Sun *et al.*
EURASIP Journal on Image and Video Processing (2022) 2022:7
https://doi.org/10.1186/s13640-022-00585-7

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Face image synthesis from facial parts

Qiushi Sun, Jingtao Guo and Yi Liu*

*Correspondence:
yiliu_bjtu@163.com
Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Abstract

Recently, inspired by the growing power of deep convolutional neural networks (CNNs) and generative adversarial networks (GANs), facial image editing has received increasing attention and has produced a series of wide-ranging applications. In this paper, we propose a new and effective approach to a challenging task: synthesizing face images based on key facial parts. The proposed approach is a novel deep generative network that can automatically align facial parts with the precise positions in a face image and then output an entire facial image conditioned on the well-aligned parts. Specifically, three loss functions are introduced in this approach, which are the key to making the synthesized realistic facial image: a reconstruction loss to generate image content in an unknown region, a perceptual loss to enhance the network's ability to model high-level semantic structures and an adversarial loss to ensure that the synthesized images are visually realistic. In this approach, the three components cooperate well to form an effective framework for parts-based high-quality facial image synthesis. Finally, extensive experiments demonstrate the superior performance of this method to existing solutions.

Keywords: Deep learning, Generative adversarial network, Image completion

1 Introduction

The rapid progress of deep convolutional neural networks (CNNs) and generative adversarial networks (GANs) [1] has led to a surge of new applications in computer vision. Among these, facial image processing has been a popular area, including subfields such as face image generation [2–4], inpainting [5–7], manipulation of expression, age and other facial attributes [8–11]. In this paper, we address an interesting yet challenging task: generating the whole face image based only on image patches of several facial parts, as shown in Fig. 1. Many works by predecessors mainly complete the missing part of the face, and the missing part is often a small part of the whole picture; however, here is the opposite situation: we need to complete the whole portrait according to the very small facial part clue. Because human facial parts are the ultimate basic feature elements in portraits, the face image synthesis methods in this paper are used in a wide range of applications, such as surgical plastic effect preview, portrait drawing, and virtual portrait synthesis. Specifically, in actual virtual portrait synthesis applications, users may prefer to use the key facial organs from different persons to generate a realistic photo of the "desired" virtual portrait. Additionally, the generated fake face retains human identity information at some level, so it can be

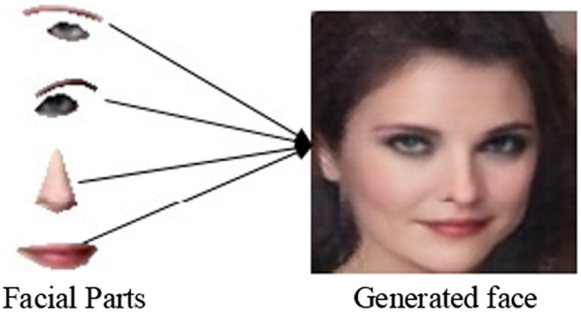

Fig. 1 A sketch of parts-based facial image synthesis

used to provide fake data for training and evaluating applications such as face recognition [12] and face tracking [13–15].

To the best of our knowledge, this work represents the few attempts to synthesize the whole face image according to the limited facial parts provided by a user. Existing work that addresses a similar but simpler problem includes methods for image inpainting [5–7] and for domain transformation.

Methods for domain transformation [16–18] synthesize the target domain from the source domain through conditional GANs. These approaches work well when there is a strong correlation between the two domains. However, when the source domain contains large missing areas, as in this case (containing only parts), these methods fail to discover relationships between the two domains well. Thus, they are unable to generate visually plausible contents for the missing regions. This is mainly because the large missing areas destroy the potential correlations between the two domains, which in turn hurts the generative performance of the model. Another relevant research field is image inpainting [5–7], which aims to synthesize visually realistic and semantically plausible pixels for missing regions that are coherent with the other parts of an image. To date, a large number of image inpainting methods have emerged due to the rapid progress of CNNs and GANs, which formulate inpainting as a conditional image generation task. These methods work well when the pixels around the missing area are known. However, when most of the pixels in the image are missing, there is less neighboring information for unknown areas in the image, and image inpainting methods fail to work well. For instance, in terms of generating the face image based on limited facial parts, these methods often create distorted structures and/or blurry textures.

To address the limitations of previous works, this paper presents a novel convolutional encoder-decoder generative network to implement face synthesis conditioned on key facial parts. It is able to synthesize high-quality facial images even conditioned on several key facial parts only. The deep network of this approach following the typical GAN structure contains a generator and a discriminator, as shown in Fig. 2. The generator network is designed to automatically align the facial parts to the precise position in a face image to generate a complete result, while the discriminator network pushes the generated results to be visually realistic. Both networks contain convolutional, BatchNorm [19], and ReLu layers. In addition, to mitigate the

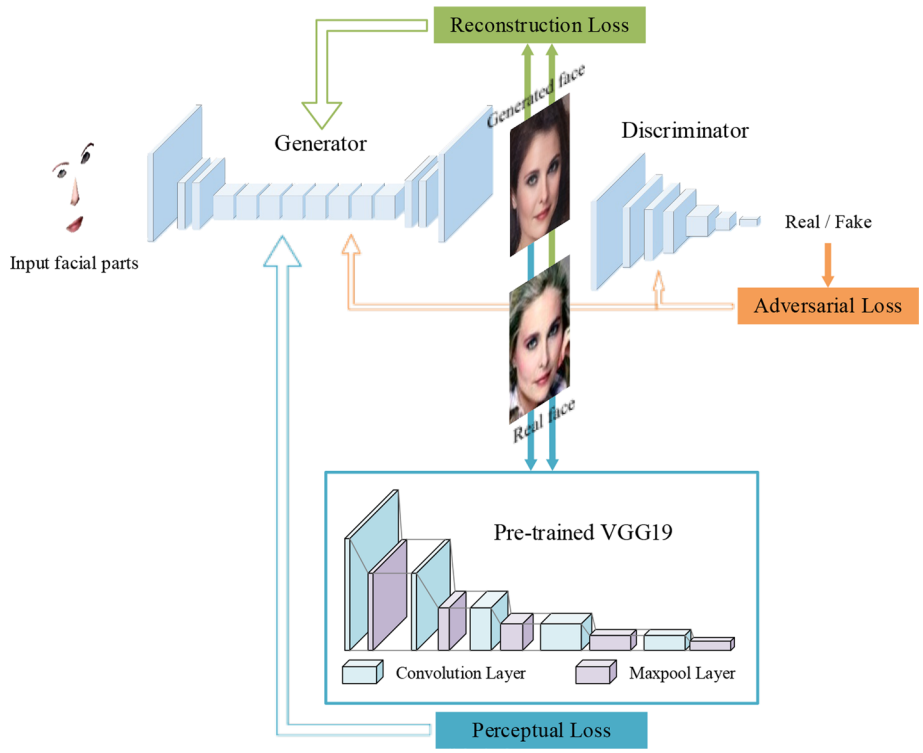

Fig. 2 Network architecture. In the architecture, the extracted human facial parts are input into the generator to generate a fake face. The discriminator is used to discriminate between real and generated faces. Reconstruction loss and perceptual loss are obtained by comparing generated faces and real faces. The reconstruction loss part is shown in green and is used to train the generator. The perceptual loss part is represented in dark blue, and its calculation uses a pretrained VGG19 network. Finally, the orange part represents the adversarial loss used to train the generator and discriminator

loss of texture information, the generator network only decreases the image resolution twice with stride convolutions. For the training process, we propose integrating the reconstruction loss, the perceptual loss [20, 21] and the adversarial loss into a unified framework to achieve the best result. Specifically, the reconstruction loss is used to generate contents in the unknown region, and the perceptual loss models the high-level semantic structure, eliminating structure distortion and texture inconsistency of the synthesized contents. Furthermore, the perceptual loss can speed up the training process of the model, with fewer training steps and better results. Finally, the adversarial loss is employed to enhance visual authenticity and ensure that the model's adversarial gaming process is ongoing.

The method in this paper performs well in face synthesis and repair and can even modify and replace facial organs. In summary, the contributions of this work are as follows:

- Face images are synthesized based on key facial parts. It brings the possibility of fusing multiple facial organs from different persons to generate realistic virtual portraits, which has great application prospects in medical facial plastic surgery, portrait drawing of suspects or virtual anchor synthesis and implementation.
- Three loss functions are introduced in our approach, which are the key to making the synthesized realistic portraits: a reconstruction loss to generate image content

in an unknown region, a perceptual loss to enhance the network's ability to model high-level semantic structures and an adversarial loss to ensure that the synthesized images are visually realistic.

- Comprehensive experiments are performed on the CelebA dataset [22], and both qualitative and quantitative results show the promising performance of this model. Moreover, further validation is performed on the Cross-Age Celebrity Dataset (CACD) [37] and Labeled Faces in the Wild Home (LFW) [38]. Our method outperforms the average performance of the state-of-the-art methods.

This paper is structured as follows. In Sect. 2, several key areas related to this research are reviewed. In Sect. 3, the novel convolutional encoder–decoder generative network to synthesize facial images based only on limited facial parts is presented. Extensive experimental results are presented in Sect. 4. Finally, the conclusion of this paper is presented in Sect. 5.

2 Related works

In this section, we review related works from three closely related areas, namely, generative adversarial networks, image translation and image inpainting.

2.1 Generative adversarial network

Generative adversarial networks (GANs) [1], as a special deep generative model, aim to model a mapping from a random vector to an image by adversarial training. A typical GAN consists of a discriminator and a generator. The generator is trained to generate fake samples from the random noise vectors. The discriminator is trained to distinguish between real samples and fake samples. This framework can be represented as a two-player min–max game with value function:

$$\min_G \max_D E_{x\sim p_{\text{data}}(x)}[\log(D(x))] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))], \tag{1}$$

where x is sampled from the real data's distribution $p_{\text{data}}(x)$, and $p_z(z)$ represents the distribution of the noise input z.

Recently, many variants of GAN have been proposed to greatly improve their performance and broaden the application scopes. Radford et al. [2] proposed deep convolutional generative adversarial networks (DCGANs), which replace fully connected layers in the original GANs with the convolutional layers in both the generator and the discriminator network. DCGANs optimize the network structure of the generator and discriminator, which can make the generator learn good representations of images and improve the stability in the adversarial training process at the same time. Another important variant is the conditional version of generative adversarial nets (CGAN) [23], which adds class information to the discriminator and generator to model conditional probability distributions. The idea of conditional image generation has also been successfully applied to face image generation [8–11], image translation [16–18], and image inpainting [5–7]. Inspired by these approaches, we propose a new GAN-based framework that is able to generate face images conditioned on a small patch of facial parts. This framework combines three loss functions, the reconstruction loss, the perceptual

loss [20, 21] and the adversarial loss, which can constrain the model to generate elegant and accurate portraits.

2.2 Image inpainting

Image inpainting aims to synthesize plausible contents for the missing regions in the image such that the completed image appears to be visually realistic. Recently, many image inpainting methods based on deep generative models have been proposed [5–7]. These methods formulate image inpainting as a conditional image generation problem, which synthesizes the contents of the missing regions in a convolutional end-end fashion. For example, context encoders [5] first introduce generative adversarial loss to train deep neural networks for the image inpainting task, where the completion network is trained by minimizing the pixelwise reconstruction loss and the adversarial loss, which can produce much sharper results and avoid blurred texture. Iizuka et al. [6] improve this work by optimizing the completion network structure to introduce a global adversarial loss, which further improves the coherency between generated and existing pixels. Nevertheless, this approach still needs to employ a Poisson blending postprocessing step to improve the visual effect of the completed image. Yu et al. [7] proposed a novel contextual attention module to capture the long-range spatial dependencies, which can eliminate the effect of invalid pixels in missing regions by borrowing or copying feature information from known regions to complete missing pixels. The methods mentioned above are designed for the scenario in which the pixels around the missing area are known. In this case, the surrounding pixels are critical to successfully generate plausible structures and textures for the missing regions. However, when the missing region in the image is large or even dominates, as in our case of generating a face image based on a few facial parts, these methods will not work well and tend to create distorted structures or blurry textures in the missing region.

2.3 Image-to-image translation

Image translation, as a common image processing task, aims to translate an input image from a source domain to a target domain. Recently, various methods [16–18, 24, 25] have been proposed to address this task due to the rapid progress of deep convolutional networks and generative adversarial nets. Instead of directly optimizing the L1 loss, which often leads to blurry images, these approaches leveraged the adversarial loss to encourage sharper results. For example, the "pix2pix" work of Isola et al. [16] first employs conditional adversarial networks to translate images from the source domain to the target domain using input–output image pairs as training data. It effectively transforms Google maps to satellite views and generates object images from sketch maps. In contrast to using paired data, unpaired image-to-image translation frameworks [24, 25] have also been proposed. CycleGAN [25] and DiscoGAN [24] show promising results on unsupervised image translation by utilizing cycle consistency. However, when the source domain and the target domain are only relevant in some local areas, such as face generation based on a few image patches of facial parts, the source and target domains have strong correlations in the facial parts region, but there is little correlation in the other regions due to the loss of large areas in the source domain. These methods easily learn the relationships in the known facial part regions of the source domain. However,

it is difficult for them to learn the relationships outside these regions, which is prone to cause instability in adversarial training, thereby creating distorted structures or inconsistent blurry textures in these areas.

2.4 Face completion

In Li's article [26], the use of two independent discriminators is proposed: a local discriminator for calculating the loss of the missing part of the face and a global discriminator for calculating the adversarial loss of the entire image. Then, the pixelwise softmax loss was used to train the generative network. As discussed by the authors, such a network has a disadvantage: it does not perform well in inpainting faces that are not aligned. The reason is that the pixelwise losses do not capture perceptual differences between output and ground-truth faces. For example, moving a face a few pixels in parallel as a new image will still be the same person compared to the original image, but their pixelwise loss may be quite large. FCENet [27] continued to use the local/global discriminator structure and introduced a facial geometry estimator to infer facial part maps and landmark heatmaps. The RGAN [28] introduced a recurrent neural network to the GAN model, which can extract multiscale features and transfer them for face completion at different feature levels. There are many methods that are not mentioned here. By studying the work of these methods, it can be found that a major challenge in face completion is that the missing parts will be blurred in the generated face. In the study of Jian et al. [29–32], the SVD method was used to enhance the face image to complete the conversion of face images from low-resolution (LR) inputs to high-resolution (HR) outputs. The recent method of Wang et al. [33] combines a variety of losses and proposes a new method of face restoration, which, in addition to dividing the face part, also introduces the concept of identity preserving. The above methods perform well in the application of face completion, face hallucination and face restoration but do not take into account an extreme case: almost the entire portrait is missing, and only part of the face organs are input.

3 Method

In this section, the proposed method for face generation conditioned on a few patches of key facial parts is described. The key facial part extraction, network architecture and loss function methods are described in detail below.

3.1 Training data preparation

This method is required to precisely extract the facial parts to achieve facial image generation given a small patch of facial parts. To achieve this goal, first, the 68 facial key points are detected using dlib [34]. The facial parts mask can be obtained by connecting all points pairs. Then, the facial parts mask is used to extract the facial parts separately, as shown in Fig. 3A. When a user wants to synthesize a whole face giving these facial parts, the model will position these facial parts to an "average face", where the facial parts are positioned in a rough position, which is used as the input of our model, as shown in Fig. 3B.

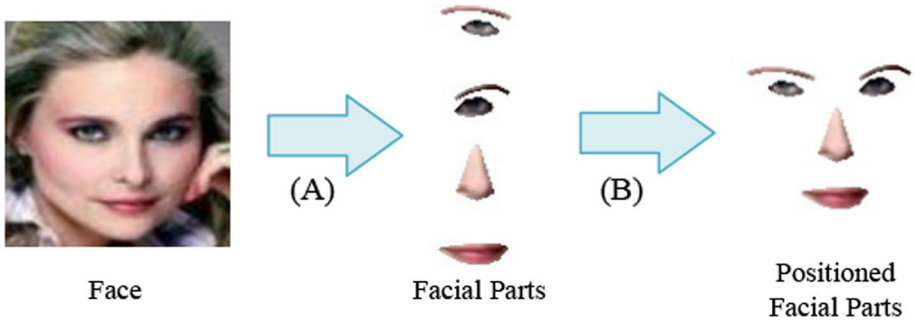

Fig. 3 Example of facial part extraction, left: input images, middle: the extracted facial parts, right: positioning facial parts to an "average face"

Table 1 Specification of the generator network

Num	Type	Kernel	Stride	Outputs
1	Input	–	–	3
1	Conv	5 × 5	1	64
1	Conv	3 × 3	2	128
1	Conv	3 × 3	1	128
1	Conv	3 × 3	2	256
6	Conv	3 × 3	1	256
1	Decon	3 × 3	2	256
1	Decon	3 × 3	1	128
1	Decon	3 × 3	2	128
1	Decon	3 × 3	1	64
1	Decon	3 × 3	1	3

Each "conv." denotes a convolutional-BatchNorm-LeakyReLu module. Each "deconv." is followed by a BatchNorm layer. The last layer uses the tanh activation function

3.2 Model architecture

Given an "average face" image, the goal is to generate the whole face that is coherent with existing facial parts, which can be regarded as a conditional image generation problem. Many previous works [5–7, 16–18] used the convolutional encoder–decoder network, jointly trained with adversarial networks to handle this task. The encoder contains a series of downsampling convolutional layers that encode the input image into a latent feature representation, and the decoder consists of several upsampling convolutional layers that decode the latent feature representation back to the original size. The more network layers there are, the stronger the learning ability, and the more information is lost through the process of downsampling and upsampling. To achieve a balance between the two, the generator network (encoder–decoder network) only employs two downsampling convolutional layers, as shown in Fig. 2, which can avoid reducing too much information. We also employ a series of convolutional blocks to enhance the generative ability of the model. For the discriminator, the input of the network is the generated face image and the real ones sampled from the training datasets. As shown in Fig. 2, the discriminator consists of five downsampling convolutional layers and a fully connected layer, and then the output features of the discriminator are processed by a sigmoid function. Unlike the generator network,

Table 2 Specification of the discriminator network

Num	Type	Kernel	Stride	Outputs
1	Input	–	–	3
1	Conv	3 × 3	2	32
1	Conv	3 × 3	2	64
1	Conv	3 × 3	2	128
1	Conv	3 × 3	2	256
1	Conv	3 × 3	2	512
1	FC	–	–	1024
1	FC	–	–	2

Except for the last type, each "conv." denotes a convolutional-LeakyReLu module. The term "FC" denotes the fully connected layer. The last layer uses the sigmoid activation function

the BatchNorm layer is not used after the convolution operation. Tables 1 and 2 show the detailed network parameters of the generator and discriminator.

3.3 Loss functions

To train the network to generate high-quality face images conditioned on key facial parts, three loss functions are jointly used: a per-pixel reconstruction loss to ensure training stability, a perceptual loss to model the high-level semantic structure for the large unknown regions and an adversarial loss of the generative adversarial network (GAN) [1] to improve the authenticity of the results.

As shown in Fig. 2, the reconstruction loss and the perceptual loss are obtained by comparing the generated fake faces and real faces, and they were used to train the encoder–decoder pairs. The adversarial loss was used to train the generator and the discriminator.

Let x be the ground-truth image; the corresponding "average face" is denoted by z. Generation G takes z as the input and generates a whole face image $\tilde{x} = G(z)$. We first define a per-pixel reconstruction loss L_{r} between the output $\tilde{x}$ and the ground-truth x, where $\|\cdot\|_2$ represents the Euclidean norm. The reconstruction loss function for the generator is formulated as follows:

$$L_{\mathrm{r}} = \left\|x - \tilde{x}\right\|_2. \tag{2}$$

Because the input image contains large missing regions, the per-pixel loss pays more attention to the low-level pixel-value differences of the reconstruction. To better reconstruct the high-level semantic structure for the large unknown regions, we employ a perceptual loss, which was first introduced by Gatys et al. [21]. This is an essential loss function for the training process that works well in our approach. Specifically, it computes the L_1 distances between x and $\tilde{x}$, but after projecting these images into a series of high-level feature spaces using a pretrained network [35], it better captures the high-level semantic structures. In terms of mathematical formulation, the perceptual loss L_{perc} based on L_1 distances is defined as formula (3):

$$L_{\mathrm{perc}} = \sum_{i=1}^{N} \left\|\Phi_i\left(x^a\right) - \Phi_i\left(x^{\hat{a}}\right)\right\|_1. \tag{3}$$

Here, Φ_i is the ith layer of a pretrained network, and N is the total number of layers. Here, we use the three layers conv1_1, conv2_1 and conv3_1 of the VGG-19 network [35] pretrained on the ImageNet dataset [36]. It is worth noting that we can use the L_2 normal form (squared Euclidean distance) or squared Frobenius norm instead of L_1 distances. Inspired by [20] and [21], a style loss can also be added to preserve the picture style, and a total variation regularization to encourage pattern smoothness in the generated faces.

However, previous work suggests that the outputs often become blurry when the reconstruction loss is used. To overcome this problem, we combine the adversarial loss with the reconstruction loss to enhance the authenticity of the output images. Here, the adversarial loss serves as a binary classifier to distinguish whether an image is real or fake, and the generator network jointly trained with adversarial loss encourages the output images to be more realistic. Formally, the adversarial loss is defined as formula (4):

$$L_{\text{adv}} = \min_G \max_D E_{x\sim p_{\text{data}}(x)}\left[\log\left(D(x)\right) + \log\left(1 - D\left(\tilde{x}\right)\right)\right]. \tag{4}$$

Collectively, the loss functions used to train the discriminator and the generator networks are formula (5) and formula (6):

$$L_D = \log\left(D(x)\right) + \log\left(1 - D\left(\tilde{x}\right)\right), \tag{5}$$

$$L_G = \lambda_r L_r + \lambda_{\text{perc}} L_{\text{perc}} + \lambda_{\text{adv}} \log\left(D\left(\tilde{x}\right)\right), \tag{6}$$

where λ_r, λ_{perc} and λ_{adv} are the weights to balance the strength of the perception loss and the adversarial loss with the reconstruction loss. In our experiments, we set up different λ_r, λ_{perc}, and λ_{adv} for ablation experiments between losses.

4 Results and discussion

In this section, we present the experimental results and evaluate the performance of our proposed method on the test set. First, we introduce the benchmark dataset used in the experiment. Second, we describe in detail the strategy of network training and the related parameter configurations. Third, to explore the practicality and robustness of our proposed model, we provide face synthesis results based on facial-part patches from a single person or from multiple persons. Finally, we document qualitative and quantitative comparison results with other image inpainting and translation algorithms to demonstrate the superior performance of the proposed method.

4.1 Benchmark dataset

We conduct our experiments on the CelebA dataset [22], which has been widely used in a variety of computer vision tasks, such as face detection, facial attribute editing and facial part localization. The CelebA dataset contains approximately 202 K facial images covering rich facial pose variations (2,025,099 images in total). In the experiment, we follow the standard split operation with 182 K images for training and 20 K for testing. As mentioned in Sect. 3, we extract facial parts using dlib [34] to generate the training and testing data. To extensively test the robustness of our method, we also introduce two datasets, CACD [37] and LFW [38], for further validation.

4.2 Training details

The proposed methods are implemented using the TensorFlow deep learning framework [39] and executed on a computer with a single NVIDIA 1080Ti GPU (12 GB). For the network training, we scale images down to the 128×128 resolution, and train the network using a batch size of 32 images. To make the training process stable and efficient, three-phase training procedures are adopted. First, for training step-1, the generator is trained for 6000 iterations using both the per-pixel reconstruction loss and the perceptual loss to obtain blurry results. Afterwards, for training step-2, the generator network is fixed, and the discriminator network is trained for 1500 iterations with the adversarial loss to learn to distinguish between real and fake samples. Finally, for training step-3, both the generation and the discriminator network are trained jointly for 50,000 ~ 100,000 iterations until the end of training. The entire training procedure takes approximately 1 day with a single NVIDIA 1080Ti GPU (12 GB), but the test procedure can be performed in real time. The detailed training procedure is shown in Algorithm 1. If the number of steps is changed in the three-step training, different results will be generated. Overall, if the number of steps in training step-1 is reduced, the model will converge more slowly, and the final result will tend to be blurry.

Algorithm 1 Three-phase training procedure.

```
while iterations t < T_train do
    Sample a minibatch of original images x and
    Generate corresponding "average face" z for each
image
    x in the minibatch
    If t < T_G then
        Fix D and update G according to Eq. (6)
    else if t > T_G and t < T_D
        Fix G and update D according to Eq. (5)
    else
        Fix G and update D according to Eq. (5)
        Fix D and update G according to Eq. (6)
    end if
    t++
end while
```

Output: G outputs the generated face images

4.3 Qualitative results

First, we use the proposed method to generate whole facial images from a few facial-part patches. Exemplar results are shown in Fig. 4. It is clear that the proposed method can not only automatically align facial parts to the precise position in a face image but also successfully synthesize visually realistic whole facial images even when

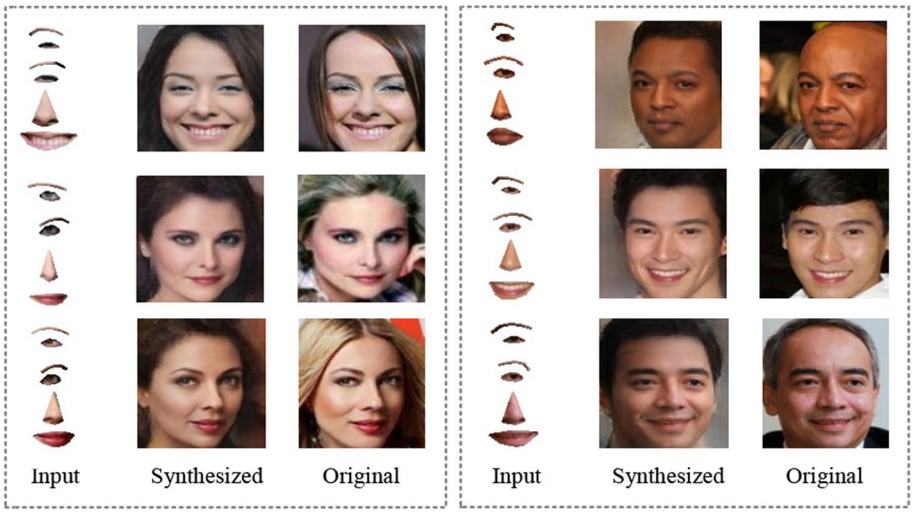

Fig. 4 Face images generated on the test dataset. In each panel from left to right are facial-part patches, our synthesized facial images and the original facial images (the ground truth)

most of the pixels in facial images are absent. The results firmly demonstrate the powerful generative capability of this approach.

Interestingly, compared to the original portrait photos, the colors and illuminance in the fake portraits generated by our method are more uniform, and there are fewer noise points. In this regard, this method can effectively remove the highlight noise and blur caused by the lighting factors in the original image, making the photo portrait more recognizable (Fig. 4).

In practical applications, users may prefer to generate realistic faces based on the key facial parts from more than one person (for example, virtual portrait synthesis). To test whether our approach could address this need, we present multiple facial parts from different persons to the algorithm and check if it could output a realistic and consistent facial image. The synthesized example images are shown in Fig. 5. The results again show that the proposed method can synthesize visually realistic images conditioned on facial parts from multiple persons. This is not a simple task, since the facial parts must be fine-tuned so that they look consistent and reasonable in one image. However, this algorithm can achieve this goal and synthesize sharp faces.

To further study the ability of our method to individually edit a certain part of the face, we fix an original face as input and replace the person's eyes with someone else's eyes, and the same for the nose and mouth. The results of this attempt are shown in Fig. 6. The human cheek and chin portion are tested in Fig. 7 for an additional test of this model. In the future, it may be possible to restore faces using only simple strokes. This can exercise the function of editing or exchanging the attributes of facial parts. We made a face synthesis matrix by exchanging the facial parts of the faces, as shown in Fig. 8. To synthesize a "synthetic face" from more than one person, we randomly combined different parts of the face of 4 people. These results demonstrate the powerful synthesis capability of our method, especially in virtual portrait synthesis.

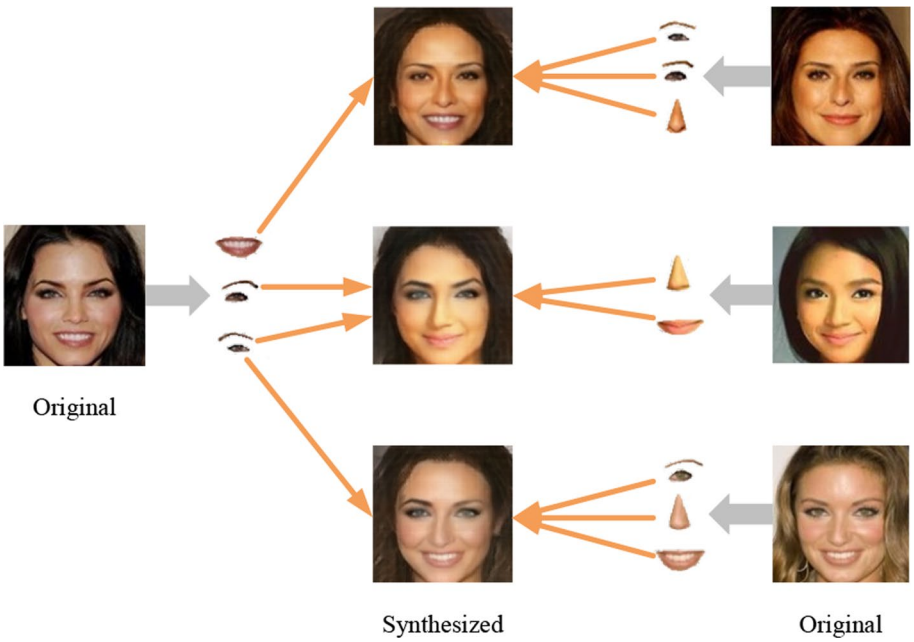

Fig. 5 Examples of synthesized images with facial parts from different persons. The original faces are shown in the left and right columns. The facial parts sent to the algorithm are indicated by arrows. The middle column shows the synthesized result images

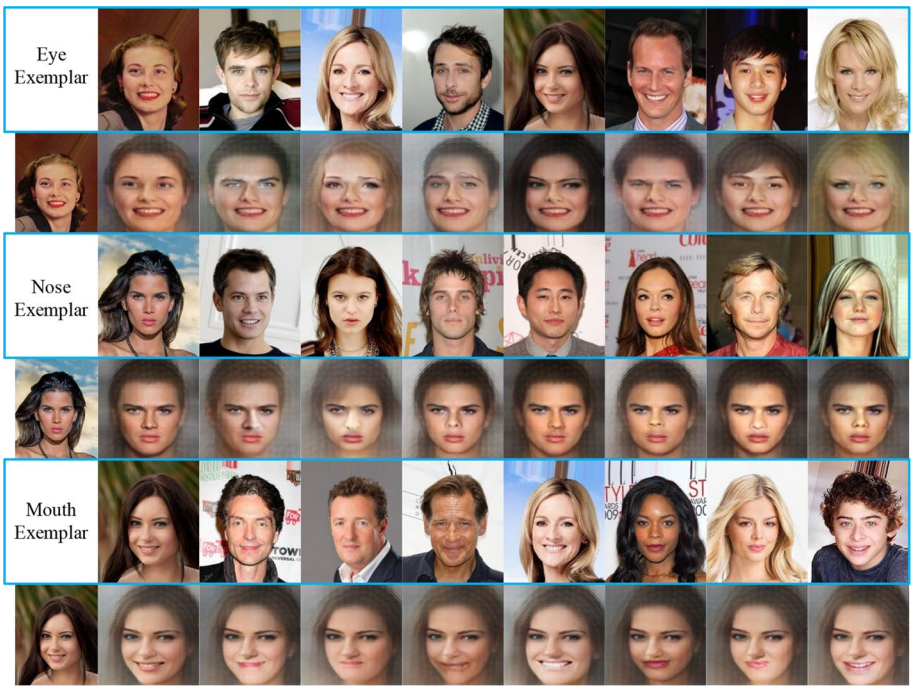

Fig. 6 Face editing by replacing parts of a person's face with someone else's. Rows 1, 3 and 5 are examples of eyes, nose and mouth, respectively. Rows 2, 4 and 6 are the synthetic results after replacing this facial part, where the first image is the reconstruction of the input image

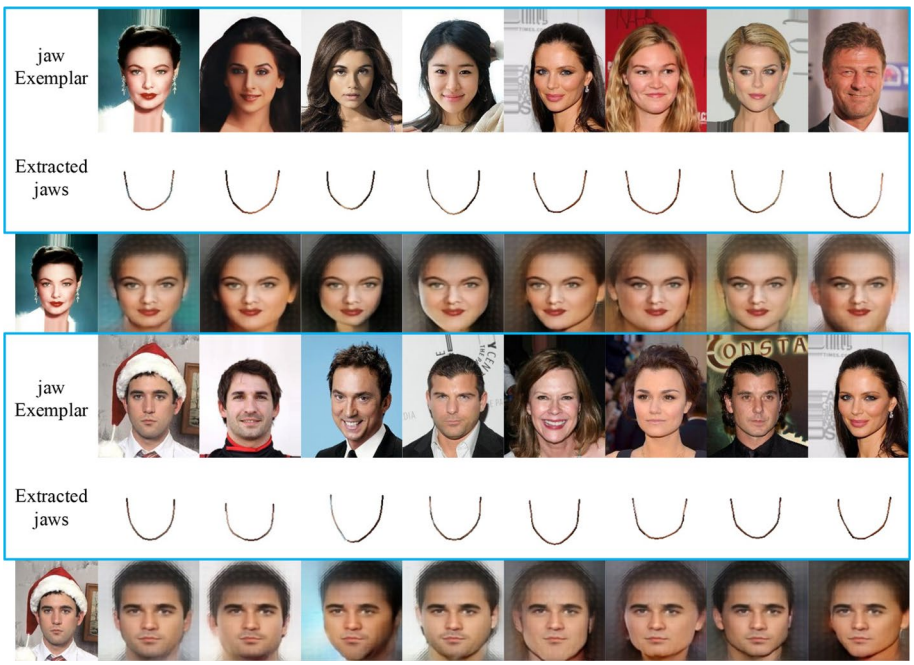

Fig. 7 Face editing by replacing jaws from different persons. Rows 1 and 4 are examples of jaws. Rows 2 and 5 are the extracted jaws. Rows 3 and 6 are the synthetic results, where the first image is the reconstruction of the input image

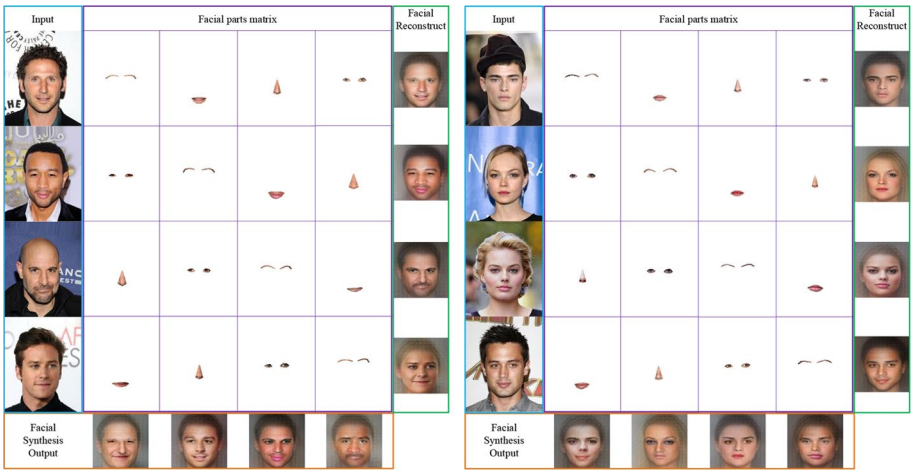

Fig. 8 Random face synthesis from facial parts. Two facial parts combined matrices are shown here. In each matrix, the first column is the input image, each row is a decomposition of the input face, the last column is the reconstruction of the input, and the last row is the portrait synthesis result

4.4 Qualitative comparisons

Facial image generation based on facial parts is a challenging computer vision task, and very few existing works have tried to address this specific task. To this end, we have found some image completion methods to participate in the comparison: PatchMatch (PM) [40], Context Encoder (CE) [5], Image Inpainting [7], pix2pix [16], and Pluralistic Image Completion (PIC) [41]. They are perhaps the most relevant ones to

our work. To make the comparison make sense, the inpainting method [7] is modified to achieve inpainting conditioned on facial parts. In addition, to make a fair comparison, we train these methods with exactly the same input configuration as our method (using CelebA datasets). The output results are shown in Fig. 9.

The PatchMatch and the Context Encoder participated in the evaluation as baseline methods. Because the problem we study may be too extreme, most of the faces are missing in the image, so many previous methods are not perfect in performing this task. The Pix2Pix algorithm can generate visually plausible face image structures and textures, but some structures are distorted, and in some areas, the textures are blurry and inconsistent with known facial parts. In addition, the input facial parts have obvious boundaries with the surrounding areas. Although the results generated by the image inpainting method do not have the boundary problem, the generated content has more serious structure distortion and texture blurring in the synthesized area. Different from Pix2Pix and image inpainting, our method generates more realistic results with fewer artifacts than the two baseline models due to the perceptual loss, which eliminates structural distortion by modeling the high-level semantic structures.

4.5 Quantitative results

In addition to the visual comparison, we also perform quantitative evaluation of different algorithms on the CelebA test dataset. Although in principle there is no good numerical metric to evaluate facial image generation results due to the existence of many possible solutions, we still report the results of three commonly used image quality assessment metrics: PSNR, SSIM and the inception score following the work of [7, 16, 42] (see the Additional file 1 for details on the measurement method). The inception score has been used for GANs to measure generated sample quality and diversity based on the inception model. The comparison results are documented

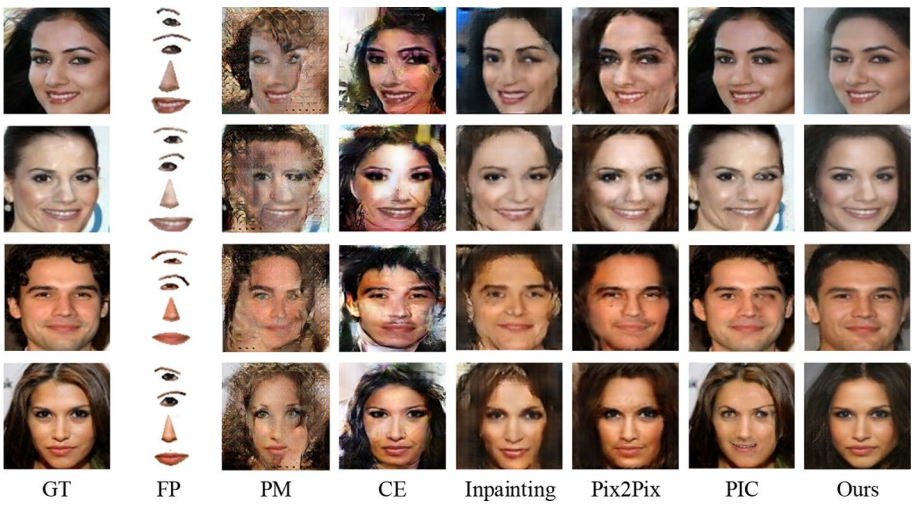

Fig. 9 Comparison of some methods for image generation based on facial parts. Columns from left to right: the original facial images (GT is short for the ground truth), the facial parts (FP), the results of the modified image completion method. PatchMatch (PM) [40], Context Encoder (CE) [5], Image Inpainting [7], pix2pix [16], Pluralistic Image Completion (PIC) [41]

Table 3 PSNR, SSIM and inception scores (IS) on 128 random test images from CelebA

Methods	PSNR	SSIM	IS
Inpainting	28.92	0.762	0.196
Pix2Pix	30.71	0.873	0.124
PIC	29.04	0.791	0.140
ours	34.38	0.956	0.063

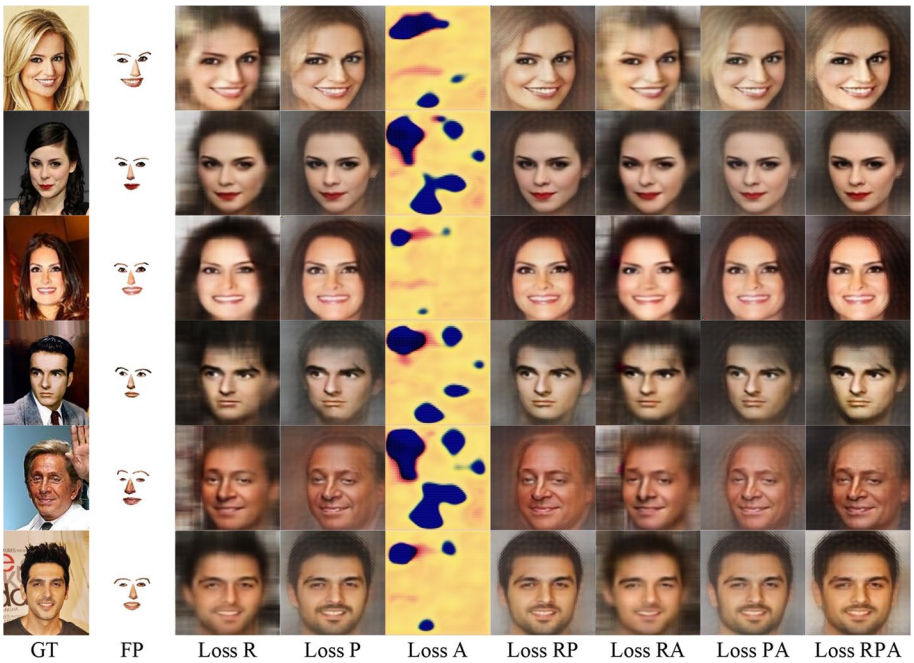

Fig. 10 The result of only using one loss and their pairwise combinations. Loss R, P, and A denote reconstruction loss, perceptual loss and adversarial loss, respectively, and Loss RP is the combination of reconstruction loss and perceptual loss. In this model, using adversarial loss alone does not make any sense

in Table 3. It is clear that our method outperforms the other methods on all three metrics.

4.6 Ablation study

In this section, we conduct ablation experiments to specifically explore the specific role of the three losses. The specific method is to shield one or two losses by changing the weight of the loss to evaluate the effects of loss individually and in combination. We tested reconstruction loss, perceptual loss and adversarial loss separately, as well as pairwise combinations between them (Fig. 10, best viewed by zooming in).

Although reconstruction loss can reconstruct the entire face image, its pixelwise properties determine its lack of generalization performance. Furthermore, images generated only by reconstruction loss may have results that look similar to the input but are prone to overfitting; simple pixel translation may lead to prediction failure. Therefore, we combine perceptual loss and adversarial loss. Among them, the adversarial loss ensures a high degree of realism of the image, making the image more natural, but cannot be

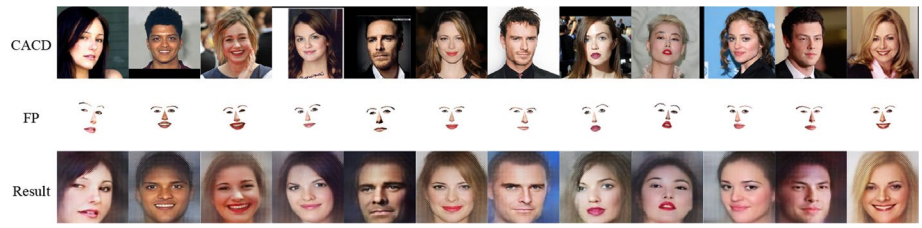

Fig. 11 Apply the model trained on CelebA to the CACD dataset. The first row is randomly selected images from CACD, and the last row is a synthetic portrait of our model. (FP is short for facial parts.)

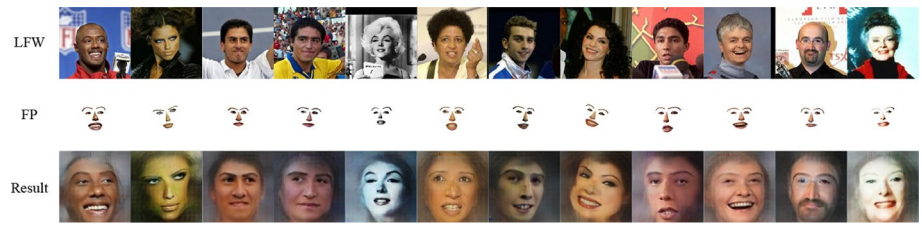

Fig. 12 Apply the model trained on CelebA to the LFW dataset. The first row is randomly selected images from LFW, and the last row is a synthetic portrait of our model. (FP is short for facial parts.)

used alone. The perceptual loss can enable the generated image to reproduce the content (features) and style of the image, which is the most important of the three losses. In particular, to remove noise and mosaics from images, we also introduce total variation regularization to reduce the spikey artifacts of generated images. In summary, the loss makes the picture clearer and more realistic, and regularization can reduce the noise and spikey artifacts of the picture.

4.7 Additional dataset

To study the generalization performance of our model, we tested the model trained on the CelebA dataset on an additional dataset. Here, we use the Cross-Age Celebrity Dataset (CACD) [37] and Labeled Faces in the Wild Home (LFW) [38] for further validation, which are widely used in face image research.

Following the method above, we trained our model on CelebA using 3 loss functions, which took approximately 30 thousand training steps. The results of applying this model to other datasets are shown in Figs. 11 and 12, and the results are not too poor. Because the size of the images in the CACD and LFW datasets is different from that in the training dataset, we uniformly scale them to the same size as the input. We found that the model performed well on the CACD dataset (Fig. 11), which may be because the CACD dataset is similar to the CelebA dataset. However, our model does not perform well on the LFW dataset (Fig. 12). One possible reason is that the portraits in the LFW dataset are much smaller, which provides less information. The face angle may be quite different, and the expressions of the characters are exaggerated. Even in this case, our model's performance is reasonable, which shows the necessity and effectiveness of the combination of the 3 losses. To test our method more extensively, we present more graphical results in the Additional file 1 and discuss some concerns not mentioned above.

5 Conclusions

In this paper, we explore the challenging task of facial image generation from facial parts. A novel end-to-end image synthesizing framework based on deep learning is proposed to address this problem. By introducing multiple loss functions in the facial image generation network, valid and visually realistic images are synthesized semantically based on only a few facial-part patches. We also demonstrate the unique ability of the proposed method to fuse multiple facial parts from different persons to generate a realistic facial image. Extensive qualitative and quantitative comparisons with two existing approaches strongly demonstrate the superiority of the proposed method. Furthermore, the proposed algorithm is highly flexible in various facial synthesis, restoration, and camouflage applications. In the future, we will explore the possibility of allowing a user to manipulate facial attributes, making the algorithm competent for generating multiple output images with different styles, etc. These extensions will greatly enhance the usefulness of the proposed algorithm in many real-world applications.

Abbreviations

CNN: Convolutional neural networks; GANs: Generative adversarial networks; DCGANs: Deep convolutional generative adversarial networks; CGAN: Conditional version of generative adversarial nets; CelebA: Celeb faces attributes dataset; CACD: Cross-age celebrity dataset; LFW: Labeled faces in the wild home; FCENet: Face completion and editing network; RNN: Recurrent neural network; RGAN: Recurrent generative adversarial network; SVD: Singular value decomposition; PSNR: Peak signal-to-noise ratio; SSIM: Structural similarity; IS: Inception score.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13640-022-00585-7.

Additional file 1. Additional results and mentioned methods.

Acknowledgements

No additional acknowledgments.

Author contributions

All authors participated in the conception and writing of the article. YL provided method guidance and writing guidance for this article. JG modified some algorithms and made many valuable suggestions, especially the evaluation part of the method. QS carried out the construction of the main frame and was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (No. 61300072, 31771475).

Availability of data and materials

In this paper, the public facial datasets CelebA, CACD, and LFW are used.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 15 September 2021 Accepted: 22 April 2022
Published online: 10 May 2022

References

1. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (2014), pp. 2672–2680.
2. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in *International Conference on Learning Representations*, ed. by Y. Bengio, Y. LeCun (2016).

3. H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in International Conference on Machine Learning (PMLR, 2019), pp. 7354–7363.
4. M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, CoRR abs/1701.07875, (2017).
5. D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: feature learning by inpainting, in IEEE Conference on Computer Vision and Pattern Recognition (IEEE Computer Society, 2016), pp. 2536–2544.
6. S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion. ACM Trans. Graph. 36, 107:1 (2017).
7. J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. S. Huang, Generative image inpainting with contextual attention, in IEEE Conference on Computer Vision and Pattern Recognition (Computer Vision Foundation / IEEE Computer Society, 2018), pp. 5505–5514.
8. Y. Choi, M.-J. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, StarGAN: unified generative adversarial networks for multi-domain image-to-image translation, in IEEE Conference on Computer Vision and Pattern Recognition (Computer Vision Foundation / IEEE Computer Society, 2018), pp. 8789–8797.
9. G. Perarnau, J. van de Weijer, B. Raducanu, J. M. Álvarez, Invertible Conditional GANs for Image Editing, CoRR abs/1611.06355, (2016).
10. S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, W. He, GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data, CoRR abs/1705.04932, (2017).
11. Z. Zhang, Y. Song, H. Qi, Age progression/regression by conditional adversarial autoencoder, in IEEE Conference on Computer Vision and Pattern Recognition (IEEE Computer Society, 2017), pp. 4352–4360.
12. Y. Feng, S. Yu, H. Peng, Y.-R. Li, J. Zhang, Detect Faces Efficiently: A Survey and Evaluations, ArXiv Preprint ArXiv: 2112.01787 (2021).
13. S. Liu, D. Liu, K. Muhammad, W. Ding, Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615 (2021)
14. S. Liu, S. Wang, X. Liu, C.-T. Lin, Z. Lv, Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**, 90 (2020)
15. S. Liu, S. Wang, X. Liu, A.H. Gandomi, M. Daneshmand, K. Muhammad, V.H.C. de Albuquerque, Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multim. **23**, 2188–2198 (2021). https://doi.org/10.1109/TMM.2021.3065580
16. P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in IEEE Conference on Computer Vision and Pattern Recognition (IEEE Computer Society, 2017), pp. 5967–5976.
17. Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation, in International Conference on Learning Representations (OpenReview.net, 2017).
18. T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, in IEEE Conference on Computer Vision and Pattern Recognition (Computer Vision Foundation / IEEE Computer Society, 2018), pp. 8798–8807.
19. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in *International Conference on Machine Learning*, ed. by F. R. Bach, D. M. Blei, Vol. 37 (JMLR.org, 2015), pp. 448–456.
20. J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in *Computer Vision*, ed. by B. Leibe, J. Matas, N. Sebe, M. Welling, Vol. 9906 (Springer, 2016), pp. 694–711.
21. L. A. Gatys, A. S. Ecker, M. Bethge, A Neural Algorithm of Artistic Style, CoRR abs/1508.06576, (2015).
22. Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in IEEE International Conference on Computer Vision (IEEE Computer Society, 2015), pp. 3730–3738.
23. M. Mirza, S. Osindero, Conditional Generative Adversarial Nets, CoRR abs/1411.1784, (2014).
24. T. Kim, M. Cha, H. Kim, J. K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in *International Conference on Machine Learning*, ed. by D. Precup, Y. W. Teh, Vol. 70 (PMLR, 2017), pp. 1857–1865.
25. J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in IEEE International Conference on Computer Vision (IEEE Computer Society, 2017), pp. 2242–2251.
26. Y. Li, S. Liu, J. Yang, M.-H. Yang, Generative face completion, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 3911–3919.
27. L. Song, J. Cao, L. Song, Y. Hu, R. He, Geometry-aware face completion and editing, in Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33 (2019), pp. 2506–2513.
28. Q. Wang, H. Fan, G. Sun, W. Ren, Y. Tang, Recurrent generative adversarial network for face completion. IEEE Trans. Multimed. **23**, 429 (2020)
29. M. Jian, K.-M. Lam, J. Dong, A novel face-hallucination scheme based on singular value decomposition. Pattern Recogn. **46**, 3091 (2013)
30. M. Jian, K.-M. Lam, J. Dong, Facial-feature detection and localization based on a hierarchical scheme. Inf. Sci. **262**, 1 (2014)
31. M. Jian, K.-M. Lam, Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. IEEE Trans. Circ. Syst. Video Technol. **25**, 1761 (2015)
32. M. Jian, C. Cui, X. Nie, H. Zhang, L. Nie, Y. Yin, Multi-view face hallucination using SVD and a mapping model. Inf. Sci. **488**, 181 (2019)
33. X. Wang, Y. Li, H. Zhang, Y. Shan, Towards real-world blind face restoration with generative facial prior, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), pp. 9168–9178.
34. D.E. King, Dlib-Ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755 (2009)
35. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *International Conference on Learning Representations*, ed. by Y. Bengio, Y. LeCun (2015).
36. B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1452 (2018)
37. B.-C. Chen, C.-S. Chen, W. H. Hsu, Cross-age reference coding for age-invariant face recognition and retrieval, in European Conference on Computer Vision (Springer, 2014), pp. 768–783.

38. G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, in Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008).
39. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: a system for large-scale machine learning, in *USENIX Symposium on Operating Systems Design and Implementation*, ed. by K. Keeton, T. Roscoe (USENIX Association, 2016), pp. 265–283.
40. C. Barnes, E. Shechtman, A. Finkelstein, D.B. Goldman, PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**, 24 (2009)
41. C. Zheng, T.-J. Cham, J. Cai, Pluralistic image completion, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 1438–1447.
42. G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in *Computer Vision*, ed. by V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss, Vol. 11215 (Springer, 2018), pp. 89–105.

Publisher's Note