Multimedia Tools and Applications (2022) 81:29785–29797
https://doi.org/10.1007/s11042-022-12912-1

Mask removal : Face inpainting via attributes

Yefan Jiang[1] · Fan Yang[2] · Zhangxing Bian[3] · Changsheng Lu[4] · Siyu Xia[1]

Received: 2 January 2021 / Revised: 9 April 2021 / Accepted: 9 March 2022 /
Published online: 5 April 2022

Abstract
Due to the outbreak of the COVID-19 pandemic, wearing masks in public areas has become an effective way to slow the spread of disease. However, it also brings some challenges to applications in daily life as half of the face is occluded. Therefore, the idea of removing masks by face inpainting appeared. Face inpainting has achieved promising performance but always fails to guarantee high-fidelity. In this paper, we present a novel mask removal inpainting network based on face attributes known in advance including nose, chubby, makeup, gender, mouth, beard and young, aiming to ensure the repaired face image is closer to ground truth. To achieve this, a dual pipeline network based on GANs has been proposed, one of which is a reconstructive path used in training that utilizes missing regions in ground truth to get prior distribution, while the other is a generative path for predicting information in the masked region. To establish the process of mask removal, we build a synthetic facial occlusion that mimics the real mask. Experiments show that our method not only generates faces more similarly aligned with real attributes, but also ensures semantic and structural rationality compared with state-of-the-art methods.

Keywords Image inpainting · Mask removal · Face attributes · GAN

1 Introduction

Image inpainting is a task of generating visually realistic content in missing regions of corrupted input images. It has a wide range of applications. For example, it allows removing unwanted objects from an image or synthesizing features under occlusion areas. Face inpainting is an interesting and challenging branch. The main challenge of face inpainting

✉ Siyu Xia
xsy@seu.edu.cn

[1] Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing, 210096, People's Republic of China

[2] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China

[3] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA

[4] College of Engineering and Computer Science, The Australian National University, Canberra, Australia

lies in that face is a region with strong structural and semantic features while the texture of the facial features is continuous and unobtrusive. These unique properties require the face inpainting not only to recover reasonable semantic information but also to ensure the continuity of texture and structure.

Recently, along with the development of deep learning, significant progress has been made in image and face inpainting [12, 17, 21, 34, 35, 39]. Although these prior works can generate proper results for the missing region, they cannot ensure high-fidelity. It is widely assumed that, once the missing region consists of a large continuous area or involves key information, it is almost impossible to guarantee the repaired results exactly match the ground truth.

Nevertheless, we can still improve the authenticity of the inpainting under reasonable assumptions. Compared with intricate natural scenes, human faces are easier to make reasonable assumptions. The human face involves exact attributes including eyes, nose, mouth, cheeks, beard, and other features. Specifically, taking mouth as an example, there are big and small kinds in size, open and close in posture. If attributes in the occluded face area are determined in advance according to ground truth, for example, the face has a beard or the mouth is open, the generated result may be more similar to the ground truth. Unfortunately, this is a paradox. The main challenge of image inpainting is the lack of information. Thus, the basic assumption of this paper is that we can obtain face attributes in advance and these face attributes are in line with ground truth.

Based on this observation, our main goal is then to ensure the repaired results meeting the basic semantics content, and visually approximate to the ground truth under the guidance of face attributes during face inpainting (as shown in Fig. 1). Wearing masks in public is now a consensus, which happens to be a practical application scenario for mask removal. Thus, we proposed a face inpainting method for mask removal accordingly. To simulate the actual scene of wearing a mask, we made a synthetic facial occlusion that mimics the real mask, resulting in 1) large-scale irregular continuous area is occluded; 2) the lower half of the face including nose, cheek, mouth, and other face details are occluded. The main contributions of our paper can be summarized as:

1. A novel face inpainting method aiming to remove face masks as well as realistically recover the missing face region under the guidance of the face attributes is proposed.
2. We design a novel dual pipeline network structure with an attention mechanism to effectively exploit both known and unknown regions' information to obtain semantic results.
3. Experiments results compared with the state-of-the-arts demonstrate that our method can achieve competitive results.

2 Related work

2.1 Automatic image inpainting

The approaches related to image inpainting can be roughly classified into two categories: patch-based or diffusion-based methods and deep learning methods.

Patch-based methods [1, 4, 8, 9, 14] fill the missing region patch-by-patch by searching and extending pixels in the undamaged area of the image. These algorithms work well on images that the unobstructed areas that have the contents for obstructed regions, which cannot be guaranteed in most real scenes. Diffusion-based methods [14, 16, 23] fill the

Fig. 1 Face image inpainting results by our approach. It takes mask-wearing faces as input and facial attributes as guidance to obtain the mask removal result

missing region by propagating the neighborhood content to the missing holes, which are robust for dealing with small holes but fail to complete complex images with semantic structures.

Deep learning methods are proposed to image inpainting that directly generates the missing regions. Pathak et al. [18] proposed a context encoder to extract features and then decode the features to reconstruct the output. Iizuka et al. [7] proposed global and local discriminators to handle high-resolution images and generate more realistic results. Yu et al. [34] proposed an end-to-end image inpainting model with contextual attention to borrowing information from surroundings. However, these methods can only solve rectangular masks. To handle irregular masks, Liu et al. [12] proposed partial convolution to merely utilize valid pixels and update the masks and weights with layers.

However, partial convolution [12] heuristically classifies all spatial locations to be either valid or invalid and the update way is un-learnable. Therefore, gated convolution [35] is proposed to learn a dynamic feature selection mechanism for each spatial location and obtain competitive results. After that, Sagong et al. [22] propose PEPSI to improve the structure of gated convolution which reduces the number of convolution operation almost by half but exhibits superior performance to other models in terms of testing time and qualitative scores.

Besides, Yi et al. [32] propose a Contextual Residual Aggregation (CRA) mechanism that can produce high-frequency residuals for missing contents by weighted aggregating residuals from contextual patches, thus only requiring a low-resolution prediction from the network. CRA enables large images (up to 8K) with considerable hole sizes to be inpainted with limited memory and computing resources, which is intractable for prior methods. The disadvantage is that it is not possible to generate multiple styles of repair images. Wu et al. [28] propose a new end-to-end and coarse-to-fine generative model through combining a local binary pattern (LBP) learning network with an actual inpainting network to tackle various unpleasant artifacts, especially in the boundary and highly textured regions. But how to select these two sub-networks is very crucial and could lead to vastly different results. Zhao et al. [38] present Unsupervised Cross-space Translation Generative Adversarial Network (UCTGAN). The network realizes one-to-one mapping between instance image space and conditional completion image space, which can significantly reduce the possibility of mode collapse and improve the diversity of restored images. But it can't inpaint images including free-form masks.

Recently, Qin et al. [19] propose one method based on weighted facial similarity for face inpainting with large missing regions. Han et al. [5] propose one face image inpainting method with evolutionary generators to overcome the gradient vanishing problem. Yang et al. [31] use paired discriminator to inpaint damaged face images to keep stronger semantic consistency. These methods achieve good image inpainting results by optimizing the structure, but they do not combine the original semantics of human face well.

2.2 Guided image inpainting

In general image inpainting tasks, input includes a corrupted image as well as a mask that indicates missing pixels. Blind image inpainting like [26] only takes corrupted images as input and adopts mask prediction network to estimated masks. However, more inpainting methods adopt additional input besides image and mask to improve inpainting results. Yu et al. [35] input user sketch to extend inpainting network as a user-guided system. Ren et al. [21] use incomplete image structure as input which is firstly reconstructed as global structure information in texture generation. Nazeri et al. [17] generate edges of missing regions of the image and use hallucinated edges as a priori to guide inpainting. Xiong et al. [30] learn to predict the foreground contour first and then repair the missing region using the predicted contour as guidance. However, these works mainly focus on vision information as additional input. To acquire more semantically accurate inpainting images, Zhang et al.

[36] propose a novel model combining the descriptive text as part of the input. In this paper, we use face attributes as an additional input to guide inpainting results.

2.3 Face generation with attributes

Face generation is a challenging problem until the rapid advancement of the generative adversarial network (GAN). The original GAN work [3] introduced a novel framework which contains two feed-forward networks, a generator G and a discriminator D. At present, countless GAN variants [11, 15, 20, 33, 40] are proposed in many computer vision applications. Thus, face generation problems obtain great development. To generate faces of different ages from an arbitrary query face without knowing its true age, CAAE [37] was proposed, which used age as attributes in training. Xiao et al. [29] proposed ELEGANT using face attributes to transfer face features between two face images. Nasir et al. [17] used textual descriptions of the face in face generation. However, in the above studies, most of input images are complete and thus semantic information could be utilized to guide image generation. While in inpainting problems, the image is corrupted and the feature is no longer complete. Nevertheless, we can still get some inspiration from above works because all of us need to use face attributes as part of input to generate human faces. Hence, how to use attributes and combine them with latent features is the key problem in face inpainting, which will be discussed in Section 3.1.

3 Approach

The detailed structure of the proposed network is shown in Fig. 2. Suppose we have a complete image I_g, which serves as ground truth. Then I_g is degraded by a piece of masked regions M and became I_m. I_m is a masked image needed to be repaired and its complement is defined as I_c. Besides, some face attributes of I_g will be considered as part of the input, named I_{attr}.

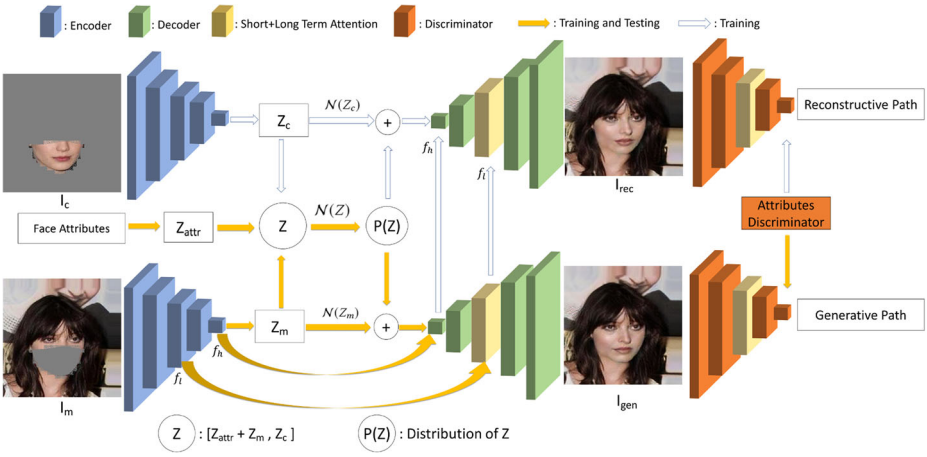


Fig. 2 Overview of our framework. The upper one is reconstructive path only used in training. The lower one is generative path used both in training and testing. Face attributes are fed to guide the inpainting results during the whole process

3.1 Dual pipeline network structure

The framework consists of two paths: a reconstructive path and a generative path. The upper path is a reconstructive path using information from the whole image including I_c and I_m. I_c is used to infer the whole image's latent information only in training. The lower path is the generative path. Like other one-path inpainting networks, it only uses I_m as input to obtain latent information and will be used both in training and testing. Both the reconstructive path and the generative path will share identical weights.

The input and output face images are 256×256 RGB images. A 5-layer neural network with residual modules is adopted as the encoder, aiming to extract the semantic information about the masked area from I_m and I_c.

$$Z_{c/m} = Enc(I_c/I_m) \tag{1}$$

where Enc denotes the image encoder. In the reconstructive path, I_c will be fed into the encoder. As I_c has the information in masked regions, Z_c denotes the effective area which represents the true characteristics of the occluded area. While in the generative path, Z_m is the output of the encoder and has the information mainly about the unmasked region of I_m.

As we use face attributes as guidance, we select 7 representative attributes including the big nose, chubby face, makeup face, gender is male, open mouth, no beard, and young face according to CelebA [13]. We define these face attributes as one vector in which 1 is true and -1 is false. The details are shown in Section 4.1. As for the input of face attributes, we also use an encoder to change the label into a feature tensor.

$$Z_{attr} = Enc_{attr}(I_{attr}) \tag{2}$$

where Enc_{attr} is the face attributes encoder and I_{attr} is an attributes vector. Z_{attr} is the feature tensor indicating unique characteristic. In practice, we ensure the Z_{attr} has the same dimension with Z_m and Z_c.

After extracting the image's high-level features of the input face in I_c and I_m, along with face attributes Z_{attr} , these tensors are concatenated and fed into a distribution network.

$$Z = [Z_{attr} + Z_m, Z_c] \tag{3}$$

$$P(Z) = \mathcal{N}(Z) \tag{4}$$

where $\mathcal{N}$ is a sampler and $P(Z)$ is distribution of Z. Consider Z_c has specific semantic feature and Z_{attr} has abstract face attributes in masked areas while Z_m only has unmasked areas information and need to infer and generate masked face, we prefer adding Z_{attr} and Z_m both in training and testing rather than concatenating to guide generative path obtain semantic face images. As Z_c is only available in training, we utilize it to get a prior distribution of effective area and assist generative path to get a distribution that is closer to the real situation. The effectiveness of this process will be further discussed in Section 4.4.

Inspired by [39], we feed the low-level image features from the encoder layer to the decoder layer in the generator through a high way path based on short+long term attention. Besides, it will also be shared with the reconstructive path in the same layer. The short+long term attention can maintain fine features in the encoder and obtain semantical features in the decoder.

3.2 Discriminator on face images

Follow the principle of GAN, the discriminator firstly require the generator to get more realistic faces. And both in the reconstructive path and generative path, the discriminator

D_r and D_g are based on LSGAN [15] which is better than original GAN according to our research.

$$\mathcal{L}_{ad}^{r} = \|D_r(I_{rec}) - D_r(I_g)\|_2 \tag{5}$$

where $\mathcal{L}_{ad}^{r}$ is the adversarial loss in reconstructive path and I_{rec} is reconstructive output image.

$$\mathcal{L}_{ad}^{g} = [D_g(I_{gen}) - 1]^2 \tag{6}$$

where $\mathcal{L}_{ad}^{g}$ is the adversarial loss in generative path and I_{gen} is generative image.

Since face attributes are used as part of the input, the output face image should also correspond to the expected attributes. To distinguish the attributes of output face images, a pre-trained face detection model named D_{attr} was build using ResNet [25] as the backbone. D_{attr} is used to detect both reconstructive and generative face images and give confidence for face attributes.

$$\mathcal{L}_{attr}^{r} = \sum_{i=0}^{6} |D_{attr}(I_{rec}^{(i)}) - I_{attr}^{(i)}| \tag{7}$$

$$\mathcal{L}_{attr}^{g} = \sum_{j=0}^{6} |D_{attr}(I_{gen}^{(j)}) - I_{attr}^{(j)}| \tag{8}$$

where $\mathcal{L}_{attr}^{r}$ is the loss for attributes in reconstructive path and $\mathcal{L}_{attr}^{g}$ is the loss in generative path. As we have 7 face attributes as input, i and j are the index for different attributes. Besides, $D_{attr}(I_{rec}^{(i)}) \in [-1, 1]$, $D_{attr}(I_{gen}^{(j)}) \in [-1, 1]$ and $I_{attr}^{(i/j)} \in \{-1, 1\}$. If the score of an attribute is 0, it means that the attribute is difficult to distinguish, and the result of image inpainting is not consistent with ground truth. Once the loss score of $\mathcal{L}_{attr}^{r}$ or $\mathcal{L}_{attr}^{g}$ reduced, it means the attributes of reconstructive or generative face image are identical with input attributes. The specific results of this discriminator will be shown in Section 4.4.

3.3 Loss function

The proposed appearance matching loss is used to constraint the appearance fields. It determines whether the constructive or generative image match ground-truth in structure and texture. It is computed as

$$\mathcal{L}_{app}^{r} = \|I_{rec} - I_g\|_1 \tag{9}$$

$$\mathcal{L}_{app}^{g} = \|I_{gen} - I_g\|_1 \tag{10}$$

Since we use the prior distribution of known regions in I_c and sub-distribution of occlusion area in I_m, KL divergence term is adopted in our network to regularize the sample function and fixed latent distribution according to [36]. In reconstructive path, posterior sampling function q is used, z is the latent vector, and $\mathcal{N}$ is the Gaussians. KL loss is formulated as:

$$\mathcal{L}_{KL}^{r} = -KL(q(z|I_c, P(Z))\|\mathcal{N}(0, 1)) \tag{11}$$

For the generative path, conditional prior and likelihood p is used:

$$\mathcal{L}_{KL}^{g} = -KL(q(z|I_c, P(Z))\|p(z|I_m)) \tag{12}$$

Overall, the total loss function consists of four groups:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{app}(\mathcal{L}_{app}^{r} + \mathcal{L}_{app}^{g}) + \lambda_{KL}(\mathcal{L}_{KL}^{r} + \mathcal{L}_{KL}^{g}) \\ & + \lambda_{ad}(\mathcal{L}_{ad}^{r} + \mathcal{L}_{ad}^{g}) + \lambda_{attr}(\mathcal{L}_{attr}^{r} + \mathcal{L}_{attr}^{g})\end{aligned} \tag{13}$$

where λ_{app}, λ_{KL}, λ_{ad}, λ_{attr} are hyperparameters. In our experiments, we set $\lambda_{app} = 20$, $\lambda_{KL} = 20$, $\lambda_{ad} = 1$ and $\lambda_{attr} = 1$.

4 Experiments

4.1 Implementation details

We evaluate our model on CelebA [13] which have the annotations for 40 attributes and 5 landmark locations.

Although the face has multiple attributes, we only need to focus on the occluded area. Note that our target is to remove mask via face attributes, we select 7 representative and generic attributes mainly in the masked region to reduce the impact of unique properties among all humans including BigNose (big nose), Chubby (chubby face), Makeup (makeup face), Male (gender is male), MouthOpen (open mouth), NoBeard (no beard) and Young (young face). Among these attributes, 1 is true and -1 is false according to the label in CelebA. During training, we use a pre-trained face detection model to distinguish the attributes of output face images, which scores between -1 and 1. If the score of an attribute is 0, it means that the attribute is difficult to distinguish and the result of image inpainting is not consistent with ground truth. We use this inference as the criterion for discriminators.

Besides, to our knowledge, there is no open-source datasets about-face wearing a mask by the time of this submission. Hence we made a synthetic facial occlusion that mimicked the real mask. Firstly, we use PRNet [2] to get face pose estimation and alignment. Secondly, we focus on the local key points based on face mask in reality and create a mask dataset shown in Fig. 3. Each face corresponds to a unique mask. Thus, we have the same amount of masks as training and testing images.

Then we randomly select 180,000 training images and 2112 testing images from CelebA and get the corresponding attributes and mask image. Both face images and mask images are

Fig. 3 Our masks dataset. Due to each face corresponds to a unique mask as input, the masks in dataset have different shapes and sizes

Table 1 Quantitative comparison with state-of-the-arts. ↓ means lower is better and ↑ means higher is better

Method	l_1 loss ↓	PSNR ↑	SSIM ↑	FID ↓
GC [35]	2.5170	29.0535	0.9103	5.8251
EC [17]	2.7296	27.3367	0.9143	9.6962
PIC [39]	2.6139	29.5075	0.9084	**3.5977**
Ours	**2.2646**	**30.2009**	**0.9180**	3.7065

Bold entries mean the best

resized into 256 × 256 while face images are RGB images and mask images are grayscale images.

Our proposed model is implemented in PyTorch. The network is trained using 256 × 256 images with batch size as 8. We use the Adam optimizer [10] with learning rate as 10^{-4}.

4.2 Quantitative results

Similar to most inpainting works, we measure the quality of our results using the following metrics: 1) mean l_1 loss; 2) peak signal-to-noise ratio (PSNR); 3) structural similarity index (SSIM) [27]. Although the rationality of pixels cannot represent semantic rationality, these

(a) Ground truth (b) Input masked face (c) GC [35] (d) EC [17] (e) PIC [39] (f) Ours

Fig. 4 The qualitative comparisons with existing state of the art methods on CelebA. Zoom in for a better view

Table 2 Quantitative comparison in ablation studies. ↓ means lower is better and ↑ means higher is better. Besides, w/o means without face attributes and w/ means with face attributes

Method	l_1 loss ↓	PSNR ↑	SSIM ↑	FID ↓
Ours	**2.2646**	**30.2009**	**0.9180**	**3.7065**
Ours w/o	2.3064	30.0246	0.9168	4.2249
Z_{swap} w/o	2.4675	29.3965	0.9146	4.3838
Z_{swap} w/	2.4551	29.5571	0.9157	5.0743

Bold entries mean the best

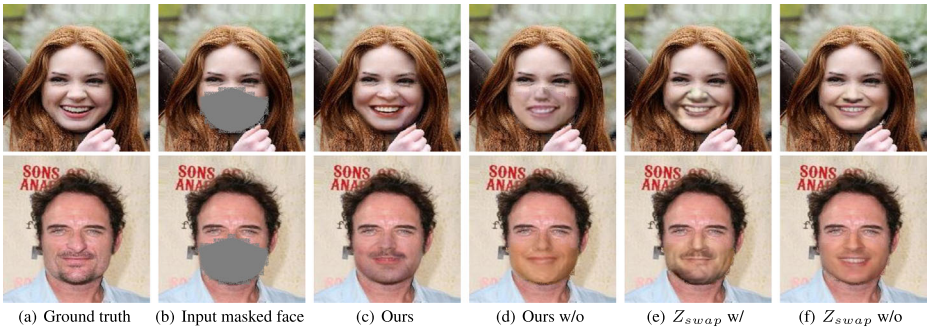
metrics could measure the distortions of the results. Besides, we also introduce Fréchet Inception Distance (FID) [6] as one of our metrics. As FID calculates the Wasserstein-2 distance between two distributions, it can indicate the perceptual quality of the results. In this paper, we use the pre-trained Inception-V3 model [24] to extract features of real and inpainted images when calculating FID scores.

The results over CelebA are reported in Table 1. It can be seen that our approach achieves better performance against other methods in most metrics, which suggests that our method could produce inpainting results with higher quality. Although in FID, PIC [39] obtains a score slightly better than ours, it is worth noting that PIC can generate multiple diverse results by sampling various results. For comparison, we use the same way mentioned in [39] to obtain quantitative measures. And we will discuss and show visualized results in Section 4.3.

4.3 Qualitative results

Next, we visually compare our model with previous state of the art methods [17, 35, 39]. Fig. 4 shows a sample of automatic inpainting results. Specifically, we use the pre-trained model of GC [35] and EC [17] because they did similar work on CelebA. As for PIC [39], we train and finetune their model in our mask dataset for a fair comparison.

It can be seen that the results of GC and EC suffer from artifacts because of the large and continuous mask as well as facial complexity. Although EC can infer edges of missing regions, our mask is a large-scale continuous occlusion and thus it is difficult for EC to infer such information and structures in the masked face. Compared with these two methods



(a) Ground truth (b) Input masked face (c) Ours (d) Ours w/o (e) Z_{swap} w/ (f) Z_{swap} w/o

Fig. 5 The qualitative comparisons in ablation studies. (From left to right) Ground truth, input masked face, results of ours, results of ours without attributes, results of Z_{swap} with attributes, and results of Z_{swap} without attributes. Zoom in for a better view

Table 3 Face attributes evaluation. The attributes include BigNose, Chubby, Makeup, Male, MouthOpen, NoBeard and Young. Among them, value close to 1 means true while -1 means false

Picture	Face attributes						
	BigNose	Chubby	Makeup	Male	MouthOpen	NoBeard	Young
Image1 (GT)	−1	−1	−1	−1	1	1	1
Image1 (Ours)	−0.354	−0.856	−0.028	−0.927	0.843	0.826	0.723
Image2 (GT)	1	−1	−1	1	1	−1	1
Image2 (Ours)	0.100	−0.707	−0.942	1.000	−0.665	−1.000	0.068

above, PIC can recover reasonable face features. However, masked faces lacks information about the mouth, nose, cheek, beard, and other details. So it is almost impossible for PIC to guarantee the correspondence of attributes between generated results and the ground truth.

In contrast, our method uses face attributes as part of the input to guide face inpainting, we can generate face more similarly aligned with real attributes as well as ensure semantic and structural rationality.

4.4 Ablation studies

We further perform experiments to study the effect of the components of our model, especially face attributes. Using the same pre-trained model, we change the test details where face attributes are abandoned. Besides, noticed that in (3), we alter the feature tensor that $Z_{swap} = [Z_m, Z_{attr} + Z_c]$. That is, masked region information is combined in the reconstructive path and no prior guidance can be used in the generative process. Although some features are shared, this way weakens the effect of face attributes. In the same way, we also maintain and abandon the face attributes in testing. The evaluation is shown in Table 2 and the results are shown in Fig. 5.

Our current method leads to significantly better results. It can be seen that without face attributes in testing, both semantic and photo-realistic features will be damaged. Although Z_{swap} can get good visual effects, its characteristics cannot match ground truth well. In short, our current method can obtain competitive results not only in visualized results but also on evaluation metrics.

As mentioned in (7) and (8), we will show our face attributes evaluation of output face images. We also use our results in Fig. 5 to give an example. The evaluation is shown in Table 3. Compare with the results of the visualization, it can be seen that some features match labels well while some cannot be well classified by discriminator because of the ambiguous labels and classifier error, which inevitably affect the inpainting results. How to improve a certain attribute will be discussed in the future work.

5 Conclusion

In this paper, we proposed a novel dual pipeline network for mask removal. From a practical perspective, we focused on face wearing a mask and used face attributes as input to guide the inpainting process. We showed that our network architecture and loss functions could use face attributes information and remove mask well. Besides, we built a mask dataset

simulating the real occlusion effect of the mask. Experiments showed that our model could obtain competitive results compared with several state-of-the-art methods.

Declarations

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

References

1. Ding D, Ram S, Rodriguez JJ (2019) Image inpainting using nonlocal texture matching and nonlinear filtering. IEEE Trans Image Process 28(4):1705–1719
2. Feng Y, Wu F, Shao X, Wang Y, Zhou X (2018) Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV
3. Goodfellow I, Pougetabadie J, Mirza M, Xu B, Wardefarley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. pp 2672–2680
4. Guo Q, Gao S, Zhang X, Yin Y, Zhang C (2018) Patch-based image inpainting via two-stage low rank approximation. IEEE Trans Vis Comput Graph 24(6):2023–2036
5. Han C, Wang J (2021) Face image inpainting with evolutionary generators. IEEE Signal Process Lett 28:190–193. https://doi.org/10.1109/LSP.2020.3048608
6. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637
7. Iizuka S, Simoserra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graph 36(4):107
8. Jin KH, Ye JC (2015) Annihilating filter-based low-rank hankel matrix approach for image inpainting. IEEE Trans Image Process 24(11):3498–3511
9. Kawai N, Sato T, Yokoya N (2016) Diminished reality based on image inpainting considering background geometry. IEEE Trans Vis Comput Graph 22(3):1236–1247
10. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization
11. Lin CH, Chang CC, Chen YS, Juan DC, Wei W, Chen HT (2019) Coco-gan: Generation by parts via conditional coordinating. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV)
12. Liu G, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B (2018) Image inpainting for irregular holes using partial convolutions. In: The european conference on computer vision (ECCV)
13. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. pp 3730–3738
14. Lu H, Liu Q, Zhang M, Wang Y, Deng X (2018) Gradient-based low rank method and its application in image inpainting. Multimed Tools Appl 77(5):5969–5993
15. Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP (2017) Least squares generative adversarial networks, pp 2813–2821
16. Mo J, Zhou Y (2019) The research of image inpainting algorithm using self-adaptive group structure and sparse representation. Clust Comput 22(3):7593–7601
17. Nazeri K, Ng E, Joseph T, Qureshi F, Ebrahimi M (2019) Edgeconnect: Structure guided image inpainting using edge prediction. In: The IEEE international conference on computer vision (ICCV) workshops
18. Pathak D, Krähenbühl P, Donahue J, Darrell T, Efros A (2016) Context encoders: Feature learning by inpainting
19. Qin J, Bai H, Zhao Y (2020) Face inpainting network for large missing regions based on weighted facial similarity. Neurocomputing 386:54–62. https://doi.org/10.1016/j.neucom.2019.12.079, https://www.sciencedirect.com/science/article/pii/S0925231219317941
20. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks
21. Ren Y, Yu X, Zhang R, Li TH, Liu S, Li G (2019) Structureflow: Image inpainting via structure-aware appearance flow. pp 181–190
22. Sagong M, Shin Y, Kim S, Park S, Ko S (2019) Pepsi : Fast image inpainting with parallel decoding network. pp 11360–11368
23. Sridevi G, Kumar SS (2019) Image inpainting based on fractional-order nonlinear diffusion for image reconstruction. Circuits Systems and Signal Processing 38(8):3802–3817

24. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
25. Visin F, Kastner K, Cho K, Matteucci M, Courville A, Bengio Y (2015) Renet: A recurrent neural network based alternative to convolutional networks arxiv: Computer Vision and Pattern Recognition
26. Wang Y, Chen Y, Tao X, Jia J (2020) Vcnet: A robust approach to blind image inpainting arxiv: Computer Vision and Pattern Recognition
27. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
28. Wu H, Zhou J, Li Y (2020) Deep generative model for image inpainting with local binary pattern learning and spatial attention
29. Xiao T, Hong J, Ma J (2018) Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: Proceedings of the European conference on computer vision (ECCV), pp 172–187
30. Xiong W, Yu J, Lin Z, Yang J, Lu X, Barnes C, Luo J (2019) Foreground-aware image inpainting. pp 5840–5848
31. Yang X, Xu P, Xue Y, Jin H (2021) Contextual feature constrained semantic face completion with paired discriminator. IEEE Access 9:42100–42110. https://doi.org/10.1109/ACCESS.2021.3065661
32. Yi Z, Tang Q, Azizi S, Jang D, Xu Z (2020) Contextual residual aggregation for ultra high-resolution image inpainting. In: Conference on computer vision and pattern recognition (CVPR)
33. Yi Z, Zhang H, Tan P, Gong M (2017) Dualgan: Unsupervised dual learning for image-to-image translation. pp 2868–2876
34. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition
35. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2019) Free-form image inpainting with gated convolution. pp 4471–4480
36. Zhang L, Chen Q, Hu B, Jiang S (2020) Neural image inpainting guided with descriptive text. arXiv:abs/2004.03212
37. Zhang Z, Song Y, Qi H (2017) Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on computer vision and pattern recognition (CVPR)
38. Zhao L, Mo Q, Lin S, Wang Z, Lu D (2020) Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
39. Zheng C, Cham TJ, Cai J (2019) Pluralistic image completion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1438–1447
40. Zhu J, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International conference on computer vision (ICCV), pp 2242–2251