The Visual Computer (2022) 38:1151–1164
https://doi.org/10.1007/s00371-021-02074-w

ORIGINAL ARTICLE

Facial expression GAN for voice-driven face generation

Zheng Fang[1] · Zhen Liu[1] · Tingting Liu[2] · Chih-Chieh Hung[3] · Jiangjian Xiao[4] · Guangjin Feng[1]

Accepted: 25 January 2021 / Published online: 22 February 2021

Abstract
Cross-modal audiovisual generation is an emerging topic in machine learning. In particular, voice-to-face is one of the most popular research branches, which aims to generate faces from human voice clips. Most recent works in voice-to-face generation do not take emotion information into account. However, it could be widely observed that expressions are the key face attributes to reconstruct sharper and more discriminative faces. In this paper, we propose a novel facial expression GAN (FE-GAN) which takes emotion and expressions into account in face generation. To achieve this goal, we use two auxiliary classifiers to learn more emotion and identity representations between different modalities, respectively. Moreover, we design two discriminators, each focusing on a different aspect of the faces, to measure identity and emotion semantic relevance in generating. The triple loss is designed to make FE-GAN robust to voice variety and keep balance in two different modalities. Extensive experiments are conducted on two real datasets to demonstrate the effectiveness of FE-GAN in both quantitative and qualitative perspectives. The experimental results show that FE-GAN can not only outperform the previous models in terms of FID and IS values, but also generate more realistic face images compared with previous models.

Keywords Expression reconstruction · Cross-model generation · Voice-to-face generation · Generative adversarial networks

1 Introduction

Cross-modal generation aims to generate data from one modality conditioned on another correlated modality, which has attracted a lot of research efforts. Early researches on cross-model generation usually generate low-dimensional data from high-dimensional data, such as voice-to-text [1,2] and image-to-text [3,4]. Recently, thanks to the rapid growth of generative adversarial networks (GANs) [5] and with increase in multi-modal datasets [6], it is possible to generate complex data from low-dimensional data, such as text-to-image [7,8] and audio-to-image [9,10]. Note that the audio and visual information are the most important perceptual modalities in our daily life. We believe that the research on cross-modal audiovisual generation can endow machines with humanized capabilities of imagination and interpretation. Here, we leverage the voices to directly generate speakers' facial images by GANs.

Many previous works have been done in solving the audio–image generation problem. Duarte et al. proposed conditional GANs (cGANs) [11] to directly generate face from voice [12]. Later, Wen et al. used an auxiliary classifier GANs (AC-GANs) [15] to directly generate face [14]. Oh et al. leveraged the encoder–decoder network to learn the cross-modal visual-audio mutual relationship, then generated the face based on the corresponding static face and voice [13]. However, generated faces from studies above are usually with certain unsatisfactory artifacts and missing parts. The reasons are twofold. First, face generation in previous works usually considers identity information of target faces but leave alone the corresponding facial expressions. It can be observed that one's expression can usually change with different emotions when she/he talks to others. Emotion would be a key to construct high-quality facial image. Second, we find that GANs with a single discriminator are not able for learning

✉ Zhen Liu
liuzhen@nbu.edu.cn

Zheng Fang
1901100018@nbu.edu.cn

1 Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China

2 College of Science and Technology, Ningbo University, Ningbo, China

3 National Chung Hsing University, Taichung City, Taiwan

4 Ningbo Institute of Materials, Chinese Academy of Sciences, Ningbo, China

the complex mapping relationships between audio and visual modalities. That is, a constraint that only allows the generated images to be on the one manifold of the truth data distribution. To improve the quality of faces generated, it is a promising way to have multiple discriminators, rather than only one, to help the generator learn from more fine-grained face features extracted from audio.

In this paper, we propose a novel model, facial expression GANs (abbr. as FE-GAN) to generate faces by given voice information. In a nutshell, FE-GAN considers emotion and identity variations from face and voice simultaneously. The existence of semantic consistency in human's voice and face [16,17] inspires us to adopt both identify labels and emotion labels for model training. Specifically, more discriminators could take the emotion and identity constraints into account so that the generator can also retain more emotion and identity characteristics. The simplified pipeline of the proposed method is shown in Fig. 1. The core of FE-GAN is composed of one generator network (G-net) and two discriminator-classifier pairs, say (C1-net, D1-net) and (C2-net, D2-net). In the generating process, the voice encoder extracts the Fbank features F_{v} from voice clip V and obtain voice embedding E_{v} by V-net. Next, taking E_{v} as an input, generator G-net generates the face image I_{G}. Finally, the two discriminators D1-net and D2-net are used to distinguish whether or not a face image is real or fake, meanwhile, the auxiliary classifier C1-net and C2-net predict its identity and emotion. This design of FE-GAN can not only learn one-to-one mapping between faces and voices but also capture various emotions of the target person that are correlated with the input speech. Our contributions can be summarized as follows:

(1) We propose an effective GAN model (FE-GAN) for cross-modal voice-to-face generation. It explores the emotional and identity relationship in cross-modal voice-to-image task and generates sharper facial images with expression. (2) We adopt two discriminators and two classifiers in GANs. They help the model generate more realistic images, and transfer the label information to generator. Besides, we explore the multiply discriminators and classifiers optimization problem, a triple loss is presented to optimize the FE-GAN. (3) We conduct qualitative and quantitative experiments on RAVDESS [18] and eNTERFACE [19] dataset, the results show that FE-GAN outperforms previous GANs methods [12,14] and achieves the best performance in the series metrics with remarkable improvements. The rest of the paper is organized as follows. Section 2 lists the previous relevant research works. Section 3 gives the technical detail of the proposed approach FE-GAN. Section 4 reports the experimental results. Section 5 concludes this paper.

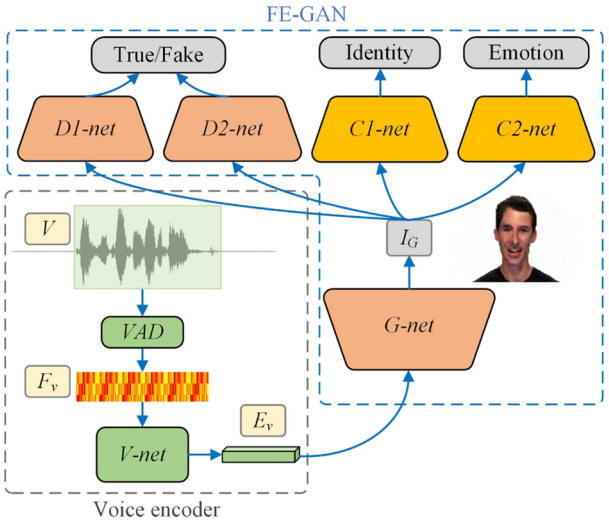

Fig. 1 The simplified pipeline of the proposed method. Our method has divided into two parts: voice encoder (gray dashed box) and FE-GAN (blue dashed box). (1) Voice encoder consists of VAD (voice activity detection) and V-net, which takes Fbank features F_{v} as input, and outputs embedding features E_{v}. (2) FE-GAN consists of five parts: G-net, C1-net, C2-net, D1-net and D2-net. The FE-GAN is used to transfer embedding E_{v} into face image I_{G}, then to predict its sources (true or fake) and categories (emotion and identity)

2 Related work

2.1 Generative adversarial networks

GANs [5] is an excellent game theory architecture. It is easily assembled with others backbone networks and mechanisms. A vanilla GANs [5] consists of two neural networks: a generator and discriminator. Given a random sample with noise, the generator attempts to generate image for fooling the discriminator. Then, the discriminator is responsible for distinguishing generated image and real image. In order to address the training instability and get high-quality generated results, many variants of GAN have been developed. For example, conditional GANs (cGANs) [11] introduces a conditional constraint to get more attribute information, the condition could be class labels, object attributes or feature embeddings. However, it will bring additional noise to network and increases extra burdens to training process. Compared to cGANs, auxiliary classifier GANs (AC-GANs) [15] leverage an additional auxiliary classifier to assist in supervising the learning process, which is share weights with discriminator and can help GANS to generate sharper images. Besides, dual discriminator GANs (D2GANs) [20] and generative multi-adversarial networks (GMANs) [21], which use multiple discriminators to improve generation performance, extend GANs architecture.

Recently, many cross-modal methods use GANs and their variants to generate face from voice [12–14,22–25]. Inspired

by the above success of GANs in cross-modal generation task, we establish our FE-GAN model based on AC-GANs [15] and D2GANs [20]. Different from the two GANs, we employ two discriminator and corresponding classifiers to guide the generator for producing photo-realistic facial images.

2.2 Audio representations selection and extraction

In human interaction, voice contains various emotions and identity information, which conveyed by linguistic information (e.g., word, sentence and language meaning) [26] and prosodic information (e.g., voice pitch, tempo, loudness and intonation) [27]. The linguistic contents are dynamic variation and highly dependent on word dictionaries and language model [26,28]. However, it is unreliable and difficult to infer speaker's emotion and identity state by linguistic features [29]. Compared to linguistic information, the prosodic information are global-level and they cannot describe the dynamic variation in voice [30]. Thus, we decide to learn audio representations from speech prosody, and transfer the emotional and identity knowledge into face images.

The quality of audio representations influences the results of generation methods. Most audio-related methods involve the analysis of a speech representations using either hand-crafted prosody features (e.g., Mel Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP), Spectrograms, Fbank and Fourier transforms), or through a neural network which indirectly learns high-level representations. Compared to these hand-crafted methods [31,32], the convolutional neural networks (CNNs) are enabled to learn robust high-dimensional features, which achieve high accuracy in emotion and identity classification [30,33,34]. Therefore, we also use CNNs as audio feature extractors (V-net) to extract emotion and identity information from prosody features. Our experiments prove that CNNs are able to learn temporal filters across features and distill an entire utterance down into a static representation by fully connected layer to model.

2.3 Audio-to-visual generation

Many methods have been proposed to reconstruct visual information from different types of audio signals. Existing studies in audio-to-visual generation mainly synthesize a specific talking face from an audio clip and a still image. For example, Chung et al. [23] use facial landmarks and voice clip to synthesize a talking face video by an encoder–decoder CNNs model. Chen et al. [22] design a cascade GANs combined RNN to learn joint features from voice clip and facial landmarks to generate talking face video on the features. Furthermore, Vougioukas et al. [24] and Yi et al. [35] consider facial expressions in generation. They adopt GANs to synthesize a talking face video from voice and image.

On the other hand, some methods try to generate lip shapes from voice to synthesize a specific identity face with lip shape. Suwajanakorn et al. [36] and Jalalifar et al. [37] use the long-short term memory (LSTM) network to generate talking mouth features from voice to synthesize a talking video of Obama conditioned on these landmarks. To improve the quality of synthesis lip, Sadoughi and Busso [38] propose cGANs to learn emotion features from the speech, and generate lip animation with different expressions. However, these methods need to parametrize the reconstructed face model a priori, this often requires post-processing using computer graphics techniques to produce realistic albeit subject-dependent results.

There are very few works try to leverage audio to directly generate facial image, which is different from the above-mentioned methods using both audio and visual modalities as inputs. Existing methods on voice-to-face generation [12–14] use CNNs to extract embedding features from input voice, then the feature is feed into the generator or decoder to generate corresponding images. Moreover, some works generate images condition on music directly [10,39]. To overcome shortcomings of the conventional cross-modal GANs model and generate more realistic face, we introduce the emotion to our facial expression GAN (FE-GAN) and perform voice-to-face generation.

3 Proposed methods

3.1 Overview of V-net and FE-GAN

This section gives the detailed architecture of V-net and FE-GAN, as shown in Fig. 2. V-net is a standard CNNs with normalization, which learns a voice embedding from speech prosody feature. FE-GAN is composed of G-net (which generates a face image from a voice embedding), D1-net with C1-net, and D2-net with C2-net (two pairs of a discriminator with its classifier to identify whether a face image is true from identity and emotion perspectives, respectively).

After extracting speakers' voice and face from videos, we can obtain the training dataset tuples of F_{v}, I_T, l_{ve}, l_{vi}, l_{fe}, l_{fi}, where F_{v} are the Fbank features extracted from speakers' voice, I_T is the face image, and l_{xy} is the label of y based on attribute x where x can be v (voice) or f (face) and y can be i (identity) or e (emotion). Given the identity label l_{vi} and the Fbank feature F_{v}, we firstly pre-train V-net to classify a person through her voice. After pre-training of V-net, the voice embeddings E_{v} of each voice can be extracted.

Subsequently, given a voice embedding E_{v} with Gaussian noise N_{g}, G-net is trained to generate the target face I_{G}. Concurrently, we use the true face I_{T} with labels (l_{fi}, l_{fe})

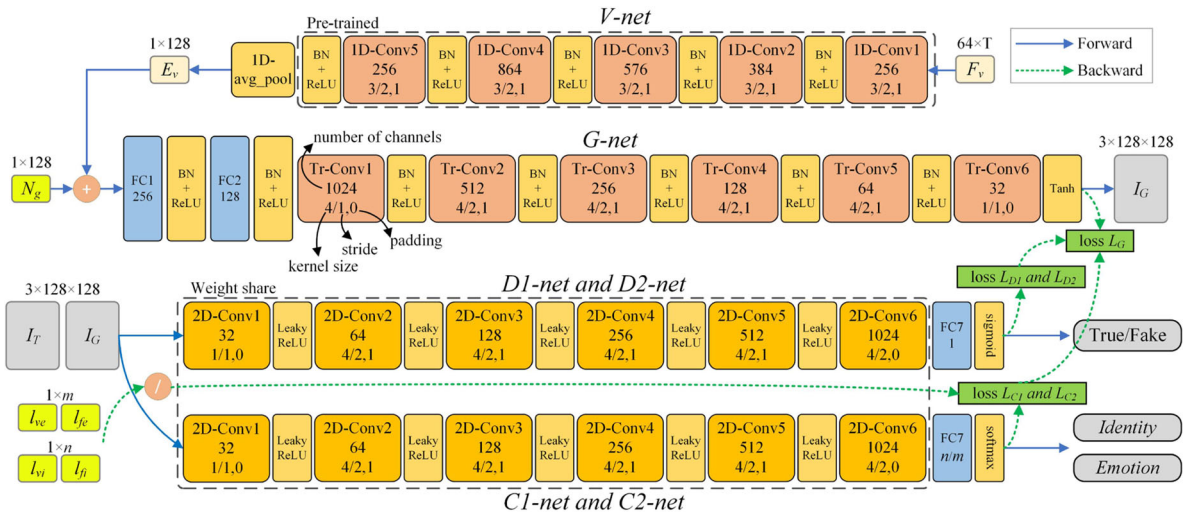

Fig. 2 The detailed architecture of V-net and FE-GAN. The symbol + represents concatenation operation; the / represents OR operation chooses the corresponding labels, and the label symbol l_{fe}, l_{fi}, l_{vi}, l_{ve} (yellow blocks) represent face emotion, face identity, voice emotion and voice identity, respectively; I_T and I_{G} (gray blocks) represent real face from dataset and fake face from G-net, respectively; blue line and green dotted line denote forward and backward propagation paths, respectively. The dimensions of input and output are denoted on top of the blocks. Besides, the loss equations L_{G}, L_{D1}, L_{D2}, L_{C1}, L_{C2} (green blocks) and other symbols are described in the rest of Sect. 3

and fake faces I_{G} with labels (l_{ve}, l_{vi}) to train the discriminators D1-net and D2-net, with the auxiliary network C1-net and C2-net. In this way, D1-net and D2-net are trained to distinguish input face image I_T or I_{G} is true or fake, respectively; C1-net and C2-net are trained to classify the emotion and identify labels of input face, respectively. Besides, the proposed triple loss combining loss equations from the generator, the discriminators and the classifiers are designed to optimize FE-GAN.

3.2 Pre-processing and V-net

We firstly use voice activity detection (VAD) module [40] for original voice to remove the silent frames (e.g., in RAVDESS datasets, the average duration time of the original voice is 3.6 s. After removing silent parts, it is shortened into 2.4 s.). Then, the voice clips are resampled at 32 kHz and a single audio channel is preserved. Next, we repeat the audio clips 3 4 times and eliminate the redundancy so that they all become 10 s long. Furthermore, Fbank features (F_{v}), MFCC and Spectrogram are calculated by fast Fourier transform with window length of 33 ms (milliseconds), and a hop length is 16 ms. In addition, we use the face detector based on Resnet-18 in Dlib [41] to detect the face regions from video, and resize them to 128∼128 pixels. To augment the training data, we use random cropping in audio features and left-right flipping in image, the cropping length is 300–800 ms.

Our V-net aim to classify features F_{v} into different identity categories and extract voice embedding features E_{v}. V-net takes $64 \times T$ (frequency $\times$ time) dimensional F_{v} as input, and outputs 1×128 dimensional features E_{v}. The top row of Fig. 2 shows the network architecture of V-net where there are 5 one-dimensional convolutional layers 1D-Conv1, 1D-Conv2, ..., and 1D-Conv5 with kernel size 3, stride 2 and padding 1 and a batch normalization (BN) operation is followed with Leaky-ReLU as the activation function. After the 5th convolutional layers, the temporal channels of F_{v} have been decimated to 256. Next, we apply the average 1D pooling layer along the temporal dimension. This allows us to efficiently aggregate information over time and makes the model applicable to input speech of varying duration. By the 1D pooling layer, V-net compresses features E_{v} to 1×128 dimensions. Besides, cross-entropy loss with softmax function is used to train V-net.

3.3 G-net

G-net will learn the emotion and identity mapping between voice embeddings and generated images so that it can generate more realistic face images to deceive discriminators. The architecture of G-net is shown in the middle row in Fig. 2. First of all, the voice embedding E_{v} is concatenated with 1×128 dimensional noise N_{g} and this concatenated embedding is mapped to $1 \times 1 \times 128$ by two fully connected layers (FC1, FC2) with BN operation and ReLU function. Then, we use 6 two-dimensional transposed convolution layers (Tr-Conv1 6) to upsample to $3 \times 128 \times 128$ dimensional I_{G}. Each layer has kernel size 4, stride 2 and padding 1 followed by

BN and ReLU. Apart from the first layer (kernel size 4, stride 1 and padding 0) and the last layer (kernel size 1, stride 1 and padding 0). The number of channels in transposed layers is 1024-512-256-128-64-32. To improve the generative capacity of G-net, we add a dropout strategy and Tanh activation function inspired by Wasserstein GANs [42].

3.4 D-net and C-net

The original AC-GANs conducts backpropagation mainly determined by one discriminator and one classifier. One discriminator only judges the images from the one perspective, but not the different semantic perspectives. Likewise, one classifier is not able to solve multi-label consistency problem. In the paper, we argue that corresponding voice and face can match with the two types semantic label. Therefore, apart from distinguishing real or fake identity attributes of the speaker from D1-net, we also distinguish real or fake emotion attributes by D2-net. To further control the label consistency in generating, we use two corresponding classifiers C1-net and C2-net to make sure the generated faces belong to the same label with input audios.

D1-net and D2-net are designed to discriminate whether the input image is real face I_T or fake I_G. In this way, the fake label and true label are, respectively, couple with I_G and I_T, then input them into D1-net and D2-net to get two scores. The two discriminators architecture is shown in bottom row in Fig. 2. They both have 6 two-dimensional convolution layers. Each layer is only followed by a Leaky-ReLU function. The number of channels in convolution layers is inverse of G-net that is 32-64-128-256-512-1024, and the other parameters like kernel size, stride are also inverse. Finally, we apply a FC7 with 1 channel and sigmoid activation function to obtain a score as the output. Besides, our discriminators base on DCGANs [43] architecture.

C1-net is emotion classifier that helps achieve expression reconstruction of the speakers. And the C2-net is identity classifier that ensures the speakers' facial identity. In other words, the emotional category of I_G and corresponding voice emotion label Lve should keep consistent, and face emotion label Lfe is consistent with the category of I_T. In addition, C1-net and C2-net share weights with the convolution layers in D1-net and D2-net, respectively. The architectures of the classifiers are similar to D1-net and D2-net, as shown in bottom row in Fig. 2, they also consist of the 6 two-dimensional convolution layers followed by Leaky-ReLU functions, a FC7 and softmax function. The FC7 of the two classifiers have i and m channels, respectively (i denotes the number of speakers, and m denotes voice emotion categories).

3.5 Triple loss

Our triple loss is composed of three parts: The G-net loss L_G, two discriminator losses L_{D1} and L_{D2}, and two classifier losses L_{C1} and L_{C2}. The generator and discriminator losses are both designed to reduce the differences between true face I_T and generated face I_G. The classifier losses target to guarantee the semantic consistency, which can control the generated faces in the specific class domains. Here, we use these losses to optimize the different parts of FE-GAN, the backpropagating paths of these losses are shown in Fig. 2.

First, we adopt the cross-entropy loss with softmax activation as losses of two classifiers. Here, the loss equations of L_{C1} and L_{C2} are defined as:

$$\begin{aligned} L_{C1} = & -\sum_{j=1}^{n} p\left(l_{\text{fi}}^{j}\right) \log\left(p\left(l_{\text{fi}}^{j}\left(\text{C1}, I_T\right)\right)\right) \\ & -\sum_{j=1}^{n} p\left(l_{\text{vi}}^{j}\right) \log\left(p\left(l_{\text{vi}}^{j}\left(\text{C1}, I_G\right)\right)\right) \end{aligned} \tag{1}$$

where $p(l)$ denotes the probability of the label l, l_{fi}^{j} and l_{vi}^{j} denotes the j-th face and voice identity labels, respectively; $l_{\text{fi}}^{j}(\text{C1}, I_T)$ and $l_{\text{vi}}^{j}(\text{C1}, I_G)$ denotes that the predicted label by C1-net given the true and generated faces are the j-th face identify label, respectively; n denotes the numbers of identity categories.

$$\begin{aligned} L_{C2} = & -\sum_{j=1}^{m} p\left(l_{\text{fe}}^{j}\right) \log\left(p\left(l_{\text{fe}}^{j}\left(\text{C2}, I_T\right)\right)\right) \\ & -\sum_{j=1}^{m} p\left(l_{\text{ve}}^{j}\right) \log\left(p\left(l_{\text{ve}}^{j}\left(\text{C2}, I_G\right)\right)\right) \end{aligned} \tag{2}$$

where $p(l)$ denotes the probability of the label l, l_{fe}^{j} and l_{ve}^{j} denotes the j-th face and voice emotion labels, respectively; $l_{\text{fe}}^{j}(\text{C2}, I_T)$ and $l_{\text{fe}}^{j}(\text{C2}, I_G)$ denotes that the predicted label by C2-net given the true and generated faces are the j-th emotion label, respectively; m denotes the numbers of emotion categories.

Then, the generator loss L_G of G-net is defined as:

$$\begin{aligned} L_G = & \frac{1}{2} E_{(e_v, N_g)\sim \text{data}} [-\log D_1\left(G\left(e_v, N_g\right)\right) \\ & -\log D_2\left(G\left(e_v, N_g\right)\right)] \end{aligned} \tag{3}$$

where D_1, D_2 and G represent the discriminators D1-net, D2-net and generator G-net, respectively; embedding feature E_v is from V-net; $G(E_v, N_g)$ takes E_v and a random noise N_g as input, and generates a fake image I_G, that is, $G\left(e_v, N_g\right) = I_G$; $D_1(G(.))$is the score assigned by discriminator D1-net,

$D_2(G(.))$ is similar to D_1 that is the score assigned by D2-net, e.g., $D_1(I_T)$ is the score from D1-net given a real image I_T.

Meanwhile, the two discriminators loss L_{D1} and L_{D2} are formulated as:

$$L_{\mathrm{D}_{i=1,2}} = E_{(I_T)\sim\mathrm{data}}\left[\log\left(D_i\left(I_T\right)\right)\right] + E_{(e_\mathrm{v},N_\mathrm{g})\sim\mathrm{data}}\left[\log\left(1-D_i\left(G\left(E_\mathrm{v},N_\mathrm{g}\right)\right)\right)\right]. \tag{4}$$

Finally, we implement cross-entropy loss with sigmoid function as loss functions L_G, $L_{\mathrm{D1}}$and L_{D2}, and our triple loss L_{triple} is a combination of the above four losses:

$$\arg\min_{\{\mathrm{G,C1,C2}\}}\max_{\{\mathrm{D1,D2}\}} L_{\mathrm{triple}} = L_\mathrm{G} + \lambda_1 L_{\mathrm{D1}} + \lambda_2 L_{\mathrm{D2}} + L_{\mathrm{C1}} + L_{\mathrm{C2}} \tag{5}$$

where λ_1 and λ_2 are the hyper-parameters to control the relative weight of L_{D1} and L_{D2}, respectively. In triple loss, the generator learns to minimize the score that can obtained by the generated I_G, then the two discriminators learn to give higher score to the real images I_T and give lower score to the generated images I_G to maximize L_{D1} and L_{D2}. Besides, the two classifiers need to minimize L_{C1} and L_{C2} between the predicted label from I_G or I_T and the target emotion and identity label.

Note that the proposed triple loss is different from the loss in Triangle GANs (Δ-GANs) [44] and Triple GANs [45], which adopts their triple loss between the input image and the reconstruction image in the image space. In this paper, we employ the two different modalities triple loss to optimize our FE-GAN. In addition, FE-GAN is trained in a semi-supervised manner. The generator and the discriminators are trained iteratively. That is, the generator is fixed, and two discriminators and two classifiers are updated once. Then, we fix the discriminator, and update the parameters of the generator.

4 Experiments

4.1 Datasets and settings

To validate the performance of FE-GAN in voice-to-face generation task, our experiments are run on two multi-modal datasets: RAVDESS [18] and eNTERFACE [18]. They are collected in lab-controlled environments where the speakers are asked to read the given sentences with certain voice emotions and facial expressions. RAVDESS consists of 1440 clips, which are expressed by 24 actors with 8 emotion categories. eNTERFACE contains 1166 clips, which are expressed by 43 speakers with 6 emotion categories. Table 1 summarizes the details of the datasets used in our work.

Our model is implemented in PyTorch and trained on Nvidia GeForce RTX 2080ti. V-net and FE-GAN are trained separately. First, using RAVDESS [18] or eNTERFACE [19] datasets, V-net is pre-trained where SGD optimizer is chosen, the batch size is 64 and the initial learning rate is 0.03 which decreases by half for every 100 epochs. Next, FE-GAN is trained with Adam optimizer, the batch size is set to 64 and the learning rate is 0.0002. In addition, the hyper-parameters λ_1 and λ_2 in triple loss is 0.7 and 0.3, respectively.

4.2 Evaluation metrics

To evaluate realism and variation of the generated images, we choose Inception score (IS) [46], Fréchet Inception Distance (FID) [47] and classification accuracy as quantitative metrics.

$$\mathrm{IS}\left(g\right) = \exp\left(E_{x\sim g} D_{KL}\left(p\left(y|x\right)\middle|\middle|p\left(y\right)\right)\right) \tag{6}$$

where $x \sim g$ represents generated images from generator; $p(y)$ and $p(y|x)$ are marginal label distribution and conditional label distribution, respectively.

FID measures the quality of an overall generative images. FID computes the Wasserstein-2 distance between the generated images and the real images in the feature space from by a pre-trained Inception-v3 network [48]. The FID is defined as follows:

$$FID\left(x,g\right) = \left\|\mu_x-\mu_g\right\|_2^2 + Tr\left(\sum_x+\sum_g-2\left(\sum_x\sum_g\right)^{\frac{1}{2}}\right) \tag{7}$$

where (μ_x, μ_g) and $(\sum_x, \sum_g)$ are the means and covariances of the images from the true dataset distribution and generator's learned distribution, respectively. The authors of FID [47] shows that FID is consistent with human judgment and more robust to noise than IS.

In our experiments, a lower IS value indicates that the model can produce the images that are less variety and not associated with voice features; a higher IS indicates that the model falls into mode collapse and the images have blurry parts. Thus, the reasonable IS of models is similar to the datasets. FID is a more confident and comprehensive metric. A lower FID value means the generated images are closer to the distribution of a dataset. In addition, to evaluate the model's performance regarding the identity and emotion preservation, we compute the emotion and identity classification accuracy by VGG-Face network [49]. The way to obtain accuracy is that the VGG-Face are pre-trained on RAVDESS or eNTERFACE dataset, and then we use the pre-trained VGG-Face on our generated results. Due to the

Table 1 Summary of datasets' sample numbers, duration time and emotion categories

Datasets	Speakers	Emotion categories (numbers)	Duration time (h)
RAVDESS	24(12 M, 12 F)	Happy (384), angry (384), sad (384), surprise (384), fear (384), disgust (384), calm (384), neutral (192)	∼ 4
eNTERFACE	43(34 M, 8 F)	Happy (213), anger (217), sad (216), surprise (216), fear (216), disgust (216)	∼ 11

M for male, F for female

previous works [12,14] lack of the emotion in generating, we are not able to compute emotion accuracy of their works.

4.3 Ablation experiments

Two ablation experiments are conducted on RAVDESS dataset to (1) find which kind of audio feature is the most suitable feature for our task, and (2) analyze the contribution of each component of our FE-GAN.

First of all, we perform ablation experiment on different audio features: MFCC, Fbank and Spectrogram. Specifically, we report IS and FID by using the same model and training method with audio features varied. Table 2 shows the quantitative results. Fbank can lead to the highest FID score (58.79) and IS score (1.71) and MFCC is in the second place with a litter higher value FID (64.35) and a lower IS (1.65). Compared to Fbank and MFCC, Spectrogram performs the worst where FID score (96.16) and IS score (1.91) are the highest IS (1.91) among all the audio features. The reason may be threefold: (1) Spectrogram is too primitive so that it may include many irrelevant emotion and identity information in audio; (2) MFCC outperforms Spectrogram, but it only retains 13-dimensional features that related to speech content, and discards some information about emotion and identity; (3) Fbank is best because it preserves more prosodic and acoustic information from the inputting voice. Figure 3 shows the qualitative results. We can observe that Fbank can obtain better generated images compared with MFCC and Spectrogram. In general, images generated by Fbank is sharper and with more distinct expressions in mouth and eye. Therefore, Fbank feature is selected in the rest of experiments.

Second, we conduct an experiment to evaluate the impact of four components in FE-GAN, say (a) C1-net, (b) C2-net, (c) D1-net and (d) D2-net. These four components share the same baseline (FE-GAN) but a particular part is abandoned. That is to say, FE-GAN is running without (a) C1-net, (b) C2-net, (c) D1-net and (d) D2-net. Table 2 shows their IS and FID scores.

(a) Influence of C1-net: If adding C1-net into the model, it can make a dramatic improvement in FID by 42.3% and in IS by 12.8%. As C1-net and D1-net share weights, the well-trained C1-net can provide the basic identity information to D1-net so that the generated images of different speakers are forced to keep the consistency of identity label between voices and images.

(b) Influence of C2-net: If adding C2-net, it can lead to further improvement in FID by 47.0% and in IS by 10.4%. The emotion feature is a special identity feature that relies on facial attributes, and the single C1-net has weak representation ability to extract emotion features. Thus, we use C2-net to learn the emotion representation from voice. The union of C1-net and C2-net can progressively reduce the collapse mode in training and improve the classification accuracy of generated images.

(c) Influence of D1-net: If using D1-net, it can improve FID by 53.8% and IS by 3.4%. The discriminator loss L_{D1} provides G-net strong guidance toward the ground-truth. Besides, C1-net shares weights with D1-net, it can optimize G-net from point of view of the identity label distribution. Therefore, G-net knows the way to learn identity semantic relevance between image and voice.

(d) Influence of D2-net: The single discriminator only discerns images by one attribution and it cannot exactly control the content of generated images. Therefore, we add two discriminators to improve distinguish ability and generation performance. Extra D2-net can supply the missing emotion information to our model. Table 3 shows that D2-net can improve 34.7% in FID and 2.3% in IS, which means that the two discriminator performs better than the single discriminator and further improve image quality. The shared weights also could help to learn better D2-net.

Finally, Fig. 4 visualizes the influence of above components. The generated images by full model have more fine-grained details and are more similar to the ground truth.

Table 2 Ablation experiments: FID and IS results of different audio features, duration time, noises and components on RAVDESS dataset

Experiments	Ablation items	FID Score	IS Score
Audio features	Spectrogram	96.16	1.91
	MFCCs	64.35	1.65
	Fbank (our method)	58.79	1.71
Network components	Without C1-net or C2-net	101.94/110.84	1.96/1.91
	Without D1-net or D2-net	127.12/90.02	1.77/1.67

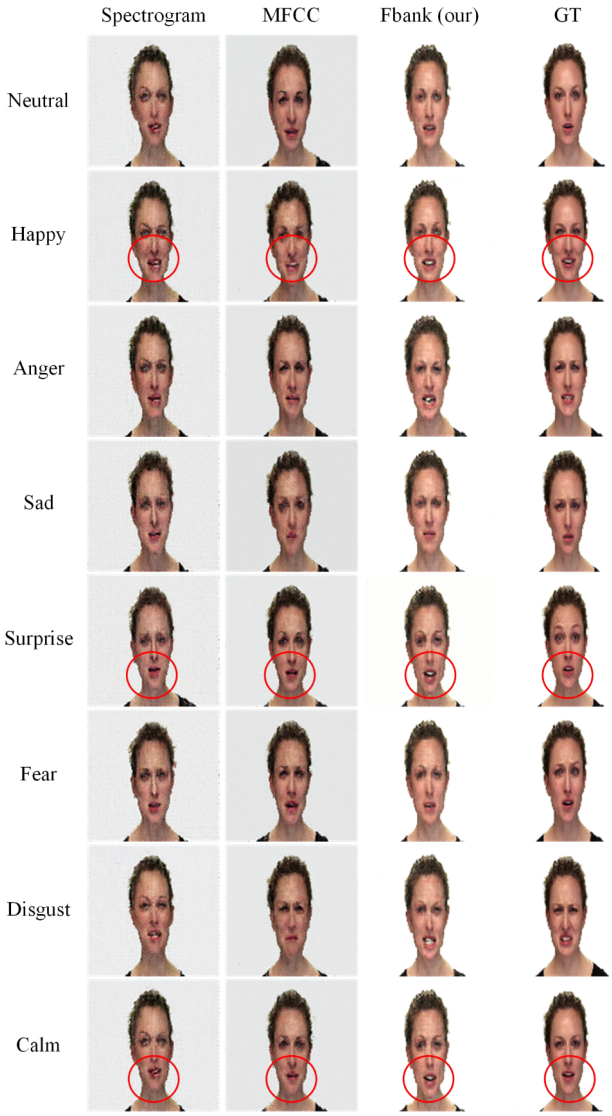

Fig. 3 Ablation experiments 1: generated images by different voice features that performed on the RAVDESS dataset. GT represents ground truth. The red circles depict the mouth regions under analysis for different expressions

4.4 Robustness tests

Robustness of FE-GAN is evaluated in this section. Two robustness experiments are conducted to verify how FID and IS scores vary when different input conditions are given: (1) the audio with different levels of noise; (2) the audio with different time durations.

Table 3 Robustness tests: FID and IS results of different duration times and noises on RAVDESS dataset

Experiments	Input items	FID Score	IS Score
Noise intensity (dB)	1	61.56	1.71
	5	68.08	1.73
	10	72.92	1.63
	25	93.57	1.43
Audio duration time (s)	1	72.70	1.67
	3	67.98	1.68
	5	64.17	1.72
	10	58.54	1.71

First of all, we study the effect between various levels of noise and quality of images. On RAVDESS dataset, we add different intensities of babble noises to voice with four Signal Noise Ratio (SNRs): 1 dB, 5 dB, 10 dB and 25 dB. The qualitative results for the experiments with added noise can be seen in Figs. 5 and 6. While the noise intensity is increased, we observe that the generated images are gradually to be blurry and unrecognizable. The reason may be that the useful features can be destroyed by noises as no identity and emotion information in noises. Moreover, the quantitative results of this experiment are reported in Table 3. We also observe that the FID and IS scores of various levels of noise gradually decrease, which are also consistent with the qualitative results.

The effect of different audio durations on FE-GAN is then evaluated. We conduct experiment on 1 s, 3 s, 5 s and 10 s voice segments. We observe that the audio duration has obvious effect on the quality of the reconstructions, as shown in Figs. 5 and 6. The qualitative results show that a longer duration of the input voice can improve the performance. For example, when using 10 s voice segments (the 4th column in Figs. 5 and 6), the generated faces are seen to be more clear, recognizable and less background noises. Furthermore, the corresponding quantitative results are also shown in Table 3. We find that feeding longer audio segments as input leads to considerable improvement in the FID and IS scores, that

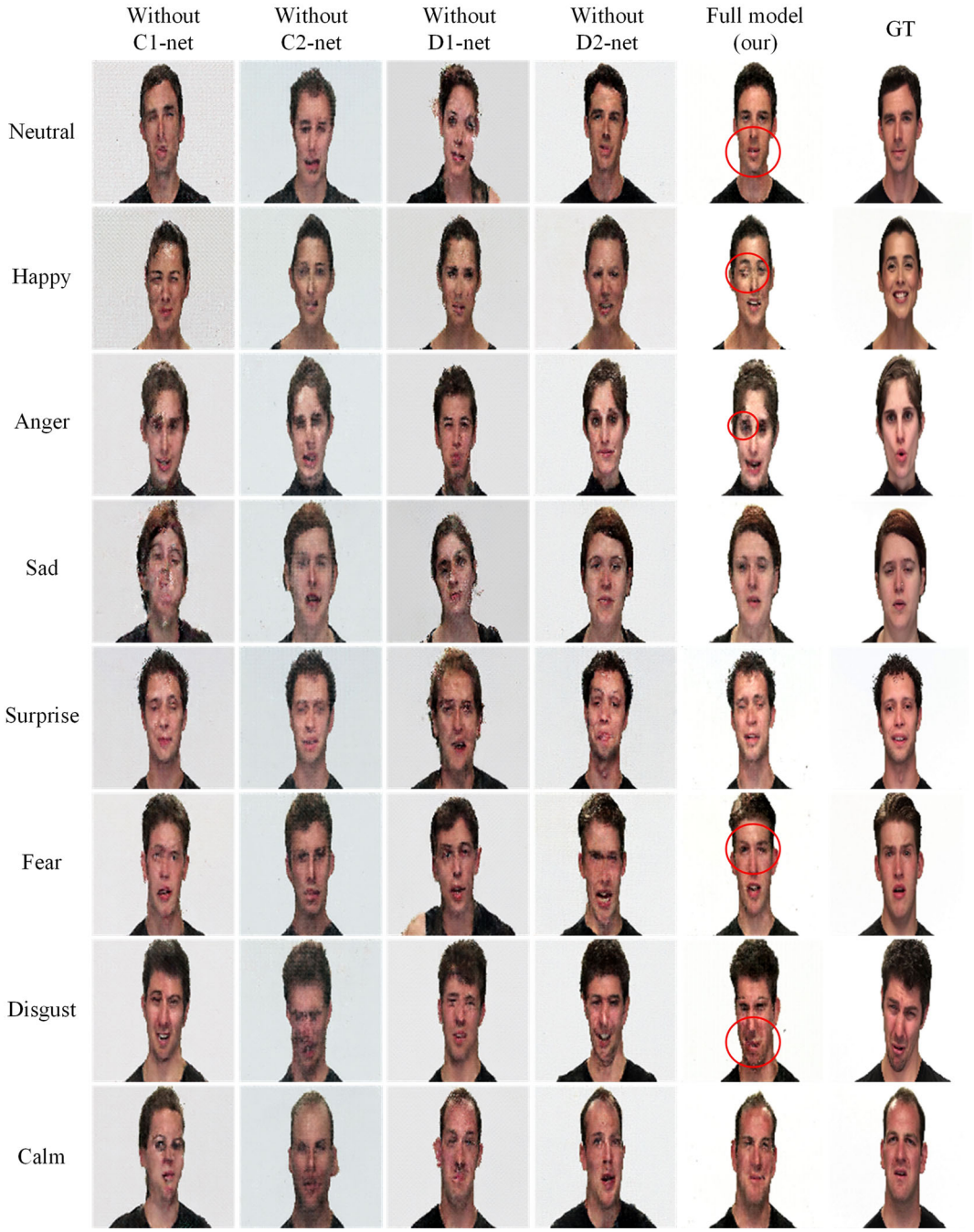

Fig. 4 Ablation experiments 2: generated images by four components that performed on the RAVDESS dataset. The red circle depicts the obscured and incorrect regions under analysis for different expressions

is, reconstructed faces capture the personal attributes and emotions better, regardless of which of the levels of noise are added.

Besides, Figs. 5 and 6 also show qualitative comparison of the effect of gender. We find that the model is able to successfully capture the latent attributes like gender, reconstructing the facial image with different voices.

4.5 Comparison to state-of-the-art

To verify the effectiveness of our FE-GAN model, we compare with two state-of-the-art methods on RAVDESS and eNTERFACE datasets, say AC-GAN [14] and CGANs [12]. Table 4 shows the comparison results of FID and IS, and Table 5 shows results of identity classification accuracy.

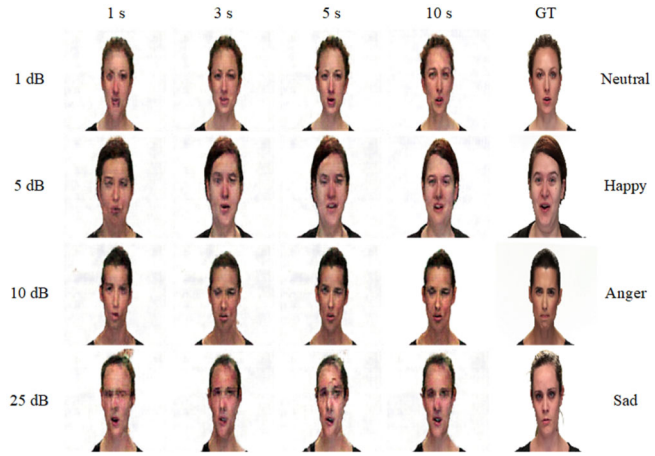

Fig. 5 Robustness tests with female speakers: generated images by voices with different noisy conditions and duration times that performed on the RAVDESS dataset. The left side represents the noise levels, and the right side represents the emotion types. Each row represents the generated faces using one of the four noisy conditions with different duration times

Fig. 6 Robustness tests with male speakers: generated images by voices with different noisy conditions and duration times that performed on the RAVDESS dataset

We first conduct comparison on the RAVDESS. Table 4 shows that FE-GAN performs better than AC-GAN, which can improve FID by 23.7% and IS by 2.8%. Compared with CGANs method, FE-GAN also improves FID by 48.1% and IS by 2.3%. As shown in Table 5, we make an improvement in training accuracy of identity by 9.7%, 15.1% compared with AC-GANs and CGANs, respectively. In the testing dataset, FE-GAN can achieve an increase of 9.8% and 18.0% compared with AC-GANs and CGANs. Besides, FE-GAN always achieves the high emotion accuracy of 95.08% in the training dataset. These quantitative results reveal that the sufficient utilization of both identity and emotion infor-

Table 4 FID and IS results of different methods on RAVDESS and eNTERFACE dataset

Methods	FID score		IS score	
	RAVDESS	eNTERFACE	RAVDESS	eNTERFACE
FE-GAN (our method)	58.79	84.58	1.71	1.89
AC-GANs [14]	77.04	94.26	1.76	1.71
CGANs [12]	113.31	129.28	1.75	1.78
Ground truth (GT)	–	–	1.71	1.94

Table 5 Identity classification accuracy of different methods on RAVDESS and eNTERFACE dataset

Methods	Training (%)		Testing (%)	
	RAVDESS	eNTERFACE	RAVDESS	eNTERFACE
FE-GAN (our method)	99.40	99.23	84.64	76.83
AC-GANs [14]	89.37	92.10	74.79	68.03
CGANs [12]	84.22	83.57	66.67	59.27

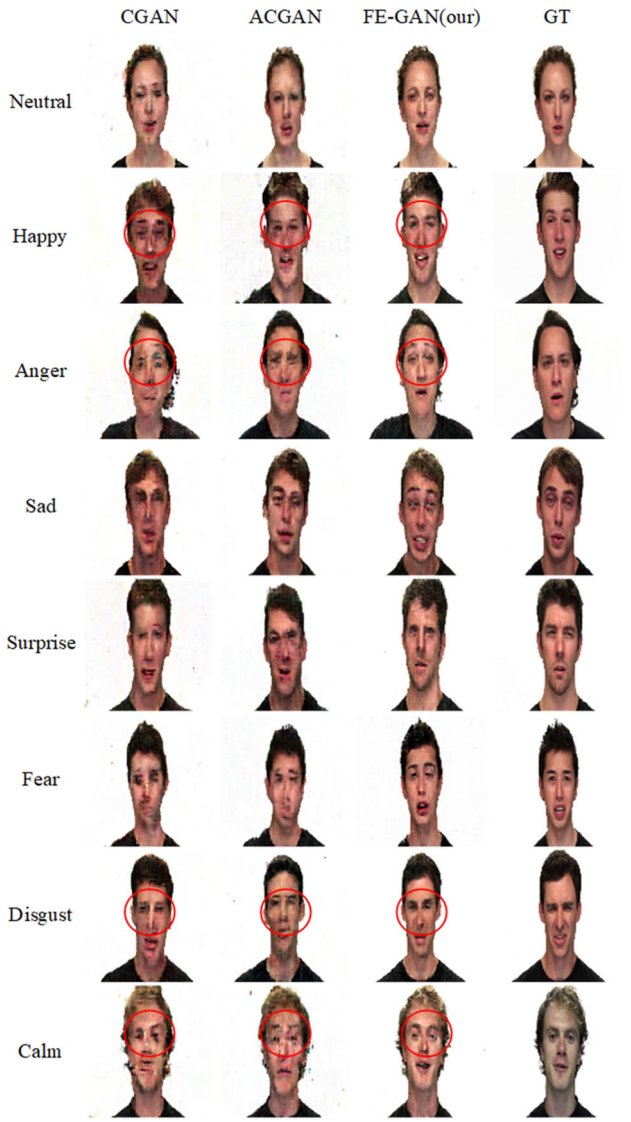

Fig. 7 Generated images of different methods on the RAVDESS dataset. The red circles depict the eye regions under analysis for different expressions

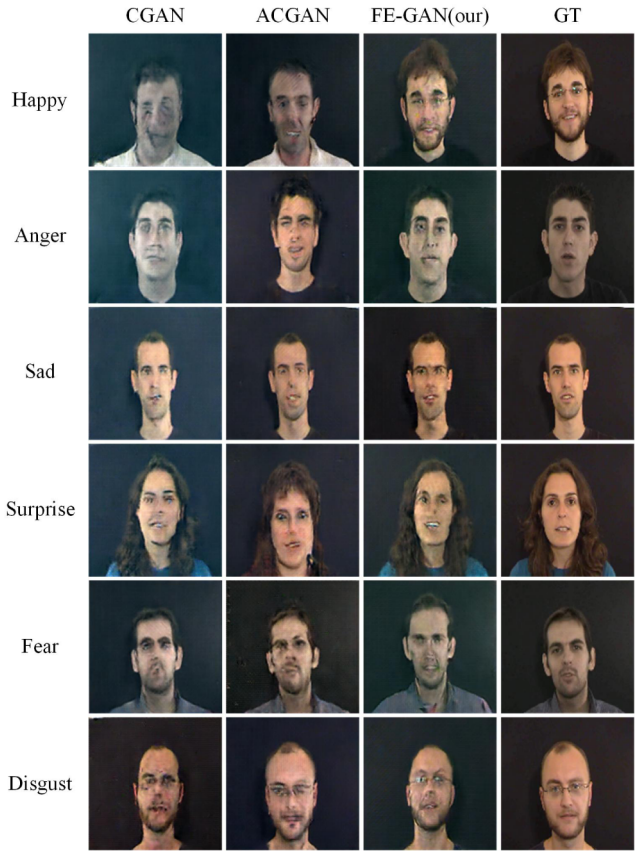

Fig. 8 Generated images of different methods on the eNTERFACE dataset

mation from voice can significantly boost the performance of classification task. Furthermore, the qualitative results of RAVDESS are as shown in Fig. 7, which also based on our FE-GAN and two competitors. It can be seen that FE-GAN can not only generate the faces with more exactly identity information, but also maintain the more expression information. Our results have less noise in background and are more realistic than without emotion samples.

To further verify the robustness of our method for voice-to-face generation, we evaluate our method on another dataset eNTERFACE and also give the comparison results in Tables 4 and 5. In Table 4, we observe that our method achieves the highest IS (1.89) and the lowest FID (84.58), demonstrating the effectiveness and robustness of our method. As shown in Table 5, FE-GAN also achieves the highest identity accuracy in training (99.23%) and testing (76.83%). FE-GAN outperforms comparison methods in eNTERFACE. However, there are still some defects in the generated images. Figure 8 shows that the faces are blurs and artifacts, even corrupted facial expressions around eyes, nose and mouth regions. Besides, our method also gets the low emotion accuracy of 73.15% on training images. This is because of the imbalanced data distribution and large variance between same class of training data. A low-quality and bad-controlled dataset may cause unstable generation results. Although the images in Fig. 8 are not sharp, we can still see that the identity of the generated images is semantically consistent with the input audio information, which means our method has captured the semantic attribution from speech features to some extent.

4.6 Limitation of FE-GAN

During our experiments, we found there are some generated images which have observable failures as shown in Figs. 3, 4 and 7. The major problems include moderate artifacts (e.g., the texture and color of face seem unnatural), loss of facial contours and details (e.g., tooth, hair and eyebrow region

are obscure or missing), and minor semantic inconsistency (e.g., compared with GT images). There are the two main reasons for these problems: (1) The intra-personal and inter-personal variances of emotion are large in datasets, which make FE-GAN hard to learn these face and voice emotion features effectively. (2) The input embedding features are only from the single modality (voice) instead of multiple modalities (voice and face). That is, a part of facial attributes is irrelevant to speakers' voices so that the generator cannot build these mapping between voice and face. Therefore, it is unable to generate high-quality tooth, hair, eyebrow and head pose by only using single modality features.

5 Conclusion

Facial expression plays an important role in high-quality face generation. Human perception is very sensitive to subtle facial expression. Therefore, without taking emotion about this face and voice into account, it is hard to generate shaper and proper face images. In this paper, we propose a novel FE-GAN to consider the emotion in voice-to-face generation problem. Specifically, audio emotion and identity are used to directly generate face images with expressions. We proposed FE-GAN which includes one generator and two discriminators with their auxiliary classifiers. The core idea is to use auxiliary classifiers to help discriminators to better identify whether a face image is generated or true based on the identity and emotion represented in this image. Therefore, the generator can be trained to generate more realistic face images. Finally, the proposed triple loss facilitates the generalization and optimization ability of the model. Experimental results show that our proposed method outperforms the state-of-the-art approaches in both quantitative and qualitative perspectives.

FE-GAN has its own limitation. Firstly, the output based on single generator has model collapse and over-fitting problems. For example, some facial identity features and emotional features cover up each other, resulting in a lot of ambiguous and pixel jittering in images, and some emotion samples are insufficient, which can affect the generation of face images. On the other hand, the model is hard to achieve the best balance between the two discriminators in training. In addition, the intensity of the expressions should be considered to further improve the quality of generated images.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant 61761166005), Ministry of Science and Technology, Taiwan (MOST 106-2218-E-032-003-MY3), and National Natural Science Foundation of Zhejiang (Grant LY20F020007), and the Ningbo Science Technology Plan projects (Grant 2019B10032) and the K.C. Wong Magna Fund in Ningbo University.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Sriram, A., Jun, H., Gaur, Y., Satheesh, S.: Robust speech recognition using generative adversarial networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5639–5643 (2018)
2. Dumpala, S.H., Sheikh, I., Chakraborty, R., Kopparapu, S.K.: A Cycle-GAN approach to model natural perturbations in speech for ASR applications. arXiv preprint arXiv:1912.11151 (2019)
3. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional gan. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2970–2979 (2017)
4. Chen, C., Mu, S., Xiao, W., Ye, Z., Wu, L., Ju, Q.: Improving image captioning with conditional generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8142–8150 (2019)
5. Goodfellow, I., Pougetabadie, J., Mirza, M., Xu, B., Wardefarley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
6. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
7. Han, F., Guerrero, R., Pavlovic, V.: CookGAN: meal image synthesis from ingredients. Computer Vision and Pattern Recognition. arXiv (2020)
8. Nasir, O.R., Jha, S.K., Grover, M.S., Yu, Y., Kumar, A., Shah, R.R.: Text2FaceGAN: face generation from fine grained textual descriptions. In: IEEE International Conference on Multimedia Big Data, pp. 58–67 (2019)
9. Qiu, Y., Kataoka, H.: Image generation associated with music data. In: Computer Vision and Pattern Recognition (CVPR), pp. 2510–2513 (2018)
10. Wan, C., Chuang, S., Lee, H.: Towards audio to scene image synthesis using generative adversarial network. In: International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 496–500 (2019)
11. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976 (2017)
12. Duarte, A., Roldan, F., Tubau, M., Escur, J., Pascual, S., Salvador, A., Mohedano, E., Mcguinness, K., Torres, J., Giroinieto, X.: Wav2Pix: speech-conditioned face generation using generative adversarial networks. In: International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 8633–8637 (2019)
13. Oh, T., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W.: Speech2Face: learning the face behind a voice. In: Computer Vision and Pattern Recognition (CVPR), pp. 7539–7548 (2019)
14. Wen, Y., Singh, R., Raj, B.: Face reconstruction from voice using generative adversarial networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 5265–5274 (2019)
15. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: International Conference on Machine Learning, pp. 2642–2651 (2017)

16. Smith, H.M.J., Dunn, A.K., Baguley, T., Stacey, P.C.: Matching novel face and voice identity using static and dynamic facial images. Atten. Percept. Psychophys. **78**(3), 868–879 (2016)
17. Nagrani, A., Albanie, S., Zisserman, A.: Seeing voices and hearing faces: cross-modal biometric matching. In: Computer Vision and Pattern Recognition (CVPR), pp. 8427–8436 (2018)
18. Livingstone, S.R., Russo, F.A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE **13**(5), 1–35 (2018)
19. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06), pp. 8–8. IEEE Computer Society (2006)
20. Nguyen, T.D., Le, T., Vu, H., Phung, D.: Dual discriminator generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), pp. 2670–2680 (2017)
21. Durugkar, I., Gemp, I., Mahadevan, S.: Generative multi-adversarial networks. In: International Conference on Learning Representations (2017)
22. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Computer Vision and Pattern Recognition (CVPR), pp. 7832–7841 (2019)
23. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? In: British Machine Vision Conference (BMVC) (2017)
24. Vougioukas, K., Petridis, S., Pantic, M.: End-to-end speech-driven facial animation with temporal GANs. In: British Machine Vision Conference (BMVC) (2018)
25. Konstantinos, V., Stavros, P., Maja, P.: Realistic speech-driven facial animation with GANs. Int. J. Comput. Vis. **8**(5), 1398–1413 (2020)
26. Watanabe, S., Kim, S., Hershey, J.R., Hori, T.: Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE J. Sel. Top. Signal Process. **11**(8), 1240–1253 (2017)
27. Chandrasekar, P., Chapaneri, S., Jayaswal, D.: Automatic speech emotion recognition: a survey. In: International Conference on Circuits, pp. 341–346 (2014)
28. Passricha, V., Aggarwal, R.K.: A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. J. Intell. Syst. **29**(1), 1261–1274 (2019)
29. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
30. Aldeneh, Z., Provost, E.M.: Using regional saliency for speech emotion recognition. In: International Conference on Acoustics, Speech and Signal Processing, pp. 2741–2745 (2017)
31. Chenchah, F., Lachiri, Z.: Acoustic emotion recognition using linear and nonlinear cepstral coefficients. Int. J. Adv. Comput. Sci. Appl. **6**(11), 1–4 (2015)
32. Waghmare, V.B., Deshmukh, R.R., Shrishrimal, P.P., Janvale, G.B., Ambedkar, B.B.: Emotion recognition system from artificial Marathi speech using MFCC and LDA techniques. In: International Conference on Advances in Communication, Network, and Computing (2014)
33. Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., Schuller, B.: Speech emotion classification using attention-based LSTM. IEEE Trans. Audio Speech Lang. Process. **27**(11), 1675–1685 (2019)
34. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. In: the Proceedings of the 22nd ACM international conference on Multimedia, pp. 801–804
35. Yi, R., Ye, Z., Zhang, J., Bao, H., Liu, Y.: Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137 (2020)
36. Suwajanakorn, S., Seitz, S.M., Kemelmachershlizerman, I.: Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph. (TOG) **36**(4), 1–13 (2017)
37. Jalalifar, S.A., Hasani, H., Aghajan, H.: Speech-driven facial reenactment using conditional generative adversarial networks. arXiv preprint arXiv:1803.07461 (2018)
38. Sadoughi, N., Busso, C.: Speech-driven expressive talking lips with conditional sequential generative adversarial networks. IEEE Trans. Affect. Comput. (2019). https://doi.org/10.1109/TAFFC.2019.2916031
39. Duan, B., Wang, W., Tang, H., Latapie, H., Yan, Y.: Cascade attention guided residue learning GAN for cross-modal translation. arXiv preprint arXiv:1907.01826 (2019)
40. Van Segbroeck, M., Tsiartas, A., Narayanan, S.S.: A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice. In: Conference of the International Speech Communication Association, pp. 704–708 (2013)
41. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
42. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein Gans. In: Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
43. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations (2016)
44. Gan, Z., Chen, L., Wang, W., Pu, Y., Zhang, Y., Liu, H., Li, C., Carin, L.: Triangle generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 5247–5256 (2017)
45. Li, C., Xu, K., Zhu, J., Zhang, B.: Triple generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 4088–4098 (2017)
46. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Neural Information Processing Systems, pp. 2234–2242 (2016)
47. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Neural Information Processing Systems, pp. 6626–6637 (2017)
48. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
49. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)

Zheng Fang received his masters degree in 2018 from Wenzhou University, China. Now, he is pursuing the PhD degree in Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His research interest is affective computing.

Zhen Liu is a professor in Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His main research interests include virtual reality and artificial intelligence.

Tingting Liu is an associate professor in College of Science and Technology, Ningbo University, China. Her research interests include virtual reality and artificial intelligence.

Chih-Chieh Hung is an assistant professor in the Department of Management Information System, National Chung Hsing University (NCHU), Taiwan. His research interests include data mining, pervasive and mobile computing, big data systems, e-commerce intelligence, and artificial intelligence.

Jiangjian Xiao is a professor at Ningbo Industrial Technology Research Institute, CAS. His research interests include computer vision, computer graphics, and visualization. He is a member of ACM and IEEE. He also is an associate editor of Machine Vision and Application Journal.

Guangjing Feng received his bachelors degree from Ningbo Institute of Technology, Zhejiang University, China, in 2018. Now he is pursuing the masters degree in Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His research interest is human action recognition.