OPEN ACCESS Freely available online
S. Liu, "Audio-Driven Talking Face Generation: A Review"
J. Audio Eng. Soc., vol. 71, no. 7/8, pp. 408–419, (2023 July/August).
DOI: https://doi.org/10.17743/jaes.2022.0081

Audio-Driven Talking Face Generation: A Review

SHIGUANG LIU,* *AES Member*
(lsg@tju.edu.cn)

College of Intelligence and Computing, Tianjin University, Tianjin, P.R. China

Given a face image and a speech audio, talking face generation refers to synthesizing a face video speaking the given speech. It has wide applications in movie dubbing, teleconference, virtual assistant, etc. This paper gives an overview of research progress on talking face generation in recent years. The author first reviews traditional talking face generation methods. Then, deep learning talking face generation methods, including talking face synthesis for a specific identity and talking face synthesis for an arbitrary identity, are summarized. The author then surveys recent detail-aware talking face generation methods, including noise based approaches, eye conversion based approaches, and facial anatomy based approaches. Next, the author surveys the talking head generation methods, such as video/image driven talking head generation, pose information–driven talking head generation, and audio-driven talking head generation. Finally, some future directions for talking face generation are highlighted.

0 INTRODUCTION

Visual and sound, two of the most important senses for human perception, play together to help one understand the world. The audio-visual synchronization is critical for one's immersion experiences [1–6]. Among the audio-visual topics, talking face generation is a recent focus of research.

Given a face image and a piece of speech audio, speech audio driven talking face generation aims to synthesize a video of the face image speaking the speech, as shown in Fig. 1. The talking face generation technology is widely used in human-computer interaction. For example, digital human technology has been applied in many scenarios, such as online customer service, virtual professors, virtual anchors, AI doctors, etc. Compared with only voice communication, people prefer to talk face to face. Additionally, this technique can also be applied to the design of auxiliary equipment for hearing impaired people, by converting speech information into visual information to help hard-of-hearing people understand the communication. As the work of video generation from images, talking face generation can expand the dimension of information and significantly improve the user experience. In some situations in which data transmission is limited, only the bandwidth of audio transmission can be supported to allow users to conduct "video" communication with the help of talking face generation.

The key to talking face generation is to improve the authenticity of the face and its motion, i.e., a good talking face generation method needs to ensure that the appearance and identity of the face in the video are consistent with the given face image, and the mouth movement matches with the input speech. However, it is challenging to generate a realistic talking face video that is indistinguishable from the real video.

Early studies [27, 8, 9] focused on how to preserve the face of the target and improve the authenticity of mouth movement. In recent years, the effect of talking face generation has been significantly improved with the continuous development of deep learning and computer vision research and researchers' attention has gradually shifted from mouth synchronization to richer facial details. Among them, eye movements are critical to the overall authenticity of the generated results.

Obviously, in order to improve the user experience and the overall authenticity of the generated video, eye movements are indispensable. Among the movements of various eye areas, winking is the most frequent and obvious action, which has the greatest impact on the overall authenticity. The common blink patterns are various, and the blink frequency is largely determined by personal habits. Therefore, simply generating random blink actions [30–32] cannot simulate all the real situations, which means that the controllability of blink generation is also necessary. However, because the speech usually does not contain blink information directly, the current deep learning architectures [53, 28, 29] are difficult to directly establish correct

*To whom correspondence should be addressed, e-mail: (lsg@tju.edu.cn). Last updated: November 30, 2022

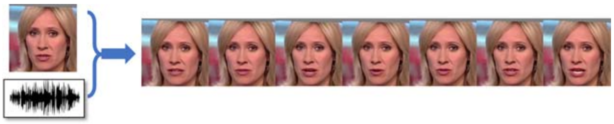

Fig. 1. An illustration of speech driven talking face generation. Left top: the face image; left bottom: the speech audio; right: the synthesized video sequence of the taking face.

speech-blink mapping, so it is unable to generate realistic eye movements.

In addition to the mouth movements and natural facial movements details (e.g., frowning, cheek movements, eye blink) synchronized with the speech audio, the personalized head movements are also important to increase the vividness of the talking face generation results. The previous work mainly focused on the generation of facial motion synchronized with speech, emphasizing on the mouth motion with high correlation with speech. Only the mouth motion was accompanied by some blink changes, but the authenticity of such a talking face was insufficient. The real speech movement is a combination of head posture change and various facial regions. It includes not only the mouth movement highly correlated with the speech audio, but also the detailed change of eyes and cheeks and head posture change with weak correlation. Therefore, adding facial detail movement and head movement, which have weak speech correlation, can improve the video realism and further enhance the user's sense of experience and realism. The talking face generation with head movement is also called talking head generation.

In the early days, the speech audio driven talking face generation technology manipulated mouth movement by building a specific face model. With the development of neural network, the method based on deep learning has gradually become the main research direction. In recent years, the Generative Adversarial Networks (GANs) have been widely used in the field of talking face generation and have achieved compelling results. At the same time, many researchers also began to carry out relevant research based on the infrastructure of the encoder-decoder. The author classified previous work into four categories, i.e., traditional talking face generation, deep learning based talking face generation, detail-aware talking face generation, and talking head generation, as shown in Fig. 2.

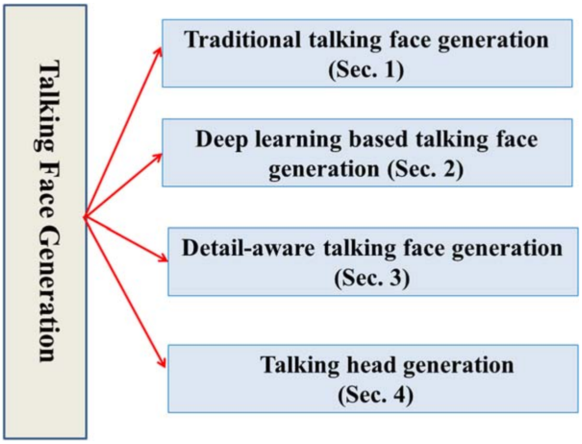
Talking Face Generation
Traditional talking face generation (Sec. 1)
Deep learning based talking face generation (Sec. 2)
Detail-aware talking face generation (Sec. 3)
Talking head generation (Sec. 4)

Fig. 2. An overview of previous talking face generation methods.

1 TRADITIONAL TALKING FACE GENERATION

Traditional talking face generation focuses on the mapping of audio features (e.g., Mel-scale frequency cepstral coefficients) to visual features (e.g., landmarks and videos) [7]. Table 1 summarizes the main traditional talking face generation methods.

Table 1. Traditional talking face generation methods.

Refs.	Year	Main points
Cao et al. [8]	2005	A visual representation: animes
Xie et al. [9]	2007	An audio-visual articulatory model based on DBN
Wang et al. [10]	2013	Statistical HMM

The relationship between language and facial motion has long been used by some computer graphics methods that assume a direct correspondence between basic speech and video units. Cao et al. [8] established a visual representation called animes, which corresponds to audio features. Under certain constraints of collaborative clarity and smoothness, the visual representation graph is searched to find the sequence that best represents a given utterance. Additionally, the system can also detect the speaker's mood and adjust the animation accordingly so that it can generate corresponding motion on the whole face. The animation sequence is time curled to match the speaking time and mixed to achieve a smooth final effect. Note that this method performs training with a speech-related facial motions database, from which a generative model of facial motion with emotion control and accurate lip-synching is generated. This method interpolates between key frames to achieve smooth motion. However, the simplification of facial dynamics may lead to unnatural lip motion.

Later, Xie et al. [9] transformed speech to mouth motion with an articulatory modelling approach. They used the audio-visual articulatory model based on the Dynamic Bayesian Network (DBN) to directly simulate the movement of the articulators, such as lips, tongue, and teeth. Note that the visual and auditory signals are synchronized by a multiple-stream structure equipped with a shared articulator layer. They further optimize the facial parameters estimated from audio using a Baum-Welch DBN inversion approach. This method indicates the synchronization between visual mouth and the speech audio and demonstrates that different articulators evolve asynchronously. It is reported that the facial parameters computed by this method is closer to the ground truth than the phonemic Hidden Markov Model (HMM) approaches.

Wang et al. [10] used audio-visual databases to train the statistical HMM of lip movement and used the trajectory generated by HMM as a guide to select the best mouth image sequence in the original training database, fusing it

Table 2. Deep learning–based talking face generation methods.

Refs.	Identity	Face region
Fan et al. [11]	Specific	Mouth
Suwajanakorn et al. [13]	Specific	Whole face
Karras et al. [7]	Specific	Whole face
Garrido et al. [27]	Arbitrary	Mouth
Zhu et al. [56]	Arbitrary	Mouth
Chung et al. [53]	Arbitrary	Mouth
Taylor et al. [21]	Arbitrary	Mouth
Chen et al. [54]	Arbitrary	Mouth
Prajwal et al. [55]	Arbitrary	Mouth
Prajwal et al. [23]	Arbitrary	Mouth
Wiles et al. [19]	Arbitrary	Whole face
Zhou et al. [20]	Arbitrary	Whole face
Song et al. [14]	Arbitrary	Whole face
Jamaludin et al. [15]	Arbitrary	Whole face
Chen et al. [29]	Arbitrary	Whole face
Pham et al. [22]	Arbitrary	Whole face
Huang et al. [24]	Arbitrary	Whole face
Yu et al. [25]	Arbitrary	Whole face
Zhou et al. [28]	Arbitrary	Whole face

with a background face video to generate the final talking face results. It is noted that this method is automatic. Given a 20-min audio/video, this method is able to produce a realistic talking face video of the same speaker synchronized with any speech audio.

Thies et al. [12] proposed a real-time facial reenactment method of a monocular target video. They used a dense photometric consistency metric to track the facial expressions of the source video and target video and then realized resynthesis through fast and effective deformation transfer between them. They retrieve the mouth interior that best matches the redirected expression from the target sequence and deform it to produce an accurate fit.

2 DEEP LEARNING–BASED TALKING FACE GENERATION

With the extensive application of convolutional neural networks (CNNs) and deep neural networks, more and more deep learning–based talking face generation methods have been developed. Table 2 summarizes the existing deep learning–based talking face generation methods.

2.1 Talking Face Generation for Specific Identity

One goal of talking face generation is to synthesize face motion of the target identity from the speech audio signal based on the specific target identity. Some methods construct 3D face models for specific identity, and then generate corresponding speech animation by manipulating the face models. Another methods train by collecting a large number of video clips of specific targets as data sets.

Fan et al. [11] introduced deep bidirectional Long Short-Term Memory (BLSTM) for talking face generation. They generated two sequences from a audio-visual talking dataset: 1) a contextual label sequence obtained through forcibly matching audio and text and 2) a visual feature by applying the Active Appearance Model on the lower face area of all training image samples. They then train a regression model with the square error of predicting visual sequence from label sequence as the loss function. It is reported this method is superior to HMM-based methods, both quantitatively and qualitatively.

Table 3. Talking head generation methods.

Methods	References
Video/Image Driven Methods	Wiles et al. 2018 [19], Zakharov et al. 2019 [36], Ha et al. 2020 [37], Koujan et al. 2020 [38], Doukas et al. 2021 [39]
Pose Information Driven Methods	Burkov et al. 2020 [42], Zhou et al. 2021 [43]
Audio Driven Methods	Karras et al. 2017 [7], Chen et al. 2020 [46], Thies et al. 2020 [44], Yi et al. 2020 [45], Zhou et al. 2020 [33]

Suwajanakorn et al. [13] learned the mapping of Obama's personal audio and mouth animation through a recurrent neural network (RNN) and matched it with a specific 3D pose. Many methods build specific 3D facial models for selected subjects and make the face move by manipulating the 3D mesh of the facial model or searching for matching models from the library. This method can produce a realistic result of Obama speaking in a target video matched a given speech audio.

Karras et al. [7] realized talking face generation in real time by exploiting CNNs to convert audio features into 3D meshes of the face model of a specific person. This mapping between visual and audio can disambiguate the expression variation in the face. However, the face sequences generated by such a model are sometimes unclear because the identity information related to the speaker's face and the information related to speech are coupled. Therefore, it is difficult to learn the generation of a talking face video in a purely data-driven way. To solve this problem, Liu et al. [16] proposed a deep network (i.e., deep voxel flow) that uses a combination of the pixel value–based method and the optical flow–based methods to predict video frames by flowing pixel values. This method can automatically synthesize new video frames both in an interpolation manner and an extrapolation manner. Some video generation methods solve similar problems by generating the whole sequence at once [17] or in small batches [18].

The above methods need to collect a large number of video frames with target identity, so talking face videos can only be generated for face images with specific target identity. This type of target-specific talking face generation method relies on the construction of face model, needs to collect a large number of speech video clips of target identity, and can only generate mouth movement of speech face for specific targets, which is difficult to adapt to the face image of any target.

2.2 Talking Face Generation for Arbitrary Identity

Some researchers have proposed various methods to generate talking face for arbitrary identity. First, researchers focused on synthesizing the lip motion of arbitrary identity. Then, they payed attention to the whole face synthesis.

2.2.1 Lip Motion-Aware Talking Face Generation

Garrido et al. [27] proposed VDub, a system of modifying the mouth motion of an actor in a face video to achieve reasonable visual alignment to a dubbed audio track. This approach depends on a fine capture of 3D facial performance and high-quality 3D reconstruction of the mouth region. It has great potential in movie and TV production dubbing. Zhu et al. [56] proposed a dynamic attention block based on lip region to retain identity information and use the features of lip movement to separate identity related features from lip related features. Chung et al. [53] proposed an end-to-end structure of two-stream ConvNet system, which can train unlabeled data to learn the mapping between speech audio and lip image.

Taylor et al. [21] presented a simple yet efficient deep learning–based method for automatic talking face synthesis. A sliding window predictor was proposed to match phoneme label input sequences with lip motion. Note that they used deep neural networks to convert phoneme sequences into shape sequences of the lower part of the face, ensuring that the change is independent of identity. This method is real time, with minimal parameter tuning, that can be integrated into the existing production framework. Chen et al. [54] realize talking face generation for any identity by learning a disentangled audio-visual representation. They defined a new loss to couple speech and mouth movement and used a GAN with three branches to judge the synchronization of lip movement and speech.

The LipGAN by Prajwal et al. [55] gets joint audio and video embedding by covering the lower part of the face and encoding speech and image during training. When a pose is given, LipGAN will give a mouth shape that matches the pose, generate lip movement according to short speech clips, and then paste the generated mouth area back to the original video seamlessly to get a speech video.

Previous work can synthesize precise mouth movement for a static image of a specific person; however, it may fail for an arbitrary identity in dynamic, unconstrained cases. Prajwal et al. [23] proposed the Wav2Lip model to accurately change the lip movement of any identity in the dynamic speaking face video by learning from a lip-sync discriminator. They also designed novel metrics to measure lip synchronization in an unconstrained speech video. Given an arbitrary long speech audio and visual facial animation, Eskimez et al. [26] generated a talking face video in an end-to-end manner. They used a mouth region mask to make the network focus on mouth movement instead of speech independent movement. A GAN training helps the quality of the resulting facial animation and the lip-audio consistency. A noise-resilient training further improves the robustness of the network to unseen speech noise. The above methods focus on generating the mouth motion of the face and only consider the mapping of speech audio to mouth motion, which therefore cannot generate a complete face animation image from the input face image and speech audio alone.

2.2.2 Whole Face-Aware Talking Face Generation

Wiles et al. [19] used streams to generate video-based high-precision arbitrary identity talking faces, and claimed to be able to generate video from audio. However, if there is no specific separation of facial and lip motion information, high-quality results cannot be generated. It can be seen that many of the above methods heavily rely on 3D face models or specific person's material data, which makes it difficult to extend to any identity. Additional training or redesigning methods and steps are required to adapt these types of method to new face images. To this end, Zhou et al. [20] proposed the VisemeNet, a deep learning–based, audio-driven, animator-centric speech synthesis approach. Its main ingredient is a three-stage LSTM network, which can achieve real-time synchronization between lip motion and a given speech audio. They used speaker independent features such as visemes and JALI (Jaw and Lip) parameters to make their methods applicable to non-specific people.

Given an arbitrary face image and a speech audio, Song et al. [14] proposed the use of conditional RNNs to produce a talking face video. The recurrent unit includes image and audio features for temporal dependency. Then, a pair of spatial-temporal discriminators and a lip-reading discriminator are built for improving the visual quality of the talking face video and increase the lip synchronization. Jamaludin et al. [15] proposed an encoder-decoder CNN model for talking face generation, which uses the joint embedding of face and audio to synthesize a taking video for a complete face. Note that this model was trained on unlabeled videos using cross-modal self-supervision. It is real-time and suitable for faces and audio unavailable in the training set.

Chen et al. [29] proposed a cross-modal talking face generation method by transferring audio to high-level facial structure (i.e., landmarks). Note that landmark is formed by key points of a face marking the positions of different facial features (e.g., corners of eyes, pupils, mouth, nose, etc.), which can be detected through the face detection algorithm. With the facial landmarks, they then produce talking face video frames in a hierarchical, cascade way. A dynamic pixel-wise loss with an attention mechanism helps the network focus on the frame area highly related to audio so as to decrease the pixel jittering phenomenon. A discriminator is used to produce a clearer visual result with better lip-audio synchronization. However, the result still has the problem of missing facial details. This method ignores other facial areas, such as eyebrows and eyes, so the overall result is still slightly stiff. This method directly provides the global region of the face image in the generation process, and it is difficult to find the relationship between audio and local facial motion in the network.

Pham et al. [22] proposed a deep learning–based talking face generation method with implicit emotional awareness in real time. They used LSTM RNNs to generate real time emotional facial animations from a given speech audio. Note that they train an LSTM network to learn a mapping from the a given audio to facial movements. Here, a set of audio features (e.g., Mel-scaled spectrogram, Mel frequency cepstral coefficients, and chromagram) are used to characterize the input speech audio, and a blend-shape model is employed to characterize the visual facial animation. In this way, the emotion can be represented by the expression weights of the face model in an implicit manner.

Because the goal of a good talking face generation method is to preserve both video quality of facial animation and lip-audio synchronization, Huang et al. [24] proposed a coarse-to-fine, tree-like architecture for directly synthesizing realistic face frames from speech audio clips. Then a video-to-word regeneration module is developed to transform the resulting talking face videos to the word space so as to match the speech audio. Such a multi-level framework can produce fine-grained talking face results, of which the visual quality and lip-audio consistency are superior to previous approaches. Existing talking face generation methods are limited by allowing only one type of information (e.g., audio, text, etc.) as input.

Given any input of a speech audio or texts, Yu et al. [25] proposed a multi-modal learning method for talking face generation with spatial-temporal dependency. They produce mouth landmarks with multiple inputs with a multimodal learning method. Then a deep learning network, Face2Vid, is designed to generate video frames–based on the above landmarks. Note that the Face2Vid used the strategy of optical flow and self-attention to maintain the temporal dependency and spatial coherency, respectively.

Zhou et al. [28] proposed a talking face generation method by adversarially disentangled audio-visual representation. They regard a talking face video sequence as a combination of subject-related identity information and speech-related information. Therefore, a network, called DVAS, is designed that can separate facial identity information and speech features (and their corresponding lip motion information) and make use of them to synthesize a talking face video with more realistic results. The mouth effect of the synthetic video of the model is fine; however, there is still a lack of handling of details such as blinking. In the pure audio-driven case, the result is basically unable to show the effect of blinking, or there are occasional extremely short actions like trying to close the eyes. However, the information separated from their method can only be used to generate mouth movements.

The above methods can generate a whole speaking face for any face and any speech audio, but they directly map speech audio to facial motion, so the generated talking face is only limited to mouth motion and miss the movement of other areas of the face (e.g., eyes, eyebrows, and cheeks). However, in the actual speaking process, different areas of the whole face have different degrees of movement, such as blinking, eyebrow picking, cheek movement, muscle line changes, etc. The face generated by ignoring these details looks unnatural.

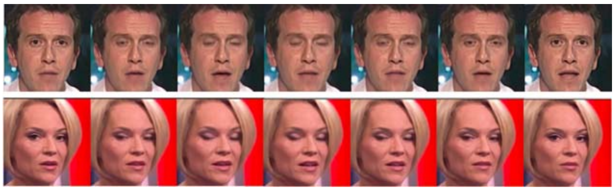

Fig. 3. Detail-aware talking face generation results [35].

3 DETAIL-AWARE TALKING FACE GENERATION

Most of the talking face generation methods focus on improving the authenticity of mouth movements and identity information. As shown in Fig. 3, a good speech-driven talking face video should include the following attributes: audio-lip synchronization, identity preservation of the target person, accurate mouth movements, and realistic eye blinks. There were few works for more facial details other than mouth motion. Recently, more and more researchers began to pay attention to tackle this issue.

3.1 Noise-Based, Detail-Aware Talking Face Generation

Vougioukas et al. [30] specially designed a speech-driven facial animation architecture based on GANs that can drive eye movement. The proposed GAN architecture consists of three discriminators to realize high-quality visual animation, audio-lip consistency, and vivid facial expressions (e.g., blinks and eyebrow movements), respectively. However, the way it generates blink is driven by randomly generated noise. This inevitably makes the blink uncontrollable and easily leads to distortion effects.

Sinha et al. [31] proposed an identity-preserving talking face generation method. They estimated person-independent facial landmarks from the speech audio using DeepSpeech feature. Then, eye blinks are produced with unsupervised learning and adapt the person-independent landmarks to person-specific landmarks to preserve the facial structure. A GAN-based architecture with attention mechanism is used to learn facial texture from person-specific facial landmarks. However, this method also used noise to drive the landmark of the eye to generate blinking.

Zeng et al. [32] used expression-tailored GAN to generate a talking face video in an end-to-end way. Different from previous methods that use face image and speech audio as input, this method uses an expression video of arbitrary identity as the source. In this method, the expression encoder and the audio encoder serve to disentangle expression-tailored representation from the source expression video and disentangle visual-audio representation, respectively. Similarly, this method applied embedded Gaussian noise on each video frame through Gate Recurrent Unit to achieve randomness of generated eye blink results.

Similar problems occur in the method of Zhou et al. [33]. Its slow blink action is uncontrollable, and the blink mode is relatively simple.

Because most of the above methods are driven by noise, their blinking cannot be controlled freely, and it is difficult to establish a correct relationship with the speech audio. The main reason why existing speech-driven talking face generation methods cannot effectively yield realistic faces with blinks is that they usually try to directly drive the animation of the entire face including eye blinks by speech audio. However, it is not easy to estimate eye movements from speech audio alone. When the model accepts phoneme level audio as input, this problem is particularly prominent, because such a short audio cannot even contain semantic information, it is impossible to successfully establish a reasonable mapping with the blink action.

3.2 Eye Conversion–Based Detail-Aware Talking Face Generation

Recently, Hao et al. [34] proposed a method for talking face generation with controllable blinking actions via eye conversion. This proposed architecture uses two phases to generate blink motion. By separating the synthesis of eye movements from the generation of speaking faces, the blink action can be freely controlled by the user, avoiding the network from establishing wrong audio-blink mapping during learning. In this architecture, they propose a blink conversion network based on the traditional cycleGAN [61], and design two conversion modules, which are respectively used to convert the input eye opening image to the half closed and closed eye blink images. The generated blinking image is replaced back to the original frame sequence through frame replacement, and the talking face video with controllable blinking action is obtained.

In view of the scarcity of half closed eye data in the training data set, they design a joint training method, which makes the blink conversion network make full use of each other's data to enhance the conversion effect of generators in the dual module training. In addition, a new face loss with mask is proposed, which are employed to partially replace the loss of identity features, so as to retain the mouth movement details to the greatest extent and strengthen the eye conversion ability.

Considering that a high frame rate video requires smoother eye movements and richer blink patterns, Liu and Hao [35] proposed a joint feature driven talking face generation architecture. This architecture uses identity encoder and speech audio encoder to extract identity features and speech features from face images and speech audio, respectively. On this basis, a new blink score is defined, which is employed to intuitively represent the degree of eyelid closure. This blink score is converted into specially designed blink features, and combined with identity features and speech audio features. Then, it is fed into the decoder, and generate a video frame where the degree of eyelid closure is directly controlled by the blink features.

In order to ensure that the blink feature has full control over the blink action, and eliminate the impact of the input face image and identity feature on the eye action, they design a GAN training model for the identity encoder to eliminate the information of eye movement from the identity feature. By learning the mapping between joint features and generated images, the architecture can not only generate videos with identity retention and mouth matching sound, but also accurately establish the relationship between blink action and blink score.

3.3 Facial Anatomy–Based Detail-Aware Talking Face Generation

Recently, Liu and Wang [48] proposed a two-stage, detail-aware talking face generation method via facial anatomy. The action units are used to enhance the relationship between the speech audio and facial animation details. The features of speech audio are mapped to action units so as to control face generation by manipulating the action units. Finally, audio-driven conversion is achieved to control action unit changes to generate talking faces. This method is able to produce more realistic talking face videos for arbitrary face with richer facial details (e.g., cheek motion and eyebrow movement) than the state-of-the-art methods.

3.4 Emotional Talking Face Generation

It is known that facial emotional plays a critical role in audio-visual communication. Besides lip movements, eye blinking, and other facial attributes, emotion is another big factor for talking face generation. Wang et al. [49] proposed MEAD (Multi-view Emotional Audio-visual Dataset), a large-scale dataset for emotional talking face generation. They also developed an emotional talking face generation baseline with explicit emotion control. Sinha et al. [50] presented a method for one-shot talking face generation with independent emotion control, that need only a single image of any arbitrary target person, a speech audio and an emotion vector as input. A graph convolutional network was used for facial detail (e.g., structure and geometry) preservation.

Eskimez et al. [51] proposed a talking face generation conditioned on categorical emotions in an end-to-end learning fashion, with a speech audio, target face image, and categorical emotion condition as input. The rendered talking face video can match with the speech audio and displaying the conditioned emotion. Recently, Ji et al. [52] realized one-shot emotional talking face generation via audio-based EAMM (Emotion-Aware Motion Model).

4 TALKING HEAD GENERATION

Talking face generation with head movement is also called talking head generation. Fig. 4 shows two talking head generations results of two given samples. It can be seen that, besides the vivid facial details, there are also vivid rhythmic head movements in the talking head generation results. According to different driving sources, talking head generation methods can be classified into video/image-driven, pose information-driven, and audio-driven methods.

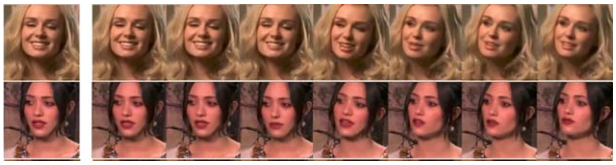

Fig. 4. Talking head generation results. In each row, the leftmost is the given face images and the right ones are the generation results [48].

4.1 Video/Image Driven Talking Head Generation

Video/image-driven talking head generation is also called face reenactment. The synthesized head pose is the same as that in the driven video or image. Video/image-driven talking head generation uses visual information to generate talking faces. The goal is to change faces, which belongs to image-to-image translation.

Wiles et al. [19] proposed a self-supervised network, i.e., X2Face, which uses embedded faces and driving vectors (e.g., video frames, pose vectors, and audio vectors) to generate a talking head video. Zakharov et al. [36] proposed a few-shot adversarial learning method to synthesize realistic talking head models with boundary conditions. After meta-learning on a large scale of talking head video dataset, this method performs few-shot and one-shot learning of talking head models of unseen persons in an adversarial training manner. It is noted that, because initializing the generator and the discriminator parameters in a person-specific way, this method is able to run using only several images.

However, this method does not adapt the boundary extracted from the source face to the head geometry of the target, resulting in the identity mismatch of the generated talking head video. Using the adaptive normalization of Zakharov et al. [36] or the deformation module of Wiles et al. [19] cannot maintain the target identity well. With the difference between the driver identity and the target identity, the identity loss is more serious.

When the target identity and the driver identity differs greatly, previous methods suffer from poor talking head video synthesis results. To this end, Ha et al. [37] proposed a network architecture called MarioNETte for high-quality talking head generation in few-shot scenarios. They introduced some components including image attention block, target feature alignment, and landmark transformer to overcome the problem of driver leaking due to identity mismatch. Moreover, by separating the expression geometry using landmark disentanglement, the identity preservation problem was overcome thanks to the landmark transformer.

To solve the temporal consistency problem of talking head generation, Koujan et al. [38] proposed a method called Head2Head, driven by 3D information, which makes use of the special structure of facial motion (e.g., mouth motion). This method greatly enhances the temporal consistency of the resulting facial reproduction, which can produce realistic results with vivid facial expression, eye gaze, and head pose. However, their method requires the use of long video clips of the target speaker to train the network. Recently, Doukas et al. [39] introduced an one-shot talking head synthesis method called HeadGAN. It synthesizes a talking head video based on 3D face representation, requiring only one reference image. This method can adapt to the geometric structure of faces from any source as it disentangles identity from expression.

Recently, talking head generation with Neural Radiance Fields (NeRFs) emerged. Given a set of 2D talking face image sets, instead of relying on extra intermediate representations (e.g., 2D landmarks or 3D face models), NeRFs directly reproduce novel views of a 3D talking head via a fully connected neural network, storing the information of 3D facial geometry and appearance in terms of voxel grids. Gafni et al. [40] introduced dynamic NeRFs to model a 4D facial avatar with a monocular facial video. This method can be used to synthesize both novel head poses and vivid facial expression changes. Guo et al. [41] proposed AD-NeRF, a method of audio-driven NeRFs model for talking head synthesis. Because of the ray dispatching strategy of NeRFs, the methods of talking face generation with NeRFs can generate high-quality facial details including teeth and hair, with great potential in practical applications.

4.2 Pose Information Driven Talking Head Generation

The pose information-driven talking head generation method performs head pose transformation according to head motion information, which needs to learn the expression of pose information and change the head posture through gesture information in existing speech videos.

Wiles et al. [19] used pose codes to control the head pose. When the pitch, yaw, and roll angles in the pose codes are changed, the head pose in the generated video frame will change accordingly. This method learns the distortion between pixels but does not define structural information (e.g., face landmarks, segmentation images, and 3D models). Burkov et al. [42] proposed a natural talking head reenactment system driven by latent pose descriptors based on the segmentation graph. As a part of the learning content of the whole reenactment system, the latent pose representation can be derived from the identity, and they can obtain a new identity through meta learning.

Zhou et al. [43] proposed a pose-controllable talking face generation framework through an implicitly modularized audio-visual representation. They designed a head pose code to obtain head motion information from additional video information, and they successfully added head motion to the talking face video results. Note that both speech content and head pose information are coded in a joint non-identity embedding space. In this way, a pose code can be learned in a modulated reconstruction framework, and the speech can be learned from the synchronization between audio and lips. This method overcomes the degradation problem caused by the inaccurate estimation information under extreme conditions. However, these pose information cannot capture eye movements, and even if additional image input is used, this method cannot effectively generate blink movements.

4.3 Audio-Driven Talking Head Generation

Audio-driven talking head generation method aims to produce personalized and naturally rhythmic talking faces with head movements based on the speech audio, which is a type of method most widely used but also the most difficult to implement. At present, most audio-driven talking head generation methods use facial landmarks or 3D face models to represent faces.

Taking audio as the only input, Zhou et al. [33] generated a talking head video from a single face image. They disentangle the content information and personalized information in the input speech audio signal, use facial landmarks as an intermediate representation to express the speaker's dynamics, and then use an image translation networks to transform facial landmarks into talking head video frames. Karras et al. [7] proposed an end-to-end data-driven learning method to directly infer the offset of vertex positions corresponding to facial expression changes from the input audio. In order to solve the influence of emotional changes on expression driven effect in the audio-driven process, the network automatically learns the hidden variables of emotional state from the data set.

Thies et al. [44] proposed Neural Voice Puppetry, an audio-driven facial reenactment network using a latent 3D talking head model space. This model learns temporal stability from the 3D representation in an implicit manner. This method can synthesize talking head videos for different people with the speech audio of any unknown source. Yi et al. [45] proposed a depth neural network model for the problem of head in-plane and out of plane rotation during head posture change. This method first reconstructs the 3D face model, then animates the face according to the parameters of speech learning, and finally re-renders it into a synthesized video frame. At the same time, they proposed a memory-augmented GAN module to fine tune these synthesized video frames into video frames with smooth background transition. It is reported that this method is able to yield talking face videos with more distinguishing head movement effects than the existing methods.

In a speech, the rhythmic head motion of a speaker conveys prosodic information. However, existing landmark or video frames guided talking head synthesis methods fail to control the head movements in the synthesized talking head video. To this end, Chen et al. [46] proposed an approach for talking head synthesis with controllable, rhythmic head motion. They decompose the head motion and facial motion, and use a 3D-aware generation network module to learn head motion and facial motion, respectively. A hybrid embedding module is used to dynamically aggregate the information of reference images, and a nonlinear synthesis module is designed to alleviate the visual discontinuity caused by head motion and facial motion.

It may lead to some problems using audio to drive facial landmarks and facial motion. First, facial landmarks are composed of key points, and only the motion information of face contour can be learned. The information of pixel changes will be lost, and the generated talking head will miss a lot of facial details. Secondly, it is difficult to generate an accurate face model from a single face image.

Recently, Chen et al. [47] provided a survey on the state-of-the-art talking-head generation methods. They also proposed new metrics for evaluating the emotional expression, semantic-level lip synchronization, and blink motion of a synthesized talking head video.

5 FUTURE DIRECTIONS

In recent years, talking face generation has made great progress and has been applied in many practical senarios such as virtual assistant, teleconferencing, and movie dubbing. However, it still has some limitations, which deserves further study in the future.

- **Dataset extension.** Because most of the currently published datasets are English videos of real faces, the effect of talking face videos generated from non-realistic faces (e.g., cartoon face images) and non-English speech audio is less satisfactory. How to expand the data set and bring more diverse training data plays a critical role in future talking face generation for more types of face images and different languages.
- **Large head movement cases.** Existing talking head synthesis methods would fail for some extreme cases (e.g., large angle change of head pose). It may be a feasible solution to combine the advantages of both facial landmarks and 3D face reconstruction to drive large head pose movement.
- **High-quality visual-audio synchronization.** The synchronization accuracy of audio and video in the generated results is an important indicator of the realism of a talking face video. Improving the synchronization of a generated video can further enhance its realism. Therefore, how to better learn stronger synchronization constraints from real videos to improve the synchronization accuracy of a generated talking face video remains a challenge.
- **More vivid facial details.** In addition to facial details and head movement, facial expression changes and eye movements can greatly enhance the realism of a talking face video. How to establish the relationship between speech audio and expression changes, eye movements, and body posture changes is a topic worthy of in-depth study, which will help introduce more vivid talking face details and can also benefit the study in the other research field (e.g., psychology).
- **Efficient talking face generation.** A lightweight talking face generation solution is very important for mobile application. Although the existing talking face generation methods can produce compelling results, there is still room for optimization of algorithm efficiency and model complexity.

- **Visual to sound.** The inverse problem of talking face generation is lip reading [57, 58]. The progress of talking face generation can provide a large number of training data for lip reading. It will also benefit the research of visual to sound [59, 60], i.e., predicting sound from a video.
- **Quantitative index.** Currently qualitative evaluation was widely used for measuring the quality of talking face generation results. The PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity) were also employed for objective evaluation of a taking face video frame. However, a quantitative index specified to talking face generation evaluation is still required.

6 CONCLUSION

This paper has provided a review on talking face generation methods in recent years. Firstly, the early traditional talking face generation approaches were reviewed. Secondly, the talking face generation methods based on deep learning were summarized and discussed. Then, the detail-aware talking face generation methods and the talking head generation methods were surveyed, respectively. The limitations of existing works and the possible research solution in the future were analyzed at last.

7 ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers for their valuable comments. This work has been funded in part by the Natural Science Foundation of China (project number 62072328). This template version was edited by Shiguang Liu.

8 REFERENCES

[1] F. Rumsey, “Audio in Multimodal Applications,” *J. Audio Eng. Soc.*, vol. 58, no. 3, pp. 191–195 (2010 Mar.).

[2] J. T. Mullin, “Synchronous Audio-Visual Display Techniques,” *J. Audio Eng. Soc.*, vol. 8, no. 4 pp. 236–243 (1960 Oct.).

[3] S. Agrawal, A. Simon, S. Bech, K. Bntsen, and S. Forchhammer, “Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences,” *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 404–417 (2020 Jun.). https://doi.org/10.17743/jaes.2020.0039.

[4] S. Agrawal, S. Bech, K. Brentsen, K. De Moor, and S. Forchhammer, “Method for Subjective Assessment of Immersion in Audiovisual Experiences,” *J. Audio Eng. Soc.*, vol. 69, no. 9, pp. 656–671 (2021 Sep.). https://doi.org/10.17743/jaes.2021.0013.

[5] S. Liu and D. Manocha, *Sound Synthesis, Propagation, and Rendering* (Springer, Cham, Switzerland, 2022). https://doi.org/10.1007/978-3-031-79214-4.

[6] S. Liu, H. Cheng, and Y. Tong, “Physically-Based Statistical Simulation of Rain Sound,” *ACM Trans. Graph.*, vol. 38, no. 4, paper 123 (2019 Jul.). doi: 10.1145/3306346.3323045.

[7] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, “Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion,” *ACM Trans. Graph.*, vol. 36, no. 4, paper 94 (2017 Jul.). doi: 10.1145/3072959.3073658.

[8] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, “Expressive Speech-Driven Facial Animation,” *ACM Trans. Graph.*, vol. 24, no. 4, pp. 1283–1302 (2005 Oct.). doi: 10.1145/1095878.1095881.

[9] L. Xie and Z. Liu, “Realistic Mouth-Synching for Speech-Driven Talking Face Using Articulatory Modelling,” *IEEE Trans. Multimed.*, 2007, vol. 9, no. 3, pp. 500–510 (2007 Apr.). https://doi.org/10.1109/TMM.2006.888009.

[10] L. Wang, X. Qian, W. Han, and F. K. Soong, “Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp. 446–449 (Makuhari, Japan) (2013 Sep.). https://doi.org/10.21437/Interspeech.2010-194.

[11] B. Fan, L. Wang, F. K. Soong, and L. Xie, “Photo-Real Talking Head With Deep Bidirectional LSTM,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4884–4888 (South Brisbane, Australia) (2015 Apr.). https://doi.org/10.1109/ICASSP.2015.7178899.

[12] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-Time Face Capture and Reenactment of RGB Videos,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395 (Las Vegas, NV) (2016 Jun.). https://doi.org/10.1109/cvpr.2016.262.

[13] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing Obama: Learning Lip Sync From Audio,” *ACM Trans. Graph.*, vol. 36, no. 4, paper 95 (2017 Jul.). https://doi.org/10.1145/3072959.3073640.

[14] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, “Talking Face Generation by Conditional Recurrent Adversarial Network,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 919–925 (Macao, China) (2019 Aug.). https://doi.org/10.24963/ijcai.2019/129.

[15] A. Jamaludin, J. Chung, and A. Zisserman, “You Said That?: Synthesising Talking Faces From Audio,” *Int. J. Comput. Vis.*, vol. 127, pp. 1767–1779 (2019 Feb.). https://doi.org/10.1007/s11263-019-01150-y.

[16] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, “Video Frame Synthesis Using Deep Voxel Flow,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 4473–4481 (Venice, Italy) (2017 Oct.). https://doi.org/10.1109/iccv.2017.478.

[17] C. Vondrick, H. Pirsiavash, and A. Torralba, “Generating Videos With Scene Dynamics,” in *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 613–621 (Barcelona, Spain) (2016 Dec.).

[18] M. Saito, E. Matsumoto, and S. Saito, “Temporal Generative Adversarial Nets With Singular Value Clipping,” in *Proceedings of IEEE International Conference on*

Computer Vision (ICCV), 2849–2858 (Venice, Italy) (2017 Oct.). https://doi.org/10.1109/iccv.2017.308.

[19] O. Wiles, A. S. Koepke, and A. Zisserman, "X2Face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 690–706 (Munich, Germany) (2018 Oct.). https://doi.org/10.1007/978-3-030-01261-8_41.

[20] Y. Zhou, Z. Xu, C. Landreth, et al., "VisemeNet: Audio-Driven Animator-Centric Speech Animation," *ACM Trans. Graph.*, vol. 37, no. 4, paper 161 (2018 Jul.). https://doi.org/10.1145/3197517.3201292.

[21] S. Taylor, T. Kim, Y. Yue, et al., "A Deep Learning Approach for Generalized Speech Animation," *ACM Trans. Graph.*, vol. 36, no. 4, paper 93 (2017 Jul.). https://doi.org/10.1145/3072959.3073699.

[22] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-Driven 3D Facial Animation With Implicit Emotional Awareness: A Deep Learning Approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2328–2336 (Honolulu, HI) (2017 Jul.). https://doi.org/10.1109/CVPRW.2017.287.

[23] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, "A Lip Sync Expert is All You Need for Speech to Lip Generation in the Wild," in *Proceedings of ACM Multimedia*, pp. 484–492 (Seattle, WA) (2020 Oct.). https://doi.org/10.1145/3394171.3413532.

[24] X. Huang, M. Wang, and M. Gong, "Fine-Grained Talking Face Generation With Video Re-Interpretation," *Vis. Comput.*, vol. 37, no. 1 pp. 95–105 (2020 Sep.). https://doi.org/10.1007/s00371-020-01982-7.

[25] M. Yu, J. Yu, M. Li, and Q. Ling, "Multi-Modal Inputs Driven Talking Face Generation With Spatial-Temporal Dependency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 203–216 (2021 Feb.). https://doi.org/10.1109/TCSVT.2020.2973374.

[26] S. Eskimez, R. Maddox, C. Xu, and Z. Duan, "End-to-End Generation of Talking Faces From Noisy Speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1948–1952 (Barcelona, Spain) (2020 Apr.). https://doi.org/10.1109/ICASSP40776.2020.9054103.

[27] P. Garrido, L. Valgaerts, H. Sarmadi, et al,, "VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 193–204 (2015 Jun.). https://doi.org/10.1111/cgf.12552.

[28] H. Zhou, Y. Liu, Z. Liu, P. Luo, abd X. Wang, "Talking Face Generation by Adversarially Disentangled Audio-Visual Representation," in *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 9299–9306 (Honolulu, HI) (2019 Jul.). https://doi.org/10.1609/aaai.v33i01.33019299.

[29] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7824–7833 (Long Beach, CA) (2019 Jan.). https://doi.org/10.1109/CVPR.2019.00802.

[30] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic Speech-Driven Facial Animation With GANs," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1398–1413 (2019 Oct.). https://doi.org/10.1007/s11263-019-01251-8.

[31] S. Sinha, S. Biswas, and B. Bhowmick, "Identity-Preserving Realistic Talking Face Generation," in *Proceedings of International Joint Conference on Neural Networks*, pp. 1–10 (Glasgow, UK) (2020 Jul.). https://doi.org/10.1109/IJCNN48605.2020.9206665.

[32] D. Zeng, H. Liu, H.-C. Lin, and S. Ge, "Talking Face Generation With Expression-Tailored Generative Adversarial Network," In *Proceedings of ACM Multimedia*, pp. 1716–1724 (Seattle, WA) (2020 Oct.). https://doi.org/10.1145/3394171.3413844.

[33] Y. Zhou, X. Han, E. Shechtman, et al., "MakeItTalk: Speaker-Aware Talking Head Animation," *ACM Trans. Graph.*, vol. 39, no. 6, paper 221 (2020 Nov.). https://doi.org/10.1145/3414685.3417774.

[34] J. Hao, S. Liu, and Q. Xu, "Controlling Eye Blink for Talking Face Generation via Eye Conversion," in *Proceedings of SIGGRAPH Asia Technical Communications*, paper 1 (Tokyo, Japan) (2021 Dec.). https://doi.org/10.1145/3478512.3488610.

[35] S. Liu, and J. Hao. "Generating Talking Face With Controllable Eye Movements by Disentangled Blinking Feature," *IEEE Trans. Vis. Comput. Graph.*, vol. 29 (2023 Aug.). https://doi.org/10.1109/TVCG.2022.3199412.

[36] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9458–9467 (Seoul, South Korea) (2019 Oct.). https://doi.org/10.1109/ICCV.2019.00955.

[37] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "MarioNETte: Few-Shot Face Reenactment Preserving Identity of Unseen Targets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 10893–10900 (New York, NY) (2020 Apr.). https://doi.org/10.1609/aaai.v34i07.6721.

[38] M. R. Koujan, M. C. Doukas, A. Roussos, and S. Zafeiriou, "Head2head: Video-based Neural Head Synthesis," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 16–23 (Buenos Aires, Argentina) (2020 Nov.). https://doi.org/10.1109/FG47880.2020.00048.

[39] M. C. Doukas, S. Zafeiriou, and V. Sharmanska, "HeadGAN: One-Shot Neural Head Synthesis and Editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14398–14407 (Montreal, Canada) (2021 Oct.). https://doi.org/10.1109/ICCV48922.2021.01413.

[40] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8645–8654 (Nashville, TN) (2021 Jun.). https://doi.org/10.1109/CVPR46437.2021.00854.

[41] Y. Guo, K. Chen, S. Liang, Y. Liu, H. Bao, and J. Zhang, "AD-NeRF: Audio Driven Neural Radiance

Fields for Talking Head Synthesis," in *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5784–5794 (Montreal, Canada) (2021 Oct.). https://doi.org/10.1109/ICCV48922.2021.00573.

[42] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky, "Neural Head Reenactment With Latent Pose Descriptors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13783–13792 (Seattle, WA) (2020 Jun.). https://doi.org/10.1109/CVPR42600.2020.01380.

[43] H. Zhou, Y. Sun, W. Wu, et al., "Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4176–4186 (Nashville, TN) (2021 Nov.). https://doi.org/10.1109/CVPR46437.2021.00416.

[44] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural Voice Puppetry: Audio-Driven Facial Reenactment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 716–731 (Glasgow, UK) (2020 Aug.). https://doi.org/10.1007/978-3-030-58517-4_42.

[45] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven Talking Face Video Generation With Learning-Based Personalized Head Pose," a*rXiv preprint arXiv: 2002.10137* (2020 May). https://doi.org/10.48550/arXiv.2002.10137.

[46] L. Chen, G. Cui, C. Liu, et al., "Talking-Head Generation With Rhythmic Head Motion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–51 (Glasgow, UK) (2020 Aug.). https://doi.org/10.1007/978-3-030-58545-7_3.

[47] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What Comprises a Good Talking-Head Video Generation?: A Survey and Benchmark," *arXiv preprint arXiv: 2005.03201* (2021 May). https://doi.org/10.48550/arXiv.2005.03201.

[48] S. Liu, and H. Wang, "Talking Face Generation via Facial Anatomy," *ACM Trans. Multimed. Comput., Commun., Appl.*, vol. 19, pp. 1–19 (2023 Feb.). https://doi.org/10.1145/3571746.

[49] K. Wang, Q. Wu, L. Song, et al., "Mead: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 700–717 (Glasgow, UK) (2020 Aug.). https://doi.org/10.1007/978-3-030-58589-1_42.

[50] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick, "Emotion-Controllable Generalized Talking Face Generation," in *Proceedings of 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1320–1327 (Vienna, Austria) (2022 Jul.). https://doi.org/10.24963/ijcai.2022/184.

[51] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech Driven Talking Face Generation From a Single Image and an Emotion Condition," *IEEE Trans. Multimed.*, vol. 24, pp. 3480–3490 (2021 Jul.). https://doi.org/10.1109/TMM.2021.3099900.

[52] X. Ji, H. Zhou, K. Wang, et al., "EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model," in *Proceedings of ACM SIGGRAPH*, paper 61 (Vancouver, Canada) (2022 Jul.). https://doi.org/10.1145/3528233.3530745.

[53] J. S. Chung, and A. Zisserman, "Out of Time: Automated Lip Sync in the Wild," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 251–263 (Taipei, Taiwan) (2017 Mar.). https://doi.org/10.1007/978-3-319-54427-4_19.

[54] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip Movements Generation at a Glance," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 520–535 (Munich, Germany) (2018 Oct.). https://doi.org/10.1007/978-3-030-01234-2_32.

[55] K. R. Prajwal, R. Mukhopadhyay, J. Philip, et al., "Towards Automatic Face-to-Face Translation," in *Proceedings of the ACM International Conference on Multimedia*, pp. 1428–1436 (New York, NY) (2019 Oct.). https://doi.org/10.1145/3343031.3351066.

[56] H. Zhu, H. Huang, Y. Li, A. Zheng, and R. He, "Arbitrary Talking Face Generation via Attentional Audio-Visual Coherence Learning," *arXiv preprint arXiv:1812.06589* (2018 Dec.). https://doi.org/10.48550/arXiv.1812.06589.

[57] J. S. Chung, and A. Zisserman, "Lip Reading in the Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 87–103 (Honolulu, HI) (2017 Jul.). https://doi.org/10.1109/CVPR.2017.367.

[58] W. Chen, X. Tan, Y. Xia, et al., "DualLip: A System for Joint Lip Reading and Generation," in *Proceedings of the ACM International Conference on Multimedia*, pp. 1985–1993 (Seattle, WA) (2020 Oct.). https://doi.org/10.1145/3394171.3413623.

[59] A. Owens, P. Isola, J. McDermott, et al., "Visually Indicated Sounds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2405–2413 (Las Vegas, NV) (2016 Jun.). https://doi.org/10.1109/CVPR.2016.264.

[60] S. Liu, S. Li, and H. Cheng, "Towards an End-to-End Visual-to-Raw-Audio Generation With GANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. pp. 1299–1312 (2022 Mar.). https://doi.org/10.1109/TCSVT.2021.3079897.

[61] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251 (Venice, Italy) (2017 Oct.). https://doi.org/10.1109/ICCV.2017.244.

THE AUTHOR

Shiguang Liu

Shiguang Liu is currently a professor with the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, P.R. China. He received a Ph.D. from the State Key Lab of CAD & CG, Zhejiang University, P.R. China. He was a visiting scholar at Michigan State University from 2010 to 2011. He was also a research associate professor at CUHK and KAIST in 2012 and 2013, respectively. His research interests include audio-visual learning, multimedia, virtual reality, etc. He has published more than 100 peer-reviewed papers in ACM TOG, IEEE TVCG, IEEE TMM, IEEE TCSVT, ACM TOMM, IEEE VR, ACM MM, IEEE ICME, etc. Dr. Liu is a Senior Member of IEEE and a Member of AES.