Computers & Graphics 115 (2023) 359–370

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CAD/Graphics 2023

Generation of virtual digital human for customer service industry

Yanan Sun [a], Zhiyao Sun [a], Yu-Hui Wen [b,*], Sheng Ye [a], Tian Lv [a], Minjing Yu [c,*], Ran Yi [d], Lin Gao [e], Yong-Jin Liu [a,*]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[b] Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[c] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[d] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[e] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100864, China

ARTICLE INFO

Article history:
Received 12 May 2023
Received in revised form 3 July 2023
Accepted 9 July 2023
Available online 13 July 2023

Keywords:
2D virtual humans
Service gestures
Emotion editing
Gesture generation

ABSTRACT

With the widespread development of digital technology, individuals' daily activities are inseparable from interaction with electronic devices. Researchers have become interested in developing novel methods, to enable users to experience social and emotional satisfaction that traditional face-to-face interaction provides. In this study, we propose a novel deep learning-based pipeline to generate virtual digital humans for customer service industry. Specifically, we propose a method to construct a database with template service actions. Furthermore, we propose a two-stage method for generating 2D virtual human videos with gestures and emotional lip-sync expressions. We have conducted qualitative and quantitative experiments on the proposed 2D virtual human video generation method. The results demonstrate that the method effectively generates high-quality virtual digital humans for the customer service industry.

1. Introduction

Virtual digital humans can be applied in many applications such as feature films, game productions, and the immersive entertainment industry. Specifically, researches have found that virtual digital humans are important for improving empathy and engagement of users in human–machine interaction [1,2]. However, due to the high technical barrier of traditional virtual human production techniques, creating a high-quality virtual human requires a significant amount of time and effort. In recent years, artificial intelligence technologies have become increasingly comprehensive, leading to significant improvements in efficiency and visual quality of virtual human synthesis.

For developing virtual digital humans based on artificial intelligence technologies, audio-driven talking face generation has been extensively studied [3,4]. However, speech-driven human body generation with flexible gestures is much less explored. Gestures performed by artificial agents are important in human–machine interaction, because they are helpful for understanding utterances [5] and improving the intimacy between humans and agents [6]. Moreover, some researches have found that gestures are correlated with emotions and personality perceptions [7,8]. Thus, it is important to explore methods for generating talking human videos with audio-synchronized and realistic gestures. A recent work *Speech2Video* has proposed to synthesize talking human videos with upper body movements (i.e, head movements and gestures) from an input speech segment [9]. However, the method needs a large amount of training data (3 h of videos for a person), which considers only neutral facial expressions and random speech gestures. In addition, the synthesized results have visual artifacts in motion details, such as missing fingers. When creating virtual humans, the degree of resemblance and the presence of flaws directly impact the user experience due to the Uncanny Valley theory [10]. Apart from the efficiency and quality obstacles of virtual digital human synthesis, another unsolved challenge lies in the emotion expressiveness. Although natural emotional reactions are essential for vivid human–machine interaction [11], only a few works have studied talking-face generation with controllable emotions [12,13].

In this paper, we present a novel deep learning-based pipeline that is able to generate virtual humans with gestures and emotional lip-sync expressions for the customer service industry. Specifically, virtual service humans exhibit behavioral logic that conforms to certain rules, and they are utilized in specific situations to guide the behavior and enhance the experience of

* Corresponding author.
E-mail addresses: sunyn20@mails.tsinghua.edu.cn (Y. Sun), sunzy21@mails.tsinghua.edu.cn (Z. Sun), yhwen1@bjtu.edu.cn (Y.-H. Wen), yec22@mails.tsinghua.edu.cn (S. Ye), lt22@mails.tsinghua.edu.cn (T. Lv), minjingyu@tju.edu.cn (M. Yu), ranyi@sjtu.edu.cn (R. Yi), gaolin@ict.ac.cn (L. Gao), liuyongjin@tsinghua.edu.cn (Y.-J. Liu).

https://doi.org/10.1016/j.cag.2023.07.018
0097-8493/

users. To achieve this, we propose a method to construct a virtual digital human database for the customer service industry. To the best of our knowledge, the database is the first to capture human videos with both template service actions and emotional talking faces that cover basic service scenarios. Based on the database, we propose a two-stage pipeline for the virtual human video generation: (1) generating an action video by synthesizing transition frames between template actions; (2) generating an emotional lip-sync face to improve the facial part in the action video. The contributions we made in this work are as follows:

- By exploring actual application scenarios of service industries, we propose a method to collect videos of virtual humans with template service actions and emotional talking faces. This data collecting method lays the foundation for subsequent research on virtual human generation for the customer service industry.
- We propose a simple yet effective two-stage pipeline for generating virtual service human videos with semantically correct gestures and emotional lip-sync face expressions.

2. Related work

Virtual humans are computer-generated characters with digital appearances that rely heavily on display devices. A complete virtual human possesses three characteristics: first, it has a human appearance, with specific appearance and gender; second, it exhibits human behavior, with the ability to express itself through language, facial expressions, and body actions; third, it has humanoid thinking, with the ability to recognize specific scenarios and even interact with real humans.

In this paper, we focus on researches of generating virtual humans with realistic human appearance and vivid expressive behavior, which lay foundation for achieving a complete virtual human with humanoid thinking. Generating a full-body virtual human requires a much higher level of details for the face than for the body. As a result, the generation of the face and the body is often considered as two distinct problems. For researches that include generating both face and body, the generation process is still divided into two modules and generated separately [14].

2.1. Action video generation

Research on generating action videos of 2D virtual humans has progressed rapidly in recent decades and has garnered attention due to its ability to quickly produce videos. Some researchers use explicit 3D representations combined with neural rendering to synthesize the final image. Shysheya et al. [15] combine classical graphics processes with deep learning methods to learn a full-body model by estimating explicit texture maps and mapping input poses to UV coordinate images. However, due to the use of static texture maps, the synthesized results lack some high-frequency details contained in the original information. Liu et al. [16] propose a dynamic texture-based approach for pose prediction, which preserves details of the human body. However, the model requires professional equipments to provide accurate 3D reconstruction results, limiting its usage scenarios.

Some action video generation approaches use image translation networks [17–20]. However, these networks often require large-scale data. Some researches [21,22] make use of several sample data, but their synthesized results of human body postures are relatively blurry. Sun et al. [23] implement a robust action video generation model with a dynamic detail generation network (D^2G-Net) based on image translation and a video rendering framework. This approach does not require high-precision 3D reconstruction or large amounts of data. Wang et al. [19] and Chan et al. [14] generate high-quality dance videos of the target person based on image translation networks.

Recently, there has been significant progress in synthesizing human actions using neural radiance fields (NeRF). Weng et al. [24] and Liu et al. [25] use monocular videos combined with NeRF to achieve arbitrary view synthesis of human characters, with the latter also able to render details such as objects and backgrounds in the scene. Işık et al. [26] propose HumanRF, a 4D dynamic neural scene representation that captures full-body appearance in motion from multi-view video input, making a significant step towards production-level quality novel view synthesis. However, fully supervised learning methods based on large amounts of data are prone to overfitting, resulting in poor performance when generating actions that differ significantly from the actions in the training set. Therefore, some methods [27–34] use carefully designed processes to make the network suitable for more general action video generation tasks. Liu et al. [35] use a 3D human body mesh reconstruction module to decouple posture and shape. However, the images generated by these general methods have many problems, such as low resolution, limb blurring, distorted or even torn moving limbs, and discontinuous inter-frame textures.

To improve the quality of generation results on fine body parts, such as fingers or clothing, some researches [36–38] use optical flow to generate target posture frames. Zhou et al. [39] go further and propose a model for generating speaking person gesture videos based on action videos. The model splits and reassembles clips from a reference video through a video motion graph encoding valid transitions between clips based on the audio. Then, it uses a posture-aware video blending network to generate transition frames. The problem of frequent ghosting in the body and background parts is alleviated by using optical flow, resulting in more natural and realistic speaking person gesture videos. Inspired by this method [39], we propose a method to generate transition frames between template actions for virtual digital humans.

2.2. Audio-driven talking face generation

The task of audio-driven talking face video generation aims to synthesize a realistic and synchronized speaking video based on a single photo or video of a person and an input audio. To achieve this goal, the current mainstream approaches generally train a network to learn the mapping between audio features and visual features. The network is then used to predict the corresponding motion features from the input audio and apply the features to the input visual data.

Some researches [3,40–42] have focused on how to generate speaking faces based on videos. Prajwal et al. [4] propose a novel model called Wav2lip, which generates speaking faces based on a short input video. The model is trained on multi-person audio-visual data and focuses on the constraints of the mouth. It adds a powerful lip synchronization discriminator, which achieves better lip synchronization than previous works. However, the resolution of the generated lip part of the face is low, which cannot meet the requirements in high-resolution scenes, and there are defects when stitching the lip part to the original video. To improve synchronization and visual quality, Wang et al. [43] have used a lip-reading expert and a novel contrastive learning.

Considering that directly establishing the mapping between audio and expressions is complex, some methods use intermediate representations such as 3D Morphable Models (3DMM) [44–46] or facial keypoints [47]. Then, based on the intermediate representations, these methods generate lip movements corresponding to the input audio. Thies et al. [44] propose Neural Voice Puppetry based on Voice Puppetry [48]. The system is

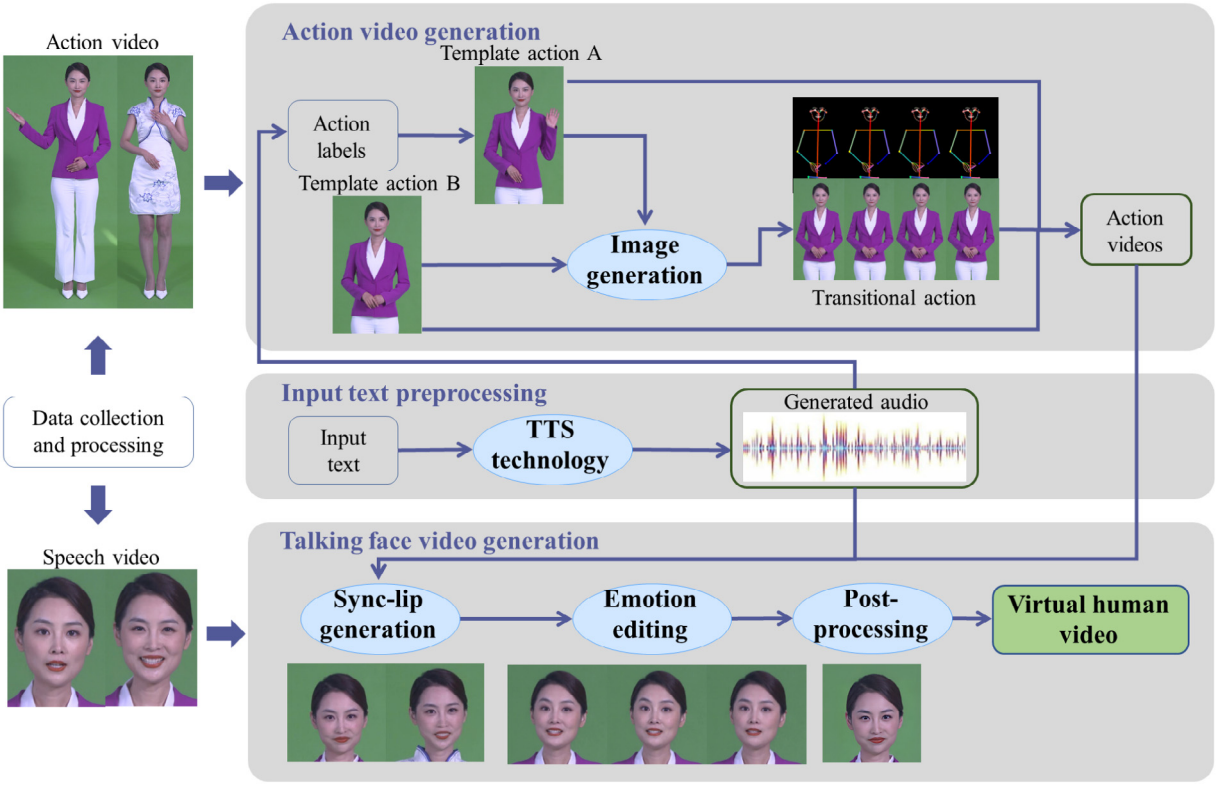


Fig. 1. Pipeline of our neural network-based 2D virtual service human video generation.

able to generate audio-driven videos of a specific person corresponding to input audio, and it reduces the scale of the required training video data to about 5 min. The AudioDVP model proposed by Wen et al. [49] uses 3D face reconstruction results to render images and can generate speaking person videos with controllable facial expressions, head positions, body postures, and lighting. Ye et al. [50] have proposed GeneFace, a generalized and high-fidelity NeRF-based talking face generation method that can generate results using out-of-domain audio, surpassing previous methods.

2.3. Facial expression editing

Face image generation refers to obtaining realistic face images through random latent code input, and common approaches use Generative Adversarial Networks (GAN) [51] to generate images. Face attribute editing refers to obtaining face images similar to the input face by modifying the attributes of the face, such as emotions, age, and facial features. Karras et al. [18] propose a face image generation network called StyleGAN, which can adjust face attributes by extracting different levels of features from different images. Furthermore, StyleGAN2 [52] improves the image quality and removes blob-shaped artifacts in generated images by adjusting the network structure. Subsequent works have achieved partial attribute editing of the face. DiscoFaceGAN [53] focuses on face attribute disentanglement, making expression, lighting, and posture attributes editable. Wu et al. [54] introduce an Expression Focal GAN (EF-GAN) that focuses on expressions, to capture better personalized features around the eyes, nose, and mouth to perform emotion editing. Ding et al. [12] propose a continuous facial expression editing method called ExprGAN. Recently, Sun et al. [13] propose a method that decouples the shape and texture of the face, resulting in a better performance.

3. Methods

In recent years, with the rapid development of deep learning technology, human video generation based on neural networks has also achieved extensive progress. Currently, some researchers have proposed virtual human generation methods that utilize Generative Adversarial Networks (GANs). These methods can automatically generate highly realistic and diverse virtual characters. However, most of them rely on a significant amount of real video data for training to generate realistic character images.

In this paper, we present a novel deep learning-based pipeline to synthesize a photo-realistic virtual digital human for customer service industry, given an input text content. Specifically, the pipeline is designed to avoid capturing a large amount of training data. To achieve this, we propose a general method to construct a virtual digital human database for customer service industry (Section 3.1). Then, we propose a two-stage pipeline for service virtual avatar generation: action video generation (Section 3.2) and emotional face generation (Section 3.3), as shown in Fig. 1. In more details, we firstly preprocess the input text and synthesize the audio using Text-to-Speech (TTS) technology. Secondly, we add corresponding action labels to the timeline based on the generated audio and obtain the target action clips accordingly. Then, we generate transition action frames between the template actions based on a conditional GAN model. All frames are spliced together to obtain the 2D virtual human action video. Next, we combine the action video with the audio and send it to the emotional lip-sync talking face generation network to obtain synchronized lip images. We use an emotion editing network and some post-processing operations, such as defect removal and super-resolution processing, to improve the quality of the faces in the action video.

Table 1
Template actions of virtual service human.

Action name	Action description
Explaining gesture	spread out hands
OK gesture	Raise left/right hand like OK
V gesture	Raise left/right hand like V-shaped
Waving gesture	Raise left/right hand and wave from side to side
Self-introduction gesture	Raise left/right hand and tap chest, then put down
Heart gesture	Raise left/right hand and express a finger heart
Guidance gesture	Raise left/right hand and spread it out
Introduction gesture	Raise left/right hand and spread it out to the upper left/right
Dropping Hands	Drop hands naturally from the state of cross-grip, then back to the state
Tilting head	Tilt head 30 degrees to left/right and back to original position
Nodding	nod head twice
Shaking	shake head twice

(a) Explaining (b) Dropping hands (c) OK gesture (d) V gesture

Fig. 2. Examples of some actions in the dataset.

(a) Positive emotion (b) Neutral emotion (c) Negative emotion

Fig. 3. Different emotions in the original recorded video.

3.1. Dataset construction

In this section, we introduce a method for constructing a virtual digital human dataset that can be applied in customer service industry, including how to collect and preprocess the raw video data.

3.1.1. Video data collection

Through an investigation of various types of service personnel, such as bank receptionists, shopping mall guides, and online anchors, we define a total of 12 categories of template actions. These template actions cover basic service scenarios, such as product introduction, welcoming guests, and simple interactions. A detailed introduction of the template actions are shown in Table 1, and some action examples are shown in Fig. 2.

To prepare for this work, professional photographers were asked to record speech and action video data of a professional anchor in a studio. The recorded video has a resolution of 3840 × 2160 and a frame rate of 30 fps. The face resolution in the video is not less than 256 × 256, and the background is a standard green screen. The environment is naturally lit, with no light spots on the anchor's body, and there will be no obvious changes in brightness during the limbs moving. The anchor should have appropriate makeup and should not have disheveled hair or bangs. The clothing should be as firm and slim as possible, with no obvious deformation during the body movement. And, no accessories that may reflect light, such as glasses or jewelry should be worn. In total, we have collected videos of the professional anchor wearing a formal suit and a cheongsam (in Fig. 2), respectively. The recorded audio-visual data has a cumulative duration of 30 min for each clothing type, including different types of actions. Moreover, the dataset includes all Chinese syllables and corpus with three emotions: positive, neutral, and negative, as shown in Fig. 3.

3.1.2. Pose data estimation

As shown in Section 3.1, we have collected some template actions of virtual digital humans for customer service industry. Then, our method generates a whole sequence of human action video frames, by synthesizing transition frames between the template actions based on a conditional GAN method with human poses as conditions. Thus, we have to collect pose data. Specifically, we utilize the OpenPose [55] toolbox to extract 2D coordinates of 25 keypoints of each human pose from action video frames. Due to self-occlusion and motion blur in the recorded video frames, we consider only keypoints with a confidence of 0.2 or higher as reliable. This approach balances limb integrity and recognition accuracy, allowing us to draw skeletal diagrams based on these reliable keypoints. In the diagrams, we use different colors to represent different bones.

3.1.3. 3D face data reconstruction

Our method generates emotional lip-sync face videos with 3DMM face data as intermediate representations. We use a 3D face reconstruction method [56] to obtain the 3DMM face coefficients $\mathcal{X} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{p}) \in \mathbb{R}^{257}$. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$ represent the shape, expression, and texture coefficients, respectively. $\boldsymbol{\gamma} \in \mathbb{R}^{27}$ represents the lighting coefficients, and $\boldsymbol{p} \in \mathbb{R}^6$ represents the pose coefficients. The facial reconstruction loss comprises three components: dense photometric alignment loss $\mathcal{L}_{photo}$, sparse landmark alignment loss $\mathcal{L}_{land}$, and statistical regularization loss $\mathcal{L}_{reg}$. The final fitting loss of the model [56] is:

$$\mathcal{L}(\mathcal{X}) = \lambda_{photo}\mathcal{L}_{photo}(\mathcal{X}) + \lambda_{land}\mathcal{L}_{land}(\mathcal{X}) + \lambda_{reg}\mathcal{L}_{reg}(\mathcal{X}) \quad (1)$$

where $\lambda_{photo} = 1.9$, $\lambda_{land} = 0.0016$, and $\lambda_{reg} = 0.0003$ [56].

The reconstructed pose-invariant face model $\mathbf{S}$ can be represented as:

$$\mathbf{S} = \bar{\mathbf{S}} + \mathbf{B}_{shape}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta} \quad (2)$$

where the average face shape $\bar{\mathbf{S}}$ and the shape basis $\mathbf{B}_{shape}$ are defined the same as in BFM09 [57], and the expression basis $\mathbf{B}_{exp}$ is consistent with FaceWareHouse [58].

For each vertex on the reconstructed model $\mathbf{S}$, we reproject it onto the image plane and rasterize the image position. Using

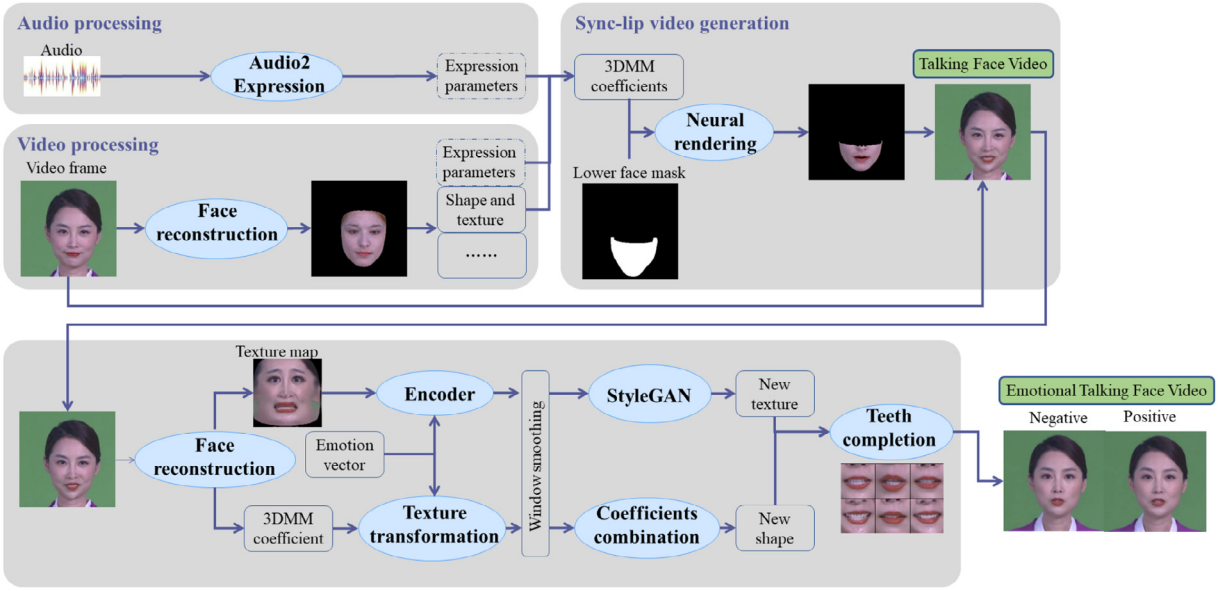


Fig. 4. Pipeline of emotional face generation.

this method, we establish the mapping relationship f from the pre-defined UV coordinates to the video frame pixels to get the texture.

3.1.4. Audio data extraction

To learn the mapping between audio and visual data, we have to further collect audio data features. Initially, we convert the audio data into Mel Frequency Cepstral Coefficients (MFCC) features. Subsequently, we feed these features into AT-net [47] to obtain the high-level features of the audio data. For each 40-millisecond input audio $\mathcal{A}_f$ at a video frame rate of 25 fps, the network extracts a 256-dimensional feature vector F.

3.2. Action video generation

Given an input text, we develop a method to generate semantically correct human action videos by synthesizing transition frames based on template actions defined in Section 3.1. Specifically, we utilize a conditional GAN [20] to generate virtual human action video frames. The conditional GAN simulates the conditional distribution of real images given the input skeletal diagram and optimizes the model performance through the minimax game between the generator and discriminator. The training dataset is defined as a set of image pairs (s_i, x_i), where s_i represents the skeletal diagram, and x_i is the corresponding real human action image, as described in Section 3.1.2. The trained conditional GAN is then used to infer the corresponding high-resolution virtual human video frames based on the input skeletal diagrams.

Our proposed method is able to generate action videos semantically conditioned on an input text with a few manually annotated action labels. Firstly, we preprocess the input text and synthesize the audio using Text-to-Speech (TTS) technology. Secondly, we add corresponding action labels to the timeline based on the input text and obtain the target action clips accordingly. Then, we use the image generation network to generate transitional action frames between the template actions. Finally, all frames are spliced together to obtain the 2D virtual human action video.

3.3. Emotional face generation

We propose an emotional lip-sync face generation method to make the virtual human suitable for expressing emotions in customer service industry. It is important for virtual service humans to express emotions properly in some application scenarios. For example, when a user makes correct manipulations during human–machine interaction, the virtual service human has to encourage the user in a positive emotion.

Inspired by existing talking face generation [49] and emotion editing methods [13], we propose to generate an emotional talking face video by first generating a neutral talking face video and then editing the emotion of the video. The input of our method includes the action video of a virtual digital human generated by the method from the last section, an audio segment and an emotion vector, which represents a neutral, positive or negative emotion. The emotion vector is represented as a one-hot vector $\boldsymbol{e}$ where only one element of a specific emotion is nonzero.

3.3.1. Talking face video generation

We utilize a parameterized 3DMM face model combined with neural rendering technology to generate virtual human lip-sync video frames [49]. The specific generation process is shown in Fig. 4. Using the 3DMM face model as an intermediate representation can effectively prevent the overfitting of correlations between audio and visual signals. We employ the 3DMM face coefficients $\mathcal{X}$ estimated from the input image I in Section 3.1.3.

In Section 3.1.4, we have extracted an audio feature vector F corresponding to each input audio segment. Inspired by Audio2Expression [49], the audio feature vector and the expressions reconstructed from the video frames are jointly used to train a network to learn the mapping from the input audio to the facial expressions. At each time step t, the audio features in the sliding window are stacked in chronological order to generate the final input feature $\mathbf{F}_t = \{F_i\}_{i=t-N_w}^{t+N_w}$, where the radius of sliding window N_w is 3, and non-existent previous or subsequent features are set to 0. The network is trained using Mean Square Error (MSE) to measure the network loss $\mathcal{L}_{exp}$:

$$\mathcal{L}_{exp} = MSE(\mathcal{H}(\mathbf{F}_t) - \delta_t) \tag{3}$$

where $\mathcal{H}$ represents the Audio2Expression network, and δ_t is the facial expression parameters obtained at time t from the reconstructed video.

Next, we perform facial rendering and stitching [49]. We use the parameters predicted by the Audio2Expression network to replace the facial expression parameters extracted from the original video frames. To render the synthesized image $\hat{I}$ corresponding to the 3D facial model parameters $\mathcal{X}$, we need to model the lighting conditions and camera position. After modeling, we calculate the coordinates $u_i(\mathcal{X})$ of each vertex $v_i \in v(\alpha, \delta)$ in the 3D model that are projected from the camera space to the 2D image space through projection Π, as well as the corresponding color $c_i(\mathcal{X})$. Finally, we feed $\{u_i(\mathcal{X})\}_{i=1}^{N_v}$ and $\{c_i(\mathcal{X})\}_{i=1}^{N_v}$ into a differentiable rasterize to render the synthesized image $\hat{I}(\mathcal{X}, \Pi)$. To minimize the generated face area as much as possible, we use a pre-trained lower face mask to extract the lower part of the face, which covers the area of the chin, lip, and part of the nose. Therefore, we can reduce the uncertainty caused by the dynamic background, and make the generated results more natural. After extracting the lower half of the face, we train the facial neural rendering network using the facial model parameters and corresponding images. The network consists of a generator G based on U-Net [59] and a discriminator D based on Patch-GAN [17]. Finally, we merge the generated lower face with the original face to get a sequence of head video frames that match the input audio.

3.3.2. Emotion editing

Our emotion editing method is designed to make facial expressions more diverse and realistic, following a previous work [13]. Specifically, we edit virtual digital human emotions by processing the shape and texture of the face, simultaneously. In more details, we adopt a face transformation network to edit the facial shape according to emotions while maintaining lip shapes. Since the geometric details in the texture are difficult to represent in the transformed shape, a texture transformation network is needed to process the texture map after the shape transformation. We utilize the StyleGAN [18], which is trained to leverage the hidden encoding obtained by projecting the input texture into a latent space and generate a new texture map based on the input encoding.

It is challenging to edit face emotions while retaining lip synchronization, due to the lack of paired training data of 3DMM face coefficients that are aligned to the same phoneme under different emotions. In this way, we train a shape transformation network with unpaired training data in a cycle-consistent manner [60]. To better optimize the generated results, in addition to using the adversarial loss $\mathcal{L}_{adv}$ commonly used in GAN networks to constrain the generator and discriminator, we also introduce a regression loss $\mathcal{L}_{reg}$ to ensure that the generated coefficients are consistent with the given emotion. In addition, we introduce a cycle-consistency loss $\mathcal{L}_{rec}$ [13] in the generator to measure the difference between the reconstructed result and the original result in 3D space. We also introduce a mouth shape preservation loss $\mathcal{L}_{mouth}$ [13] to constrain the generated lip shape to be similar to the original lip shape, as well as a regularization loss $\mathcal{L}_r$ to constrain the facial deformation. For the discriminator, in addition to the adversarial loss $\mathcal{L}_{adv}$ and the regression loss $\mathcal{L}_{reg}$, we also add a gradient penalty loss [61] $\mathcal{L}_{gp}$. Therefore, the objective functions for the generator and discriminator are:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{reg}\mathcal{L}_{reg}^G + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{mouth}\mathcal{L}_{mouth} + \lambda_r\mathcal{L}_r \tag{4}$$

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{reg}\mathcal{L}_{reg}^D + \lambda_{gp}\mathcal{L}_{gp} \tag{5}$$

where we set $\lambda_{reg} = 20$, $\lambda_{rec} = 5e^3$, $\lambda_{mouth} = 1$, $\lambda_r = 1e^3$, and $\lambda_{gp} = 10$ in all experiments following a previous work [13].

Compared with coarse face geometric shape, face textures have better appearance details to reflect emotions. In this way, we train the StyleGAN [18] to generate high-quality texture maps. Then, we train an encoder that makes it possible to control the texture map generation based on an emotional vector [62]. Finally, we calculate the editing direction for each emotion based on the encoding in the StyleGAN latent space. To further improve the lip synchronization and frame consistency of the generated results, we use window smoothing, facial blending operations, and a teeth completion operation to refine the texture of the mouth area and make the results more realistic.

Our proposed method is able to edit face emotions with a talking face video and an emotional vector as input. First, we use the texture map and the 3DMM coefficients $\mathcal{X}$ of the face in our constructed database. Then, we use the shape transformation network to process the coefficients of the 3DMM based on the input emotion vector while maintaining the lip movements. We also edit the texture map by modifying the encoding in the StyleGAN [18] latent space to capture more details based on the emotion vector. Next, we combine the modifications on the coefficients and the texture map to obtain a new face. The result is further smoothed between frames using a window smoothing module and optimized using a teeth completion module to enhance the realism of the video frames. The specific generation process is shown in Fig. 4.

4. Experiments

4.1. Implementation details

We implement our proposed method in PyTorch on a server with the Intel Xeon Gold 6126 CPU and NVIDIA Titan RTX GPUs. It takes about 8 h and 10 h to train the action video generation module and the talking face generation submodule on one GPU, respectively. For the emotional editing submodule, it takes 14 h to train the StyleGAN on four GPUs. Moreover, it takes 10 h to train the corresponding encoder of the StyleGAN and a shape transformation network on one GPU.

In the inference stage, it takes 0.4s for the method to generate a transition action frame of the virtual human video. When editing the emotion to negative or positive, 256 × 256 talking face videos are generated at a rate of about 10 fps. The frame rate can be further improved by generating and editing a batch of frames instead of one frame at a time, to satisfy realtime requirements of some applications, such as realtime human–machine interaction.

4.1.1. Action video generation module

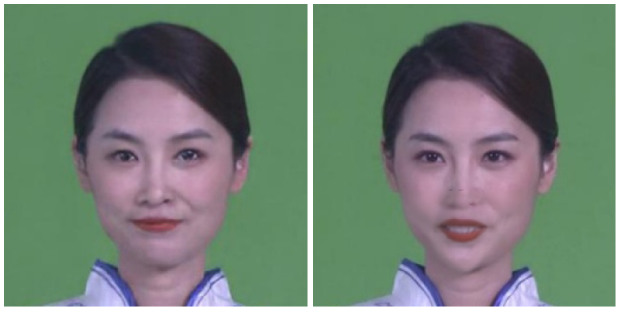
In order to generate realistic video frames, we use the pix2pixHD model [20], which achieves high realism and resolution based on the pix2pix model [17]. The pix2pix model comprises a generator G and a discriminator D. The role of G is to translate conditional labels into realistic images, while D aims to distinguish between real and generated images. The performance of the pix2pix model is optimized through the minimax game between the generator and discriminator, which can be expressed as:

$$\min\max \mathcal{L}_{GAN}(G, D) \tag{6}$$

where the objective function $\mathcal{L}_{GAN}(G, D)$ is:

$$\mathbb{E}_{(\mathrm{s},\mathrm{x})\sim p_{data}(\mathrm{s},\mathrm{x})}[\log D(\mathrm{s}, \mathrm{x})] + \mathbb{E}_{\mathrm{s}\sim p_{data}(\mathrm{s})}[\log(1 - D(\mathrm{s}, G(\mathrm{s})))] \tag{7}$$

The pix2pix model incorporates control information to guide the learning direction and generate images with controllable attributes. However, the generated images may not meet the requirements for generating high-resolution realistic images. The pix2pixHD model improves upon this by introducing a coarse-to-fine generator, a multi-scale discriminator architecture, and a robust adversarial learning objective function. The objective



(a) Head in original video frame (b) Head in generated video frame

Fig. 5. Comparison of real and generated head.

function of the proposed action image generation model can be expressed as:

$$\min_G \left(\left(\max_D \sum_k \mathcal{L}_{GAN}\left(G, D_k\right) \right) \right) + \lambda_{FM} \sum_k \mathcal{L}_{FM}\left(G, D_k\right) + \lambda_P \left(\mathcal{L}_P\left(G\left(s_i\right), x_i\right)\right) \tag{8}$$

where we set $\lambda_{FM} = 10$ and $\lambda_P = 10$ following a previous work [20] . k is the number of multi-scale discriminators, and it is set to 3. $\mathcal{L}_{FM}\left(G, D_k\right)$ is the feature matching loss, and $\mathcal{L}_P\left(G\left(s_i\right), x_i\right)$ is the perceptual reconstruction loss.

4.1.2. Emotional face generation module

Talking face video generation submodule. As depicted in Fig. 5, there are some problems in the generated results when compared to the original real video. Firstly, although we consider the possible flaws in the chin during the stitching process by selecting the appropriate facial mask shape, the flaws in the middle face cannot be resolved by adjusting the facial mask. As shown in Fig. 5(b), there is a small gray line in the middle of the nose. Although this flaw may not be noticeable at the first glance, it can cause a sense of unreality when it appears intermittently in dynamic and continuous video frames. Additionally, upon comparing the face image of the original video frame in Fig. 5(a) with the generated face image in Fig. 5(b), we observe a significant difference in clarity between the generated lower face and the original one. The generated result has blurred facial details and a decrease in the character's makeup. To address these problems, we apply some post-processing techniques to the generated head video frames.

To address the flaws in the center of the face, we can assume that the position of the flaw is in the center of the image because when cropping the head from the original video frame, the size and position of the bounding box need to ensure that the face is in the center of the cropped image. We can process the image pixels in each column that meets this condition: if there is a significant color jump in a segment of pixels (less than or equal to 3 pixels) compared to the color of the adjacent pixels above and below, we consider that there is a stitching flaw here and fill it with the interpolation of the color in the adjacent pixels above and below. To improve the resolution of the generated face and make it clearer, after repairing the flaws, we use the Generative Facial Prior GAN (GFP-GAN) [63] to further process the images. The GFP-GAN network comprises two parts: a U-Net for removing image degradation, flaws, and other problems, and a pre-trained GAN network for generating detailed faces.

Emotion editing submodule. Our emotion editing submodule contains a shape transformation network, a texture transformation network, a window smoothing operation, a facial blending operation and a teeth completion operation. In the following, we will provide a detailed introduction of these subnets and operations.

Shape transformation network: we use a traditional GAN network consisting of a generator and a discriminator. The input of the generator G includes an emotion vector $\boldsymbol{e}$, face shape and expression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, with a total of $n_c + n_e$ dimensions, where n_c is the dimension of the concatenated coefficients $\boldsymbol{c} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, and n_e is the number of all emotions except neutral emotion. The input of the discriminator D is the output of the generator or the concatenated $\boldsymbol{c}$ reconstructed from the real image. The output of the discriminator includes two parts: D_{rf} judges whether the input signal is real, and D_{reg} judges whether the emotion matches the target emotion. By inputting the coefficients $\boldsymbol{c} = (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{n_c}$ and the target emotion vector $\boldsymbol{e} \in \mathbb{R}^{n_e}$ into the trained generator G, we can obtain the predicted result $G(\boldsymbol{c}, \boldsymbol{e}) \in \mathbb{R}^{n_c}$. The edited shape can be reconstructed using the 3DMM coefficients $\boldsymbol{c}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}') = G(\boldsymbol{c}, \boldsymbol{e})$ and the original pose parameters $\boldsymbol{p}$.

For the generator and the discriminator, the regression loss can be written as:

$$\mathcal{L}_{reg}^{G} = \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}} \|\boldsymbol{e} - D_{reg}(G(\boldsymbol{c}, \boldsymbol{e}))\|^2 \tag{9}$$

$$\mathcal{L}_{reg}^{D} = \mathbb{E}_{\boldsymbol{c},\tilde{\boldsymbol{e}}} \|\tilde{\boldsymbol{e}} - D_{reg}(\boldsymbol{c})\|^2 \tag{10}$$

where $\tilde{\boldsymbol{e}}$ represents the real emotion vector corresponding to the shape coefficients $\boldsymbol{c}$.

In addition, we introduce a cycle-consistency loss $\mathcal{L}_{rec}$ [13], a mouth shape preservation loss $\mathcal{L}_{mouth}$ [13], and a regularization loss $\mathcal{L}_r$ as described in Section 3.3. The losses are defined as follows:

$$\begin{aligned}\mathcal{L}_{rec} =& \mathbb{E}_{\boldsymbol{c},\boldsymbol{0},\boldsymbol{e}} \|\mathbf{V}_{\mathbf{S}(\boldsymbol{c})} - \mathbf{V}_{\mathbf{S}(G(G(\boldsymbol{c},\boldsymbol{e}),\boldsymbol{0}))}\|_2^2 + \\ & E_{\boldsymbol{c}^*,\boldsymbol{e}^*,\boldsymbol{0}} \|\mathbf{V}_{\mathbf{S}(\boldsymbol{c}^*)} - \mathbf{V}_{\mathbf{S}(G(G(\boldsymbol{c}^*,\boldsymbol{0}),\boldsymbol{e}^*))}\|_2^2\end{aligned} \tag{11}$$

where $\mathbf{S}(\boldsymbol{c})$ represents the face shape with the parameters $\boldsymbol{c}$, and $\mathbf{V}_{\mathbf{S}(\boldsymbol{c})}$ is its vertex vector form. For each $\boldsymbol{v}_i \in \mathbf{V}_{\mathbf{S}(\boldsymbol{c})}$, $\boldsymbol{v}_i = (x, y, z)$ represents the coordinates of the ith vertex of the mesh $\mathbf{S}(\boldsymbol{c})$. In this equation, $\boldsymbol{c}$ represents the coefficients for the neutral expression, $\boldsymbol{c}^*$ represents the coefficients corresponding to the non-neutral expression $\boldsymbol{e}^*$, and $\boldsymbol{0}$ represents the zero vector corresponding to the neutral emotion.

$$\begin{aligned}\mathcal{L}_{mouth} =& \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}} \|(\mathbf{V}_{\mathbf{S}_u(\boldsymbol{c})} - \mathbf{V}_{\mathbf{S}_d(\boldsymbol{c})}) - (\mathbf{V}_{\mathbf{S}_u(G(\boldsymbol{c},\boldsymbol{e}))} - \mathbf{V}_{\mathbf{S}_d(G(\boldsymbol{c},\boldsymbol{e}))})\|_2^2 + \\ & \mathbb{E}_{\boldsymbol{c}^*} \|(\mathbf{V}_{\mathbf{S}_u(\boldsymbol{c}^*)} - \mathbf{V}_{\mathbf{S}_d(\boldsymbol{c}^*)}) - (\mathbf{V}_{\mathbf{S}_u(G(\boldsymbol{c}^*,\boldsymbol{0}))} - \mathbf{V}_{\mathbf{S}_d(G(\boldsymbol{c}^*,\boldsymbol{0}))})\|_2^2\end{aligned} \tag{12}$$

where $\mathbf{S}_u(\boldsymbol{c})$ and $\mathbf{S}_d(\boldsymbol{c})$ represent the vector forms of the keypoints in the upper and lower lip regions of the face model $\mathbf{S}(\boldsymbol{c})$, respectively.

$$\mathcal{L}_r = \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}} \|\mathbf{V}_{\mathbf{S}((\boldsymbol{\alpha},\boldsymbol{0}))} - \mathbf{V}_{\mathbf{S}((\boldsymbol{\alpha}',\boldsymbol{0}))}\|_2^2 \tag{13}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ represent the shape coefficients corresponding to $\boldsymbol{c}$ and $G(\boldsymbol{c}, \boldsymbol{e})$, respectively. In this loss, the expression parameters are set to $\boldsymbol{0}$.

Texture transformation network: we utilize a StyleGAN [18], which is trained to leverage the hidden encoding obtained by projecting the input texture into a latent space and generate a new texture map based on the input encoding. In traditional StyleGAN, the given latent code $\mathbf{z}$ in the latent space $\mathcal{Z}$ can be mapped to an intermediate code $\mathbf{w}$ through a mapping function $f : \mathcal{Z} \rightarrow \mathcal{W}$, which is then transformed into n styles through a set

of affine transformations $\{A_i|i = 1, 2, \ldots, n\}$. The entire process can be expressed as:

$$z\mathbf{w} = f(\mathbf{z}) \tag{14}$$

$$\hat{t} = \text{stylegan}(A_1(\mathbf{w}), A_2(\mathbf{w}), \ldots, A_n(\mathbf{w})) \tag{15}$$

where $\hat{t}$ represents the generated image. To enhance the accuracy of image reconstruction, we stack n distinct $\mathbf{w}$ in the texture transformation network to transform the image into the extended latent space $\mathcal{W}+$, and introduce an encoder to regress the image to the latent space. The complete process of transformation and reconstruction can be expressed as:

$$\mathbf{W} := \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\} = \text{enc}(t) \tag{16}$$

$$\hat{t} = \text{stylegan}(A_1(\mathbf{w}_1), A_2(\mathbf{w}_2), \ldots, A_n(\mathbf{w}_n)) \tag{17}$$

Window smoothing operation: we utilize a Hanning window weight to smooth the 3DMM coefficients and the latent codes of $\boldsymbol{I}^{t-1}$, $\boldsymbol{I}^{t}$, and $\boldsymbol{I}^{t+1}$ within the window for frame $\boldsymbol{I}^{t}$.

Facial blending operation: we use a soft mask to seamlessly blend the rendered face with the original background. In this mask, the values near the edge of the face gradually increase from 0 to 1 to achieve a smooth transition.

Teeth completion operation: since the 3DMM model does not include the internal structure of the mouth, it leaves a blank area in the mouth region. Therefore, we introduce a teeth completion network [13]. By reconstructing and rendering the face area, we can obtain paired training data with and without teeth, which is used to train the encoder to map the latent codes of images without teeth to those with teeth. Finally, we use a projection transformation to align the mouth area and fill the texture of the teeth area to the blank area.

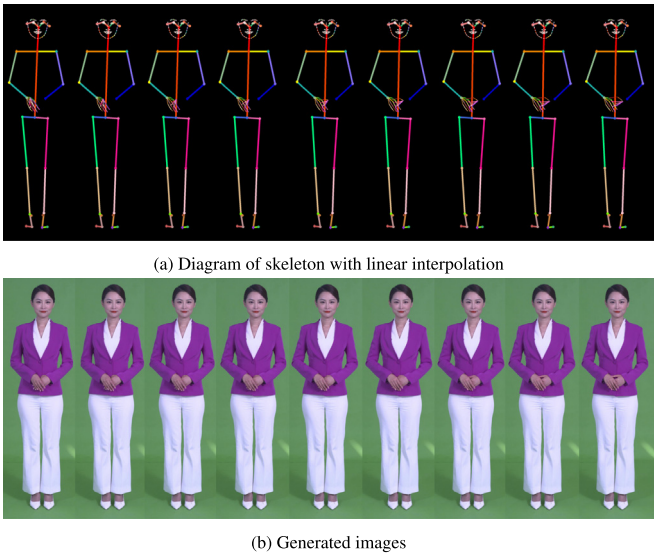
4.2. Result and discussion

We evaluate the performance of each module in our neural network-based 2D virtual digital human generation pipeline and analyze the experimental results. We show the quality of our results through qualitative visual performance and quantitative metric measurements.

4.2.1. Action video generation module

To generate videos that closely resemble real recordings, the focus of action video frame generation is on producing smooth transitions between different types of template actions. Initially, we perform linear interpolation between two consecutive template actions decided by the input text, to obtain a set of smooth transition skeletal diagrams. Subsequently, we use the trained action video generation model to synthesize the corresponding frames of the skeletal diagrams. Fig. 6 illustrates an example of a smooth gesture transition. It is evident that the character's right arm gradually moves away from the body, and the left thumb moves from an obscured state to a visible state. Other parts of the body, such as facial expressions and the clothing, exhibit good consistency within the motion clip. By employing a similar transition strategy, our proposed method can achieve seamless transitions between different template actions, and generate virtual human videos with semantically correct gestures that are helpful for customer service.

In Fig. 6, the clothing is a stiff uniform, and the overall texture is relatively simple. However, for the cheongsam, the fabric is soft and the texture patterns are complex, so it is difficult to learn the conditional distribution of human action frames conditioned on skeletal diagrams. Consequently, there may be sudden changes in the generation process. Fig. 7 displays two consecutive frames



(a) Diagram of skeleton with linear interpolation

(b) Generated images

Fig. 6. Example of smooth action transition.

Fig. 7. Example of sudden changes in non-rigid material.

in the video generated using the character wearing a cheongsam, and there is an abrupt change in the texture of the clothing, which affects the overall visual result.

To quantitatively evaluate the effectiveness of the generated video frames, we introduce three important metrics commonly used to evaluate generation results: Structural Similarity (SSIM) [64], Peak Signal to Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [65], and Fréchet Inception Distance (FID) [66].

The SSIM metric is used to measure the similarity between two images from the perspectives of brightness, contrast, and structure. Since humans are less sensitive to changes in brightness and color than to changes in edges and textures, this metric mainly compares the similarity of the image structure. The PSNR metric is a commonly used method for measuring image quality [67]. If the PSNR value is higher than 33 dB, it means that the image is very close to the original image. A value between 33-30 dB indicates that the image has some degree of distortion but is still acceptable. A value less than 30 dB indicates poor image quality. This metric is based on the mean square error (MSE).

However, the aforementioned metrics evaluate the entire image equally and may not accurately simulate human perception. To address this issue, Zhang et al. [65] proposed a deep visual

Table 2
Quantitative measurement results of action video generation module.

	Cheongsam	Formal suit
SSIM↑	0.9812	**0.9886**
PSNR↑	36.945	**42.050**
LPIPS↓	0.0055	**0.0036**
FID↓	20.269	**10.159**

feature-based evaluation metric called LPIPS. This metric provides similar results to human judgments in some image pairs where other metrics struggle to distinguish the similarity. This demonstrates that learning-based perceptual similarity metrics are closer to human perception in judging image similarity.

The FID metric shares a similar idea to LPIPS, as it also uses features for evaluation. This metric uses the Inception-v3 model to calculate the insight score. By comparing the distribution distance of the real image set and the generated image set in the feature space, it can accurately evaluate the quality of GAN-generated images.

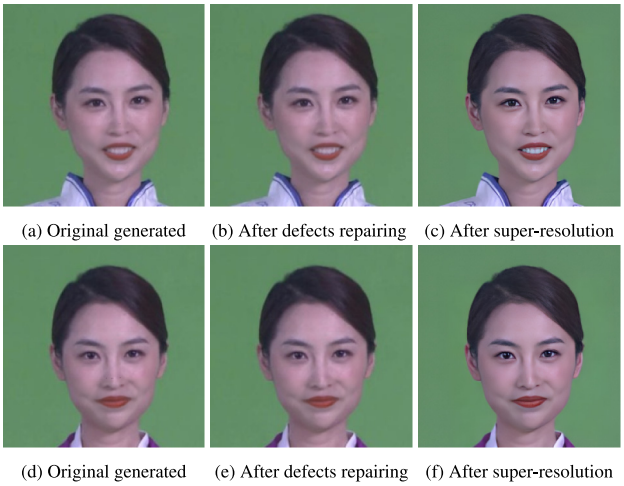
Table 2 summarizes the quantitative evaluation results of the action video generation module. For the SSIM and PSNR metrics, higher values indicate better performance, while for the LPIPS and FID metrics, lower values indicate better performance. The best metric values are shown in bold. The results indicate that our proposed method has a better performance on the character wearing the formal suit than the cheongsam in terms of all evaluation metrics . The wrinkles and complex textures do not move coherently with the cheongsam, thus leading to artifacts such as temporally instability of fine-scale details, as shown in Fig. 7. However, the generated results of the character wearing the cheongsam have excellent quality in terms of the PSNR $>$ 33dB [67]. This demonstrates the effectiveness of our method for processing challenging data with non-rigid materials and complex textures.

4.2.2. Emotional face generation module

The emotional face generation module has two submodules for generating talking face videos and editing emotions. Firstly, we evaluate the performance of the talking face video generation submodule. When interacting with virtual humans or watching videos of virtual humans, users naturally focus their attention on the faces of the character because the virtual human "speaks" in accordance with the input audio. The fineness and naturalness of the talking faces are important for the realism of the virtual human. Therefore, in the talking face video generation submodule, we introduce defect repair and super-resolution operations to improve the quality of the generated results. Images before the defect repair operation, as shown in Figs. 8(a) and 8(d), have shallow gray lines near the tip of the nose due to stitching defects. As shown in Figs. 8(b) and 8(e), after defect repair, the generated results become more natural, and there is no sudden change in the continuous video frames. The images after super-resolution processing are shown in Figs. 8(c) and 8(f), in which the character's facial features are clearer, and details such as makeup and hair are more realistic.

To quantitatively evaluate the talking face video generation submodule, we use the PSNR, SSIM, and FID metrics mentioned above, as well as the Identity Preservation (ID) metric, Cumulative Probability of Blur Detection (CPBD) [68], Lip-Sync Error-Distance (LSE-D), and Lip-Sync Error-Confidence (LSE-C) [69]. We use these metrics to evaluate the generated results before and after the defect repair and super-resolution processing operations.

The CPBD metric is a no-reference blur measurement metric that estimates the probability of detecting blur in each edge of the image using a probability model and calculates the cumulative



(a) Original generated (b) After defects repairing (c) After super-resolution
(d) Original generated (e) After defects repairing (f) After super-resolution

Fig. 8. Comparison of images after post-processing steps.

Table 3
Quantitative measurement results of talking face video generation submodule.

	Origin	Defect repair	Super-Resolution	Input
SSIM↑	0.6654	**0.6655**	0.6541	N/A
PSNR↑	18.237	**18.237**	17.940	N/A
FID↓	43.047	**39.826**	42.752	N/A
ID↑	**0.7936**	0.7935	0.7735	N/A
CPBD↑	0.0723	0.0719	**0.1252**	0.1386
LSE-D↓	8.6656	**8.6331**	8.7106	8.0712
LSE-C↑	5.2342	**5.2754**	5.1929	6.9971

probability. The LSE-D and LSE-C metrics measure the degree of matching between lip movements and audio pronunciation in videos. They use error distance and confidence probability to evaluate the synchronization between the video content and the audio. The ID metric evaluates whether the identity features of the character are well preserved during the generation process by calculating the cosine similarity between the average features of real images and generated images in ArcFace [70]. For the CPBD, LSE-C, and ID metrics, higher values indicate better performance, while for the LSE-D metric, lower values indicate better performance.

Table 3 summarizes the quantitative evaluation results of the talking face video generation submodule. The best values are shown in bold. The images after the defect repair operation are significantly better than the images after the other two stages in terms of the FID metric. This is because the defect repair operation smooths the gray dotted lines in the face, making the character features more prominent, and the overall changes to the image are relatively small. The difference between the original video image and the image after defect repair in the SSIM, PSNR, ID, LSE-D, and LSE-C metrics is very small. This is because manipulating a small number of pixels in a few frames in a continuous video is not enough to affect the pixel-by-pixel calculation metrics (SSIM and PSNR), nor it is enough to affect the metrics that judge identity preservation and lip-sync synchronization in the overall image (ID, LSE-D, and LSE-C).

Although the images after super-resolution processing have a slight decrease in the ID, LSE-D, and LSE-C metrics, there is a significant improvement in the CPBD metric. This is because the super-resolution network is trained on high-resolution image datasets and can learn prior knowledge from the network during the training phase to fill in or even add details in the original blurry areas, making the processed images perform well in CPBD. The changes to the image are relatively large, which lead to a loss of information in character identity and slight changes

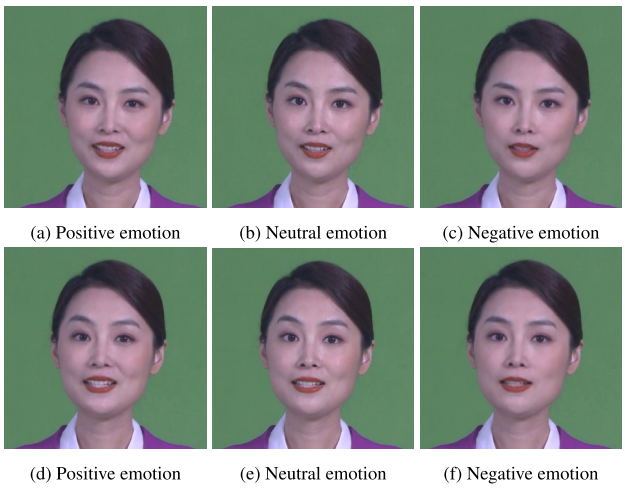


Fig. 9. Different emotions after editing (based on real neutral emotions).

in lip structure. However, in practical applications, humans can easily discriminate the features added by the super-resolution network, such as makeup, highlights on the face, and hair. People can still recognize the character's identity even if the makeup is heavier or there are more hairs on the forehead. Therefore, the introduction of the super-resolution post-processing module improves the effectiveness of our 2D virtual human generation method from a practical application perspective.

Next, we evaluate the performance of the emotion editing submodule, which aims to edit the shape and texture of the face to express emotions, while maintaining the synchronization of the lip movements. The positive and negative emotions edited from the neutral emotion are shown in Fig. 9. Figs. 9(a), 9(b), and 9(c) all show an open mouth shape, and the positive emotion after editing has a larger mouth shape and a slightly upward in the corners of the mouth, while the negative emotion has a smaller mouth shape and a more solemn expression. Figs. 9(d), 9(e), and 9(f) all show a slight upward tilt posture with an open mouth. In addition to the difference in the size of the mouth opening, the positive emotion after editing has slightly squinted eyes compared to the neutral emotion, simulating the state of the eyes bending when a real human smiles. For the negative emotion, the most obvious change is in the eyebrows. When the virtual human tilts her head, she has a slight expression of raising her eyebrows, which appears abrupt in the negative emotion. After editing from neutral emotion to negative emotion, the eyebrows are more symmetrical.

We compare the results edited by our method with those edited by ExprGAN [12], a continuous expression editing method proposed by Ding et al. The comparison results are presented in Fig. 10. Our method produces images with more intense emotions, higher clarity, and greater realism than ExprGAN. Furthermore, our results have a superior effect on the teeth area, with clearer textures, due to the inclusion of a teeth completion operation designed to fill in teeth textures.

Table 4 summarizes the quantitative evaluation results of our emotion editing method and ExprGAN on the data of the character in a formal suit. The table highlights the best values in bold. It can be seen that the generated results of our emotion editing method are better than the generated results of ExprGAN in all metrics. This is because our method decouples the shape and texture of the face and performs editing operations separately, making the results more natural. The CPBD, LSE-D, and LSE-C metrics are close to the input video, indicating that the method has a small impact on the clarity of the image and the lip-sync while editing the emotion. When comparing emotions internally, it can be found that the editing effect of positive emotions is better than that of negative emotions, which is because the features of positive emotions are more obvious, such as the upward turn in the corners of the mouth, the enlargement of the mouth shape, the slight bending of the eyes, and the upward movement of the cheekbones. However, the features of negative emotions are more likely to reflect in the overall expressions of the facial features, which are more difficult to learn and are prone to affecting other parts during editing.

Table 4
Quantitative measurement results of emotion editing submodule.

	ExprGAN		Our method		Input
	Positive	Negative	Positive	Negative	
SSIM↑	0.9135	0.9200	**0.9708**	0.9700	N/A
PSNR↑	27.928	28.765	**38.563**	38.298	N/A
FID↓	37.741	43.798	**8.8982**	11.946	N/A
ID↑	0.9228	0.9078	**0.9778**	0.9754	N/A
CPBD↑	0.0471	0.0495	**0.1284**	0.1274	0.1386
LSE-D↓	9.7377	9.7010	8.2500	**8.1539**	8.0713
LSE-C↑	4.8333	4.7124	**6.6612**	6.6268	6.9972

5. Conclusion

Our method uses Text-to-Speech technology and action labels obtained from the audio to drive the neural network to generate body movements and lip movements. The resulting head image can be further processed through emotion editing and super-resolution networks to generate a realistic virtual human video that is lip-synced with the audio and matched with the movements. Qualitative and quantitative analyses have demonstrated the irreplaceable role of each module in our method. The action generation module is the basis, constructing a character action video that matches the audio content through action labels. This video is also used as the background video for the lip shape video generation module, which generates lip movement synchronized with the audio through matching with the 3DMM coefficients, giving the virtual human the ability to speak. The addition of other modules optimizes the overall effect.

Our system can generate high-resolution videos that are close in quality to real recorded videos, with a rich database of common service actions. The videos generated by our system have almost no sudden changes. Our system is suitable for most video generation tasks in service scenarios.

CRediT authorship contribution statement

Yanan Sun: Writing – original draft, Methodology. **Zhiyao Sun:** Conceptualization, Methodology. **Yu-Hui Wen:** Conceptualization, Writing – original draft, Methodology, Supervision, Project administration. **Sheng Ye:** Writing – original draft, Methodology, Validation. **Tian Lv:** Methodology, Investigation. **Minjing Yu:** Writing – review & editing, Validation, Project administration. **Ran Yi:** Writing – review & editing, Investigation. **Lin Gao:** Writing – review & editing , Visualization. **Yong-Jin Liu:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

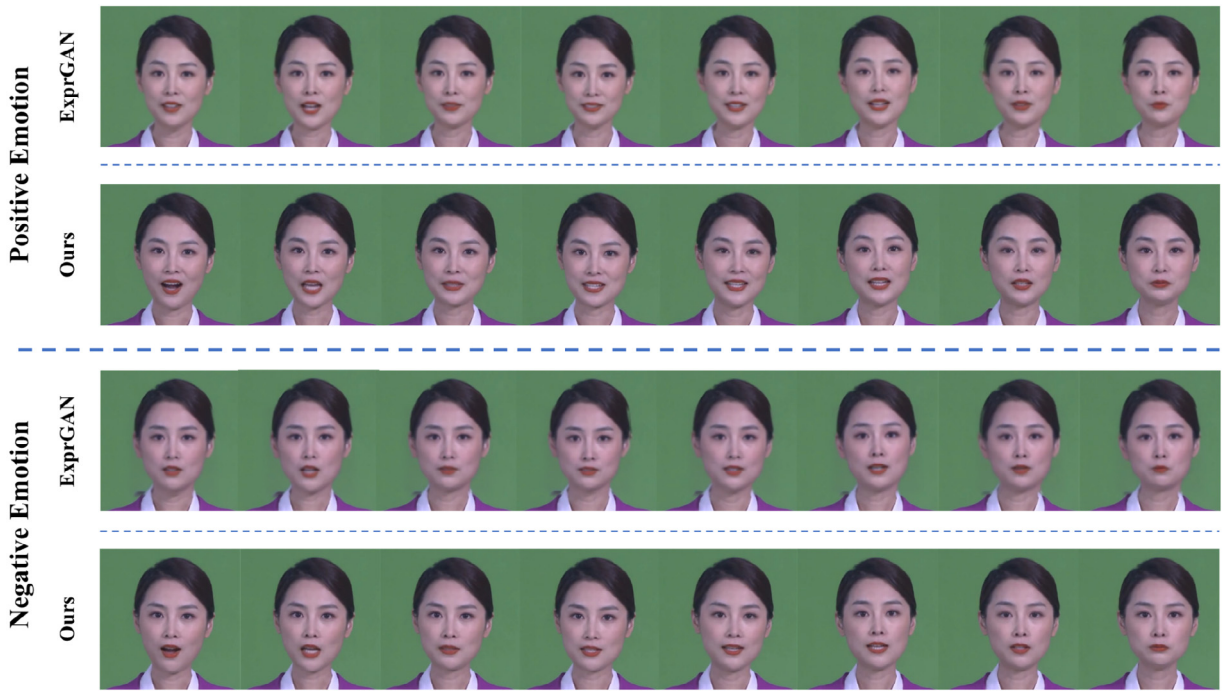


Fig. 10. Comparison results of edited emotions.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (62202257, 62002258), Beijing Jiaotong University Youth Elite Project (2023XKRC045), Shanghai Sailing Program (22YF1420300), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2023.07.018.

References

[1] Paiva A, Leite I, Boukricha H, Wachsmuth I. Empathy in virtual agents and robots: A survey. ACM Trans Interact Intell Syst (TiiS) 2017;7(3):1–40.
[2] Kimani E, Parmar D, Murali P, Bickmore T. Sharing the load online: Virtual presentations with virtual co-presenter agents. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems. 2021, p. 1–7.
[3] Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio. ACM Trans Graph 2017;36(4):1–13.
[4] Prajwal K, Mukhopadhyay R, Namboodiri VP, Jawahar C. A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia. 2020, p. 484–92.
[5] Bremner P, Pipe AG, Melhuish C, Fraser M, Subramanian S. The effects of robot-performed co-verbal gesture on listener behaviour. In: Humanoids. IEEE; 2011, p. 458–65.
[6] Wilson JR, Lee NY, Saechao A, Hershenson S, Scheutz M, Tickle-Degnen L. Hand gestures and verbal acknowledgments improve human-robot rapport. In: ICSR. Lecture notes in computer science, vol. 10652, Springer; 2017, p. 334–44.
[7] Castillo G, Neff M. What do we express without knowing? Emotion in gesture. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems. 2019, p. 702–10.
[8] Smith HJ, Neff M. Understanding the impact of animated gesture performance on personality perceptions. ACM Trans Graph 2017;36(4):1–12.
[9] Liao M, Zhang S, Wang P, Zhu H, Yang R. Speech2Video with 3D skeleton regularization and expressive body poses. 2020, CoRR abs/2007.09198.
[10] Mori M, MacDorman KF, Kageki N. The uncanny valley [from the field]. IEEE Robot Autom Mag 2012;19(2):98–100.
[11] Wang K, Wu Q, Song L, Yang Z, Wu W, Qian C, et al. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: Computer vision–ECCV 2020: 16th European conference. Springer; 2020, p. 700–17.
[12] Ding H, Sricharan K, Chellappa R. Exprgan: Facial expression editing with controllable expression intensity. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32, No. 1. 2018.
[13] Sun Z, Wen YH, Lv T, Sun Y, Zhang Z, Wang Y, et al. Continuously controllable facial expression editing in talking face videos. 2022, arXiv preprint arXiv:2209.08289.
[14] Chan C, Ginosar S, Zhou T, Efros AA. Everybody dance now. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 5933–42.
[15] Shysheya A, Zakharov E, Aliev KA, Bashirov R, Burkov E, Iskakov K, et al. Textured neural avatars. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 2387–97.
[16] Liu L, Xu W, Habermann M, Zollhöfer M, Bernard F, Kim H, et al. Neural human video rendering by learning dynamic textures and rendering-to-video translation. IEEE Trans Vis Comput Graphics 2020;PP:1.
[17] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 1125–34.
[18] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 4401–10.
[19] Wang TC, Liu MY, Zhu JY, Liu G, Tao A, Kautz J, et al. Video-to-video synthesis. In: Proceedings of the 32nd international conference on neural information processing systems. 2018, p. 1152–64.
[20] Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 8798–807.
[21] Wang TC, Liu MY, Tao A, Liu G, Kautz J, Catanzaro B. Few-shot video-to-video synthesis. In: Proceedings of the 33rd international conference on neural information processing systems. 2019, p. 5013–24.
[22] Zakharov E, Shysheya A, Burkov E, Lempitsky V. Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9459–68.
[23] Sun YT, Huang HZ, Wang X, Lai YK, Liu W, Gao L. Robust pose transfer with dynamic details using neural video rendering. IEEE Trans Pattern Anal Mach Intell 2022;45(2):2660–6.
[24] Weng CY, Curless B, Srinivasan PP, Barron JT, Kemelmacher-Shlizerman I. Humannerf: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 16210–20.
[25] Liu JW, Cao YP, Yang T, Xu EZ, Keppo J, Shan Y, et al. HOSNeRF: Dynamic human-object-scene neural radiance fields from a single video. 2023, arXiv preprint arXiv:2304.12281.

[26] Işık M, Rünz M, Georgopoulos M, Khakhulin T, Starck J, Agapito L, et al. Humanrf: High-fidelity neural radiance fields for humans in motion. ACM Trans Graph 2023;42(4):1–12. http://dx.doi.org/10.1145/3592415.
[27] Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L. Pose guided person image generation. Adv Neural Inf Process Syst 2017;30.
[28] Neverova N, Guler RA, Kokkinos I. Dense pose transfer. In: Proceedings of the European conference on computer vision. 2018, p. 123–38.
[29] Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J. Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 8340–8.
[30] Esser P, Sutter E, Ommer B. A variational U-net for conditional appearance and shape generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 8857–66.
[31] Ma L, Sun Q, Georgoulis S, Van Gool L, Schiele B, Fritz M. Disentangled person image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 99–108.
[32] Li Y, Huang C, Loy CC. Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 3693–702.
[33] Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N. Animating arbitrary objects via deep motion transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 2377–86.
[34] Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N. First order motion model for image animation. In: Proceedings of the 33rd international conference on neural information processing systems. 2019, p. 7137–47.
[35] Liu W, Piao Z, Min J, Luo W, Ma L, Gao S. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 5904–13.
[36] Siarohin A, Sangineto E, Lathuiliere S, Sebe N. Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 3408–16.
[37] Siarohin A, Woodford OJ, Ren J, Chai M, Tulyakov S. Motion representations for articulated animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 13653–62.
[38] Zheng H, Chen L, Xu C, Luo J. Unsupervised pose flow learning for pose guided synthesis. 2019, arXiv preprint arXiv:1909.13819.
[39] Zhou Y, Yang J, Li D, Saito J, Aneja D, Kalogerakis E. Audio-driven neural gesture reenactment with video motion graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 3418–28.
[40] Bregler C, Covell M, Slaney M. Video rewrite: Driving visual speech with audio. In: Proceedings of the 24th annual conference on computer graphics and interactive techniques. 1997, p. 353–60.
[41] Ezzat T, Geiger G, Poggio T. Trainable videorealistic speech animation. ACM Trans Graph 2002;21(3):388–98.
[42] Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman DB, et al. Text-based editing of talking-head video. ACM Trans Graph 2019;38(4):1–14.
[43] Wang J, Qian X, Zhang M, Tan RT, Li H. Seeing what you said: Talking face generation guided by a lip reading expert. 2023, arXiv preprint arXiv:2303.17480.
[44] Thies J, Elgharib M, Tewari A, Theobalt C, Nießner M. Neural voice puppetry: Audio-driven facial reenactment. In: 16th European conference on computer vision. Springer; 2020, p. 716–31.
[45] Yi R, Ye Z, Zhang J, Bao H, Liu YJ. Audio-driven talking face video generation with learning-based personalized head pose. 2020, arXiv preprint arXiv:2002.10137.
[46] Karras T, Aila T, Laine S, Herva A, Lehtinen J. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Trans Graph 2017;36(4):1–12.
[47] Chen L, Maddox RK, Duan Z, Xu C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 7832–41.
[48] Brand M. Voice puppetry. In: Proceedings of the 26th annual conference on computer graphics and interactive techniques. 1999, p. 21–8.
[49] Wen X, Wang M, Richardt C, Chen ZY, Hu SM. Photorealistic audio-driven video portraits. IEEE Trans Vis Comput Graphics 2020;26(12):3457–66.
[50] Ye Z, Jiang Z, Ren Y, Liu J, He J, Zhao Z. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. 2023, arXiv preprint arXiv:2301.13430.
[51] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020;63(11):139–44.
[52] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 8110–9.
[53] Deng Y, Yang J, Chen D, Wen F, Tong X. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 5154–63.
[54] Wu R, Zhang G, Lu S, Chen T. Cascade ef-gan: Progressive facial expression editing with local focuses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 5021–30.
[55] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 7291–9.
[56] Deng Y, Yang J, Xu S, Chen D, Jia Y, Tong X. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019.
[57] Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T. A 3D face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE international conference on advanced video and signal based surveillance. IEEE; 2009, p. 296–301.
[58] Cao C, Weng Y, Zhou S, Tong Y, Zhou K. Facewarehouse: A 3d facial expression database for visual computing. IEEE Trans Vis Comput Graphics 2013;20(3):413–25.
[59] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference. Springer; 2015, p. 234–41.
[60] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 2223–32.
[61] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein gans. Adv Neural Inf Process Syst 2017;30.
[62] Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, Shapiro S, et al. Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 2287–96.
[63] Wang X, Li Y, Zhang H, Shan Y. Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 9168–78.
[64] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–12.
[65] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 586–95.
[66] 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. Adv Neural Inf Process Syst 2017;30.
[67] Ennaji Y, Boulmalf M, Alaoui C. Experimental analysis of video performance over wireless local area networks. In: 2009 International conference on multimedia computing and systems. IEEE; 2009, p. 488–94.
[68] Narvekar ND, Karam LJ. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). IEEE Trans Image Process 2011;20(9):2678–83.
[69] Chung JS, Zisserman A. Out of time: automated lip sync in the wild. In: Computer vision–ACCV 2016 workshops: ACCV 2016 international workshops. Springer; 2017, p. 251–63.
[70] Deng J, Guo J, Xue N, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 4690–9.