IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 30, 2017, accepted January 30, 2018, date of publication February 8, 2018, date of current version April 4, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2803787

From Eyes to Face Synthesis: a New Approach for Human-Centered Smart Surveillance

XIANG CHEN[1], LINBO QING[1], (Member, IEEE), XIAOHAI HE[1], (Member, IEEE), JIE SU[1], AND YONGHONG PENG[2], (Member, IEEE)

[1]College of Electronics and Information Engineering, Sichuan University, Chengdu 610065 China
[2]Faculty of Computer Science, University of Sunderland, Sunderland SR6 0DD, U.K.

Corresponding author: Linbo Qing (qing_lb@scu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471248, in part by the Chengdu Science and Technology Project under Grant 2016-XT00-00015-GX, in part by the Scientific Research Project of Sichuan Education Department under Grant 18ZB0355, and in part by the Technology Program of Public Wellbeing of Chengdu under Grant 2015-HM01-00293-SF.

ABSTRACT With the popularity of surveillance cameras and the development of deep learning, significant progress has been made in the field of smart surveillance. Face recognition is one of the most important yet challenging tasks in human-centered smart surveillance, especially in public security, criminal investigation and anti-terrorism, and so on. Although, the state-of-the-art algorithms for face recognition have achieved dramatically improved results and have been widely applied in authentication scenario, the occlusion problem on face is still one of the critical issues for personal identification in smart surveillance, especially in the occasion of terrorist searching and identification. To address this issue, this paper proposed a new approach for eyes-to-face synthesis and personal identification for human-centered smart surveillance. An end-to-end network based on conditional generative adversarial networks (GAN) is designed to generate the face information based only on the available data of eyes region. To obtain photorealistic faces and identity-preserving information, a synthesis loss function based on feature loss, GAN loss, and total variation loss is proposed to guide the training process. Both the subject and objective experimental results demonstrated that the proposed method can preserve the identity based on eyes-only data, and provide a potential solution for the identification of person even in the case of face occlusion.

INDEX TERMS Face synthesis, Face recognition, conditional GAN, smart surveillance, video surveillance, image generation.

I. INTRODUCTION

With the popularity of surveillance cameras, video surveillance plays an increasingly important role for social public safety. To enable smart surveillance systems, researchers began to introduce computer vision technology to surveillance system supporting effective analysis and investigation. Furthermore, with the advent of deep learning, much progress has been made in the field of computer vision, including video surveillance. The current smart surveillance system has 4 major objectives: (i) object detection, (ii) object tracking, (iii) object classification/identification (include face recognition), behavior and (iv) activity analysis [1]. Their relations are demonstrated in Fig. 1.

These four tasks are related to each other, from low-level task to high-level task. For human-centered smart surveillance, especially for the scenarios of public security, criminal investigation and anti-terrorism, the identification of person in the video scene is one of the most significant yet challenging tasks. With the application of deep learning, face recognition has made a significant progress in recent years, such as Gaussianface [2], Deepid [3], Deepface [4], Facenet [5]. In recent studies, the recognition rate of the algorithm has even exceeded that of humans on public face dataset like labeled faces in the wild home (LFW) [2], [3]. However, the high recognition rate of these face recognition methods was obtained on the public dataset, where the quality of images are usually good, with slightly inclined and obscured. These methods normally require: the image with full faces appeared. In typical smart surveillance systems, there are numerous disturbing factors for face recognition. Tilted and blocked faces lead to some issues in existing algorithms. Particularly, the occlusion problem is hard to solve with the

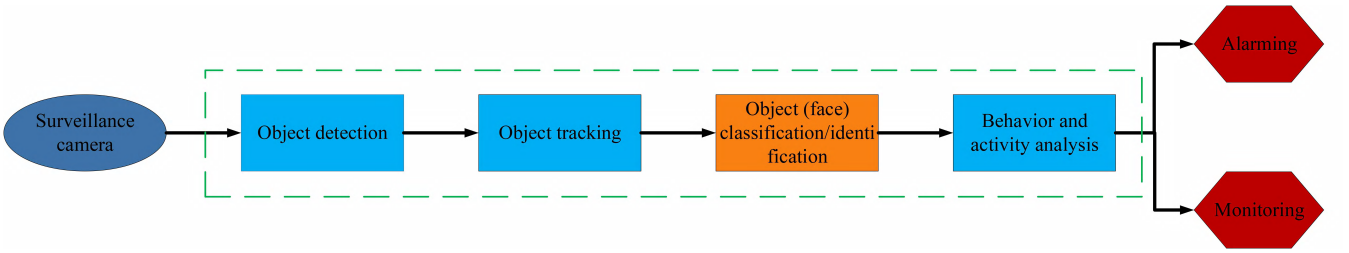

FIGURE 1. Four tasks in smart surveillance system.

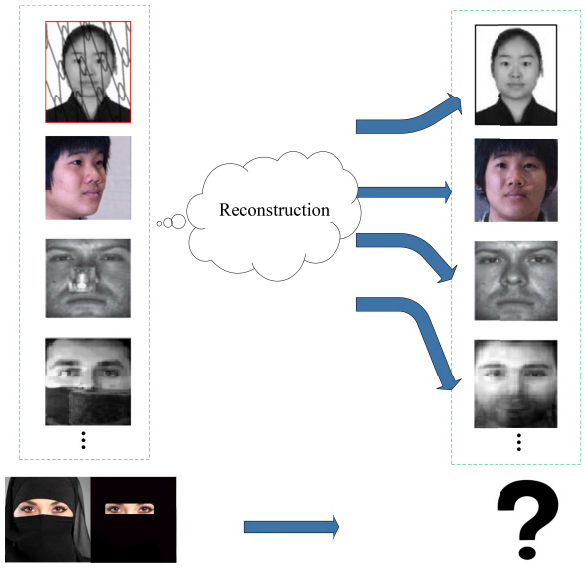

FIGURE 2. Face generation based on partial infomation.

existing face recognition algorithms. In the field of public safety, criminal investigation and anti-terrorism for instance, it is normal that the terrorists wear masks or other obstructions.

To deal with the issue of faces information loss from the video and to make the surveillance system smarter for better human identification, there are some studies turning to synthesize new faces from the original faces and perform the face recognition task based on the synthesized images. These studies are used as preprocess methods, which can effectively improve the recognition rate of face recognition.

Fig. 2 shows four typical cases of face image with information-missing. From top to bottom, the first face is blocked with occlusions incurred by random meshes, the second appears to be a side face, in the third image the nose is blocked by another image, and in the fourth image the face is blocked by a half-mask. There are some studies in the literature have made improvements to these problems (results on the right of figure) [6]–[9]. However, when it comes to the case that only eyes information available as shown in the fifth row, the existing methods mentioned above fail to work. This study is aimed at providing new solution to achieve better performance in human identification where only eyes information is available. We proposed a new approach to synthesize face from eyes region only (as shown in the fifth row in Fig. 2) for personal identification in human-centered smart surveillance system. In the proposed approach, an end-to-end neural network based on conditional GAN is designed to directly develop a mapping between eyes image and face image, based on which new face images can be synthesized for preserving the photorealistic and identity information. A specific loss function for face synthesis is proposed to effectively persist the consistency between synthesized faces and ground-truth faces. Both subjective and objective experimental results show that the synthesized faces are realistic and with high consistency to the ground-truth, with the average cosine similarity in celebfaces attributes (CelebA) and LFW dataset are both above 80%.

The rest of the paper is organized as follows. The recent related works on image inpainting and face synthesis are introduced in section II. Section III illustrates the network for eyes-to-face synthesis and its loss function. In section IV, we explained the process of eyes-to-face dataset production and experimentally verify our method for face synthesis. Finally, conclusion and future works are made in Section V.

II. RELATED WORK

Eyes to face synthesis can be considered as a generalized problem of reconstructing image from partially available information. So, face synthesis from eyes is similar to a classic task called image inpainting, which has found in many applications from the restoration of damaged paintings and photographs to the removal/replacement of selected objects [10], [11]. The blocked faces can be regarded as the face images needed to be restoration, and the restoration is the process of removal occlusion on face. Then the restored faces can be used for face recognition. Some methods have been developed to remove sparse occlusion or small rectangular occlusion [12]–[14], and use it for face recognition. The methods developed in the literature can only deal with simple occlusion issues, and the results are not satisfactory in case when only the eyes information is available.

With the rapid development of deep learning, great progress has been made in image inpainting. Based on an end-to-end network, face inpainting is able to restore higher quality faces and solve more complex problems [15]. Especially, as GAN was presented in 2014 by Goodfellow [16], it has

been successfully applied to vision tasks including image inpainting [17], super-resolution [18] and style transfer [19]. Deepak Pathak' work [17] proposed a new network called Context Encoders to produce a plausible hypothesis for the missing part(s), which are covered by a white square in the images. However, the results of Context Encoders are still not satisfactory because there are always light white square or ghosting artifacts in the synthesis results. Raymond A. Yeh's work [20] and yang's work [21] processed improved networks, but there are still some blur and ghosting region in the recovered images.

Apart from the approach of image inpainting, other face synthesis methods based on generative model have been provided to achieve better performance in case of more information loss. Some interesting generative model has been proposed based on GAN [22], [23] to generate photorealistic faces, which take a vector of random noise (e.g., a fixed dimensional uniform distribution noise) as the models' input. Since these methods' inputs are random noise, their output faces' identity is unsure or absolutely new. Though it is an interesting application to synthesize photorealistic faces from random noise, its uncertain inputs and outputs cannot satisfy the needs for face recognition. Some improved methods have been proposed to address this issue, by means of introducing conditions in the inputs to constrict the uncertain outputs. Conditional face synthesis methods were thus developed based on conditional GAN [7], [19], [24], [32]. Lu *et al.* [19] and Li *et al.* [32] both focus on the problem of generating human face pictures with specific attributes (e.g., eyeglasses and smiling). Huang's work [7] presented a novel network to generate a photorealistic front view from a profile face image and then use the front view for face recognition. However, the network is only designed for face rotation and not appropriate to eyes-to-face task. To our best knowledge, there is no study concerning on face generation from the data of eyes.

III. APPROACH

The aim of face synthesis from eyes is to synthesize a photorealistic and identity preserving face image I^F from an eye image I^E, by forming a mapping function $F(x)$ (as shown in Fig. 3). $F(x)$ can synthesize the corresponding face from a given eyes images. It can be formulated as follows:

$$I^F = F(I^E). \tag{1}$$

We use deep neural networks to realize such a mapping function $F(x)$. To train such a network, pairs of corresponding $\{I^E, I^F\}$ from multiple identities y are required during the training phase. Both the input I^E and output I^F come from pixel spaces.

Specifically, we model the synthesis function using an encoder-decoder network G_{θ_G} that is parametrized by θ_G. The network's parameters θ_G are optimized by minimizing a specifically designed synthesis loss L_{syn}. For a training set with N training pairs of $\{I_n^F, I_n^E\}$, the optimization problem

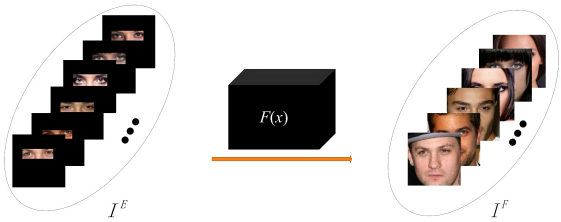

FIGURE 3. Mapping from eyes space I^E to face space I^F.

can be formulated as follows:

$$\hat{\theta}_G = \frac{1}{N}\underset{\theta_G}{\operatorname{argmin}} \sum_{n=1}^{N} L_{syn}(G_{\theta_G}(I_n^E), I_n^F), \tag{2}$$

where L_{syn} is defined as a weighted sum of several losses that jointly constrain an image to reside in the desired manifold. We will postpone the detailed description of all the individual loss functions later.

A. NETWORK ARCHITECTURE

The network proposed in this study is shown in Fig. 4. The entire network is a conditional GAN network, including two parts: generator and discriminator. The generator takes an eyes image as input and synthesize a corresponding face image. The discriminator is used to discriminate the synthesized faces and ground-truth faces. Generator and discriminator are trained together. The min-max two-player game provides a simple yet powerful way to estimate target distribution and generate novel image samples [22].

1) FULL NETWORK DESIGN

In order to generate more realistic human faces, conditional GAN was utilized to design the eyes-to-face synthesis network. It contains two parts, including generator and discriminator. The generator is the main part, which generate coincident faces from the input eyes. We use an end-to-end network called U-net [25], [26] as generator G_{θ_G}, which will be described further later. To incorporate prior knowledge of the synthesis faces' distribution into the training process, we further introduce a discriminator D_{θ_D} to distinguish ground-truth face images I^F from synthesized ones $G_{\theta_G}(I^E)$. In the discriminator, we use a general convolutional neural network (CNN). We merge synthesized faces and ground-truth together in the color channel, then take the merged images as the inputs of a 5-layers-CNN. Finally, the discriminator create new feature maps, which are used for discriminative loss. We train D_{θ_D} and G_{θ_G} to optimize the following min-max problem:

$$\min_{\theta_G}\max_{\theta_D} \mathbb{E}_{I^F\sim P(I^F)} \log D_{\theta_D}(I^F) + \mathbb{E}_{I^E\sim P(I^E)} \log(1 - D_{\theta_D}(G_{\theta_G}(I^E))), \tag{3}$$

where $P(I^E)$ and $P(I^F)$ are both fixed distributions, and $I^F \sim P(I^F)$ means I^F meets the distribution of $P(I^F)$. Solving the

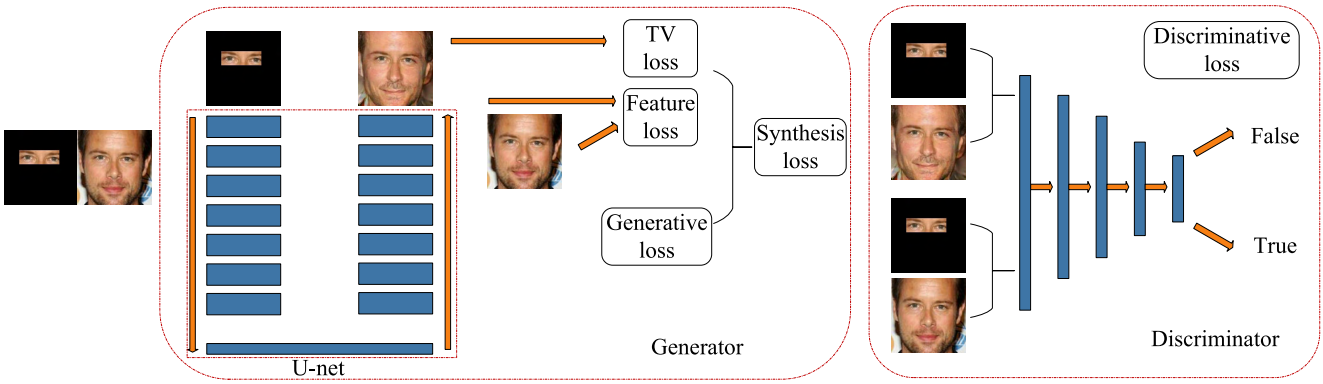

FIGURE 4. The proposed network for face synthesis from eyes.

min-max problem is to push the outputs of the generator to match the target distribution of the training synthesized faces, leading to photorealistic synthesis with appealing high frequency details.

2) GENERATOR DESIGN

In the generator, we use the U-net for the synthesis of face images. Fig. 4 shows the U-net network (the network in generator), which has become a common neural network in image segmentation and image generation [26], [27]. In the proposed network, the input is a 256×256 pixels eyes image and the output is a face image with the same size. The U-net's encoder and decoder process are totally symmetric. Both the encoder and decoder have 8 layers, and every layer includes a convolutional layer and a batch-normalize layer. The size of convolution filters is 4×4, and its strider is 2. Therefore, the feature map's size will become half of its original size in encoder, and it is no need to use a pooling layer to modify feature map size. In decoder, every deconvolution layer is just opposite, every layers' feature maps are twice the size of its previous layer. This mechanism contributes to losing less image details than using pool layers to reduce the image size.

B. LOSS FUNCTION

Generally speaking, L1 or L2 distance loss [26], [17] is the first option for an end-to-end network where the output is a image, like image segmentation and image translation. In addition, to get full perception of images, perceptual loss, content loss and style loss are also used, which are firstly introduced in image style transfer. However, for eyes-to-face task, we find that L1/L2 loss in image is not suitable.

Instead, the proposed method uses the synthesis loss function to guide the generator's training, which is a weighted sum of 3 individual loss functions, including feature loss, generative loss and total variation loss. Feature loss helps to preserve the consistency between synthesized faces and ground-truth faces, and generate photorealistic faces. Generative loss is the inherent loss of GAN and total variation (TV) loss contributes to better detail. We will give a detailed description in the following sections.

1) FEATURE LOSS

Inspired by image style transfer and uncertainty of image generation, we introduce a loss function called "feature loss" to guide the training based on Upchurch' work [28]. Different from some end-to-end tasks such as image translation, semantic segmentation and image inpainting. We focus more on the face consistency between synthesized faces and ground-truth. In other word, we focus on better face recognition rather than reconstruct original face. In pixel space, natural images lie on an (approximate) non-linear manifold [28]. Non-linear manifolds are locally Euclidean, but globally curved and non-Euclidean [28]. Bengio *et al.* [29] hypothesize that convnets linearize the manifold of natural images into a (globally) Euclidean subspace of deep features. Hence we can calculate the difference between synthesized faces and ground-truth in a linear way in feature space generated from a pre-trained convolutional model. We input the synthesized faces and ground-truth faces together into a pre-trained VGG19 model [30], and then calculate the L1 loss of every feature maps between synthesis faces and ground-truth faces. Finally we sum all the L1 loss in feature maps and use it as the feature loss. The feature loss will guide the generator to persist the identification of the synthesized faces. The feature loss of a face image takes the form:

$$L_{feature} = \sum_{i=1}^{3} \frac{1}{Wi \times Hi} \sum_{x=1}^{Wi} \sum_{y=1}^{Hi} |Ii_{x,y}^{pred} - Ii_{x,y}^{gt}|. \quad (4)$$

Noted that the equation is simplified, which meets the situation of just a feature map in each layer. In fact, there are many feature maps in every layer in VGG19. For simplicity, we choose 3 feature layers in VGG19 convolutional layer to calculate the L1 distance loss, whose name are Conv3-1, Conv4-1 and Conv5-1, which can be seen in Fig. 5. $Ii_{x,y}^{pred}$ is pixel value at points (x,y) in synthesized faces, $Ii_{x,y}^{gt}$ is pixel value at points (x,y) in ground-truth. And the Ii^{pred} and Ii^{gt} is the relatively feature maps in VGG19 models, which

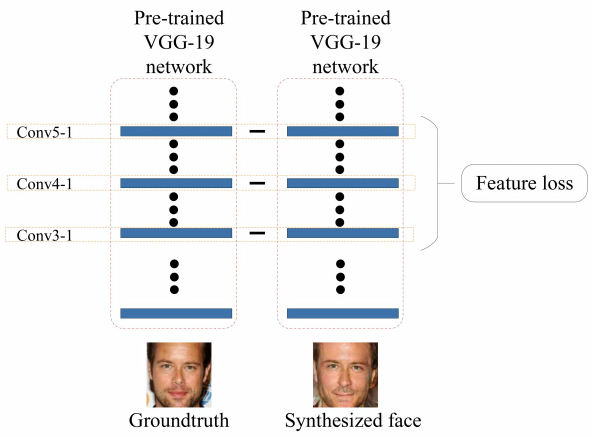

FIGURE 5. Feature loss based on pre-trained VGG19 network.

is already trained on Imagenet dataset. $L_{feature}$ is used for guiding the training process, telling the network what's the deference between synthesized faces and ground-truth face in feature space, which helps avoid abnormal face results.

2) GAN LOSS

Our network is based on conditional GANs. For a GAN networks, the generative loss and discriminative loss are most important. In this paper, the generative loss L_G and discriminative loss L_D can be represented as follows:

$$L_G = E_{x\sim P_{data}(x),z\sim P_z(z)}[-\log G(x,z)], \tag{5}$$

$$L_D = E_{y\sim P_{data}(y)}[-\log D(x,y)] + E_{x\sim P_{data}(x),z\sim P_z(z)}[\log(1-D(x,G(x,z)))]. \tag{6}$$

In our network, L_G is one of the synthesis losses guiding the training process. Besides, L_D loss is used for training discriminator and serves as a supervision to push the synthesized image to reside in the manifold of face images. It can prevent blur effect and produce visually pleasing results [7].

3) TOTAL VARIATION LOSS

Finally total variation loss L_{tv} [31] is also included, in order to avoid a mutation in the synthesis faces. This loss is also used for image style transfer initially. The L_{tv} can constrain the pixel changes in the generated results y and encourages smoothness, which can be expressed as follows:

$$L_{tv} = \sqrt{(y_{i+1,j}-y_{i,j})^2 + (y_{i,j+1}-y_{i,j})^2}. \tag{7}$$

4) FINAL SYNTHESIS OBJECTIVE FUNCTION

The final synthesis loss function is a weighted sum of the losses defined above:

$$L_{syn} = \lambda_1 L_{feature} + \lambda_2 L_G + \lambda_3 L_{tv}. \tag{8}$$

IV. EXPERIMENTAL RESULTS AND DISSCUSION

The experimental settings for subjective and objective evaluation of the proposed method are first discussed in this section. Qualitative and quantitative results of the synthesis experiments are presented for evaluation.

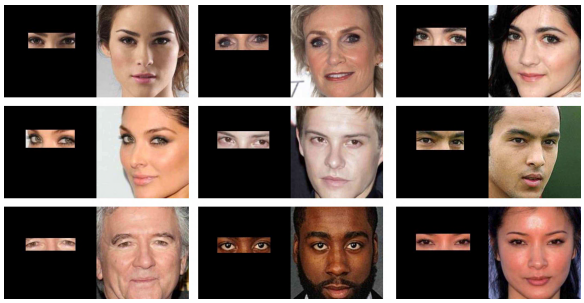

FIGURE 6. Examples in eyes-face dataset, include side faces and front view faces, across different ages, color, face expression, poses and gender.

We evaluate our method based on 3 datasets, including CelebA, LFW and a dataset created by ourselves. There are more face image data in CelebA than the LFW, with better quality. Since there is no off-the-shelf dataset for eyes-to-face synthesis, we firstly made an eyes-face dataset based on CelebA dataset. The dataset includes: different resolution, age, gender, color of skin, facial expression and poses faces. A sample of eyes-face images is shown in Fig. 6. To establish the dataset we firstly we extract the faces using the Open Source Computer Vision Library (OpenCV) from the original image; secondly, we reshape faces images' size to 256×256, and then extract eyes' region in the face; finally, we place the eyes images and the corresponding face images together as image pairs with the eyes in the left and face in the right, obtaining a size of 512×256 image. After removing some images whose eyes were detected by mistake, we finally obtained 13220 eyes-face pair images. In our experiments, we randomly selected 12605 images as training dataset, with the rest 615 images as test dataset. Through the same methods, we also create a dataset based on LFW. Besides, we create a new dataset which contain blocked faces images.

The experiments are design as follows:

- To verify the effect of feature loss, we compare the results synthesized by L1 loss and feature loss(other losses in synthesis loss are the same).
- To evaluate the stability and generalization of our methods, we test our methods on the aforementioned datasets.
- To verify the faces consistency objectively and demonstrate how the proposed work can conserving the human identity, we calculate the euclidean distance and cosine similarity between synthesized faces and ground-truth based on a pretrained face verification network.

The proposed network is trained on a single TitanX GPU for 400 epochs. In all our experiments, we empirically set $\lambda_1 = 0.05$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, and the learning rate of generator and discriminator are both 0.0002.

A. COMPARISON OF THE OBJECTIVE FUNCTION

Different from existing work, the eyes-to-face synthesis for face recognition dose not require that the reconstructed faces

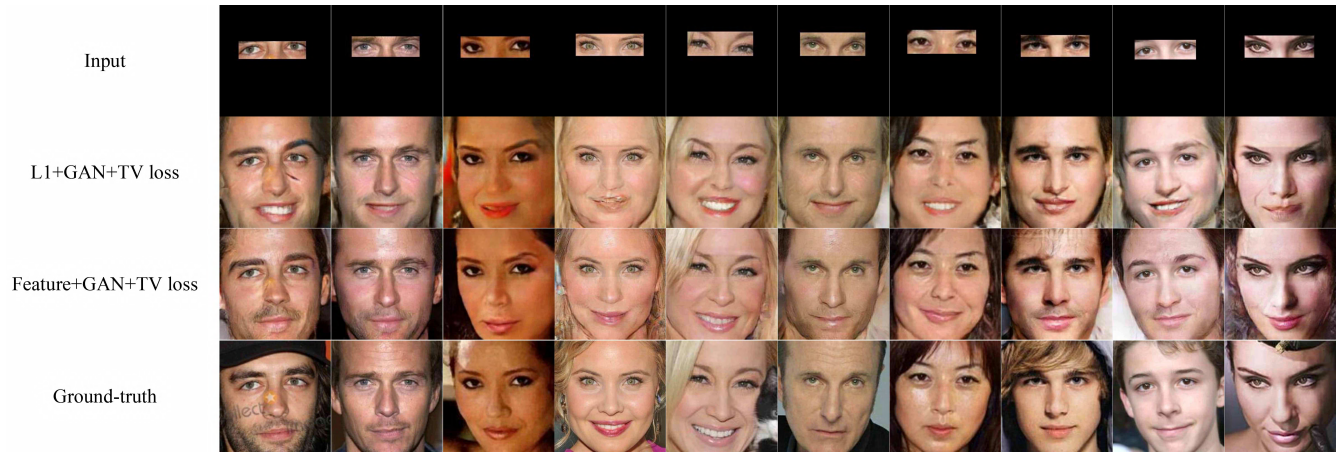

FIGURE 7. Comparison between L1 loss synthesized results and feature loss synthesized results.

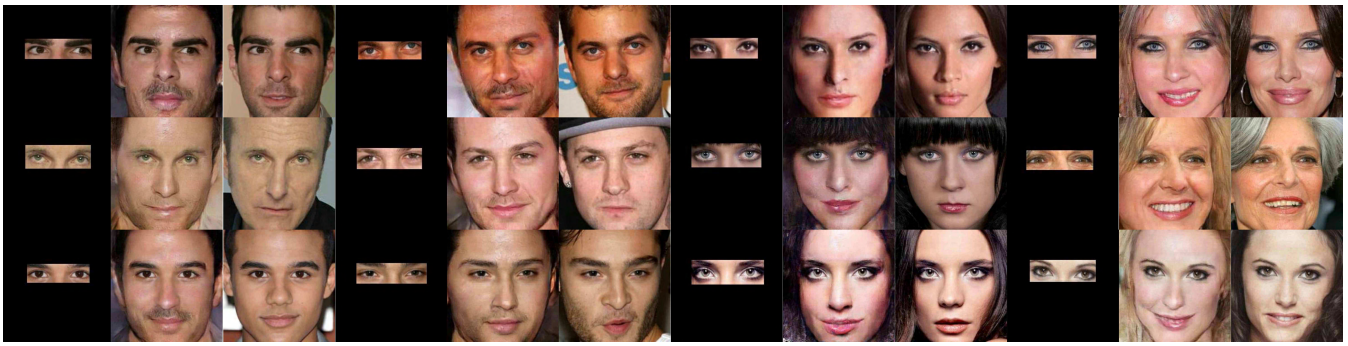

FIGURE 8. Synthesis results by our method on CelebA dataset. For every image triple, the left column is input eyes, the middle column is synthesized faces and the right column is ground-truth.

are same as ground-truth in every pixel values. Instead, we aim to generate a coincident face with the ground-truth for better face recognition.

Previous work demonstrated that L1 distance loss encourages less blurring than L2 [26]. In order to verify the priority of feature loss, we compare L1 loss and feature loss. The results presented in Fig.7 is the comparison based on CelebA's test dataset. Two networks are both trained for 400 epochs and they are no different except L1/feature loss. The feature loss show achieving more satisfactory results comparing to L1 loss. In addition, when we combined L1 loss, the synthesized faces always appear ghosting and unnatural, with the training loss L_{syn} hard to be reduced.

B. EVALUATE STABILITY AND EFFECTIVENESS OF FACE SYNTHESIS

1) FACE SYNTHESIS ON CELEBA

We first test our method on the CelebA test dataset. The synthesized faces are shown in Fig. 8. To illustrate the performance, we show the synthesized results of 6 man and women which showed that our method can generate photorealistic faces.

The synthesized faces do not appear totally same as the ground-truth faces. However, the two faces are like each other to some extent. This illustrates that the proposed method can extract consistent attributes of faces, including gender, color of skin, angles, local facial parts (e.g., eyes, nose and mouth), emotion, facial expression and so on. For example, when the person in ground-truth is smiling, the generated face is also smiling.

2) FACE SYNTHESIS ON LFW

To evaluate the stability and generalization of the proposed approach, we apply the model trained based on CelebA training dataset to synthesize faces in LFW dataset. We employ the same method to process the original LFW dataset and get 6730 eyes-face samples. Compared with CelebA, the LFW dataset are with relatively worse quality. The input eyes, synthesis faces and the ground-truth faces are presented in Fig.9. It can be seen that most of LFW faces are not as clear as CelebA images. The synthesized faces' quality is not so well as the results in CelebA faces. However, it is seen that the synthesized faces' quality is as the same level of ground-truth images. We hypothesize that the quality of the synthesized faces is dependent of the quality of input eyes. Hence, high-resolution eyes would generate high-resolution faces, and low resolution eyes synthesize low-resolution faces. Though the model was trained based on CelebA faces, it can produce

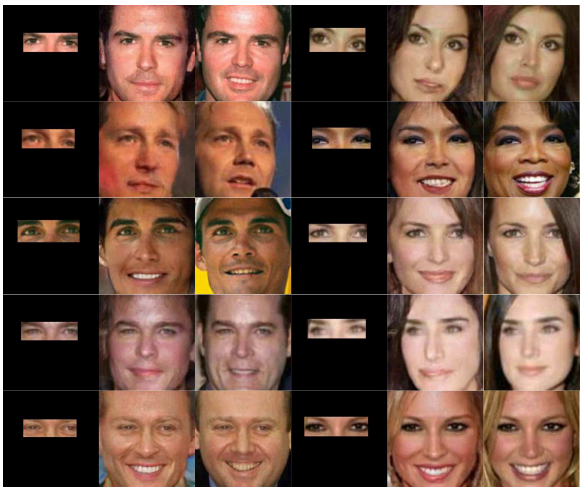

FIGURE 9. Examples of our methods' results in LFW dataset. For every image triple, the left column is input eyes, the middle column is synthesized faces and the right column is ground-truth.

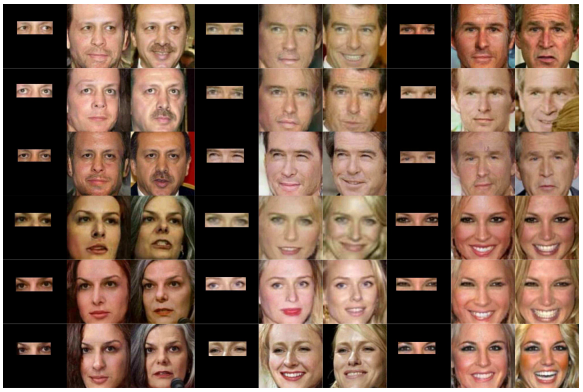

FIGURE 10. Examples of different poses' results. For every image triple, the left column is input eyes, the middle column is synthesized faces and the right column is ground-truth.

competitive results on different data such as LFW, which demonstrate the robustness of the proposed method.

3) FACE SYNTHESIS ON MULTIPLE POSES

To further demonstrate robustness of the proposed method under the condition of different poses, we conducted experiments on face image with different poses. The source data of face images is from LFW dataset, as there are many different pictures in some LFW persons' file. We choose 6 persons in LFW dataset, and for each person we choose 3 different poses faces for generation. As shown in Fig.10, the synthesized faces' poses are almost the same as the ground-truth images, regardless of the incline angle. Moreover, the generated faces from same person are similar to each other. It should be noted that our method does not rely on any 3D knowledge, the inference and synthesis is made through purely data-driven learning.

4) FACE SYNTHESIS ON OUR OWN BLOCKED DATASET

In order to demonstrate the practical value of the proposed method in applications, we collected a new face dataset and processed it like before. It includes 200 half masked or blocked faces, with half masked faces as the main parts. For samples in this dataset, we do not actually have the corresponding ground-truth for the eyes since the faces are all blocked. Visual inspection on the results shown in Fig.11, it shows that our method obtained satisfactory synthesized faces, demonstrating the stability and generalization of our method.

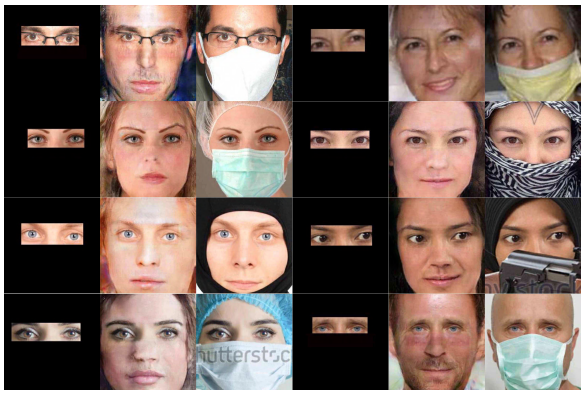

FIGURE 11. Results of blocked faces from our own dataset. For every image triple, the left column is input eyes, the middle column is synthesized faces and the right column is ground-truth.

C. EVALUATE CONSISTENCY OF FACE SYNTHESIS(IDENTITY PRESERVING)

In order to evaluate the consistency of synthesized faces and ground-truth faces objectively, we design 3 assessment criterias, based on the idea of [28]. As it is improper to compare the difference in pixel space, we extended it to define the assessment in a feature space with a pre-trained classification CNN network. We use a pre-trained face recognition model to assess the difference between generated faces and label faces. We apply the Facenet model [5] to extract 128-byte feature vectors of synthesized faces and ground-truth faces, then calculate the euclidean distance and cosine similarity between them. As suggested in Facenet [5], the threshold value to verify a person is set to be 1.1. When the euclidean distance between two faces is smaller than 1.1, they are considered to be the same person. Based on this, we design another criterion called synthesis accuracy percentage, which is the probability of the consistent faces in the entire faces dataset. We experimented on CelebA and LFW dataset, the results are presented in Table. 1. We also compare the L1 loss and feature loss with such objective criterion in Table. 2. The overall euclidean distance for CelebA dataset in Fig.12. It is seen that the synthesized faces are of high consistency to the ground-truth faces. The average cosine similarity between synthesized faces and ground-truth in CelebA and LFW are both higher than 80%. The LFW dataset is created by the same method with CelebA. Nevertheless, with larger scale sample (6730) and worse quality, our method still obtained a satisfactory result. When choosing 1.1 as the threshold, the synthesis accuracy percentage is higher than 91%,

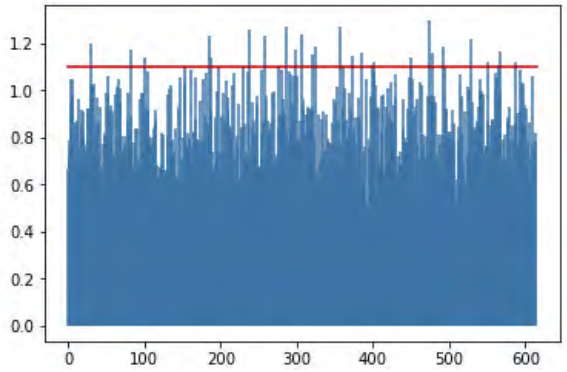

FIGURE 12. Euclidean distance between synthesized faces and their ground-truth faces in 615 CelebA test dataset.

TABLE 1. Objective results for CelebA and LFW dataset.

Criterion	CelebA	LFW
Average euclidean distance	0.8002	0.8686
Average cosine similarity	83.26%	80.51%
Synthesis accuracy percentage	95.29%(615)	91.87%(6730)

TABLE 2. Objective results for comparison between L1 loss and feature loss.

Criterion	L1 loss	Feature loss
Average euclidean distance	0.8458	**0.8002**
Average cosine similarity	81.30%	**83.26%**
Synthesis accuracy percentage	89.43%(615)	**95.29%(615)**

which means that we could get a recognition accuracy higher than 91%. Besides, the objective comparison results between two loss function are in accord with subjective results, according to Table. 2.

D. ALGORITHMIC ANALYSIS

In this section, We mainly analyze the overall performance of our approach and some limitations. As is shown before, our method can synthesize in essence of the attributes with ground-truth, include: gender, color of skin, facial expression, face poses and so on. Hence, our methods also can be a basis for gender classification, facial expression classification and other face related work. We further compare the similarity between the synthesized faces and the ground-truth faces by euclidean distance and discrete cosine distance. The results show that the proposed method can provide a potential solution for face recognition.

Of course, there are still several problematic synthesized results. Some poor quality synthesized faces are presented in Fig.13. Through analyses on a large number of synthesized faces, we find that our method still have some limitations, which can be further improved. Firstly, when the angle of side face is big and close to 90°, the intercepted eyes region is incomplete, this would lead to blur or ghosting results.

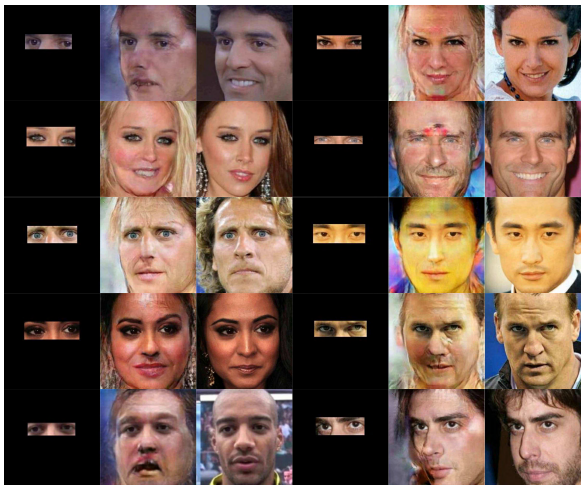

FIGURE 13. Some typical poor quality synthesized faces. For every image triple, the left column is input eyes, the middle column is synthesized faces and the right column is ground-truth.

Secondly, based on the face images of European and American people in the training dataset, the synthesized faces of Asian and African people is not very well. This can be enhanced by expand training dataset. Furthermore, when there are abnormal color or resolution as inputs, the synthesized faces is shown also to be abnormal.

V. CONCLUSION AND FUTURE WORK

To tackle the occlusion in face recognition, we proposed a new approach for eyes-to-face synthesis based only on the eyes information, specifically for human identification in human-centered smart surveillance system. To realize such a task, we design an end-to-end neural network based on conditional GAN. We use U-net to synthesize faces from eyes, and improve the training process by means of conditional GAN. To enhance the synthesis, we further introduce a synthesis loss which is a sum of feature loss, generative loss and total variation loss in the training process. The feature loss compare the difference of them in feature space, which contributes to identity-preserving and more photorealistic faces. Experimental results demonstrate that the proposed method not only presents compelling perceptual results but also have important robustness, which enables the development of human-centered smart surveillance system, e.g. for the scenarios of criminal identification and tracking. Our methods also can be a basis for gender classification, facial expression classification, face analysis and other face related work. In the future, we will introduce more information to synthesize more realistic and consistent faces, e.g. facial contour information from masked persons.

REFERENCES

[1] A. B. Hamida, M. Koubaa, H. Nicolas, and C. B. Amar, "Video surveillance system based on a scalable application-oriented architecture," *Multimedia Tools Appl.*, vol. 75, no. 24, pp. 17187–17213, 2016.
[2] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with GaussianFace," in *Proc. AAAI*, 2015, pp. 3811–3819.
[3] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. CVPR*, Jun. 2014, pp. 1891–1898.

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, Jun. 2014, pp. 1701–1708.
[5] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, Jun. 2015, pp. 815–823.
[6] S. Zhang, R. He, Z. Sun, and T. Tan, "DeMeshNet: Blind face inpainting for deep MeshFace verification," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 637–647, Mar. 2018.
[7] R. Huang *et al.*, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. ICCV*, Oct. 2017.
[8] Y. F. Yu *et al.*, "Discriminative multi-scale sparse coding for single-sample face recognition with occlusion," *Pattern Recognit.*, vol. 66, pp. 302–312, Apr. 2017.
[9] G. Gao, J. Yang, X.-Y. Jing, W. Yang, D. Yue, and F. Shen, "Learning robust and discriminative low-rank representations for face recognition with occlusion," *Pattern Recognit.*, vol. 66, pp. 129–143, Jun. 2017.
[10] M. Bertalmio, G. Sapiro, C. Ballester, and V. Caselles, "Image inpainting," in *Proc. Int. Conf. Comput. Graph. Interactives Techn.*, 2000, vol. 4. no. 9, pp. 417–424.
[11] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 460–489, 2005.
[12] Z. Mo, J. P. Lewis, and U. Neumann, "Face inpainting with local linear representations," in *Proc. BMVC*, 2004, pp. 1–10.
[13] R. Min and J. L. Dugelay, "Inpainting of sparse occlusion in face recognition," in *Proc. ICIP*, 2011, pp. 1425–1428.
[14] N. Batool and R. Chellappa, "Detection and inpainting of facial wrinkles using texture orientation fields and Markov random field modeling," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3773–3788, Sep. 2014.
[15] M. Jampour, C. Li, L.-F. Yu, K. Zhou, S. Lin, and H. Bischof, "Face inpainting based on high-level facial attributes," *Comput. Vis. Image Understand.*, vol. 161, pp. 29–41, Aug. 2017.
[16] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. NIPS*, vol. 3. 2014, pp. 2672–2680.
[17] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. CVPR*, 2016, pp. 2536–2544.
[18] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, 2016, pp. 4681–4690.
[19] Y. Lu, Y.-W. Tai, and C.-K. Tang. (2017). "Conditional CycleGAN for attribute guided face image generation." [Online]. Available: http://arxiv.org/abs/1705.09966
[20] 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. CVPR*, 2017, pp. 6882–6890.
[21] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. CVPR*, Jul. 2017, pp. 4076–4084.
[22] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. NIPS*, 2015, pp. 1486–1494.
[23] D. Berthelot, T. Schumm, and L. Metz. (2017). "BEGAN: Boundary equilibrium generative adversarial networks." [Online]. Available: http://arxiv.org/abs/1703.10717
[24] M. Mirza and S. Osindero. (2014). "Conditional generative adversarial nets." [Online]. Available: https://arxiv.org/abs/1411.1784
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[26] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Jul. 2017, pp. 5967–5976.
[27] X. Di, V. A. Sindagi, and V. M. Patel. (2017). "GP-GAN: Gender preserving GAN for synthesizing faces from landmarks." [Online]. Available: https://arxiv.org/abs/1710.00962
[28] P. Upchurch *et al.*, "Deep feature interpolation for image content changes," in *Proc. CVPR*, 2016, pp. 6090–6099.
[29] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai, "Better mixing via deep representations," in *Proc. ICML*, 2012, pp. 552–560.
[30] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[31] J. Johnson, A. Alahi, and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.
[32] M. Li, W. Zuo, and D. Zhang. (2016). "Convolutional network for attribute-driven and identity-preserving human face generation." [Online]. Available: http://arxiv.org/abs/1608.06434

XIANG CHEN received the B.S. degree in electronic information engineering from Sichuan University, Chengdu, China, in 2016, where he is currently pursuing the master's degree. He has acted as the main participant in several projects in the areas of artificial intelligence, e.g., (Identification of Unsound Kernels in Wheat and Suspects Tracking in Surveillance System). His research interests include machine learning, computer vision, deep learning, video caption, and face synthesis.

LINBO QING (M'16) received the B.S. degree in electronic information science and technology and the Ph.D. degree in communication and information system from Sichuan University, China, in 2003 and 2008, respectively. He is currently an Associate Professor with the College of Electronics and Information Engineering, Sichuan University. His main research interests include image processing, video coding and transmission, and information theory.

XIAOHAI HE (M'16) received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in biomedical engineering from Sichuan University, Chengdu, China, in 1985, 1991, and 2002, respectively. He is currently a Professor with the College of Electronics and Information Engineering, Sichuan University. His research interests include image processing, pattern recognition, computer vision, image communication, and software engineering. He is a Senior Member of the Chinese Institute of Electronics. He is an Editor of the *Journal of Information and Electronic Engineering* and the *Journal of Data Acquisition & Processing*.

JIE SU received the B.S. degree in electronic science and technology from the Chengdu University of Information Technology in 2012, and the M.S. degree in signal and information processing from Sichuan University, Chengdu, China, in 2015, where she is currently pursuing the Ph.D. degree in communication and information system. Her research interests include image processing, machine learning, pattern recognition, computer vision, and deep learning.

YONGHONG PENG (M'02) is currently a Professor of data science and the Leader of data science research with the University of Sunderland, U.K. His research areas include data science, machine learning, data mining, and artificial intelligence. He is a member of Data Mining and Big Data Analytics Technical Committee of the IEEE Computational Intelligence Society. He is also a Founding Member of the Technical Committee on Big Data of the IEEE Communications and an Advisory Board Member of the IEEE Special Interest Group on Big Data for Cyber Security and Privacy. He is the Chair of the Big Data Task Force. He is an Associate Editor for the IEEE TRANSACTION ON BIG DATA, and an Academic Editor of *PeerJ* and *PeerJ Computer Science*.