The Visual Computer (2022) 38:2539–2554
https://doi.org/10.1007/s00371-021-02129-y

ORIGINAL ARTICLE

Multi-view face generation via unpaired images

Shuai Wang[1] · Yanni Zou[1] · Weidong Min[2,3] · Jiansheng Wu[1] · Xin Xiong[1]

Accepted: 28 March 2021 / Published online: 13 April 2021

Abstract

Multi-view face generation from a single image is an essential and challenging problem. Most of the existing methods need to use paired images when training models. However, collecting and labeling large-scale paired face images could lead to high labor and time cost. In order to address this problem, multi-view face generation via unpaired images is proposed in this paper. To avoid using paired data, the encoder and discriminator are trained, so that the high-level abstract features of the identity and view of the input image are learned by the encoder, and then, these low-dimensional data are input into the generator, so that the realistic face image can be reconstructed by the training generator and discriminator. During testing, multiple one-hot vectors representing the view are imposed to the identity representation and the generator is employed to map them to high-dimensional data, respectively, which can generate multi-view images while preserving the identity features. Furthermore, to reduce the number of used labels, semi-supervised learning is used in the model. The experimental results show that our method can produce photo-realistic multi-view face images with a small number of view labels, and makes a useful exploration for the synthesis of face images via unpaired data and very few labels.

Keywords Multi-view face generation · Generative adversarial net · Adversarial autoencoder · Semi-supervised learning

1 Introduction

Multi-view face image synthesis, which generates images from different views for a given face image, is an interesting but challenging question. It has been widely applied in various domains such as unconstrained face recognition and computer graphics. Two conditions need to be satisfied for a single view generating multi-view face images. The first is that the generated images should be the same identity as the input image. The second is that the same view of different identities should be consistent.

In order to satisfy these conditions, a number of approaches are presented by researchers. These methods are roughly classified into two categories: 3D face model-based methods [1–7] and deep learning methods [8,9]. The method based on 3D face model synthesizes face images with new angles by establishing 3D face model as a reference and fitting model. For example, Blanz et al. [1] proposed a representative method which first employed a face database to construct an average face deformation model; then, the model is matched with the given new face image; finally, the images of new angles are fitted by continuously modifying the parameters of model. Hang et al. [4] proposed a novel unsupervised framework, which rotated faces in the 3D space back and forth, re-rendered them to the 2D plane in a strong self-supervision manner, and generated the final image by Pix2Pix [7]. RigNet [5] provided a face rig-like control over a pretrained and fixed StyleGAN via a 3DMM. The network is trained in a self-supervised manner, without the need for manual annotations. Although these methods based on 3D model are effective, their synthesis results are not realistic. The method based on deep learning makes the model learn the abstract representations of the identity and view of the input image by training neural network, and then get the multi-view images by using feature fusion. For example, Ghodrati et al. [8] input a pair of face images with the same identity but different views and view labels into the network. Then, images of different views are obtained by image coding, attribute vector coding, feature map fusion and image decoding. Finally, a clearer picture is obtained by image generation refinement. Zhu et al. [9] pro-

✉ Weidong Min
minweidong@ncu.edu.cn

[1] School of Information Engineering, Nanchang University, Nanchang 330031, China

[2] School of Software, Nanchang University, Nanchang 330047, China

[3] Jiangxi Key Laboratory of Smart City, Nanchang 330047, China

posed multi-view perceptron to disentangle the identity and view representations of the input images. The identity features and the view representation are learned by determining the deterministic hidden neurons and random hidden neurons from perceptron. Images from different views are generated by fusing different view representations and identity characteristics. Fu et al. [10] first predict the boundary image of the target face in a semi-supervised way, modeling pose and expression jointly, and then utilize the predicted boundary to perform refined face synthesis. Their method has achieved good effect, but it is necessary to input a pair of images during training. Furthermore, the fine details are often missed from faces which are synthesized from methods based on convolutional neural network and deep neural network mentioned above.

In generative adversarial network (GAN) [11], clear and authentic samples can be produced by simulating data distribution according to decision theory and game theory, which also make it an impressive achievement in multi-view generation. These GAN-based methods [12–21] usually resort to images from different views of the same identity (x_i, x_j) during training. The identity and view representations are first disentangled in the latent space; then, identity representation is input into the generator under the constraint of another view label $\boldsymbol{v_j}$ to generate image $\widetilde{x_j}$ of the same identify but different view, and then, the discriminator is trained to distinguish $\widetilde{x_j}$ from real image x_j. Not only view labels but identity labels are used in these models during training. In addition, TP-GAN [14] and LB-GAN [15] also need to label the eyes and mouth of the face image to get specific local texture information. Hu et al. [16] propose CAPG-GAN to generate both neutral and profile head pose face images. The head pose information is encoded by facial landmark heatmaps. A couple-agent discriminator is introduced to reinforce on the realism of synthetic arbitrary view faces. Besides the generator and conditional adversarial loss, CAPG-GAN further employs identity preserving loss and total variation regularization to preserve identity information and refine local textures, respectively. Sanchez et al. [17] propose a “recurrent cycle consistency loss” which for different sequences of target attributes minimizes the distance between the output images, independent of any intermediate step. However, this approach requires the use of paired images and facial landmark annotations to train the model. Studies [18–21] are able to learn meaningful latent spaces, explaining generative factors of variation in the data. However, to the best of our knowledge, there has been no work explicitly disentangling the latent space for object geometry of GANs. These methods based on GAN have high demand on the collection and labeling of data sets and takes a lot of manpower and time. For example, in reference [22], in order to collect the Multi-PIE, 337 subjects were recorded using a hardware synchronization network consisting of 15 high-quality cameras and 18 flashes, and then, all images were labeled with their identity, illumination, pose and expression. CycleGAN [23] aims to solve the problem of the need to use paired images in the training of traditional image style migration networks. However, on the fine-grained task of human face, the details of facial features will bring great challenges to the generation task. Without optimizing pixel-level loss, it is easy to cause face image distortion.

To reduce labor and time costs, paired images using is avoided and the dependence on labels are reduced in this paper. The encoder and discriminator were trained so that the encoder can learn the identity and view representations of the input image in the manifold space, and then, these low-dimensional codes were input into the generator to get high-dimensional data. At last, by training the generator and discriminator, a realistic face image can be reconstructed from the generator. In order to produce realistic faces with different views meanwhile preserve the identity features, two adversarial networks are applied to the encoder and generator, respectively. To further reduce the dependency on the view labels, another discriminator is imposed on the encoder, which force its output to follow the categorical distribution. Overview of our proposed method is shown in Fig. 1. During training, encoder E maps the input image to identity representation $\boldsymbol{z}$ and view representation $\tilde{\boldsymbol{v}}$, respectively. Discriminator D_z forces $\boldsymbol{z}$ to follow the uniform distribution. Discriminator D_v forces $\tilde{\boldsymbol{v}}$ to follow the categorical distribution, and make the view representation more accurate by minimizing the cross-entropy of the fake view label and the real view label. The generator G reconstructs the image using $\tilde{\boldsymbol{v}}$ and $\boldsymbol{z}$, the reconstructed result and the real image are, respectively, connected with $\tilde{\boldsymbol{v}}$ and then input into the discriminator D_{img} for similarity judgment. During the test, multiple one-hot vectors were imposed on the identity representation in the latent space, and the vectors representing the views were connected with the identity representation, respectively; then, they were input into generator to synthesize multi-view images while preserving the identity features.

In summary, we have following contributions:

(1) Our method does not resort to paired images during training and does not depend on the identity information of face images.
(2) Semi-supervised learning is used to further reduce the number of used labels. Only a small number of view labels are required for training.
(3) Experimental results show that our network effectively disentangles the identity and view representations, and generates multiple-view face images while preserving the identity features.

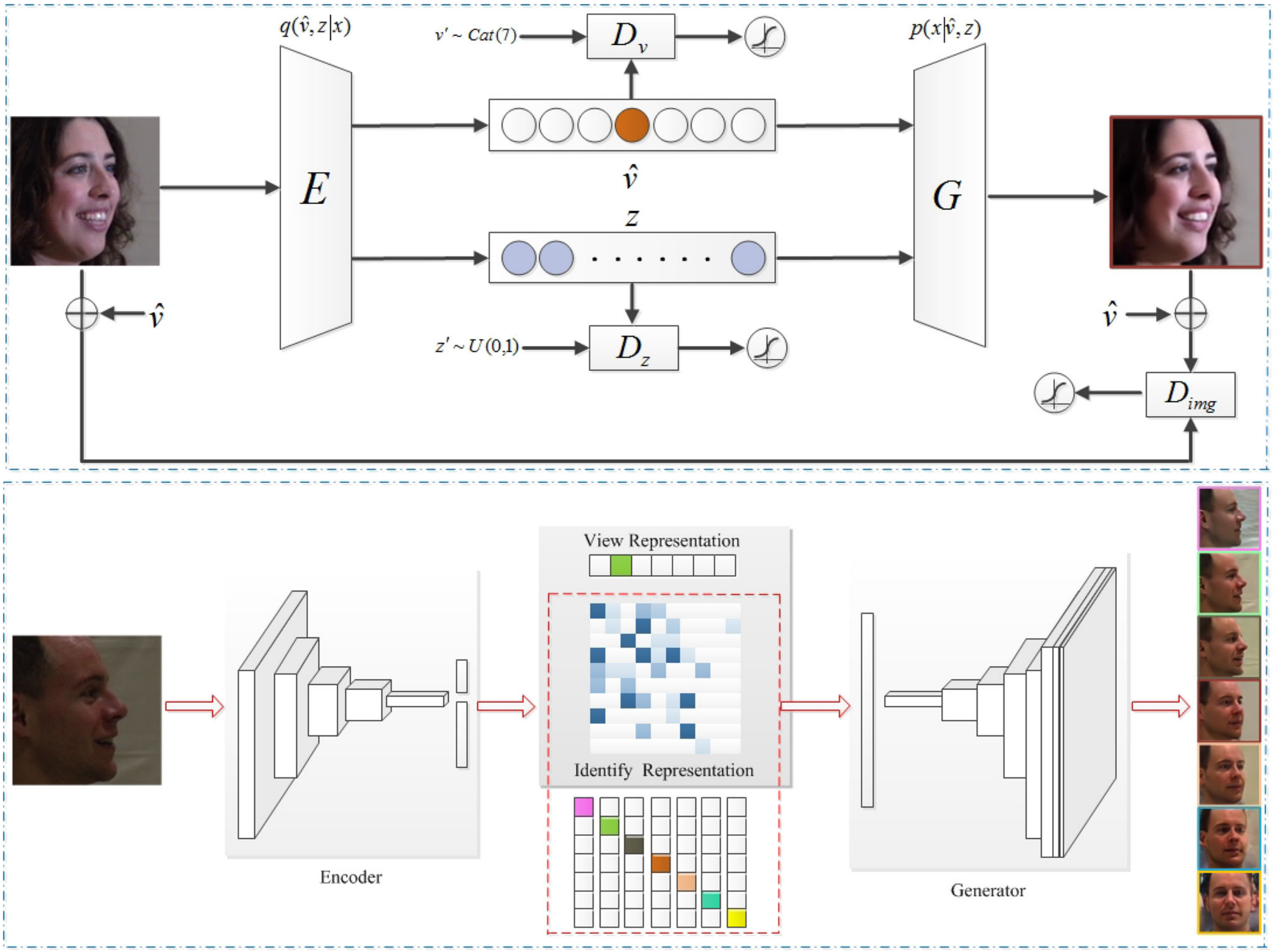


Fig. 1 Overview of our method

2 Related work

2.1 Generative adversarial networks

Generative adversarial network has achieved great success in image generation and has received extensive attention. It samples from the complex probability distribution by training the discriminator and the generator in turn to compete against each other. However, GAN generates an image from random noise and cannot control the output image. In recent years, GAN's architecture has been continuously improved. CGAN [24] attempted to add additional conditional information to the generator and discriminator to guide the training of the two models of GAN. CC-GAN [25] used semi-supervised learning to repair missing parts of the image. Each image generated by the generator of AC-GAN [26] has a category label, and the discriminator also gives two probability distributions for the source and category labels. InfoGAN [27] obtained decomposable feature representations through unsupervised learning. GAN can generate clear images, but only in fairly small resolutions and with somewhat limited variation. The samples generated by variational autoencoder (VAE) [28] are very close to the original image, but usually blurred. CVAE-GAN [29] combined the advantages of both to produce realistic and diverse samples. Backpropagation is performed in VAE using KL divergence, so the exact functional form of the prior distribution is required. Adversarial autoencoder (AAE) [30] only needs to sample from the prior distribution, and then make the prior distribution fit the real distribution through adversarial training. To avoid using KL divergence, AAE is used in our method.

2.2 Face frontalization

Face frontalization is a technique for synthesizing frontal images using face images from other angles, which is very helpful for improving face recognition rate, and has been extensively studied. The existing face frontalization methods can be classified into three categories: 3D-based methods [31–34], statistical methods [35,36] and deep learning methods [37–40]. Zhu et al. [32] established corresponding relation between 2D face and 3D model at different angles according to the key point matching rule, and then, the 3D model is established to eliminate the effects of angles and generate frontal images. Sagonas et al. [36] consider that frontal image has the smallest rank in all the different poses, they obtained frontal images by minimizing nuclear norm and the matrix L1 norm accounting. Yin et al. [39] proposed a method which combined 3D morphable model (3DMM) [41] and GAN. In their works, 3D model is first used to get general information, and then, 3DMM coefficients and original image are input into GAN to generate detailed frontal face images. Compared with methods mentioned above, our method uses two discriminators to ensure the accuracy of identity and view.

2.3 Representation learning

Effective representation of learning samples can simplify the difficulty of data processing. Autoencoder obtained the effective representation of the learning sample by minimizing the reconstruction error, which compresses the input data into a latent representation and then reconstructs the output through the characterization, and usually used for data denoising and visual dimension reduction. Adversarial autoencoder [30] employed an adversarial strategy based on autoencoder, and training a discriminator to distinguish whether the sample is from the latent code of autoencoder or the user-defined prior distribution. Eduardo et al. [42] showed that models with adversarial network can improve the quality of representation learning. Jirui et al. [43] proposed a multi-view predictive latent space learning model, which learns a latent representation by maximizing the correlation between the feature space where all the feature vectors exist and latent space where latent vectors exist. Lample et al. [44] presented a new approach to generate variations of images by changing attribute values, which generates realistic images of high resolution without needing to apply a GAN to the decoder output. Tang et al. [45] proposed a method for expressive style transfer. Liu et al. [46] applied GAN to video-to-video translation. Huang et al. [47] proposed a method of learning the pooling scheme to learn high-level features of face images. Zhu et al. [9] proposed multi-view perceptron (MVP) to disentangle the identity and view representations. The deterministic hidden neurons and the random hidden neurons are used to learn the identity features and capture the view representation, respectively. DR-GAN [12] used an adversarial strategy based on MVP, which made the model have better representation ability and obtained high-quality face synthesis images. Tian et al. [13] proposed CR-GAN, which used a two-pathway learning scheme to learn complete representations. Our approach is most relevant to references [9,12], but different. The method proposed in the literature [12] cannot learn the conditional representation of the input images. In the literature [9], the identity and view representations are solved using different neurons, while we used adversarial training to make the learned data representation more accurate, this is because in the process of adversarial learning, the training sample is no longer the original sample, but the original sample and counter sample. This is equivalent to adding the generated confrontation samples into the training set as new training samples and treating them equally. Then, with more and more training of the model, the accuracy of the learned data representation will increase, and the robustness of the model to the adversarial examples will also increase.

3 Proposed method

In this section, we first introduce our network structure, which consists of an encoder, a generator and three discriminators. In Sect. 3.2, objective function we used is described. Section 3.3 introduces our semi-supervised learning approach and shows the training details.

3.1 Model structure

3.1.1 Encoder

In the process of generating multi-view face images, it is difficult to directly manipulate the face images in high-dimensional space, and such high-dimensional data are not required to reflect the identity and view information of the face; therefore, it is necessary to map the high-dimensional data of the face image to the low-dimensional latent vector in the latent space.

The encoder is used to learn the features of face images. It uses a convolutional neural network to encode a face image, and outputs latent variables that represent facial features. In order to obtain a useful feature representation, the model proposed in this paper uses an encoder based on a convolutional neural network. The specific structure is shown in Table 1, which is mainly composed of 5 convolutional layers and 2 fully connected layers. In order to reduce the amount of calculation and increase the training speed, the image size is scaled to 128 × 128, and the identity representation and view representation of the facial image identity features are obtained through the convolutional layers and the fully connected layers.

In the process of generating multi-view face images, how to change the view while retaining identity features is very critical. According to the theory of manifold learning, the data we observe are actually mapped from a low-dimensional manifold to a high-dimensional space. Assuming that the input face image is located on a low-dimensional manifold, moving the sample along the manifold can realize the

Table 1 Structure of the encoder

Layer	Filter/stride	Output size
conv1	5 × 5/2	64 × 64 × 64
conv2	5 × 5/2	32 × 32 × 128
conv3	5 × 5/2	16 × 16 × 256
conv4	5 × 5/2	8 × 8 × 512
conv5	5 × 5/2	4 × 4 × 1024
fc1	-	50
fc2	-	13/9

angle of view conversion. The model proposed in this paper converts the face image into two latent spaces through two fully connected layers of the encoder, so as to obtain a low-dimensional representation of the identity and view of the face image, which can reduce the accumulated error of the reconstructed image; this is because the model proposed in this paper transforms the feature into a hidden space. Specifically, this is achieved by having a full connection layer with bias, thereby reducing errors accumulated by rebuilding the view. The weight of the full connection layer can be expressed as W=[w_1,w_2,...w_v], and v represents the view label, and each view corresponds to a weight. Therefore, the process of view conversion from a to b can be expressed as e^{ab}=[$e_1{}^{ab}$, $e_2{}^{ab}$, ...$e_v{}^{ab}$]^T , and then, the process of angle conversion can be expressed as z^b=z^a+Wz^{ab}. In this paper, considering the image reconstruction effect and computational cost, the sampling parameters were selected for the subsequent experiments, and finally, the dimension of facial identity representation was set to 50. In addition, in order to keep the dimension consistent with the view label vector, the dimension represented by the view is set to 13 or 9 (13 or 9 multi-view images are output). The pixels of the input image are normalized to (-1, 1), the activation function of convolutional layers is ReLU, and the activation functions of the two fully connected layers are tanh and softmax, respectively.

3.1.2 Generator

The input of the generator is the latent variable representing the face features and the one-hot encoding representing the view label. Through the deconvolution process, the multi-view face images are output, and the identity of the face is maintained while converting the perspective of the face. Specifically, the latent variables and one-hot encoding are connected to input generator, so that the generator can generate images according to certain rules. The generator used in this paper consists of 1 fully connected layer and 7 deconvolution layers, whose structure is shown in Table 2.

Table 2 Structure of the generator

Layer	Filter/stride	Output size
fc	–	16384
deconv1	5 × 5/2	8 × 8 × 512
deconv2	5 × 5/2	16 × 16 × 256
deconv3	5 × 5/2	32 × 32 × 128
deconv4	5 × 5/2	64 × 64 × 64
deconv5	5 × 5/2	128 × 128 × 32
deconv6	5 × 5/1	128 × 128 × 16
deconv7	5 × 5/1	128 × 128 × 3

Table 3 Structure of D_z and D_v

Layer	Filter/stride	Output size
fc1	–	64
fc2	–	32
fc3	–	16
fc4	–	1

Table 4 Structure of D_{img}

Layer	Filter/stride	Output size
conv1	5 × 5/2	64 × 64 × 16
conv2	5 × 5/2	32 × 32 × 32
conv3	5 × 5/2	16 × 16 × 64
conv4	5 × 5/2	8 × 8 × 128
fc1	–	1024
fc2	–	1

3.1.3 Discriminator

The purpose of the discriminator D_z is to force the output of the encoder to follow a priori distribution. In order to achieve this, D_z is trained to discriminate the output z of the encoder and the sampling of the prior distribution, and the encoder is trained to generate the latent variable z that can deceive D_z. The prior distribution used here is uniform distributed.

The discriminator D_v also makes the output of the encoder obey the category distribution through the same confrontation process. Since unlabeled samples and labeled samples are independently sampled from the same data with the same distribution, the information about the data distribution contained in the unlabeled samples is beneficial to modeling. In order to make full use of unlabeled data and reduce the model's dependence on labels, this paper uses an additional discriminator D_v in the model to make it a semi-supervised model.

The discriminator D_{img} forces the face image generated by the generator to be more realistic, and when training, the view label is applied to D_{img}, which can ensure the consistency of the perspective of the image generated by the generator. Although the identity of the face image can be ensured by minimizing the distance between the input and output images, there is no guarantee that the images in the test set can also achieve good results on the model, because the images in the test set are not used during training. Therefore, the loss in pixels during reconstruction can only be generated by interpolation to the image closest to the image in the training set, and the discriminator D_{img} can avoid this situation, making the model perform well on the test set.

The structures of discriminators D_z, D_v and D_{img} are shown in Tables 3 and 4, respectively. The output layers of D_z and D_v use sigmoid as the activation function. The activation functions of the remaining layers are LeakyReLu, the activation function of the first fully connected layer (fc1) of D_{img} is LeakyReLu, and the activation function of the second fully connected layer (fc2) is sigmoid.

3.1.4 Compared with the structure of DR-GAN and CR-GAN

Figure 2 shows DR-GAN [12], CR-GAN [13] and the network structure mentioned in this article. The network structure of this article is different from DR-GAN in three points:

(1) DR-GAN uses view coding as a priori condition to guide the generation of face images, adds noise data to the hidden layer after the generator and encoder and expands the discriminator used to judge the true and false images to the classifier for image classification, pose estimation and face recognition tasks. Although face image classification, pose estimation and face recognition are three very related tasks, due to the different degrees of difficulty of these three tasks, simply using a discriminant network cannot well balance the three, increasing the complexity of the model. Therefore, in the network structure used in this paper, the view classification network and the encoder are in the same structure. In this way, the view classification network and the face generation network can be mutually restricted and promoted, and the synthesis rate can be improved while improving the view recognition rate;

(2) In this paper, the discriminator D_z is applied to the encoder to ensure a smooth transition in the latent space, and DR-GAN does not use this confrontation strategy;

(3) The DR-GAN discriminator inputs images of two different views that need the same identity. (One is a real image and the other is a reconstructed image.) Tags and the network in this article only need one kind of view label. Through training, the encoder can untie the identity representation and view representation, so that the generator can reconstruct a realistic image.

CR-GAN adds a path to DR-GAN to ensure that the network learns a complete representation. In addition to the above three points, the network structure of this paper is also different from CR-GAN: CR-GAN uses dual paths to ensure that the network can also generate good results in the test set, and this paper achieves the same purpose by training the discriminator D_{img}.

3.2 Objective function

In general, the generated images need to meet three requirements: (1) The input face and the output face should keep the identity characteristics unchanged. (2) The same view of

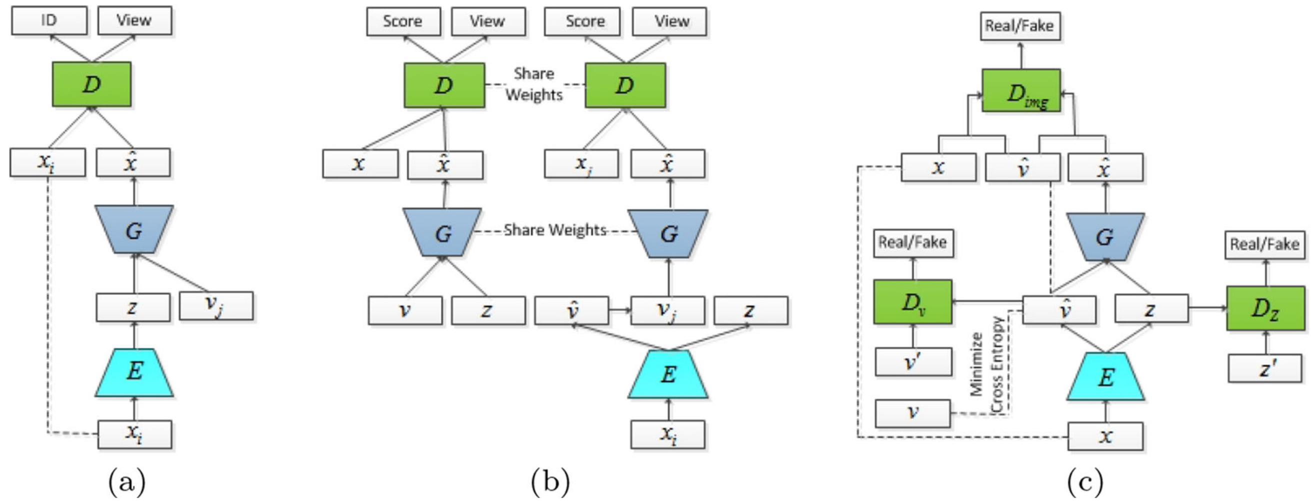


Fig. 2 Comparsion of DR-GAN(**a**), CR-GAN(**b**) and our method(**c**)

different identity faces should be consistent. (3) The output face should be realistic.

To obtain the face's identity representation, the input face image x is mapped to be the feature vector $\boldsymbol{z}$ by the encoder E. Then, the feature vector $\boldsymbol{z}$ and the sample obtained from the prior distribution are input into the discriminator D_z, the encoder E and D_z are trained by the min–max game, which force $\boldsymbol{z}$ to gradually close to the prior distribution:

$$\begin{aligned}\min_{E}\max_{D_z} V(E, D_z) &= \mathbb{E}_{z' \sim p_z(z)}[\log D_z(z')] \\ &+ \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_z(E(x)))]\end{aligned} \tag{1}$$

where $p_z(z)$ denotes the prior distribution, $p_{data}(x)$ denotes the data distribution of the real input image, $E(x) = z \in \mathbb{R}^n$, and n denotes the dimension of the face feature.

Similarly, to get a view representation of the face, encoder E map the face image x to be the view vector $\tilde{\boldsymbol{v}}$, we input the $\tilde{\boldsymbol{v}}$ and the sample obtained from the prior distribution into the discriminator D_v, and train E and D_v by the min–max game:

$$\begin{aligned}\min_{E}\max_{D_v} V(E, D_v) &= \mathbb{E}_{v' \sim p_v(v)}[\log D_v(v')] \\ &+ \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_v(E(x)))]\end{aligned} \tag{2}$$

where $p_v(v)$ denotes the prior distribution, $E(x) = \tilde{\boldsymbol{v}} \in \mathbb{R}^n$,$n$ denotes the number of views. Unlabeled data and labeled data are alternately input into the encoder until all the labeled data are used. The cross-entropy of the output $\tilde{\boldsymbol{v}}$ of the encoder E and the real label v is calculated by

$$H(v, \tilde{v}) = -[v \log \tilde{v} + (1 - v) log(1 - \tilde{v})] \tag{3}$$

We optimize E by minimizing cross-entropy to reduce classification error. Note that the very low classification error has slight effect on the generated results due to the large number of training samples.

To make the reconstructed image more realistic, we input the feature vector $\boldsymbol{z}$ and the corresponding view vector $\tilde{\boldsymbol{v}}$ into G to generate $\tilde{x}$; then, $(\tilde{x}, \tilde{\boldsymbol{v}})$ and $(x, \tilde{\boldsymbol{v}})$ are input into the discriminator D_{img}, and the role of $\tilde{\boldsymbol{v}}$ is the same as that of the label in supervised learning. G and D_{img} can be trained by

$$\begin{aligned}\min_{G}\max_{D_{img}} V(G, D_{img}) &= \mathbb{E}_{x, \tilde{v} \sim p_{data}(x, \tilde{v})}[\log D_{img}(x, \tilde{v})] \\ &+ \mathbb{E}_{x, \tilde{v} \sim p_{data}(x, \tilde{v})}[\log 1 - D_{img}(G((z, \tilde{v}), \tilde{v})))]\end{aligned} \tag{4}$$

In order to ensure that after the encoder and the generator, the output face image $\tilde{x}$ is located in the face manifold space; during training, the output face image $\tilde{x}$ and the input face image x share the identity characteristics of the face and view information; therefore, to ensure that x and $\tilde{x}$ are similar, you also need to calculate:

$$\min_{E,G} L(x, G(E(x), \tilde{v})) \tag{5}$$

where L represents the L2 norm; it is defined as follows:

$$\| x, \tilde{x} \| = \sqrt{\sum_{i=1}^{n} (x_i - \tilde{x}_i)^2} \tag{6}$$

3.3 Semi-supervised learning

With the development of data collection and storage technologies, it is convenient to collect large amounts of data, but only a small percentage of the data can be correctly labeled. To further reduce the number of labels used, semi-supervised learning is used in our model. Firstly, the output of the encoder $\tilde{\boldsymbol{v}}$ and the random sampling of the category distribution are input discriminator D_v. The encoder E can generate view labels, and the D_v can distinguish between real label and forecast label. D_v and E are updated by Formula (2) in this process. Secondly, in order to reduce classification error, E is updated by Formula (3) when labeled data is inputted. The training details is shown in Algorithm 1, we take advantage of unlabeled data by this strategy. E will be a good view estimator after several iterations. Unlike most of existing semi-supervised generative adversarial networks [25,26,48–50], our discriminators only judge true or false and do not output categories.

When training the model proposed in this paper, the labeled data are trained first with supervised learning method. Then the initial classifier obtained by training is used to predict the unlabeled data, and the data with high confidence and its annotation are added to the labeled data to retrain the classifier. Through the method introduced in Sect. 3.2, in the process of training, the discriminator D_z and D_v guide the latent variables $\boldsymbol{z}$ and the label variables $\tilde{\boldsymbol{v}}$ approaching the uniform distribution and category distribution, respectively. In other words, the two discriminators guide the output from the neural network before the full connection layer of the encoder to two different distributions. Therefore, the two distributions will weaken each other's regularization ability and then affect the convergence rate of the model. In order to reduce this effect, when the cross-entropy of the real label and the predicted label are reduced to a certain degree, the encoder parameters are no longer updated by Formula (3). Through this strategy, the unlabeled data are fully utilized. After several iterations, the encoder gradually acquires the ability of view estimation.

Algorithm 1 Semi-supervised training with unpaired data

```
Input: Training set of size $N$, where the number of labeled images is $M$, the number of iterations $T$, and batch size $m$;
Output: Trained network $E$, $D_z$, $D_v$, $G$ and $D_{img}$;
for $t = 1$ to $T$ do
    for $j = 1$ to $N/m$ do
        Take $m$ samples from the training set$\{x^i\}_{i=1}^m$;
        Sampling $m$ samples from the prior distribution of latent variables$\{z_p{}^i\}_{i=1}^m$;
        Sampling $m$ samples from the prior distribution of label variables$\{v_p{}^i\}_{i=1}^m$;
        if $j < M/m$ then
            Get the real sample label;
            Update parameters of the encoder $E$ by minimize the Formula (3);
        end if
        Input $\{x^i\}_{i=1}^m$ into the encoder to generate latent variable $\{\boldsymbol{z}^i\}_{i=1}^m$ and label variable $\{\tilde{\boldsymbol{v}}^i\}_{i=1}^m$
        Update parameters of $D_z$, $D_v$ and $E$ by the Formula (1) and Formula (2);
        Input $\{\boldsymbol{z}^i\}_{i=1}^m$ and $\{\tilde{\boldsymbol{v}}^i\}_{i=1}^m$ into generator $G$, and output generated sample $\{\tilde{x}^i\}_{i=1}^m$
        Update parameters of $D_{img}$ and $G$ by the Formula (4);
    end for
end for
```

4 Experiments

In this section, we first introduce the data sets we use and the experimental details. Then, to demonstrate the effectiveness of our method, DR-GAN [12] and CR-GAN [13] are compared with our method.

4.1 Experimental data and parameter settings

Multi-PIE [22] is a face data set established by Carnegie Mellon University in the USA, where "PIE" refers to the abbreviation of pose, illumination and expression. Developed on the basis of. Multi-PIE was collected in a restricted environment. The camera collected face images of 337 volunteers under 43 different lighting and 13 different shooting angles. Each face image contains at least 4 different expressions. Each volunteer's head image contains 13 yaw angles within ±90° (15° for every two attitudes).

300W-LP [51] is a 3DMM [41] label obtained by the 3DDFA team based on the existing AFW, IBUG, HEPEP and FLWP and other 2D face alignment data sets through 3DMM fitting. A large pose 3D face alignment data set obtained by mirroring.

In this paper, the face images of 249 volunteers (a total of 129480 images) of session1 in the Multi-PIE database are used for the experiment, of which 103584 images are used for training, and the remaining images are used for testing. Only the labels of 3000 images are used for training. In this paper, 122450 images in the 300W-LP data set are used for the experiment, of which 97960 images are used to train the model, and the remaining images are used as the test set. Only 2500 images are labeled during training. In order to compare with the experimental results of CR-GAN, only the images with yaw angle within ±60° in the 300W-LP data set are used, and they are dispersed into 9 intervals. The usage of the two data sets is shown in Table 5. It should be noted that DR-GAN and CR-GAN need to divide the training set test set according to identity. For example, in Multi-PIE, CR-GAN uses 200 identities for training, and the remaining identities are used for testing. The method proposed in this paper does not require identity tags during training, so there is no identity requirement for the training set images.

The preprocessing in this experiment includes face detection and image normalization. Since the research and application of face area detection has been very mature and is not the focus of this article, this paper uses reference [52] to detect and crop face images, and normalize the images to 128 × 128. The preprocessed images of Multi-PIE and 300W-LP are shown in Figs. 3 and 4, respectively.

Multi-PIE and 300W-LP preprocessed training set images were, respectively, input into the model for training. The hyperparameters are set as follows: batch size = 100, using Adam optimizer [53] as the optimizer algorithm, learning rate = 0.0002, momentum = [0.5,0.999]. The GPU used in the experiment was Nvidia Quadro P4000 with 8GB memory.

4.2 Comparison results

We have made qualitative and quantitative evaluations of our methods; three aspects are considered: the visual quality, the identity preserving property and the view preserving property. In addition, we give some synthesis results under different illuminations. Furthermore, we show how the model behaves with a varying number of unlabeled data.

Visual quality. Figures 5 and 6, respectively, show the image reconstruction results of the proposed method on two data sets. Figure 7 shows the face correction results of DR-GAN, CR-GAN and the proposed method. The method

Table 5 Multi-PIE and 300W-LP data set usage

Data set	Training set	Test set	Training/test	Number of views	Number of labels
Multi-PIE	103584	25896	1/4	13	3000
300W-LP	97960	24480	1/4	9	2500

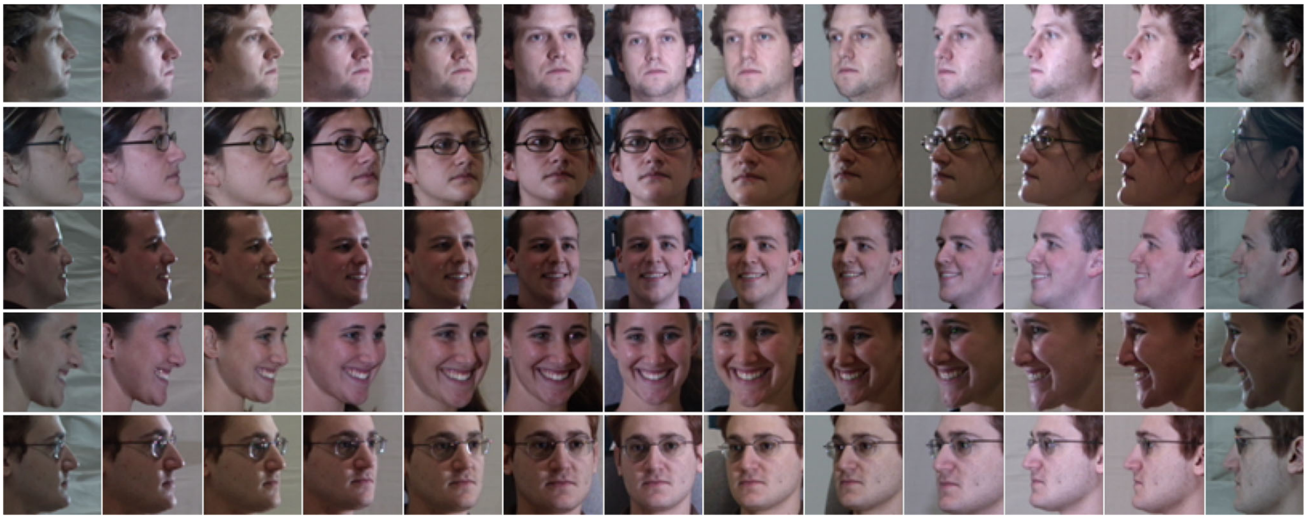


Fig. 3 Example images from Multi-PIE

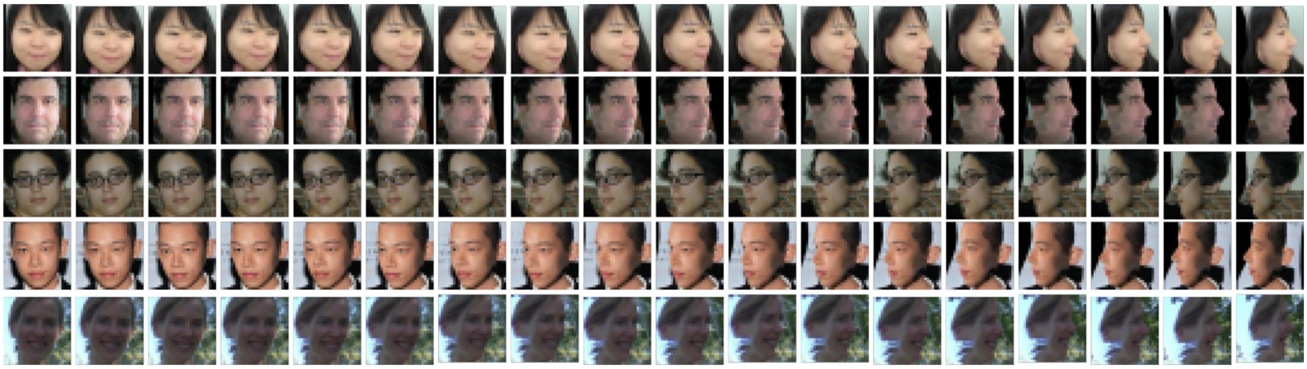
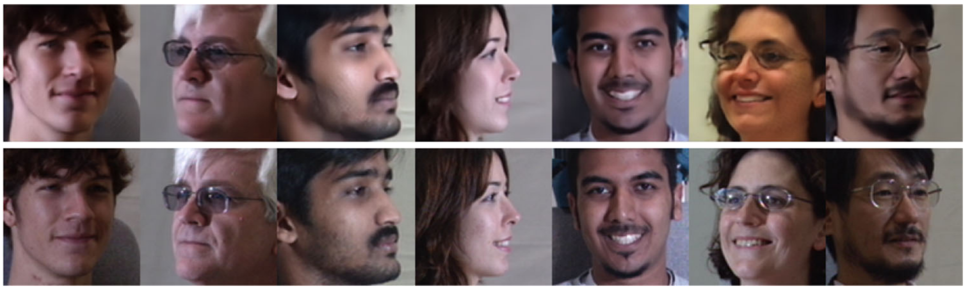


Fig. 4 Example images from 300W-LP

Fig. 5 Reconstruction results on Multi-PIE. The first row is the reconstructed image, and the second row is the image in the data set



proposed in this paper generates a realistic face image that is very similar to the image in the data set. CR-GAN and DR-GAN also have good results, but these two models require paired images and a large number of labels during training. In addition, the method proposed in this paper also has a good effect when inputting large-scale face images. Figure 8 shows the results of the method in this paper. The generated image is very similar to the input image, and there are continuous angle changes. This shows that the model proposed in this paper not only unlocks the identity representation and view representation of face images, but also can synthesize realistic face images. Figure 9 shows the results of CR-GAN and this method on 300W-LP. The method in this paper can synthesize high-quality images, but the CR-GAN-synthesized image is far from the real image, and it is easy to produce distortion.

Identity preserving property. To evaluate identity preserving property of our model, we randomly select 10 views for each identity on Multi-PIE session1 (249 identities), and input all generated images of the same identity into FaceNet [54] to calculate the L2 distance between each two images. The L2 distance reflects the similarity of the face, the faces

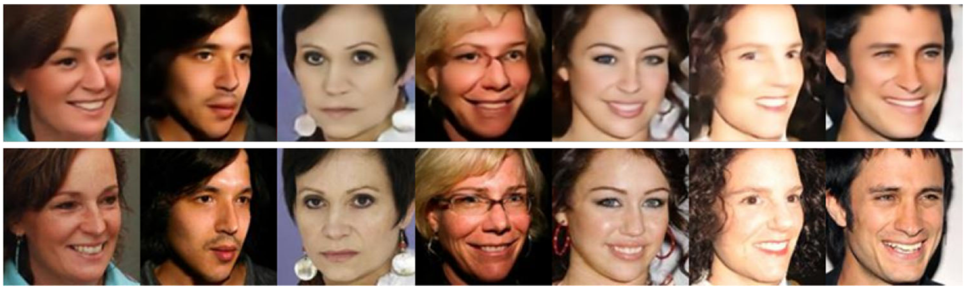


Fig. 6 Reconstruction results on 300W-LP. The first row is the reconstructed image, and the second row is the image in the data set

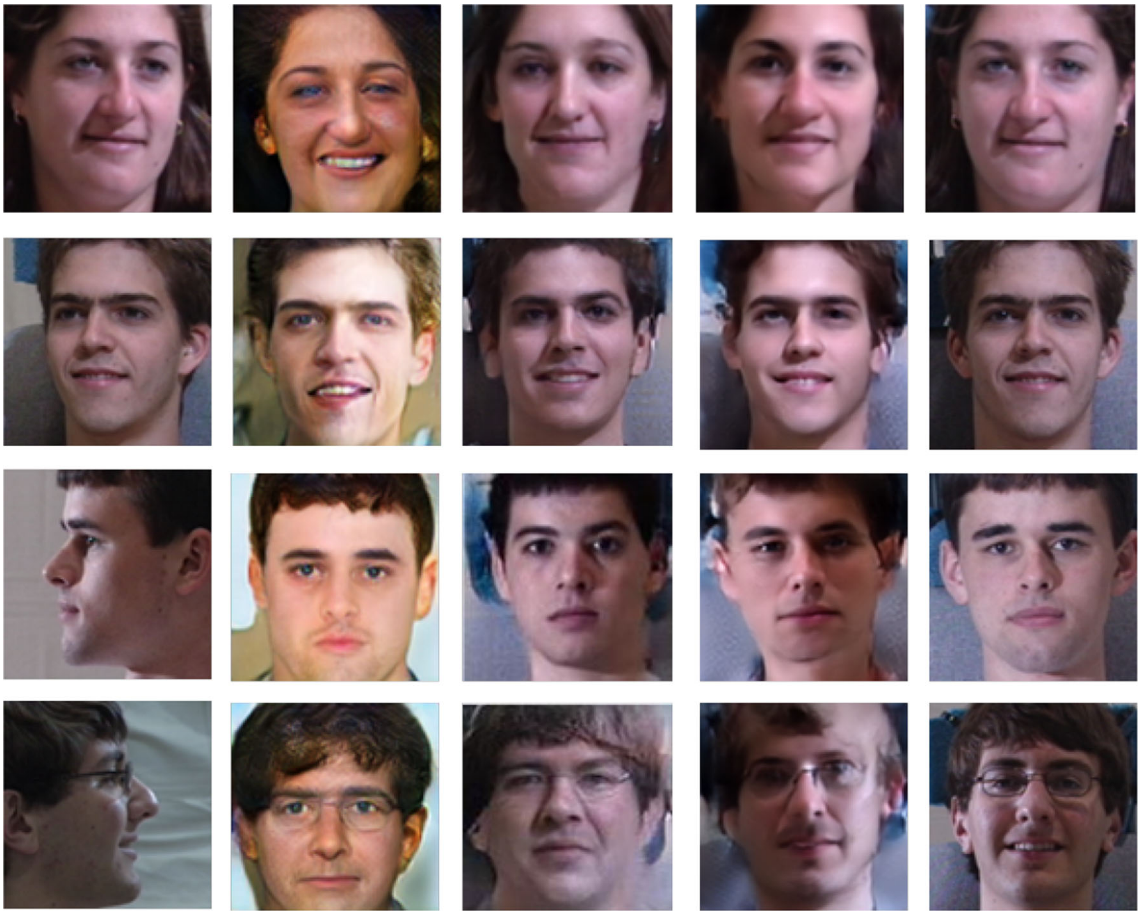


Fig. 7 Results of face frontalization of DR-GAN(Col2), CR-GAN(Col3) and our method(Col4). The first column is the input image, and the fifth column is the corresponding frontal face image

of different views of the same identity should have a small L2 distance, and there should be a large L2 distance between different identity faces. The mean and variance of the L2 distances for DR-GAN, CR-GAN and our method are shown in Table 6. Besides, we compare Frechet Inception Distance (FID)[55] with CR-GAN and DR-GAN. We calculate the FID between the real faces and the synthesized faces. The results are shown in Table 7. Tables 6 and 7 show that our method has a small gap with DR-GAN and CR-GAN. It should be noted that our method does not use identity labels, and each face generates 13 images of different views, while DR-GAN and CR-GAN only generates 9 images in the case of using identity label; that is to say, in this statistic, the results of our method include images of 13 views generated from large pose face images, and large pose face images generated from other views.

View preserving property. To evaluate the view preserving property of our model, we use the A third-party head pose estimator (THPE)[1]. to calculate the yaw angle of the real images and the images generated by our model on Multi-PIE. THPE can only calculate the yaw angle within ±45°,

[1] https://github.com/guozhongluo/head-pose-estimation-and-face-landmark

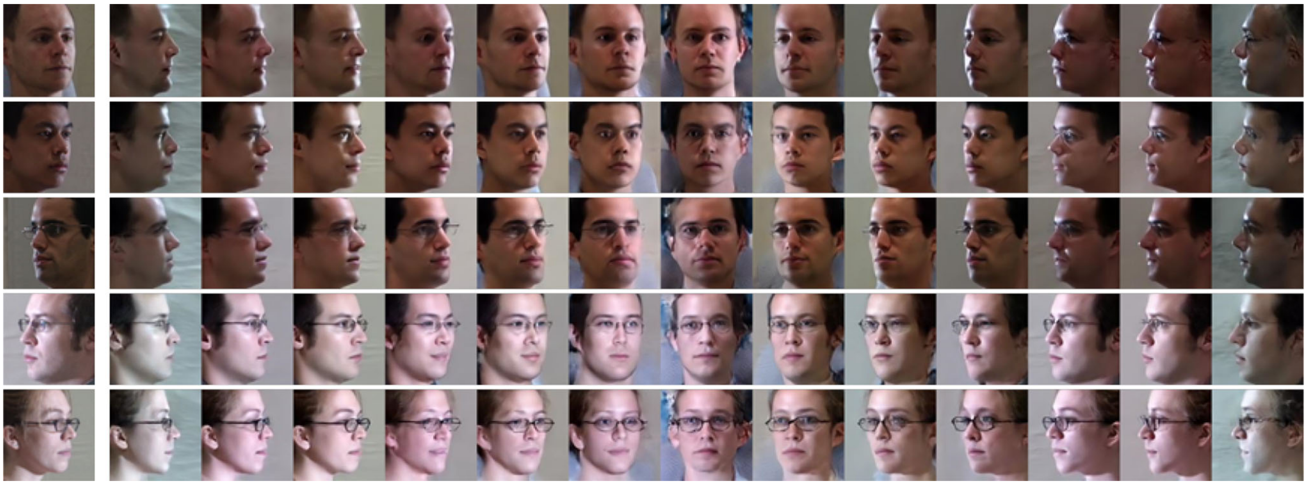


Fig. 8 Results of our method; the first image in each row is the input image

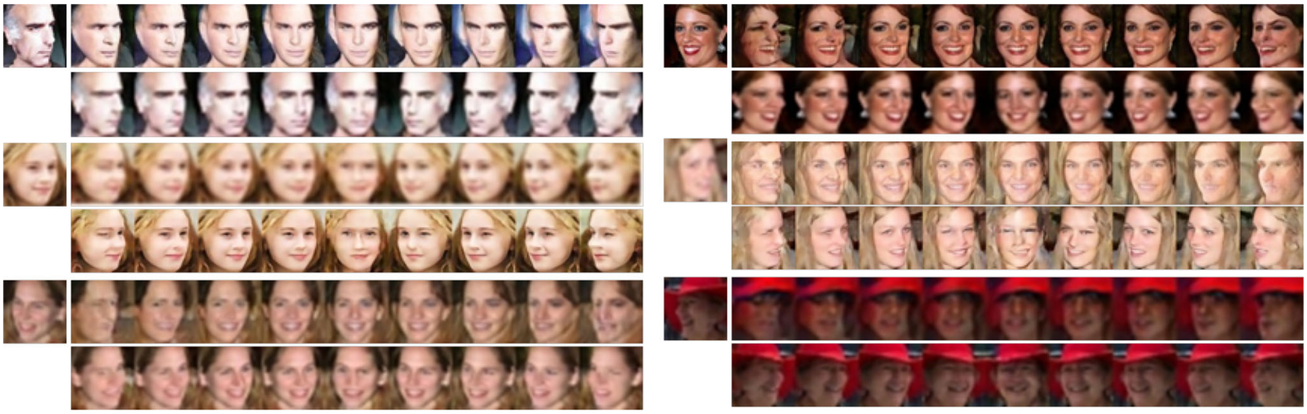


Fig. 9 Results of CR-GAN and our method on 300W-LP. Images synthesized by CR-GAN (Row 1) are quite different from real images and are prone to distorted, and our method (Row 2) can synthesize high-quality images.

Table 6 Identity similarities between real and generated images

	Mean	Variance
CR-GAN[13]	0.871	0.057
DR-GAN[12]	0.914	0.051
Ours	0.939	0.060

Table 7 FID comparisons with CR-GAN [13], DR-GAN [12] and our method on the Multi-PIE database

	CR-GAN	DR-GAN	Ours
FID	16.93	20.54	21.19

Table 8 Mean head pose estimation (in degree) comparisons with CR-GAN [13], DR-GAN [12] and our method on the Multi-PIE database predicted by THPE

	±45°	±30°	±15°	0°
Genuine data	38.50	29.18	16.76	1.17
CR-GAN	38.94	29.07	16.82	1.40
DR-GAN	38.53	29.22	16.91	1.33
Ours	37.40	29.20	17.75	1.95

so we only tested the face image in this range. Table 8 shows the average yaw angles of real images and images generated by CR-GAN, DR-GAN and our method, respectively. The results show that there are small mean head pose estimation errors between the multi-view images generated by our model and the real images, and the results between our method and the CR-GAN and DR-GAN are very close. Note that our method is based on semi-supervised learning, using only a very small number of views labels.

Synthesis results under different illuminations. In order to compare the synthesis results of our method under different illuminations, we selected the face images of 100 identities in the Multi-PIE data set under 5 illumination conditions (No. 00–No. 04) as the test set. Figure 10 shows the synthesis

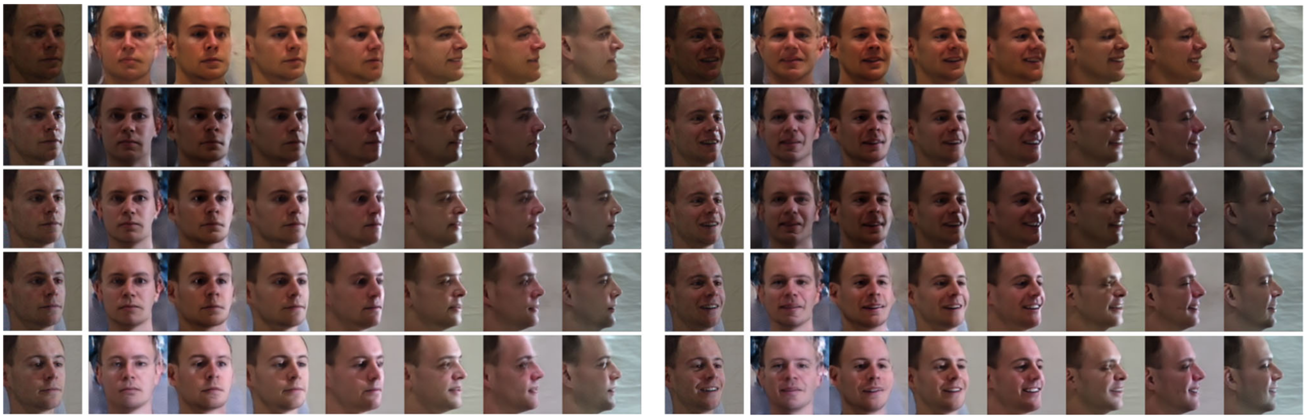


Fig. 10 Synthesis results under different illuminations

Table 9 FID and mean head pose estimation errors of the images generated by our method under different illuminations

	00	01	02	03	04
FID	21.15	21.14	21.20	21.24	21.22
Mean head pose estimation error	0.019	0.018	0.020	0.021	0.019

Table 11 Results in cross-database experiments

Database	color FERET	Pointing'04
FID	44.79	37.02
Mean head pose estimation error	0.033	0.021

results of the same identity under 5 illumination conditions (the same identity contains two expressions under each illumination). It can be seen that the our method can generate high-visual-quality images under different illumination conditions. Table 9 shows the FID and mean head pose estimation errors between the real faces and the synthesized faces under 5 lighting conditions. (Mean head pose estimation error is the difference between the real image head pose and the generated image head pose calculated by THPE when the yaw angle is 30°.) The results show that the FID and mean head pose estimation errors of the images generated by our method under different illuminations is very close.

The model behaviors with a varying number of unlabeled data. For semi-supervised learning, the more the labeled data, the better the performance of the model. Therefore, in practical application, we are concerned about how to make use of a limited amount of labeled samples to achieve optimal model performance. In order to explore the influence of the number of unlabeled samples on the model accuracy in the semi-supervised learning method, this experiment divides the training set images into two parts. The first part is labeled data, with a total of 3000 images. The second part is data without labels, with the number of images set as 0, 6000, 8000, 10000, 12000 and 15000, respectively. Six training sets were, respectively, inputted into the model for training, where the epoch was set as 20. The results are shown in Table 10. It can be seen that when the number of samples without labels increases from 0 to 6000, the cross-entropy of the model decreases rapidly. With the increase of the number of samples without labels, the cross-entropy of the model decreases gradually. In addition, we input 1000 randomly selected images from the test set into 6 models obtained by training for testing. Experimental results show that models trained with more unlabeled samples had smaller FID value and mean head pose estimation error. These results indicate that the unlabeled samples have a positive effect on the training effect of the model, and also prove the effectiveness of the semi-supervised model designed in this paper.

Cross-database experiment. In order to verify the performance of our model in cross-database experiments, we randomly selected 500 images from color FERET [56] and Pointing'04 [57], respectively, and input them into the model trained by Multi-PIE data set to calculate FID and mean head pose estimation error. The results are shown in Table 11. Figure 11 shows the multi-view face images generated by our

Table 10 Results with a varying number of unlabeled data

Number of unlabeled data	0	6000	8000	10000	12000	15000
Cross-entropy	0.5108	0.2533	0.2139	0.1835	0.1512	0.1076
FID	66.41	51.92	49.20	38.77	35.48	31.05
Mean head pose estimation error	0.837	0.501	0.412	0.371	0.325	0.244

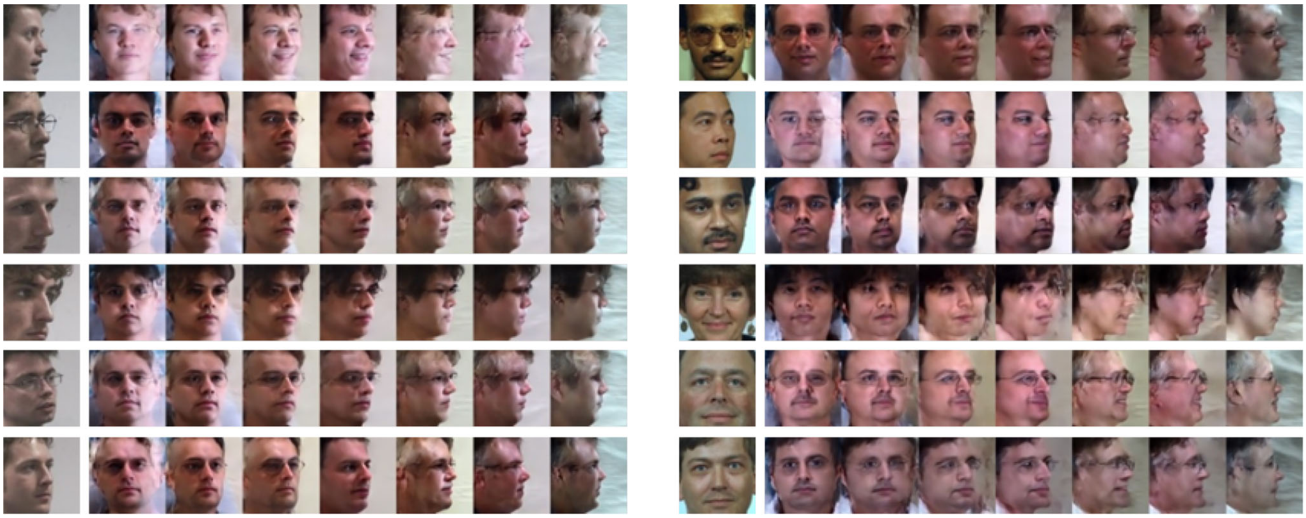


Fig. 11 Results in cross-database experiments. The first image in each row is the input image. The left is the result on Pointing'04, and the right is the result on color FERET

method in cross-database experiments. These results show that there are small mean head pose estimation errors between multi-view images generated by our model and real images, which indicates that our model has reliable head pose estimation ability. However, the face output is relatively fuzzy, and the similarity between the synthetic face and the input face is very different. This is because the proposed model uses a single path, i.e., using an encoder to map the input image to a latent space, and then reconstructs the image through a generator. The proposed model may lack the generalization ability because with limited number of training samples, the output of the encoder only constitutes the subspace of the latent variable of the face image. This may make the generator only "see" a part of the face image representation. When the sample of other data set is inputted for testing, the latent variable of the face image of the input encoder may be outside the subspace, causing the generator to reconstruct different face images.

5 Conclusions and discussion

A method for generating multiple views via unpaired images is presented in this paper. Our method is based on adversarial autoencoder and generative adversarial network. The identity and the view representations are disentangled, and the realistic face image can be reconstructed by training the five sub-networks. During the test, multiple one-hot vectors are imposed on the identity representation, so that the generated images not only preserve the identity characteristics, but also have a continuous view variation. Compared with other multi-view face generation methods, our method does not need to use paired face images in training, does not rely on the identity label of the data set and only needs a few view labels.

However, our method has undoubtedly increased the complexity of the model. Compared with DR-GAN [12] and CR-GAN [13], FLOPs and parameters have increased, so compression of the model should be considered in the future. In addition, the method proposed in this paper focuses on solving the problem that pairs of images and a large number of labels must be used in the traditional training models, and does not pay attention to the generalization of the model. Therefore, the model proposed in this paper does not perform well in cross-database experiments. To enhance the generalization of the model, we will try to add another generation path in the future research to ensure the integrity of the latent space. What is more, the literature [16] employs identity preserving loss to preserve identity information. Specifically, they choose a pretrained Light CNN as identity preserving network, it makes the same subject form a compact cluster with small intra-class distances and variances in embedding space. In this paper, we use Formula (5) to preserve identity information. The literature [16] uses a more sophisticated approach to preserve identity information, and produces realistic images. The method proposed in this paper focuses on solving the problem that pairs of images and a large number of labels must be used in the traditional training model, we will try to add an identity preserving loss as the literature [16] in the future research to make generated images more realistic.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 62076117, Grant 61762061 and 61862044, the Natural Science Foundation of Jiangxi Province, China, under Grant 20161ACB20004, and Jiangxi Key Laboratory of Smart City under Grant No. 20192BCD40002.

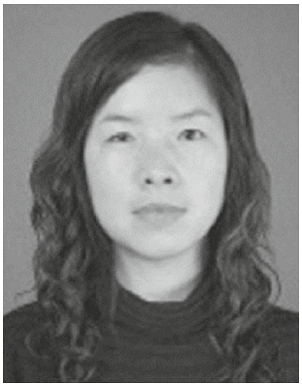
References

1. Blanz, V., Vetter, T., Rockwood, A.: A morphable model for the synthesis of 3D faces. ACM Siggraph. **187–194**, (2002)
2. Luo, J., Juyong, Z., Bailin, D., et al.: 3D face reconstruction with geometry details from a single image. IEEE Trans. Image Process. **27**, 4756 (2018)
3. Booth, J., Roussos, A., Ponniah, A., et al.: Large scale 3d morphable models. Int. J. Comput. Vision. pp. 1-22 (2017)
4. Zhou, H., Liu, J., Liu, Z., et al.: Rotate-and-render: unsupervised photorealistic face rotation from single-view images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
5. Tewariet, A., al.: StyleRig: Rigging StyleGAN for 3D Control over Portrait Images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
6. Ma, M., Peng, S., Hu, X.: A lighting robust fitting approach of 3D morphable model for face reconstruction. V. Computer **32**(10), 1223 (2016)
7. Isola, P., Zhu, J, Y., Zhou, T., et al.: Image-to-image translation with conditional adversarial networks[J]. (2016)
8. Jia, X., Ghodrati, A., Pedersoli, M., et al.: Towards automatic image editing: learning to see another you. In: British Machine Vision Conference (BMVC). (2016)
9. Zhu, Z., Luo, P., Wang, X., et al.: Multi-view perceptron: a deep model for learning face identity and view representations. In: Annual Conference on Neural Information Processing Systems (NIPS), pp. 217-225 (2014)
10. Fu, C., et al.: High fidelity face manipulation with extreme pose and expression. arXiv:1903.12003. (2019)
11. Goodfellow, I.J., Pouget, J., Mirza, M., et al.: Generative adversarial nets. In: annual conference on neural Information processing systems (NIPS), pp. 2672-2680. (2014)
12. Tran, L., Yin, X., Liu, X.: Disentangled representation learning GAN for pose-invariant face recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), volume 3, page 7. (2017)
13. Tian, Y., Peng, X., Zhao, L., et al.: CR-GAN: Learning complete representations for multi-view generation. In: International joint conference on artificial intelligence (IJCAI), pp. 942-948. (2018)
14. Huang, R., Zhang, S., Li, T., et al.: Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: IEEE International Conference on Computer Vision (ICCV). (2017)
15. Cao, J., Hu, Y., Yu, B., et al.: Load balanced gans for multi-view face image synthesis. arXiv preprint arXiv:1802.07447. (2018)
16. Hu, Y., Wu, X., Yu, B., et al.: Pose-guided photorealistic face rotation. In: IEEE conference on computer vision and pattern recognition (CVPR). (2018)
17. Sanchez, E., et al.: A recurrent cycle consistency loss for progressive face-to-face synthesis. in FG. (2020)
18. Donahue, C., Lipton, Z. C., et al.: Semantically decomposing the latent spaces of generative adversarial networks. In: International Conference on Learning Representations (ICLR). (2018)
19. Chen, M., Denoyer, L., et al.: Multi-view data generation without view supervision. In: International Conference on Learning Representations (ICLR). (2018)
20. Emily, L., Denton., Vighnesh, B.: Unsupervised learning of disentangled representations from video. In: Annual conference on neural iInformation processing ystems (NIPS). (2017)
21. Deng, Y., Yang, J., Chen, D., et al.: Disentangled and controllable face image generation via 3D imitative-contrastive learning. In: IEEE conference on computer vision and pattern recognition (CVPR). (2020)
22. Gross, R., Matthews, I., Cohn, J., et al.: Multi-pie. Image V. Comput. **28**(5), 807–813 (2010)
23. Zhu, J.Y., Park, T., Isola, P., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision (ICCV). (2017)
24. Mirza, M., Osindero, S.: Conditional generative adversarial nets. Comput. Sci. **2672–2680**, (2014)
25. Denton, E., Gross, S., Fergus, R.: Semi-supervised learning with context-conditional generative adversarial networks. arXiv preprint arXiv:1611.06430. (2016)
26. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: International conference on machine learning (ICML), pp. 2642-2651. (2017)
27. Chen, X., Duan, Y., Houthooft, R., et al.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Annual Conference on Neural Information Processing Systems (NIPS). (2016)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International conference on learning representations (ICLR). (2014)
29. Bao, J., Chen, D., Wen, F.: Cvae-gan: fine-grained image generation through asymmetric training. In: IEEE International Conference on Computer Vision (ICCV). (2017)
30. Makhzani, A., Shlens, J., Jaitly, N., et al.: Adversarial autoencoders. In: International Conference on Learning Representations (ICLR). (2016)
31. Hassner, T., Harel, S., Paz, E., et al.: Effective face frontalization in unconstrained images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4295-4304. (2015)
32. Zhu, X., Lei, Z., Yan, J., et al.: High-fidelity pose and expression normalization for face recognition in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
33. Li, S., Liu, X., Chai X., et al.: Morphable displacement field based image matching for face recognition across pose. In: European conference on computer vision (ECCV). (2012)
34. Bas, A., Smith, W., Bolkart, T., et al.: Fitting a 3d morphable model to edges: a comparison between hard and soft correspondences. In: Asian conference on computer vision (ACCV), pp. 377-391. (2016)
35. Sagonas, C., Panagakis, Y., Zafeiriou, S.: Robust statistical face frontalization. In: IEEE international conference on computer vision (ICCV). (2015)
36. Sagonas, C., Panagakis, Y., Zafeiriou, S., et al.: Robust statistical frontalization of human and animal faces. Int. J. Comput. **122**(2), 270–291 (2017)
37. Yang, J., Reed, S., Yang, M.H., et al.: Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In: Annual conference on neural information processing systems (NIPS), pp. 1099-1107. (2015)
38. Yim, N.J., Jung, N.H., Yoo, B.I., et al.: Rotating your face using multi-task deep neural network. In: IEEE conference on computer vision and pattern recognition (CVPR), pp. 676-684. (2015)
39. Yin, X., Yu, X., Sohn, K., et al.: Towards large-pose face frontalization in the wild. In: IEEE international conference on computer vision (ICCV). (2017)
40. Sanghoon, K., Jinmook, L., Kyeongryeol, B., et al.: Low-power scalable 3-d face frontalization processor for cnn-based face recognition in mobile devices. IEEE Journal on Emerging and Selected Topics in Circuits and Systems. (2018)
41. Blanz, V., Vetter, T., S. : Face recognition based on fitting a 3d morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003)
42. Eduardo, P., Carlos, C.: Unsupervised learning for concept detection in medical images: a comparative analysis. Appl. Sci. **8**, 1213 (2018)
43. Jirui, Y., Ke, G., Pengfei, Z, K., et al.: Multi-view predictive latent space learning. Pattern Recognition Letters. (2018)

44. Guillaume, L., Neil, Z., Nicolas, U., et al.: Fader networks: manipulating images by sliding attributes. In: Annual conference on neural information processing systems (NIPS). (2017)
45. Tang, Y., Han, X., Li, Y., et al.: Expressive facial style transfer for personalized memes mimic. V. Comput. **35**(6–8), 783–795 (2019)
46. Liu, H., Li, C., Lei, D., et al.: Unsupervised video-to-video translation with preservation of frame modification tendency. V. Comput. **36**, 2105 (2020)
47. Huang, R., Ye, M., Xu, P., et al.: Learning to pool high-level features for face representation. V. Comput **31**, 1683 (2015)
48. Odena, A.: Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583. (2016)
49. Salimans, T., Goodfellow, I.J., Zaremba, W., et al.: Improved techniques for training gans. In: Annual Conference on Neural Information Processing Systems (NIPS). (2016)
50. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. In: International Conference on Learning Representations (ICLR). (2016)
51. Zhu, X., Lei, Z., Liu, X, Shi H., et al.: Face alignment across large poses: a 3D solution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
52. Viola, P., Jones, M.: Fast and robust classification using asymmetric AdaBoost and a detector cascade. Adv. Neural Inf. Process. Syst (2002)
53. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). (2015)
54. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815-823. (2015)
55. Martin, H., Hubert, R., Thomas, U., et al.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Adv. Neural Inf. Process. Syst. (2017)
56. Phillips, D.P., Moon, H., et al.: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
57. GourierD, N., Hall, Crowley, J.: Estimating face orientation from robust detection of salient facial structures. in Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures, pp. 17-25. (2004)

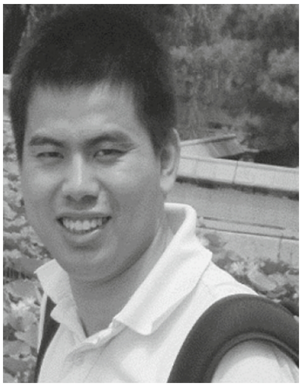
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Shuai Wang received the B.E. degree in digital media technology from Xi'an University of Technology, China in 2015. He is currently pursuing the master's degree in computer vision at Nanchang University, China.

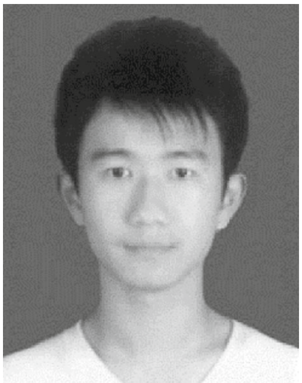


Yanni Zou received the Ph.D. degree from Nanchang University, Nanchang, China in 2016. She has been with the School of Information Engineering, Nanchang University, Nanchang 330031 China since 2017. Her research interests include virtual surgery, image processing, biomechanical calculation etc. She has authored in various journals, such as IEEE Transactions on cybernetics, etc.

Weidong Min received the B.E., M.E. and Ph.D. degrees in computer application from Tsinghua University, China in 1989, 1991 and 1995, respectively. He is currently a Professor and the Dean, School of Software, Nanchang University, China. He is an Executive Director of China Society of Image and Graphics. His current research interests include image and video processing, artificial intelligence, big data, distributed system and smart city information technology. Since 2015 he has been a Professor with Nanchang University, China. From 2011 to 2014 he cooperated with School of Computer Science & Software Engineering, Tianjin Polytechnic University, China. From 1998 to 2014 he worked as a Senior Researcher and Senior Project Manager at Corel and other companies in Canada. From 1995 to 1997 he was a Post-Doctoral Researcher at the University of Alberta, Canada. From 1994 to 1995 he was an Assistant Professor at Tsinghua University, China.



Jiansheng Wu received the B.S. and Ph.D. degrees from the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China, in 2009 and 2015, respectively. He joined the School of Information Engineering, Nanchang University, in 2015. His current research interests include machine learning and data mining, especially focusing on large scale data clustering.



Xin Xiong received the M.E. degree in control theory and control engineering from Nanchang Hangkong University, China in 2015. He is currently pursuing the Ph.D. degree at Nanchang University, China. His current research focuses on computer vision.