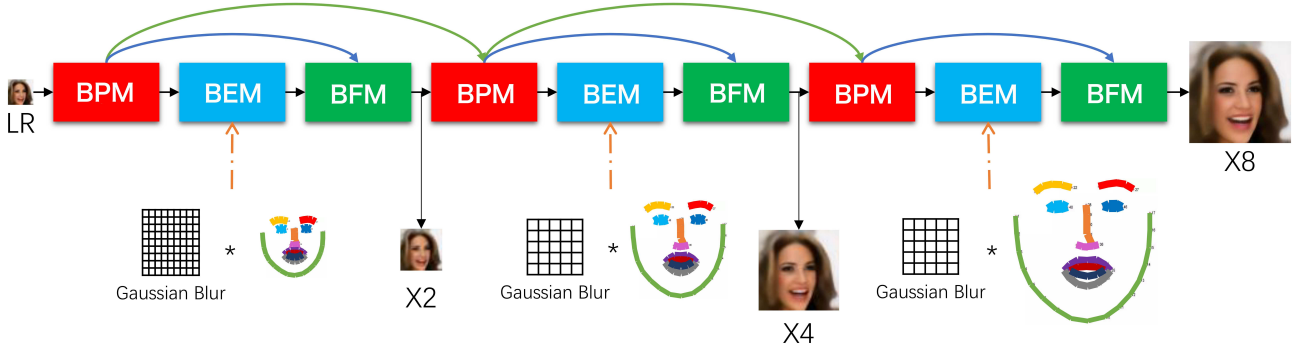


Figure 2. Pipeline of our proposed MSFSR model. ‘BPM’ is responsible for magnifying spatial resolution of LR input. ‘BEM’ extracts enhanced facial boundaries from previous outputs, and ‘BFM’ combines outputs from ‘BPM’ with enhanced facial boundaries together to generate super-resolved image. The green lines indicate shortcut connections between different stages. The orange arrows indicates extra supervision on enhanced facial boundaries.

been proposed to produce photo-realistic face images. The minority of these approaches [2, 12] adopted various kinds of facial attribute vectors as facial features. However, these approaches merely focus on specific facial attributes, which result in hallucinating wrong faces. The majority of FSR approaches are based on diverse structural facial representations such as facial parsing maps [4], facial landmark points [3] and dense correspondence fields [7]. As depicted in Figure 1 (e), (f), these prior-based approaches are able to generate face images with better visual quality. Unfortunately, Figure 1 (e), which is generated by a network that utilizes facial landmark points, suffers from distortion on important facial components such as eyes. The FSR network with facial parsing maps [4] reconstructs face images with accurate localization of each facial component, whereas these images lose inner structural details on the nose bridge, which is demonstrated in Figure 1 (f). To solve these problems, novel face structural representations should be explored to improve FSR performance.

Wu *et al.* [13] first adopted 13 facial boundary lines in the task of face alignment in the wild. Considering the superiority of facial boundary lines in the face alignment task, we manage to expand the robustness of facial boundary lines in face contour representation to facilitate the FSR task. This is because the facial boundary lines carry more semantic guidance to contour than pure landmark points due to the inherent continuity. In our design, we simplify the definition of original 13 facial boundaries and remove the ambiguous definitions to fit LR images in the task of FSR. Figure 1 (g), (h) confirms the effectiveness of the proposed 11 enhanced facial boundaries. Based on these considerations, we design a novel FSR framework i.e., a multi-stage network (MSFSR) containing three basic modules optimized for FSR. The three basic modules are Basic Pre-process Module (BPM), boundary estimation module

(BEM) and boundary fusion module (BFM). In addition, coarse-to-fine supervision and cross-stage shortcut connections are proposed to further improve the performance of MSFSR.

In summary, the contributions of this study are mainly in three aspects:

(1) We propose the enhanced facial boundaries as a new facial structural representation in the task of FSR. The enhanced facial boundaries are formed by connecting facial landmark points according to their semantic meanings. The continuity of the connected boundaries attaches richer semantic information and remains robust to faces with large poses and variations.

(2) We design three basic modules for the task of FSR: BPM, BEM, and BFM. The BPM gets rid of pre-defined upscaling operations with its post upscaling design, which improves the efficiency of the network. The BEM estimates enhanced facial boundaries directly from face images in a unified framework, and the BFM attaches channel-wise attention to fusing facial boundaries and feature maps, which fully explores the relationship between different latent spaces.

(3) We introduce the MSFSR network to improve the quality of reconstructed images. We integrate the enhanced facial boundaries with three basic modules into a multi-stage network design and propose a coarse-to-fine supervision to constrain the fineness of facial boundaries at different stages numerically. The proposed network not only estimates accurate facial representations, but also improves the fidelity of the reconstructed images effectively. Our approach achieves the state-of-the-art performance in both quantitative and and qualitative results.

2. Related Works

General image super-resolution. General super-resolution approaches can be divided into two main categories: traditional approaches and deep learning approaches. The traditional algorithms have been around for decades, but they are out-performed by the deep convolutional neural networks (CNN). Dong *et al.* [14] first proposed the Super-Resolution Convolutional Neural Network (SRCNN) to learn a mapping between LR and HR images. Lim *et al.* [1] introduced an Enhanced Deep Residual Network for Single Image Super-Resolution (EDSR). Kim *et al.* [15] proposed Deep Recursive Convolutional Network (DRCN) to break down the harder SR problem into a set of simpler ones. In addition, Ledig *et al.* [16] proposed Super-Resolution Generative Adversarial Network (SRGAN) to generate photo-realistic SR images.

All aforementioned approaches can be applied to all types of images, but they did not introduce any extra prior knowledge. Even though these approaches outperforms classical approaches with higher PSNR and SSIM scores, the recovered face images still suffer from mismatched textures and blurry facial components, which makes them impossible for high-level applications.

Face super-resolution. Since CNNs showed their great success in the field of general image SR, the FSR has also benefited from the widespread use of CNNs. Zhu *et al.* [7] designed the Deep Cascaded Bi-Network (CBN) for face super-resolution, which can be used for hallucinating unconstrained poses and with LR faces. Chen *et al.* [4] introduced an End-to-End Learning Face Super-Resolution Network with Facial Priors (FSRNet), which integrates the FSR task with the face alignment task into a unified structure. Other approaches have also been adopted into FSR. Li *et al.* [17] introduced Guided Face Restoration Network (GFRNet), which includes a sub-network to correct poses and a sub-network to intergrade degraded observation and warped guidance to produce the definitive restoration result. The concept of the generative adversarial network (GAN) has also been extensively used in the field of face hallucination. Xin *et al.* [2] first adopted a Discriminative Generative Network for Ultra-Resolving Face Images (URDGN). Hsu *et al.* [18] proposed a Siamese GAN (SiGAN) to reconstruct high-quality HR faces with corresponding identities. Zhang *et al.* [19] defined an identity loss to assess the differences between a hallucinated face and its corresponding HR face. Bulat *et al.* [10] proposed Super-FAN, which integrates face super-resolution and landmark localization into a single end-to-end system. Kim *et al.* [3] designed the Progressive Face Super-Resolution (PFSR) to introduce progressive training technique into reconstructing Extremely LR images (PFSR).

Although these approaches generate high-fidelity face images, the results suffer from performance degradations in

dealing with extreme LR images and cannot remain friendly in both annotation or computational cost. On the contrary, the proposed MSFSR adopts the enhanced facial boundaries and attaches coarse-to-fine supervision to generate enhanced facial boundaries. They carry abundant semantic information in assisting FSR. Furthermore, we decompose the one-time projection on extreme large upscale factor into a multi-stage, small upscale factor SR process.

3. Multi-Stage Face Super-Resolution

There are three aspects related to the main improvement in the MSFSR. First, the enhanced facial boundaries are proposed and applied after careful consideration of the deficiency in existing FSR approaches. With the enhancement process, the enhanced facial boundaries are more effective in representing the face structure. Second, three flexible blocks are designed and optimized for FSR. Finally, a coarse-to-fine supervision and the shortcut connections are introduced to the end-to-end framework of MSFSR.

3.1. Analysis of Facial Representations

As mentioned in the introduction, both facial landmark points and facial parsing maps have their limits in representing facial structures. To improve the performance of FSR approaches, we introduce the enhanced facial boundaries as a novel facial feature.

Most recent FSR approaches [10, 3, 4, 20] estimated facial features with variants of hourglass structure [21]. However, the hourglass structure is incapable of estimating com-

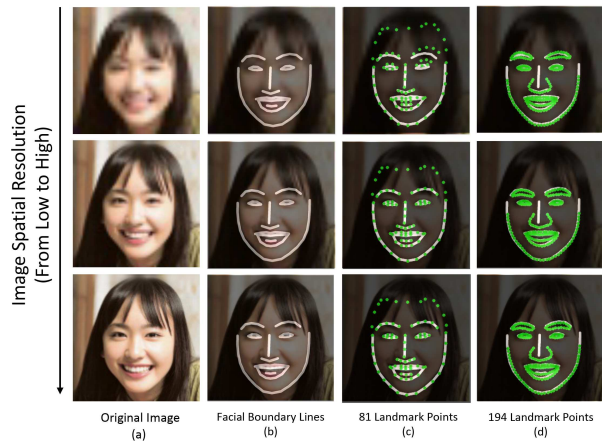


Figure 3. (a) shows face images at different resolution scales. (b) illustrates our proposed enhanced facial boundaries, which are manually generated from landmark points. (c) displays 11 enhanced facial boundaries and 81 landmark points on the same face image. (d) demonstrates 11 enhanced facial boundaries and 194 landmark points on the same face image.

plex representations due to its model capacity. The main challenge of discovering appropriate facial representation information lies in discovering a facial feature with rich semantic information and low complexity.

Facial landmark point is one of the most commonly used facial features in FSR approaches. The semantic information carried by landmark points is highly related to the density of landmark annotations.

Unfortunately, estimating dense landmark points requires HR face images that are unavailable in the FSR task. [20] adopts facial parsing maps to retain additional semantic information from its inherent properties on pixel-level annotations. However, facial parsing maps are very expensive to obtain and they tend to disjoint facial components while ignoring the inner geometric structure of facial components (e.g., nose bridge).

Compared with the above facial representation, enhanced facial boundaries show their potential in describing geometric information of the facial structure. As demonstrated in Figure 3, the enhanced facial boundaries contains not only almost all the information of facial landmark points, but also much more semantic information in the connected lines, which indicates that the basic facial geometric structures can be completely represented by enhanced facial boundaries. Therefore, we choose enhanced facial boundaries as face representation instead of facial landmarks.

In the setting of the original facial boundary lines introduced by [13], contours on human eyes are divided into upper boundaries and lower boundaries. The separation benefits the accurate localization of landmark points around eyes in HR. Unfortunately, the contours of human eyes are reduced to a few number of pixels in LR images, which make it difficult to estimate accurate facial structures. Therefore, we choose to form closed loops on human eyes. With our novel enhanced facial boundaries (e.g., face contours, eyebrows, nose bridge, eyes and mouth contours), we can effectively model geometric structures of faces.

In addition, as shown in Figure 3, facial landmark points are likely to shift along boundaries with spatial resolution of face images decreases. On the contrary, enhanced facial boundaries still remain stable in dealing with different resolution scales.

Due to the robustness in representing facial structures, we choose enhanced facial boundaries as facial structural feature in our FSR network.

3.2. Details of Multi-Stage Face Super-Resolution

The pipeline of our MSFSR model is shown in Figure 2. The network includes three individual modules: *Basic Pre-process Module (BPM)*, *Boundary Estimation Module (BEM)* and *Boundary Fusion Module (BFM)*. In addition, we introduce *Residual Channel Attention Block (RCAB)* [22] as a replacement for all vanilla residual blocks.

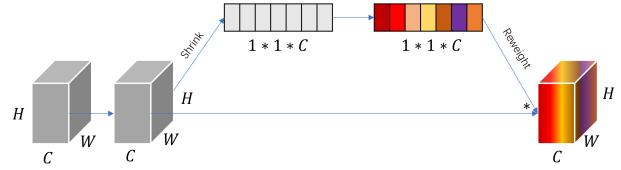


Figure 4. The design of RCAB [22]. The weights on feature map channels are calculated by convolution layers. They adjust weights automatically.

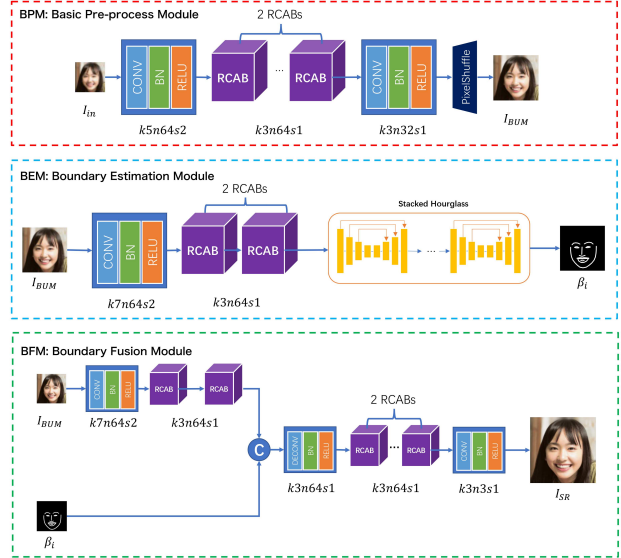


Figure 5. The overview of ‘BPM’, ‘BEM’ and ‘BFM’. ‘k5n64s2’ indicates the kernel size of 5, the number of feature maps to be 64 and stride at 2. The enhanced facial boundaries are estimated with 11 channels, but for intuitive visualization, we display all boundary lines in one heatmap.

Basic Pre-process Module (BPM). The structure of BPM is displayed in Figure 5. BPM plays the role of increasing spatial resolution of LR inputs. In the BPM design, we optimize ESPCN [23] with RCABs. The design can easily change a little with the last pixel-shuffle layer to deal with different upscale factors. Moreover, BPM allows our network to succeed in end-to-end training and testing. The objective function l_U is as follows:

$$l_U = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} |I_{BPM} - I_{gt}|, \quad (1)$$

where W or H indicates the width or height of LR inputs, r denotes the upscale factor of BPM. The mean absolute error (MAE) between upscaled image I_{BPM} and ground truth I_{gt} is calculated.

Benefiting from the design of the network structure and

loss functions, BPM reaches a perfect balance between computational cost and reconstruction quality.

Boundary Estimation Module (BEM). We use a stacked hourglass structure H to predict pre-defined K facial boundaries directly from the output of the first module I_{BPM} . The predicted K facial boundaries are presented as a stack of heatmaps β_i :

$$[\beta_1, \beta_2, \dots, \beta_K] = H_s(H_{s-1} \dots (H_2(H_1(I_{BPM}))))), \quad (2)$$

where s denotes the number of hourglass blocks.

The BEM does not require any pretraining. The mean squared error (MSE) between estimated facial boundaries and ground truth facial boundaries is calculated as Equation 3:

$$l_E = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H |\beta_i - \hat{\beta}_i|^2, \quad (3)$$

where $\{\beta_i\}_{i=1}^K$ means estimated boundary lines and $\{\hat{\beta}_i\}_{i=1}^K$ is ground truth boundary lines.

Boundary Fusion Module (BFM). Finally, BFM takes in β_i and I_{BPM} with its two-branch design and hallucinates face images from these cues. We use a series of RCABs to project our I_{BPM} into high-dimensional representations V_j :

$$V_j = R_n(R_{n-1}(R_{n-2} \dots R_2(R_1(F(I_{BPM}))))), \quad (4)$$

where n denotes the number of RCAB blocks and F means mapping from images to basic feature maps.

Afterwards, we concatenate β_i and V_j into a decoder-like network for recovering final SR images I_{sr} . We also introduce RCAB in the decoder to better exploit interdependencies among facial boundaries and high-dimensional representations. BFM measures the distance between final output I_{sr} and ground truth I_{gt} with MAE loss:

$$l_F = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H |F(R_2(R_1(F^{-1}(V_j, \beta_k)))) - I_{gt}|, \quad (5)$$

where $F^{-1}(V_j, \beta_k)$ denotes the deconvolution layer, $R_1(\cdot)$ or $F(\cdot)$ is the RCAB or convolution layer respectively.

3.3. Coarse-to-Fine Boundary Supervision

In FSR, facial structural information, which can be seen as context information, plays an essential role in recovering face images from very small LR inputs. Different from face alignment tasks, the inputs of FSR are under a low-resolution scale, making it more challenging to obtain sufficient information across different resolution stages. Therefore, we design a coarse-to-fine supervision to constrain the

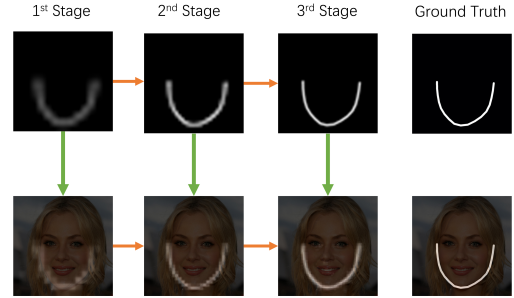


Figure 6. Illustration of coarse-to-fine boundary supervision. For better visual display, we only display one enhanced facial boundary line on our image. On the first stage, we use thicker boundary lines for better guiding component localization in LR images. When the stage goes deeper and the image spatial resolution gets larger, our enhanced facial boundaries get thinner and clearer.

fineness of enhanced facial boundaries at different resolution scales.

Due to the multi-stage FSR network design, the spatial resolution of face images increases step by step. On earlier stages, the downsampling procedure of hourglass decreases the image size into extremely low resolution, leading to the difficulty of estimating exact enhanced facial boundaries. As demonstrated in Figure 2, the coarse-to-fine supervision is realized by applying different gaussian kernels at different stages to control the fineness of facial boundaries numerically. The thickness of ground-truth facial boundary lines is 2. With the application of gaussian kernels, the thickness of facial boundary lines expands. On earlier stages, we apply a larger kernel on the ground-truth facial boundaries. However, on latter stages, the ground-truth facial boundaries are applied with smaller kernels. Figure 6 demonstrates an example of the coarse-to-fine supervision. With the additional supervision on the enhanced facial boundaries, our network produces more realistic face images.

3.4. MSFSR-GAN

In recent years, GAN-based approaches have achieved a good visual effect for image generation. Inspired by SR-GAN [16], we introduce an extra discriminator into distinguishing the super-resolved images from real HR images.

Our discriminator network consists of 8 convolution layers. We use a convolution layer with kernel size of 1 to replace a fully convolution layer in order to reduce computational cost. The objective function of the adversarial network D is expressed as follows:

$$l_{GAN} = E[\log D(I_{gt})] + E[\log(1 - D(I_{sr}, I_{gt}))], \quad (6)$$

where E is the expectation of the log probability while D

denotes the generative model.

Meanwhile, we also introduce the perceptual loss l_{VGG} for further improvement on perceptual quality. The perception loss is formulated as follows:

$$l_{VGG} = \frac{1}{W_k H_k} \sum_{x=1}^{W_k} \sum_{y=1}^{H_k} (\Phi_k(I_{gt})_{x,y} - \Phi_k(I_{sr})_{x,y})^2, \quad (7)$$

where Φ_k denotes pre-trained VGG16 [24] features at a given layer k .

On early stages of our network, increasing spatial resolution is the priority task of FSR and the GAN-based network introduces extra noises and distortions. Therefore, we only apply GAN loss and perceptual loss on the final stage of our network. The overall loss function is shown as follows:

$$l_{total} = \begin{cases} l_U + l_E + l_F + \alpha l_{GAN} + \gamma l_{VGG}, & s = n; \\ l_U + l_E + l_F, & otherwise, \end{cases} \quad (8)$$

where α denotes coefficient of GAN loss and γ is the trade-off between perceptual loss and others. In our network, we set $\alpha = 1 \times 10^{-3}$ and $\gamma = 6 \times 10^{-3}$.

4. Experiments

4.1. Implementation Details

Datasets. We use CelebAMask-HQ [25], which provides accurate ground-truth face parsings on CelebA-HQ [26] dataset, to evaluate our model performance. In addition, Helen [27] and WFLW [13] are applied to demonstrate the superiority of the proposed approach on public datasets comprehensively. We follow the preprocess procedure of algorithm [4] on CelebAMask-HQ dataset. For the preprocess procedure on Helen and WFLW, we adopt the latest MaskGAN [25] to generate ground-truth face parsings. For CelebAMask-HQ, we randomly select 17,000 images for training and the remaining 13,000 images for testing. For Helen dataset, we randomly choose 1,200 images for training and the rest 400 images for testing. As for WFLW dataset, we select 2,000 images for training and leave the rest 600 images for testing. In this way, all the training sets are not overlapped with these testsets.

Training Settings. We firstly use MTCNN [28] network to crop main faces on Helen [27] and WFLW [13] datasets. Together with CelebAMask-HQ [26] face images, we resize them to 128×128 as ground-truth. In addition, we further resize these images with bilinear downsampling into 64×64 as targets of stage-2, 32×32 as targets of stage-1 and 16×16 as LR inputs. We use a 68-point landmark detection method [29] to generate ground-truth landmark points. By connecting them based on facial semantical meanings, 11 enhanced facial boundaries are regarded as ground-truth boundaries

(e.g., face outer contours are generated by connecting Point 0 to Point 18). Detailed information of connecting landmark points to boundary lines can be found in our project site ¹.

We implement our model using Pytorch [30] on a computer with 32GB memory. The initial learning rate of our network is set to 1×10^{-4} and halved at different iterations (15k, 30k, 45k, 90k, 180k). We adopt Adam [31] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size is set to 16 and the whole network takes about 48 hours to train with one 1080Ti GPU.

4.1.1 Evaluation Protocols

For fair comparison, we evaluate our model using the standard image SR metrics: PSNR, SSIM [32] and Perception Index (PI) [33]. In addition, face SR aims at dealing with the domain-specific problem, so we adopt the Face Similarity (FS) calculated by the open-platform Face++’s face comparing API [34] for evaluating the quality of recovered face images.

4.2. Ablation Study

4.2.1 Effectiveness of Enhanced Facial Boundaries

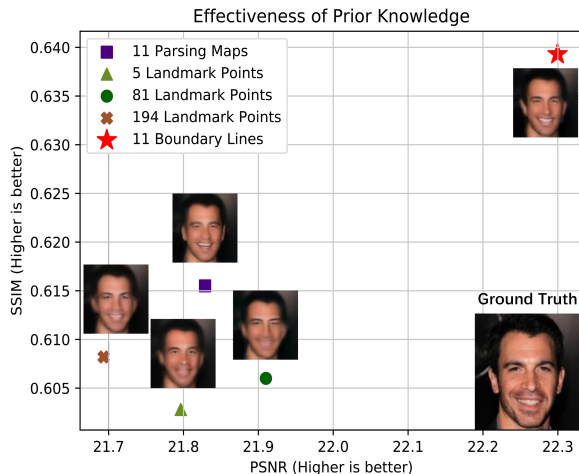


Figure 7. The effectiveness of different facial structural representations on CelebA-HQ dataset. The result is evaluated on the network without GAN.

To validate the effectiveness of the enhanced facial boundaries, we compare the performance of FSR network with different structural representations in Figure 7: 11 enhanced facial boundaries (denoted as red star of Figure 7), 81 landmark points (denoted as green node of Figure 7), 194 landmark points (denoted as brown cross of Figure 7) and 11 face parsing maps (denoted as purple square of Figure

¹<https://github.com/cydiachen/MSFSR/>

7). For fair comparison, we design a simplified one-stage FSR network, which only changes BPM’s last Pixel-Shuffle layer to directly output $8\times$ images.

Figure 7 shows the performance of our network with different types of facial structural representations. It can be observed that:

(1) When the number of landmark points increases, our network generates better face images. Moreover, facial parsing maps provide pixel-level information of facial structures, the network with facial parsing maps outperforms those with facial landmark points. The phenomenon indicates a positive correlation between the performance of the network and the semantic information carried by facial representations.

(2) The stacked hourglass design shows its own limit in estimating complex facial representations. The network performs well on facial representations of 5 landmark points, 11 facial parsing maps and 81 landmark points. However, the structure suffers from severe performance degradation when dealing with dense landmark points like 194.

(3) Compared with other facial representations, our network with enhanced facial boundaries achieves the best performance in both PSNR and SSIM criteria. Therefore, we adopt enhanced facial boundaries as facial structural representation in our FSR model.

4.2.2 Effectiveness of Coarse-to-Fine Supervision

To testify the effectiveness of coarse-to-fine supervision for our multi-stage network, we make performance comparison between our MSFSR and MSFSR without coarse-to-fine strategy. As demonstrated in Table. 1, we can see that the network without coarse-to-fine supervision (setting-6) is inferior to the network with coarse-to-fine supervision. We conduct several experiments to verify the best settings for coarse-to-fine supervision. As described in Section 3.3, we should apply a large gaussian kernel on the first stage to relieve the difficulty of estimating accurate facial boundaries from a tiny face image. When the spatial resolution of the face image increases, the size of gaussian kernel gradually

Settings	1	2	3	4	5	6
Kernel Size 1	11	9	9	9	5	0
Kernel Size 2	9	5	5	5	3	0
Kernel Size 3	5	5	3	0	0	0
PSNR	22.8	24.2	23.2	23.1	22.9	22.4

Table 1. PSNR performance of our 3-stage MSFSR with different supervision strategies on CelebA-HQ dataset at $8\times$ SR. The kernel size controls the fineness of supervision and a larger value indicates a coarser setting. The last column with kernel size of 0 denotes no supervision on the multi-stage network.

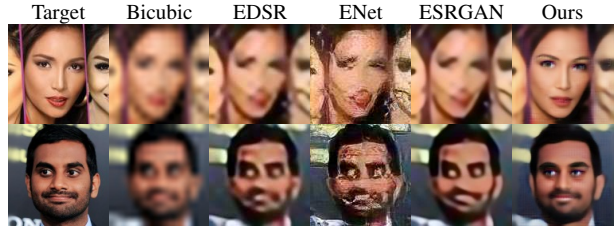


Figure 8. Qualitative comparisons with the state-of-the-art general image SR approaches on CelebA-HQ, Helen and WFLW datasets at $8\times$ SR. The ‘ENet’ is short for EnhanceNet [35].

decreases. As shown in Table. 1, we can see that either setting-1 or setting-5 degrades the performance compared with other settings. On the contrary, when the kernel size at the last stage of the network increases (setting-2, setting-3 and setting-4), the performance of the network is improved. The phenomenon demonstrates that an appropriate setting of supervision also helps the performance improvement on the multi-stage network. In our network setting, we use gaussian kernel size of 9 on stage-1, kernel size of 5 on stage-2 and kernel size of 5 on stage-3.

4.2.3 Comparison with State-of-the-Art Approaches

In this section, we compare our MSFSR network with state-of-the-art FSR algorithms on the testset of CelebA-HQ, WFLW and Helen separately. We choose Bicubic, EDSR [1], EnhanceNet [35] and ESRGAN [36] as general image SR baselines. We reimplement URDGN [2], FSRNet [4] and PFSR [3] with Pytorch[30] as FSR baselines. The reimplemented approaches are trained with the same dataset as our network. The comprehensive experimental evaluation on our test set with $8\times$ upscaling factors shows that the proposed MSFSR and MSFSR-GAN can generate more realistic and visually pleasing textures compared to the state-of-the-art approaches. We also display more results in our supplementary material.

As demonstrated in Figure 8, general SR approaches simply increase the spatial resolution of input images. EDSR [1] tends to generate face images with distorted geometric structures. EnhanceNet [35] introduces a large number of false details. The artifacts are mainly caused by the ignorance of facial geometric structures. By contrast, face-specific approaches reconstruct more realistic faces. As shown in Figure 9, FSRNet [4] generates realistic features on human faces, but some facial components are misplaced. The performance degradation of FSRNet should be attributed to the insufficient ability of the unified framework in predicting complex face representations and the additional errors introduced by the predefined upscaling operation. URDGN [2] tends to hallucinate faces with wrong identities. URDGN shares the same network structure with

Dataset	Bicubic	EDSR	EnhanceNet	ESRGAN	URDGN	FSRNet	PFSR	Ours	Ours-GAN
CelebA-HQ	22.39 / 0.60	23.02 / 0.63	20.64 / 0.46	22.26 / 0.62	19.70 / 0.49	21.32 / 0.67	23.87 / 0.69	26.60 / 0.80	25.87 / 0.78
Helen	22.59 / 0.62	23.18 / 0.64	20.67 / 0.46	21.43 / 0.60	18.88 / 0.45	22.07 / 0.60	22.92 / 0.65	25.59 / 0.75	24.12 / 0.71
WFLW	22.32 / 0.60	22.99 / 0.64	20.57 / 0.46	21.63 / 0.61	19.03 / 0.46	22.08 / 0.60	23.13 / 0.65	26.63 / 0.73	25.43 / 0.72

Table 2. PSNR, SSIM for the state-of-the-art approaches on CelebA-HQ, Helen and WFLW at $8\times$ SR. Red text indicates the best, blue text indicates the second best and green text indicates the third best performance.

SRGAN [16]. The extracted facial attributes are generated by the network automatically from the distribution of training samples as they are not robust in modeling facial structures. The PFSR [3] tends to generate unpleasant facial components with wrong textures, especially on human eyes. This artifacts are mainly caused by the absence of several facial landmark points. Conversely, the proposed MSFSR generates more accurate face images without artifacts on human faces. The qualitative comparisons with the state-of-the-art approaches demonstrate the effectiveness of our enhanced facial boundaries and multi-stage design.

We evaluate the results by SR metrics: PSNR (evaluated on the luminance channel in YCbCr color space) and SSIM. The quantitative results on different competitors in different public datasets are displayed in Table. 2. The proposed MSFSR shows no extra bias on particular public datasets. Comparing with the state-of-the-art approaches, our MSFSR achieves the highest scores on PSNR and SSIM.

We also evaluate the results with Perception Index (PI) and Face Similarity (FS). The PFSR ranks first, and the URDGN ranks second in PI scores. However, the recovered textures of PFSR and URDGN are visually unpleasant and implausible as shown in Figure 9, which indicates that the PI score is not suitable to assess the quality of reconstructed face images.

As shown in Figure 9, the proposed methods outperform other FSR approaches in FS scores with a large margin. The FS scores measure the similarity between reconstructed faces and ground-truth faces. The proposed methods rank the first and the second on line 1 and line 2. The FSRNet [4] ranks the third in face images on line 1 and line 2, but it also generate face images with wrong identities and blurry facial details. The PFSR [3] ranks the second on line 3. The result shows that the progressive design of FSR network can effectively improve the fidelity of reconstructed face images in the task of FSR.

5. Conclusion

In this paper, we have presented a novel MSFSR model for FSR. To the best of our knowledge, this model is the first FSR network that adopts the enhanced facial boundary lines as an accurate facial representation. Moreover, we design three optimized modules to improve the capability of FSR model. To further improve our method, we propose a



















Target	FSRNet	URDGN	PFSR	Ours	Ours-GAN
					
PSNR/SSIM	21.28/0.60	19.42/0.48	22.91/0.65	24.43/0.71	23.42/0.67
PI/FS	23.5/77.28	21.9/42.02	21.7/76.44	22.5/85.32	21.7/83.77
					
PSNR/SSIM	21.22/0.59	20.30/0.48	23.15/0.65	24.90/0.72	24.80/0.69
PI/FS	23.6/82.49	22.4/64.49	22.3/63.53	23.3/90.28	22.4/88.68
					
PSNR/SSIM	19.15/0.55	18.54/0.49	22.50/0.70	23.91/0.75	23.81/0.74
PI/FS	23.3/36.69	22.1/53.52	22.0/62.50	23.1/59.60	22.7/65.61

Figure 9. Qualitative comparisons with the state-of-the-art face SR approaches on CelebA-HQ, Helen and WFLW datasets at $8\times$ SR. Red text indicates the best, blue text indicates the second best and green text indicates the third best performance.

multi-stage network with coarse-to-fine supervision on facial boundaries. Based on the enhanced facial boundaries and optimized network structure, we obtain state-of-the-art performance and generate more realistic face images with precise facial details in comparison to other SR algorithms at $8\times$ upscale factors.

Acknowledge

The work was supported in part by the National Natural Science Foundation of China under Grant 61901117, Grant U1805262, Grant 61701118, Grant 61571128, and Grant 61871131.

References

- [1] 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 conference on computer vision and pattern recognition workshops (CVPRW)*, pages 136–144, 2017.
- [2] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *European conference on computer vision*, pages 318–333. Springer, 2016.

- [3] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019.
- [4] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2018.
- [5] L. Chen, J. Pan, and Q. Li. Robust face image super-resolution via joint learning of subdivided contextual model. *IEEE Transactions on Image Processing*, 28(12):5897–5909, 2019.
- [6] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (9):1167–1183, 2002.
- [7] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *European conference on computer vision*, pages 614–630. Springer, 2016.
- [8] Mengyan Li, Yuechuan Sun, Zhaoyu Zhang, and Jun Yu. A coarse-to-fine face hallucination method by exploiting facial prior knowledge. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 61–65. IEEE, 2018.
- [9] Junjun Jiang, Yi Yu, Jinhui Hu, Suhua Tang, and Jiayi Ma. Deep cnn denoiser and multi-layer neighbor component embedding for face hallucination. *arXiv preprint arXiv:1806.10726*, 2018.
- [10] 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.
- [11] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 908–917, 2018.
- [12] Cheng-Han Lee, Kaipeng Zhang, Hu-Cheng Lee, Chia-Wen Cheng, and Winston Hsu. Attribute augmented convolutional neural network for face hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 721–729, 2018.
- [13] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2138, 2018.
- [14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision (CVPR)*, pages 184–199. Springer, 2014.
- [15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.
- [16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- [17] 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.
- [18] Chih-Chung Hsu, Chia-Wen Lin, Weng-Tai Su, and Gene Cheung. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 2019.
- [19] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 183–198, 2018.
- [20] 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 (ECCV)*, pages 217–233, 2018.
- [21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision (ECCV)*, pages 483–499. Springer, 2016.
- [22] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.
- [23] Wenzhe Shi, Jose Caballero, Ferenc Huszár, 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, 2016.
- [24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [25] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*, 2019.
- [26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [27] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.
- [28] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded

- convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
- [29] Davis E. King. Dlib c++ library. <http://dlib.net/>.
- [30] Nikhil Ketkar. Introduction to pytorch. In *Deep learning with python*, pages 195–208. Springer, 2017.
- [31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [33] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *The European Conference on Computer Vision (ECCV) Workshops*, 2018.
- [34] Megvii. Face++ compare api. <https://www.faceplusplus.com/face-comparing/>.
- [35] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.
- [36] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision (ECCV) Workshops*, 2018.