Facial Expression Synthesis by U-Net Conditional Generative Adversarial Networks

Xueping Wang
Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
Beijing, China
xpwang@buaa.edu.cn

Weixin Li*
Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
Beijing, China
weixinli@buaa.edu.cn

Guodong Mu
Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
Beijing, China
muyouhang@buaa.edu.cn

Di Huang
Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
Beijing, China
dhuang@buaa.edu.cn

Yunhong Wang
Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
Beijing, China
yhwang@buaa.edu.cn

ABSTRACT

High-level manipulation of facial expressions in images such as expression synthesis is challenging because facial expression changes are highly non-linear, and vary depending on the facial appearance. Identity of the person should also be well preserved in the synthesized face. In this paper, we propose a novel U-Net Conditioned Generative Adversarial Network (UC-GAN) for facial expression generation. U-Net helps retain the property of the input face, including the identity information and facial details. We also propose an identity preserving loss, which further improves the performance of our model. Both qualitative and quantitative experiments are conducted on the Oulu-CASIA and KDEF datasets, and the results show that our method can generate faces with natural and realistic expressions while preserve the identity information. Comparison with the state-of-the-art approaches also demonstrates the competency of our method.

KEYWORDS

Facial expression synthesis; generative adversarial network (GAN); identity preserving

ACM Reference Format:
Xueping Wang, Weixin Li, Guodong Mu, Di Huang, and Yunhong Wang. 2018. Facial Expression Synthesis by U-Net Conditional Generative Adversarial Networks. In *ICMR '18: 2018 International Conference on Multimedia Retrieval, June 11–14, 2018, Yokohama, Japan.* , 8 pages. https://doi.org/10.1145/3206025.3206068

*indicates the corresponding author.

ICMR '18, June 11–14, 2018, Yokohama, Japan

ACM ISBN 978-1-4503-5046-4/18/06…$15.00
https://doi.org/10.1145/3206025.3206068

1 INTRODUCTION

Facial expression synthesis (FES) generally aims to render a face image of a neutral expression with different universal ones, i.e., happiness, sadness, fear, anger, disgust, and surprise. FES has received increasing attention in recent years from both the academia and industry, not only due to its scientific challenges, but also for its wide-range of applications, e.g., human-computer interactions, animation and facial reenactment [30].

The key issue in FES is to transfer expressions by reconstructing texture and deforming shape of the given face, and this is a rather difficult task as the expressions produced are expected to be natural and with typical texture and shape patterns, in particular in the presence of variations of ambient illumination, head pose, image resolution, etc. For example, surprise usually makes the mouth open and the eyes become bigger, while disgust only slightly varies facial geometry but tends to present more detailed wrinkles on the areas around the nose and between the eyebrows. Furthermore, corresponding input and output faces are required to share the same identity, i.e. preserving identity cues after expression editing. In literature, there exist a number of investigations which focus on this topic, and the last two decades have witnessed its development from hand-crafted methods to deep learning based ones.

Specifically, the most majority of the early studies are hand-crafted, and they manipulate faces by warping facial components according to the difference calculated from training samples [17, 18, 38]. Although they sometimes demonstrate promising results, even of high-resolution, modeling differences between expressions needs paired data where one subject has the samples of different expressions. Meanwhile, the face warping phase depends heavily on locations of fiducial landmarks, whose detection accuracies are limited by the discriminative power of traditional features. More importantly, the identity clue in the given face cannot be well protected due to global or local texture warping. Later, deep learning techniques have also been explored in this topic. High-order Boltzmann machine [26] and Flow Variational Auto-Encoder (FVAE) [35] are attempted; however, their faces synthesized are not visually satisfying, mainly because identity cues are lost more or less, making

the input and output faces do not look like each other. Expression Generative Adversarial Network (ExprGAN) is proposed for photo-realistic facial expression [5]. It explicitly emphasizes the necessity of identity in FES and adds an identity constraint term in the loss function, aiming to balance expression generation and identity preservation. The results achieved are largely superior to those in the previous work, but the basic encoder-decoder generator framework leaves much space for further improvement.

In this paper, we propose a novel GAN based FES apporach that takes a neutral face as input and synthesizes the ones of the same person, but with different universal expressions. Instead of the encoder-decoder network used in generator [5], we make use of the U-Net model. As U-Net shares low level information between input and output images by skip connections, it helps retain the property of the input, thus better preserving identity information in the faces of synthesized expressions. At meantime, we present a new identity preserving loss, which not only minimizes the similarity between the same person, but maximizes that between different ones as well, to further enhance the proposed model. Finally, we embed the Auxiliary Classifier GAN (AC-GAN) to the framework so that one-to-many (neutral to various expressions) synthesis can be reached simultaneously. Extensive experiments are carried out on the Oulu-CASIA and KDEF databases, and the results are state-of-the-art in terms of both expression editing as well as identity protection, which clearly indicates the effectiveness of the proposed method.

The remainder of this paper is organized as follows. We describe the related work in the next section. The proposed method is introduced in Section 3. Experimental results are shown and analyzed in Section 4. Finally, we conclude the paper in Section 5.

2 RELATED WORK

2.1 Facial Expression Synthesis

In recent years, facial expression synthesis has been widely studied in the field of computer vision. The traditional methods mainly include geometry-controlled image warping based approaches [18, 38] and morphing-based approaches [2, 24, 28]. The methods based on geometry controlled image warping use the feature difference vector of the source and target expression images to generate feature positions for a new face. They can only capture the feature motions of the face but ignore the facial detail changes brought by the changes of facial expressions. Morphing-based approaches [2, 24, 28], can only generate expressions in-between the given expression through interpolation and cannot be used to generate expressions for a new face. The deep learning-based methods can be divided into three subclasses from the perspective of generative modeling, namely Deep Belief Network (DBN) [26, 29], variational autoencoders (VAE) [3, 6, 13, 33, 35, 40], and generative adversarial networks (GAN) [4–7, 11, 14, 16, 23].

In the first subclass, Susskind et al. [29] used a deep belief net to generate expressions for a given identity with elementary facial expressions such as "raised eyebrows". Reed et al. [26] proposed a higher-order Boltzmann machine which incorporates multiplicative interactions to effectively disentangle the variation of facial expression and identity. However, the generated images of these methods are low-resolution with size of 48 x 48, which are not visually satisfying.

In the second subclass, Cheung et al. [3] used augmented autoencoders with a supervised cost and an unsupervised cross-covariance penalty to reatain subject identity in faces. Ghodrati et al. [32] adopted a convolutional encoder-decoder architecture to generate a similar and plausible facial image, based on some desired attributes. Then the images are refined by the refinement network using convolutional filters. Though this model can alter a face without human intervention and with noticeable accuracy, the synthesized facial images has low resolution (32×32) and tend to be blurry. Larsen et al. [14] combined variational autoencoder and generative adversarial networks together to encode, and generate face images with feature-wise error measures. And the latent image representation could disentangle factors of variation, such as simple arithmetic applied in the learned latent space producing images which refect changes in some attributes. Yan et al. [33] interpreted a facial image as a composite of the foreground and background, and adopted a layered generative model to generate object images from high-level description based on images and texts, by using a conditioned variational auto-encoder. Yeh et al. [35] proposed Flow Variational Autoencoder (FVAE) model to edit the facial expression using latent vector arithmetic. Although the generated face image has high resolution, paired data of one subject with different expressions are needed to obtain the face expression interpolation. Zhou et al. [40] proposed the conditional difference adversarial autoencoder (CDAAE) to generate a face image with a target emotion or facial action unit (AU) label. In order to preserve the person identity, a feedforward path was added to an autoencoder model connecting low level features in the encoder to features at the corresponding level in the decoder. Though the method preserved the person identity information, the generated facial images are of low resolution (64×64) and some expression details and hair information are lost.

In the third subclass, Li et al. [16] presented a deep convolutional network model to generate a facial image conditioned on the source input image and the reference attribute, which can preserve the referenced attribute and the same or similar identity of the input image. Choi et al. [4] utilized a mask vector method that enables a novel generative adversarial network named StarGAN to control all available domain labels and translate the input image and domain information into corresponding domain. Olszewski et al. [23] proposed a method to transfer the face of a single RGB target image to a source video sequence with the same expression. Kaneko et al. [11] proposed a conditional filtered generative adversarial network (CF-GAN) to generate or edit an image while intuitively controlling large variations of an attribute. Ding et al. [5] proposed an Expression Generative Adversarial Network (ExprGAN) for photo-realistic facial expression editing to control both the type of the expression and its intensity simultaneously. However, the images this method generates tend to be a little blurry, do not have much expression details and miss the original person identity.

2.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [7] are a type of parametric method that has been widely applied and studied for image

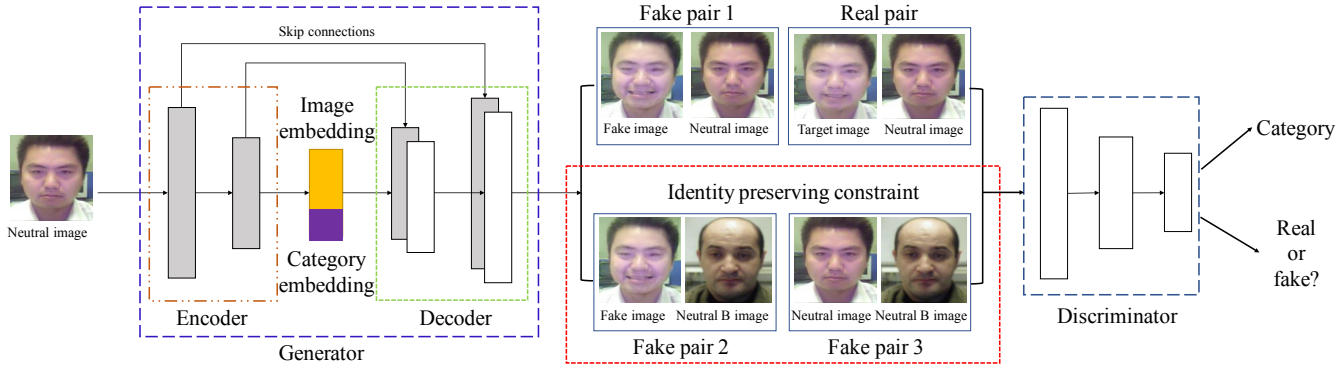


Figure 1: Illustration of the proposed UC-GAN model. Our model contains two parts: the generator G and the discriminator D.

synthesis. The main idea is to train paired generator and discriminator networks at the same time, where the goal of the discriminator is to classify between "real" images and generated "fake" images, and the generator aims to fool the discriminator so that the generated images are indistinguishable from real images. Once trained, the generator can be used to synthesize images driven by a compact vector of noise. Compared to the blurry and low-resolution outcome from other methods, GANs can produce sharp and plausible images. Conditional GAN (cGAN) is an extension of GAN with some conditioned information settings [21, 25]. Ledig et al. [15] presented SRGAN, a generative adversarial network (GAN) for image super-resolution (SR) with a perceptual loss function, which is capable of generating photo-realistic natural images for 4 × upscaling factors. Isola et al. [10] explored GANs in the image conditional setting for image-to-image translation in many tasks. This paper uses U-Net model as the generator which shares the low-level information between an input image and a corresponding output image. Odena et al. [22] proposed an auxiliary classifier GAN (AC-GAN) to generate globally coherent high resolution ImageNet samples, where every generated sample has a corresponding class label. Yang et al. [34] presented a novel face age progression method based on GAN. In order to achieve and enhance the aging details, a pyramidal adversarial discriminator was used to convey the high-level age-specific features. Our method is partly inspired by these works.

3 APPROACH

In this section, we describe the proposed UC-GAN for facial expression generation. We first introduce the cGAN, which our UC-GAN is built on. Then we explain the framework and formulation of UC-GAN in detail.

3.1 Conditional Generative Adversarial Networks

The conditional GAN (cGAN) [21] is an extension of the GAN model allowing the generation of images with some extra information y, such as class labels or data from other modalities. The structure of cGAN is just like GAN, including two "adversarial" models: a generative model G and a discriminative model D. G is optimized to capture the real data distribution based on random noise z which is hard for D to discriminate from real data x. D estimates the probability that a sample comes from the training data rather than G. D and G play the two-player minimax game with value function $V(D,G)$. The objective function is defined as

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

3.2 U-Net Conditioned Generative Adversarial Networks (UC-GAN)

The proposed UC-GAN model is illustrated in Fig. 1. As shown in the figure, the generator G is a U-Net conditioned model, including an encoder and a decoder. Different from the traditional U-Net model, the U-Net conditioned model transmits the encoded features, along with the conditional information, to the decoder. The UC-GAN takes neutral images I_N and target category embedding (labels) c as input and outputs a high confidence target facial image of the same person. Here we directly encode RGB facial images instead of latent random noises for two reasons. Firstly, if the GAN model generates images from random noises, the output images may not be controlled. This is undesirable in facial expression synthesis, where we have to ensure that the output face preserves the identity of the input face. Secondly, the conditioned images as a meaningful latent space, can maintain the sharp textures and person identity of the input facial images.

As shown in Fig. 1, the neutral image I_N is first encoded by an encoder E, yielding a facial identity embeddings $E(I_N)$. After combining it with the category embedding c, we now can have the GAN model that can handle multiple expressions at the same time. However, a new problem emerges: the model starts to confuse and mix the expression categories, generating faical images that don't look like any of the target emotions. Inspired by the AC-GAN model [22], we add the multi-class category loss to supervise the discriminator to penalize such scenarios, by predicting the facial expression categories of the generated facial images. The encoder maps the same facial expression into the same vector. The decoder, on the other hand, will take both the facial identity embedding and category embedding to generate the target facial expression.

The discriminator D aims to distinguish between target image and the generated image, and classify expression.

Identity Preserving Loss. Whether a GAN model works well mainly depends on the cooperation between the generator and discriminator. If the discriminator is too weak, it also results in low performance. Whereas, a comparatively strong discriminator can achieve more reasonable and nautral-looking facial expression details and clear facial profile [34]. For facial expression generation method, the person identity should not be changed while generating different facial images. So in terms of identity, it is desired to have the distance between the generated image and input image smaller than that between the generated image and another different person's image. As shown in Fig. 1, the generated fake images should have the same identity as the input, and far away from the "neutral B image". To solve this problem, we add two more fake pairs to enhance the ability of discriminator to preserve the person identity: the first pair consists of the generated facial image $G(I_N, c)$ and another person's neutral face image I_{BN}; the second one consists of the input neutral image I_N and another person's neutral face image I_{BN}. The discriminator D also performs multi-class classification to classify the facial expression category. The identity preserving loss is defined as

$$L_{ip} = 2 * L_{sce}\left(D\left(I_N, I_T\right), 1\right) + L_{sce}\left(D\left(I_{BN}, G\left(I_N, c\right)\right), 0\right) + L_{sce}\left(D\left(I_{BN}, I_T\right), 0\right) \quad (2)$$

where L_{sce} denotes the sigmoid cross-entropy loss function. I_T is the real target facial expression image. The value 1 means the pair discriminated is a positive sample for D, and the value 0 means the pair discriminated is a negative sample for D. We multiply 2 to the loss of real pair to balance the number of positive and negative samples.

Category Loss. We use six basic expression categories in this paper. The network D also measures whether an image belongs to a specific fine-grained expression category. Here we use a standard method for classification. The network D takes I_N as input and outputs a six-dimensional vector, which is then turned into class probabilities through a softmax function. The category loss function is defined as

$$L_c = L_{sce}\left(D\left(I_N, I_T\right), c\right) + L_{sce}\left(D\left(I_N, G\left(I_N, c\right)\right), c\right). \quad (3)$$

Adversarial Loss. In order to generate realistic facial expression images indistinguishable from real facial images, we utilize the adversarial loss

$$L_{adv}^D = L_{sce}\left(D\left(I_N, I_T\right), 1\right) + L_{sce}\left(D\left(I_N, G\left(I_N, c\right)\right), 0\right) \quad (4)$$

$$L_{adv}^G = L_{sce}\left(D\left(I_N, G\left(I_N, c\right)\right), 1\right) \quad (5)$$

where L_{adv}^D and L_{adv}^G denote the loss function for the discriminator D and the generator G, respectively. D tries to distinguish generated images $G(I_N, c)$ from real image while G produces the facial image $G(I_N, c)$ conditioned on both input image I_N and target label c.

Reconstruction Loss. Though G is trained to generate realistic images with correct target label through minimizing the adversarial and category losses, the generated images cannot be guaranteed to have fine texture details compared to target images. To alleviate this problem, we apply a reconstruction loss to the generator, which is defined as

$$L_{L1} = \left\| G\left(I_N, c\right) - I_T \right\|_1 \quad (6)$$

where $G(I_N, c)$ is the generated images based on input I_N and corresponding target expression label c. We adopt the $L1$ distance rather than $L2$ in our reconstruction loss for less blurring effects [10].

Full Objective Function. The final objective loss function is defined as

$$G^* = L_{adv}^D + L_{adv}^G + \lambda_1 L_{L1} + \lambda_2 L_c + \lambda_3 L_{ip} \quad (7)$$

where λ_1, λ_2, and λ_3 are hyper-parameters that balance the importance of reconstruction, category, and identity preserving losses, respectively.

Network architecture. For the model architecture, we mainly use the U-Net based architecture [27], which can help ensure the low-level information to be shared between the input and output. The only difference from traditional U-Net is that we combine the category embedding after encoding. The conditioned U-Net model connects each layer i and layer n-i, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer n-i. The classifier function and check function of D share parameters except the output layer. This parameter sharing scheme enables the networks to leverage their common information such as features at low-level layers that are close to the data layer, hence helping train model effectively. In addition, it also minimizes the number of parameters and adds minimal complexity to the standard GAN.

4 EXPERIMENTS

4.1 Dataset

Two datasets are mainly used in our experiment, i.e. Oulu-CASIA and KDEF.

Oulu-CASIA. The Oulu-CASIA NIR&VIS facial expression database (Oulu-CASIA) [39] contains 80 subjects with six expressions, i.e., angry, disgust, fear, happy, sad and surprise. The whole image sequences are obtained under three different illumination conditions: dark, strong and weak. In the experiment, we only use the videos with strong illumination captured by a VIS camera, in which the first frame is always neutral and the last frame has the peak expression. For each expression squence of each subject, only the first and the last three frames are used, and we combine the first neutral images with the last three frames respectively to generate three image pairs. The total number of image pairs is 1440. Training and test sets are divided based on identity, with 1296 for training and 144 for testing. We align the faces by Multi-task CNN algorithm [37], and crop them.

KDEF. The Karolinska Directed Emotional Faces (KDEF) [19] dataset includes totally 4900 pictures of human facial expressions. It contains 70 individuals, each displaying 7 different emotional expressions including the neutral expression. Each expression is photographed (twice) from 5 different angles. Each subject appears in two series. In the experiment, we only choose the frontal facial images. So for one subject in each series, we only have seven facial images. Then we combine the neutral expression with the universal expressions. The total number of image pairs is 840. Training and test sets are divided based on identity, with 756 for training and 84 for testing.

4.2 Implementation Details

The UC-GAN mainly contains two parts: a generator G with an encoder and a decoder, and a discriminator D. The encoder contains a traditional convolution layer and seven encoder layers where the numbers of channels are 64, 128, 256, 512, 512, 512, 512, 512, respectively. The traditional convolution layer is 5×5 stride 2 convolution. The encoder layer in a block is composed of Leaky ReLU [20], 5×5 stride 2 convolution and batch normalization [9]. The decoder includes eight decoder layers. The decoder layer in a block is composed of 5×5 deconvolutional operation with stride 2, batch normalization, skip connection and dropout. For the first three decoder layers, the dropout probability is 0.5. For each decoder layers except the last one, the decoder layers have skip connections with the corresponding encoder layers. With skip connections, the decoder's numbers of channels are 1024, 1024, 1024, 1024, 512, 256, 128, and 3. D has two functions: one is for checking if the images are true or false, and the other is for expression classification. Both functions share the first four encoder layers blocks with 64, 128, 256, and 512 channels. Then it is branched to two heads, one for checking and one for classification.

We conduct a ten-fold person-independent cross validation for each dataset, so the number of training set and test set is 9:1. During training, the generator takes a neutral face image and a corresponding label information as input, and the output is a generated expression face image conditioned on the expression label. The input and output face images are 256×256 RGB images $I \in \mathbb{R}^{256\times256\times3}$. The discriminator takes image pairs $(I_N : I_T)$ as positive sample pairs, and $(I_N : G(I_N, c))$, $(I_{BN} : G(I_N, c))$, $(I_{BN} : I_T)$ as negative sample pairs.

On both datasets, we use the Adaptive Moment Estimation (ADAM) [12] optimizer with $\beta_1 = 0.5$. The initial learning rate is set to 0.0002 for both generator and discriminator. The images are scaled to 256×256 and the pixel values are normalized into the range $[-1, 1]$. We empirically set the batch size to 1, and the weights are set as $\lambda_1 = 100$, $\lambda_2 = 1$, $\lambda_3 = 1$, respectively. During testing, we randomly sample a neutral image and set an expression label. The model is implemented using Tensorflow [1].

4.3 Facial Expression Synthesis

We train our model on two datasets mentioned before for facial expression synthesis, respectively. The expression of the input face is fixed as neutral, and the target expression can be any from the six expressions (i.e. happy, surprise, fear, disgust, angry, and sad).

4.3.1 *Qualitative Evaluation.*

To show that our facial expression synthesis method can generate natural facial expressions while preserving the identity, we conduct the qualitative experiment.

Facial expression generation results. The first column shows the input faces, and the rest columns show our results for different expressions. From Fig. 3, we can see that our method can produce natural and reasonable facial expressions, even with variations in race, gender, etc. Our results also preserve the identity information of the input face sufficiently.

Comparison with other methods. We also compare our method with one state-of-the-art method for face expression synthesis, i.e. ExprGAN [5]. The Oulu-CASIA dataset is used for the comparison experiment. The comparison results are shown in Fig. 2. The first column shows input neutral faces for different methods. The other six columns show the generated results for the six expressions. The ground-truth images are also included in the Fig. 2 (in the last row). As shown in Fig. 2, compared with ExprGAN, the proposed UC-GAN has better realistic expression generation results, while the personal identities of the input faces are also well maintained. ExprGAN fails to preserve the personal identity effectively in the synthesis results, and the output faces are not clear.

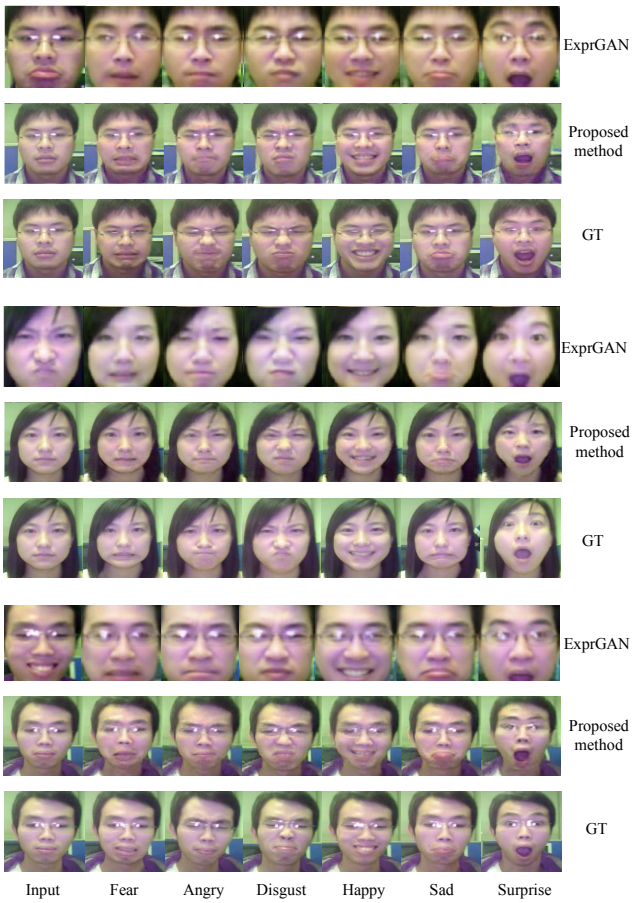
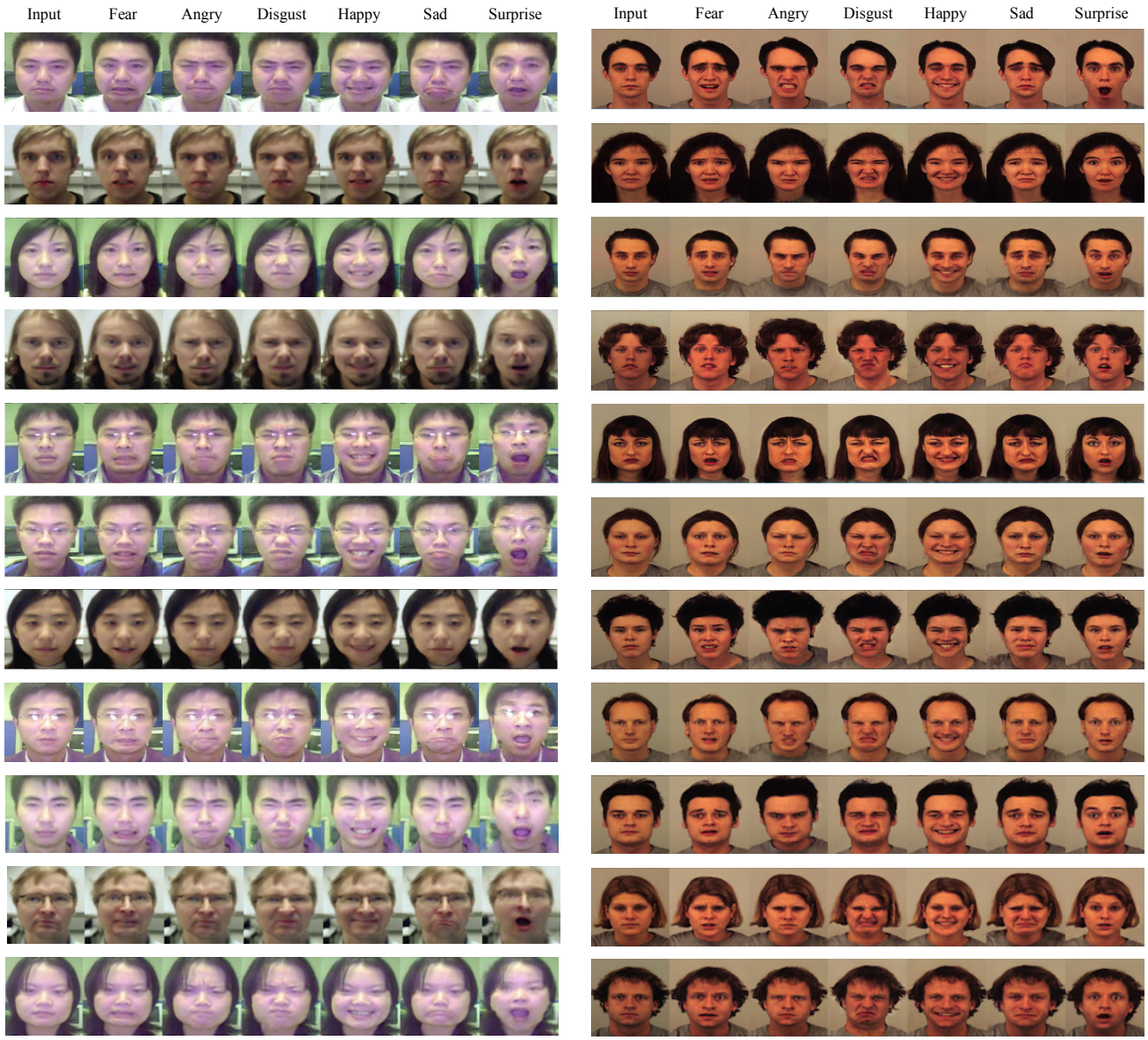


Figure 2: Comparison results of different facial expression synthesis methods on Oulu-CASIA. The ground-truth images (GT) are shown along with the synthesis results.

4.3.2 *User Study.*

We also evaluate our method by user study. Facial expression synthesis aims at generating natural and reasonable expressions while remaining his/her identity intact. So in the user study, we ask the volunteers to check whether the generated images have realistic and reasonable expressions, and whether they can still be recognized as the same person. There are 20 participants in this experiment. Each is shown with several pairs of images and is asked to give a score ranging from 1 to 10 to evaluate whether the generated



(a) Oulu-CASIA (b) KDEF

Figure 3: Facial expression synthesis results of our method on the Oulu-CASIA and KDEF datasets.

images are reasonable or not and decide whether they are images from the same person. The average scores of user study are 76.82% and 82.75% for the Oulu-CASIA and KDEF datasets, respectively, which demonstrate that the proposed UC-GAN generates natural and reasonable expressions, even though the number of subjects in the training datasets is small. And 79.64% of the generated faces are judged to be the same person in the input image, which shows that our method can preserve the person identity effectively.

One thing we want to discuss is that even though our method can synthesize natural faces with reasonable expressions, the synthesized expressions may not be personalized. For instance, the synthesized faces in Fig. 4 do not look exactly the same as the ground-truth images (the expression intensities and patterns are different), even though they are perceived to be natural and reasonable. This is also a general problem for most learning based approaches, which tend to learn the most-common expression patterns from the training data and apply them to the test faces. But personalized results are

highly demanded in practice. So personalized facial expression synthesis, which has more practical applications, can be one of our future research directions.

Figure 4: Example facial expression synthesis results on Oulu-CASIA and their corresponding ground-truth images (GT).

4.4 Contribution of Identity Preserving Loss

In this subsection, we demonstrate the discriminator's ability to make generated faces identity-preserving and natural via an ablation study. In detail, we compare the performance of our method with and without the identity preserving loss. Fig. 5 shows the qualitative evaluation results, in which the faces generated using our method with the identity preserving loss are more natural and identity-preserving, compared to those generated using the method without the loss.

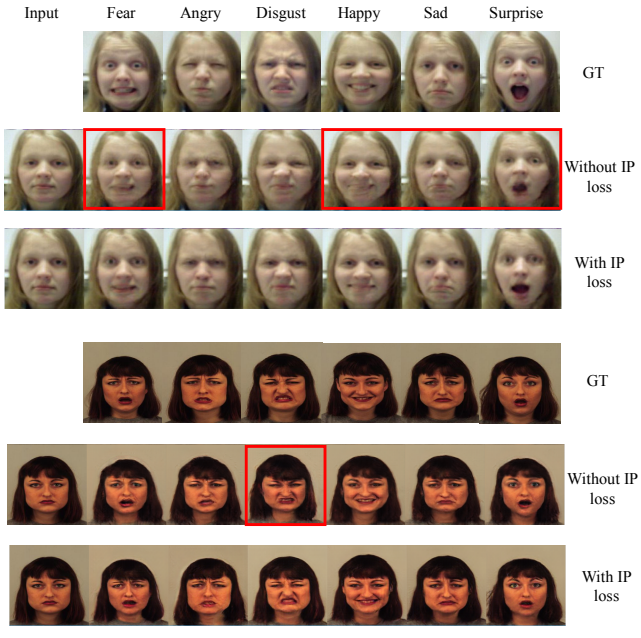


Figure 5: Comparison results of our method with or without identity preserving loss on the Oulu-CASIA and KDEF datasets. The top three rows are the comparison results from Oulu-CASIA, and the rest are from KDEF. The red boxes mark some synthesis results which are blurred or lose the identity information.

Table 1: Face verification results on Oulu-CASIA and KDEF

Dataset name	Dataset accuracy	Without IP loss[1]	With IP loss[1]
Oulu-CASIA	96.18%	92.26%	95.38%
KDEF	96.07%	93.88%	95.98%

[1] IP is the abbreviation of identity preserving.

In order to evaluate the performance of the proposed identity preserving loss quantitatively, we adopt the automatic face verification method by Wen [31] to check if the identity information of the input face is well preserved during expression synthesis. We use their model, which is trained on the CASIA-webface dataset [36], to extract facial features for face verification. A traditional shallow neural network is further used for verification, which includes two convolution layers and the dimension of feature maps for each layer is 128 and 2, respectively. Using this face verification algorithm and feature extracting method, the face verification accuracy on LFW dataset [8] is 98.50%. Inspired by [34], we perform comparisons between the input and the generated facial images for the six expressions, and the statistical analysis among the generated facial images is conducted. So in Oulu-CASIA, we have 144×21 pairs in total. As shown in Table 1, the verification rate of our method with identity preserving loss on Oulu-CASIA rises about 3.12% compared with the method without the loss. The dataset accuracy on the Oulu-CASIA dataset, which measures the accuracy of the verification method using the ground-truth faces with expressions, is 96.18% . For KDEF, there are 98×21 pairs of verifications, and the verification rate rises from 93.88% to 95.98% after adding the identity preserving loss. The dataset accuracy tested on this algorithm is 96.07%. These results quantitatively demonstrate the ability of identity preservation of the proposed method.

5 CONCLUSION

In this paper we propose a U-Net Conditioned Generative Adversarial Networks (UC-GAN) for facial expression synthesis. U-Net is adopted for its ability in keeping the low-level information of the input in the output face image. We also propose a person identity preserving loss to minimize the distance between the same person and maximize that between different ones. This loss constraint further enhances the distinction ability of the discriminator, which is good for producing identity-preserving faces with vivid expression details. Extensive experiments are conducted on the Oulu-CASIA and KDEF datasets. The proposed method outperforms the state-of-the-art method in qualitative evaluations. We also analyze the proposed method comprehensively by ablation studies using qualitative and quantitative evaluations. Both qualitative and quantitative experimental results demonstrate that our method, especially with the identity preserving loss added, can generate satisfactory face images that are natural-looking, and identity-preserving.

ACKNOWLEDGMENTS

This work is funded by the National Natural Science Foundation of China (No. 61673033 and No. 61421003).

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (Mar 2016).

[2] Thaddeus Beier and Shawn Neely. 1992. Feature-based Image Metamorphosis. *SIGGRAPH Comput. Graph.* 26, 2 (Jul 1992), 35–42.

[3] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. 2015. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583v4* (Jun 2015).

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2017. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *arXiv preprint arXiv:1711.09020* (Nov 2017).

[5] Hui Ding, Kumar Sricharan, and Rama Chellappa. 2018. ExprGAN: Facial Expression Editing with Controllable Expression Intensity. In *2018 Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

[6] Jon Gauthier. 2014. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester* 2014, 5 (2014), 2.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*. 2672–2680.

[8] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.

[9] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37. 448–456.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-To-Image Translation With Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976.

[11] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. 2017. Generative attribute controller with conditional filtered generative adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7006–7015.

[12] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (Dec 2014).

[13] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (Dec 2013).

[14] Anders Boesen Lindbo Larsen, SÃ¿ren Kaae SÃ¿nderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48. 1558–1566.

[15] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 105–114.

[16] Mu Li, Wangmeng Zuo, and David Zhang. 2016. Deep Identity-aware Transfer of Facial Attributes. *arXiv preprint arXiv:1610.05586* (Oct 2016).

[17] James Jenn-Jier Lien, Takeo Kanade, Jeffrey F Cohn, and Ching-Chung Li. 2000. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems* 31, 3 (May 2000), 131–146.

[18] Zicheng Liu, Ying Shan, and Zhengyou Zhang. 2001. Expressive Expression Mapping with Ratio Images. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 271–276.

[19] Daniel Lundqvist, Anders Flykt, and Arne Öhman. 1998. The Karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet* (1998).

[20] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, Vol. 1. 3.

[21] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (Nov 2014).

[22] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional Image Synthesis with Auxiliary Classifier GANs. In *2017 Thirty fourth International Conference on Machine Learning (ICML)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. 2642–2651.

[23] Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. 2017. Realistic Dynamic Facial Textures from a Single Image using GANs. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 5439–5448.

[24] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. 2005. Synthesizing Realistic Facial Expressions from Photographs. In *ACM SIGGRAPH 2005 Courses*. Article 9.

[25] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (Nov 2015).

[26] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. 2014. Learning to Disentangle Factors of Variation with Manifold Interaction. In *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32. 1431–1439.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *2015 eighteenth International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 234–241.

[28] Steven M. Seitz and Charles R. Dyer. 1996. View Morphing. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. 21–30.

[29] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. 2008. Generating facial expressions with deep belief nets. In *Affective Computing*.

[30] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.

[31] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). 499–515.

[32] Marco Pedersoli Xu Jia, Amir Ghodrati and Tinne Tuytelaars. 2016. Towards Automatic Image Editing: Learning to See another You. In *Proceedings of the British Machine Vision Conference (BMVC)*. Article 101, 11 pages.

[33] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision – ECCV 2016*. Amsterdam, The Netherlands, 776–791.

[34] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. 2017. Learning Face Age Progression: A Pyramid Architecture of GANs. *arXiv preprint arXiv:1711.10352* (Nov 2017).

[35] Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. 2016. Semantic Facial Expression Editing using Autoencoded Flow. *arXiv preprint arXiv:1611.09961* (Nov 2016).

[36] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (Nov 2014).

[37] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (Oct 2016), 1499–1503.

[38] Qingshan Zhang, Zicheng Liu, Gaining Quo, Demetri Terzopoulos, and Heung-Yeung Shum. 2006. Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Computer Graphics* 12, 1 (Jan-Feb 2006), 48–60.

[39] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikäinen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (Aug 2011), 607–619.

[40] Yuqian Zhou and Bertram Emil Shi. 2017. Photorealistic Facial Expression Synthesis by the Conditional Difference Adversarial Autoencoder. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 370–376.