IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON BIG DATA LEARNING AND DISCOVERY

Received January 13, 2019, accepted February 7, 2019, date of publication February 14, 2019, date of current version March 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899466

Generating Photographic Faces From the Sketch Guided by Attribute Using GAN

JIAN ZHAO, XIE XIE, LIN WANG, (Member, IEEE), MENG CAO, AND MIAO ZHANG
School of Information Science and Technology, Northwest University, Xi'an 710127, China

Corresponding author: Lin Wang (wanglin@nwu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772421 and Grant 61503300, and in part by the Foundation of Key Laboratory of Space Active Opto-Electronics Technology of Chinese Academy of Sciences under Grant AOE-2016-A02.

ABSTRACT From a sketch image or text description, generating a semantic and photographic face image has always been an extremely important issue in computer vision. Sketch images generally contain only simple profile information but not the detail of the face. Therefore, it is difficult to generate facial attributes accurately. In this paper, we treat the sketch to face the problem as a face hallucination reconstruction problem. In order to solve this problem, we propose an image translation network by exploiting attributes with the generated adversarial network. And it can significantly contribute to the authenticity of the generated face by supplementing sketch image with the additional facial attribute feature. The generator network is composed of a feature extracting network and downsampling–upsampling network, both networks use skip-connection to reduce the number of layers without affecting network performance. The discriminator network is designed to examine whether the generated faces contain the desired attributes or not. In the underlying feature extraction phase, our network is different from most attribute-embedded networks, we fuse the sketch images and attributes perceptually. We set the network sub-Branch A and B, which receive a sketch image and attribute vector in order to extract low-level profile information and high-level semantic features. Compared with the state-of-the-art methods of image translation, the performance of the proposed network is excellent.

INDEX TERMS Face hallucination, GAN, face generation, attribute-embedded, skip-connection.

I. INTRODUCTION

It is an extremely challenging problem in computer vision to generate a corresponding image based on a simple text descriptions or sketch, which has many practical applications such as criminal investigation and game character creation. Recently, there are many studies on trying to solve this problem. For text to image task, Reed *et al.* [1] based on deep convolutional generative adversarial networks (DCGAN) [2] proved that GAN can generate images effectively conditioned on text descriptions. Similarly, Zhang *et al.* [3] proposed a stacked generative adversarial networks (StackGAN) for synthesizing photo-realistic images from text. For sketch to image task, Isola *et al.* [4] proposed a pixel to pixel image translation network which opened the image to image research boom, there is an interesting demo using edge to generate shoes (cat). Lu *et al.* [5] posed the image generation problem as an image completion problem, with sketch providing a weak contextual constraint. Same as analyzing user preferences from a large amount of data [6], [7], we can train a network to generate impressive photographic results from a large number of sketch-face image pairs.

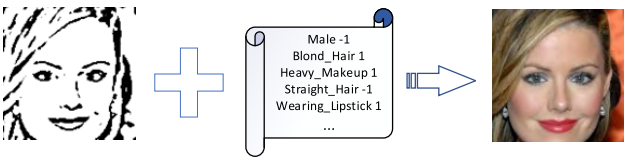


FIGURE 1. A single sketch and a simple text description are possible to generate a photographic face. Through some preliminary experiments, we found that sketch images contain rich profile information, while attribute vectors contain high-level semantic information about texture details, colors, and so on. If you combine the two information, the sketch provides rough outline information, and the attribute provides local texture details so that you can generate real photographic faces.

Similar to DCGAN, it is possible to generate a certain kind of image, which is completely random and uncontrollable. However, for generating a face image, we hope to generate what we want, e.g., a certain kind of profile, hairstyles, ruddy faces, etc. That is, we need to achieve a diverse and controllable generation process as shown in Fig. 1.

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

FIGURE 2. Face hallucination super-resolution reconstruction is the process of completing (a), and our sketch to face is the process of completing (b). In our analysis, the two processes are similar. Among them, the left side is the hallucination face, the middle is the ground-truth face, and the right side is the sketch face.

Inspired by Create Anime girl [8], the detailed attributes of the face can be known from the text description, and the sketch can provide the corresponding positional composition of the text. And we can generate more realistic images if we have both sketch and text. We propose a method which can use GAN to generate photographic face from the sketch guided by attribute.

Face hallucination reconstruction [9]–[11] is a special image super-resolution reconstruction technique which is reconstructs the hallucinated low-resolution face into a high-resolution face. Hallucination face reconstruction is important in video semantic recognition [12]–[14]. The low-resolution image is considered as the input of hallucination face reconstruction which contains low frequency components of color and profile. In addition, the difference between the real HR image and the LR image provides high frequency components: detail. Correspondingly, for sketch to face problem, we think that the sketch image provides low-level profile information and the attribute provides high-level semantic information such as color and detail (as shown in Fig. 1 and 2). Thus, the above two problems are special image translation problems, but the input is different depending on the task. Because of the similarities mentioned above, we treat sketch to face problem as a hallucination face reconstruction problem in this paper. Our network is exploiting the attribute information to reconstruct a color photographic face from a sketch image. It is that attributes are used to generate high-level semantic information.

Our network is a standard generated adversarial network [15], which has two sub-networks of generator and discriminator as shown in Fig. 3. The generators can be divided into feature extraction module G_f and face reconstruction module G_r. And feature extraction module G_f is composed of branch A and B. The reason to adopt such a structure is that, the overlap between branches is very small. And the sketch image provides low frequency components such as textures, while the attribute provides high frequency components such as detail and color. By visualizing the convolution operation shown in Fig. 4, the overlap between branch A and B is very small.

GAN have recently shown that they have a significant performance on image generation. The current trend is that, performance can be improved by increasing the number of layers. However, the most distressing thing is the rapid increase of computational complexity with the increase of the number of layers and the emergence of various problems. Skip-connection can greatly reduce the number of network layers without affecting network performance. Because of the sketch to face task is regarded as face hallucination super-resolution reconstruction in our network, in order to make our network not as deep as those methods such as Residual-Net [16] and Densenet [17], skip-connection [18] is essential. This technique can bring better performance to our network with less network layers. The specific comparison experiment is put in Section 4.

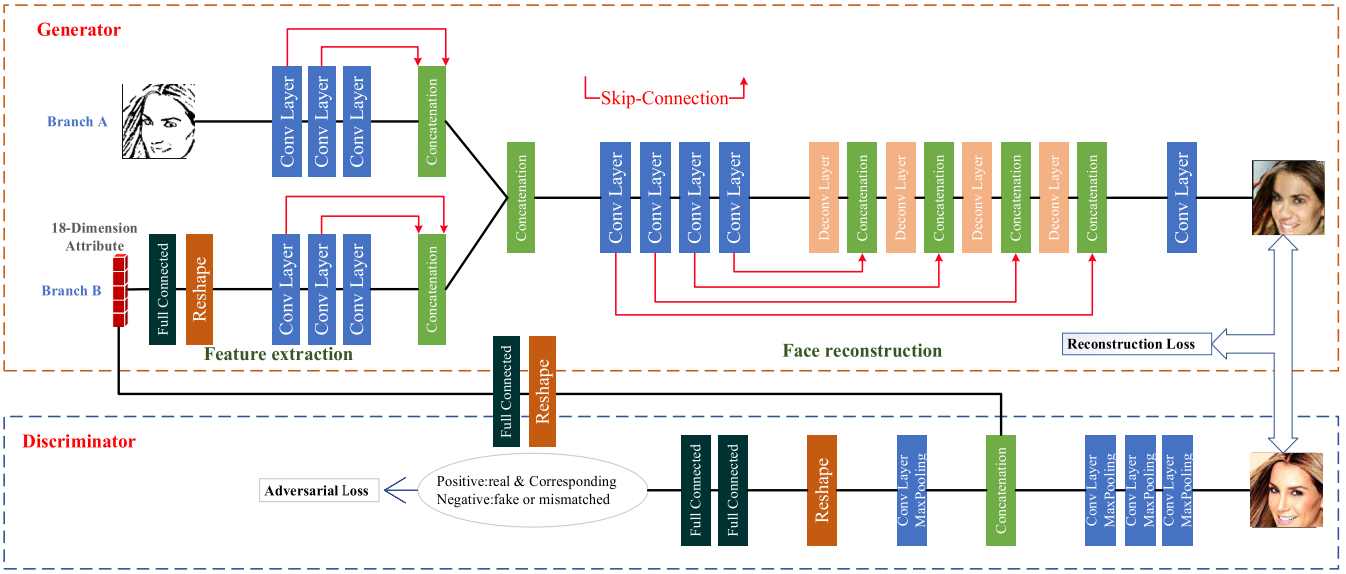
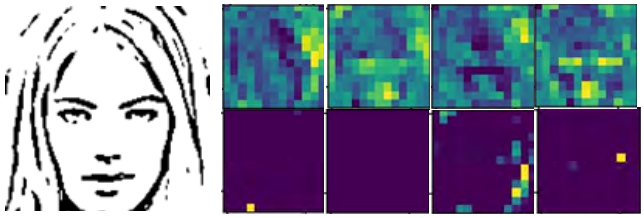


FIGURE 3. As shown above, this is the structure of the complete network we proposed. Our network is divided into two parts: the generator and the discriminator. The input to the generator is sketch image which is the input of branch A and the attribute vectors which is the input of branch B. The face image is generated after passing through the generator network, and the difference from the ground-truth face image constitutes Reconstruction Loss. The discriminator loss is Adversarial Loss. The loss function is used to update the parameters of the generator network and the discriminator network.



(a) (b)

FIGURE 4. We use the method of visualizing the feature map to visualize the feature maps of branch A and branch B after convolution, and arbitrarily select from among the 16 feature maps. We can find very little overlap between them. (a) is a sketch image, and (b) is an extracted feature map.

And the main contributions of this paper are summarized as follows:

First, according to our analysis, face hallucination super-resolution reconstruction has a great similarity to sketch of face. Therefore, we regard sketch of face as a special face hallucination super-resolution reconstruction. Due to the change of our task, we propose a new GAN-based network structure to solve it.

Second, in the underlying feature extraction phase, our network is different from most attribute-embedded networks. Those methods simply expand the dimension of attribute and reshape its size, and then concatenate it to a convolution layer output. High-level semantic features are extracted from the attribute vector and then reshaping it into the same size as the input sketch image. Such a network setting can perceptually learn from the attributes to advanced semantic features.

Third, the skip-connection structure is adopted in both feature extracting network and downsampling-upsampling network of the generator network, which can greatly reduces the number of our network layers without affecting network performance.

II. RELATED WORK

A. GENERATIVE ADVERSARIAL NETWORK

Goodfellow *et al.* [15] introduced the GAN framework which could train two convolution neural networks at the same time as generator and discriminator. The training process alternates to optimize the generator and discriminator, which can compete with each other. The generator learns to generate samples that can fool the discriminator. The discriminator learns to distinguish real data and data which generated. Through training, the data generated by this method is very real. Radford *et al.* [2] proposed DCGAN that make GAN stable to train in most settings. Arjovsky *et al.* [19] purposed a new method that used the distance of real and generated data distribution called Wasserstein GAN which makes training processes of GAN more stable. There are now a number of GANs [31], [32] that can handle a variety of different tasks.

B. FACE HALLUCINATION RECONSTRUCTION

Face hallucination is a special kind of single image super-resolution reconstruction. Traditional face hallucination and deep learning-based face hallucination are two main types of division. For traditional face hallucination, Wang, and Tang [20] reconstructed HR by using a linear mapping between the input LR image and HR image. Liu *et al.* [21] purposed a global face model learning by principal component analysis (PCA). Baker and Kanade [22] reconstructed high-frequency details of aligned LR face images by searching the best mapping between LR and HR patches. For deep learning-based face hallucination, many methods [9]–[11], [33] are based on GAN by using the attributes of the face to achieve the task.

C. IMAGE TO IMAGE & TEXT TO IMAGE

For paired data, CGAN proposed by Isola *et al.* [4] can be used for multiple tasks in image-to-image translation, such as semantic label to image, map to aerial photo, edge to photo, and so on. But in reality, it is difficult to obtain pairs of images. Therefore, Zhu *et al.* proposed CycleGAN [23], DualGAN [24], and DiscoGAN [25] in order to solve the problem of image to image translation without paired images. Johnson *et al.* [26] also proposed style transfer.

Reed *et al.* [1] solved the text to image task by adding random noise of DCGAN and the text descriptions that you want to generate. Zhang *et al.* [3] proposed StackGAN for synthesizing photo-realistic images from text, which can generate corresponding high-resolution photo-level images from text in two stages. Di and Patel [27] implemented text-sketch-face by using VAEs [34] and GANs.

III. PROPOSED METHOD

Similar to face hallucination reconstruction, we propose a new network based on GAN. In the feature extraction stage, we extract profile information and high-level semantic information from sketch images and attribute vectors. A combination of GAN and Skip connection layers are used in all the generator. Each output of the convolution layer is passed to the back convolution layer and simultaneously concatenated to the next concatenation layer.

A. OVERALL FRAMWORK

The network we proposed is a standard GAN, that is, a network of discriminators and a generator network. The input of the generator in our network is a sketch image with the size of $128 \times 128 \times 3$ and an 18-dimensional attribute vector. The discriminator input is a $128 \times 128 \times 3$ photographic face image corresponding to the sketch image and the same 18-dimensional attribute vector as the generator network. The difference is that the attribute vector in the generator is reshaped into the same size as the sketch image after the fully connected layer, and the discriminator network is reshaped and then concatenate it to the third convolution layer output. This is to extract high-level semantic features in the build network better. And skip connection in the network is adopted to generate a strategy. And the details are given in Section 4.

B. NETWORK ARCHITECTURE

1) GENERATOR NETWORK

Our network has two inputs: I^s and A., The input of branch A, I^s, is a sketch image with the size of $128 \times 128 \times 3$, and three 3×3 convolution layers are used to extract low-level features such as facial profile. Here, the output of the three convolution layers is concatenated together by using skip-connection. A is the 18-dimensional attribute vector, which is used to represent some facial attributes in the corresponding image. Using the fully connected layer expands its dimension from 18 to 41592 ($128 \times 128 \times 3$) and then we reshape it to the same size as the sketch image. The following convolution process is the same as branch A. The outputs of the two sub-branches concatenated together and are sent to the next downsampling-upsampling network. Meanwhile, a technique of skip-connection is adopted, and a photographic face image is generated by C64-C128-C256-C512-DC512-DC256-DC128-DC64-C3, where C represents convolution layer and DC represents deconvolution. We use PReLU [28] activation function after each layer in convolution and deconvolution layer except for face reconstruction which uses tanh.

2) DISCRIMINATOR NETWORK

Similar to Reed *et al.* [1], we want to exploit the attribute information of the sketch image and then generate the corresponding photographic face. It is mean that we need the discriminator network to distinguish the real image and the corresponding attribute pair. Specifically, in order to train the discriminator network, we take ground-truth face image f and their corresponding ground-truth attributes a as positive sample pairs $\{f, a\}$. Negative data is generated by generator face $\hat{f}$ and their ground-truth attributes a as well as real faces and mismatched attributes $\hat{a}$. Therefore, the negative sample pairs consist of both $\{\hat{f}, a\}$ and $\{f, \hat{a}\}$.

C. OBJECTIVE FUNCTION

Our objective function consists of two parts, namely, reconstruction loss and adversarial loss. Since we regard the facial sketch problem as a face hallucination super-resolution reconstruction problem, we also use the common pixel-wise Euclidean distance loss as our reconstruction loss [16]. The loss is used to measure the difference between generated face and ground-truth face. Adversarial loss is used in our objective function to make the generated face more realistic, closer to the photo level, and constraint with attribute. The loss function is used to update the parameters of the generator network and the discriminator network.

Thus, our objective function is defined as follows:

$$L = L_{\text{Reconstruction}} + \lambda L_{\text{adv}} \tag{1}$$

According to experience, we set the value of $\lambda = 0.01$.

1) ADVERSARIAL LOSS

According to the settings of our discriminator network, our positive sample pairs are $\{f, a\}$, while negative sample pairs are $\{\hat{f}, a\}$ and $\{f, \hat{a}\}$.

Therefore, the adversarial loss of our network is defined as follows:

$$\begin{aligned} L_{adv}(f, \hat{f}, a) &= -E[\log D_\psi(f, a)] \\ &\quad -E[\log(1 - D_\psi(\hat{f}, a) + \log(1 - D_\psi(f, \hat{a})] \\ &= -E_{(f_i,a_i)\sim p_{(f,a)}}[\log D_\psi(f_i, a_i)] \\ &\quad -E_{(f_i,a_i)\sim p_{(f,a)}}[\log(1 - D_\psi(\hat{f}_i, a_i) + \log(1 - D_\psi(f_i, \hat{a}_i)] \end{aligned} \tag{2}$$

where f is the ground-truth of face image I_f, $\hat{f}$ represents generated face which is the output of input sketch image I_s after the generator G_ω, that is $\hat{f} = \mathrm{G}_\omega(I_s, a)$. a is the attribute of face, $\hat{a}$ is mismatched attribute of face. ω and ψ represent the parameters of the generator and discriminator networks. $D_\psi(f, a)$, $D_\psi(\hat{f}, a)$, and $D_\psi(f, \hat{a})$ are the output of the discriminator network.

2) RECONSTRUCTION LOSS

As described in Section 1, face hallucination reconstruction and sketch to face have similarity. We penalize pixel-wise Euclidean distance between generated face and the corresponding ground-truth face as follows:

$$\begin{aligned} L_{\text{Reconstruction}}(a, I_s, I_f) &= ||G_\omega(I_s, a) - I_f||_2^2 \\ &= ||\hat{f} - f||_2^2 \\ &= ||\hat{f}_i - f_i||_2^2 \end{aligned} \tag{3}$$

Therefore, our objective function is defined as formula 4:

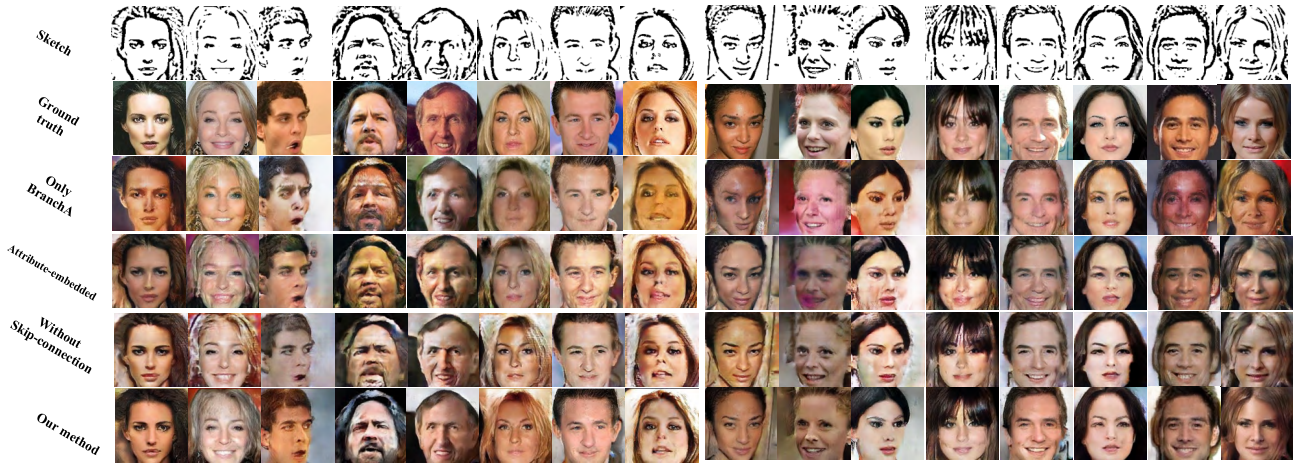
$$\begin{aligned} \max_\omega \min_\psi \frac{1}{N} \sum_{i=1}^{N} ||\hat{f}_i - f_i||^2 + \lambda(&-E_{(f_i,a_i)\sim p_{(f,a)}}[\log D_\psi(f_i, a_i)] \\ &-E_{(f_i,a_i)\sim p_{(f,a)}}[\log(1 - D_\psi(\hat{f}_i, a_i) \\ &+ \log(1 - D_\psi(f_i, \hat{a}_i)]) \end{aligned} \tag{4}$$

According to Eq. 4, the parameters ω of the generator can be maximized and the parameters ψ of the discriminator can be minimized. Through optimization, And the network can achieve excellent performance by optimize the objective function (4).

D. IMPLEMENTATION DETAILS

The detailed architectures of the generator and discriminator networks are illustrated in Fig. 3. We implement our model by using the TensorFlow. We use the Adam optimizer [29] with a learning rate of 0.0002 and a beta of 0.5 for both the generator and discriminator network. We adopt the dataset which is mentioned (contain 100K train images) in Section 4 and it is selected and cropped from CelebA. The network is trained with a batch size of 64 and epochs of 200, which takes a week for training on a single Nvidia 1080Ti GPU.

In order to test the performance of our network, we compare it with the state-of-the-arts technologies. At the same time, we carry out a lot of self-contrast tests for the feature extraction network which is divided into two branches and the



(a)

(b)

FIGURE 5. For Fig. 5 (a), after comparing the changed network with our network, it is obvious that our network is excellent in the task of sketch to face. Only Branch A represents a network without branch B, which is, a face is generated only by sketch, without reference to the attribute vector. Attribute-embedded represents concatenate attribute traditionally. Without skip-connection means that skip connect is removed from our network, and Ours represents the complete network proposed by us. For Fig5.(b), we have chosen several face images generated from our network, and we can see that we have preserved the attributes and texture details that should be generated, and remained diversity as well.

application of skip-connection. Experimental results show that our network can yield excellent performance, and the sub-branches and skip-connection techniques are effective to improve the performance of our network.

IV. EXPERIMENTS

A. DATASET

Similar to [18], we use the Celebrity Face Attributes (CelebA) dataset [30] to train and test our model. The dataset is consisted of 200K face images, where each image has 40 binary attributes classes. We select 100K images from CelebA as train dataset, and 10K images from celebA as test dataset. We detect and crop face from those original images to the shape of 128 $\times$ 128. The XDoG method is used to the extract sketch from cropped faces. According to our cropping database, we notice that some attributes are out of the clipping range or do not contain semantic information, and some attributes can be clearly perceived from the sketch image from the pre-training point of view. Even if the attribute is changed, the generated face cannot be changed, for example, wearing_Earrings, wearing_Necktie, wearing_eyeglasses. Hence, we choose 18 attributes, such as gender, hair color, and beard information from 40 attributes in CelebA.

B. EVALUATION ON OUR OWN NETWORK

Our network applies the sub-branch structure in the feature extraction phase. The previous algorithms only used branch A from image to image, or used branch B alone from the text to image algorithm. Our network is a task for implementing sketch to image. And it is similar to [1], attribute vectors are added to the generation network. We find that the addition of attributes can make the generated face image behave better in terms of face texture and color, and we have a more subtle structure when adding attribute vectors. Meanwhile, the skip-connection structure added to the generator network makes our network even better.

In Fig. 5(b), some sample images are randomly selected and several corresponding generated face images are randomly selected. From Fig. 5(b), we can see that our generated results are very good in generating texture details and

color information, while ensuring the diversity of generated images.

1) WITHOUT SUB-BRANCH

Our network proposes a sub-branch structure to extract features. Through our analysis in Section 3, sketch image, the input of branch A, provides profile information, and attribute input of branch B provides high-level semantic information such as the texture and color. The specific network performance is shown in the sub-branch as Fig. 5. In order to verify that our network is performing well, we also conduct the following comparative tests:

a: ONLY BRANCH A

We compare our network with the traditional network from image to image, which only has the image input in branch A, instead of the input of the attribute vector in branch B [2]. In branch A of Fig. 5, it can be clearly seen that such a network performs poorly, and the generated image is roughly similar to the input sketch image, and the difference between texture features and color of generation is obvious.

b: CONCATENATE ATTRIBUTE TRADITIONALLY

Similar to [1], [3], and [27], we embed the attribute vector. The difference is in our proposed network, we expand the attribute vector to 41592 dimensions first. Then, after the reshape operation, it becomes the same size as the sketch image. Finally, it joins generator, so that the high-level semantic information can be added to from the outline information provided by the sketch image. Such an operation can extract high-level semantic information in the attribute vector during the feature extracting stage better. Attribute-embedded is shown in Fig. 5, and we can see that our fusion method is better than traditional concatenate attribute.

2) WITHOUT SKIP-CONNECTION

In order to make our network not as deep as those methods such as Residual-Net and Densenet, the skip-connection is essential. In order to test the effect of skip-connection in our network, we conduct a comparative test to remove all skip-connections in the network. As is shown in Fig. 5, we can see that, the performance of the network is very poor without the skip-connection. The details of the face are very unnatural, and the colors are not coordinated.

3) LOCAL REGION

Through several experiments, we find that our network is very good for the generation of some details, especially the generation of facial colors, such as whether lipstick or heavy make is worn. As shown in Fig. 6, we randomly select a few test results. It can be clearly seen that we are more realistic in generating local details and colors, and the effect is even better.

C. COMPARING WITH STATE-OF-THE-ARTS

Pixel to pixel image translation [4] proposed by Isola *et al.* can be used for edge to cat. It is the same applicable in

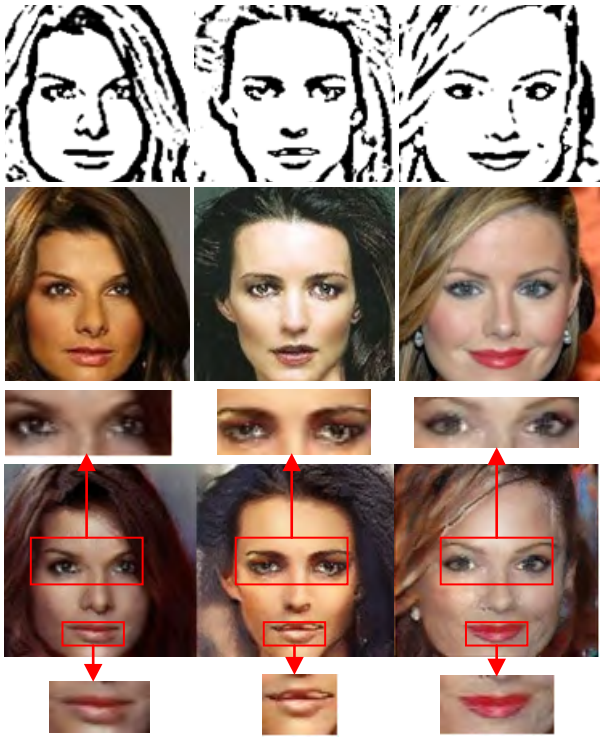


FIGURE 6. We randomly extracted several generated images from the sample. It can be seen that the proposed network generates obvious realistic facial texture details, which are represented by eyes and eyelashes, as well as color information of the lips. The top is the sketch image, the middle is the ground-truth, and the bottom is the face image which is our purposed network generated face.

FIGURE 7. The comparison between our proposed network and pix2pix and Lu *et al.* [5] is shown above. Where (a) is the sketch image, (b) is ground-truth, (c) is generated by pix2pix, (d) is generated by Lu *et al.* [5], and (e) is generated by our method.

sketch to face, and also shows a relatively good effect. Lu *et al.* [5] treat the image generation as image completion and sketch providing a weak contextual constraint, which is a good implementation of sketch to image. Although there are already many ways to implement the task of sketch to image, compared with our network, the performance of details and colors are lagging behind the performance of our network. Especially the second picture in Fig. 7, which is a man with sunglasses, the best way to generating the sunglasses is

our network. Although sunglasses are generated in (c), it is obvious that the generated abruptness is not as natural as that produced by our method, and no sunglasses are generated at all in (d), although it is difficult. This may be due to the fact that our network has joined the guidelines for the attribute, while the other two methods have not. Fig. 7 shows that performance of our network compared with state-of-the-arts. It is clear that our network is excellent in generating of facial texture details and color. Fig. 7 (e) is the generated face by our method, (d) and (e) are respectively Pix2Pix and Lu *et al.* [5].

In Fig. 7, the face image generated by (e) and the ground-truth in (b) remain intact in terms of profile features. The generation of color information also appears to be naturally smooth, although is not exactly the same as ground-truth (this is the generated image after all). In terms of local detail generation, it is also very similar to (b).

TABLE 1. Analysis of network performance with PSNR and SSIM.

METHOD	PSNR/SSIM
PIX2PIX	14.5834/0.5682
LU *ET AL.* [5]	15.9439/0.5785
Only BranchA	14.8937/0.5765
ATTRIBUTE-EMBEDDED	12.9434/0.5365
WITHOUT SKIP-CONNECTION	13.4379/0.5471
OURS	16.3069/0.5790

Since our network is based on a framework of image super-resolution reconstruction, we can apply similar evaluation metrics to illustrate the effectiveness of our method. In Table 1, we use image super resolution evaluation metrics which named peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate the face image.

It can be seen that our method has achieved good results. However, compared with the current image super-resolution algorithm, the effects of PSNR and SSIM are poor. The can't be said that our network is not good, just because the face image generated from the sketch image is very random. The also ensures the diversification of the generated results.

V. CONCLUSION

In image to image and text to image, there are already many excellent algorithms and networks, and they all can generate clear images. However, in the field of face generation, it is difficult to generate a satisfactory face due to the specialty of the face which need more texture details and color information. Regarding the sketch to face problem as face hallucination super-resolution reconstruction, we propose a more suitable network according to the change of tasks. We adds attribute information into feature extracting, and makes better use of these advanced semantic information. Meanwhile our network applies skip-connection skills. These all make the generated face image more realistic and closer to the photographic. And the effect of the proposed method is excellent, especially in the generation of local features.

REFERENCES

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, May 2016, pp. 1681–1690. [Online]. Available: https://arxiv.org/abs/1605.05396

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

[3] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Aug. 2017, pp. 5907–5915. [Online]. Available: https://arxiv.org/abs/1612.03242

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134. [Online]. Available: https://arxiv.org/abs/1611.07004

[5] Y. Lu *et al.*, "Image generation from sketch constraint using contextual GAN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 639–648. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-01270-0

[6] Z. Cheng, Y. Ding, L. Zhu, and M. Kankanhalli, "Aspect-aware latent factor model: Rating prediction with ratings and reviews," in *Proc. Int. World Wide Web Conf. Steering Committee (WWW)*, Apr. 2018, pp. 639–648.

[7] Z. Cheng, J. Shen, L. Nie, T. S. Chua, and M. Kankanhalli, "Exploring user-specific information in music retrieval," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Aug. 2017, pp. 655–664.

[8] Y. H. Jin. *MakeGirlsMoe*. Accessed: Dec. 15, 2018. [Online]. Available: https://make.girls.moe/

[9] X. Yu and F. Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, Feb. 2017, pp. 4327–4333.

[10] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 614–630. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46454-1_37

[11] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5367–5375. [Online]. Available: https://ieeexplore.ieee.org/document/8100053

[12] Z. Zeng, Z. Li, D. Cheng, H. Zhang, K. Zhan, and Y. Yang, "Two-stream multirate recurrent neural network for video-based pedestrian reidentification," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3179–3186, Jul. 2018.

[13] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1617–1632, Aug. 2017.

[14] M. N. Luo, X. J. Chang, L. Q. Nie, Y. Yang, and G. Alexander Hauptmann, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, Feb. 2018.

[15] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 2672–2680.

[16] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654. [Online]. Available: https://arxiv.org/abs/1511.04587

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708. [Online]. Available: https://arxiv.org/abs/1608.06993

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Oct. 2015, pp. 234–241.
[19] M. Arjovsky, S. Chintala, and L. Bottou. (Jan. 2017). "Wasserstein GAN." [Online]. Available: https://arxiv.org/abs/1701.07875
[20] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005.
[21] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vis.*, vol. 71, no. 6, pp. 2035–2049, Jun. 2015.
[22] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2002, pp. 1167–1183. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/CVPR.2000.854852
[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial network," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232. [Online]. Available: https://arxiv.org/abs/1703.10593
[24] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2849–2857. [Online]. Available: https://arxiv.org/abs/1704.02510v1
[25] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Aug. 2017, pp. 1857–1865. [Online]. Available: https://arxiv.org/abs/1703.05192
[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 694–711. [Online]. Available: https://arxiv.org/abs/1603.08155v1
[27] X. Di and V. M. Patel. (2017). "Face synthesis from visual attributes via sketch using conditional VAEs and GANs." [Online]. Available: https://arxiv.org/abs/1801.00077
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[29] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[30] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2015, pp. 3730–3738.
[31] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4790–4798. [Online]. Available: https://arxiv.org/abs/1606.05328
[32] X. Wang and A. Gupta, A, "Generative image modeling using style and structure adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 318–335. [Online]. Available: https://arxiv.org/abs/1603.05631
[33] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 690–698. [Online]. Available: https://arxiv.org/abs/1708.03132v1
[34] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Dec. 2015, pp. 3483–3491.

Authors' photographs and biographies not available at the time of publication.