When Face Completion Meets Irregular Holes: an Attributes Guided Deep Inpainting Network

Jie Xiao
Sun Yat-Sen University
xiaoj45@mail2.sysu.edu.cn

Dandan Zhan
Sun Yat-Sen University
zhandd3@mail2.sysu.edu.cn

Haoran Qi
Sun Yat-Sen University
qihr3@mail2.sysu.edu.cn

Zhi Jin*
Sun Yat-Sen University
Guangdong Provincial Key Laboratory of Fire Science and Technology
jinzh26@mail.sysu.edu.cn

ABSTRACT

Lots of convolutional neural network (CNN)-based methods have been proposed to implement face completion with regular holes. However, in practical applications, irregular holes are more common to see. Moreover, due to the distinct attributes and large variation of appearance for human faces, it is more challenging to fill irregular holes in face images while keeping content consistent with the rest region. Since facial attributes (*e.g.*, gender, smiling, pointy nose, etc.) allow for a more understandable description of one face, they can provide some hints that benefit the face completion task. In this work, we propose a novel attributes guided face completion network (**AttrFaceNet**), which comprises a facial attribute prediction subnet and a face completion subnet. The attribute prediction subnet predicts facial attributes from the rest parts of the corrupted images and guides the face completion subnet to fill the missing regions. The proposed AttrFaceNet is evaluated in an end-to-end way on commonly used datasets CelebA and Helen. Extensive experimental results show that our method outperforms state-of-the-art methods qualitatively and quantitatively especially in large mask size cases. Code is available at https://github.com/FVL2020/AttrFaceNet.

CCS CONCEPTS

• **Computing methodologies** → *Image processing*.

KEYWORDS

face inpainting; irregular holes; attribute prediction

ACM Reference Format:
Jie Xiao, Dandan Zhan, Haoran Qi, and Zhi Jin. 2021. When Face Completion Meets Irregular Holes: an Attributes Guided Deep Inpainting Network. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475466

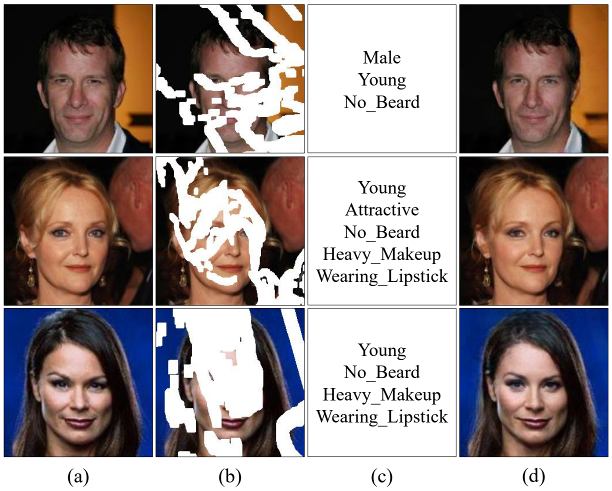


Figure 1: Face completion results of our method. The images in each row from left to right are (a) ground truth image, (b) corrupted image, (c) predicted attributes and (d) image completed by AttrFaceNet, respectively. Noting that we only list the predicted attributes with confidence greater than 0.8 in the cases where there are too many positive attributes predicted.

1 INTRODUCTION

Image inpainting, which is also known as image completion, is intended to generate visually plausible pixels in the missing parts of corrupted images, meanwhile, to make sure the generated content being coherent with the unmasked region. Hence, it has many important applications. For example, the unpleasant object in an image can be masked and then removed during the inpainting process [4]. Image inpainting can also restore the cracks on old photos [2] and editing images as needed [1]. Moreover, it can be used as a pre-processing step for face recognition when the face is occluded by some objects.

As a sub-branch of image inpainting, face completion focuses on filling holes of corrupted face images, which is more difficult than completing general images. Although the human face has a symmetric structure, the texture details of the corresponding left

[*] Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475466

and right parts are obviously different, as no one has exactly the same left and right eyes. This hinders the application of traditional methods, since most of the traditional methods [1, 35] search for similar patches from the unmasked region to synthesize the missing parts. Moreover, face images often contain large appearance variations (*e.g.*, expressions and different viewpoints) and the facial components are tightly correlated to each other. It not only needs to guarantee the low-level pixel consistency of color and texture but also requires to satisfy attribute consistency between the inpainted part and the rest observed part.

Over the past few years, deep learning (DL)-based inpainting methods have been developed in succession and achieve significant improvement [15, 28, 29]. However, these methods mainly focus on face completion with regular holes [10, 20, 34], which usually locates at the center of images. It may cause the completion network overfitting to regular holes and become a barrier for applying in real-world cases, since irregular holes are more common in practical application, for example, the cracks on old photos. Although some methods adopt different types of predicted information as prior knowledge to assist the face completion, such as edges [20], facial landmarks [29], and semantic maps [17], these prior knowledge is either low-level features which are vulnerable to noise or high-level features which are lack of details. However, facial attributes providing an abstraction between the low-level features and the high-level labels, which can be regarded as mid-level representations and have been successfully used for face recognition [5, 33]. Using attributes as the prior knowledge can benefit the face completion task (see Figure 1).

To address the above issues, we propose a novel facial attributes guided face completion network named **AttrFaceNet**. The proposed AttrFaceNet is composed of two subnets. One is for facial attribute prediction while the other one is for face completion. The attribute prediction subnet learns to output 38 predicted facial attributes (*e.g.*, male, smiling, eyeglasses, *etc.*) from the corrupted image. For better prediction, these 38 attributes are firstly divided into nine groups according to their characteristics or locations on the face and then are predicted by nine branches respectively. Moreover, due to intrinsic relations between these attributes, *e.g.*, "Male" is inherently related to "No Beard", a fully connected layer is added at the end of this subnet to collaboratively fuse the nine group attributes. The face completion subnet is a U-Net [23] architecture, which firstly downsamples the feature maps and then upsamples them to the original size. During the upsampling process, the former predicted attributes are also upsampled through deconvolutional layers and then injected as guidance into the face completion subnet at every upsampling scale.

The main contributions of this work can be summarized as follows:

- To the best of our knowledge, the proposed AttrFaceNet is the first attempt that utilizes predicted facial attributes as guidance to explicitly assist face completion in an end-to-end way. With these distinct attributes, both color and structure discrepancy can be well relieved.
- An attribute prediction subnet is employed to predict facial attributes from corrupted images directly while the intrinsic relations between attributes are considered.
- To complete face images, the predicted attributes are merged into face completion subnet at multiple feature scales, so that to provide comprehensive guidance.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related works. Section 3 introduces the proposed AttrFaceNet in detail. Section 4 shows the extensive experimental results and ablation analysis to evaluate our model. Section 5 concludes the paper and introduces our future work.

2 RELATED WORK

2.1 Face inpainting

Some typical DL-based inpainting methods involve training deep CNNs combined with generative adversarial learning [7] to predict each pixel of the missing area. Context Encoders proposed by Pathak *et al.* [22] is the pioneer attempt to apply CNN with generative adversarial learning to entire images. However, without enough attention to the local details, the recovered regions of output images often contain blurriness. Thereafter, local discriminator proposed by Iizuka *et al.* [11] aims to better generate more realistic details. PConv proposed by Liu *et al.* [18] is specially designed for filling irregular holes, which only adopts valid pixels in each convolution to alleviate artifacts by updating a binary mask according to handcrafted rule. Xie *et al.* [27] improved PConv and proposed learnable bidirectional attention maps to force network focus on filling irregular holes. In addition, gated convolution designed by Yu *et al.* [30] learns soft mask automatically from data rather than updating based on handcrafted rules to implement free-form image inpainting.

Compared with general images, face completion is a much more challenging task owing to the diverse appearance and complex structure of the human face. Li *et al.* [15] firstly applied a deep generative adversarial network (GAN) for face completion task which consists of an encoder-decoder generator and two discriminators (global and local) to generate missing content. Afterward, Nazeri *et al.* [20] utilized hallucinated edges as prior knowledge to guide inpainting while Yang *et al.* [29] employed predicted facial landmarks. However, when the irregular holes become larger, this prior knowledge is prone to be mispredicted thus to degrade inpainting performance. Zhou *et al.* [34] proposed a dual spatial attention (DSA) module with oracle supervision for face completion. Li *et al.* [16] elaborately designed DF-GAN for completing face images. However, they focus on filling either rectangular holes or structured occlusions, which may cause performance degradation when meeting irregular holes.

2.2 Facial attribute recognition

The DL-based methods for facial attribute recognition boomed shortly after CelebA dataset [19] was released. Liu *et al.* [19] proposed a method that combines two networks which respectively work for locating face region and extracting attribute features to collaboratively recognize facial attributes. However, due to the employment of SVM [3] as classifiers at the last procedure, the whole pipeline is not in an end-to-end manner. MOON proposed by Rudd *et al.* [24] is the first to recognize multiple attributes via a single network. However, it involves tremendous parameters. MCNN with fewer parameters proposed by Hand *et al.* [8] took

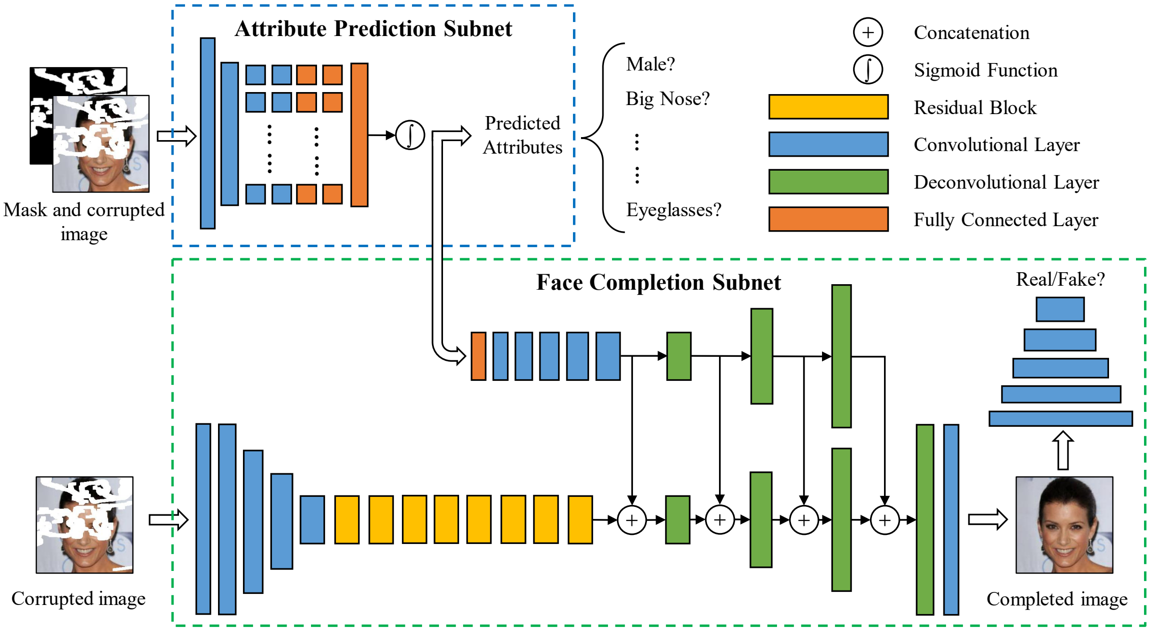


Figure 2: Overview of our proposed AttrFaceNet, which involves an attribute prediction subnet and a face completion subnet. The attribute prediction subnet is responsible for predicting facial attributes from corrupted images. The face completion subnet is responsible for generating visually realistic content and filling the missing part with predicted attributes as guidance.

implicit and explicit relationships between facial attributes into consideration and an auxiliary network was added at the tail of the trained MCNN to learn score-level attribute correlation. Besides, MCNN inspired us to build our attribute prediction subnet. After that, the performance improvement of facial attribute prediction has entered a plateau.

3 PROPOSED ATTRFACENET

The framework of our proposed AttrFaceNet is illustrated in Figure 2, which consists of two subnets: attribute prediction subnet and face completion subnet. More specifically, the attribute prediction subnet fulfills the task of predicting facial attributes from corrupted images, while the face completion subnet utilizes the predicted attributes as guidance to generate missing content. Let A represent the attribute prediction subnet, while G and D represent the generator and discriminator of the face completion subnet, respectively.

3.1 Attribute prediction subnet

Before providing the network details, the used denotations are introduced. Suppose that I_m denotes the corrupted image, M denotes the mask (i.e., the hole area), A_{out} denotes the predicted attributes and A_{gt} denotes the ground truth attributes. M is a binary mask with 1 labels missing region and 0 labels background. Both A_{out} and A_{gt} are vectors with 38 elements. More specifically, $A_{out} = \{A^1_{out}, A^2_{out}, ..., A^{38}_{out}\}$ and $A^i_{out} \in [0, 1]$, which means the predicted attributes are with the confidence values between 0 and 1, while $A_{gt} = \{A^1_{gt}, A^2_{gt}, ..., A^{38}_{gt}\}$ and $A^i_{gt} \in \{0, 1\}$, which means A_{gt} is a boolean value. 0 represents the ground truth attribute is not contained in the face and 1 represents the attribute is definitely contained in the face. The attribute prediction network A takes I_m together with M as the input and outputs A_{out}, which can be formulated as

$$A_{out} = A(I_m, M). \tag{1}$$

Referring to the upper half of Figure 2, I_m and M are concatenated as the subnet input and go through two convolutional layers to share the low-level information. After that, considering different types of facial attributes have an impact on different locations of the human face, the 38 attributes are divided into nine groups [8]: **Gender**, **Nose**, **Mouth**, **Eyes**, **Face**, **AroundHead**, **FacialHair**, **Cheeks** and **Fat**. Hence, each group has at least one attribute, *e.g.*, the **Nose** group has two attributes “Big Nose” and “Pointy Nose”. Correspondingly, the middle part of the attribute prediction subnet is split into nine branches to separately predict the facial attributes. Specifically, every branch contains two convolutional layers and two fully connected layers to convert feature maps into vectors. As analyzed before, facial attributes are tightly correlated with each other, thus not only the explicit relationships between facial attributes should be taken into consideration, but also the implicit ones. Therefore, a fully connected layer is added immediately after the nine branches to fuse the information from different groups. Finally, a sigmoid function is employed at the tail of the subnet so that to guarantee $A^i_{out} \in [0, 1]$. It is worth noting that A_{out} is not binarized to hard label (either 0 or 1) for assigning different predictive confidence to different attributes.

For attribute prediction subnet, we choose binary cross-entropy as our loss function $\mathcal{L}_A$, which is defined as

$$\mathcal{L}_A = -\frac{1}{N_A} \sum_i A^i_{gt} log A^i_{out} + (1 - A^i_{gt}) log(1 - A^i_{out}), \tag{2}$$

where N_A is the number of facial attributes.

3.2 Face completion subnet

Let I_{out} denote the final completed image and I_{gt} denote the ground truth image. The face completion subnet adopts generative adversarial learning, where G learns to generate visually realistic I_{out} and D learns to discriminate I_{out} and I_{gt}. The face completion subnet takes I_m and A_{out} as input and returns I_{out}, which can be written as

$$I_{out} = G(I_m, A_{out}) \odot M + I_m \odot (1 - M), \tag{3}$$

where $\odot$ is element-wise multiplication.

As illustrated in the bottom half of Figure 2, our generator G is within an encoder-decoder architecture with residual blocks in the middle part. The encoder continuously downsamples the input four times to learn latent representations, which are then passed through eight residual blocks. Thereafter, the decoder undertakes the function of upsampling by deconvolutional layers to the original resolution and completing the human face from the learned latent representations. Meanwhile, A_{out} is converted to 2D feature maps through a fully connected layer and five convolutional layers to match dimensions for later concatenation. These 2D feature maps are then upsampled by deconvolutional layers similarly and injected into the decoding process at every feature scale to assist the generation of the final output I_{out}. The discriminator D is a 5-layer CNN which takes I_{out} and I_{gt} as the input and discriminates the input face image as real or fake. The output of D is the probability of the input face image being a real face image.

The face completion subnet is trained with joint loss constraints, which consist of ℓ_1 loss, adversarial loss [7], perceptual loss [12] and style loss [6]. The ℓ_1 loss, also known as pixel reconstruction loss, is a commonly used loss to reduce pixel-level differences and formulated as

$$\mathcal{L}_{\ell_1} = ||I_{out} - I_{gt}||_1. \tag{4}$$

The adversarial loss [7] assures the generated face be visually realistic, which is defined as

$$\mathcal{L}_{adv} = \mathbb{E}[logD(I_{gt})] + \mathbb{E}[log(1 - D(I_{out}))]. \tag{5}$$

The perceptual loss [12] is employed to reduce the perceptual differences between I_{out} and I_{gt} and can be computed by

$$\mathcal{L}_{perc} = \frac{1}{N}\sum_i ||\phi_i(I_{out}) - \phi_i(I_{gt})||_1, \tag{6}$$

where ϕ_i represents the feature map of i-th activation layer of a pretrained network, which is the activation map of the `relu1_1`, `relu2_1`, `relu3_1`, `relu4_1` and `relu5_1` layer of the pretrained VGG-19 [25] network in our implementation. The style loss [6] is also computed from the corresponding feature maps utilized in perceptual loss, and the formulation is

$$\mathcal{L}_{style} = \frac{1}{N}\sum_i ||G_i(I_{out}) - G_i(I_{gt})||_1, \tag{7}$$

where G_i denotes the Gram matrix constructed from ϕ_i. The total loss for face completion subnet $\mathcal{L}_I$ is the weighted summation of the above losses, which can be written as

$$\mathcal{L}_I = \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style}. \tag{8}$$

In our experiments, we set $\lambda_{\ell_1} = 1$, $\lambda_{adv} = 0.1$, $\lambda_{perc} = 0.1$ and $\lambda_{style} = 250$ empirically.

4 EXPERIMENTS

4.1 Implementation details

Our AttrFaceNet is trained on CelebA dataset [19], which contains more than 200,000 face images and each face is annotated with 40 binary attributes. It is worthy of note that we crop and resize each image to 256×256 as pre-processing, thus "Necklace" and "Necktie" attributes are abandoned and only 38 attributes are retained. The irregular masks obtained from [18] are distributed in six different intervals and augmented by four rotations and a horizontal reflection. We follow the official CelebA training set together with 60% augmented irregular masks as our training data (162,770 images) and the official CelebA validation set together with 20% augmented irregular masks as our validation data (19,867 images). For testing, to ensure the one-to-one correspondence between images and masks, the rest 20% augmented irregular masks with the same number of images selected from the official CelebA test set are combined as our test set (19,200 images). Moreover, 2,060 images with no more than two faces per image selected from Helen dataset [14] are also employed as our test set. We implement the proposed AttrFaceNet based on PyTorch [21] and the network parameters are optimized by Adam [13] with a learning rate of 1×10^{-4}. Furthermore, the attribute prediction subnet and the face completion subnet are separately trained to convergence at first with 30 and 40 epochs, respectively, and then jointly trained in an end-to-end manner with 30 epochs so as to fully enhance the performance of both subnets in attributes prediction and face completion.

4.2 Comparisons with state-of-the-arts

Baselines We compare the face completion performance of our proposed AttrFaceNet with five state-of-the-art baseline methods in both quantitative and qualitative aspects:

- PConv [18]: a generative method where the convolutional layers are specially designed for filling irregular holes.
- PEN-Net [31]: a deep generative model with U-Net architecture comprises of a pyramid-context encoder and a multi-scale decoder.
- PICNet [32]: a probabilistically principled network with a reconstructive path and a generative path to generate diverse plausible results.
- LaFIn [29]: a method which employing predicted facial landmarks as prior to complete the human faces.
- EC [20]: a generative model which generate the edge maps first and then complete corrupted image with the help of hallucinated edges.

These above inpainting methods either focus on face completion or irregular holes, thus be selected as the benchmark methods. Moreover, for fair comparisons, these models are retrained and tested on the same training and test datasets as we used based on their official implementation except for PConv, whose official source codes are not available yet and we implement it according to their paper.

Table 1: Quantitative comparisons of our AttrFaceNet with state-of-the-art methods on CelebA and Helen datasets. Noting that the FID results on Helen dataset are not provided since the number of images is not sufficient enough to compute an accurate FID. The best and second best results are marked in red and blue, respectively.

Metric	Mask	CelebA						Helen					
		PConv	PEN-Net	PICNet	LaFIn	EC	Ours	PConv	PEN-Net	PICNet	LaFIn	EC	Ours
ℓ_1 loss	0-10%	0.77	1.25	1.20	0.90	0.53	0.59	0.94	1.47	1.34	1.08	0.63	0.69
	10-20%	1.87	2.46	2.30	1.83	1.37	1.44	2.44	3.05	2.73	2.23	1.72	1.76
	20-30%	3.18	4.19	3.91	3.06	2.49	2.56	4.11	5.12	4.61	3.77	3.09	3.14
	30-40%	4.60	6.26	6.00	4.51	3.80	3.80	6.24	8.14	7.70	5.76	5.10	5.13
	40-50%	6.29	8.80	8.60	6.26	5.43	5.33	8.62	11.30	10.89	8.09	7.74	7.39
	50-60%	9.52	13.15	14.27	9.59	8.89	8.24	13.93	17.15	18.90	12.80	12.58	12.44
PSNR	0-10%	38.10	36.58	37.06	39.04	41.13	40.79	36.41	34.47	35.59	36.51	38.96	38.39
	10-20%	32.83	30.83	31.54	33.25	34.84	34.79	29.90	28.04	28.61	30.14	31.73	31.62
	20-30%	29.60	27.27	27.92	29.85	31.03	31.13	26.96	24.99	25.43	27.20	28.42	28.27
	30-40%	27.33	24.79	25.13	27.37	28.37	28.63	24.25	22.20	22.32	24.53	25.24	25.19
	40-50%	25.35	22.70	22.89	25.38	26.10	26.48	22.28	20.35	20.40	22.55	22.88	23.06
	50-60%	22.35	20.10	19.50	22.57	22.74	23.46	18.95	17.71	16.95	19.66	19.74	19.74
SSIM	0-10%	0.9917	0.9916	0.9919	0.9936	0.9961	0.9956	0.9895	0.9886	0.9901	0.9918	0.9945	0.9936
	10-20%	0.9780	0.9750	0.9768	0.9826	0.9877	0.9876	0.9759	0.9684	0.9734	0.9792	0.9833	0.9834
	20-30%	0.9592	0.9468	0.9506	0.9658	0.9733	0.9742	0.9448	0.9324	0.9400	0.9553	0.9623	0.9617
	30-40%	0.9373	0.9121	0.9160	0.9453	0.9551	0.9582	0.9078	0.8774	0.8854	0.9233	0.9301	0.9279
	40-50%	0.9104	0.8677	0.8728	0.9158	0.9312	0.9373	0.8621	0.8204	0.8278	0.8841	0.8820	0.8876
	50-60%	0.8431	0.7773	0.7663	0.8606	0.8630	0.8853	0.7421	0.6856	0.6699	0.7944	0.7717	0.7716
FID	0-10%	3.17	3.30	3.22	3.02	3.03	3.04	–	–	–	–	–	–
	10-20%	3.64	4.91	3.89	3.15	3.08	3.06	–	–	–	–	–	–
	20-30%	4.52	9.38	5.57	3.61	3.41	3.38	–	–	–	–	–	–
	30-40%	5.44	17.41	8.67	4.26	3.95	3.78	–	–	–	–	–	–
	40-50%	6.65	29.52	13.53	5.14	4.51	4.31	–	–	–	–	–	–
	50-60%	8.42	42.14	20.91	7.08	6.50	5.76	–	–	–	–	–	–

4.2.1 *Quantitative results.*

For quantitative comparison, we choose ℓ_1 loss, PSNR, SSIM [26], and FID [9] as the evaluation metrics, where PSNR and SSIM are the higher the better while ℓ_1 loss and FID are the lower the better. Table 1 reports the results of various methods for different sizes of masks on CelebA and Helen datasets. Noting that the FID results on Helen dataset are not provided since the number of images is not sufficient enough to compute an accurate FID.

It can be seen from Table 1 that our AttrFaceNet outperforms the state-of-the-art methods and the margin becomes even larger in large mask cases on CelebA. PEN-Net and PICNet suffer from low performance since both of them mainly concentrate on filling rectangular holes, they are not capable of handling irregular holes even after retraining. Compared with PEN-NET and PICNet, the maximum ℓ_1 loss/PSNR(dB)/SSIM/FID improvement of our method can be 4.91/3.36/0.1080/36.38 and 6.03/3.96/0.1190/15.15 both at 50-60% mask size. PConv is proposed especially for irregular holes, thus the performance is slightly better. Nevertheless, it is difficult for PConv to complete human faces with pleasant results. When comparing with PConv, the maximum ℓ_1 loss/PSNR(dB)/SSIM/FID improvement of our method is 1.28/1.11/0.0422/2.66 at 50-60% mask size. LaFIn and EC obtain similar results with our AttrFaceNet when the mask size is relatively small. However, when the mask size gets larger, the predicted facial attributes bring much more benefit for face completion than facial landmarks and edge maps.

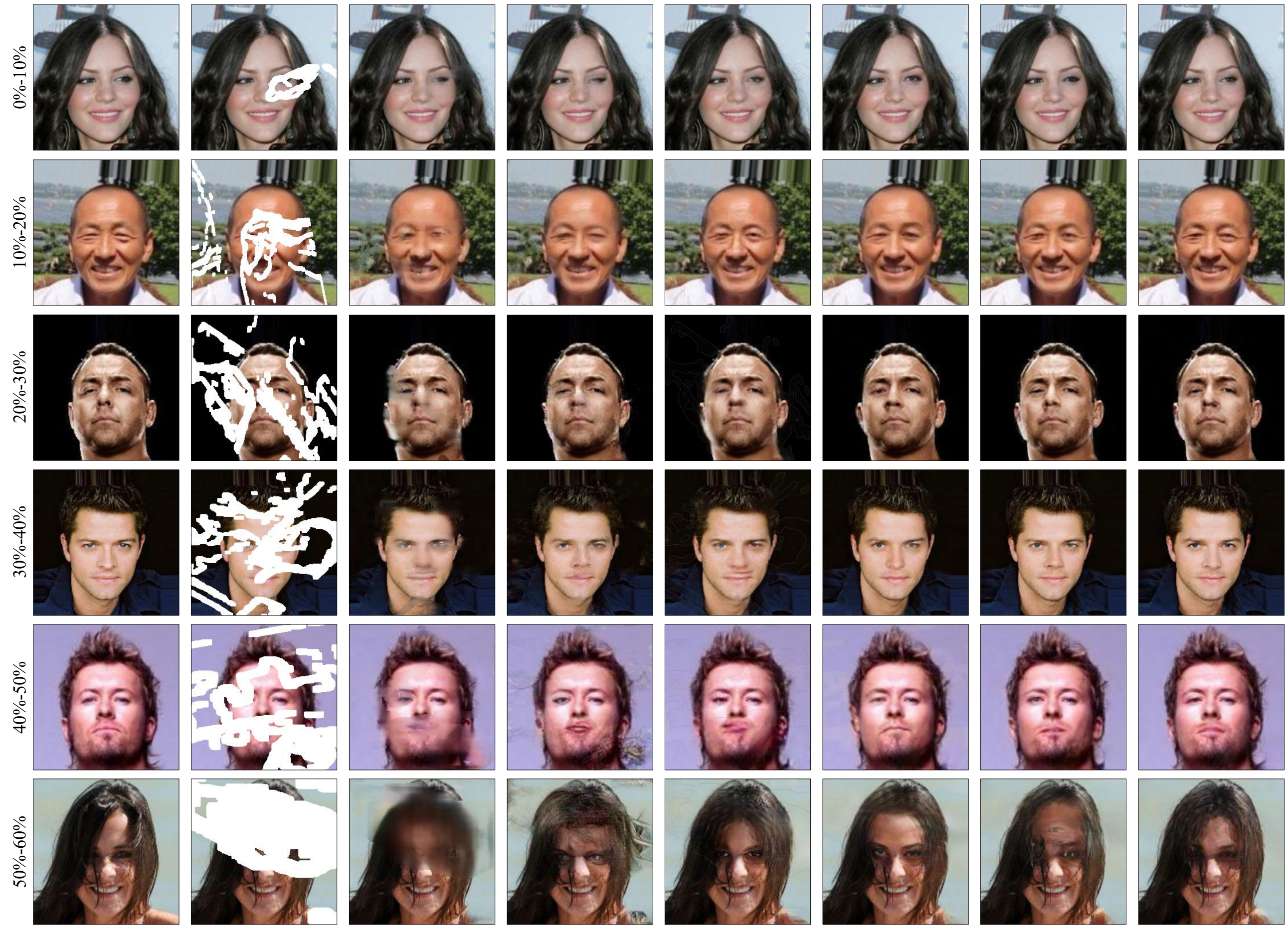
Similar results can be found on the Helen dataset in Table 1, and the proposed AttrFaceNet achieves favorable results compared with other methods. Since Helen dataset contains many cases of non-close-up form and multiple faces per images, which makes edge prediction easier than attribute prediction, EC obtains slightly better performance than ours by employing predicted edge information of the missing region as the prior knowledge.

4.2.2 *Qualitative results.*

We select one image from each mask size interval and the qualitative comparisons are illustrated in Figure 3. It can be seen that the visual results of AttrFaceNet contain plausible content with finer details. By comparison, PEN-Net and PICNet tend to generate blurry content, which significantly degrades the quality of completed images. PConv is able to handle the blurring problem, however, it suffers from obvious artifacts on the mask edges which makes the generated part and the rest part are not well fused. For LaFIn and EC, visually realistic results can be obtained when the mask size is smaller than 50% of the image size. However, when taking a closer look at the generated region, especially the eyes part, the left and the right eyes are not coherent, e.g., the eyes in the image selected from mask size 40%-50%, which further demonstrates that employing facial landmarks and edges as prior knowledge will not be effective enough in keeping attribute consistency of human faces. Paying attention to the last row in Figure 3, it can be found that LaFIn and EC fail to complete face since it is hard to predict reasonable landmarks and edge maps for guidance when the mask size is large.

4.3 Ablation study

To further evaluate our AttrFaceNet, we conduct ablation analysis on the effectiveness of predicted attributes, the robustness of



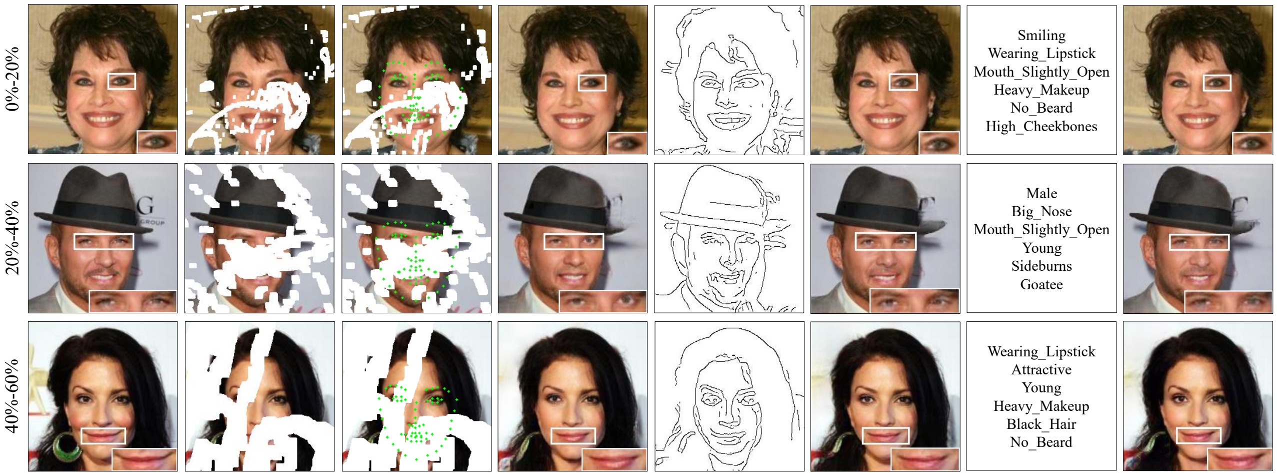
(a) GT (b) Input (c) PEN-Net (d) PICNet (e) PConv (f) LaFIn (g) EC (h) Ours

Figure 3: Qualitative comparisons of our AttrFaceNet with state-of-the-art methods. The images in each row from left to right are (a) ground truth image, (b) corrupted image, (c) image completed by PEN-Net [31], (d) image completed by PICNet [32], (e) image completed by PConv [18], (f) image completed by LaFIn [29], (g) image completed by EC [20] and (h) image completed by our AttrFaceNet, respectively. Best viewed with zoom-in.

predicted attributes, the effect of reversing attributes and the effectiveness of joint training. All these ablation experiments are conducted on the same training and testing datasets.

4.3.1 Effectiveness of predicted attributes. To verify the effectiveness of the attribute prediction subnet in the face completion network, we compare the inpainting performance of the networks without attributes (w/o Attr), with random attributes (w RAttr), with predicted attributes (w PAttr), with ground truth attributes (w GTAttr) and with predicted attributes combined with joint training strategy (Ours). In the comparison of different types of attributes, the attribute prediction subnet is trained first and then is fixed in the training and testing of the face completion subnet. The results are listed in Table 2. Comparing to the results of "w/o Attr", "w RAttr" achieves worse performance due to the negative effects caused by the randomly predicted attributes, and "w PAttr" obtains better performance due to the positive effects caused by the accurate predicted attributes. This demonstrates that the attribute information really affects the performance of face completion. Although the face completion network guided by ground truth attributes ("w GTAttr") obtains slightly better results than that of the face completion network guided by predicted attributes ("w PAttr"), the gap can be neglected and can be compensated by the benefits of joint training strategy ("Ours"). Since the ground truth attributes are hard constraints, where the value of each attribute is either 1 or 0, while the value of the predicted attributes is soft constraints and is between 0 and 1.

4.3.2 Robustness of predicted attributes. To analyze the robustness of predicted attributes, we conduct the comparison between different predicted information which is used as the prior knowledge in the face completion network. The selected predicted information is facial landmarks from LaFIn and the edge maps from EC. The corresponding inpainting results under different mask sizes are shown in Figure 4. LaFIn fails to predict correct eye landmarks on the first and second images. Therefore, it easily generates the left and right eyes with different sizes and shapes, which makes the completed face images implausible. Moreover, EC tends to generate



(a) GT (b) Input (c) Landmarks (d) LaFIn (e) Edges (f) EC (g) Attributes (h) Ours

Figure 4: Comparisons of different predicted information and their corresponding inpainting results. The images in each row from left to right are (a) ground truth image, (b) corrupted image, (c) predicted facial landmarks, (d) image completed by LaFIn [29], (e) predicted edges, (f) image completed by EC [20], (g) predicted attributes and (h) image completed by our AttrFaceNet, respectively.

(a) GT (b) Input (c) GTAttr (d) Reversed

Figure 5: Visual effects of reversing specific attribute. The images in each row from left to right are (a) ground truth image, (b) corrupted image, (c) completed images guided by ground truth attributes and (d) completed images guided by reversed attributes, respectively.

discontinuous edges in the mask area, which results in insufficient guidance for the face completion task. While assisted by the accurately predicted attributes, our AttrFaceNet is able to recover realistic details even with large masks (see the curvature of the mouth in the last row of Figure 4).

Table 2: Ablation analysis of the effectiveness of predicted attributes. The best and second best results are marked in red and blue, respectively.

Metric	Mask	CelebA				
		w/o Attr	w RAttr	w PAttr	w GTAttr	Ours
ℓ_1 loss	0-10%	0.60	0.63	0.62	0.61	0.59
	10-20%	1.51	1.54	1.50	1.49	1.44
	20-30%	2.68	2.74	2.68	2.65	2.56
	30-40%	3.97	4.09	3.99	3.95	3.80
	40-50%	5.57	5.74	5.59	5.53	5.33
	50-60%	8.60	8.81	8.64	8.54	8.24
PSNR	0-10%	40.37	40.24	40.42	40.53	40.79
	10-20%	34.41	34.25	34.40	34.53	34.79
	20-30%	30.78	30.65	30.74	30.88	31.13
	30-40%	28.30	28.16	28.24	28.36	28.63
	40-50%	26.16	26.04	26.12	26.23	26.48
	50-60%	23.16	23.08	23.12	23.22	23.46
SSIM	0-10%	0.9955	0.9952	0.9953	0.9954	0.9956
	10-20%	0.9866	0.9862	0.9865	0.9869	0.9876
	20-30%	0.9726	0.9716	0.9720	0.9728	0.9742
	30-40%	0.9555	0.9540	0.9546	0.9558	0.9582
	40-50%	0.9339	0.9317	0.9324	0.9345	0.9373
	50-60%	0.8801	0.8768	0.8774	0.8810	0.8853
FID	0-10%	3.04	3.04	3.03	3.05	3.04
	10-20%	3.09	3.08	3.09	3.07	3.06
	20-30%	3.42	3.46	3.43	3.42	3.38
	30-40%	3.92	3.95	3.89	3.91	3.78
	40-50%	4.55	4.70	4.54	4.48	4.31
	50-60%	6.21	6.34	6.19	6.02	5.76

4.3.3 Effect of reversing attributes. To figure out how the attributes affect the final results, we randomly reverse different numbers of GT attributes from A^i_{gt} to $1-A^i_{gt}$ while keeping the input image and mask the same. It can be seen from Table 3 that more attributes were reversed, worse PSNR results were obtained. Meanwhile, reversing attributes also results in significant changes on the corresponding region of the completed faces but keeps other parts unchanged. When the "Mouth Slightly Open" attribute is reversed, only the mouth region changes accordingly (see the first row of Figure 5). Similarly, a pair of eyeglasses appears in the eye region when the "Eyeglasses" attribute is reversed (see the second row of Figure 5). Both the quantitative and visual results strongly prove the importance of attribute guidance for face completion.

4.3.4 Effectiveness of joint training. The proposed AttrFaceNet achieves better performance when larger area is masked, which is shown in Table 1. Better results obtained in harder cases not only

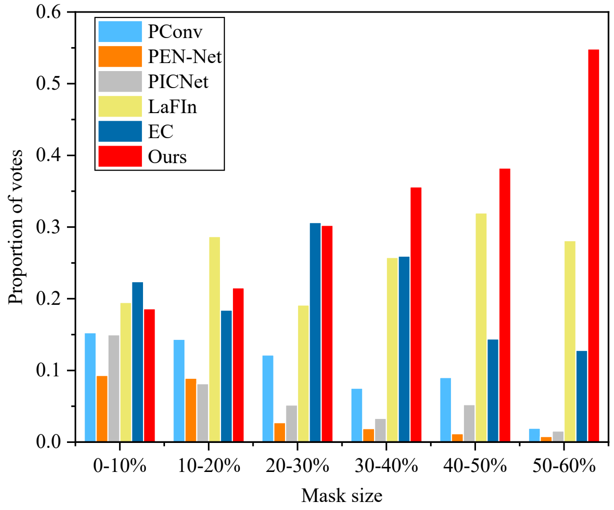
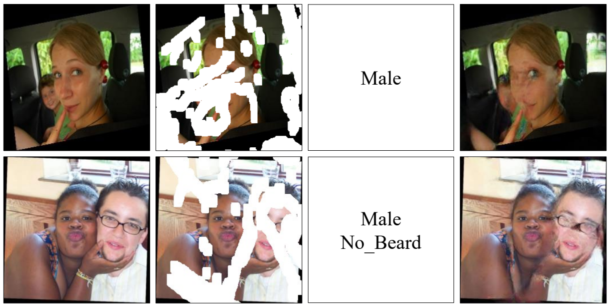


Figure 6: User study of AttrFaceNet.



(a) GT (b) Input (c) Attributes (d) Ours

Figure 7: Failure cases of our AttrFaceNet on Helen dataset. The images in each row from left to right are (a) ground truth image, (b) corrupted image, (c) predicted attributes and (d) image completed by our AttrFaceNet, respectively.

benefit from the robustness of predicted attributes, but also the joint training strategy. When the face images are severely corrupted by a large mask, it is almost impossible to predict all 38 attributes correctly. However, the joint training strategy makes the two subnets coordinate and help each other thus to alleviate the difficulties. More specifically, the attribute prediction subnet is not independent since it also receives information from the face completion subnet in the joint training stage. Table 4 demonstrates the accuracy of attribute prediction subnet at different epoch numbers especially in larger mask cases, and the results validate that the joint training strategy truly helps the attribute prediction.

4.4 User study

To better evaluate the results of our AttrFaceNet, we conducted an experiment where participants are asked to vote for the most realistic image from six options generated by different methods. We randomly select 50 faces for each model at different mask size and a total of 300 images are shown to 86 participants. The proportions of votes for different methods at different mask sizes are illustrated in Figure 6. In small mask cases, six methods achieve similar proportions. With the mask size increased, the superiority of AttrFaceNet is gradually emerging. When the face image is severely corrupted, our AttrFaceNet outperforms the other five methods with large margins.

Table 3: The PSNR(dB) results of different numbers of GT attributes reversed on CelebA dataset.

Metric	Mask	Number of Reversed Attributes				
		0	8	16	32	38
PSNR(dB)	0-10%	40.53	40.48	40.44	40.23	40.04
	10-20%	34.53	34.45	34.39	34.16	33.98
	20-30%	30.88	30.79	30.72	30.49	30.33
	30-40%	28.36	28.27	28.19	27.97	27.82
	40-50%	26.23	26.13	26.06	25.84	25.71
	50-60%	23.22	23.09	23.00	22.81	22.70

Table 4: The accuracy of attribute prediction subnet at different epochs during the joint training.

Mask	Epoch Number of Joint Training			
	0	10	20	30
30-40%	87.81%	87.92%	88.11%	88.22%
40-50%	87.78%	87.82%	88.09%	88.16%
50-60%	87.00%	87.11%	87.23%	87.26%

4.5 Failure cases

Some typical failure cases are shown in Figure 7. In our method, the facial attribute labels are provided by CelebA dataset, where most of the face images are presented in a close-up way and the face occupies a large area of the whole image. However, the face only occupies a small area of the whole image in Helen dataset, which causes the scale differences of facial attributes. Therefore, when the image contains small face area, the facial attributes of the corrupted image are hardly to be predicted accurately (see the first image in Figure 7). In addition, there are several images with multiple human faces. When completing one face, the other ones may bring negative effects on facial attribute prediction and cause the missing prediction of some attributes. For example, for the second image in Figure 7, the facial attribute "Eyeglasses" is missing.

5 CONCLUSION

In this paper, we propose a novel deep network named AttrFaceNet to complete face images with predicted attributes as guidance. The proposed AttrFaceNet consists of an attribute prediction subnet and a face completion subnet, where the former one predicts 38 facial attributes with confidence values from corrupted images and the later one generates the missing content with aid of predicted attributes. Extensive experiments verify the superiority of our method over state-of-the-art methods that our AttrFaceNet can well complete missing content with plausible textures. In our future work, we would like to solve the face size and multiple faces caused problems in the failure cases.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant No. 62071500, 61701313; Shenzhen Science and Technology Program under Grant No. 2021A26.

REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ToG* 28, 3 (2009), 24.

[2] Rong-Chi Chang, Yun-Long Sie, Su-Mei Chou, and Timothy K Shih. 2005. Photo defect detection for image inpainting. In *ISM*. 5–pp.

[3] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[4] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. 2004. Region filling and object removal by exemplar-based image inpainting. *TIP* 13, 9 (2004), 1200–1212.

[5] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. 2012. Discovering localized attributes for fine-grained recognition. In *CVPR*. 3474–3481.

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *CVPR*. 2414–2423.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*. 2672–2680.

[8] Emily M Hand and Rama Chellappa. 2017. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI*. 4068–4074.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*. 6626–6637.

[10] Yibing Song Wei Huang Hongyu Liu, Bin Jiang and Chao Yang. 2020. Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations. In *ECCV*.

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ToG* 36, 4 (2017), 1–14.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*. 694–711.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. 2012. Interactive facial feature localization. In *ECCV*. 679–692.

[15] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017. Generative face completion. In *CVPR*. 3911–3919.

[16] Zhihang Li, Yibo Hu, Ran He, and Zhenan Sun. 2020. Learning disentangling and fusing networks for face completion under structured occlusions. *PR* 99 (2020), 107073.

[17] Haofu Liao, Gareth Funka-Lea, Yefeng Zheng, Jiebo Luo, and S Kevin Zhou. 2018. Face completion with semantic knowledge and collaborative adversarial learning. In *ACCV*. 382–397.

[18] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *ECCV*. 85–100.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.

[20] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. 2019. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *ICCVW*.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 8026–8037.

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *CVPR*. 2536–2544.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. 234–241.

[24] Ethan M Rudd, Manuel Günther, and Terrance E Boult. 2016. Moon: A mixed objective optimization network for the recognition of facial attributes. In *ECCV*. 19–35.

[25] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *TIP* 13, 4 (2004), 600–612.

[27] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. 2019. Image inpainting with learnable bidirectional attention maps. In *ICCV*. 8858–8867.

[28] Jie Yang, Zhiquan Qi, and Yong Shi. 2020. Learning to Incorporate Structure Knowledge for Image Inpainting.. In *AAAI*. 12605–12612.

[29] Yang Yang, Xiaojie Guo, Jiayi Ma, Lin Ma, and Haibin Ling. 2019. LaFIn: Generative Landmark Guided Face Inpainting. *arXiv preprint arXiv:1911.11394* (2019).

[30] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *ICCV*. 4471–4480.

[31] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. 2019. Learning pyramid-context encoder network for high-quality image inpainting. In *CVPR*. 1486–1494.

[32] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic Image Completion. In *CVPR*. 1438–1447.

[33] Jingjing Zheng, Zhuolin Jiang, Rama Chellappa, and Jonathon P Phillips. 2014. Submodular attribute selection for action recognition in video. In *NeurIPS*. 1341–1349.

[34] Tong Zhou, Changxing Ding, Shaowen Lin, Xinchao Wang, and Dacheng Tao. 2020. Learning Oracle Attention for High-fidelity Face Completion. In *CVPR*. 7680–7689.

[35] Daniel Zoran and Yair Weiss. 2011. From learning models of natural image patches to whole image restoration. In *ICCV*. 479–486.