The Visual Computer (2021) 37:2855–2867
https://doi.org/10.1007/s00371-021-02236-w

ORIGINAL ARTICLE

Unsupervised face super-resolution via gradient enhancement and semantic guidance

Luying Li[1] · Junshu Tang[1] · Zhou Ye[2] · Bin Sheng[1] · Lijuan Mao[3] · Lizhuang Ma[1,4]

Accepted: 27 June 2021 / Published online: 23 July 2021

Abstract

Face super-resolution aims to recover high-resolution face images with accurate geometric structures. Most of the conventional super-resolution methods are trained on paired data that is difficult to obtain in the real-world setting. Besides, these methods do not fully utilize facial prior knowledge for face super-resolution. To tackle these problems, we propose an end-to-end unsupervised face super-resolution network to super-resolve low-resolution face images. We propose a gradient enhancement branch and a semantic guidance mechanism. Specifically, the gradient enhancement branch reconstructs high-resolution gradient maps, under the restriction of two proposed gradient losses. Then the super-resolution network integrates features in both image and gradient space to super-resolve face images with geometric structure preservation. Moreover, the proposed semantic guidance mechanism, including a semantic-adaptive sharpen module and a semantic-guided discriminator, can reconstruct sharp edges and improve local details in different facial regions adaptively, under the guidance of semantic parsing maps. Qualitative and quantitative experiments demonstrate that our proposed method can reconstruct high-resolution face images with sharp edges and photo-realistic details, outperforming the state-of-the-art methods.

Keywords Unsupervised face super-resolution · Facial semantic priors · Gradient enhancement

1 Introduction

Image super-resolution (SR) aims to reconstruct high-resolution (HR) images from the observed low-resolution (LR) inputs. Face super-resolution, also known as face hallucination, is a special case of image super-resolution. It has attracted increasing attention for its widespread application in surveillance [1,2], photo restoration [3], face recognition [4,5], etc. Most notably, it is difficult to capture LR-HR image pairs of human faces in the real-world setting, which poses challenges to the face super-resolution task. Hence, in this work, we aim at recovering the corresponding HR face images from LR inputs via unsupervised learning.

A great number of deep-learning methods have been proposed to reconstruct HR images from LR inputs. Recent works [6–8] mostly apply generative adversarial networks (GAN) [9] to recover photo-realistic HR images. ESRGAN [7] introduces a perceptual loss [10] that is calculated in high-level feature space to improve the perceptual quality. SPSR [8] utilizes gradient maps of LR and HR images to provide structural priors for the super-resolution process. These methods have shown good performance in conventional SR tasks. However, they are not competitive when super-resolving face images and they cannot tackle the SR tasks without paired data.

The difficulties of unsupervised face super-resolution lie in the following aspect. First, the LR face images of tiny

✉ Lijuan Mao
maolijuan@sus.edu.cn

✉ Lizhuang Ma
ma-lz@cs.sjtu.edu.cn

Luying Li
liluying@sjtu.edu.cn

Junshu Tang
tangjs@sjtu.edu.cn

Zhou Ye
yezhou@cls.cn

Bin Sheng
shengbin@cs.sjtu.edu.cn

1 Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

2 Shanghai CLS Fintech Co., LTD, Shanghai 200030, China

3 Shanghai University of Sport, Shanghai 200438, China

4 School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

scale provide less information compared to the ordinary LR images. Second, the lack of paired data makes the training process unstable and hard to train. Third, the facial geometric structures and identity information should be reconstructed correctly in the HR outputs.

To tackle these difficulties, several methods extract facial prior knowledge, such as landmarks [11,12], parsing maps [12,13], and facial attributes [14], to recover HR face images while preserving facial structures. Also, to overcome the lack of paired data, an intermediate LR domain [15,16] is introduced for the transformation from LR domain to HR domain. However, these methods still cannot reconstruct photo-realistic high-resolution face images, particularly in an unsupervised manner.

To this end, we propose an unsupervised face super-resolution network **(GESGNet)** with gradient enhancement and semantic guidance. We propose a gradient branch that reconstructs HR gradient maps of face images. Furthermore, we propose a statistical gradient loss and a pixel-wise gradient loss to encourage the reconstruction. Then the super-resolution network concatenates features in image space and that in gradient space to super-resolve face images while maintaining geometric structures.

Moreover, we propose a semantic guidance mechanism. Specifically, to further retain facial geometric structures, we propose a semantic loss by calculating semantic maps through a pre-trained face parsing network. We also propose a semantic-adaptive sharpen module to sharpen and enhance details adaptively, under the guidance of semantic maps. Besides, a semantic-guided discriminator that discriminates on different facial components is proposed to generate diverse details.

The main contributions of this paper are as follows:

- We propose an unsupervised face super-resolution network **(GESGNet)** to reconstruct high-resolution face images. To the best of our knowledge, this is the first attempt to employ facial semantic priors and gradient information for unsupervised face SR task.
- We propose a gradient enhancement branch and two gradient losses to recover HR gradient maps. The extracted gradient features can encourage to super-resolve images with accurate geometric structures and sharp edges.
- We propose a semantic guidance mechanism including a semantic-guided discriminator and a semantic-adaptive sharpen module, which can further preserve geometric structures and generate diverse details for different facial components.
- We implement detailed experiments on our constructed dataset. The qualitative and quantitative results show that our method can recover photo-realistic HR face images and outperforms state-of-the-art methods.

2 Related work

2.1 Unsupervised image super-resolution

Image super-resolution aims at recovering HR images from the LR counterparts, which has become a significant task in the field of computer vision for its widespread application in surveillance [1,2], image enhancement [17], medical imaging [18], face recognition [4,5], etc. Earlier works utilized prediction-based methods [19], edge-based methods [20,21], statistical methods [22,23], and patch-based methods [24,25] to reconstruct HR images. Recently, with the rapid development of deep learning techniques, a large number of deep-learning based super-resolution methods [6,7,26,27] have been proposed and shown impressive performance. Dong et al. [26] proposed SRCNN, firstly employed CNN-based methods for image super-resolution task. Ledig et al. [6] proposed a generative adversarial network with a perceptual loss to reconstruct photo-realistic HR images. The adversarial learning was also adopted in Enhancenet [27] and ESRGAN [7], demonstrating the powerful ability of GAN models for image super-resolution task. Though these GAN-based super-resolution methods can recover high-fidelity HR images, they tend to generate geometric distortions and unsharp edges. To address this issue, Ma et al. [8] proposed a gradient-guided SR method. They reconstructed HR gradient maps from gradient maps of LR images to provide structural priors for the image super-resolution process. Encouraged by their success, we introduce a gradient branch and propose two gradient losses to preserve geometric structures and generate sharp edges.

However, note that most SR methods super-resolve images with paired data, which is difficult to obtain in real-world setting. To address this issue, several researchers proposed unsupervised image super-resolution methods. Among them, Yuan et al. [28] proposed a cycle-in-cycle network structure. They mapped the input domain into a noise-free LR domain through the first CycleGAN-based network and then transformed the intermediate domain to the HR domain through the second network. Based on [28], Zhang et al. [29] proposed progressive multiple cycle-in-cycle networks, which can generate clear structures and reasonable textures. Fritsche et al. [15] treated the low and the high image frequencies separately by applying the pixel-wise loss only on low frequencies while adversarial loss only on high frequencies. They introduced an intermediate LR domain to divide the SR process into the first unsupervised stage and the second supervised stage. Zhou et al. [16] employed an intermediate LR domain as the previous works and proposed a color-guided domain mapping to alleviate the color shift in domain transformation. Although these intermediate LR domains play an important role in the learning process of unsupervised image super-resolution, the transformation from input LR domain to intermediate LR

domain is extremely difficult for face super-resolution due to the tiny scale of inputs. Thus, instead of the intermediate LR domain, we first convert the input LR domain into an intermediate HR domain and then convert it into the real HR domain.

2.2 Face super-resolution

Face super-resolution is a special case of image super-resolution, which requires facial prior knowledge to reconstruct accurate geometric structures and diverse facial details. Several attempts have been made to utilize facial prior knowledge for face super-resolution, such as facial component heatmaps, facial landmarks, identity attributes, and semantic parsing maps. Choudhury et al. [30] detected facial landmarks first and then searched for the matching facial components from a dictionary of training face images. Yu et al. [31] estimated facial component heatmaps and then concatenated the heatmaps with image features in the super-resolution network. Chen et al. [12] utilized two branches to extract image features, and estimate facial landmarks and parsing maps, respectively. Then the extracted image features and facial prior knowledge were combined and sent to the decoder to reconstruct HR images. Bulat et al. [32] proposed a face-alignment branch that localized facial landmarks on the reconstructed images to enforce facial structural consistency between the LR images and the reconstructed HR images. Yin et al. [11] proposed a joint network for face super-resolution and alignment, where these two tasks shared deep features and benefited each other. Yu et al. [14] encoded LR images with facial attributes when super-resolving images, and then embedded attributes into the discriminator to examine whether the reconstructed images contain desired attributes or not. Xin et al. [33] extracted facial attributes as semantic-level representation and then combined them with pixel-level texture information to recover HR images. Wang et al. [34] proposed a network that took both facial parsing maps and LR images as inputs to reconstruct HR images. Zhao et al. [13] jointly trained a face super-resolution network and a face parsing network. They extracted facial priors through a semantic attention adaptation module that bridged the two networks.

These methods can reconstruct high-quality HR face images, outperforming generic SR methods, which indicates the significance of facial prior knowledge for face super-resolution. However, most of these methods employ facial priors by designing an auxiliary network or training multiple tasks jointly, which requires more computational resources. Besides, the extracted facial priors are mainly used for structure preservation, not used for generating diverse details among various facial components. Instead, we propose a semantic guidance mechanism, where the semantic parsing maps are calculated through a pre-trained facial parsing network. Specifically, our proposed semantic-guided discriminator, semantic-adaptive sharpen module, and semantic loss can reconstruct accurate geometric structures and generate diverse details for different facial components.

3 Proposed method

3.1 Overview

The aim of our method is to reconstruct corresponding high-resolution face images from low-resolution inputs on unpaired data. The overall framework of our proposed method is shown in Fig. 2. The unpaired dataset consists of LR images $I_x \in \mathcal{X}$ and HR images $I_y \in \mathcal{Y}$. To reconstruct HR images in an unsupervised manner, we designed a cycle network structure that consists of two generators, G_{ZY} and G_{YZ}, a gradient branch G_{gra}, three discriminators, D_Y, D_Z and D_{sm}, as well as a pre-trained upsample generator G_{XZ}.

For a given LR image $I_x \in \mathcal{X}$, it is firstly upsampled to I_z by a pre-trained ESRGAN model G_{XZ}. To effectively learn the geometric representation, we propose a gradient branch G_{gra} and two gradient loss functions. G_{gra} takes the gradient map of I_z as input to reconstruct a high-resolution gradient map. Then, G_{ZY} concatenates feature maps from the gradient branch G_{gra} to reconstruct image $\hat{I}_y$ with gradient information. Moreover, to recover HR images with sharp edges, we propose a semantic-adaptive sharpen module (SASM), which is embedded into G_{ZY}. The proposed SASM sharpens facial components with different degrees according to semantic parsing maps, and thus can sharpen different facial regions adaptively.

The discriminators distinguish the synthesized data from the real data to improve the reconstructing ability of generators. In particular, we propose a semantic-guided discriminator D_{sm} that can discriminate on different regions, respectively, under the guidance of semantic parsing maps. In this way, D_{sm} enables G_{ZY} to reconstruct HR images with diverse details in different facial components.

3.2 Gradient enhancement branch

Generating sharp edges and fine-grained details is important but challenging when super-resolving images. Most of the previous works [6,7,27] try to improve sharpness and fidelity through optimization in image space. However, these methods still cannot reconstruct sharp edges and details as that in real HR images. Gradient maps of images can reflect the sharpness of edges. We find that there are huge differences between gradient maps of LR images and that of HR images, as shown in Fig. 1. The gradient maps of HR images are with clearer edges and stronger contrast between the high and the low intensity. Thus, we hope to utilize gradient information

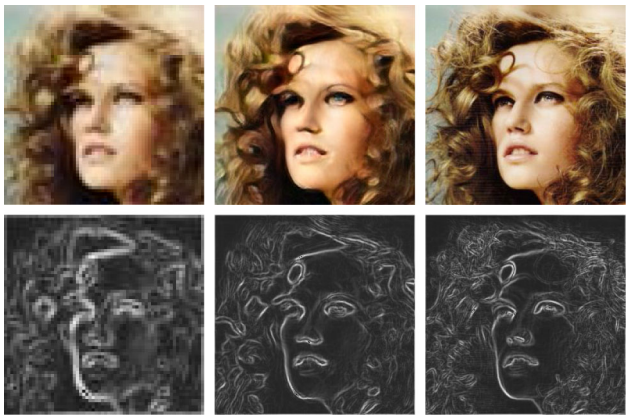

Fig. 1 Visualization of gradient maps. From left to right, we show the image and its gradient map in LR domain, intermediate domain, and HR domain

to guide face super-resolution. Ma et al. [8] built a gradient branch for supervised SR, which shows effectiveness in preserving geometric structures and edge sharpness. Encouraged by [8], we build a gradient enhancement branch G_{gra} as shown in Fig. 2, which takes LR gradient maps of $I_z \in \mathcal{Z}$ as input and estimates HR gradient maps. Then the super-resolution network G_{ZY} integrates the gradient features and the previous image features to reconstruct super-resolution images $\hat{I}_y = G_{ZY}(I_z)$.

The gradient map $\mathcal{G}(I_z)$ of an image $I_z \in \mathcal{Z}$ can be described as

$$\begin{aligned} \nabla_h(I_z) &= I_z(\mathsf{x}+1,\mathsf{y}) - I_z(\mathsf{x}-1,\mathsf{y}) \ , \\ \nabla_v(I_z) &= I_z(\mathsf{x},\mathsf{y}+1) - I_z(\mathsf{x},\mathsf{y}-1) \ , \\ \nabla(I_z) &= (\nabla_h(I_z), \nabla_v(I_z)) \ , \\ \mathcal{G}(I_z) &= \|\nabla(I_z)\|_2 \ , \end{aligned} \tag{1}$$

where (x, y) are pixel coordinates of image I_z.

Since G_{gra} aims to estimate gradient maps of images in real HR domain, we first propose a statistical gradient loss to make the estimated gradient maps $G_{gra}(\mathcal{G}(I_z))$ have the same intensity distribution as the gradient maps $\mathcal{G}(I_y)$ of real HR images I_y. The statistical gradient loss is formulated as

$$\mathcal{L}_{gra\_s} = \mathbb{E}_{I_z, I_y}\left[\|\mathcal{H}\left(G_{gra}(\mathcal{G}(I_z))\right) - \mathcal{H}\left(\mathcal{G}(I_y)\right)\|_1\right] \ , \tag{2}$$

where $I_z \in \mathcal{Z}$, $I_y \in \mathcal{Y}$, and $\mathcal{H}(\cdot)$ is the intensity histogram of gradient map.

Besides, the estimated gradient maps $G_{gra}(\mathcal{G}(I_z))$ should retain geometric structures as $\mathcal{G}(I_z)$. Hence, we propose a pixel-wise gradient loss, which is formulated as

$$\mathcal{L}_{gra\_p} = \mathbb{E}_{I_z}\left[\|G_{gra}(\mathcal{G}(I_z)) - \mathcal{G}(I_z)\|_1\right] \ . \tag{3}$$

The combination of $\mathcal{L}_{gra\_s}$ and $\mathcal{L}_{gra\_p}$ enables the estimated gradient maps $G_{gra}(\mathcal{G}(I_z))$ to have the similar intensity and distribution as gradient maps of real HR images. In this way, our proposed method can reconstruct HR images as

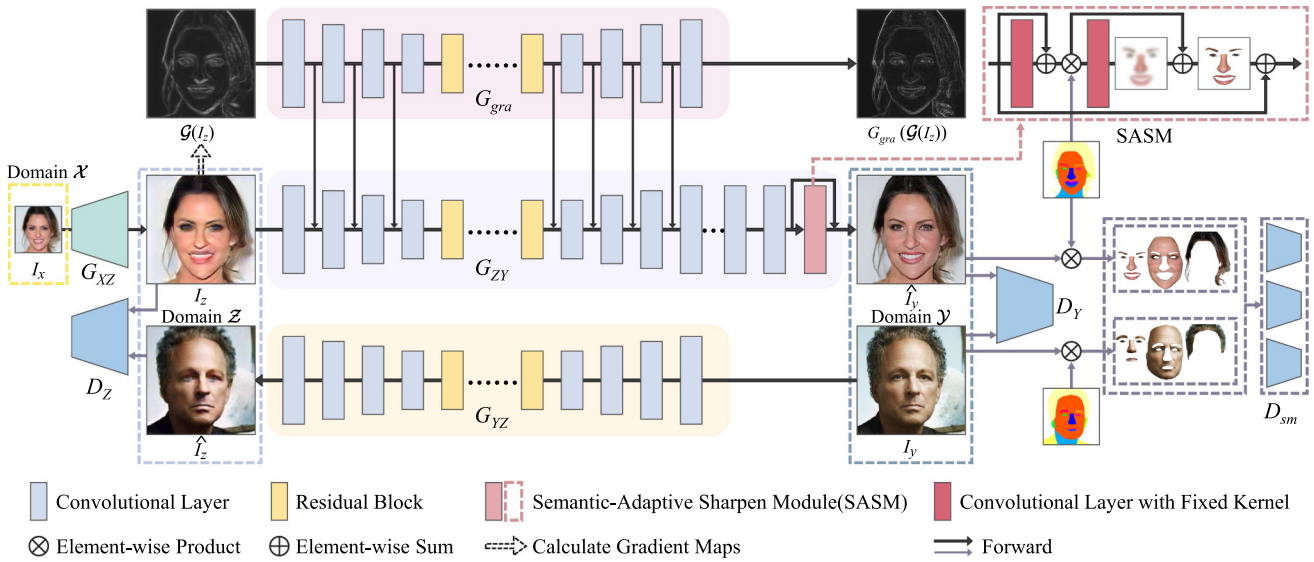

Fig. 2 Overall framework of our proposed method. Given an input LR image I_x, we aim to recover the corresponding HR image $\hat{I}_y$. G_{XZ} converts I_x to I_z. Then G_{YZ} and G_{ZY} enable unsupervised transformation between domain $\mathcal{Z}$ and $\mathcal{Y}$. We propose a gradient branch G_{gra}, a semantic-adaptive sharpen module (SASM), and a semantic-guided discriminator D_{sm} to reconstruct photo-realistic HR face images with geometric structure preservation

sharp as real HR images, while preserving geometric structures simultaneously.

3.3 Semantic guidance mechanism

For stable unsupervised transformation from domain $\mathcal{Z}$ to domain $\mathcal{Y}$, we apply an adversarial loss [9], a cycle loss [35], and an identity loss [35], which are defined as

$$\begin{aligned}\mathcal{L}_{adv} &= \mathbb{E}_{I_y}\left[\|D_Z(G_{YZ}(I_y)) - 1\|_2\right] \\ &\quad + \mathbb{E}_{I_z}\left[\|D_Y\left(G_{ZY}(I_z)\right) - 1\|_2\right], \end{aligned} \tag{4}$$

$$\begin{aligned}\mathcal{L}_{cyc} &= \mathbb{E}_{I_y}\left[\|G_{ZY}\left(G_{YZ}(I_y)\right) - I_y\|_1\right] \\ &\quad + \mathbb{E}_{I_z}\left[\|G_{YZ}\left(G_{ZY}(I_z)\right) - I_z\|_1\right], \end{aligned} \tag{5}$$

$$\mathcal{L}_{idt} = \mathbb{E}_{I_y}\left[\|G_{ZY}(I_y) - I_y\|_1\right] + \mathbb{E}_{I_z}\left[\|G_{YZ}(I_z) - I_z\|_1\right]. \tag{6}$$

However, the reconstructed geometric structures are easy to distort and blur during the unsupervised learning process. To address this issue, we propose a semantic guidance mechanism to preserve geometric structures with the help of facial semantic parsing maps.

Unsupervised semantic loss. To accurately preserve geometric structures and generate clear boundaries during unsupervised super-resolution, we propose a semantic loss, which is defined as

$$\begin{aligned}\mathcal{L}_{sm} &= \mathbb{E}_{I_z}\left[\|\psi\left(G_{ZY}(I_z)\right) - \psi(I_z)\|_1\right] \\ &\quad + \mathbb{E}_{I_y}\left[\|\psi\left(G_{YZ}(I_y)\right) - \psi(I_y)\|_1\right], \end{aligned} \tag{7}$$

where $\psi(\cdot)$ is the output semantic maps from a pre-trained facial parsing network [36], the parameters of which are fixed in our training process. Our proposed $\mathcal{L}_{sm}$ is beneficial for preserving semantic structures in the transformation between domain $\mathcal{Z}$ and $\mathcal{Y}$.

Semantic-adaptive sharpen module. To eliminate blur and further enhance sharpness in reconstructed images, we propose a semantic-adaptive sharpen module.

In order to sharpen images, some previous works [38,39] introduce unsharp masking (USM) sharpening method. For a given image I, they first implement Gaussian blur on I, and then subtract the blurring result from I. As shown in Fig. 3, the result of USM is much sharper than the original image. The USM sharpening process can be described as

$$\begin{aligned}\hat{I} &= \frac{I - \omega * I_{blur}}{1 - \omega} \\ &= (1 + \lambda_s)I - \lambda_s I_{blur}, \end{aligned} \tag{8}$$

where I_{blur} is the image after Gaussian blur, ω is the coefficient, and $\lambda_s = \frac{\omega}{1-\omega}$.

Encouraged by the success of USM method, we propose a semantic-adaptive sharpen module (SASM) to sharpen reconstructed images. We utilize a convolutional layer with fixed kernel to implement Gaussian blur. Besides, in order to sharpen different facial components with different degrees, we sharpen the components in various regions, respectively. The sharping parameter of each region is learnable during the training process. In this way, the reconstructed images can be sharpened adaptively for different regions.

Specifically, as shown in Fig. 4, the semantic-adaptive sharpen module consists of two Gaussian blurring layers, $\mathcal{B}_1$ and $\mathcal{B}_2$, which can generate a blurry image using the output from the previous module. Given the image feature map $I_z^{'}$, the first convolutional layer generates its blurry result $\mathcal{B}_1(I_z^{'})$. The first-step sharpening result $I_z^{''}$ can be calculated by subtracting $\mathcal{B}_1(I_z^{'})$ from $I_z^{'}$. Then we divide $I_z^{''}$ into different facial regions through the element-wise product of $I_z^{''}$ and its parsing map $\psi(I_z^{''})$. Each region is fed into the second convolutional layer to get its sharpening result. Finally, we combine these sharpened regions with $I_z^{''}$ to obtain the final result of SASM.

The improved result $I_z^{'''}$ of semantic-adaptive sharpen module can be described as

$$\begin{aligned} I_z^{''} &= (1 + \lambda_s) \cdot I_z^{'} - \lambda_s \mathcal{B}_1(I_z^{'}), \\ I_z^{'''} &= \sum_{i=1}^{n} \psi_i(I_z^{''}) \cdot ((1 + \alpha_i) \cdot I_z^{''} - \alpha_i \mathcal{B}_2(I_z^{''})), \end{aligned} \tag{9}$$

where $\psi_i(\cdot)$ is the i-th region in parsing maps and α_i is a learnable parameter. The hyper-parameter λ_s is set as 0.4.

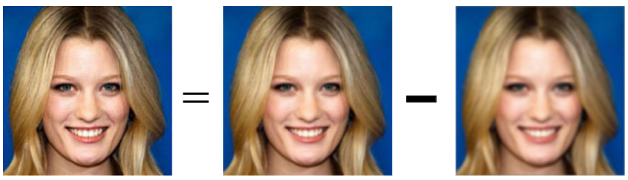

Fig. 3 Unsharp masking (USM) sharpening method. From left to right, we show the result of USM method, the original image, and the blurry image. The USM result can be calculated by subtracting the blurry image from the original image, as shown in Eq. 8

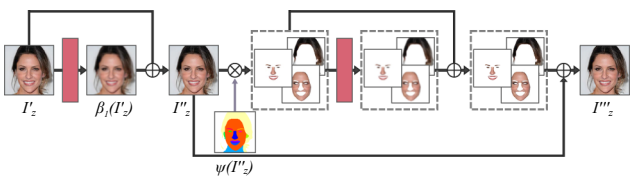

Fig. 4 Semantic-adaptive sharpen module (SASM). The SASM module consists of two convolutional layers to implement unsharp masking (USM) sharpening method. The facial parsing map $\psi(I_z^{''})$ instructs SASM to sharpen different regions with different degrees adaptively

The semantic-adaptive sharpen module sharpens different facial regions adaptively and hence improves visual quality remarkably.

Semantic-guided discriminator. There are characteristic texture details in different facial components. In order to recover diverse texture details, we propose a semantic-guided discriminator D_{sm}, which can discriminate on different facial components with different receptive fields under the guidance of facial parsing maps $\psi(\cdot)$.

The discriminator loss of D_{sm} (including D^1_{sm}, D^2_{sm},..., and D^n_{sm}) is formulated as:

$$\mathcal{L}_{D\_sm} = \mathbb{E}_{I_z, I_y}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\|D^i_{sm}(\psi_i(I_y)\cdot I_y) - 1\|_2 + \|D^i_{sm}(\psi_i(G_{ZY}(I_z))\cdot G_{ZY}(I_z))\|_2\Big)\right], \tag{10}$$

where n is the number of parsing regions, and $\psi_i(\cdot)$ is the i-th region in facial parsing maps. Then D_{sm} improves the adversarial learning by providing an extra adversarial loss for G_{ZY}:

$$\mathcal{L}^{sm}_{adv} = \mathbb{E}_{I_z}\left[\frac{1}{n}\sum_{i=1}^{n}\|D^i_{sm}(\psi_i(G_{ZY}(I_z))\cdot G_{ZY}(I_z)) - 1\|_2\right]. \tag{11}$$

By incorporating these losses, the full objective is defined as

$$\min_{\{G_{YZ}, G_{ZY}, G_{gra}\}}\max_{\{D_Y, D_Z, D_{sm}\}} \mathcal{L} = \mathcal{L}_{adv} + \lambda_1\mathcal{L}^{sm}_{adv} + \lambda_2\mathcal{L}_{cyc} + \lambda_3\mathcal{L}_{idt} + \lambda_4\mathcal{L}_{sm} + \lambda_5\mathcal{L}_{gra\_s} + \lambda_6\mathcal{L}_{gra\_p}, \tag{12}$$

where the hyper-parameters $\lambda_{(\cdot)}$ control the importance of each loss term.

4 Experiments

4.1 Datasets and implementation details

We build an unpaired dataset from CelebA-HQ dataset [40] for unsupervised face super-resolution. We first select 2000 images in different identities and bicubically downscale them to the size of 256×256 as HR images of the training dataset. Then we select other 2000 images and bicubically downscale them to the size of 64×64 as LR images of the training dataset. The LR images and HR images are in different identities. Then we randomly select 500 images from CelebA-HQ dataset and downscale them to the size of 256×256 and 64×64 as testing dataset. The constructed dataset and codes can be found in .

In our experiments, the super-resolution scale factor is set as $\times 4$. The hyper-parameters of loss terms are empirically set as: $\lambda_1 = 0.1$, $\lambda_2 = 10$, $\lambda_3 = 0.5$, $\lambda_4 = 0.4$, $\lambda_5 = 50$, and $\lambda_6 = 0.5$. All experiments are trained for 4×10^5 iterations on an Ubuntu18.04 server with a Intel Core i7-9700K CPU at 3.60GHz and a Nvidia RTX 2080Ti GPU. Our model is implemented using Pytorch. The optimizer is Adam [41] with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The initial learning rate is 2×10^{-4} and halved after 2×10^5 iterations. The training process of our method takes about 40 hours. The number of parameters of each module is shown in Table 2.

As for network structures, G_{gra} and G_{YZ} are of the same network structure. It consists of four pairs of up-sample and down-sample convolutional layers, as well as nine residual blocks. In addition to the above network layers, G_{ZY} consists of an RRDB block [7], three additional convolutional layers, and a proposed semantic-adaptive sharpen module. Discriminator D_{sm} consists of three basic discriminators that process on different facial regions. For discriminator D_Z, D_Y, and each basic discriminator in D_{sm}, we follow the PatchGAN discriminator structure of Pix2Pix [42].

4.2 Qualitative results

We compare our proposed method with several state-of-the-art unsupervised super-resolution methods: ZSSR [37], DSGAN [15], and CinCGAN [28]. The illustration of comparison with other methods is shown in Fig. 5. We can observe that our proposed GESGNet can reconstruct more realistic and high-fidelity face images and preserve finer details than the others. ZSSR reconstructs HR images with low-quality and coarse details. DSGAN produces distorted geometric structures, obvious artifacts, and unnatural colors. CinCGAN can recover images with satisfactory quality, but the lack of semantic restriction leads to blurry facial geometric structures. Compared to these methods, our method can reconstruct photo-realistic HR images approximating to the HR ground truth. The reconstructed images of our method are with accurate geometric structures and very clear boundaries among facial components. Besides, our reconstructed images are with diverse and fine-grained details, such as textured hairs, natural skins, and sharp edges.

We also compare our method with a state-of-the-art supervised super-resolution method: ESRGAN [7]. ESRGAN can reconstruct acceptable HR results, but the generated facial components, such as eyes and hairs, are not realistic enough. Besides, ESRGAN is trained in a supervised manner and shows weak performance if there is a large gap between LR domain and HR domain. In our method, we first use pre-trained ESRGAN to transform the LR domain $\mathcal{X}$ into an intermediate HR domain $\mathcal{Z}$, then we focus on improving unsupervised super-resolution performance. Compared to ESRGAN, our proposed method can not only tackle unsupervised super-resolution, but also reconstruct more photo-realistic and fine-grained HR images. The qualitative comparison indicates that our proposed method can recon-

struct HR images with better visual quality than the other methods.

4.3 Quantitative results

To quantitatively evaluate our method, we utilize several popular metrics: Peak Signal to Noise Ratio (PSNR), Structural similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [43]. Moreover, we apply an effective face alignment model FAN [44] to evaluate the performance of preserving identity information. We utilize FAN to extract 68 landmarks from reconstructed images and ground truth HR images, then calculate mean square error (MSE) among the coordinates. We compare our method with several unsupervised super-resolution methods, ZSSR, DSGAN, and CinCGAN, as well as a supervised method, ESRGAN.

As shown in Table 1, our proposed GESGNet is superior to ZSSR, DSGAN, and CinCGAN in all metrics. Our method shows good performance on PSNR metric, indicating that our method can reconstruct high-quality images with pixel-wise accuracy. The highest SSIM values demonstrate that our method can preserve the best geometric structures when super-resolving face images. Our method also achieves the best performance on LPIPS metric. This indicates that our method can super-resolve images with the best perceptual quality. The best alignment MSE shows that our method can preserve identity information and retain facial geometric structures much more accurately than other methods.

ESRGAN is a supervised super-resolution method. It achieves the best PSNR score, indicating its good pixel-wise performance. Compared to ESRGAN, our method obtains better SSIM, LPIPS, and MSE scores, which indicates our proposed method can reconstruct HR images with overall

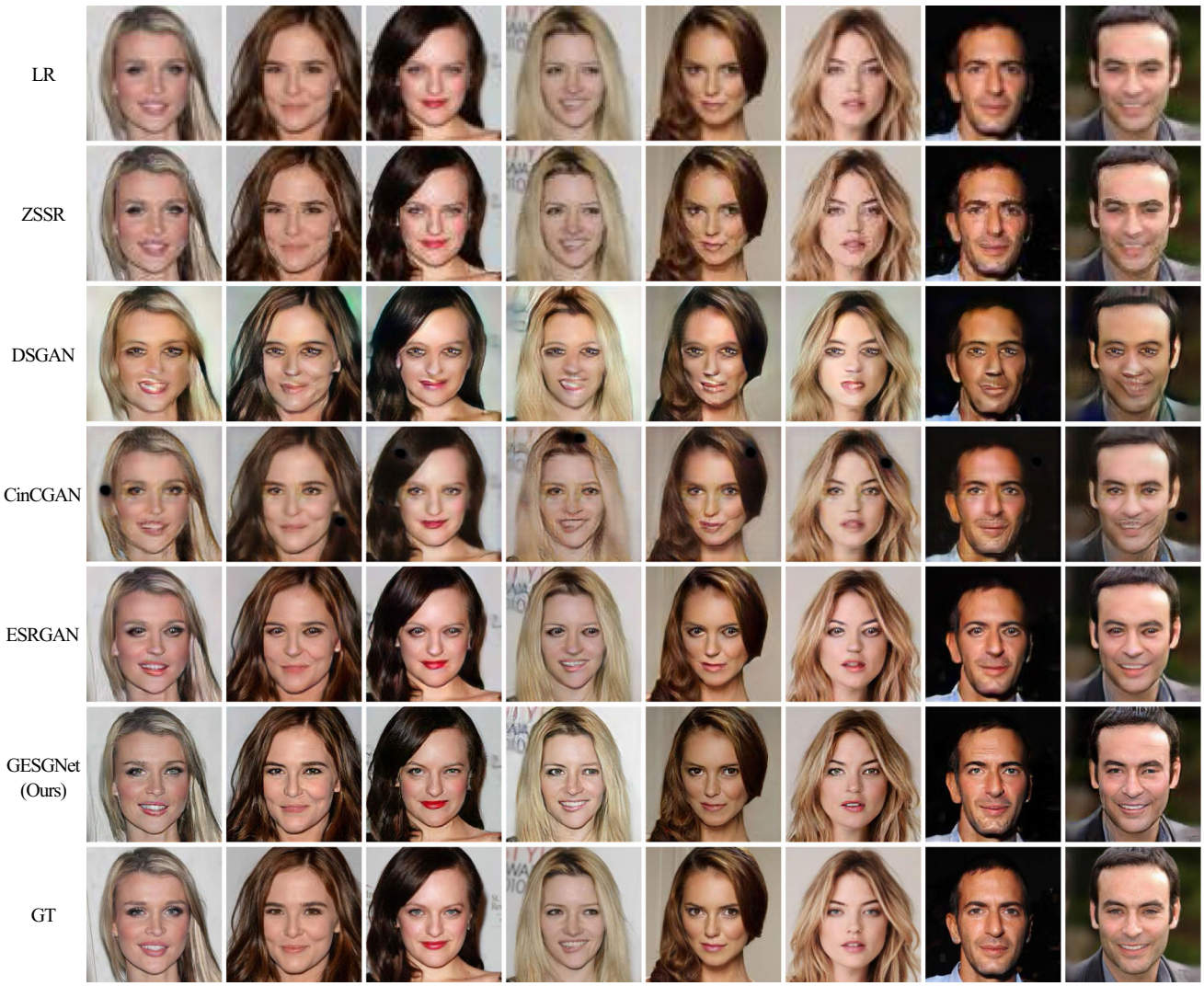

Fig. 5 Comparison of super-resolution results. From up to down, we show the input LR images, results of ZSSR [37], DSGAN [15], CinCGAN [28], ESRGAN [7], our proposed method, and the ground truth HR images

Table 1 Quantitative results of ZSSR [37], DSGAN [15], CinCGAN [28], ESRGAN [7], and our proposed method

Method	Metric				Time
	PSNR↑	SSIM↑	LPIPS↓	MSE↓	
ZSSR	29.785	0.661	0.374	24.080	0.031s
DSGAN	28.664	0.601	0.241	18.590	0.067s
CinCGAN	29.248	0.640	0.252	12.664	0.094s
ESRGAN	**30.120**	0.669	0.195	5.9561	0.048s
GESGNet (Ours)	29.831	**0.673**	**0.181**	**5.367**	0.126s

The best result of each metric is shown in bold

Table 2 The number of training parameters

Module	G_{ZY}	G_{YZ}	G_{gra}	D_Z	D_Y	D_{sm}
Para. ($\times 10^6$)	50.29	45.59	45.59	2.76	2.76	8.29

higher quality. The reconstructed HR images of our method are more photo-realistic and with more accurate geometric structures.

Moreover, we compare the computational efficiency of our proposed method with ZSSR [37], DSGAN [15], CinCGAN [28], and ESRGAN [7]. As shown in the last column of Table 1, our method consumes only a little longer run-time, but can reconstruct HR images with significantly higher quality than other methods.

4.4 Ablation study

In order to validate the effectiveness of the proposed method, we conduct several ablation studies. We take a CycleGAN model [35] with L_{adv}, L_{cyc}, and L_{idt} as baseline. The implementation details of ablation study can be found in Table 3.

The qualitative results of ablation study are shown in Fig. 6. It is obvious that the baseline model generates low-quality results with distortions and artifacts. By comparing (a) and (b), we can observe that L_{sm} contributes to accurate geometric structures and eliminate distortions on local details, such as eyes and mouths in (a). The results of (c) are much sharper and more high-fidelity than (b), which indicates that the gradient branch and two gradient losses contribute to sharp edges, and improve overall performance. The comparison between (c) and (d) shows that D_{sm} benefits diverse and fine-grained details, such as thin hairs and skin textures. The sharper facial components in (e) show that the semantic adaptive sharpness module can further sharpen and improve the reconstructed results.

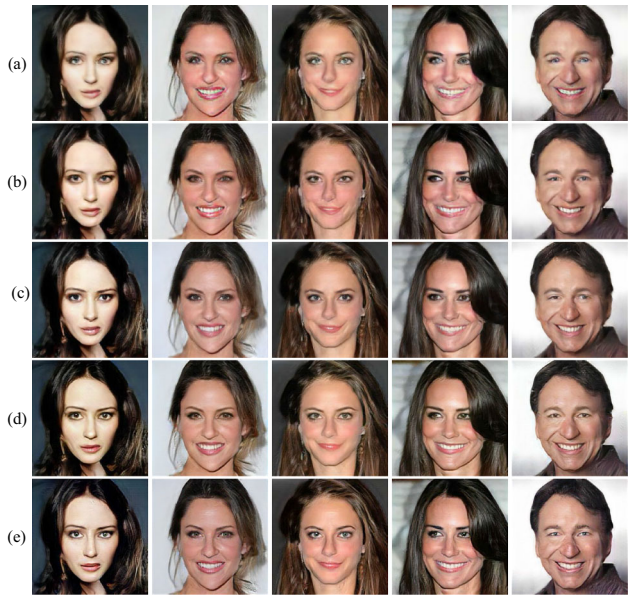

Fig. 6 SR results of our proposed method and its variants

The quantitative results of ablation study are shown in Table 3. We can observe that semantic loss improves the performance on SSIM and alignment MSE metrics significantly,

Table 3 Quantitative results of our proposed method and its variants

No.	Method		Metric			
	Module	Loss	PSNR↑	SSIM↑	LPIPS↓	MSE↓
(a)	Baseline		**30.186**	0.647	0.212	6.029
(b)	Baseline	L_{sm}	29.711	0.670	0.210	5.672
(c)	Baseline+G_{gra}	$L_{sm} + L_{gra\_s} + L_{gra\_p}$	29.476	**0.682**	0.194	5.413
(d)	Baseline+G_{gra} + D_{sm}	$L_{sm} + L_{gra\_s} + L_{gra\_p} + L_{adv}^{sm}$	29.650	0.675	0.186	5.468
(e)	Baseline+G_{gra} + D_{sm}+SASM	$L_{sm} + L_{gra\_s} + L_{gra\_p} + L_{adv}^{sm}$	29.831	0.673	**0.181**	**5.367**

The best result of each column is shown in bold

Fig. 7 SR results on real-world images. From up to down, we show the input LR images, results of ZSSR [37], DSGAN [15], CinCGAN [28], ESRGAN [7], and our proposed method

indicating its ability to maintain geometric structures. Note that the statistical gradient loss and pixel-wise gradient loss are applied simultaneously for gradient branch G_{gra}. Hence, we evaluate the effectiveness of G_{gra} and the two gradient losses together. By comparing (b) and (c), we can observe that G_{gra} and the two gradient losses improve perceptual quality and geometric structures. The better LPIPS score in (d) and the better MSE score in (e) show that the proposed semantic guidance mechanism is beneficial for preserving perceptual consistency and geometric structures.

4.5 Experiments on real-world images

We also implement experiments on real-world images from FDDB dataset [45]. Because there is no ground truth, we only show qualitative comparisons.

As shown in Fig. 7, Our method shows good performance on real-world images. The results of ZSSR are of low quality. DSGAN introduces artifacts in reconstructed results. CinCGAN reconstructs HR images with severe distortions; some facial components are even in the wrong position. The performance of ESRGAN on real images is not as well as that in Sect. 4.2, due to the large domain gap in real-world setting. In contrast, our proposed method can generate photo-realistic and visually reasonable results, with very few artifacts.

5 Application

5.1 Post-generation image enhancement

Recently, a large number of image generation methods [42, 46–48] have been proposed, which can generate high-quality images with fine-grained details. However, it requires expensive computational resources to generate high-resolution images directly through these complicated networks. To tackle this problem, several researchers [49,50] employ super-resolution methods as post-process enhancement tools to generate high-quality images with low resource consumption. Since our proposed method can reconstruct photo-realistic high-resolution images, it can be used as a post-process enhancement tool for image generation tasks. We conduct image-generation experiments and then super-resolve the generated images through our proposed method to validate its ability for post-process enhancement.

Experimental setting and results. We generate face images through StyleGAN2 [47] and super-resolve the generated images through our proposed GESGNet as a post-process enhancement. First, We train StyleGAN2 on CelebA-HQ dataset [40] to generate face images with the size of 256×256 and 64×64, respectively. All of experiments are conducted with 1.0×10^5 iterations. Then we super-resolve images with the size of 64×64 to the size of 256×256 through our trained GESGNet model, the training details of which can be found in Sect. 4.1.

Evaluation. Because there is no ground truth when generating images through StyleGAN2, we only show qualitative results. Figure 8a shows images generated by StyleGAN2 with size of 64×64. Figure 8b shows images in row (a) super-resolved by our proposed GESGNet. Figure 8c shows images generated by StyleGAN2 with size of 256×256. As shown in Fig. 8b, we can observe that the images enhanced by our method are almost photo-realistic as the high-resolution images generated directly by StyleGAN2. Note that generating images with the size of 256×256 by StyleGAN2 directly consumes about twice time as generating images with the size of 64×64. This demonstrates that our proposed GESGNet can be used as a post-generation image enhancement tool, which saves computational resources while enhancing image quality significantly.

5.2 Low-resolution face recognition

Face recognition has been a popular task in the field of computer vision for several decades. However, most state-of-the-art face recognition methods achieve good performance on datasets with high-resolution images. The recognition accuracy of these methods decreases dramatically in some practical applications, such as video surveillance, because the

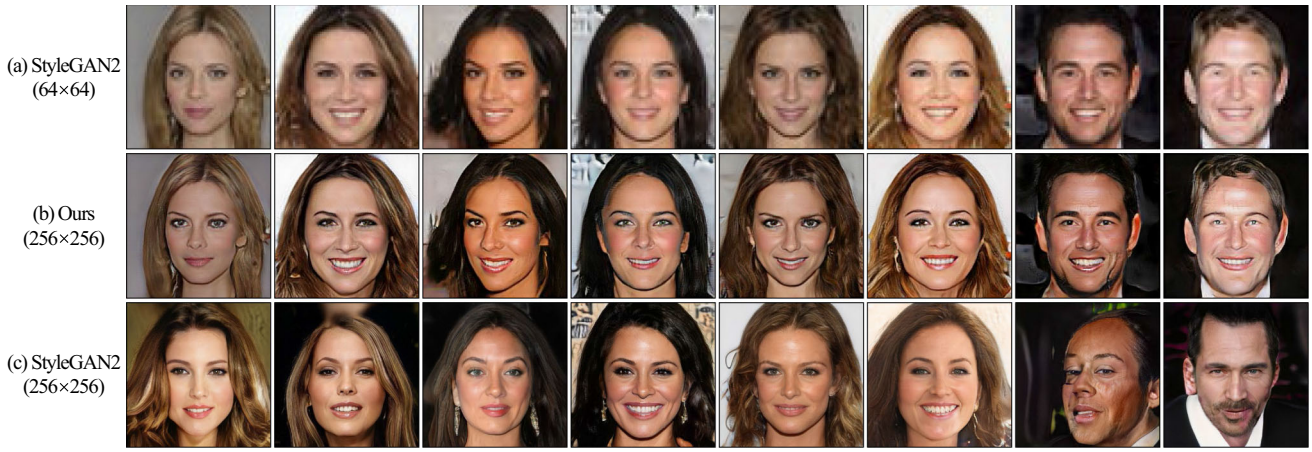

Fig. 8 Qualitative results of our proposed method as a post-process enhancement tool. From up to down, we show the images generated by StyleGAN2 with size of 64 × 64, the former images with a post-process through our method, and the images generated by StyleGAN2 with size of 256 × 256

input images are of low resolution. To address this issue, several efforts [51–53] have been made to super-resolve images before face recognition, which successfully improves the performance for low-resolution face recognition.

Our proposed GESGNet can reconstruct photo-realistic high-resolution face images from the low-resolution counterparts, and thus can boost face recognition performance on low-resolution face images. To demonstrate its performance, we conduct face recognition experiments on low-resolution face images, original high-resolution face images, and reconstructed high-resolution face images by our proposed GESGNet method and other SR methods.

Experimental setting and results. We perform super-resolution and face recognition experiments on LFW dataset, images of which are resized to 256 × 256 as original high-resolution images and 64×64 as low-resolution images. First, we select 6409 images of 3705 identities in LFW dataset [54] as training dataset for face super-resolution, while other 1403 images of 685 identities as testing dataset. Second, we train ZSSR, DSGAN, CinCGAN, ESRGAN, and our proposed GESGNet on the training dataset. All experiments are conducted with the same implementation details as that in Sect. 4.1. Then we evaluate the super-resolution performance of the above methods on the testing dataset. Afterward, we employ a pre-trained state-of-the-art face recognition model (SphereFaceNet [55]) to conduct face recognition on low-resolution images, original high-resolution images, and reconstructed high-resolution images by the above super-resolution methods, respectively. We compute the cosine distance of extracted features to evaluate face recognition accuracy.

Evaluation. Table 4 shows the performance comparisons of our proposed method and other super-resolution methods for low-resolution face recognition. We compare the face recognition accuracy on LR images, original HR images, and reconstructed HR images by ZSSR, DSGAN, CinCGAN, ESRGAN, and our proposed GESGNet. We can observe that face recognition accuracy on low-resolution (LR) images is much lower than that on original high-resolution (HR) images. Several super-resolution methods, including ZSSR, CinCGAN, ESRGAN, and our method GESGNet, can improve face recognition performance by reconstructing HR face images from LR inputs. However, the face recognition accuracy on HR images reconstructed by DSGAN is even lower than the accuracy on LR images, because the facial geometric structures are distorted in super-resolution process. We can observe that our proposed method achieves the highest face recognition accuracy. This demonstrates that our method can preserve geometric structures and identity information, and thus significantly improves low-resolution face recognition performance.

Table 4 Face recognition accuracy on LFW dataset [54]. From up to down, we show the face recognition accuracy on LR images, original HR images, as well as reconstructed HR images by ZSSR [37], DSGAN [15], CinCGAN [28], ESRGAN [7], and our proposed method

Method	Accuracy (%)
Original HR images	99.1
LR images	58.3
ZSSR	76.3
DSGAN	53.7
CinCGAN	77.3
ESRGAN	82.4
GESGNet (Ours)	**86.7**

The best result is shown in bold

6 Conclusion

In this paper, we have proposed an unsupervised face super-resolution network with gradient enhancement and semantic guidance. A gradient enhancement branch is proposed to generate sharp edges and preserve structures with the restriction of statistical gradient loss and pixel-wise gradient loss. Furthermore, a semantic guidance mechanism, including a semantic-adaptive sharpen module, a semantic-guided discriminator, and a semantic loss, is proposed to further preserve geometric structures and generate diverse details. Experiments show that our GESGNet can reconstruct photo-realistic high-resolution face images, significantly outperforming state-of-the-art methods.

Acknowledgements This work is supported by the National Key Research and Development Program of China (No. 2019YFC1521104), National Natural Science Foundation of China (No. 61972157), the Economy and Informatization Commission of Shanghai Municipality (No. XX-RGZN-01-19-6348), and Fundamental Research Funds for the Central Universities (No. 2021QN1072).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zhang, L., Zhang, H., Shen, H., Li, P.: A super-resolution reconstruction algorithm for surveillance images. Signal Process. **90**(3), 848–859 (2010)
2. Nie, Yongwei, Xiao, C., Sun, H., Li, P.: Compact video synopsis via global spatiotemporal optimization. IEEE Trans. Visual. Comput. Graphics **19**(10), 1664–1676 (2012)
3. Amaranageswarao, G., Deivalakshmi, S., Ko, S.-B.: Joint restoration convolutional neural network for low-quality image super resolution. Vis. Comput., pp. 1–20 (2020). https://doi.org/10.1007/s00371-020-01998-z
4. Zou, W.W.W.: Very low resolution face recognition problem. IEEE Trans. Image Process. **21**(1), 327–340 (2011)
5. Wang, Z., Miao, Z., Wu, Q.M.J., Wan, Y., Tang, Z.: Low-resolution face recognition: a review. Vis. Comput. **30**(4), 359–386 (2014)
6. Ledig, C., Theis, L., Huszar, F., Caballero, J., 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, pp. 4681–4690 (2017)
7. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change L. C., Esrgan: Enhanced super-resolution generative adversarial networks, In: Proceedings of the European Conference on Computer Vision, pp. 0–0 (2018)
8. Ma, C., Rao, Y., Cheng, Y., Chen, C., Lu, J., Zhou, J.: Structure-preserving super resolution with gradient guidance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7769–7778 (2020)
9. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Bing, X., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets, In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
10. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 694–711 (2016)
11. Yin, Y., Robinson, J., Zhang, Y., Fu, Y.: Joint super-resolution and alignment of tiny faces. Proc. AAAI Conf. Artif. Intell. **34**, 12693–12700 (2020)
12. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: end-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2492–2501 (2018)
13. Zhao, T., Zhang, C.: Saan: semantic attention adaptation network for face super-resolution. In: 2020 IEEE International Conference on Multimedia and Expo. IEEE, pp. 1–6 (2020)
14. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 908–917 (2018)
15. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop. IEEE, pp. 3599–3608 (2019)
16. Zhou, Y., Deng, W., Tong, T., Gao, Q.: Guided frequency separation network for real-world super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 428–429 (2020)
17. Wen, Y., Sheng, B., Li, P., Lin, W., Feng, D.D.: Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. IEEE Trans. Image Process. **28**(2), 994–1006 (2019)
18. Huang, Y., Shao, L., Frangi, A. F.: Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6070–6079 (2017)
19. Keys, Robert: Cubic convolution interpolation for digital image processing. IEEE Trans. Acoust. Speech Signal Process. **29**(6), 1153–1160 (1981)
20. Fattal, R.: Image upsampling via imposed edge statistics. In: ACM SIGGRAPH 2007 papers, pp. 95-es. (2007)
21. Freedman, Gilad, Fattal, R.: Image and video upscaling from local self-examples. ACM Trans. Graph. (TOG) **30**(2), 1–11 (2011)
22. Xiong, Z., Sun, X., Feng, W.: Robust web image/video super-resolution. IEEE Trans. Image Process. **19**(8), 2017–2028 (2010)
23. Zhang, H., Yang, J., Zhang, Y., Huang, T. S.: Non-local kernel regression for image and video restoration. In: European Conference on Computer Vision. Springer, pp. 566–579 (2010)
24. Freeman, William T., Jones, Thouis R., Pasztor, Egon C.: Example-based super-resolution. IEEE Comput. Graph. Appl. **22**(2), 56–65 (2002)
25. Chang, H., Yeung, D-Y., Xiong, Y.: Super-resolution through neighbor embedding. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. IEEE, vol. 1, pp. I–I (2004)
26. Dong, C., Loy, C. C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision. Springer, pp. 184–199 (2014)
27. Sajjadi, M.S.M, Scholkopf, B., Hirsch, M.: Enhancenet: single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4491–4500 (2017)
28. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 701–710 (2018)

29. Zhang, Y., Liu, S., Dong, C., Zhang, X., Yuan, Y.: Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. IEEE Trans. Image Process. **29**, 1101–1112 (2019)
30. Choudhury, A., Segall, A.: Channeling mr. potato head-face super-resolution using semantic components. In: Southwest Symposium on Image Analysis and Interpretation. IEEE **2014**, 157–160 (2014)
31. Yu, X., Fernando, B., Ghanem, Bernard, P., Fatih, H., Richard: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 217–233 (2018)
32. Bulat, A., Tzimiropoulos, G.: Super-fan: integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 109–117 (2018)
33. Xin, J., Wang, N., Gao, X., Li, J.: Residual attribute attention network for face image super-resolution. Proc. AAAI Conf. Artif. Intell. **33**, 9054–9061 (2019)
34. Wang, C., Zhong, Z., Jiang, J., Zhai, D., Liu, X.: Parsing map guided multi-scale attention network for face hallucination. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2518–2522 (2020)
35. Zhu, J.-Y., Park, T., Isola, P., Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
36. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 325–341 (2018)
37. Shocher, A., Cohen, N., Irani, M.: Zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3118–3126 (2018)
38. Cao, Gang, Zhao, Yao, Ni, Rongrong, Kot, Alex C.: Unsharp masking sharpening detection via overshoot artifacts analysis. IEEE Signal Process. Lett. **18**(10), 603–606 (2011)
39. Peng, K.-S., Lin, F-C., Huang, Y-P., Shieh, H.-P.D.: Efficient super resolution using edge directed unsharp masking sharpening method. In: IEEE International Symposium on Multimedia. IEEE **2013**, 508–509 (2013)
40. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196 (2017)
41. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
42. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
43. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3118–3126 (2018)
44. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1021–1030 (2017)
45. Jain, V., Learned-Miller, E.: Fddb: a benchmark for face detection in unconstrained settings. Tech. Rep, UMass Amherst technical report (2010)
46. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
47. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
48. Li, L., Tang, J., Shao, Z., Tan, X., Ma, L.: Sketch-to-photo face generation based on semantic consistency preserving and similar connected component refinement. Vis. Comput., pp. 1–18, (2021). https://doi.org/10.1007/s00371-021-02188-1
49. Anokhin, I., Solovev, P., Korzhenkov, D., Kharlamov, A., Khakhulin, T., Silvestrov, A., Sergey, N., Victor, L., Gleb, S.: High-resolution daytime translation without domain labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7488–7497 (2020)
50. Damer, N., Boutros, F., Saladie, A. M., Kirchbuchner, F., Kuijper, A.: Realistic dreams: cascaded enhancement of gan-generated images with an example in face morphing attacks. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–10 (2019)
51. Biswas, Soma, Aggarwal, Gaurav, Flynn, Patrick J., Bowyer, Kevin W.: Pose-robust recognition of low-resolution face images. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 3037–3049 (2013)
52. Chen, J., Chen, J., Wang, Z., Liang, C., Lin, C.-W.: Identity-aware face super-resolution for low-resolution face recognition. IEEE Signal Process. Lett. **27**, 645–649 (2020)
53. Hennings Y., Pablo H,. Baker, S., Vijaya, K.: BVK: simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8 (2008)
54. Huang, G.B, Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
55. Liu, W., Wen, Y., Yu, Z., Li, Ming, R., Bhiksha, S., Le: Sphereface: deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212–220 (2017)

Luying Li received the B.Eng. degree in Computer Science and Technology from Ocean University of China in 2018. She is now a Ph.D. candidate in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. Her research interests include computer vision and image processing.

Junshu Tang received the B.Eng. degree in Computer Science and Technology from the Xidian University, China in 2019. She is now a Ph.D. candidate in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. Her research interests are image synthesis and facial attribute edit.

Zhou Ye received his master's degree of Software Engineering from Nanjing University. He is now the CTO of Shanghai CLS Fintech CO., Ltd. His research interests include computer vision and deep learning.

Bin Sheng is a professor of Shanghai Jiao Tong University. He received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong. His current research interests include machine learning, virtual reality, and computer graphics.

Lijuan Mao received her PhD degree from East China Normal University 2002. She is currently a professor at Shanghai University of Sport. Her research interests include sports kinesiology, physical education and sports training.

Lizhuang Ma is a distinguished professor of Shanghai Jiao Tong University. He is the recipient of national outstanding Youth Foundation. He received the Ph.D. degree in Zhejiang University in 1991. His research interests include computer vision, computer graphics and computer-aided design.