Neurocomputing 437 (2021) 95–106

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Facial image inpainting using attention-based multi-level generative network

Jie Liu, Cheolkon Jung *

School of Electronic Engineering, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 3 February 2020
Revised 18 December 2020
Accepted 28 December 2020
Available online 12 January 2021
Communicated by Steven Hoi

Keywords:
Face inpainting
Multi-level generator
Attention mechanism
Edge-preserving loss

ABSTRACT

Facial image inpainting is a challenging task since the facial images lose much content causing blur and unnaturalness. In this paper, we propose facial image inpainting using attention-based multi-level generative network. We adopt multi-level feature processing to reduce the training and testing time while enhancing the performance by reducing the number of channels of each convolutional layer in the generator. We combine attention with multi-level feature processing to maintain soft correlation with the surrounding content. For network optimization, we utilize two kinds of loss functions: content and texture. Content loss consists of mean absolute error (MAE) and edge-preserving losses, and produces inpainting results close to the ground truth. Texture loss consists of adversarial and perceptual losses to fine-tune the texture synthesis. Besides, we use the edge-preserving loss to take edge and patch similarity. Comparative experiments indicate that the proposed method produces photo-realistic and plausible inpainting results in random masks as well as outperforms the state-of-the-art ones in terms of quantitative measurements and subjective evaluations.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Recently, deep convolutional neural networks (CNN) and generative adversarial network (GAN) have been widely applied to many tasks, and there are remarkable improvements in image processing including single image super-resolution, image denoising, and image inpainting. Image inpainting is a task to generate a complete output given an incomplete image with a large hole or multiple holes as shown in Fig. 1. It is a key problem in computer vision, which is often used to recover objects or synthesize the missing parts in an image. Researchers have achieved good performance in image inpainting with the help of CNN [1] and GAN [2–5]. Unlike general image inpainting approaches to generating suitable content of missing regions similar to the background from image patches, facial image inpainting needs plausibly semantic knowledge about the target object for photo-realistic outputs because the missing regions often include key components (*e.g.* eyes, nose, and mouth). In general, face image completion is an important step in many image processing applications [6–8]: face alignment, face recognition, facial expression analysis, and face parsing.

Up to the present, most traditional methods for facial image inpainting are based on diffusion [9–11], patch matching [12–14] and big dataset [15]. Pathak et al. [1] proposed contextual encoder-decoder structure for image inpainting. This is the first work to introduce CNN into image inpainting, and they achieved remarkable performance improvement on Paris StreetView dataset. Li et al. proposed a face completion method with the help of semantic parsing network [4]. However, this method still has some limitations such as blurry artifacts in facial image inpainting causing unpleasant visual effects. A number of image inpainting methods have been proposed, and most of them are not applicable to facial image inpainting. If GFC [4] and CA [3] are light-weight, their results are not photo-realistic and often cause artifacts. Zhang et al. [16] proposed a mesh-based structure for facial image inpainting. Song et al. [17] utilized facial landmark and parsing maps for face completion. Some models with a little bigger network produce good inpainting results, such as generative multi-column convolutional neural networks (GMCNN) [18] and multi-level generative network (MLGN) [19] whose parameters are 12.5 M and 15 M, respectively. A light-weight network for facial image inpainting is required.

For facial image inpainting, the recovery of facial structures is important in facial image inpainting along with texture details. Faces have a standard pattern that consists of eyebrows, eyes, nose, and mouth. Motivated by multi-level feature processing, we divide features into three scale branches: main branch, middle spatial resolution (MSR) branch, and low spatial resolution (LSR) branch. The

* Corresponding author.
E-mail addresses: jieliu543@gmail.com (J. Liu), zhengzk@xidian.edu.cn (C. Jung).

https://doi.org/10.1016/j.neucom.2020.12.118
0925-2312/© 2021 Elsevier B.V. All rights reserved.

Fig. 1. Facial image inpainting by the proposed method. The resolution of the input image is 256 × 256. **First and forth column**: Masked images by a hole or multi holes. **Second and fifth column**: Generated images by the proposed method. **Third and sixth column**: Ground truth. From the first to the third rows, the masks with 256 × 256 size are located at random position, while the last row is randomly masked with three holes.

multi-level generative network adopts multiple scales of feature maps and concatenates them to capture multi-scale feature information. We use a multi-level architecture in the generator to enhance facial structures with texture details. Moreover, we utilize an adversarial loss to fine-tune texture details. To reduce the time costs of training and testing, we propose multi-level generator based on attention for facial image painting. To maintain the relationship with the surrounding content, we introduce an attention mechanism into MSR and LSR branches. Moreover, we adopt an attention-based generative network, in which the attention module plays a role of carrying contextual information from long-range pixels dependencies based on patch representation similarity. From [20], the criss-cross block is a more effective and efficient attention mechanism. As illustrated in Fig. 2, the input is the feature map, while each pixel of the input is put into two 1 × 1 convolution layers to generate two feature maps, **K** and **Q**. After performing an affinity operation on **K** and **Q**, we produce the final attention map by aggregation operation and element-wise addition. Regarding loss function, we utilize edge-preserving loss for structure matching and patch similarity. We also utilize adversarial and perceptual losses to get photo-realistic textures. Therefore, the proposed model outperforms the others in completing the details of facial components and achieves a balance between high-quality inpainting results and the number of parameters. Fig. 3 illustrates the network structure of the proposed generator, while Fig. 6 shows the network architecture of the proposed discriminator. We adopt GAN as a basic framework where the generator is a multi-level feature fusion network (Fig. 3) and and the discriminator has two branches: global branch and local branch (Fig. 6).

In our previous paper [19], we provided preliminary results of facial image inpainting using multi-level generative network (MLGN). Although our model has achieved competitive performance in facial image inpaining, it needs to effectively capture the information from feature maps. In this extended paper, we combine a criss-cross block as attention with MLGN to maintain soft correlation with the surrounding content. Moreover, multi-level architecture and criss-cross attention are very effective for facial image inpainting. Faces have a common structure with distinct features (mouth, eyes, and nose), and thus the multi-level architecture is highly applicable to the facial image inpainting. We provide a novel discriminator that consists of global and local branches in total adversarial loss and thus improve details in facial image inpainting.

Compared with existing methods, main contributions of this paper are as follows:

- We build a multi-level generative network (MLGN) to capture multiscale feature information and enhance representation ability.
- We combine attention mechanism with MLGN to maintain soft correlation with the surrounding content.
- We utilize two kinds of losses: content and texture. Content loss reconstructs global facial content, while texture loss enhances details in facial images.

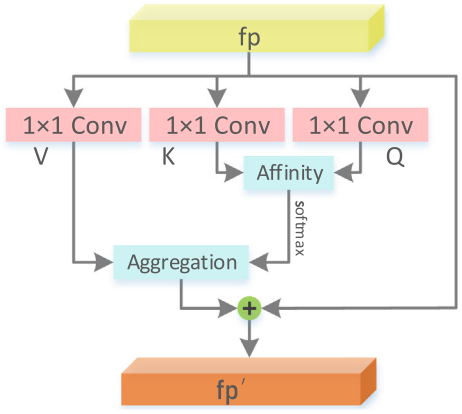

Fig. 2. Criss-cross block. The input **fp** (feature maps) and the output fp′ (feature maps with attention) are both shown as the shape of $R^{H\times W\times C}$.

2. Related work

2.1. Image generation by GAN

Generative adversarial network (GAN) [21] achieves a considerable breakthrough in image generation by producing realistic images. In general, GAN consists of one generator and one discriminator, whose training strategy is adversarial learning. Radford et al. proposed deep convolutional generative adversarial network (DCGAN) [22] for applicability as general image representations, and this approach has achieved plausible results in image generation. Then, several works have utilized GAN to perform texture synthesis and detail enhancement. For instance, Yi et al. adapted cycle-consistent adversarial networks [23] for unpaired image-to-image translation. Chen et al. [24] proposed two-way GAN, in which they used U-Net [25] as a global generator using an adaptive weighting scheme on WGAN [26]. Chen et al. [27] proposed an asymptotic residual back-projection network (RBPNet) for face super-resolution from a very low-resolution image. They progressively learned feature details from residual and edge maps and then combined them with GAN loss to fine-tune the network.

2.2. Image inpainting

Traditional image inpainting approaches are not effective for facial image completion because it is difficult to find the best patch

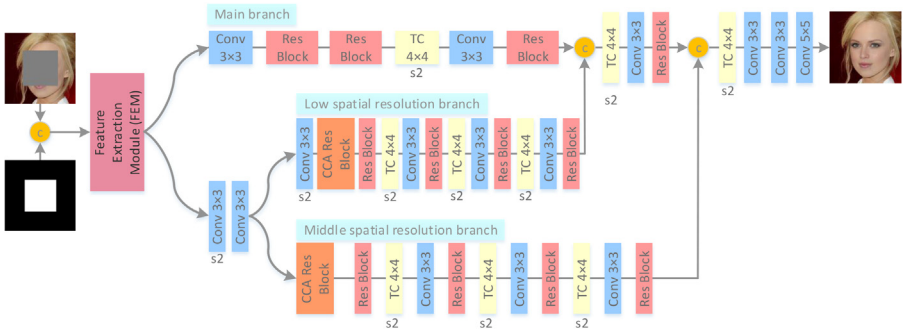

Fig. 3. Network structure of the proposed generator. The generator is used to fill the masked image with resolution 256×256. Instance normalization and leaky ReLU are following after each convolution and transposed convolution except the first and final convolution layers. The feature map sizes of the main, low and middle branches are 32×32, 8×8 and 16×16, respectively.

from a masked facial image. Based on CNN, Pathak et al. [1] proposed contextual encoder-decoder to reproduce the Paris StreetView dataset. They achieved outstanding performance for low-resolution image inpainting at that time. Yang et al. [2] offered multi-scale neural patch synthesis to process high-resolution image inpainting. They handled global content and local texture by downsizing the image to 3 scales (including the size of the original image) and combining holistic content loss and local texture loss. Yu et al. [3] proposed a novel deep generative approach based on a two-stage feed-forward fully convolutional neural network with contextual attention (CA). They processed images with multi-hole at arbitrary positions causing shape deformation in face completion. For face completion, Li et al. [4] proposed a method with the condition of semantic information constraint (GFC), which estimated reliable pixels while keeping the consistency between local and global contents by combining reconstruction loss, global and local adversarial losses, and semantic parsing loss. Yan et al. [28] proposed Shift-Net for inpainting based on U-Net architecture that established feature relationship between existing and missing regions with a custom shift-connection layer. Wang et al. [18] suggested a multi-column convolutional neural network for image inpainting (GMCNN) and introduced implicit diversified Markov random fields (ID-MRF) as regularization to find the nearest neighbors for patches in the generated region. This approach produced fairly good visual quality with a large number of parameters. Liu and Jung [19] proposed a multi-level generative network (MLGN) to synthesize a high-quality image with inputting incomplete facial image. Liu et al. proposed a structure-guided network architecture for image inpainting to keep consistency between the structure and its neighborhoods [29]. They utilized a tensor tree rank to build a globally multi-dimensional structure with total variation minimization to keep locally patch-wise smoothness for inpainting.

2.3. Criss-cross module

Non-local operation [30] was originally proposed for video classification and static image recognition, which achieved outstanding performance. It was computed as a weighted sum of all features at a position. Non-local matching was also successful in many image processing tasks, such as inpainting [12] and texture synthesis [31]. Although non-local matching brought high quality, it had a large amount of calculation. Huang et al. proposed criss-cross matching [20] as an attention module for semantic segmentation. In addition, the criss-cross non-local attention module needs less calculation, which can balance the ability of capturing semantic information and the time cost. Moreover, the criss-cross non-local operation was calculated in the horizontal and vertical directions.

3. Proposed method

In this work, we propose facial image inpainting using attention-based multi-level generative network [19]. We adopt the multi-level generator based on attention to reduce the number of trainable parameters. Attention enhances features by considering similarity, and thus it reduces the number of layers in three branches. Thus, the training time of the proposed generator is less than recent methods based on GAN. To keep the integrity of local–global contents between the filled region and the background, we introduce the global and local discriminator, which is also used in the Iizuka et al.'s work [32]. Moreover, we adopt the attention mechanism to establish the best match. We train the proposed network with a combination of content and texture losses. The texture loss consists of a perceptual loss and two adversarial losses, which maintains global consistency. We form the content loss based reconstruction (MAE) loss and edge-preserving loss, which recovers facial edges and makes the generated faces more plausible.

3.1. Network architecture

3.1.1. Generator

The generator is composed of convolution and transposed convolution operators, and the network structure is shown in Fig. 3. We adopt multi-level feature processing and criss-cross attention operations to produce perceptually realistic results based on deep features. As shown in Fig. 3, the input is defined as follows:

$$\mathrm{I}_{\mathrm{masked}} = \mathrm{I}_{\mathrm{gt}} \odot (\mathbf{1} - \mathrm{I}_{\mathrm{mask}}), \tag{1}$$

$$\mathrm{I}_{\mathrm{in}} = [\mathrm{I}_{\mathrm{masked}}, \mathrm{I}_{\mathrm{mask}}] \tag{2}$$

where I_{gt} and $\mathrm{I}_{\mathrm{mask}}$ are the ground truth and the mask image, respectively. Then, the masked image $\mathrm{I}_{\mathrm{masked}}$ is produced by I_{gt} and $\mathrm{I}_{\mathrm{mask}}$, which is defined in Eq. 1, in which $\odot$ denotes Hadamard product. For Eq. (2), we firstly take $\mathrm{I}_{\mathrm{masked}}$ concatenated with $\mathrm{I}_{\mathrm{mask}}$ in the channel dimension as inputs, and then we feed I_{in} into feature extraction module (FEM) as shown in Fig. 3. The feature extraction module (FEM) is shown in Fig. 4a. FEM contains convolution with the stride of 2, instance normalization, followed by each convolution except the first layer, leaky ReLU as activation function, and residual block as shown in Fig. 4b. FEM is used to synthesize the common feature maps for three branches: main, LSR, and MSR. In addition, the generator is optimized and trained by the total loss function calculated by the output image $\mathrm{I}_{\mathrm{out}} = G(\mathrm{I}_{\mathrm{in}})$ and I_{gt}, where G is the generator.

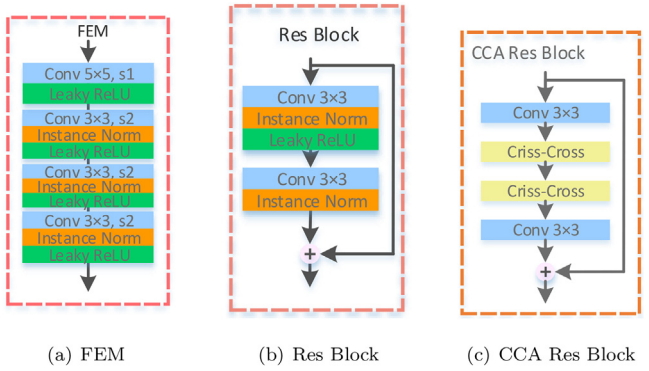

Fig. 4. (a): Network architecture of the feature extraction module (FEM). (b): The architecture of the residual block (ResBlock). (c): CCA Residual block in LSR and MSR, in which Criss-Cross is criss-cross non-local calculation.

3.1.2. Multi-level feature processing

Pyramid network [33] is a convolution neural network, which takes advantage of multi-level feature processing. It uses pyramid pooling to capture contextual information. Wu et al. [34] utilized the pyramid network for UAV image dehazing to get global prior information. Different from it, we utilize convolution with stride 2 to keep structures and details by integrating feature information. Therefore, we build low spatial resolution (LSR) branch and middle spatial resolution (MSR) branch in the proposed generator to extract feature maps with different scales and fuse information components with different feature resolutions. In practice, the multi-level integration increases the number of parameters than single-level integration. It is a trade-off between the performance and the number of parameters from the network structure. As shown in Fig. 3, the main branch assists in exploring high-resolution features. In the generator, we use some residual blocks in three branches to reuse features from the previous ones. In branches, the instance normalization and leaky ReLU are also utilized after each convolution operation. We utilize the transposed convolution as the up-sampling operation. After each transposed convolution, one general convolution and a residual block are added to mitigate grid artifacts caused by transposed convolution except the last two up layers. Furthermore, we perform down-sampling through convolution with stride 2.

3.1.3. Attention mechanism

Attention is used to retain more details and collect more useful contextual information. In the attention module, each pixel in feature maps has the most similar relationship through convolution mapping. Since the attention mechanism uses global information more effectively, we utilize it to reduce the number of parameters in the generator decreasing calculations. We first establish the relationship at the element of location y with neighboring elements in the horizontal and vertical directions (the same as the criss-cross non-local module [20] as illustrated in Fig. 5,6) by affinity, and then use a softmax layer to produce the attention map. The affinity is defined as follows:

$$u_{x,y} = Q_y \cdot \alpha_{x,y}^T \tag{3}$$

where $\alpha_{x,y}$ is a vector with the same row and column as Q_y at position y, and $u_{x,y}$ is obtained by the elementwise multiplication of Q_y and $\alpha_{x,y}$ (Q_y and $\alpha_{x,y} \in \mathbb{R}^{(H+W-1)\times(H+W)}$). It is an affinity operation as shown in Fig. 2. The position y is required to calculate attention. In addition, $u_{x,y} \in U$ collects the correlation and then the attention map $M \in \mathbb{R}^{(H+W-1)\times W\times H}$ is updated by U after a softmax layer. $\mathbf{V}$ represents feature adaption, which is generated by the feature map with 1×1convolution layer. The out feature map $\mathrm{fp}\prime$ is produced by aggregation operation as follows:

$$\mathrm{fp}_{y}\prime = \sum_{x\in|V_{y}\prime|} M_{x,y} V_{x,y}\prime + \mathrm{fp} \tag{4}$$

where $V_{x,y}\prime$ is a vector that consists of the row and column elements on y. $M_{x,y}$ is the channel x at position y, and fp is the input feature map. Meanwhile, the attention mechanism includes two convolution layers and two criss-cross attention layers as shown in Fig. 4c, and efficiently enhances content information in LSR and MSR.

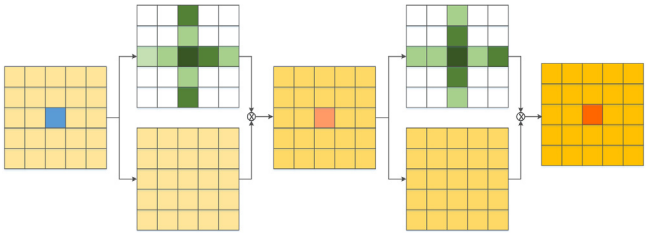

Fig. 5. Illustration of the criss-cross non-local module [20].

3.1.4. Discriminator

The discriminator determines whether the input is real or fake, while the generator tries to cheat the discriminator, whose inputs are the output of the generator and the ground truth image. As for the global and local adversarial learning, the global image and the cropped image are put into the global and local branches, respectively, and then the final layer identifies the authenticity after concatenating them as illustrated in Fig. 6. In the updating phase, the discriminator promotes to alleviate the difference in data distribution between the generated data and the ground truth at each step of the training. In this scenario, the generator and the discriminator are optimized alternately on the fly.

3.2. Loss function

3.2.1. Reconstruction loss

In terms of pixel-level supervision, we use l_1 norm as the reconstruction loss $\mathscr{L}_{\mathrm{R}}$ to measure the distance between the ground truth I_{gt} and the generated image $\mathrm{I}_{\mathrm{out}}$. $\mathscr{L}_{\mathrm{R}}$ is represented by mean absolute error (MAE) as follows:

$$\mathscr{L}_R = \mathscr{L}_{\mathrm{MAE}} = ||\mathrm{I}_{\mathrm{gt}} - \mathrm{I}_{\mathrm{out}}||_1 \tag{5}$$

The MAE loss contributes to pixel-wise correction, but it may fail to capture texture details. Therefore, we introduce the following adversarial, perceptual and edge-preserving losses into the proposed network.

3.2.2. Adversarial loss

In this work, least squares loss [35] updates the discriminator and the generator. LSGAN [35] not only enhances the stability in the training process but also leverages to improve the generator performance with the aid of more gradients. Discriminator needs to distinguish real classes of the generated image and the corresponding ground truth in training discriminator. Thus, the loss function of the discriminator as follows:

$$\mathscr{L}_D^{\mathrm{LSGAN}} = \mathbb{E}_{\mathrm{I}_{\mathrm{gt}}\sim P}\left[\left(D(\mathrm{I}_{\mathrm{gt}}-1)\right)^2\right] + \mathbb{E}_{\mathrm{I}_f\sim Q}\left[\left(D(\mathrm{I}_f-0)\right)^2\right] \tag{6}$$

where $I_f = G(\mathrm{I}_{\mathrm{in}})$, also the same as $\mathrm{I}_{\mathrm{out}}$; P and Q are data distributions of I_{gt} and I_f, respectively. The generator devotes itself to making I_f to close to I_{gt}. Therefore, the adversarial loss function for generator is expressed as follows:

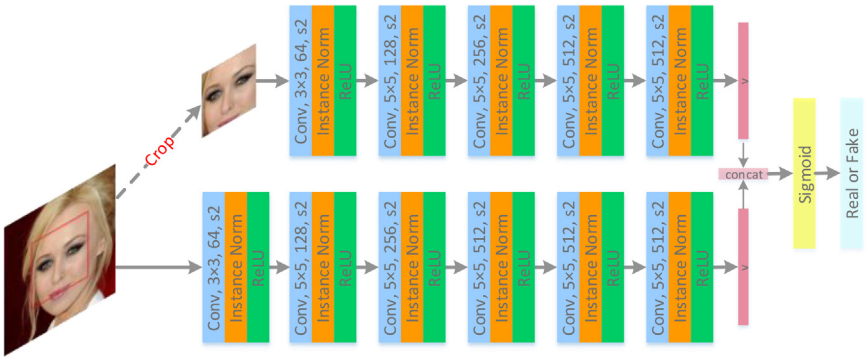

Fig. 6. Network architecture of the proposed discriminator. We combine convolution, instance normalization (IN), and ReLU activation into the discriminator. All convolutional layers are followed by IN and ReLU except the last layer, and sigmod is used in the last layer as activation function.

$$\mathscr{L}_G^{\text{LSGAN}} = \mathbb{E}_{\mathrm{I}_f \sim \mathrm{Q}}\left[\left(D(\mathrm{I}_f) - 1\right)^2\right] \tag{7}$$

3.2.3. Perceptual loss

To capture more contextual texture information, we introduce content loss [36] as perceptual loss from a pre-trained VGG-19 [37] model at the latent space. Specifically, we compute normalized squared Euclidean distance on the feature maps produced by 5 layers, which are $relu1\_1, relu2\_1, relu3\_1, relu4\_1$, and $relu5\_1$, and the loss is defined as follows:

$$\mathscr{L}_{\text{per}} = \mathbb{E}\left\|E\left(\mathrm{I}^l_{\text{out}}\right) - E\left(\mathrm{I}^l_{\text{gt}}\right)\right\|_2^2 \tag{8}$$

where E expresses the average of features, and l is feature maps of l-th layer and the total layers $L = 5$. In a convolutional network, different depth layers have different feature details represent, which is also similar to [38,39].

3.2.4. Edge-preserving loss

Total variation (TV) loss is often used for smoothing the edge, for example, that TV loss function acts as the image fidelity term in Aly et al.'s work [40]. Bastian et al. [41] also used it on the denoising task. However, it cannot be good at the high-quality image since it smoothes textures. To keep the photo-realistic skin and reasonable edge, we introduce edge-preserving loss [42] to fine-tune the generator by learning the relationship of the target and the generated image. For the edge value of each position, i-definition d_i is as follows:

$$d_i(I) = \text{mean}\sum_{j \in P(i)}\left|\sum_c \left(I_{i,c} - I_{j,c}\right)\right| \tag{9}$$

where I is the image to calculated edge; $P(i)$ is the 7×7 patch of the position i in this image; and c is each channel of the RGB color space.

To keep textures in the generated region, the modified edge-preserving loss function needs all edge points instead of important edge points [42], which is defined as:

$$\mathscr{L}_{\text{edge}} = \text{mean}\left\|d(I_{\text{out}}) - d\left(I_{\text{gt}}\right)\right\|_2^2 \tag{10}$$

3.2.5. Total loss function

For generator, we combine the loss functions with different weights as the total loss in the training phase as follows:

$$\begin{aligned}\mathscr{L}_G C &= \mathscr{L}_{\text{content}} + \mathscr{L}_{\text{texture}} \\ &= \left(\lambda_r \mathscr{L}_{MAE} + \lambda_e \mathscr{L}_{\text{edge}}\right) \\ &\quad + \left(\lambda_{Gl}\mathscr{L}_{Gl}^{\text{LSGAN}} + \lambda_{Gg}\mathscr{L}_{Gg}^{\text{LSGAN}} + \lambda_p \mathscr{L}_{\text{per}}\right)\end{aligned} \tag{11}$$

where $\lambda_r, \lambda_e, \lambda_{Gl}, \lambda_{Gg}$ and λ_p are set to $1, 0.001, 0.005, 0.005$, and 15, respectively.

For discriminator, its loss function is simply defined as:

$$\mathscr{L}_D = \left(\lambda_{Dl}\mathscr{L}_{Dl}^{\text{LSGAN}} + \lambda_{Dg}\mathscr{L}_{Dg}^{\text{LSGAN}}\right) \tag{12}$$

where λ_{Dl} and λ_{Dg} are both set 0.5.

4. Experimental results

We evaluate the proposed method on two datasets of CelebA [43] and CelebA-HQ [44]. There are $202,559$ images in CelebA, and we use $162,770$ images for training to optimize parameters, $19,867$ images for validation in the training phase, and $19,962$ images for tests, in which the image resolution is 178×218. As for CelebA-HQ dataset, whose image is 1024×1024, we set $38,000$ images for training and $2,000$ for testing, which is the same as Yu et al.'s work [3]. For all experiments, we train and test the proposed network by 256×256 images with one 128×128 hole at a random position. The input of generator is the masked image $\mathrm{I}_{\text{masked}}$ concatenated with mask I_{mask} in Eq. (2) for both training and testing phases.

4.1. Training details

For training the network, the batch size is 14 with one GTX 1080ti GPU. Each image is resized to 256×256 by bicubic interpolation, and there is no data argumentation. The proposed algorithm is trained with 40 epochs for CelebA dataset and 80 epochs for CelebA-HQ dataset from scratch with the learning rate 2×10^{-4}, which decreases by the factor 0.5 for every 1×10^5 iterations. Adam [45] optimizer acts as an assistant in updating the discriminator and the generator.

4.2. Performance comparison

To show the performance of the proposed network, we perform quantitative measurements, visual comparison and subjective evaluations. For comparison, baselines are as follows:

- GFC [4]: a method focusing on face completion. Since GFC is only trained and tested on CelebA dataset, we compare the proposed network with GFC on it.
- CA [3]: an inpainting method with two-stage via contextual attention in feature maps. We compare the proposed network with CA on CelebA and CelebA-HQ datasets. Moreover, we perform subjective evaluations on CelebA-HQ dataset because the data initialization is different on CelebA dataset with the other methods.

- GMCNN [18]: a multi-column network that uses implicit diversified Markov random field (ID-MRF) term as regularization. We perform the comparison on CelebA-HQ dataset.
- MLGN [19]: a multi-level generative network via fusing the features with three resolutions. Multi-level branches enhance different kinds of feature information.

For a fair comparison, we choose the same datasets and mask the same region for each facial image. Moreover, we train and test the pre-trained models for GFC and MLGN on the images which are directly cropped to 128×128 resolution and resized to 256×256. Thus, we retrain MLGN, named MLGN-Ori, as shown in Fig. 7. However, the retrained GFC does not perform better than the pre-trained model, and thus we use the original model for tests.

4.2.1. Quantitative comparison

For quantitative evaluation, we introduce PSNR, structural similarity (SSIM), Fréchet Inception Distance (FID) [46], and Learned Perceptual Image Patch Similarity (LPIPS) [47] as evaluation metrics. PSNR and SSIM measure content and structure reconstructions, respectively, while FID evaluates the similarity of data distributions between the completed images and the ground truth by a pre-trained Inception-V3 model [48]. LPIPS computes the perceptual similarity between the generated and target images. We provide the evaluation results in Tables 1 and 2. We use the original model of MLGN [19] without retraining. From the tables, it can be observed that the proposed network achieves competitive performance in facial image inpainting. However, MLGN [19] obtains the best score in LPIPS metric on CelebA dataset. For MLGN, we

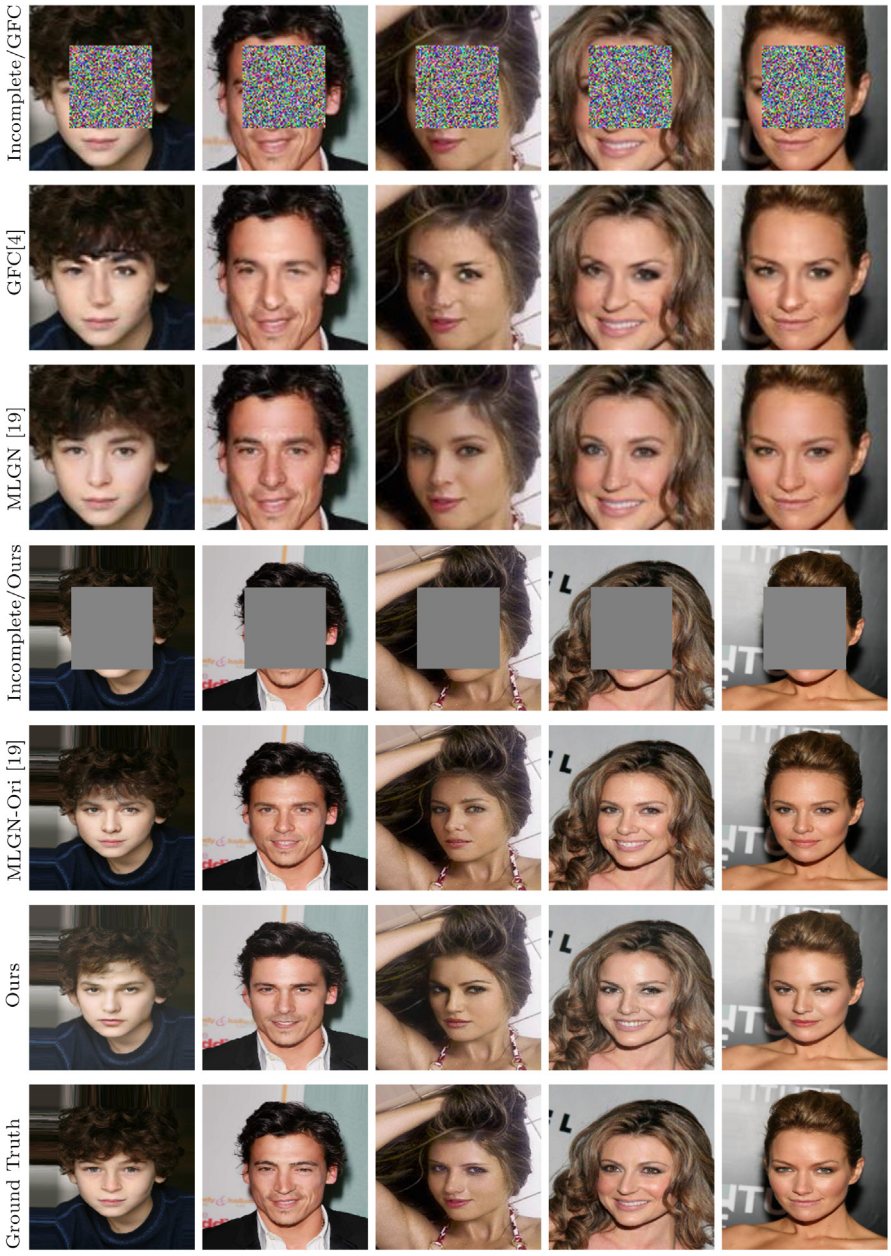

Fig. 7. Experimental results on CelebA Val dataset. MLGN: MLGN trained and tested with 128×128 and resized to 256×256 resolution. MLGN-Ori: MLGN [19] trained and tested with 178×218 and then cropped and resized to 256×256 resolution. In Incomplete/Ours, the masked images are the input for MLGN-Ori and the proposed method.

Table 1
Performance comparison between different methods with the center position mask in terms of PSNR, SSIM, FID, and LPIPS. For PSNR and SSIM the higher the better, while for FID and LPIPS the lower the better. Here, MLGN-Ori [19] is the retrained MLGN with the resolution of 256 × 256, and MLGN is tested and trained with the same resolution as MLGN-Ori.

Test Datasets	CelebA Test			
Method	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
GFC [4]	26.43	0.9191	-	0.0732
MLGN [19]	27.99	0.9302	-	0.0563
CA [3]	25.24	0.8849	5387	0.0732
MLGN-Ori [19]	27.23	0.9120	3.36	**0.0393**
Ours	**27.48**	**0.9329**	**3.21**	0.0452
Test Datasets		CelebA Val		
GFC [4]	26.46	0.9192	-	0.0736
MLGN [19]	28.03	0.9301	-	0.0567
CA [3]	25.24	0.8844	5.84	0.0742
MLGN-Ori [19]	27.24	0.9115	3.27	**0.0398**
Ours	**27.46**	**0.9328**	**3.15**	0.0455

Table 2
Performance comparison among different methods in terms of PSNR, SSIM, FID, and LPIPS on CelebA-HQ dataset. We test the incomplete image with one center 128 × 128 mask. The runtime means average time for test images (unit: ms).

Test Dataset	CelebA-HQ				
Method	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓	Runtime
CA [3]	23.98	0.8589	13.69	0.0752	49
GMCNN [18]	25.32	0.8849	7.48	0.0518	33
MLGN [19]	26.26	0.9004	8.92	0.0471	24
Ours	**27.18**	**0.9202**	**7.14**	**0.0416**	16

directly crop 128 × 128 resolution on each image, thus leading to small face occlusion. As mentioned earlier, we resize the original images as shown in Fig. 7, and use them for tests, called MLGN-Ori. That is, the face occlusion is large for MLGN-Ori and the proposed method. Thus, MLGN achieves higher PSNR than MLGN-Ori and the proposed method as shown in Table 1. However, the test results by MLGN-Ori and the proposed method provide performance comparison in the same testing condition. We provide the runtime comparison in Table 2. As shown in the figure, the runtime is about 16 ms for a 256 × 256 facial image by the proposed method, while it is about 24 ms by MLGN. The proposed method is faster than MLGN in the same resolution of test images. Since the proposed method combines multi-level generator with attention, it has less layers than MLGN. For each layer, it has the same kernel size, thus resulting in less parameters. In addition, the sizes of feature maps are the same for the proposed method and MLGN, thus leading to less parameters and FLOPs.

4.2.2. Visual comparison

The inpainting results on CelebA and CelebA-HQ datasets are shown in Figs. 7 and 8, respectively. Even in the same resolution of facial images, the quality of CelebA-HQ dataset is better than that of CelebA dataset due to the camera performance. Thus, the quality improvement by the proposed method is more obvious on CelebA-HQ dataset (Fig. 8) than CelebA (Fig. 7), e.g. eyes and noses of the last column. In addition, the proposed method keeps the consistency between the generated region and the background. In CelebA dataset, the training data size for the proposed method is smaller than that for MLGN. It is because the training data for MLGN contain a big size of faces and we remove them for training. We retrain MLGN on the reduced training data for the proposed method. Moreover, only a GAN loss is used for MLGN because it is good for image generation. However, there are some object boundaries in the masks, which cannot be recovered by the GAN loss. Thus, we put the global images and the cropped images into the global and local branches in the discriminator, respectively. As shown in the figures, the proposed method successfully fills the masks in the textural area located at the mask boundaries, thus producing plausible and realistic results in visual quality. It can be observed that if one eye is in the hole, the other eye is produced with the same color and shape. For the lost nose and mouth, the proposed method produces photo-realistic and natural-looking facial images. Besides, the proposed method keeps the consistency between the generated region and the background. Thus, the proposed method produces clear eyes and makes beard and hair of the man natural.

4.2.3. Subjective evaluation

In terms of the perceptual quality, we perform subjective evaluations on the Google Forms platform and the results are shown in Table 3 and Fig. 9. We select 20 pairs of each dataset for comparison, where each pair contains two generated images by two different methods and the incomplete inputs are with the same masked region. Then, there are 20 volunteers randomly selected to vote the more photo-realistic and natural image from each pair. We shuffle the order in each pair of images. It can be concluded from Table 3 and Fig. 9 that the proposed approach achieves a positive impact on perceptual quality. As shown in Figs. 10 and 11, our method produces good inpainting results at random mask and multiple holes.

4.3. Ablation study and analysis

To show the effectiveness of each component in the proposed method such as multi-level structure and loss function, we perform ablation study on CelebA-HQ dataset on the network architecture and loss function. Moreover, we test the different impacts of the content loss and the perceptual loss on image inpainting.

4.3.1. Only main and middle branches

We remove the low spatial resolution (LSR) branch to verify the contribution of the LSR branch to the proposed network. Theoretically, losing the LSR branch reduces global visual quality. Table 4 shows that the absence of LSR slightly degrades the performance in all metrics and greatly influences PSNR performance.

Fig. 8. Experimental results on CelebA-HQ dataset.

Table 3
Subjective evaluation results. The values represent the percentage that the proposed method is better.

	CelebA-Test	CelebA-Val	CelebA-HQ
Ours > CA [3]	–	–	100%
Ours > GMCNN [18]	–	–	74%
Ours > MLGN-Ori [19]	81.75%	82%	84%

4.3.2. Only main and low branches

We remove the middle spatial resolution (MSR) branch to see the contribution of the MSR branch to the proposed network. MSR is good for the visual quality, especially in image structure, but SSIM is lower without MSR.

4.3.3. Only main branch

The resolution of the least feature maps is 32×32 in the main branch. Without LSR and MSR, the network is a simple encoder-decoder architecture for image inpainting. It is worse in each metric than the network without LSR or MSR. Therefore, both LSR and MSR are useful for facial image completion.

4.3.4. Without attention mechanism

We only perform experiments by the proposed network without criss-cross non-local attention (CCA) blocks. As shown in Table 4, all values of the proposed network without CCA are the worst in performance. The results indicate that attention has the most significant impact on performance. This is because the attention mechanism is able to establish an effective pixel correlation between feature maps.

4.3.5. Without edge-preserving loss

The edge-preserving loss keeps the local structure while removing the boundaries in the generated images caused by the masked boundary, thus making results more natural. In Table 5, SSIM is greatly affected but the others are only little affected. In Figs. 7 and 8, there is no unnatural boundary in the masked edge in our results. Therefore, the edge-preserving loss is beneficial to the structure of the generated images.

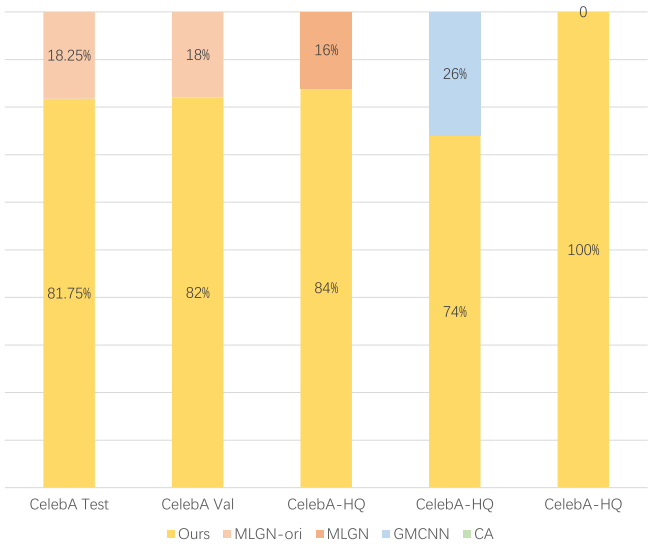

Fig. 9. Subjective evaluation results. The values represent the percentage of each method chosen as the best.

4.3.6. Without perceptual loss

The ablation study with and without perceptual loss is shown in Table 5, where each metric value without perceptual loss is higher than the one without edge-preserving loss. Thus, the impact of edge-preserving loss is more than that of perceptual loss on the performance.

4.3.7. Without perceptual and edge-preserving losses

When we use only MAE and adversarial losses without perceptual and edge-preserving ones to optimize the network, the performance is the worst as shown in Table 5.

4.3.8. Style loss instead of perceptual loss

The style loss [49] has been used in MLGN [19] from a pre-trained VGG-19 [37] model the same as the Zhou et al.'s work [38]. Specifically, we compute gram matrices on the feature maps produced by 5 layers, which are $relu1\_1, relu2\_1, relu3\_1, relu4\_1$, and $relu5\_1$. The style loss is defined as follows:

$$\mathscr{L}_{\text{style}} = \sum_{l=1}^{L} w_l \left\| \text{Gram}(\mathrm{I}_{\text{out}}) - \text{Gram}(\mathrm{I}_{\text{gt}}) \right\|_2^2 \tag{13}$$

where w_l is the weight of the style loss about the nth layer, and w_l = [0.244, 0.061, 0.015, 0.004, 0.004], which is also similar to [38,39]. Gram matrix is calculated by the Hermitian matrix of inner products, and thus it can be expressed $\text{Gram}(F) = F_l^T F_l$, where F_l is feature maps of the lth layer. The weight of the style loss for total loss function is 100 when using the style loss instead of the perceptual loss, whose weight is adjusted according to the value of each loss. By observing the experimental results of MLGN [19], there are some unnatural distortions in the generated images as shown

Fig. 10. Facial inpainting results on CelebA-HQ dataset with multiple holes by the proposed method.

Fig. 11. Facial inpainting results with a random mask on CelebA-HQ dataset. The odd rows are the input and the even rows are the inpainting results by the proposed method.

Table 4
Ablation study of the network architecture on CelebA-HQ dataset in terms of PSNR, SSIM, and LPIPS. We use incomplete images with one center 128 × 128 hole for the testing phase. LSR: low spatial resolution. MSR: middle spatial resolution. CCA: criss-cross non-local attention.

Strategy	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
w/o LSR	27.08	0.9200	7.22	0.0420
w/o MSR	27.10	0.9195	7.19	0.0435
only main branch	27.05	0.9182	7.83	0.0445
w/o CCA	25.18	0.8920	7.36	0.0591
All	**27.18**	**0.9202**	**7.14**	**0.0416**

Table 5
Ablation study of loss functions in terms of PSNR, SSIM, and LPIPS on CelebA-HQ dataset. We test the incomplete images with one center 128 × 128 hole.

Strategy	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
w/o edge	25.69	0.8698	8.09	0.0603
w/o perceptual	25.98	0.9097	7.48	0.0550
w/o edge and perceptual	23.73	0.8685	9.82	0.0868
Style instead of perceptual	26.49	0.9137	7.16	0.0461
All	**27.18**	**0.9202**	**7.14**	**0.0416**

in Figs. 7 and 8. Accordingly, we use content loss as perceptual loss instead of style loss. We experiment content loss and style loss separately, and the results are shown in the last two rows in Table 5. As shown in the table, they achieve similar performance in FID, but the style loss degrades the performance in the other metrics.

4.4. Analysis and discussion

Fig. 10 provides some inpainting results generated by the proposed method whose inputs are with multiple holes. When one eye is occluded but the other is not occluded, the synthesized output keeps natural and photo-realistic. The generated key components (nose, eyebrow, and mouth) are plausible as shown in Fig. 11, where the masked region is at random position with 128 × 128. The results indicate that the proposed method is effective for facial image inpainting in the network architecture and loss function.

However, there are some limitations in the proposed method. In the fourth column of the first and second rows in Fig. 10, the generated edges of the hat are unnatural and the synthesized hair is not similar to the original hair of the input image. It seems that the training dataset is not able to cover them for inpainting.

5. Conclusion

In this paper, we have proposed facial image inpainting using attention based multi-level generative network. For generator, we have combined the attention mechanism with multi-level fea-

ture processing to keep soft correlation with the surrounding content reducing the number of parameters. For discriminator, we have utilized global and local branches to distinguish real and fake images. Last but not least, we have introduced edge-preserving loss to hold local edge consistency and effectively learn it for facial image inpainting. Experimental results demonstrate that the proposed method outperforms state-of-the-art ones in terms of quantitative measurements and subjective evaluations. Various ablation studies further verify the effectiveness of the multi-level generative network, attention, and loss functions.

CRediT authorship contribution statement

Jie Liu: Data curation, Methodology, Software, Writing - original draft. **Cheolkon Jung:** Supervision, Conceptualization, Visualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61872280) and the International S&T Cooperation Program of China (No. 2014DFG12780).

References

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
[2] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, H. Li, High-resolution image inpainting using multi-scale neural patch synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6721–6729.
[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514.
[4] Y. Li, S. Liu, J. Yang, M.-H. Yang, Generative face completion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3911–3919.
[5] X. Niu, B. Yan, W. Tan, J. Wang, Effective image restoration for semantic segmentation, Neurocomputing 374 (2020) 100–108.
[6] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural network-driven feature learning method for multi-view facial expression recognition, IEEE Trans. Multimedia 18 (12) (2016) 2528–2536, https://doi.org/10.1109/TMM.2016.2598092.
[7] H. Li, J. Sun, Z. Xu, L. Chen, Multimodal 2d+3d facial expression recognition with deep fusion convolutional neural network, IEEE Trans. Multimedia 19 (12) (2017) 2816–2831, https://doi.org/10.1109/TMM.2017.2713408.
[8] S. Agarwal, D.P. Mukherjee, Synthesis of realistic facial expressions using expression map, IEEE Trans. Multimedia 21 (4) (2019) 902–914, https://doi.org/10.1109/TMM.2018.2871417.
[9] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, Image inpainting, in: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, 2000, pp. 417–424.
[10] D. Liu, X. Sun, F. Wu, S. Li, Y.-Q. Zhang, Image compression with edge-based inpainting, IEEE Trans. Circ. Syst. Video Technol. 17 (10) (2007) 1273–1287.
[11] S. Esedoglu, J. Shen, Digital inpainting based on the mumford–shah–euler image model, Eur. J. Appl. Math. 13 (4) (2002) 353–370.
[12] C. Barnes, E. Shechtman, A. Finkelstein, D.B. Goldman, Patchmatch: a randomized correspondence algorithm for structural image editing, ACM Trans. Graph. 28 (3) (2009) 24:1–24:11..
[13] S. Darabi, E. Shechtman, C. Barnes, D.B. Goldman, P. Sen, Image melding: combining inconsistent images using patch-based synthesis, ACM Trans. Graph. 31 (4) (2012), 82–1.
[14] J.-B. Huang, S.B. Kang, N. Ahuja, J. Kopf, Image completion using planar structure guidance, ACM Trans. Graph. 33 (4) (2014) 129.
[15] J. Hays, A.A. Efros, Scene completion using millions of photographs, ACM Trans. Graph. 26 (3)..
[16] S. Zhang, R. He, Z. Sun, T. Tan, Demeshnet: blind face inpainting for deep meshface verification, IEEE Trans. Inf. Forensics Secur. 13 (3) (2018) 637–647.
[17] L. Song, J. Cao, L. Song, Y. Hu, R. He, Geometry-aware face completion and editing 33 (2019) 2506–2513.
[18] Y. Wang, X. Tao, X. Qi, X. Shen, J. Jia, Image inpainting via generative multi-column convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2018, pp. 329–338.
[19] J. Liu, C. Jung, Facial image inpainting using multi-level generative network, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2019, pp. 1168–1173.
[20] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: criss-cross attention for semantic segmentation, arXiv preprint arXiv:1811.11721..
[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[22] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434..
[23] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: unsupervised dual learning for image-to-image translation, Proceedings of the IEEE Conference on Computer Vision (2017) 2868–2876.
[24] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, Y.-Y. Chuang, Deep photo enhancer: unpaired learning for image enhancement from photographs with gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6306–6314.
[25] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.
[26] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, in: arXiv:1701.07875, 2017..
[27] X. Chen, X. Wang, Y. Lu, W. Li, Z. Wang, Z. Huang, Rbpnet: an asymptotic residual back-projection network for super-resolution of very low-resolution face image, Neurocomputing 376 (2019) 119–127.
[28] Z. Yan, X. Li, M. Li, W. Zuo, S. Shan, Shift-net: image inpainting via deep feature rearrangement, Proceedings of the European Conference on Computer Vision (2018) 1–17.
[29] Y. Liu, Z. Long, C. Zhu, Image completion using low tensor tree rank and total variation minimization, IEEE Trans. Multimedia 21 (2) (2019) 338–350, https://doi.org/10.1109/TMM.2018.2859026.
[30] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[31] A.A. Efros, T.K. Leung, Texture synthesis by non-parametric sampling, Proceedings of the IEEE Conference on Computer Vision (1999) 1033.
[32] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Trans. Graph. 36 (4) (2017) 107:1–107:14..
[33] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[34] Y. Wu, Y. Qin, Z. Wang, X. Ma, Z. Cao, Densely pyramidal residual network for uav-based railway images dehazing, Neurocomputing 371 (2020) 124–136.
[35] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, Proceedings of the IEEE Conference on Computer Vision (2017) 2813–2821.
[36] A. Sanakoyeu, D. Kotovenko, S. Lang, B. Ommer, A style-aware content loss for real-time hd style transfer, Proceedings of the European Conference on Computer Vision (2018) 698–714.
[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1701.07875, 2014..
[38] Y. Zhou, Z. Zhu, X. Bai, D. Lischinski, D. Cohen-Or, H. Huang, Non-stationary texture synthesis by adversarial expansion, ACM Trans. Graph. 37 (4) (2018) 49:1–49:13..
[39] L.A. Gatys, A.S. Ecker, M. Bethge, A neural algorithm of artistic style, arXiv preprint arXiv:1508.06576..
[40] H.A. Aly, E. Dubois, Image up-sampling using total-variation regularization with a new observation model, IEEE Trans. Image Process. 14 (10) (2005) 1647–1659.
[41] B. Goldluecke, D. Cremers, An approach to vectorial total variation based on geometric measure theory, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 327–333.
[42] Q. Fan, J. Yang, D. Wipf, B. Chen, X. Tong, Image smoothing via unsupervised learning, ACM Trans. Graph. 37 (6) (2018) 259:1–259:14..
[43] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, Proceedings of the IEEE Conference on Computer Vision (2015) 3730–3738.
[44] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: Proc. ICLR, 2018..
[45] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, CoRR abs/1412.6980.
[46] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6626–6637.
[47] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[49] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, Proceedings of the European Conference on Computer Vision (2016) 694–711.

Jie Liu received the B.S. degree in electronic engineering from Xidian University, China, in 2017. She is currently pursuing the master degree in the same university. Her research interests include image processing and computer vision.

Cheolkon Jung is a Born Again Christian. He received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sungkyunkwan University, South Korea, in 1995, 1997, and 2002, respectively. He was a Research Staff Member with Samsung Advanced Institute of Technology, Samsung Electronics, South Korea, from 2002 to 2007. He was also a Research Professor with the School of Information and Communication Engineering, Sungkyunkwan University, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently a Full Professor and the Director of the Xidian Media Laboratory. His main research interests include image and video processing, computer vision, pattern recognition, machine learning, computational photography, video coding, virtual reality, information fusion, multimedia content analysis and management, and 3DTV.