Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048703

ARTICLE

Attention-Enhanced Voice Portrait Model Using Generative Adversarial Network

Jingyi Mao, Yuchen Zhou, Yifan Wang, Junyu Li, Ziqing Liu and Fanliang Bu*

School of Information Network Security, People's Public Security University of China, Beijing, 100038, China

*Corresponding Author: Fanliang Bu. Email: bufanliang@sina.com

Received: 15 December 2023 Accepted: 28 February 2024 Published: 25 April 2024

ABSTRACT

Voice portrait technology has explored and established the relationship between speakers' voices and their facial features, aiming to generate corresponding facial characteristics by providing the voice of an unknown speaker. Due to its powerful advantages in image generation, Generative Adversarial Networks (GANs) have now been widely applied across various fields. The existing Voice2Face methods for voice portraits are primarily based on GANs trained on voice-face paired datasets. However, voice portrait models solely constructed on GANs face limitations in image generation quality and struggle to maintain facial similarity. Additionally, the training process is relatively unstable, thereby affecting the overall generative performance of the model. To overcome the above challenges, we propose a novel deep Generative Adversarial Network model for audio-visual synthesis, named AVP-GAN (Attention-enhanced Voice Portrait Model using Generative Adversarial Network). This model is based on a convolutional attention mechanism and is capable of generating corresponding facial images from the voice of an unknown speaker. Firstly, to address the issue of training instability, we integrate convolutional neural networks with deep GANs. In the network architecture, we apply spectral normalization to constrain the variation of the discriminator, preventing issues such as mode collapse. Secondly, to enhance the model's ability to extract relevant features between the two modalities, we propose a voice portrait model based on convolutional attention. This model learns the mapping relationship between voice and facial features in a common space from both channel and spatial dimensions independently. Thirdly, to enhance the quality of generated faces, we have incorporated a degradation removal module and utilized pretrained facial GANs as facial priors to repair and enhance the clarity of the generated facial images. Experimental results demonstrate that our AVP-GAN achieved a cosine similarity of 0.511, outperforming the performance of our comparison model, and effectively achieved the generation of high-quality facial images corresponding to a speaker's voice.

KEYWORDS

Cross-modal generation; GANs; voice portrait technology; face synthesis

1 Introduction

Voice portrait technology aims to analyze the voice of a speaker through cross-modal intelligent techniques, seeking the mapping relationship between facial features and voice, and generating facial images that correspond to the respective identity. This technology involves biology, speech analysis, image generation, cross-modal and other multi-disciplinary fusion technology, which has a wide range of applications in the fields of public security, judicial identification, medical diagnosis and personalized entertainment. For example, for the actual combat needs of public security, typical network fraud shows a trend of frequent and high incidence around the world, which has become a worldwide crime problem. When the police obtain the audio samples of the criminal suspect, they can use voice portrait technology to generate facial images similar to the identity characteristics of the criminal suspect and provide important clues. On the other hand, the technology can also be used to generate the personalized digital image of the speaker from voice, enhance the sense of experience of human-computer interaction, and provide new possibilities for personalized entertainment and other fields. In this paper, we analyze the cross-modal correlation between the speakers' speech and their facial features. We aim to reconstruct an image with the facial characteristics of the speaker by inputting a short segment of the speaker's voice.

There is a close relationship between human voice and face. From a person's voice, we can extract many biometric features such as age, gender, emotion, race, lip movement, weight, skin color, facial shape, etc. [1,2]. Recently, some research work based on deep learning is exploring the relationship between voice and faces. In the research of speech and face matching, Wen et al. [3] have achieved voice-face cross-modal matching by mapping voice and face modes to common covariates and learning their common embedding. With the in-depth study of the relationship between speech and face, Kameoka et al. [4] proposed a method based on auxiliary classifier, which makes use of the correlation between speech and face to generate a face image that matches the input speech. In addition, Duarte et al. [5] proposed a Conditional Generative Adversarial Network (CGAN) model to generate face pixels directly from speech. Although the above work can obtain reasonable face portrait effect, there is still a big challenge in finding the corresponding feature mapping between voice and face. Therefore, we discuss how to combine the attention mechanism with the generative and discriminative modules in the network model to further analyze the key pixel blocks of voice and face to improve the effect of the model.

Generative Adversarial Networks (GANs) [6] have advantages in effectively generating the desired samples and eliminating deterministic biases [7]. Based on these characteristics, GANs has become one of the mainstream methods in the field of image generation. In this paper, we propose a new attention-enhanced voice portrait model using Generative Adversarial Network (AVP-GAN). This model only needs to use voice as input and does not need any a priori knowledge to generate human face. Specifically, we first input a segment of the speaker's voice into the voice encoder to extract the Mel-frequency cepstral coefficients (MFCCs) [8], which includes rich facial attribute information. Subsequently, we input the representation containing rich identity information into the generative model. To achieve a better match between the face images generated through voice and real face images in more facial details, we combine the Convolutional Block Attention Module (CBAM) [9] with the Deep Convolutional Generative Adversarial Network (DCGAN). This combination involves adaptive scaling in both spatial and channel dimensions. This enables the discriminator to better assist the generator in creating more distinct and discriminative regions. Finally, to enhance the quality of the generated face images, we employ a pre-trained model to perform restoration on the generated face images. Overall, our main contributions can be summarized as follows:

- We have designed a novel attention-enhanced voice portrait model using generative adversarial network (AVP-GAN). Specifically, our AVP-GAN model consists of three modules: The voice encoder module, the face generation module, and the image enhancement module. This model is capable of generating facial images that align with the identity features of a speaker based on their voice.

- We combine the feature extraction ability of the convolutional network with the generative network, and apply spectral normalization in the network structure. This ensures that the discriminator satisfies Lipschitz continuity by constraining the degree of sharp changes in the function, and helps prevent issues such as mode collapse during training, making the model more stable.

- We combine the CBAM with the GANs and apply it to the voice portrait task. By incorporating adaptive scaling in both spatial and channel dimensions, the network is capable of capturing more facial feature details, thereby enhancing the face generation performance of the voice portrait model.

- To address the issue of low fidelity in generated images, we have introduced a degradation removal module and utilized pre-trained facial GANs as facial priors for facial restoration. This approach aims to enhance the clarity of the generated facial images. Experimental results demonstrate that our voice portrait model exhibits favorable face generation performance.

The rest of this paper is organized as follows: Section 2 briefly introduces related work on cross-modal generation techniques for voice portrait. Section 3 describes the details of our proposed AVP-GAN model. Section 4 presents the qualitative and quantitative evaluation of our model's performance. Section 5 concludes the full paper.

2 Related Works

2.1 Voice Representation Learning

Voice portrait technology aims to explore the relationship between human voice and face. Voice is produced in the human vocal tract, which includes channels and cavities connecting the lungs to the external environment. In the process of voice production, the movement of the tongue, lips, chin and so on will change the shape of the sound channel, resulting in a resonance chamber with different shape contraction separation. The air from the lungs is converted into a series of periodic pulses by the vocal cords and resonates within these resonance chambers, producing an auditory voice characterized by spectral peaks and troughs. At present, a large number of studies have shown that voice signals have direct or indirect influence on many human biometric parameters. For example, the increase of age [10,11] will affect the harmonic structure of the voice signal, and the body size of the speaker can be inferred by analyzing the pronunciation method of the voice [12]. In the field of learning voice representations, a method known as Speech Enhancement GANs (SEGAN) [13] works end-to-end with original audio, capable of learning from various speakers and noise types, and incorporating them into the network's parametrization. In order to adapt to different data distribution, Hong et al. [14] explored the method of Audio Albert to self-supervise the speech representation in the original speech and compress the speech waveform into a vector containing high-level semantic information from the original speech. Yi et al. [15] proposed a self-supervised pre-training architecture to encode speech through a multi-layer convolutional neural network. MFCCs is proposed based on the auditory characteristics of human ear, and shows a good application prospect in speaker verification task and voice recognition.

2.2 Research on the Correlation between Voice and Facial Features

In recent years, the task of finding the relationship between voice and facial features is becoming a hot research topic. In 1916 [16], Swift indirectly hinted at the connection between speech and human body parameters in an article. It is clearly pointed out in the article that voice quality is solely determined by bone structure, an inherited characteristic, leading to the conclusion that sound quality is also a hereditary attribute. There is also a great correlation between the shape of the lips and the speaker's pronunciation habits. In fact, the shape of the vocal tract can be determined according to the pulse reflection of the lips [2]. For example, British people generally have thinner lips, which is attributed to the reduced use of retroflex consonants in the pronunciation of British English. There is less variation in lip movements, and the mouth typically tilts backward, appearing flatter. Additionally, the upper lip often exhibits a converging motion. In contrast, for Korean pronunciation, to distinguish between different vowels, Koreans need to mobilize the muscles in the oral cavity more frequently. This leads to the appearance of many Koreans having a puckered mouth shape.

As early as in 2006, a study by Rosenblum et al. [17] used point-light technology experiments to show that visible voice motion is capable of supporting cross-modal speaker matching, suggesting that it is possible for humans to utilize the relationship between faces and voices for speaker identification. With the continuous development of machine learning, Nagrani et al. [18] proposed a machine learning algorithm for speech and face association learning, which introduces the task of cross-mode matching between face and voice, and uses Convolutional Neural Networks (CNNs) architecture to solve this problem. Kameoka et al. [4] combined StarGAN and Conditional Variational Autoencoder (CVAE) to construct a cross-modal voice conversion (VC) model, which is used to generate facial images that match the speech features of the input speech. Wen et al. [19] proposed a new framework of adaptive identity weights for voice-face association, which obtained better results on the retrieval task of voice portrait. Although there is a large semantic gap between voice and human face, the above studies have confirmed that there is a great correlation between voice and face. In our voice portrait method, we have chosen to process the voice signal using the log-mel spectrum, which is based on the characteristics of human auditory perception, in the hope of better extracting the features of the audio signal.

2.3 Cross-Modal Face Generation from Voice

Voice and image are two direct and important ways of human communication. In the field of image processing, image enhancement and segmentation are very important for processing and analyzing images. Wang et al. [20] used GANs to propose a blind face restoration model, achieving effective facial image enhancement. Linguo et al. [21] proposed a method called Fuzzy Multilevel Image Thresholding Based on Improved Coyote Optimization Algorithm, which showed strong advantages in image segmentation. In the field of voice-to-face generation, human hearing and vision are the main basis for the brain to obtain external information, respond, judge and make decisions. There is a close relationship between them, and both can provide biological attribute monitoring information data for the other party. Speech2Face [22] used the videos of millions of speakers on the internet to map speech sonogram features to real face features in high-dimensional space, thus training a depth neural network model that can reconstruct the speaker's facial image through speech. However, because a single convolution layer is generally unable to capture long-distance features, the traditional CNNs may lead to poor image quality. In order to meet this challenge, Wang et al. [23] proposed a new residual speech portrait model based on attention mechanism, thus improving the effect of the model. Duarte et al. proposed a deep neural network Wav2Pix [5] for cross-modal visual synthesis. The model does not need any prior knowledge and is trained from the beginning in an end-to-end manner. The GANs is used to synthesize the speaker's face image under a given original speech waveform. At the same time, Bragin et al. [24] designed an automatic encoder neural network including speech encoder and facial decoder to realize the task of reconstructing human face from speech. In addition to the two-dimensional static voice portrait task, Wu et al. [25] designed a cross-modal three-dimensional facial generation model using 3D Morphable Models. This model achieves a rough reconstruction of a 3D face based on sound.

2.4 GAN in Face Generation

GANs is a powerful tool for unsupervised image generation, which has been widely studied in the field of computer vision, and its core idea comes from zero-sum game in game theory. The traditional GANs consists of generator network (Generator) and discriminant network (Discriminator). Its goal is to train the discriminator's ability to distinguish between real samples and false samples through against training, so that the generator can generate realistic data samples that can deceive the discriminator. The loss function is as follows:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data(x)}}\left[logD\left(x\right)\right] + E_{z\sim p_{z(z)}}\left[\log(1 - D\left(G(z)\right))\right] \tag{1}$$

where $p_{data(x)}$ represents the distribution of real data, p_z represents the distribution of original noise, $G(z)$ represents the generated mapping function, and $D\left(x\right)$ represents the discriminant mapping function.

Since being proposed by Ian Goodfellow and others, GANs have gained widespread attention. GANs-based models are widely used in a variety of tasks, including handwritten number generation, style transfer [26], face image generation [27] and so on. Aiming at the application of GANs in the face image generation task, Yeh et al. [28] filled and repaired the damaged face image, and achieved good results. In 2017, Karras et al. [29] used GANs to generate new facial images with celebrity facial features by using celebrity faces as input. Huang et al. [30] used GANs to generate frontal face images with different angles, which can be used in face verification system. However, the current application of GANs in the field of voice portraits is limited, and their training exhibits instability. Moreover, existing methods still face challenges in generating high-quality facial images. Therefore, we employ a GANs based on attention mechanisms, applied to the task of voice portrait, to generate facial images that align with the different speech features.

Overall, we have summarized and analyzed the current methods for voice portraits, as outlined in Table 1.

Table 1: Current methods for voice portraits

Voice portrait	Methods	Strengths	Weaknesses
Speech2Face	CNN + DNN	1. Modeling attributes explicitly is not necessary. 2. Unable to maintain facial similarity.	1. Limited image quality. 2. Static facial images can be generated.
AR-SPM	CNN + Attention mechanism	Better accuracy in gender and age recognition.	Limited image quality.
Wav2Pix	CGAN	1. Established a new dataset. 2. Generates faces from the voice without any conditions.	1. Insufficient research on cross-modal correlation. 2. The generated image quality is limited.
Voice2Face	GAN	It can achieve the generation of facial images from voice in an unsupervised learning manner.	1. Limited image quality. 2. Model training exhibits instability.
Cross-modal perceptionist	GAN+3D Morpable models	It can predict the 3D geometric shape of a face based on voice.	Limited similarity.
SF2F	CNN + GAN	1. Established a high-quality face database. 2. Improved facial similarity through enhanced training.	1. Limited discussion on voice embedding methods. 2. Limited image quality.

3 Method

Our method aims to solve a challenge related to voice-face cross-modal tasks, which involves drawing clear facial features of a person based on their voice features. CNNs exhibit strong performance in feature extraction, and GANs include a generator (D) and a discriminator (G), which play games with each other and train to find and achieve Nash equilibrium by training to achieve convergence, and are widely used in image generation and image transformation tasks. We combined CNNs with GANs, and used the structure of CNNs and added Spectral Normalization [31] in both the generator network and discriminator network of GANs to improve the training stability of the model and the quality of the generated results. CBAM extracts more details by adaptively scaling the spatial and channel features. In this research, we adopted a combined approach of using the CBAM with a deep convolutional GANs. Additionally, we incorporated a degradation removal module [20] and a pre-trained image enhancement model serving as a facial prior. This combination aims to achieve higher-quality facial image generation from speaker's voice, resulting in improved model performance. In this section, we will provide a detailed introduction to our newly proposed AVP-GAN model.

3.1 Symbolic Conventions

First of all, we have stipulated some symbolic representations. In this paper, we use a set of voice-face pairs as the training data set, where we use the symbol $P = \{P_1, P_2, ..., P_m\}$ to represent the face image data and the symbol $V = \{V_1, V_2, ..., V_n\}$ to represent the voice data. Here, m and n denote the number of face and voice samples in the training set, respectively. Moreover, we use the symbol Y to represent the identity of the speakers in the training set, $Y = \{Y_1, Y_2, ..., Y_T\}$, and T represents the total number of different speakers in the training set, where m, n, and T can be unequal from each other. The identity label of the speaker corresponding to the face is denoted as $\mathrm{Y}^{\mathrm{P}} = \{Y_1^P, Y_2^P, ..., Y_T^P\}$, and $Y_i^P \in \mathrm{Y}$, similarly, the identity label of the speaker corresponding to the speech data is denoted as $\mathrm{Y}^{\mathrm{V}} = \{Y_1^V, Y_2^V, ..., Y_T^V\}$ and $Y_i^V \in \mathrm{Y}$. We define the relevant symbols as shown in Table 2.

Table 2: Related symbols and their meanings in this paper

Serial number	Symbol	Notation
1	m	Number of face images in the training set.
2	P_m	Face Image Label.
3	n	Number of speech data in the training set.
4	V_n	Voice Data Label.
5	T	Number of different speakers trained to concentrate.
6	Y_T	Speaker Identity Label.
7	Y_i^P	Real speaker identity labels corresponding to face data.
8	Y_i^T	Real speaker identity labels corresponding to voice data.
9	$F \in R^{C*H*W}$	Input characteristics in CBAM.
10	$M_C \in R^{C*1*1}$	One-dimensional convolution of channel attention modules.
11	$M_S \in R^{1*H*W}$	Two-dimensional convolution of spatial attention modules.
12	F'	Output results in the process of CAM.
13	F''	Output results in the process of SAM.
14	$\otimes$	Hadamard product.
15	σ	Sigmoid function.
16	MLP	Multilayer Perception.

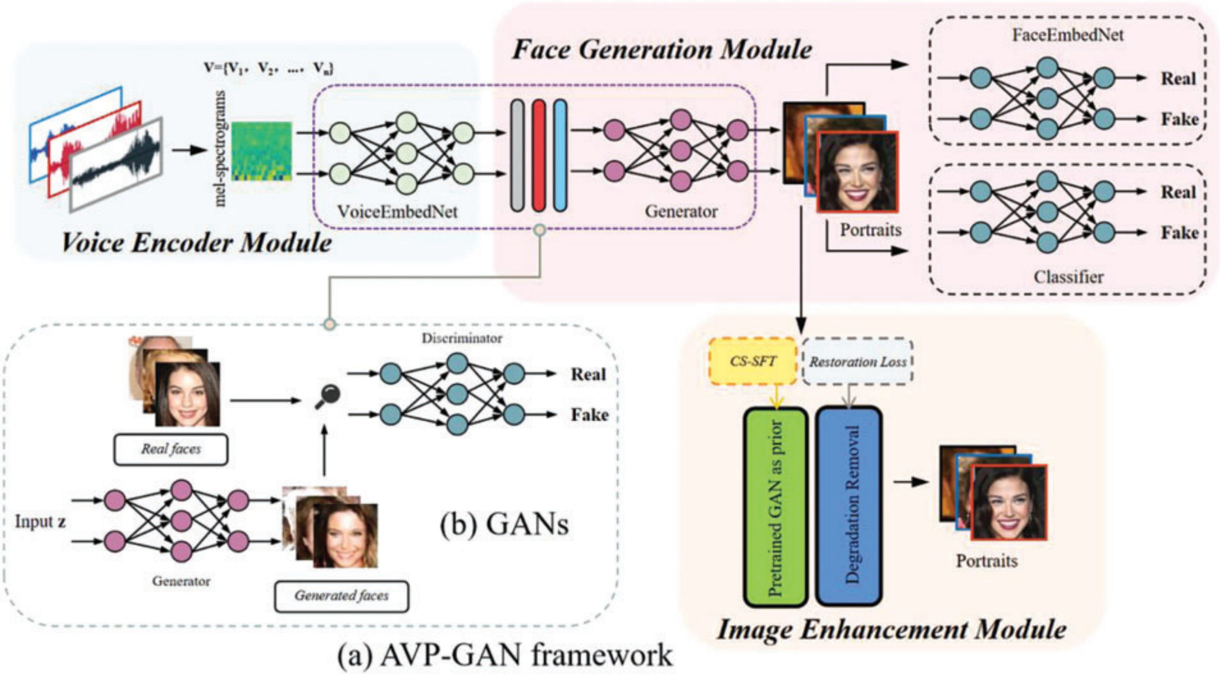
(Continued)

Table 2 (continued)

Serial number	Symbol	Notation
17	$W_0 \in R^{C/r*C}$ $W_1 \in R^{C*C/r}$	MLP weights, are shared for both inputs and the ReLU activation function is followed by W_0.
18	f^{7*7}	Represents a convolution operation with the flitter size of 7∗7.
19	$IC(x)$	Identity card of $\boldsymbol{x}$.
20	L_D, L_C, L_G	Denote the loss functions of the discriminator, classifier and generator, respectively.

3.2 Overall

Fig. 1 shows the framework of our proposed AVP-GAN. Due to the inherent correlation between the way a person voices and their facial features, our objective is to generate a face that corresponds to the speaker's voice. To achieve this goal, we propose a new AVP-GAN model. Specifically, our model consists of three modules: The voice encoder module, the face generation module, and the image enhancement module. We integrate the CBAM with the GANs to focus on extracting correlational features between the voice and facial vectors in a shared mapping space. Additionally, we introduce a degradation removal module in our AVP-GAN model to enhance the clarity of generated images. This effectively produces higher-quality facial images that align with the characteristics of the given voice. More details to be described in the ensuing subsections.

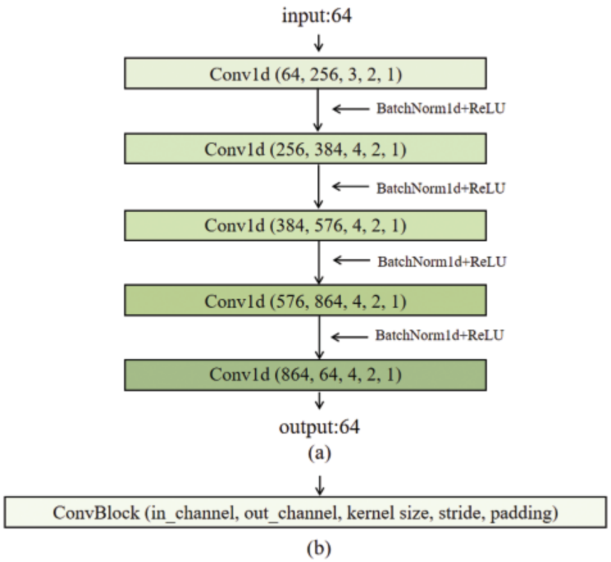


(a) AVP-GAN framework

Figure 1: (a) Overall structure of AVP-GAN. Specifically, our AVP-GAN model consists of three modules: The voice encoder module, the face generation module, and the image enhancement module. (b) Traditional GANs framework

In summary, first, input the speaker's voice into the Voice Encoder for voice encoding. Secondly, the CBAM and the GANs are combined to generate facial images corresponding to the identity of the speaker through training. Thirdly, the Channel-Split Spatial Feature Transform (CS-SFT) is used to balance the fidelity of the images, and the image enhancement model based on facial priors is applied to obtain higher quality images. In our AVP-GAN model, there are two discriminators: One judges whether the generated image is a facial image, and the other assesses the similarity between the generated facial image and the voice features.

For the voice coder module, we normalize the extracted 64-dimensional logarithmic Mel spectrogram, clip the voice segment to around 8 seconds, and input it into the voice coder network for feature extraction. The voice coder network is a one-dimensional convolutional neural network designed to extract features from voice for the purpose of creating a voice profile model. The structure of the voice encoder is illustrated in Fig. 2, where *k, s*, and *p* represent kernel size, stride, and padding, respectively.


input:64
Conv1d (64, 256, 3, 2, 1)
BatchNorm1d+ReLU
Conv1d (256, 384, 4, 2, 1)
BatchNorm1d+ReLU
Conv1d (384, 576, 4, 2, 1)
BatchNorm1d+ReLU
Conv1d (576, 864, 4, 2, 1)
BatchNorm1d+ReLU
Conv1d (864, 64, 4, 2, 1)
output:64
ConvBlock (in_channel, out_channel, kernel size, stride, padding)
(a)

(b)

Figure 2: (a) Structure diagram of the voice encoder network; (b) Internal structure of the ConvBlock (the internal structure of all network structure diagrams in this article is consistent with the one shown above)

3.3 Generation Module and Image Enhancement Module

CBAM is a lightweight convolutional attention module that can perform adaptive rescaling of spatial and channel features, and combining CBAM with GANs can enhance the discrimination of salient regions and extract richer detailed features [32]. The CBAM consists of two processes: One is to compress the spatial dimension while keeping the channel dimension unchanged, i.e., the Channel Attention Module (CAM); and the other is to compress the channel dimension while keeping the spatial dimension unchanged, i.e., the Spatial Attention Module (SAM). Fig. 3 illustrates the entire process of CBAM. The CBAM formula is shown below, where the input feature $F \in \mathrm{R}^{C*H*W}$, $M_C \in \mathrm{R}^{C*1*1}$ is the one-dimensional convolution of the channel attention module and $M_S \in \mathrm{R}^{1*H*W}$ is the two-dimensional convolution of the spatial attention module.

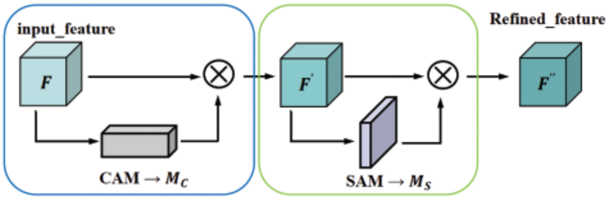


Figure 3: The overall process of the convolutional block attention module (CBAM)

For CAM:

$$\begin{aligned} M_C(\mathrm{F}) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathrm{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathrm{F}))) \\ &= \sigma(W_1(W_0(\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}})) + W_1(W_0(\mathrm{F}^{\mathrm{c}}_{\mathrm{max}}))), \end{aligned} \tag{2}$$

$$\mathrm{F}' = M_C(\mathrm{F}) \otimes \mathrm{F}, \tag{3}$$

For SAM:

$$\begin{aligned} M_S\,(\mathrm{F}) &= \sigma\,(\mathrm{f}^{7*7}([\mathrm{AvgPool}(\mathrm{F})); \mathrm{MaxPool}(\mathrm{F})])) \\ &= \sigma\left(\mathrm{f}^{7*7}\left(\left[\mathrm{F}^{\mathrm{S}}_{\mathrm{avg}}; \mathrm{F}^{\mathrm{S}}_{\mathrm{max}}\right]\right)\right) \end{aligned} \tag{4}$$

$$F'' = M_s(F') \otimes F' \tag{5}$$

Note that the MLP weights, $W_0 \in R^{C/r*C}$ and $W_1 \in R^{C*C/r}$, are shared for both inputs and the ReLU activation function is followed by W_0.

Generator Network Structure The network structure of generator G is composed of multiple layers of 2D transposed convolution. CBAM are inserted into the network structure of the G to focus on the important features of the face extracted from the voice signal, thereby improving the expressive power of the generator. Specifically, the network structure of the G is shown in Fig. 4. In the figure, except for the last layer, the ReLU activation function is added after each layer of two-dimensional transposed convolution, and Spectral Normalization is added in the second to fifth layers. The G needs to be able to generate a face portrait consistent with the identity features of the voice, as follows:

$$\mathrm{Y}^{\mathrm{V}} = \mathrm{IC}\,(\mathrm{V})\,, \mathrm{Y}^{\mathrm{P}} = \mathrm{IC}(\mathrm{P}) \tag{6}$$

Y represents the identity of an entity providing voice or facial data. Y^V denotes the real identity of the voice subject, and Y^P represents the real identity of the facial subject. IC () denotes the function that maps a voice or facial record to its corresponding identity.

Discriminator Network Structure The discriminator D includes a discriminator and a classifier. One of the discriminator serves to identify the authenticity of the image, and define labels for the real photo and the generated photo, respectively; another discriminator, which we also call a classifier, is used to verify that the generated faces are matched with real faces. The network of the D consists of multiple layers of 2-dimensional convolution, except for the last layer, where a LeakyReLU activation function is added after each layer of 2-dimensional convolution. In addition, Spectral Normalization is added to each layer of 2D convolution except for the first and last layers. The use of spectral normalization in the D can ensure the Lipschitz continuity of the network, which limits the drastic degree of function variation in the network, and makes the model more stable. The structure of the D network is shown in Fig. 5.

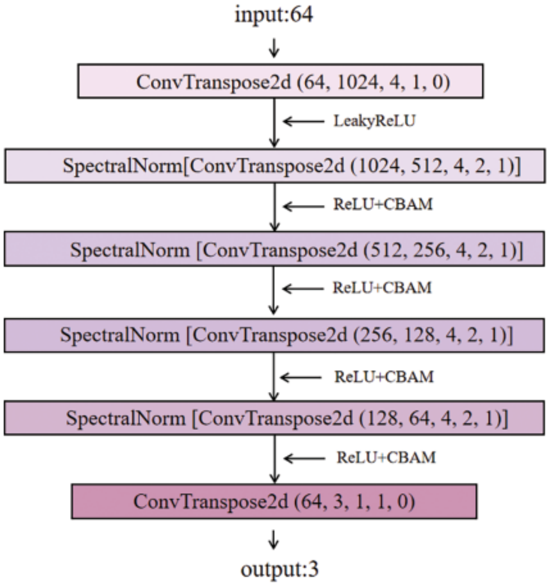

input:64
ConvTranspose2d (64, 1024, 4, 1, 0)
LeakyReLU
SpectralNorm[ConvTranspose2d (1024, 512, 4, 2, 1)]
ReLU+CBAM
SpectralNorm [ConvTranspose2d (512, 256, 4, 2, 1)]
ReLU+CBAM
SpectralNorm [ConvTranspose2d (256, 128, 4, 2, 1)]
ReLU+CBAM
SpectralNorm [ConvTranspose2d (128, 64, 4, 2, 1)]
ReLU+CBAM
ConvTranspose2d (64, 3, 1, 1, 0)
output:3

Figure 4: The Generator network structure diagram

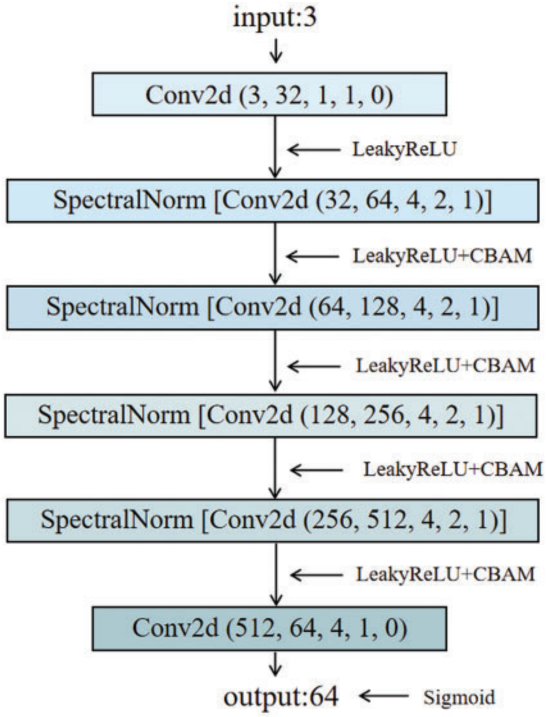

input:3
Conv2d (3, 32, 1, 1, 0)
LeakyReLU
SpectralNorm [Conv2d (32, 64, 4, 2, 1)]
LeakyReLU+CBAM
SpectralNorm [Conv2d (64, 128, 4, 2, 1)]
LeakyReLU+CBAM
SpectralNorm [Conv2d (128, 256, 4, 2, 1)]
LeakyReLU+CBAM
SpectralNorm [Conv2d (256, 512, 4, 2, 1)]
LeakyReLU+CBAM
Conv2d (512, 64, 4, 1, 0)
output:64
Sigmoid

Figure 5: Discriminator network structure diagram

Image Enhancement Module The purpose of the image enhancement module is to improve the clarity of the generated facial images by estimating a new, high-quality image based on the already generated ones. Our image enhancement module draws inspiration from the work of Wang et al. [20], incorporating a degradation removal module and a pretrained face GANs as a facial prior. The pretrained face GANs model was fine-tuned on the Voxceleb facial dataset. Through the face repair of the image generated by the generation module, the quality of the generated face image is further improved and the credible details are restored. The loss function used is consistent with that of Generative Facial Prior using GANs (GFP-GAN), comprising four parts: Adversarial loss, reconstruction loss, facial component loss, and identity preserving loss. The loss function is as follows:

Reconstruction Loss:

$$L_{rec} = \lambda_{l1}||\hat{\boldsymbol{y}} - \boldsymbol{y}||_1 + \lambda_{per}||\varnothing(\hat{\boldsymbol{y}}) - \varnothing(\boldsymbol{y})||_1 \tag{7}$$

where $\varnothing$ is the pretrained VGG-19 network [33]. λ_{l1} and λ_{per} denote the loss weights of L1 and perceptual loss, respectively, $\lambda_{l1} = 0.1$, $\lambda_{per} = 1$.

Adversarial Loss:

$$L_{adv} = -\lambda_{adv}E_{\hat{y}}softplus(D(\hat{\boldsymbol{y}})) \tag{8}$$

where D denotes the discriminator and λ_{adv} represents the adversarial loss weight, $\lambda_{adv} = 0.1$.

Facial Component Loss:

$$L_{comp} = \sum_{ROI}\lambda_{local}E_{\hat{y}ROI}\left[\log\left(1 - D_{ROI}\left(\hat{\boldsymbol{y}}_{ROI}\right)\right)\right] + \lambda_{fs}||Gram\left(\varphi\left(\hat{\boldsymbol{y}}_{ROI}\right)\right) - Gram(\varphi(\boldsymbol{y}_{ROI}))||_1 \tag{9}$$

where ROI is region of interest from the component collection {eyes and mouth}. D_{ROI} is the local discriminator for each region. φ denotes the multi-resolution features from the learned discriminators. λ_{local} and λ_{fs} denote the loss weights of local discriminative loss and feature style loss, repectively, $\lambda_{local} = 1$, $\lambda_{fs} = 200$. In Eq. (9), the first part represents the discriminator loss in the adversarial loss, and the second part is the feature style loss. The Gram matrix statistics are typically effective in capturing essential information about the problem.

Identity Preserving Loss:

$$L_{id} = \lambda_{id}||\mu(\hat{\boldsymbol{y}}) - \mu(\boldsymbol{y})||_1 \tag{10}$$

where μ represents face feature extractor. λ_{id} denotes the weight of identity preserving loss, $\lambda_{id} = 10$. The identity fidelity loss employs the pretrained facial recognition ArcFace model, enforcing that the restoration results maintain a small distance from the ground truth in a compact deep feature space.

The overall model objective is a combination of the above losses:

$$L_{total} = L_{rec} + L_{adv} + L_{comp} + L_{id} \tag{11}$$

3.4 Loss Function

The GAN-based face image generation method first generates the calculation of the loss function in the discriminator. Since the discriminator typically outputs a binary judgment of true or false, our AVP-GAN model uses binary cross-entropy loss function, in order to make the generated data distribution of the generator output closer to the real data distribution. Initially, the parameters of the generator are frozen, and the discriminator is trained to judge the ability of both real data and generated data, updating the parameters of the discriminator. We define the following loss function:

$$L_D = \left[\sum\nolimits_{i=1}^{n} ln(1 - D(G(p_i)) + \sum\nolimits_{i=1}^{m} ln(D(p_i))\right] \cdot w_n \tag{12}$$

$$L_C = \left[\sum\nolimits_{i=1}^{m} y_p \cdot \ln\left(C\left(p_i\right)\right)\right] \tag{13}$$

$$L_G = \left[\sum\nolimits_{i=1}^{n} y_v \cdot lnC\left(p_i\right) + \sum\nolimits_{i=1}^{m} \ln(D(p_i))\right] \cdot w_n \tag{14}$$

where L_D, L_C, L_G represents the loss functions of the discriminator (D), classifier (C), and generator (G). Here, w_n represents the weight, which is weighted for each sample and is set to a default value of 1 in this paper. $G(p_i)$ represents the generated face portrait by the generator. y_p represents the binary classification of the face portrait, which indicates whether the generated face image $\mathrm{p_i}$ is the same as the real identity face image y_i. If they are the same, $y_p = 1$; otherwise, $y_p = 0$. y_v represents the binary classification of the voice data, which indicates whether the voice identity label v_i is the same as y_i. If they are the same, $y_v = 1$; otherwise, $y_v = 0$.

The algorithm description of the AVP-GAN is shown in Algorithm 1.

Algorithm 1: AVP-GAN Algorithm

Inputs: A set of voice recordings with identity labels (V, Y^V). A set of labeled face images with identity labels (P, Y^P). A voice embedding network F_e (v; θ_e) trained on V for speaker recognition tasks. θ_e is fixed during training. Randomly initialized parameters θ_g (for the generator), θ_d (for the discriminator), θ_c (for the classifier).

Output: The parameters θ_g.

```
while not converge do
    Randomly sample a minibatch of n voice recordings $\{v_1, v_1, \ldots, v_n\}$ from $V$
    Randomly sample a minibatch of m face images $\{p_1, p_2, \ldots, p_m\}$ from $P$
    Apply CBAM to the discriminator
            - CAM: $M_C(\mathrm{F}) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathrm{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathrm{F})))$
            - SAM: $M_S(\mathrm{F}) = \sigma(\mathrm{f}^{7*7}([\mathrm{AvgPool}(\mathrm{F})); \mathrm{MaxPool}(\mathrm{F})]))$
    Update the discriminator $P_d(p; \theta_d)$ by ascending the gradient
            $\nabla\theta_d(\sum_{i=1}^{n} \log\left(1 - P_d\left(\widehat{p_i}\right)\right) + \sum_{i=1}^{m} \log P_d(f_i))$
    Update the classifier $P_c(p; \theta_c)$ by ascending the gradient ($a[i]$ indicates the $i$-$th$ element of
    vector $a$)
            $\nabla\theta_c(\sum_{i=1}^{m} \log P_c\left(p_i\right)[y_i^p])$
    Apply CBAM to the generator
    Update the generator $P_g\left(p; \theta_g\right)$ by ascending the gradient
            $\nabla\theta_g(\sum_{i=1}^{n} \log P_c(P_g\left(v_i\right))[y_i^v] + \sum_{i=1}^{m} \log P_d(P_g\left(P_e\left(v_i\right)\right)))$
end while
Apply image enhancement model to enhance the generated face images
    - Reconstruction Loss: $L_{rec} = \lambda_{l1}||\hat{y} - y||_1 + \lambda_{per}||\varnothing(\hat{y}) - \varnothing(y)||_1$
    - Adversarial Loss: $L_{adv} = -\lambda_{adv}E_{\hat{y}}softplus(D(\hat{y}))$
    - Facial Component Loss: $L_{comp} = \sum_{ROI} \lambda_{local}E_{\hat{y}ROI}\left[\log\left(1 - D_{ROI}\left(\hat{y}_{ROI}\right)\right)\right] +$
                              $\lambda_{fs}||Gram\left(\varphi\left(\hat{y}_{ROI}\right)\right) - Gram(\varphi(y_{ROI}))||_1$
    - Identity Preserving Loss: $L_{id} = \lambda_{id}||\mu(\hat{y}) - \mu(y)||_1$
    - Loss function: $L_{total} = L_{rec} + L_{adv} + L_{comp} + L_{id}$
```

4 Experiments

To assess the effectiveness of the proposed method in this paper, this section provides detailed explanations of the dataset used, experimental details, evaluation metrics, and experimental results. The specific experimental details are as follows.

4.1 Datasets and Experimental Details

In the training process, we utilized a dataset of 1251 speech samples from Voxceleb [18,34] and obtained corresponding facial datasets from the VGG face [35] dataset. The datasets corresponding to two identities include approximately 150,000 speech segments and facial images from 1225 different speakers. Following the data processing approach outlined in previous work [36], the speech was randomly cropped into audio segments of approximately 8 seconds. We extracted 64-dimensional logarithmic MEL spectrograms and cropped RGB facial images to a size of 3 × 64 × 64. This study follows the partition method of Nagrani et al. [18], where the training set, validation set, and test set are mutually exclusive. The dataset partitioning is detailed in Table 3.

Table 3: Dataset used in our experiments

	Train	Validation	Test	Total
# of identities	924	112	189	1225
VoxCeleb dataset				
# of portraits	106584	12533	20455	135972
VGGFace dataset				
# of voice segments	113322	14182	21850	149354

We used Pytorch to implement our methods and the AVP-GAN network proposed in this paper is trained for 50,000 epochs using the Adam optimizer, with a learning rate of 0.0002, β_1 of 0.5%, and β_2 of 0.999%. All experiments were conducted on a computer with a 12th Gen Intel (R) Core (TM) i7-12700H CPU, 16 GB RAM, and NVIDIA GeForce RTX 3070Ti GPU.

4.2 Evaluation Metrics

To assess the performance of our model in the task of voice portrait generation, we chose Voice2Face [36] as the baseline model for comparison. We evaluated our proposed AVP-GAN model from both qualitative and quantitative perspectives. In terms of qualitative evaluation, we trained the Voice2Face model and our model on our equipment, comparing the generated facial images from both models. Both models were tested using approximately 6-s and 8-s voice data. For quantitative evaluation, we employed the Face Cosine Similarity [37] as the assessment metric. A higher Face Cosine Similarity indicates that the generated facial images are closer to the real facial images. Specifically, we utilized Arcface [38] for facial image preprocessing. After performing facial recognition and feature extraction on the image data, we analyzed the facial image matrices and computed the cosine distance between the corresponding vectors of two compared images. The formula for calculating facial cosine similarity is as follows:

$$cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}\left(A_i^2\right)} \times \sqrt{\sum_{i=1}^{n}\left(B_i^2\right)}} \tag{15}$$

where suppose that A and B are the corresponding vectors in the original face picture and the generated face picture, respectively, and θ is the angle between vector A and vector B in space. The smaller the angle θ, the larger the cosine value, indicating a higher cosine similarity.

4.3 Experimental Results

4.3.1 Qualitative Evaluation

We input speaker voice segments into our model and conduct tests on varying durations of voice recordings for the same speaker's speech segments. Fig. 6 displays the test results for six randomly selected audio recordings from different speakers. In each example, we present facial images generated using different durations of speaker voice—specifically, 2, 4, 6, and 8 s. The qualitative results indicate that, when given voice segments longer than 6 s as input, our model's output facial images tend to stabilize, depicting clearer facial features and facial expressions.

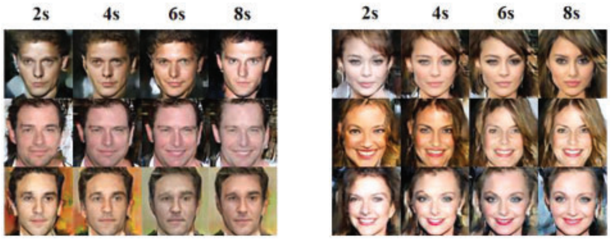


Figure 6: Facial images generated from voice segments of different durations

Fig. 7 presents a comparison between the original real images, the intermediate images (without image enhancement), and the final facial images generated by our AVP-GAN. It is evident that some facial details are not well-reconstructed in the intermediate images. The purpose of the image enhancement module is to better restore the details of the face generated from the voice and to improve the clarity of the generated facial images.

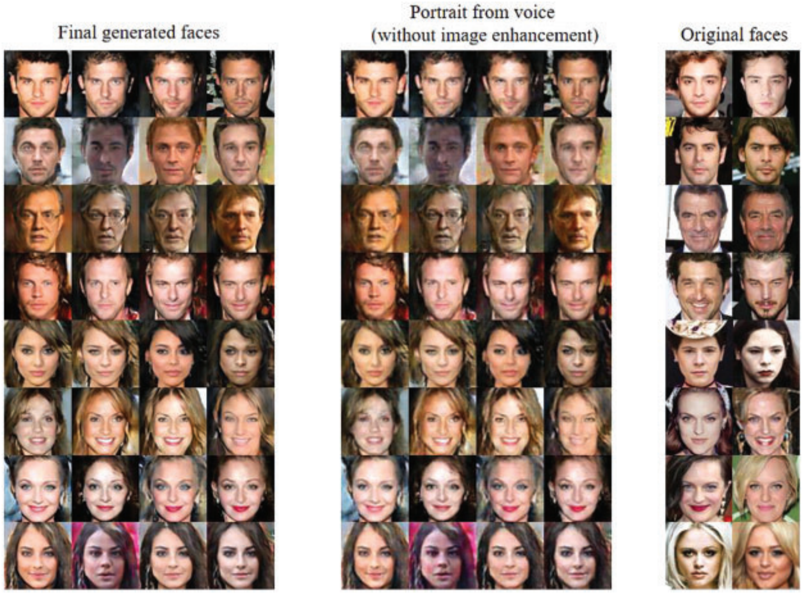


Figure 7: Comparison of the final face, intermediate face and original face

Choosing Voice2Face as the comparative model, both models were tested using 8-s audio data. Fig. 8 illustrates examples of the generated images from the comparative model and our AVP-GAN.

It can be observed that the face images generated by our method exhibit higher quality, with facial features closer to real faces.

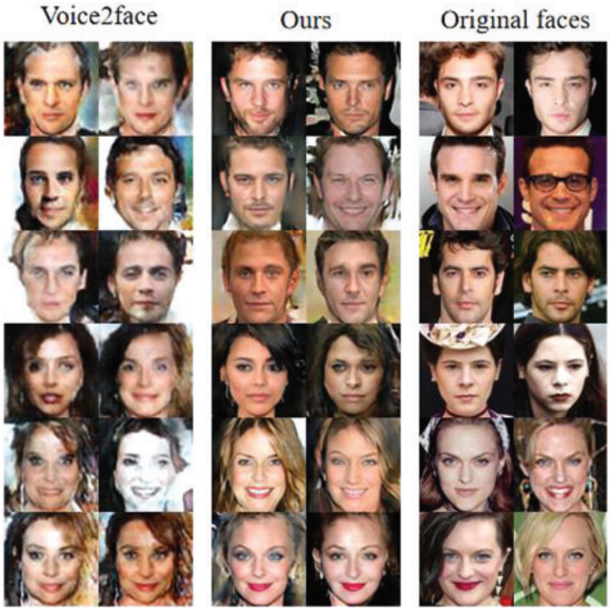


Figure 8: Comparison of results between our AVP-GAN model and the Voice2Face model

Fig. 9 displays a comparison between the facial images generated by our AVP-GAN model on the test set and real facial images. We randomly selected seven sets of male and seven sets of female results for presentation. From the images, it can be observed that our model is able to accurately capture features such as gender, age, face shape, facial features, and expressions for individuals.

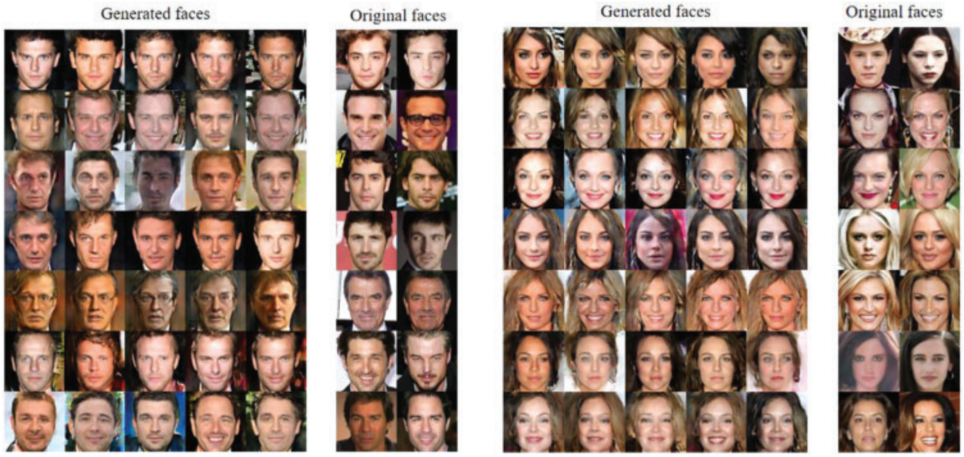


Figure 9: Comparison of 8-s voice-generated faces *vs*. reference faces

Through the analysis of qualitative results, we found that due to the significant variability in attributes such as hairstyle, makeup, and background, which have less correlation with the speaker's voice, it is generally challenging to establish a strong connection between these attributes and the voice of the speaker. Therefore, our model combines the attention mechanism with GANs to seek information about facial shape, facial features, and gender contained in the voice. From the comparison results between the generated faces and the reference faces, it can be seen that our proposed AVP-GAN model can establish a correlation between voice and attributes such as facial shape, expression, and gender, and generate a face image that is consistent with the identity of the speaker.

4.3.2 Quantitative Evaluation

Due to the difficulty in accurately capturing objective features such as hairstyle, background, and makeup through sound, directly comparing the similarity of two faces may result in significant errors. Therefore, we consider employing the following method for quantitative assessment of the depicted faces, initially recognizing facial features, followed by calculating facial cosine similarity. We generated faces using 4 and 8 s voice data, respectively, and calculated the cosine feature similarity between the original faces and the generated faces. This allowed for a quantitative assessment of the generated facial images. Simultaneously, we conducted ablation experiments on the 8 s voice dataset, evaluating the generated facial images with and without the inclusion of the CBAM attention module to verify its impact within the GANs. The quantitative evaluation results are presented in Table 4.

Table 4: Quantitative assessment results

Setting		Similarity
Method	Len.	Cosine
Voice2Face	4 s	0.408
AVP-GAN	4 s	0.471
Voice2Face	8 s	0.436
AVP-GAN (no CBAM)	8 s	0.447
AVP-GAN	8 s	**0.511**

The results indicate that the similarity of generated facial images from the 4 s audio dataset is lower compared to the similarity of generated facial images from the 8 s audio dataset. On the 8-s voice dataset, the model enhanced with the CBAM exhibited significantly higher cosine similarity metrics for generated images, compared to the model without the CBAM. This indicates that the CBAM effectively enhances the feature extraction capability of GANs. Furthermore, the cosine similarity metric between the facial images generated by our proposed AVP-GAN and the real images surpassed that of the baseline model Voice2Face on the 8-s voice dataset. This indicates that our AVP-GAN model possesses the ability to generate relatively high-quality static facial images.

5 Conclusion

In this research, we explored a relatively novel audio-visual cross-modal challenge: How to infer facial features from the voice of a person. To enhance the performance of the voice portrait model and increase the similarity between the generated and original faces, we innovatively introduced a new deep GANs model for voice portrait generation, known as the AVP-GAN, incorporating a convolutional attention mechanism. By integrating CBAM with the convolutional GANs, our model can better capture features that align with both the voice and facial characteristics. Additionally, we incorporated an image enhancement module, further improving the clarity of the generated facial images. The experiments demonstrate that our AVP-GAN can generate facial images that better align with the original facial features and exhibit higher clarity. We chose cosine similarity as the quantitative evaluation metric, and the quantitative results indicate that our AVP-GAN achieves a cosine similarity of 0.511, surpassing the performance of the comparison model. Furthermore, our AVP-GAN model demonstrates superior capability in generating improved facial images on an 8-s audio dataset, affirming the effectiveness of our approach without the need for any prior information. However, continuous improvement is still required in generating facial images that accurately capture the identity features of the speaker. For instance, our model has potential for refinement in exploring the correlation between speech and skin tone, and has yet to investigate the effects of voice portraits based on ethnicities from regions such as Asia and Africa. In future work, we plan to explore both the attributes of voice features and the search for cross-modal correlations between speech and facial features, with the goal of further improving the accuracy of generating facial features.

Acknowledgement: We would like to express our sincere gratitude to Fanliang Bu for his patient guidance and meticulous care during the writing of our paper. We are especially grateful to Dr. Yuchen Zhou for his assistance in writing our paper. Furthermore, we would like to extend our heartfelt thanks to the reviewers for their valuable suggestions.

Funding Statement: This work was supported by the Double First-Class Innovation Research Project for People’s Public Security University of China (No. 2023SYL08).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, methodology, formal analysis, writing-original draft, validation: Jingyi Mao; conceptualization, supervision, writing-review & editing: Fanliang Bu; formal analysis, validation, writing-editing: Yuchen Zhou; investigation, resources, validation, writing-editing: Yifan Wang; investigation, resources, writing-editing: Junyu Li and Ziqing Liu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this article are freely available in the mentioned references.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Singh, “Relations between voice and profile parameters,” in *Profiling Humans from Their Voice*, vol. 41. Singapore: Springer, pp. 85–120, 2019.

[2] X. Qin, N. Li, C. Weng, D. Su, and M. Li, “Cross-age speaker verification: Learning age-invariant speaker embeddings,” arXiv preprint arXiv:2207.05929, 2022.

[3] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, “Disjoint mapping network for cross-modal matching of voices and faces,” arXiv preprint arXiv:1807.04836, 2018.

[4] H. Kameoka, K. Tanaka, A. V. Puche, Y. Ohishi, and T. Kaneko, “Crossmodal voice conversion,” arXiv preprint arXiv:1904.04540, 2019.

[5] A. C. Duarte *et al.*, “WAV2PIX: Speech-conditioned face generation using generative adversarial networks,” in *ICASSP*, Brighton, UK, May 2019, pp. 8633–8637.

[6] I. Goodfellow *et al.*, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020. doi: 10.1145/3422622.

[7] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *ACM Comput. Surveys*, vol. 54, no. 2, pp. 1–38, 2021. doi: 10.1145/3439723.

[8] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *2010 4th Int. Conf. on Signal Process. and Commun. Sys.*, Gold Coast, QLD, Australia, IEEE, 2010, pp. 1–5.

[9] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19.

[10] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, vol. 6, pp. 22524–22530, 2018. doi: 10.1109/ACCESS.2018.2816163.

[11] P. H. Ptacek and E. K. Sander, "Age recognition from voice," *J. Speech Hear. Res.*, vol. 9, no. 2, pp. 273–277, 1996. doi: 10.1044/jshr.0902.273.

[12] R. Singh, B. Raj, and D. Gencaga, "Forensic anthropometry from voice: An articulatory-phonetic approach," in *2016 39th Int. Conv. on Inf. and Commun. Technol., Electron. and Microelectron. (MIPRO)*, Opatija, Croatia, IEEE, 2016, pp. 1375–1380. doi: 10.1109/MIPRO.2016.7522354.

[13] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," arXiv preprint arXiv:1703.09452, 2017.

[14] Z. Hong *et al.*, "When hearing the voice, who will come to your mind," in *2021 Int. Joint Conf. on Neural Netw. (IJCNN)*, Shenzhen, China, IEEE, Jul. 2021 pp. 1–6. doi: 10.1109/IJCNN52387.2021.9534208.

[15] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2.0 to speech recognition in various low-resource languages," arXiv preprint arXiv:2012.12121, 2020.

[16] W. B. Swift, "The possibility of voice inheritance," *Rev. Neurol. Psychiatry*, vol. 14, pp. 103–122, 1916.

[17] L. D. Rosenblum, N. M. Smith, S. M. Nichols, S. Hale, and J. Lee, "Hearing a face: Cross-modal speaker matching using isolated visible speech," *Percept. Psychophys.*, vol. 68, pp. 84–93, 2006. doi: 10.3758/BF03193658.

[18] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, Salt Lake City, Utah, USA, 2018, pp. 8427–8436.

[19] P. Wen, Q. Xu, Y. Jiang, Z. Yang, Y. He and Q. Huang, "Seeking the shape of sound: An adaptive framework for learning voice-face association," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2021, pp. 16347–16356.

[20] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2021, pp. 9168–9178.

[21] L. Li, L. Sun, Y. Xue, S. Li, X. Huang and R. F. Mansour, "Fuzzy multilevel image thresholding based on improved coyote optimization algorithm," *IEEE Access*, vol. 9, pp. 33595–33607, 2021. doi: 10.1109/ACCESS.2021.3060749.

[22] T H. Oh *et al.*, "Speech2Face: Learning the face behind a voice," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 7539–7548.

[23] J. Wang, X. Hu, L. Liu, W. Liu, M. Yu and T. Xu, "Attention-based residual speech portrait model for speech to face generation," arXiv preprint arXiv:2007.04536, 2020.

[24] A. K. Bragin and S. A. Ivanov, "Reconstruction of the face image from speech recording: A neural networks approach," in *2021 Int. Conf. on Qual. Manage., Transport and Inf. Security, Inf. Technol. (IT&QM&IS)*, Yaroslavl, Russian Federation, 2021, pp. 491–494. doi: 10.1109/ITQMIS53292.2021.9642810.

[25] C. Y. Wu, C. C. Hsu, and U. Neumann, "Cross-modal perceptionist: Can face geometry be gleaned from voices?" in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, New Orleans, Louisiana, 2022, pp. 10452–10461.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, Yaroslavl, Russian Federation, 2019, pp. 4401–4410.

[27] Y. Bai, T. Ma, L. Wang, and Z. Zhang, "Speech fusion to face: Bridging the gap between human's vocal characteristics and facial imaging," in *Proc. 30th ACM Int. Conf. on Multimedia*, Melbourne, VIC, Australia, pp. 2042–2050, Oct. 2022. doi: 10.1145/3503161.3547850.

[28] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, Honolulu, Hawaii, USA, 2017, pp. 5485–5493.

[29] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[30] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. on Comput. Vis.*, Venice, Italy, 2017, pp. 2439–2448.

[31] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," arXiv preprint arXiv:1802.05957, 2018.

[32] B. Ma, X. Wang, H. Zhang, F. Li, and J. Dan, "CBAM-GAN: Generative adversarial networks based on convolutional block attention module," in *Artif. Intell. and Security: 5th Int. Conf.,* New York, NY, USA, Springer, Jul. 2019, pp. 227–236.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[34] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.

[35] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proc. British Mach. Vis. Conf. 2015*, British Machine Vision Association, 2015, pp. 1–12.

[36] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," in *Adv. in Neural Inf. Process. Syst. 32 (NeurIPS 2019)*, 2019.

[37] H. Wang, *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, Salt Lake City, Utah, USA, 2018, pp. 5265–5274.

[38] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, Long Beach, CA, 2019, pp. 4690–4699.