IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 9, 2019, accepted October 22, 2019, date of publication October 25, 2019, date of current version November 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949614

Face Inpainting via Nested Generative Adversarial Networks

ZHIJIANG LI[1], HAONAN ZHU[1], LIQIN CAO[1], LEI JIAO[2], (Senior Member, IEEE), YANFEI ZHONG[3], AND AILONG MA[3]

[1]School of Printing and Packaging, Wuhan University, Wuhan 430072, China
[2]Department of Information and Communication Technology, University of Agder, 4879 Grimstad, Norway
[3]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China

Corresponding author: Liqin Cao (clq@whu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB10046 and Grant 2017YFB0504202, in part by the Fundamental Research Funds for the Central Universities under Grant 2042018kf0229, and in part by the National Natural Science Foundation of China under Grant 41671441.

ABSTRACT Face inpainting aims to repaired damaged images caused by occlusion or cover. In recent years, deep learning based approaches have shown promising results for the challenging task of image inpainting. However, there are still limitation in reconstructing reasonable structures because of over-smoothed and/or blurred results. The distorted structures or blurred textures are inconsistent with surrounding areas and require further post-processing to blend the results. In this paper, we present a novel generative model-based approach, which consisted by nested two Generative Adversarial Networks (GAN), the sub-confrontation GAN in generator and parent-confrontation GAN. The sub-confrontation GAN, which is in the image generator of parent-confrontation GAN, can find the location of missing area and reduce mode collapse as a prior constraint. To avoid generating vague details, a novel residual structure is designed in the sub-confrontation GAN to deliver richer original image information to the deeper layers. The parent-confrontation GAN includes an image generation part and a discrimination part. The discrimination part of parent-confrontation GAN includes global and local discriminator, which benefits the reconstruction of overall coherency of the repaired image while obtaining local details. The experiments are executed over the publicly available dataset CelebA, and the results show that our method outperforms current state-of-the-art techniques quantitatively and qualitatively.

INDEX TERMS Face inpainting, deep neural network, nested GAN.

I. INTRODUCTION

Face inpainting is a challenging task of recovering details of facial features on high-level image semantics. It can be applied in many face recognition occasions, such as wearing sunglasses, microphone occlusion during performance, and covering mask. The purpose of inpainting technology is to repair the broken part of the image with known image information. The most important goal of this task is to avoid introducing noise into non-repaired areas and to generate reliable repaired areas. Based on this technique, noise, hiatus and scratch can be removed.

Because of the strong correlation between pixels in one image, lost image information can be restored as much as possible based on undamaged or occluded area of the image and its pattern priori. During inpainting process, the content information of the whole image is considered, including low-level texture information and high-level semantic information. Traditional inpainting methods rely on low level cues to find best matching patches from the uncorrupted sections in the same image [1]–[3]. These methods work well for background completions and repetitive texture pattern. However, low level features are limited for face inpainting task as face image consists of many unique components, and inpainting process needs to be carried out with a high-level semantic level [4]–[6]. The traditional methods based on finding patches with similar appearance patches does not always perform well.

Rapid progress in deep convolutional neural networks (CNN) and generative adversarial networks (GAN) [7]

The associate editor coordinating the review of this manuscript and approving it for publication was Chunbo Xiu.

inspired lots of studies [6], [8]–[10] to restore damaged images. The GAN model [6], [10], [11], [43] is proposed to deal with both low-level textural features and high-level semantic features, which can complete the blanks in the images. However, one of the essential challenges about inpainting via GAN model is that the reconstructed area is blurry compared to global image [11]. The reason is that the output of model approximates to the global loss minimum, which will make intensity of output vague. To tackle this problem, a complete training framework based on nested generator adversarial network (NGAN) is proposed in this paper. This generation network includes a sub-confrontation GAN and a parent-confrontation GAN. Applying sub-confrontation GAN, the location of missing area is found and rough result is obtained. To avoid the loss of defect area information and the degradation of the GAN, our model adopts residual structure to jointly transmit features in different layers to a deeper network. In order to solve the puzzle of ambiguity of repairing region, a special residual transfer connection is utilized for four times in sub-confrontation generation network, which can reduce loss in convolutional network transmission process. In the parent confrontation GAN, the global and local discriminators are combined to capture both local continuity of image texture and pervasive global features in images, which aims to achieve high-quality local repair area and overall coordination.

We evaluate our method using CelebA [12] dataset compared with other state-of-the-art methods. The contributions of our work are summarized as follows:

- A *NGAN based framework* is proposed for face inpainting, which is a combination of a sub-confrontation and a parent-confrontation network. The networks produce a priori semantic constraint to reduce model collapse.
- A *novel residual connection structure* is introduced in sub-confrontation generation network, which is beneficial to generate high-quality details for facial image with mask and eliminate ambiguity.
- *Local discriminator* and *global discriminator* are combined in our framework, which can ensure global consistency of inpainting results and guarantee the details of the local inpainting area.

The remaining of the paper is organized as follows. Section II presents a short review of relevant and recent image inpainting techniques. The details of NGAN method are presented in Section III. Section IV shows the experimental results before we conclude the paper in Section V.

II. RELATED WORK

As an important branch of digital image processing, the research of image inpainting is extensive. Methods for image inpainting fall mainly into two categories: copy-paste and learning-based.

Copy-paste inpainting methods are based on the information relations between damaged areas and known areas in the image and migrates the surrounding information to the blank area. The idea of diffusion model is to iteratively propagates the underlying texture information of known image areas to damaged unknown areas [1]. The basic principle of this type of models is from the thermal diffusion equations in physics [2], [13]–[15]. Another type of inpainting approaches based on geometric image variational model imitates the process of image restoration by hand [16]–[19]. During the processing of this method, the universal function is determined based on the data prior distribution, and the defect area is repaired using the established model. These copy-paste image restoration techniques have achieved good results in smooth and continuous small-scale damaged images. However, when the loss area is large-scale, or the texture is rich and complex, the diffusion or image data model will not be able to accurately describe the lost information, resulting in unnatural and unclear results.

To solve the above problem, texture synthesis technology was presented [20]. The texture blocks with appropriate size are determined, and the missing area is synthesized by the similarity of blocks texture. Image energy optimization [1], [3], [21], [22] was introduced to measure texture proximity, and image gradient was integrated to the distance measurement between reconstructed texture [23]. Texture measurement was extended to include image segmentation and texture generation [24]. This type of approaches can improve efficiency and achieve a real-time image restoration through the patch-match algorithms. In addition, some methods for automatically estimating the structure of the scene have also been proposed [25]–[29]. These methods improve the quality of image completion by preserving important structures, such as points of interest [30], lines [31] and perspective distortion [32]. However, the image structure guidance is a heuristic constraint based on a particular type of scenes, and it is limited to a specific structure. For different images, distinct guiding rules of image results need to be designed, and these rules cannot be applied to arbitrary images. Besides, these approaches are difficult to reconstruct semantic information because they only fix the underlying texture.

Although copy-paste methods have a good performance in image restoration, it is difficult to produce textures that are not in the original picture. In order to obtain more information, a large images database was used [33]. However, compared with the general method, the premise that the database contains a large number of similar or same scenarios greatly limits its applicability.

With the development of deep neural network, deep-learning based methods are introduced to predict the unavailable content and achieve semantic inpainting results. The convolutional neural network-based image inpainting method [8] can obtained pleasing result for small occluded areas. It was applied to repair missing data from MRI and PET [9]. Generative Adversarial Net (GAN) based on dualistic game theory combined with convolutional network [6], [10], [34]–[36] could bring out very real impaired images. In these networks, the input image data includes mask areas which are to be repaired. These mask areas must be manually annotated in the real word. To address this time consumption limitation, a novel end-to-end network is proposed in [6], [43], which doesn't need an additional mask as the input information.

Although deep-learning approaches consider both content texture and semantic feature and have a good performance in image inpainting, some features are easily lost, resulting in reconstructing unreasonable structures, such as over-smoothed and/or blurry [12]. Especially, the distorted structures or blurry textures inconsistent with surrounding areas will be produced [36]. In this paper, we propose NGAN for semantic face inpainting. In our model, a nested GAN structure is introduced to constraint generation process and reduce noise introduction. A new residual connection is constructed to transmit missing information caused by network forward propagation process to deeper layers. The global and local discriminators are combined to reconstruct the overall coherency image and to obtain local details.

III. APPROACHES

A. GAN REVIEW

GAN model was proposed by Goodfellow *et al.* [7], which consists of two parametrized deep neural nets: generator, G, and discriminator, D. G maps a random vector z, sampled from a prior distribution p_z, to the image space while D maps an input image to a likelihood. The target of G is to produce images that are realistic enough, while D discriminates between the image generated from G, and the real image, x, sampled from the data distribution p_{data}.

The G and D networks are trained by optimizing the loss function:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}}(x)[\log D(x)] + \mathbb{E}_{z\sim p_z}(z)[1 - D(G(z))]. \quad (1)$$

The generator is trained to acquire minimum loss while the discriminator is trained to acquire maximum loss. The loss eventually approaches 0.5 when the training process finishes.

B. NESTING STRUCTURE OF GAN

GAN is an unsupervised learning model, which can generate clear and realistic images [6]. We introduce a generative CNN model and a training procedure for the hole filling in face images problem. Our network consists of a nested structure including two different generation networks, which are called sub-confrontation generation network and parent-confrontation generation network.

The sub-confrontation generation network identifies the location of image defects, which can preserve the original information in the non-repaired area of the image. After the confrontation training, code generator can produce robust coding information, which will be decode to generate output image. In addition, the residual structure and the dilated convolutional structure are adopted in code generator of sub-confrontation generation network to improve the local details of the output image. Meanwhile, the coding information are used as a priori semantic constraint to reduce model collapse.

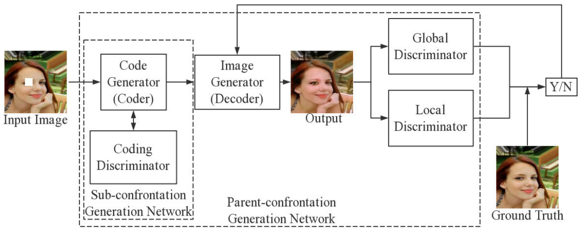


FIGURE 1. The sub-confrontation generation network consisted by a code generator and a code discriminator. The parent-confrontation generation network has two parts: generation part and discrimination part. The output of the global and local discriminators are fed back into the image generator. The output of the coding discriminator is fed back into the code generator.

The parent-confrontation generation network has two parts: generation part and discrimination part. The parent-confrontation generation network takes the corrupted image and tries to reconstruct the repaired image. The generation part uses the coding information of sub-confrontation generation network to recover the input image through multiple convolutional layers. Unlike traditional networks, the discrimination part consists two different scales discriminators. The overall structure of our framework is shown in Fig. 1.

1) SUB-CONFRONTATION GENERATION NETWORK

The sub-confrontation generation network is consisted by code generator and code discriminator, and the corruption image z is used as the input. After antagonistic training, the code generator produce code information, z', which is judged by code discriminator to be the same classification as ground truth coding. This network can obtain the ability to extract the robust features of the damaged image. Furthermore, it is also a prior constraint on the image generator, which effectively reduces the collapse of the generator pattern.

The code generator is trained with an additional code discriminator and can learn the features of the occlusive image and retain the semantic information of the original image as much as possible during the coding process. Code generator and code discriminator form an antagonistic structure and are iterated alternately until obtaining consistent coding for the corruption image and the corresponding ground truth.

The code generator consists 5 convolutional layers and 5 dilated convolutional layers. Dilated convolution can increase the receptive field of the network without increasing the number of model parameters [37], which will be analyzed in detail in III-D. To avoid losing information in calculating, a specially designed residual connection is applied, which transports original information to deep layers. The novel residual connection structure will be described in III-C. The code discriminator consists of three convolutional layers and one fully connected layer. The structure of sub-confrontation generation network is shown in Fig. 2.

2) PARENT-CONFRONTATION GENERATION NETWORK

The parent-confrontation generation network has two parts: generation part and discrimination part. The generation part

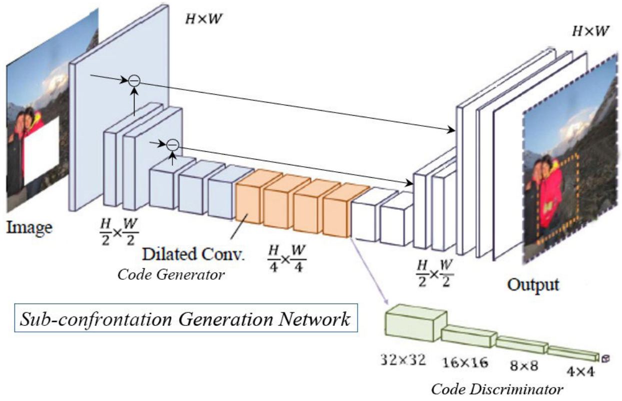

H×W
H×W
Image
$\frac{H}{2}\times\frac{W}{2}$
Dilated Conv.
$\frac{H}{4}\times\frac{W}{4}$
Code Generator
$\frac{H}{2}\times\frac{W}{2}$
Output
Sub-confrontation Generation Network
32×32
16×16
8×8
4×4
Code Discriminator

FIGURE 2. The structure of sub-confrontation generation network consisted by code generator and code discriminator.

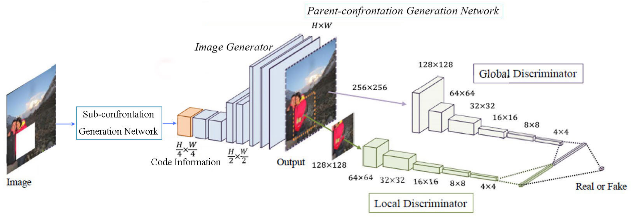


FIGURE 3. The parent-confrontation generation network in our framework consists of a complex generator and two discriminators. The defective image is as the input of the generation part. The semantic repaired is carried out by the network and the result of reconstruction is the output, which will be evaluated by the discrimination part.

is composed of a code generator of sub-confrontation generation network and an image generator. After coding process, the image generator reconstructs broken image from code information. Using the encoded information as input instead of the image directly can improve the robustness of our model and reduce model collapse.

The discriminator part consists of a global discriminator and a local discriminator. The global discriminator judges the authenticity of the whole image and enforce global consistency on a large scale. Different from global discriminator, the local discriminator only constrains the richness of image detail information and local coherency. Both discriminator networks have similar network structures, which are spliced together and produced by the fully connected hierarchy. The generation part and the discrimination part form a confrontation structure. Through the confrontation training, the generation part can reconstruct pleasing image. The structure of the parent-confrontation generation network is shown in Fig. 3.

C. NOVEL RESIDUAL CONNECTION STRUCTURE

When the information is propagated forward between layers, the size of the feature map decreases by the convolution kernel with stride 2 or larger, which will result in losing of detail texture information and degradation of the generated image. In addition, using the activation function will lose the information of original image. For example, for a single image x_0 through a convolutional network, the previous convolutional feed-forward network connects the output of the l^{th} layer to the $(l+1)^{th}$ layer, which applies the following layer transition: $x_{(l+1)} = H_l(x_l)$. In order to achieve sparse network connections and avoid losing negative value, we usually use leaky ReLU function, which reduces the negative value response of the former feature map. However, leaky ReLU function does not completely reflect the impact of information loss on image generation.

To tackle this limitation, our method takes advantage of the original available data using novel residual connection structure. Residual network structure was proposed in [38], which added a skip-connection that bypassed the non-linear transformations with an identity function:

$$x_{l+1} = H_l(x_l) + x_l, \tag{2}$$

where x_l is output of the l^{th} layer, $x_{(l+1)}$ is output of the $(l+1)^{th}$ layer, $H_l(\cdot)$ is Leaky ReLU function. Using residual block structure, the original information can be delivered to deep layer, which cannot only retain details of the input image but also avoid introducing noise. In [39], an improvement was made to reduce the number of residuals and improve network performance. However, these two residual connections can only transport the original information directly. This will transfer all the information from the damaged area of the image to the deeper layers and degrades the quality of the resulting image. Therefore, it is necessary to change the residual connection structure in order to pass only valid information. In our approach, we improved the structure in [39] and change the residual structure as follow:

$$x_{l+1} = K_l(x_l) + \phi_l(x_l), \tag{3}$$

where $K_l(x_l)$ is processing function in l^{th} layer, which is down-sample operation by convolutional layers and pooling layers in our network; $\phi_l(\cdot)$ is a robust information extractor for the output of l^{th} layer, which can filter out the original information lost through layers. The extractor function $\phi_l(\cdot)$ is defined:

$$\phi_l(t) = t - K_l^*(K_l(t)), \tag{4}$$

where $K_l^*(\cdot)$ is an up-sample operation by single convolutional layer structure. The original information t is implicitly and adaptively transported to $K_l^*(K_l(t))$ and the interpolation between feature map in deeper layer and feature map in shallow layer is the missing information in feed forward network. By transmitting the missing information $\phi_l(t)$ to deeper layers, our approach takes advantage of more primitive and useful semantic information as well as the ability to generate reliable results.

We compared different residual connection patterns and presented the experimental results in Fig.4. The result illustrates that our method is outperform than those in [38] and [39]. The inpainting images based on our connection pattern have clear details of the left eye and good consistency with the right eye, and the details of results based on the other

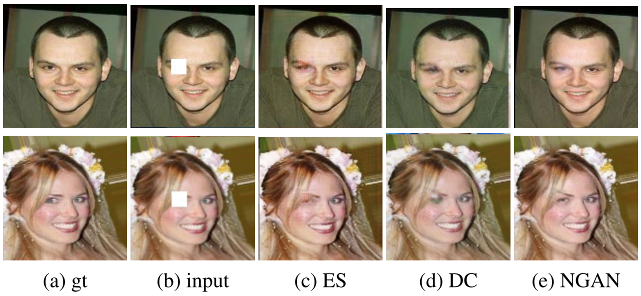


FIGURE 4. Result images using different residual structure. Left to right: (a) the ground truth from CelebA, (b) input images, (c) results based on Element-wise Sum [38], (d) results based on Depth Concatenation [39], and (e) results based on our proposed approach.

two methods are obviously missing. The repaired results based on element-wise sum [38] have significant mosaic effect, while the results using depth concatenation [39] are seriously blurred in detail.

D. DILATED CONVOLUTION

When repairing large missing regions in an image, the network needs to have a large area of receptive field. Using large convolutional kernel or deeper network will increase the parameters and make training process more difficult. To eliminate this disadvantage, dilated convolution [37] is introduced to our network. As there are some zero units in large kernel, dilated convolutional layers can obtain large receptive field without increasing the parameters.

The receptive field of dilated convolutional layer is:

$$f_i = f_{i-1} + \prod_{i=1}^{k-1} S_i \times (K_i - 1), \tag{5}$$

where f_i is the receptive field of the i^{th} layer, f_{i-1} is the receptive field of the $(i-1)^{th}$ layer, K_i is the size of kernel of the i^{th} layer and S_i is the extension rate of the i^{th} layer. When the size of convolution kernel is fixed, the size of receptive field for neural network increases exponentially with the number of layers.

The increase of receptive field by using extended convolution also introduces the problem of gridding effect [44], which may note be good for learning. Because the local information is completely missing and the information can be irrelevant across large distances, the design paradigm of Hybrid Dilated Convolution (HDC) [44] is adopted to solve the gridding problem. There can be no common divisor greater than 1 for the expansion rate of adjacent layers, which ensures that every pixel in the receptive field participates in the calculation. The convolutional rate is selected to follow the zigzag structure design as [1, 2, 5], subject to the following rules:

$$M_i = \max[M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i], \tag{6}$$

where r_i is the dilation rate of the i^{th} layer, M_i is the max dilation rate of the i^{th} layer. As shown in Fig.5, when the dilated rates is [1, 2, 5], all pixels that participate in the calculation and the gridding effect are completely eliminated. The shallow receptive field has a checkerboard effect. As the number of layers increases, the receptive field gradually tends to concentric circles.

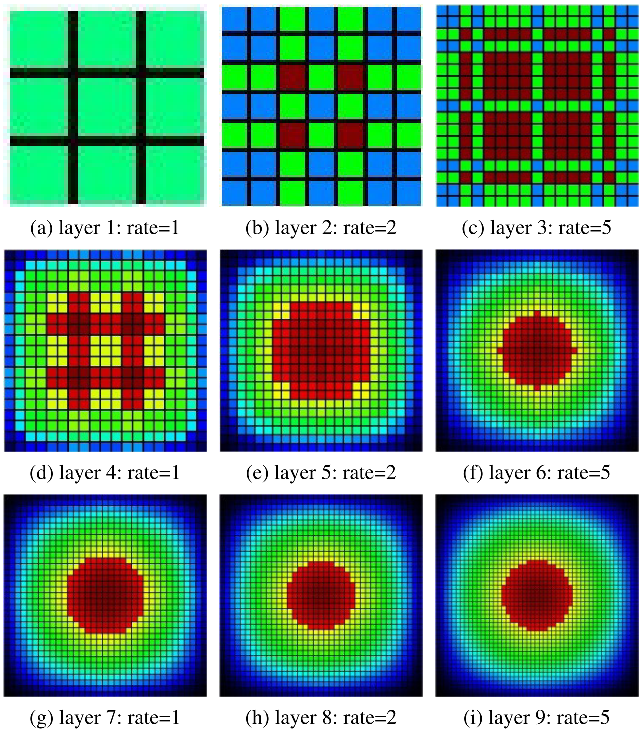


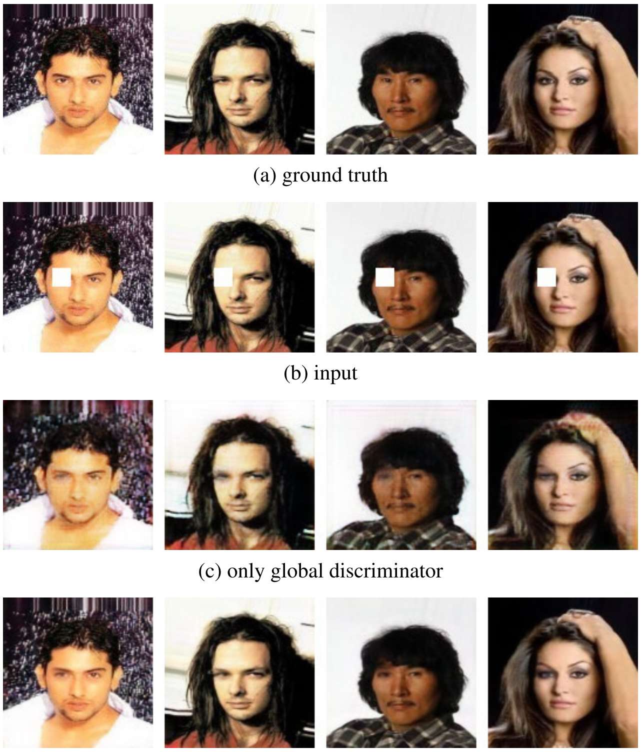
FIGURE 5. Illustration of the solution of the gridding problem. The receptive field of sequence of 9 convolutional layers has dilation rates of [1, 2, 5], respectively with kernel size 3 × 3. The times of pixel counted are represented by the color depth.

E. DISCRIMINATION PART

To achieve clarity of detail and overall consistency at the same time, an additional local discriminator for detail discrimination is used in our network. The input of the local discriminator is only the local area of the image and the discriminator only distinguishes the details. We compare the results between only using global discriminator and using both global discriminator and local discriminator in Fig. 6. From the results, the left eyes based on global discriminator tend to be blurred, while the results based on our method are clear and real, and have good consistency with the surrounding areas and the right eyes. The results deduce that our method has good performance in fixing the details and generates detailed and consistent reconstructed images regardless of the mask location.

F. OBJECTIVE FUNCTION

At the training stage, we use a combination of five loss functions. They are optimized jointly via back propagation using RMSProp optimizer [41]. In addition, an adaptive



(a) ground truth

(b) input

(c) only global discriminator

(d) NGAN

FIGURE 6. The comparison of results using only global discriminator and our proposed method. (a) Ground truth, (b) input images, the sequence (c) images that are based on only global discriminator, and the sequence (d) images that are based on our proposed method.

step-by-step function is trained in our model. We describe each loss function briefly as follows.

Code generator loss is the entropy deviation of information between the input and output of code generator. Even though it forces the network to produce a blurry output, it guides the network to roughly predict the robust information and the location of corrupted area. The code generator loss comes from the reconstruction loss of the structure of code generator, coding loss and the loss of GAN with code discriminator. It is back-propagated through the code generator and defined as:

$$\begin{aligned}\mathcal{L}_{encoder} &= MSE(\mathbb{C}(X), \mathbb{C}(X')) + MSE(X, D_{code}(\mathbb{C}(X)))\\ &\quad - E_{X'\sim P_{X'}}[\log(1 - D_{code}(\mathbb{C}(X')))], \qquad (7)\end{aligned}$$

where X is the ground truth, X' is the reconstructed image, $\mathbb{C}(\cdot)$ is the output of the code generator, and $D_{code}(\cdot)$ is the output of code discriminator, $MSE(\cdot, \cdot)$ is the pixel-wise mean square error between two images.

Code discriminator loss is the distance between synthesized image and ground truth.It is back-propagated through code discriminator. It is the discriminator loss in GAN:

$$\begin{aligned}\mathcal{L}_{code-dis} &= -E_{X'\sim P_{data}}[\log D_{code}(X')]\\ &\quad + E_{\mathbb{C}(X)\sim P_{\mathbb{C}}}[\log(1 - D_{code}(\mathbb{C}(X)))], \qquad (8)\end{aligned}$$

Image generator loss is the loss in coding reconstruction process and generator loss of the generative adversarial neural network. It is back-propagated through image generator and defined as:

$$\mathcal{L}_{gen} = MSE(X, Y) - E_{z'\sim P_{z'}}[\log(1 - D_{GL}(G(\mathbb{C}(X'))))], \qquad (9)$$

where $G(\cdot)$ is the output of image generator, Y is the reconstruction result of broken image, $D_{GL}(\cdot)$ is the sum result of global discriminator and local discriminator.

Global discriminator loss and **local discriminator loss** compute the accuracy of distinguishing synthesized image and ground truth. Global discriminator calculates based on whole image while local discriminator calculates only based on reconstructed area. They are back-propagated through the global discriminator and the local discriminator separately. They are defined respectively by:

$$\begin{aligned}\mathcal{L}_{glo-dis} &= -E_{Y\sim P_{data}}[\log D_G(Y)]\\ &\quad + E_{\mathbb{C}(X')\sim P_{\mathbb{C}(X')}}[\log(1 - D_G(G(\mathbb{C}(X'))))], \qquad (10)\\ \mathcal{L}_{loc-dis} &= -E_{y\sim P_{data}}[\log D_L(y)]\\ &\quad + E_{\mathbb{C}(x')\sim P_{\mathbb{C}(x')}}[\log(1 - D_L(G(\mathbb{C}(x'))))], \qquad (11)\end{aligned}$$

where x' is the missing area of corrupted image, y is the corresponding region in ground truth, $D_G(\cdot)$ is the result of global discriminator, $D_L(\cdot)$ is the result of local discriminator.

IV. EXPERIMENTAL RESULTS

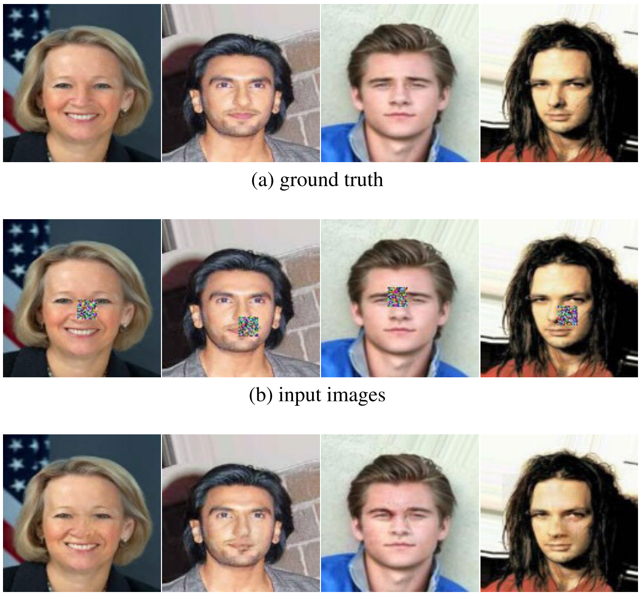
A. IMPLEMENTATION

In our work, we utilize the architecture of deep convolutional GAN (DCGAN) to train the five parts of the model. The implementation environment of the experiment is TensorFlow 1.14.0, CUDA 10.0.130, Indel(R) Core(TM) i7-6700K CPU and NVIDIA GeForceGTX1080. Our NGAN is trained on the CelebFaces Attributes (CelebA) dataset, which consists of 10,177 identities with 202,599 face images. By adding occlusion to the original face image as the input of the missing image to be repaired, the fabrication of occlusion dataset CelebA-Mask was realized. We randomly selected 10 percentage of CelebA-Mask as the test set and the remaining 90 percentage of the images as the training set. The activation function of our network uses Leaky ReLU, which can reduce the loss of information caused by negative shielding.

Our framework consists of five neural networks, and the network is constrained step-by-step during training. The training process is as follows:

1) Train code generator and image generator only;
2) Fix code generator and image generator, and train code discriminator, local discriminator and global discriminator. The number of iterations is fixed so that the training degree of the discriminators is close to that of the code generator and image generator.
3) Alternately train code generator, image generator and discriminators using the training method of GAN.

Fig. 7 shows our face completion results on the CelebA dataset. The essential facial component in an image of

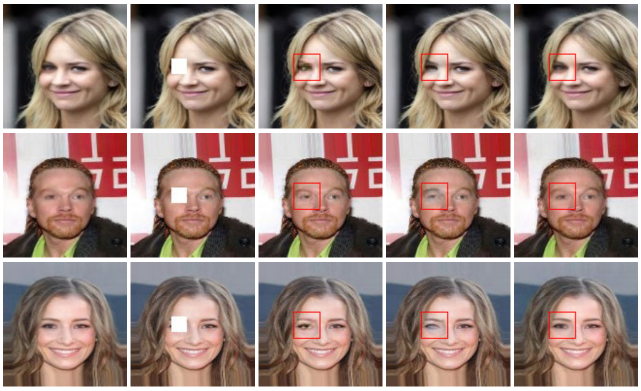


(a) ground truth

(b) input images

(c) results of our method

FIGURE 7. Example results of our proposed method. (a) Ground truth, (b) input corrupted images, (c) the inpainting results.



(a) GT (b) corrupted (c) [45] (d) [46] (e) NGAN

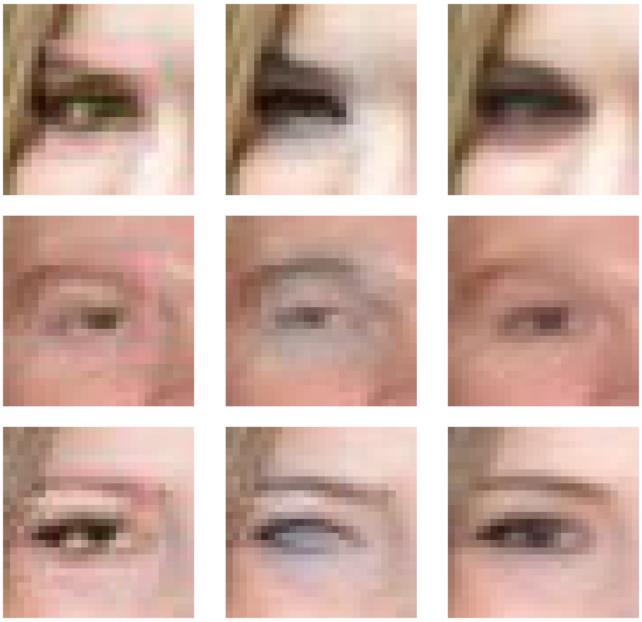
FIGURE 8. Example results: (a) ground truth, (b) corrupted images, (c) results based on [41], (d) results based on [36], (e) results based on our method. Red square areas are enlarged and shown in Fig. 9.

128 × 128 pixels is missing randomly and the size of missing area is 16 × 16 pixel.

B. QUALITATIVE EVALUATION

To evaluate the effectiveness of our model, the comparison of results with [41] and [36] are presented in Fig. 8. The occlusive area of the input map contains a wealth of semantic information, which is very different from simple texture repair work.

The results demonstrated that our approach performs best in terms of overall consistency and detail repair. In terms of detail repair, we can easily find that the repaired area of the left eye using [41] and [36] methods are blurred, while our method achieved much natural and clear details. In terms of overall consistency, the left eye and the right eye of the results using the compared methods are not consistent, while our results show that the left eye and the right eye remained coherently. The enlarged results of the inpainting area are shown in Fig. 9. The area reconstructed by the method in [41] has a clear border, and the hue of the complementary area is different from that of other parts of the face. Likewise the hue of area reconstructed by the method of [36] is different. Our method does not have this drawback, and overall consistent tone of face skin is obtained.



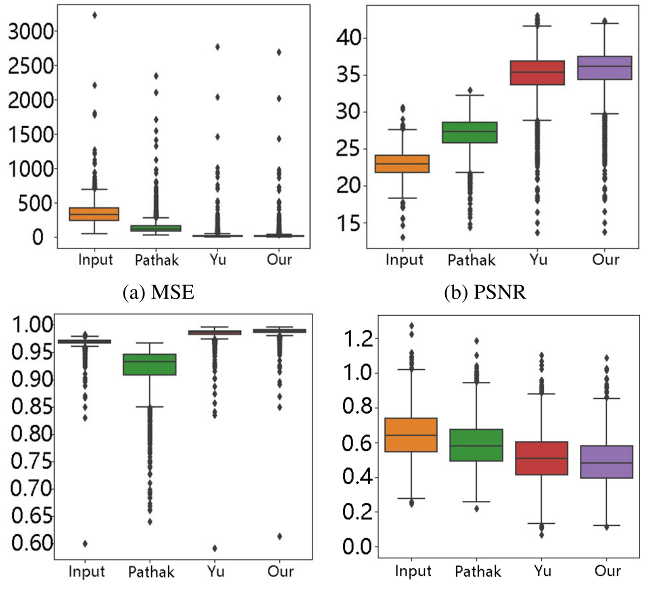
(a) [45] (b) [46] (c) NGAN

FIGURE 9. Enlarged results of the inpainting area. (a) [41], (b) [36], (c) our method.

C. QUANTITATIVE EVALUATION

In order to estimate our model quantitatively, we tested it on the whole CelebA-Mask dataset. The mean square error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and FACENET distance [42] of the repaired image are calculated and compared with the other two methods. The results of our method (end-to-end and with manual processing) are compared with Pathak *et al.* [41] and Yu *et al.* [36], and the evaluation results are shown in Fig. 10. The manual processing means replacing the non-mask region of the output with the original pixels with corresponding position manually. The corresponding results are shown in Fig. 10 as "Our".

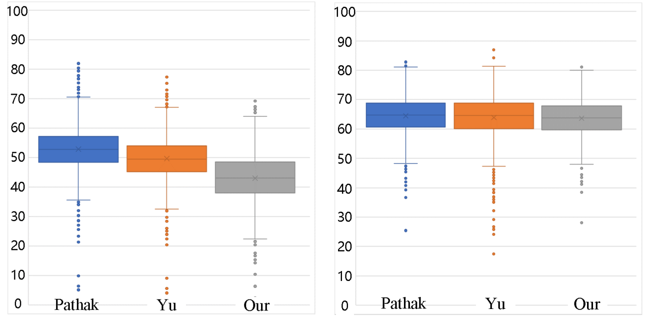
As shown in the figure, our method performs well on MSE, PSNR, SSIM, and FACENET, especially for the end-to-end results with post-processing. It deduces that our model can leverage the repaired detail, surrounding consistency, and structural reduction with less artifacts. The method directly coping raw image patches has lower MSE and FACENET, and higher PSNR and SSIM, which indicates that the repaired area of the output by our method has a high consistency with ground truth and has better facial similarity in the task of facial semantic repair. The improvement of the quality of our



(a) MSE (b) PSNR

(c) SSIM (d) FACENET

FIGURE 10. **Comparison of inpainting results with Pathak [41], Yu [36]. (a) MSE, (b) PSNR, (c) SSIM, (d) FACENET. Statistics are based on CelebA-Mask dataset with size 128 × 128 pixel. In each figure, from left to right: input image with mask, Pathak [41], Yu [36], our method (end-to-end) and the result with manual processing of our method.**



(a) $\ell_1(\%)$ (b) $\ell_2(\%)$

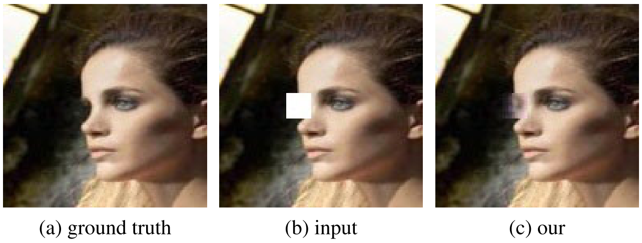
FIGURE 11. **Comparison $\ell_1(\%)$ and $\ell_2(\%)$ on CelebA-Mask with [36], [41] and our method.**

method compared with manual processing also demonstrates that the end-to-end model still introduces little noise to output images to maintaining the overall consistency and preserving the surrounding information of input image.

In addition, to compare the deductive repair capability of the network, we report our evaluation in terms of mean ℓ_1 error and mean ℓ_2 error for the results of the repaired region. The statistics distributions of [36], [41] and our method with ground truth are evaluated and compared in Fig. 11. The lower values of ℓ_1 and ℓ_2 in our method indicate that the repaired area by the proposed method is more similar to the ground truth in statistics. In addition, it is also verified that in subjective experiments, our method achieves detail-rich textures and better performance in the repaired areas.

D. LIMITATION

Although our model is able to generate semantically plausible and visually pleasing content, there is some limitations. We implement various data to test and verify the effectiveness



(a) ground truth (b) input (c) our

FIGURE 12. **Example results: (a) ground truth, (b) input images, (c) illustrate of our method.**

and robustness of our method. In the experiments in Fig. 12, our model fails to reconstruct the image for profile images. Due to the limitation of the training data, this method only works with rectangular patches (16 × 16 in this work). In the future work, we plan to merge expression detection and face position detection to our framework to address this issue.

V. CONCLUSION

In this paper, we present a novel deep generative model-based approach that improves the quality of reproducing filled regions while exhibits fine details. Our network employs a novel nesting structure to find the location of missing area and to reduce mode collapse as a prior constraint. A residual structure is adopted to deliver richer original image information to the deeper layers. Both qualitative and quantitative experiments show that our method performs well in fine details and global uniformity and can achieve end-to-end repair of defect images without additional semantic information.

REFERENCES

[1] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, Aug. 2003.

[2] M. Bertalmio, G. Sapiro, C. Ballester, and V. Caselles, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, vol. 4, no. 9. Reading, MA, USA: Addison-Wesley, 2000, pp. 417–424.

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, Jul. 2009.

[4] Y. Chen and H. Hu, "An improved method for semantic image inpainting with GANs: Progressive inpainting," *Neural Process. Lett.*, vol. 49, no. 3, pp. 1355–1367, 2019.

[5] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," 2016, *arXiv:1607.07539*. [Online]. Available: https://arxiv.org/abs/1607.07539

[6] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5485–5493.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[8] R. Köhler, C. Schuler, B. Schölkopf, and S. Harmeling, "Mask-specific inpainting with deep neural networks," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, Sep. 2014, pp. 523–534.

[9] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, vol. 17, no. 3. Cham, Switzerland: Springer, 2014, pp. 305–312.

[10] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1202–1206.

[11] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*. [Online]. Available: https://arxiv.org/abs/1901.00212

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3730–3738.

[13] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *J. Vis. Commun. Image Represent.*, vol. 12, no. 4, pp. 436–449, 2001.

[14] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1200–1211, Aug. 2001.

[15] A. Levin, A. Zomet, and Y. Weiss, "Learning how to inpaint from global image statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 305–312.

[16] J. Shen and T. F. Chan, "Mathematical models for local nontexture inpaintings," *SIAM J. Appl. Math.*, vol. 62, no. 3, pp. 1019–1043, 2002.

[17] J. Shen, S. H. Kang, and T. F. Chan, "Euler's elastica and curvature-based inpainting," *SIAM J. Appl. Math.*, vol. 63, no. 2, pp. 564–592, 2001.

[18] L. A. Vese and T. F. Chan, "A multiphase level set framework for image segmentation using the Mumford and Shah model," *Int. J. Comput. Vis.*, vol. 50, no. 3, pp. 271–293, Dec. 2002.

[19] S. Esedoglu and J. Shen, "Digital inpainting based on the Mumford–Shah–Euler image model," *Eur. J. Appl. Math.*, vol. 13, no. 4, pp. 353–370, 2002.

[20] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1033–1038.

[21] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 463–476, Mar. 2007.

[22] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[23] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 82-1–82-10, 2012.

[24] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: Image and video synthesis using graph cuts," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 277–286, 2003.

[25] J. Jia and C.-K. Tang, "Image repairing: Robust image synthesis by adaptive ND tensor voting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2003, pp. I-643–I-650.

[26] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[27] J. Kopf, W. Kienzle, S. Drucker, and S. B. Kang, "Quality prediction for image completion," in *ACM Trans. Graph.*, vol. 31, no. 6, 2012, Art. no. 131.

[28] K. He and J. Sun, "Statistics of patch offsets for image completion," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, Oct. 2012, pp. 16–29.

[29] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Image completion using planar structure guidance," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 129.

[30] I. Drori, D. Cohen-Or, and H. Yeshurun, "Fragment-based image completion," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 303–312, Jul. 2003.

[31] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum, "Image completion with structure propagation," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 861–868, 2005.

[32] D. Pavić, V. Schönefeld, and L. Kobbelt, "Interactive image completion with perspective correction," *Vis. Comput.*, vol. 22, nos. 9–11, pp. 671–681, 2006.

[33] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. no. 4.

[34] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, *arXiv:1803.07422*. [Online]. Available: https://arxiv.org/abs/1803.07422

[35] A. Lahiri, A. Jain, D. Nadendla, and P. K. Biswas, "Improved techniques for GAN based facial inpainting," 2018, *arXiv:1810.08774*. [Online]. Available: https://arxiv.org/abs/1810.08774

[36] J. Yu, L. Zhe, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[38] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5353–5360.

[39] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[40] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*. [Online]. Available: https://arxiv.org/abs/1207.0580

[41] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2536–2544.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[43] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: https://arxiv.org/abs/1511.06434

[44] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1454–1460.

ZHIJIANG LI received the B.Sc. and M.Sc. degrees in printing engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998, 2001, and 2005, respectively. Since 2005, he has been with Wuhan University. From 2015 to 2016, he was a Visiting Research Fellow with the School of Design, University of Leeds, U.K. He is currently an Associate Professor and the Head of the Department of Printing Engineering, School of Printing and Packaging, Wuhan University. He is the author of three books and more than 40 articles and holds seven patents and six software. His research interests include color vision, image processing, and image reproduction.

HAONAN ZHU received the B.E. degree from Wuhan University, China, in 2018. He is currently pursuing the master's degree with Wuhan University. His research interests include image processing and color science.

LIQIN CAO received the M.Sc. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2009, respectively. From 2015 to 2016, she was a Visiting Research Fellow with the Department of Information and Communication Technology, University of Agder (UiA), Grimstad, Norway. She is currently with the School of Printing and Packaging, Wuhan University. Her research interests include image processing, computer vision, and remote sensing data processing.

LEI JIAO received the B.E. degree in telecommunication engineering from Hunan University, Changsha, China, in 2005, the M.E. degree in communication and information system from Shandong University, Jinan, China, in 2008, and the Ph.D. degree in information and communication technology from the University of Agder (UiA), Grimstad, Norway, in 2012. He is currently an Associate Professor with the Department of Information and Communication Technology, UiA. His current research interests include resource allocation in wireless communications, smart grid, and reinforcement learning.

YANFEI ZHONG received the B.S. degree in information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 2002 and 2007, respectively. Since 2007, he has been with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, where he is currently a Full Professor. He has published more than 80 research articles, including more than 40 peer-reviewed articles in international journals, such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS PART B, and *Pattern Recognition*. His research interests include multi- and hyperspectral remote sensing data processing, high resolution image processing and scene analysis, and computational intelligence.

AILONG MA received the B.S. degree from the China University of Petroleum, Qingdao, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017. He is currently a Research Associate with Wuhan University. His major research interests include remote sensing image processing, evolutionary computing, and deep learning.