Face inpainting via Learnable Structure Knowledge of Fusion Network

You Yang[1,2], Sixun Liu[2*], Bin Xing[3] and Kesen Li[2]
[1] National Center for Applied Mathematics in Chongqing. Chongqing 401331, China.
[e-mail: 565357950@qq.com]
[2] School of Computer and Information Science, Chongqing Normal University. Chongqing 401331, China.
[e-mail: 1224371511@qq.com]
[3] National Engineering Laboratory of Industrial Big-Data Application Technology. Beijing 100041, China
[e-mail: Xing.bin@hotmail.com]
[*]Corresponding author:Sixun Liu

Received April 9, 2021; revised September 10, 2021; revised November 4, 2021; revised Janaury 3, 2022; accepted February 22, 2022; published March 31, 2022

Abstract

With the development of deep learning, face inpainting has been significantly enhanced in the past few years. Although image inpainting framework integrated with generative adversarial network or attention mechanism enhanced the semantic understanding among facial components, the issues of reconstruction on corrupted regions are still worthy to explore, such as blurred edge structure, excessive smoothness, unreasonable semantic understanding and visual artifacts, etc. To address these issues, we propose a Learnable Structure Knowledge of Fusion Network (LSK-FNet), which learns a prior knowledge by edge generation network for image inpainting. The architecture involves two steps: Firstly, structure information obtained by edge generation network is used as the prior knowledge for face inpainting network. Secondly, both the generated prior knowledge and the incomplete image are fed into the face inpainting network together to get the fusion information. To improve the accuracy of inpainting, both of gated convolution and region normalization are applied in our proposed model. We evaluate our LSK-FNet qualitatively and quantitatively on the CelebA-HQ dataset. The experimental results demonstrate that the edge structure and details of facial images can be improved by using LSK-FNet. Our model surpasses the compared models on L1, PSNR and SSIM metrics. When the masked region is less than 20%, L1 loss reduce by more than 4.3%.

Keywords: Face Inpainting, Image Edge, Gated Convolution, Region Normalization，Prior Knowledge.

This work is supported partially by the Chongqing graduate scientific research and innovation project (Grant No. YKC20038), the 13th five-year plan of Chongqing education science planning (Grant No. 2019-GX-10), and the teaching reform project of Chongqing society of higher education (Grant No. CQGJ19B22).

http://doi.org/10.3837/tiis.2022.03.007 **ISSN : 1976-7277**

1. Introduction

Image inpainting is a task which aim to recover the missing region of an image with the known information of the uncorrupted region or other information. Its goal includes mainly three holdings: reconstructed image reasonable, visual continuity and scene consistency. As a branch of image inpainting, face inpainting pays more attention to the rationality of semantics and structure, which is even more challenging. Different from other branches of image inpainting, the face structure pattern contains rich semantic representation such as eyes, mouth, nose, etc. Both low and intermediate-level visual features of the known region are not enough to infer the missing part of valid semantic features [1, 2], so it is unable to model the face geometric structure.

In the field of computer vision, there are two types of image inpainting methods: traditional and deep learning-based methods. Traditional image inpainting methods have some limitations which the model can only obtain low-level pixel features, so it can't capture the high-level semantics [3-9]. When filling missed facial details and complex scenes, it would lead to serious failure. The deep learning methods can carry out meaningful learning from the dataset through the network, and reconstruct the corrupted image in an end-to-end way. This method can fill the missing region semantically.

Some deep learning-based methods of face inpainting fail to separate the structure and texture information effectively, which will result in fuzzy edge structure and excessive smoothness. In order to solve these problems, the existing deep learning methods introduce prior knowledge to model the corrupted face. GFC is the first to use the prior knowledge of facial structure as loss constraint to assist reconstruction on missing regions, which can generate more reasonable missing part [10]. However, it’s not suitable to fill irregular hole. To make full use of the geometric priors of facial structures, FCEN takes the thermal graph and segmentation graph of facial key points as prior knowledge for inpainting network to constrain network training. However, the inaccuracy of the thermal map detection, it is easy to cause the unconformity betw**e**en the reconstructed details and the original image [11]. EdgeConnect proposes a two-stage training model composed of edge and image generator network. The edge generator restores the edge contour of the missing region, and the image generator network takes the restored edge information as prior to fill the missing region, but it still exists texture blur [12].

In order to further solve the problems described above, we proposed LSK-FNet model that combines edge generator network and face inpainting network to learns a prior knowledge for image inpainting. Similar to the idea proposed in BrainIoT [43], learning the edge information in the LSK-FNet model is based on the relationship between facial images and edge information. To improve the performance of image inpainting, we regard the learned edge information as the prior knowledge for facial image inpainting. The edge generation network is integrated with gated convolution to generate more accurate prior knowledge, while the face inpainting network uses prior knowledge, gated convolution and region normalization to generate fine details for missing region. Our main contributions are presented as follows:

- We proposed an image inpainting framework with learnable structure knowledge, which consists of edge generator and face inpainting. Edge generator could learn edge structure through a GAN combined with gated convolution and residual blocks. As prior information, edge structure is integrated into the face inpainting network. This essential two-step framework makes the reconstructed image structure more reasonable.

- We use gated convolution in both network of edge generation and face inpainting. As a learnable mask updating mechanism for improving the structural integrity of missing regions, gated convolution can not only select features according to background, mask and sketch, but also utilize some channels' dynamic semantics features.
- We integrate region normalization, a learnable normalization method, into the face inpainting network. It normalizes the known region and the missing region respectively, which can effectively solve the problem of mean and variance shift, and hold the global and local structural consistency.

2. Related Work

Image inpainting The traditional image inpainting based on patch-based method [3] seeks the best matching blocks through known regions and relies on appropriate example patches in the context, which is unable to create new objects or textures. In recent years, many deep learning-based methods have been proposed. Context Encoder was the earliest model that used generative adversarial network to fill missing image. It adopts encoder-decoder structure, which can reasonably predict the image. But, it's difficult to maintain global consistency [13]. Iizuka et al., used local and global discriminators respectively to generate real alternative content for the missing region and maintain the consistency of the image. They also use dilated convolution to increase the receptive field to improve the continuity of the missing edge [14]. In order to generate more realistic images with visual effects, some generation models usually incorporate spatial attention mechanism into the generator to improve the restoration effect of texture [15-21]. Since vanilla convolution uses the same convolution operation for all pixels resulting in fuzzy results when filling irregular missing hole, Liu et al., proposed partial convolution, which uses a binary mask to control the convolution area, so that the convolution only depends on the effective pixel. However, all channels in each layer share the same weight, and the feature extraction is not flexible enough [22]. Therefore, Yu et al., used gated convolution to extend the feature selection mechanism of each layer in the network to learn position information, and further generalize the idea of partial convolution to make the restored image more consistent with the real structure [23].

Face inpainting Due to the complexity and diversity of facial structures, traditional inpainting methods are no longer applicable. Li et al., proposed a semantically interpreted facial generation network, which used semantic segmentation to divide the face image into 11 regions. It used the semantic gap between the restored image and the real image as the semantic loss regularization term to constrain the network training [10]. Zheng et al., took the improved variational autoencoder as generation network, which can produce a variety of completion results while ensuring the generation quality [25]. Zhou et al., took facial information as prior knowledge and added multiple discriminators to these areas to improve the accuracy of face inpainting [41]. Li et al., proposed an improved GAN to restore the self-collected dataset of high-resolution human face. It output better human visual performance image when restore with irregular holes using the global average pooling in their network [42]. But the method is applied to mobile phone, not suitable for our digital document manufacturing application.

Structure Information for Inpainting Current image inpainting methods usually use different types of structural knowledge as prior knowledge to assist image inpainting [10, 11, 19, 26-30]. SPG-Net method predicted semantic segmentation labels of missing regions as structural information to constrain training. However, areas with similar semantic labels may have different textures (for example, windows and walls of the same building). This makes it

difficult to recover the image [19]. Li et al., proposed progressive reconstruction of visual structure (PRVS), which gradually integrates structural information into image features to output more structured images [28]. Nazeri et al., added edge generation to image completion. First, the complete edge information is predicted by the edge generation network. Then it is integrated into the image inpainting network, so that the image inpainting network can perceive the structural information [12]. Xiong et al., proposed an image inpainting model based on foreground perception, which divided the model into multiple subnets and gradually guide the completion of the image [26]. The model could perceive salience knowledge. The model first learns to predict the foreground contour, and then uses the predicted contour as a guide to draw in the missing region. These four methods use the structural information of incomplete images as priori knowledge to improve the accuracy of image inpainting. But there is still potential to improve performance.

3. Our Approach

3.1 LSK-FNet

Our approach, we called LSK-FNet, is based on a two step's GAN. As shown in **Fig. 1**, the overall model includes two networks: edge generation network and face inpainting network. Each network is composed of a generator and a discriminator, named G and D respectively. Mask, corrupted edge and incomplete grayscale images are put to the edge generation network simultaneously to produce the predicted edge. Predicted edge (training) and truth edge are used to train the D1(the discriminator of edge generation network). In the face inpainting network, predicted edge (trained) and corrupted face image are used to produce the output with G2(the generator of face inpainting network). Truth image and inpainted image (training) is used to train the D2(the discriminator of face inpainting network). Finally, output the inpainted image (trained). Both the edge generation network and the face inpainting network are GAN essentially. Intuitively, face inpainting guided by the edges is more effect than nothing guided. Because face detail is generally contained in these edges, hence edges could be considered as prior knowledge, which indicate the structure information and the information about the way how to fill the holes.

Gated convolutions with residual blocks are fused in the both of the edge generation network and the face inpainting network. The gated convolution has the capability of a learnable mask updating mechanism, so the edge generation network fused with gated convolution could discovery the hidden information among these visual features. The face inpainting network fused with gated convolution could combines background, mask and sketch, also consider the semantic segmentation of some channels, which can improve the structural integrity of missing regions.

Region normalization (RN) is used to promote performance of the face inpainting network. As a typical normalization feature across spatial dimension, previous image inpainting methods apply usually feature normalization (FN) in their networks. Shifts of mean and variance will occur due to the normalization impact ignoring of the input image corrupted regions. RN can effectively solve the problems of these two shifts. RN is a learnable way, which can normalize the known region and the missing region respectively. RN has the advantage to hole the global and the local structure consistency simultaneously.

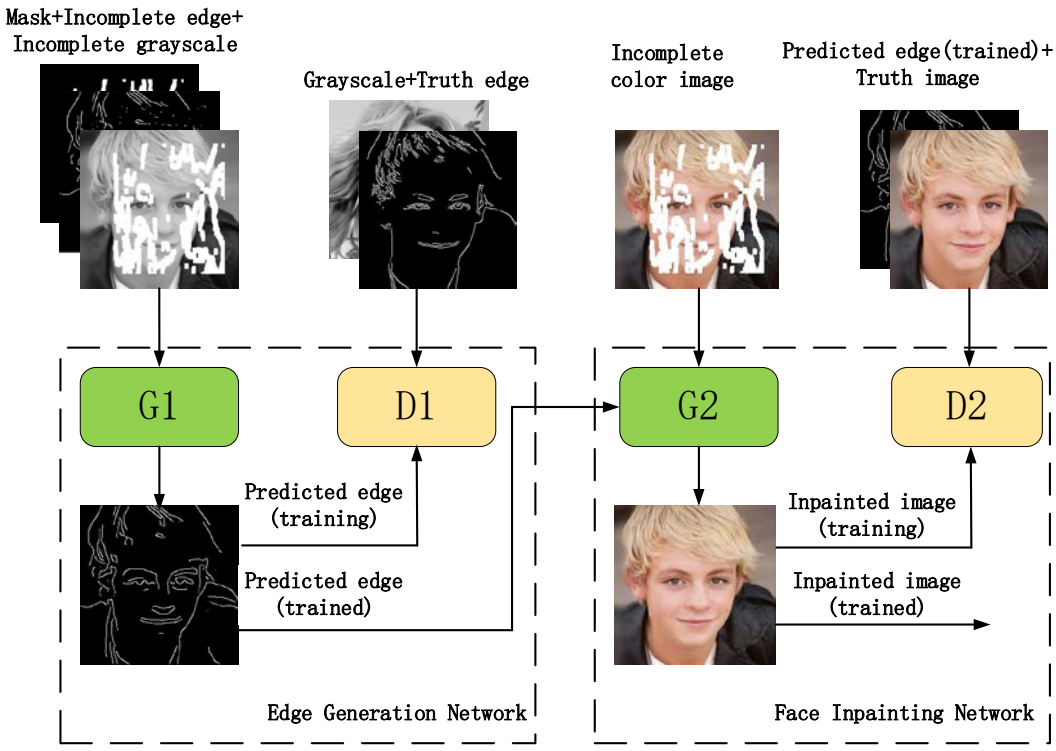


Fig. 1. Overall architecture of our proposed model

3.2 Edge Generation Network

Gated convolution can automatically learn mask update strategies from data [23]. It not only can dynamically identify the position of effective pixels in the image but also excel at transforming corrupted and uncorrupted regions distinctively to help the fusion of them. The formular defined as:

$$\begin{cases} \text{Gating}_{y,x} = \sum\sum \text{W}_g \cdot \text{I} \\ \text{Feature}_{y,x} = \sum\sum \text{W}_f \cdot \text{I} \\ \text{O}_{y,x} = \Phi(\text{Feature}_{y,x}) \odot \sigma(\text{Gating}_{y,x}) \end{cases}$$

Where σ is sigmoid function thus the output gating values are between zeros and ones. Φ can be any activation functions (such as, ReLU, ELU and LeakyReLU). W_g and W_f are two different convolutional filters that extract meaningful information though element-wise multiplication on effective pixels and image features. Both image and mask are input into gated convolution to train synchronously, rather than individually in vanilla convolution. This makes the results of image inpainting more accuracy.

As shown in **Fig. 2**, edge generation network consists of Generator (G1) and Discriminator (D1). Specifically, the generator is composed of encoder that down-sample twice, followed by eight residual blocks [34] and the decoder that up-sample twice. Residual blocks can avoid gradient diffusion problem caused by deeper networks. Each layer of the network uses gated convolutions, instead of vanilla convolution, to study semantic segmentation in some channels, which can not only learn to select different features along with mask, sketch, and background,

but also generate more plausible inpainting results. The discriminators based on a 70×70 PatchGAN [33] architecture is used to estimate the validity of edge map and parameters, and intensify ability of discrimination for image. Generally, spectral normalization is employed in all edge generation networks [35].

D1 is used to verify whether the predicted edge (training) is accuracy. The performance of discriminator for edge images will be improved via continuous learning. We used Canny edge detector [12] to extract truth edge information of completed images. Next, the extracted truth edge information combined with predicted edge (training) are input into D1 to improve the ability of discrimination. Through repeated learning, the final output of edge image is close to the real edge.

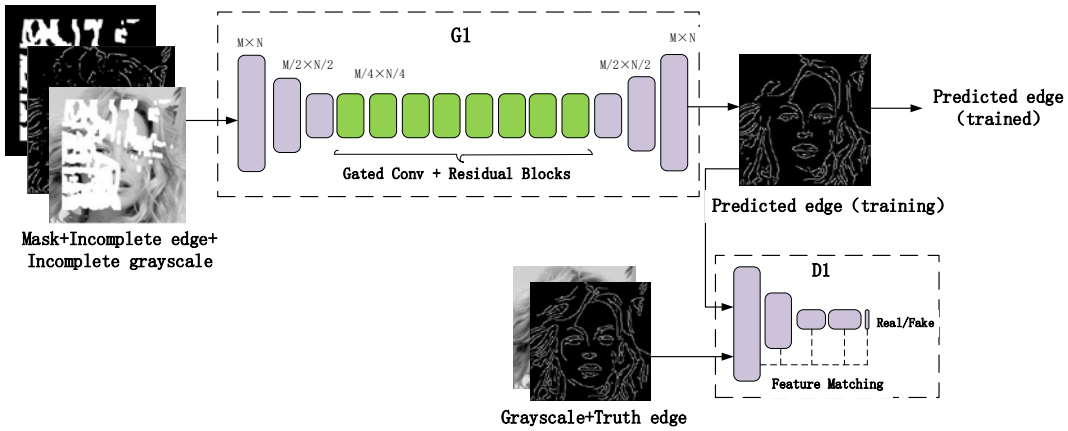


Fig. 2. Edge Generation Network

Let T_{gt} be ground truth images. Their edge map and grayscale counterpart will be denoted by C_{gt} and T_{gray}, respectively. Image mask M denotes a precondition (1 for the missing region, 0 for background). ⊙ denotes element-wise multiplication. G1 is the generator of edge information. In the edge generator, we use the corrupted grayscale $\widetilde{T}_{gray} = T_{gray} \odot (1 - M)$ as input, and its corrupted edge is $\tilde{C}_{gt} = C_{gt} \odot (1 - M)$. The generator predicts the edge map for the masked region.

$$C_{pred} = G1(\widetilde{T}_{gray}, C_{gt}, M) \tag{1}$$

We use C_{gt} and C_{pred} conditioned on T_{gray} as inputs of the discriminator that predicts whether or not a predicted edge image is real. The loss function is constructed to train the generative adversarial network to obtain the edge generation network. The adversarial loss is defined in formula (2) and the feature matching loss is defined in formula (3).

$$\mathcal{L}_{adv1} = \mathbb{E}_{(C_{gt},\ T_{gray})} \log\left[D1\left(C_{gt},\ T_{gray}\right)\right] + \mathbb{E}_{T_{gray}} \log\left[1 - D1\left(C_{pred},\ T_{gray}\right)\right] \tag{2}$$

$$\mathcal{L}_{FM} = \mathbb{E}\left[\sum_{i=1}^{L} \frac{||D1^{(i)}(C_{gt}) - D1^{(i)}(C_{pred})||_1}{N_i}\right] \tag{3}$$

Where L is the final convolution layer of the discriminator, $D1^{(i)}$ is the number of elements in the i'th activation layer, and $N_i$is the activation in the i'th layer of the discriminator. The edge generator network is comprised of an adversarial loss and feature-matching loss which are defined as

$$\min_{G1} \max_{D1} \mathcal{L}_{G1} = \min_{G1}(\lambda_{adv1} \max_{D1}(\mathcal{L}_{adv1}) + \lambda_{FM}\mathcal{L}_{FM}) \tag{4}$$

Where λ_{adv1} and λ_{FM} are regularization parameters. In our experiments, we set λ_{adv1}= 1 and λ_{FM}= 10, which has been shown to be effective in previous work [12].

3.3 Face Inpainting Network

Yu et al., proposed that RN divides spatial pixels into different regions according to the input mask, and computes the mean and variance in each region for normalization (shown in **Fig. 3**) [31]. Where N, C, H, Ware batch size, channel quantity, height, and width respectively. The blue and black boxes denote the corrupted and uncorrupted data respectively, which will be normalized separately later. There are two kinds of RN for image inpainting network: (1) Basic RN (RN-B), In the early layers of the network, the input image has massive corrupted areas, which lead to serious mean and variance shift problems, which are solved by RN-B via separately normalizing corrupted and uncorrupted regions. (2) Learnable RN (RN-L), it is difficult to obtain a region mask from the original mask after passing through several convolutional layers as the corrupted regions are gradually fused. RN-L solve the difficulty by learning to detect potentially corrupted regions via utilizing the spatial relationship of the input feature and generates a region mask for RN. Furthermore, RN-L can also boost the combination of corrupted and uncorrupted regions by global affine transformation. RN-L not only solves the mean and variance shift problem, but also enhance the reconstruction of corrupted regions. RN-L is suitable for latter layers of the network.

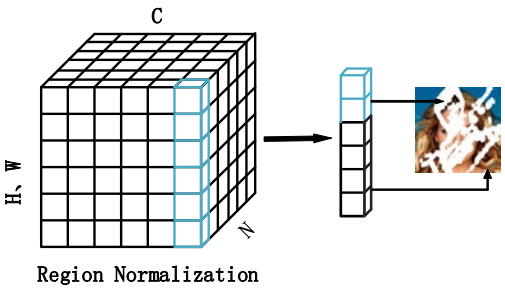


Fig. 3. Region Normalization

The architecture of face inpainting network (shown in **Fig. 4**) consist of G2 and D2, which is similar to edge generation network. In the face inpainting network, the generators consist of encoders that down-sample twice, followed by eight residual blocks and decoders that up-sample images back to the original size. Gated convolution is used in the residual block of the generator to extract the input image features, and it can learn constantly to distinguish corrupted and uncorrupted regions. Negative effects of corrupted region on face inpainting can be reduced. The color and detailed texture structure of image inpainting are more reasonable. Finally, the quality of the image inpainting is improved. RN will be used in the network. It uses spatial relations of input features to estimate potential corrupted regions to generates region mask and performs global affine transformation to enhance their fusion. Then, mean and variance shift problem are solved and reconstruction of corrupted region is promoted.

We take corrupted image and predicted edge (trained) which generated by edge generator network and taken as prior knowledge as input for G2, and then the completed image output by G2 together with uncorrupted original image is input into discriminator D2. By constantly comparing and updating the parameters of discriminator, the image inpainting ability of G2 is improved. The real image is taken as the input in the discriminator. With massive training, the completed image will be more similar to real images.

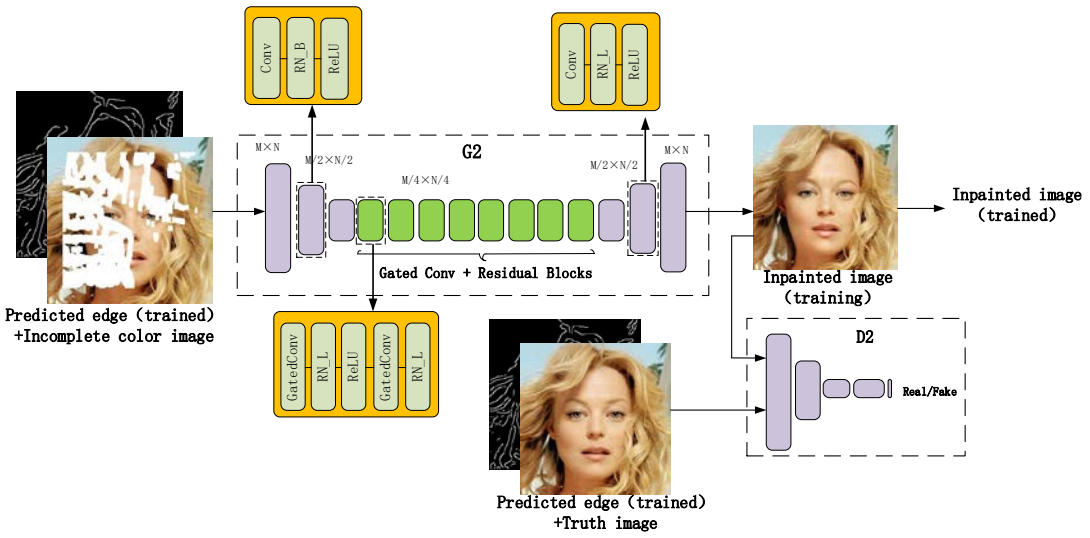


Fig. 4. Face Inpainting Network

The image completion network uses the incomplete color image, i. e. $\widetilde{T}_{gt} = T_{gt} \odot (1 - M)$ and the edge map as input. The edge map is come from the edge generator network, i.e., $C_{con} = C_{gt} \odot (1 - M) + C_{pred} \odot M$. The network returns a color image T_{pred}, with missing regions filled in, that has the same resolution as the input image.

$$T_{pred} = G2(\widetilde{T}_{gt}, C_{con}) \tag{5}$$

In image inpainting network, Loss function consists reconstruction loss, adversarial loss, perceptual loss, and style loss. These loss factors are used to train GAN for obtaining better performance. The adversarial loss is defined as:

$$\mathcal{L}_{adv2} = \mathbb{E}_{(T_{gt},\ C_{con})} \log[D2(T_{gt},\ C_{con})] + \mathbb{E}_{C_{con}} \log[1 - D2(T_{pred},\ C_{con})] \tag{6}$$

Perceptual loss [25] penalizes results that are not perceptually similar to labels by defining a distance measure between activation maps of a pre-trained network. The loss defined as:

$$\mathcal{L}_{prec} = \mathbb{E}[\sum_{i=1}^{L} \frac{||\phi_i(T_{gt}) - \phi_i(T_{pred})||_1}{N_i}] \tag{7}$$

Where ϕ_i is the activation map of the i'th layer of a pretrained network [36]. These activation maps are also used to compute style loss [37] which measures the differences between covariances of the activation maps. Given feature maps of sizes $C_j \times H_j \times W_j$, style loss is computed by

$$\mathcal{L}_{style} = \mathbb{E}_j[||G_j^{\phi}(T_{pred}) - G_j^{\phi}(T_{gt})||_1] \tag{8}$$

Where G_j^{ϕ} is a $C_j \times C_j$ Gram matrix constructed from activation maps ϕ_j in style loss. Discriminator in face inpainting network, integrated with reconstruction loss, adversarial loss, perceptual loss, and style loss, can discriminate the completed image. The total loss function is defined as:

$$\mathcal{L}_{G2} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{adv2}\mathcal{L}_{adv2} + \lambda_p\mathcal{L}_{perc} + \lambda_s\mathcal{L}_{style} \tag{9}$$

Where λ_{L1} , λ_{adv2}, λ_p and λ_s are regularization parameters. In our experiments, we set $\lambda_{L1} = 1$, $\lambda_{adv2} = \lambda_p = 0.1$ and $\lambda_s = 250$,which has been shown to be effective in previous work [12].

4. The Experimental Results and Analysis

Our LSK-FNet model is implemented in PyTorch. Its computational hardware mainly includes Intel CPU E5 (2.60GHZ) and GTX1080ti GPU. We conduct extensive experiments on high-quality human face dataset CelebA-HQ [38]. The image size in this dataset is 256*256. All the images are split into 28K, 1K and 1K for training, validation and testing respectively. We used 12K irregular mask images of 256*256 for training and evaluation. The mask images are split into 10K, 1K and 1K for training, validation and testing respectively. We use the Adam optimizer [39] with β1= 0 and β2= 0.9 to train our model. In edge generation network, the weights of loss function are set to λ_{adv1}=1 and λ_{FM}=10 respectively. In face inpainting network, the weights of the loss function are λ_{L1}=1, $\lambda_{adv2} = \lambda_p$=0.1 and λ_s=250.

We use the commonly-used L1 loss, peak signal-to-noise ratio (PSNR), and structural similarity (SSIM) as evaluation metrics. As gated convolution applied in irregular corrupted images, the algorithm is evaluated on irregular mask. The irregular dataset is grouped into five intervals along with the mask region, i.e.,10%-20%, 20%-30%, 30%-40% and 40%-50%, to observe image inpainting results of mask in each interval. We compare our method to the following five methods.

- Contextual Attention (CA): Method proposed by Yu et al [15].
- Generative Multi-column Convolutional Neural Networks (GMCNN): Method proposed by Wang et al [40].
- EdgeConnect (EC): Method proposed by Nazcri et al [12].
- Pluralistic Image Completion (PIC): Method proposed by Zheng et al [25].
- Region Normalization (RN): Method proposed by Yu et al [31].

In the LSK-FNet model, the edge knowledge of corrupted images are extracted by edge generation network，shown in **Fig. 5(c)**. The prior knowledge generated in the edge generation network is integrated into face inpainting network for image inpainting, shown in **Fig. 5(c)**. **Fig. 5** shows a sample of images which are generated by our model. The restoration effect of corrupted face image is determined by the generated edge graph. For example, the eyebrow direction and the hair in **Fig. 5(d)** was restored guided the edge graph in **Fig. 5(c)**.

(a) Input image (b) Corrupted image (c) Predict edge (d) Inpainted image

Fig. 5. Inpainted instance produced by our framework.

4.1 Quantitative Comparisons

We give the results of quantitative comparisons in **Table 1**. The compared models include: CA, GMCNN, EC, PIC and RN. The second column is the region of irregular masks at testing

time. The statistics are based on random masks with mask ratio 0%-50%. Three commonly used metrics are used: L1 loss, PSNR and SSIM. Symbol ↓ represents lower is better. Symbol ↑ represents higher is better. RN did not provide pre-training model for this dataset, so we represented it on the CelebA-HQ dataset according to the code published by the author. PIC could produce multiple results, so we selected the one which is closest to the Ground Truth. According to the results of the **Table 1** the method presented in this paper is superior to the other five methods. Compared to the EC, when corrupted area is less than 20%, L1 loss reduce by more than 4.3%. When corrupted area is less than 40%, the LSK-FNet model shows its superiority. On the whole, the method in this paper is superior to its comparison methods.

Table 1. Proposed framework is compared to other methods on CelebA-HQ.

	Mask	CA	GMCNN	EC	PIC	RN	Ours
L1↓	10% - 20%	0.294	0.290	0.208	0.366	0.335	**0.199**
	20% - 30%	0.320	0.316	0.245	0.373	0.346	**0.235**
	30% - 40%	0.313	0.340	0.278	0.400	0.354	**0.267**
	40% - 50%	0.368	0.364	0.308	0.408	0.359	**0.304**
PSNR↑	10% - 20%	30.141	29.472	33.672	33.948	15.486	**34.030**
	20% - 30%	25.691	24.937	29.372	29.560	15.290	**29.631**
	30% - 40%	22.944	22.260	26.683	26.561	15.005	**27.027**
	40% - 50%	21.003	20.401	24.490	24.382	14.577	24.442
SSIM↑	10% - 20%	0.934	0.929	0.960	0.960	0.803	**0.963**
	20% - 30%	0.873	0.868	0.922	0.922	0.764	**0.923**
	30% - 40%	0.805	0.805	0.876	0.874	0.710	**0.877**
	40% - 50%	0.734	0.738	0.822	0.819	0.644	0.819

4.2 Qualitative Comparisons

As shown in **Fig. 6**, the restored images of different methods are compared without any post-processing. The size of the corrupted area increases from top to bottom. Due to the integration of gated convolution and RN into our model, it can be seen that our model has the best performance in texture consistency near the boundary and color fusion. Besides, our model is also good at keeping the structure consistency even better than EC [12]. The images restored by CA [15], GMCNN [40] and RN [31] show visual artifacts, unclear edge structure and other features especially shown in R5 and R7. When the mask is small (e.g., R1 and R2), the completed results of EC [12], PIC [25] and our method are not much different from those of the input image, but ours has more detailed edge structure. For inpainting, CA [15], PIC [25], employ attention modules to learn contextual information. However, they still generate semantically inconsistent structures or textures because of the unreasonable attention scores. From **Fig. 6**, we can intuitively see that the LSK-FNet model restore effect is better in image rationality and visual continuity.

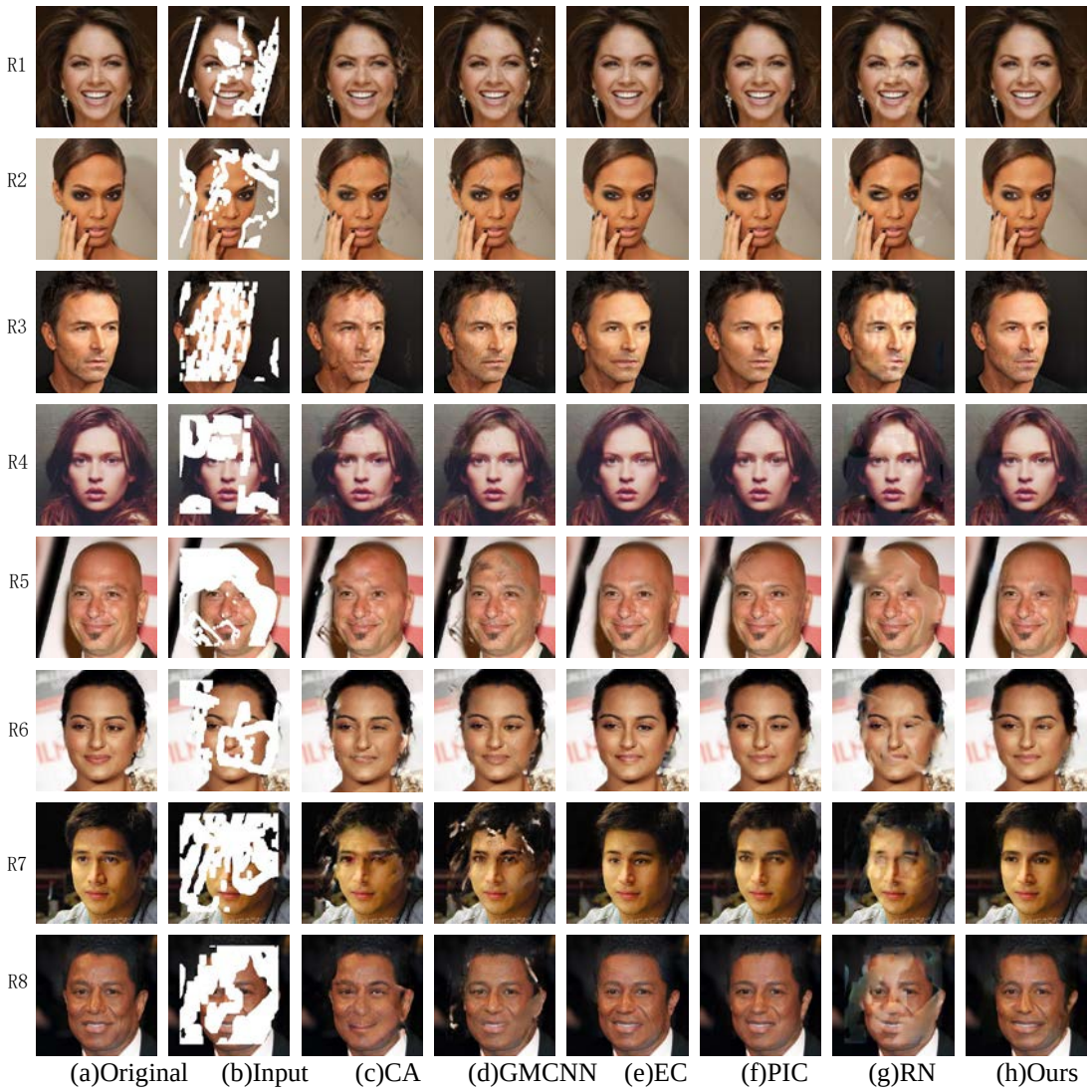
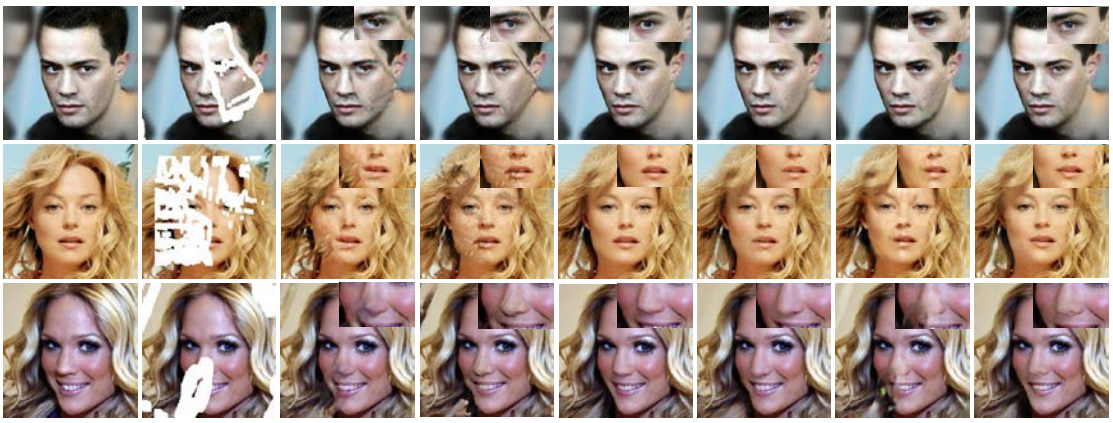


Fig. 6. Qualitative results of different inpainting methods on the CelebA-HQ dataset with irregular mask. (Top to bottom) The mask sizes are different: R1-R2(range from 10% to 20%), R3-R4(range from 20% to 30%), R5-R6(range from 30% to 40%), for R7-R8(range from 40% to 50%)

Fig. 7 shows the details in the facial image when applied different methods of completion. From the enlarged details of images, it can be seen that the completed images would generate visual artifacts when applied CA, GMCNN and PIC methods. In addition, these methods may generate artifacts or blurry effect on facial components such as the eyes in the first row. From the restore of facial details, we can see that the method in our paper is more reasonable in detail and texture.



(a) Original (b) Input (c)CA (d)GMCNN (e)EC (f)PIC (g)RN (h) Ours

Fig. 7. Detail comparison of different image inpainting methods

4.3 Ablation Study

LSK-FNet Without Gated Convolution or Region Normalization To verify the effectiveness of our framework on image inpainting, firstly we training the model without applying Gated Convolution or Region Normalization. The split of images and irregular mask on the dataset is same as section 4. To show noticeably better results, we use irregular mask to train and test with other conditions unchanged. Note that, image inpainting effect does not improve apparently when solely adding gated convolution or RN in the stage of filling the missing regions (Results shown in **Table 2**).

Table 2. Comparison results over CelebA-HQ with irregular mask. Our framework is compared with models of applying single gated convolution or RN. The third, fourth and fifth columns indicate specifications without gated convolution, without RN and both with gated convolution and RN, respectively.

	Mask	× Gated Conv	× RN	Ours
L1↓	10% - 20%	0.211	0.204	**0.199**
	20% - 30%	0.260	0.245	**0.235**
	30% - 40%	0.296	0.281	**0.267**
	40% - 50%	0.332	0.319	**0.304**
PSNR↑	10% - 20%	33.961	33.800	**34.030**
	20% - 30%	29.470	29.376	**29.631**
	30% - 40%	26.902	26.875	**27.027**
	40% - 50%	24.331	24.360	**24.442**
SSIM↑	10% - 20%	0.962	0.961	**0.963**
	20% - 30%	0.922	0.921	**0.923**
	30% - 40%	0.075	0.874	**0.877**
	40% - 50%	0.812	0.813	**0.819**

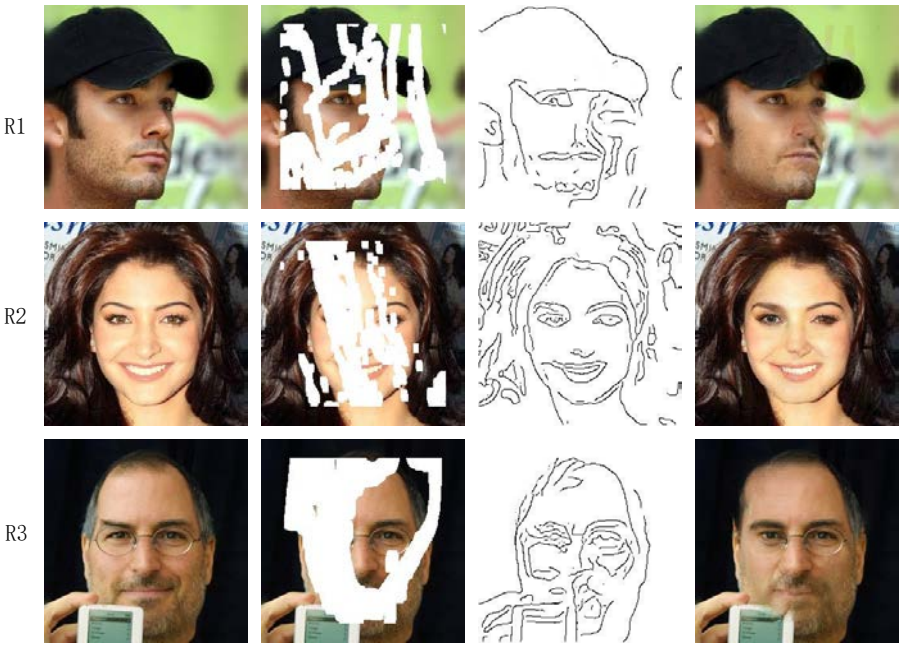
From the **Table 2**, we make a further discussion as follows: (1) Image quantity can be improved by using gated convolution or RN, but ineffective for corrupted regions. (2) The results are better when we use RN and gated convolution used in network simultaneously. (3) The performance on L1 loss do not surpass EC when RN or gated convolution are singly used.

However, we achieve significant improvement on performance with RN and gate convolution. (4) SSIM is improved no matter applied RN and gate convolution separately or together when the corrupted region is relatively small especially less than 20%.

4.4 Limitations

Due to the framework proposed in this paper is two-stage, the predicted edge image in the first stage has a great impact on the effect of image inpainting in the second stage. When the edge generation network fails to produce relevant edge information, the restore effect will be imperfect.

Fig. 8 shows the limitations of this method. It can be seen that the restoration effect of **Fig. 8 (d)** has a certain relationship with the edge generation graph of **Fig. 8 (c)**. When the corrupted face is a side face or large portion of the face image is missing, the restore effect will be imperfect. In the fourth image of R3, because the edge of high-texture region could not be described in the edge generation stage, the restored image was asymmetric at the position of the glasses.



(a) Input image (b) Corrupted image (c) Predicted edge (d) Inpainted image

Fig. 8. Limitations of the method

In the future work, we plan to investigate better edge generation network. Effective edge prior information will improve the performance of image inpainting. We believe that with an improved edge generation system, our model can be extended to very high-resolution inpainting.

5. Conclusion

In the paper, we proposed a Learnable Structure Knowledge of Fusion Network (LSK-FNet), which obtains prior knowledge with edge generation network for image inpainting. The overall model is two-stage: edge generation and face inpainting. Firstly, the feature selection function

of gated convolution makes the edge information more accurate. Secondly, we utilize the generated prior knowledge and corrupted images as input to the face inpainting network which is integrated with gated convolution. Moreover, RN is adopted to solve the mean and variance shift problems, which has been proved to be effective. Furthermore, Our LSK-FNet also improved the global and local structural consistency of images. Extensive qualitative and quantitative analyses show that our model achieve more coherent, fine-detailed and sharper results compared with some state-of-the-art methods.

References

[1] M. Bertalmio, G. Sapiro, V. Caselles, et al., "Image inpainting," in *Proc. of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, New York, USA, ACM, 417-424, 2000.

[2] J. B. Huang, S. B. Kang, N. Ahuja, et al., "Image completion using planar structure guidance," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1-10, 2014. Article (CrossRef Link)

[3] C. Barnes, E. Shechtman, Finkelstein A, et al., "PatchMatch: a randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1-11, July 27, 2009. Article (CrossRef Link)

[4] S. Darabi, E. Shechtman, C. Barnes, et al., "Image melding: combining inconsistent images using patch-based synthesis," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1-10, July 2012. Article (CrossRef Link)

[5] A. Criminisi, P. Perez, K. Toyama, "Object removal by exemplar-based inpainting," in *Proc. of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceeding*, Madison, WI, USA, pp. II-II, 2003. Article (CrossRef Link)

[6] Z. Xu, J. Sun, "Image inpainting by patch propagation using patch sparsity," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1153-1165, May 2010. Article (CrossRef Link)

[7] O. Le Meur, J. Gautier, C. Guillemot, "Examplar-based inpainting based on local geometry," in *Proc. of 2011 18th IEEE International Conference on Image Processing*, Brussels, Belgium, pp. 3401-3404, 2011. Article (CrossRef Link)

[8] Z. Qiang, L. He, D. Xu, "Exemplar-based pixel by pixel inpainting based on patch shift," *Communications in Computer and Information Science*, vol. 772, pp. 370-382, 2017. Article (CrossRef Link)

[9] A. V. D. Oord, N. Kalechbrenner, K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. of 33rd International Conference on Machine Learning, ICML 2016*, vol. 4, pp. 2611-2620, 2016.

[10] Y. Li, S. Liu, J. Yang et al., "Generative face completion," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5892-5900, 2017. Article (CrossRef Link)

[11] L. Song, J. Cao, L. Song, et al., "Geometry-aware face completion and editing," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 2506-2513, 2019.

[12] K. Nazeri, E. Ng, T. Joseph, et al., "EdgeConnect: structure guided image inpainting using edge prediction," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, South Korea, pp. 3265-3274, 2019. Article (CrossRef Link)

[13] D. Pathak, P. Krähenbühl, J. Donahue, et al., "Context Encoders: feature learning by inpainting," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2536-2544, 2016. Article (CrossRef Link)

[14] S. Iizuka, E. Simo-Serra, H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1-14, 2017. Article (CrossRef Link)

[15] J. Yu, Z. Lin, J. Yang, et al., "Generative image inpainting with contextual attention," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5505-5514, 2018. Article (CrossRef Link)

[16] Y. Zeng, J. Fu, H. Chao, et al., "Learning Pyramid-Context Encoder network for high-quality image inpainting," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR), Long Beach, CA, USA, pp. 1486-1494, 2019. Article (CrossRef Link)
[17] M. Sagong, Y. Shin, S. Kim, et al., "PEPSI: fast image inpainting with parallel decoding network," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 11352-11360, 2019. Article (CrossRef Link)
[18] C. Xie, S. Liu, C. Li, et al., "Image inpainting with learnable bidirectional attention maps," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, pp. 8857-8866, 2019. Article (CrossRef Link)
[19] Y. H. Song, C. Yang, Y. J. Shen, et al., "SPG-Net: segmentation prediction and guidance network for image inpainting," in *Proc. of 29th British Machine Vision Conference*, *BMVC 2018*, 2018. Article (CrossRef Link)
[20] N. Wang, J. Li, L. Zhang, et al., "MUSICAL: Multi-Scale image contextual attention learning for inpainting," in *Proc. of Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3748-3754, August 2019. Article (CrossRef Link)
[21] Y. G. Shin, M. C. Sagong, Y. J. Yeo, et al., "PEPSI++: fast and lightweight network for image inpainting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 252-265, Jan. 2021. Article (CrossRef Link)
[22] G. Liu, F. A. Reda, K. J. Shih, et al., "Image inpainting for irregular holes using partial convolutions," in *Proc. of 15th European Conference on Computer*, LNCS, vol. 11215 pp. 89-105, 2018. Article (CrossRef Link)
[23] J. Yu, Z. Lin, J. Yang, et al., "Free-Form image inpainting with gated convolution," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, pp. 4470-4479, 2019. Article (CrossRef Link)
[24] L. Haofu, Gareth, Z. Yefeng, et al., "Face completion with semantic knowledge and collaborative Adversarial Learning," in *Proc. of Computer Vision-ACCV 2018*, pp. 382-397, 2018. Article (CrossRef Link)
[25] C. Zheng, T. Cham, J. Cai, "Pluralistic image completion," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 1438-1447, 2019. Article (CrossRef Link)
[26] W. Xiong, J. Yu, Z. Lin, et al., "Foreground-Aware image inpainting," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 5833-5841, 2019. Article (CrossRef Link)
[27] Y. Ren, X. Yu, R. Zhang, et al., "StructureFlow: image inpainting via structure-aware appearance flow," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, pp. 181-190, 2019. Article (CrossRef Link)
[28] J. Li, F. He, L. Zhang, et al., "Progressive reconstruction of visual structure for image inpainting," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, pp. 5961-5970, 2019. Article (CrossRef Link)
[29] J. Yang, Z. Qi, Y. Shi, "Learning to incorporate structure knowledge for image inpainting," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12605-12612, 2020. Article (CrossRef Link)
[30] J. Li, N. Wang, L. Zhang, et al., "Recurrent feature reasoning for image inpainting," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 7757-7765, 2020. Article (CrossRef Link)
[31] T. Yu, Z. Guo, X. Jin, et al., "Region normalization for image inpainting," in *Proc. of the AAAI Conference on Artificial Intelligence*. vol. 34, no. 7, pp. 12733-12740, 2020. Article (CrossRef Link)
[32] Z. Guo, Z. Chen, T. Yu, et al., "Progressive image inpainting with full-resolution residual network," in *Proc. of ACM International Conference on Multimedia*, New York, NY, USA, pp.2496–2504, 2019. Article (CrossRef Link)
[33] P. Isola, J. Zhu, T. Zhou et al., "Image-to-Image translation with conditional adversarial networks," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5967-5976, 2017. Article (CrossRef Link)

[34] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016. Article (CrossRef Link)
[35] T. Miyato, T. Kataoka, M. Koyama, et al., "Spectral normalization for generative adversarial networks," in *Proc. of International Conference on Learning Representations*, 2018. Article (CrossRef Link)
[36] K. Simonyan, A. Zisscrman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of 3rd International Conference on Learning Representations*, 2015. Article (CrossRef Link)
[37] L. A. Gatys, A. S. Ecker, M. Bethge, "Image style transfer using convolutional neural Networks," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2414-2423, 2016. Article (CrossRef Link)
[38] T. Karras, T. Aila, S. Laine, et al., "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. of in International Conference on Learning Representations*, 2018. Article (CrossRef Link)
[39] D. P. Kingma, J. Ba, "Adam: a method for stochastic optimization," in *Proc. of International Conference on Learning Representations*, 2015. Article (CrossRef Link)
[40] W. Yi, T. Xin, X. Qi, et al., "Image inpainting via generative multi-column convolutional neural networks," *Neural Information Processing Systems foundation*, pp. 331-340, 2018.
[41] T. Zhou, C. Ding, S. Lin, et al., "Learning oracle attention for high-fidelity face completion," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 7677-7686, 2020. Article (CrossRef Link)
[42] Xuewei Li, Xueming Li, X. Zhang, et al., "A method of inpainting moles and acne on the high-resolution face photos," *IET Image Processing*, vol. 15, no. 3, pp. 833-844, February 2021.
[43] H. Yang, J. Yuan, C. Li et al., "BrainIoT: Brain-Like Productive Services Provisioning with Federated Learning in Industrial IoT," *IEEE Internet of Things Journal*, vol. 9, no. 3, pp. 2014-2024, 2022. Article (CrossRef Link)

You Yang received the Ph.D. degree in computer application technology from Beihang University, Beijing, China, in 2010. He is currently an associate professor of National Center for Applied Mathematics in Chongqing. His research interests include computer vision and document image processing.

Sixun Liu received the bachelor's degree in Computer Science and Technology from Chengdu College of Arts and Sciences, Chengdu, China, in 2019. She is currently a postgraduate of computer technology with the School of Computer and Information Science, Chongqing Normal University, Chongqing, China. Her research interest is mainly image inpainting.

Bin Xing received the Ph.D. degree in mechanical data processing from Ecole Centrale Paris, France, in 1998. He is currently the Chief Scientist of National Engineering Laboratory of Industrial Big-Data Application Technology and Industrial Big-Data Innovation Center of Chongqing. His research interests include the knowledge engineering and the intelligent decision in industry.

Kesen Li received the bachelor's degree in Internet of Things from Hubei University of Technology, China, in 2019. He is currently a postgraduate of computer technology with the School of Computer and Information Science, Chongqing Normal University, Chongqing, China. His research interest is mainly computer vision.