Photo-Realistic Talking Face Generation Under Latent Space Manipulation

Ridwan Salahudeen, Wan-Chi Siu, *Life Fellow, IEEE*, and H. Anthony Chan, *Life Fellow, IEEE*

***Abstract*—This paper focuses on generating photo-realistic talking face videos by leveraging on semantic facial attributes in a latent space and capturing the talking style from an old video of a speaker. We formulate a process to manipulate facial attributes in the latent space by identifying semantic facial directions. We develop a deep learning pipeline to learn the correlation between the audio and the corresponding video frames from a reference video of a speaker in an aligned latent space. This correlation is used to navigate a static face image into frames of a talking face video, which is moderated by three carefully constructed loss functions, for accurate lip synchronization and photo-realistic video reconstruction. By combining these techniques, we aim to generate high-quality talking face videos that are visually realistic and synchronized with the provided audio input. Our results were evaluated against some state-of-the-art techniques on talking face generation, and we have recorded significant improvements in the image quality of the generated talking face video.**

***Index Terms*—Talking face generation, latent space, deep learning, multimedia applications.**

I. Introduction

IMAGINE a scenario where you are watching the news, and the anchor's face suddenly morphs into your favourite celebrity, delivering the news in their unique style. With customizable media, this could be a reality [1]. You could choose the faces that appear on your TV screen, making the news more engaging and relevant to your interests. Similarly, imagine you have an important online video conference, but you are feeling under the weather and do not want to get dressed up. With customizable media, you could present yourself as a well-dressed professional without ever leaving your bed. This technology could revolutionize the way we conduct virtual meetings, making them more comfortable, convenient, and efficient.

These scenarios are not just about convenience. They are also essential in situations where there is limited network bandwidth. By minimizing the transmitted packets to just audio signals, we can reduce the load on the network while still maintaining high-quality communication [2], [3]. This means that even in low-bandwidth situations, we can still have engaging and meaningful conversations, without sacrificing the quality of communication.

To address this issue, deep learning technology is used to learn the correlation between speech sequences and their corresponding video frames. Hence, static face photos can be animated into talking face in a TV report or in a virtual meeting, using the speaker's speech. To have a talking face with realistic head pose, eye blink, etc. We aim to capture a personalized talking style of speakers by using their old video to learn the correlation between their speech segments and their visual facial styles in an aligned latent space. The training videos is pre-process by separately extracting the audio signals (in segments) and the video frames, and ensuring the audio segments matches the video frames by using appropriate Fourier Transform's parameters and Frame Per Second (FPS) respectively. They are then mapped into their individual latent spaces but ensuring they both have the same dimensions of 512 to learn the desired correlation. Finally, our pipeline uses a fine-tuned StyleGAN[4] generator to generate the talking face video frames using the summed latent codes of the audio embedding, the facial styles (detected from consecutive video frames) and the reference static image, as detailed in Section III. The approach is compared with several standard methods, and its effectiveness is assessed using multiple quantitative metrics. Our specific contributions are summarized as follows:

1) We formulate an algorithm to get semantic facial attributes in a latent space. These attributes could be mouth direction in the latent, such that it can be added to any arbitrary latent code (representing a particular face) to give a new latent code that will generate the same face with the mouth open or closed based on the magnitude of the manipulation.
2) We utilize the old video of a speaker to capture their talking style by learning the correlation between their speech segments and visual facial styles in an aligned latent space.
3) We propose an end-to-end deep learning pipeline to detect motion directions from consecutive frames of a reference video in the form of motion latent variables, and use the variables to navigate the latent code of a static face image into frames of a talking face video.
4) We device three loss functions (synchronization, reconstruction and contrastive losses) as the learning schemes

Received 30 May 2024; revised 10 October 2024; accepted 29 November 2024. Date of publication 11 December 2024; date of current version 12 June 2025. This work was supported in part by the Saint Francis University under Grant ISG200206; in part by UGC, Hong Kong, SAR, under Grant UGC/IDS(C)11/E01/20; and in part by The Hong Kong Polytechnic University. *(Corresponding author: Wan-Chi Siu.)*

Ridwan Salahudeen and H. Anthony Chan are with the School of Computing and Information Sciences, Saint Francis University, Hong Kong, China (e-mail: rsalahudeen@cihe.edu.hk; hhchan@sfu.edu.hk).

Wan-Chi Siu is with the Department of Electrical and Electronics Engineering, The Hong Kong Polytechnic University, Hong Kong, China, and also with the School of Computing and Information Sciences, Saint Francis University, Hong Kong, China (e-mail: wan.chi.siu@polyu.edu.hk).

Digital Object Identifier 10.1109/TCE.2024.3516387

that handles accurate lip synchronization using the audio input and the video reference for photo-realistic video reconstruction.

The remainder of this article is arranged as follow: prior works related to talking face generation are discussed in Section II, while we detailed our proposed method in Section III of the paper. The experimental work and the results obtained were discussed in Section IV and V before concluding the paper in Section VI.

II. Related Works

Talking Face Generation is a hot topic in the field of Artificial Intelligence that aimed to synthesis realistic talking face video of a static image corresponding to a driving modality such as audio, video or text. In this section, we review prior works relating to this task. We particularly look at works that focus on lip synchronization, latent code manipulation, audio-driven talking face generation and video-driven talking face generation.

A. Lip Sync Video Generation

The advances in deep learning is pushing the boundary for Generative AI and has prompted a lot of research interest in learning the lip-sync correlation between audio segments and corresponding video frames. Most studies for lip-sync learning are GAN-based models focusing on audio-visual correlation without any intermediation. In 2019, Jamaludin et al. [5] laid the groundwork for this area of research in the form of encoder-decoder network to learn the joint embedding of audio and face representations. In the same year, Song et al. [6] introduced a lip-reading discriminator to guide lip motion generation. In 2020, Wav2Lip [7] demonstrates that a lip synchronization discriminator is most valuable in a talking face generation while Vougioukas et al. [8] adopted an autoregressive temporal GAN in achieving more coherent sequence generation in a talking face. However, in order to achieve realistic talking face generation, more facial dynamics such as eye blink and head pose need to be incorporated. We therefore aimed to learn these facial dynamics from a reference video of a speaker.

B. Latent Code Manipulation

Generative Adversarial Networks (GANs) have emerged as a ground-breaking deep learning technique that has revolutionized the field of generative modeling. The power of GANs lies in their ability to generate realistic images, which are almost indistinguishable from real ones. StyleGAN [4] is a type of GAN that has gained significant attention in recent years due to its ability to generate highly detailed and realistic images with unprecedented control over the style and content of the generated images. One of the unique features of StyleGAN is its ability to allow for precise manipulation of the latent code, which controls the style of the generated image. The latent code is essentially a vector in a high-dimensional space that controls various aspects of the generated image, such as the pose, expression, color, and texture. By manipulating the values of the latent code, users can generate images with different styles, expressions, and features. Previous works from Shen et al. [9] proposed a model to interpret latent semantics learned by GANs for semantic face editing called InterFaceGAN. The works of Image2StyleGAN [10] and Tov et al. [11] addresses how real images can be inverted into StyleGAN latent space for possible semantic manipulations. This concept can be incorporated in a talking face generation pipeline so as to allow a seamless manipulation of some semantics such as head pose and facial expression.

C. Audio-Driven Talking Face Generation

The task of lip syncing, as explored in studies such as [8], [12], [13], [14], involves using audio input to drive the animation of a static face into a talking face video, ensuring that the audio sounds are perfectly synchronized with the visual lip movements in the generated video. This task is highly challenging due to multiple requirements that need to be met, including maintaining high visual quality, accommodating multiple speakers, achieving lip synchronization with unseen speakers, and accounting for dynamic facial expressions. To address the complex challenges of lip syncing, many studies have decomposed the problem into a two-stage reconstruction process, utilizing intermediary representations such as facial landmark prediction (as seen in MakeItTalk [14]) or 3D pose parameters (as seen in Zhou et al. [12]). While these are effective to some extent, these intermediary representations have limitations in capturing the dynamic facial gestures of different speakers in an end-to-end learning framework. To overcome these limitations, our approach is to utilize an aligned latent space of audio and video representations to learn these dynamic facial gestures. By leveraging this aligned latent space, we can directly learn the complex mapping between audio and visual information without the need for intermediate representations.

D. Video-Driven Talking Face Generation

The study conducted by Siarohin et al. 2019 [15] demonstrated the possibility of extracting motion flows from a video and applying them to warp a static image, thereby generating a video with impressive visual quality. This approach has inspired several subsequent studies, including [16], [17], [18], [19] that have adopted similar principles. For instance, Tripathy et al. [16] leveraged Action Units (AU) and pose angles obtained from a driving video to facilitate face re-enactment in a Generative Adversarial Network (GAN) setting. Their framework comprises of a two-stage model that operates in a self-supervised manner. On the other hand, Siarohin et al. 2021 [17] explored animating objects consisting of different parts using an unsupervised approach by inferring motions from a driving video on the key regions of the object. Wang et al. [19] contextualizes this approach in video conferencing application by simply using a driving video to animate a source image. Most recently, DaGAN [18] introduces a depth-aware attention network to provide a dense 3D facial geometry used to guide the generation of motion fields. Part of our approach builds upon this idea by using motion flows and styles from a driving video to learn motion dynamics in

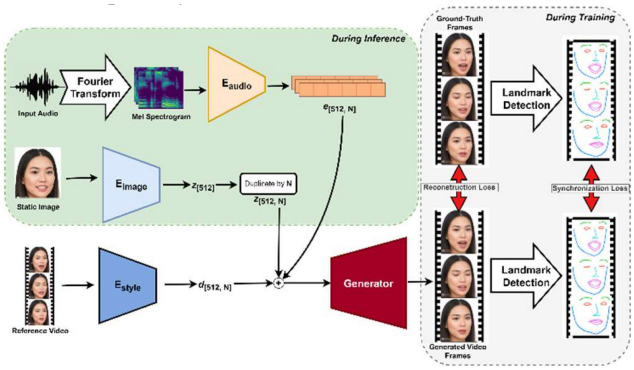

Fig. 1. Our proposed model consists of three (3) encoders: audio encoder E_{audio}, image encoder E_{image} and style encoder E_{style}. On the left-hand side, the first row is the E_{audio} encoder for input audio. The second row is the E_{image} encoder for inputting static image whereas the last row is E_{style} encoder for reference video input. On the right-hand side, comparison is done between the ground truth and generated video frames.

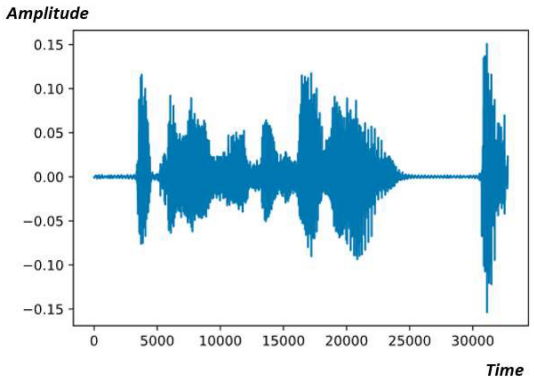

Fig. 2. Audio signal representation in a waveform.

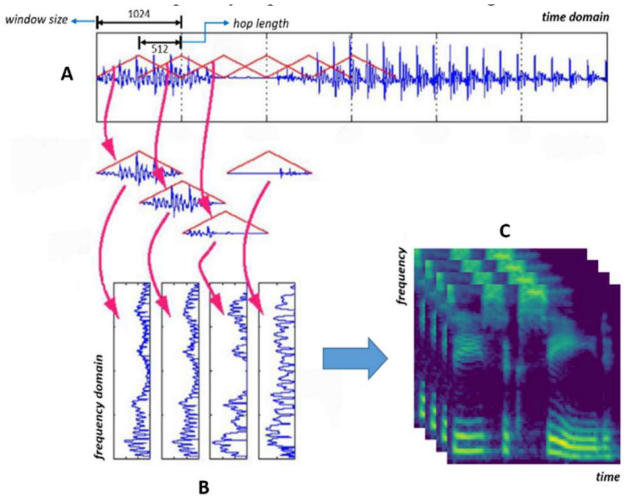

Fig. 3. Audio segmentation procedure. (A) The original audio in time-domain wave signal. (B) Cutting the original wave signal into overlapping segments and transforming them to frequency domain using Short-Time Fourier Transform. (C) Transforming each segment into mel-spectrogram, which is a frequency-time domain visual representation of the original signal.

talking face generation, such as head pose, eye blinks [20], and emotions [21].

III. Proposed Method

We presents an innovative deep learning approach for generating photo-realistic talking face videos from static face photos. In this section, we describe the steps taken to achieve this goal, including the use of speaker's speech to learn lip synchronization and video reference to learn other facial dynamics such as head pose, eye blink, etc.

The pipeline of our proposed architecture (Fig. 1) takes three inputs: an audio wave, a static image, and a reference video. The audio is segmented and transformed into mel-spectrograms using Fourier Transform. Each segment is then encoded by the audio encoder E_{audio} into a 512-dimensional latent code, resulting in N segments of 512 dimensions each. The image encoder, E_{image}, encodes the static image into a 512-dimensional latent code, which is duplicated N times. The style encoder E_{style} encodes each frame of the reference video into N 512-dimensional latent codes. The latent codes from the three encoders are summed element-wise for each frame and passed to a StyleGAN generator to produce the talking face video frames. During training, the generated frames are compared with the ground truth to measure reconstruction loss, and the detected facial landmarks of both the generated and ground truth frames are compared in the synchronization loss.

A. Problem Formulation

Given a static image $\boldsymbol{I}$, an audio wave sequence $\boldsymbol{A} = \{\mathbf{a}_i$, for $i = 1, 2, 3, \ldots$ to $L\}$ of acoustic features (which could be the audio part of a sample video), and a style reference sequence $\boldsymbol{R} = \{\mathbf{r}_i$, for $i = 1, 2, 3, \ldots$ to $N\}$ video frames, our method aims to generate a realistic talking face video $\boldsymbol{V} = \{\mathbf{v}_i$, for $i = 1, 2, 3, \ldots$ to $N\}$ with lip-synced frames reflecting the speaker's talking style.

B. Fourier Transformation

The audio input to our proposed pipeline are wave signals $\boldsymbol{A} = \{\mathbf{a}_i$, for $i = 1, 2, 3, \ldots$ to $L\}$ representing amplitudes along a time domain as shown in Fig. 2. These are continuous values showing how the amplitude of the sound wave changes over time. Hence, it is called a waveform or time-domain representation of a sound signal.

To make this waveform useful for a deep learning model, we need to transform it from its time-domain representation to a visual representation of its frequency content over time known as mel-spectrogram shown in bottom-right of Fig. 3.

To achieve this, we have to perform Short-Time Fourier Transform (STFT) as in Equation (1), by firstly segmenting the audio waveform into frames using a window size (w) and a hop (or overlapping) length (l) such as 1024 and 512 bits respectively for w and l as illustrated in Fig. 3.

$$\boldsymbol{A}_i = \{\boldsymbol{a}_i \text{ for } i = l - w \text{ to } l + w\} \tag{1}$$

This gives us an audio segment $\boldsymbol{A}_i$ with index i. This will form N segments, for $\{i = 1, 2, 3, \ldots$ to N$\}$, where N $= \lfloor$(L$-w + l)/l\rfloor + 1$. The frequency content over time, that is, the mel-spectrogram of each segment i is then computed using the Fourier Transform to obtain time-frequency representation of the signals.

C. Audio Encoder

The mel-spectrograms from the previous step are 2D image representations of the audio signal that provides the distribution of energy in the signal across the frequency scale over time in just 1D image channel, in contrast to RGB images. We make use of Convolutional Neural Network (CNN), Equation (2), to embed them into vectors of 512 dimension to align them with StyleGAN latent space.

$$\boldsymbol{e} = E_{audio}(\boldsymbol{A}_i) \tag{2}$$

D. Image Encoder

We intend to encode any given static image containing human face into StyleGAN latent space by generating a 512 dimensional vector representing the latent code for the given image, Equation (3).

$$z = E_{image}(\boldsymbol{I}) \tag{3}$$

To achieve this, two pre-trained networks (StyleGAN generator and vgg-16) were used to do the image-to-latent code inversion. Initially, we can pass a random Gaussian vector of 512 dimension to the StyleGAN generator to produce an image. We then pass both the generated image and the target image to the vgg-16. The feature maps at the last convolutional layer of the vgg-16 for both images are extracted and compared. This comparison tells us how similar the generated image is compared with the target image. The latent code is then updated and passed to the StyleGAN generator again, this process is repeated for a number of steps, usually 1000 steps are sufficient to get an optimized latent code representation of our target image.

E. Style Encoder

A video reference is passed into the model as indicated at the bottom-left of Fig. 1, through an encoder, Equation (4), which down-samples the pixel maps of the individual frames of the video using CNN.

$$\boldsymbol{d} = E_{style}(\boldsymbol{R}) \tag{4}$$

The style transition between consecutive frames make up the motion latent variables that can be learnt using Long Short-Time Memory (LSTM) network. The outputs of the LSTM for each frame is then mapped to 512 dimension using Multi-Layer Perceptron (MLP) in order to align it to the same embedding as the audio and static image inputs.

F. Generator

The generator shown at the bottom-middle of Fig. 1 is a pre-trained StyleGAN generator capable of producing images when Gaussian vector of 512 dimension is passed to it. The generator, Equation (6), receives the summed latent code of the aligned vectors of: 1) the audio embedding; 2) the static image, which is duplicated to N number of identified frames; and 3) the reference video motion latent per frame. The latent codes of these inputs are summed to give a new latent code with the same dimension as the three inputs.

$$\boldsymbol{m} = \boldsymbol{e} + z + \boldsymbol{d} \tag{5}$$

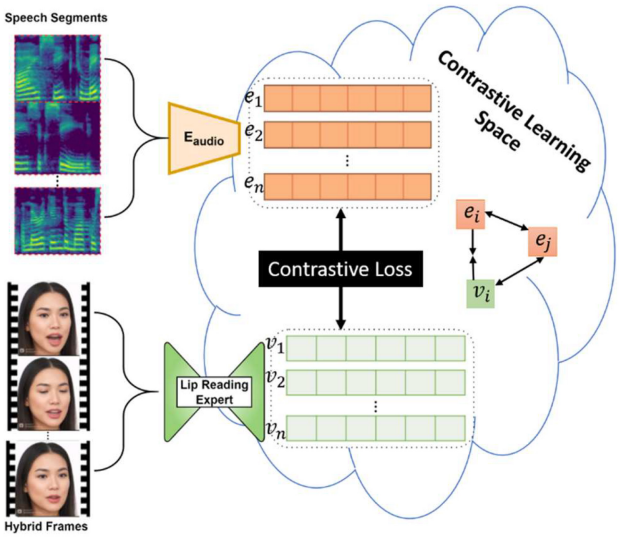

Fig. 4. Contrastive Learning to align the embedded audio segment e_i with the embedded corresponding video frame v_i.

$$\boldsymbol{V} = g(\boldsymbol{m}) \tag{6}$$

The generator can then be fine-tuned to generate realistic video frames corresponding to the identity of the static image as well as in synchronization to both the audio's lip-sync and the reference video's facial attributes. To further enhanced the lip to speech synchronization of the generator, we make use of a contrastive learning framework [22] that leverages a pre-trained lip reading expert model. This expert model provides a strong prior on the relationship between lip movements within the video frames and their corresponding speech segments. Let us elaborate the contrastive learning approach as shown below.

G. Contrastive Learning for Enhanced Lip Synchronization

This method aims to align the audio embedding with their corresponding visual context features, ensuring accurate synchronization between the generated video and the input audio. The process is as follows:

1. Lip Reading Expert: We have to employ a pre-trained lip reading model [23] to encode both the ground truth and generated frames. This model is capable of extracting detailed visual features related to lip movements and facial expressions.

2. Hybrid Frame Encoding: To create a robust learning environment, we randomly replace some frames from the generated sequence with ground truth frames. This results in a hybrid sequence of frames that includes both generated and ground truth images. These hybrid frames are then encoded using the lip reading expert [23], producing a set of visual embedding.

3. Audio Encoding: The audio input is segmented and transformed into mel-spectrograms using Fourier Transform. Each segment is encoded by the audio encoder (E_{audio}) into a 512-dimensional latent code, capturing the acoustic features of the speech.

4. Projection into Contrastive Learning Space: The visual embedding from the hybrid frames and the audio embedding from the audio encoder are projected into a shared contrastive

Fig. 5. Sample images of mouth close and mouth open pairs for two individuals from the processed VoxCeleb2 videos.

learning space. In this space, a contrastive loss function is used to optimize the alignment of the embedding.

In Fig. 4 the audio segment represented by $\boldsymbol{e}_i$ and its corresponding video frame represented by $\boldsymbol{v}_i$ are projected into the contrastive learning space. During training, a contrastive loss is used to minimize the distance between the audio embedding and their corresponding visual context features (as positive pairs, $\boldsymbol{e}_i$ and $\boldsymbol{v}_i$), and maximize the distance between the audio embedding and non-corresponding visual features (as negative pairs, $\boldsymbol{e}_i$ and $\boldsymbol{v}_j$). Note that positive means the audio segment matches the video frame whereas negative means otherwise.

IV. Experimental Work: Latent Direction Identification

Using VoxCeleb2 dataset, we borrowed inspirations from [9], [10], [24] to carefully select pairs of video frames corresponding to the same person with two extreme ends of a facial semantic. For instance, a video frame of mouth open and another video frame of mouth close for same individual as shown in Fig. 5. We manually observed the identified frames of mouth close/open and selected those which other features of the face appear most similar except the mouth. Those with the most similar appearance but alternating mouth open and close were selected as pairs. The same approach was used to create pairs of eye and head movement semantics.

After manually collecting these frames, using 118 different videos from the test set of VoxCeleb2 dataset, we grouped the frames separately for mouth open and mouth close into folders and named the frames with the ID of its original video from VoxCeleb2 with a description of "mouth_open" or "mouth_close" respectively.

For each ID, we embedded the frame for "mouth_open" and "mouth_close" separately into StyleGAN latent space using the following algorithm:

a. Detect a face in the given image. That is, given an image of any dimension, we use a facial landmark detector to detect the rectangular coordinates (x_1, x_2, y_1, y_2) of any available face(s) in the image.
b. Align the face to fit the dataset of StyleGAN. The detected face is up-sampled to fit the 1024x1024 dimension of FFHQ dataset used in training StyleGAN.
c. Project the resulting image to StyleGAN latent space as specified in Alaluf et al. [25]. The StyleGAN latent dimension is 512, so to map our 1024x1024 aligned image to 512, we start with a random Gaussian vector of 512 dimension. We pass the vector to a pre-trained StyleGAN generator and compared the generated image with the aligned image by passing both images into a pre-trained vgg-16 network. We have to measure the perceptual error between the feature maps at the last convolutional layer of the vgg-16 network for the two images, we then use the loss value to update the latent code. We repeat this optimization process for 1000 steps in order to get an optimal latent code representing the given video frames.
d. Take the differences between the optimized latent codes of mouth opened and mouth closed frames for each individual in the collection.
e. Take the average of the differences of all the individual pairs to give a generalized latent code representing mouth-open direction in the StyleGAN latent space.

The same process is repeated to obtain eye-blink and head movement latent directions.

A. Experimental Result

We used the identified latent directions to manipulate the latent codes of some sample faces used in our experiment as seen in Fig. 6. The manipulation was achieved, as formulated in Equation (7), by adding to the latent code of a particular face, the latent code representing a direction (e.g., mouth direction) magnified with a scalar value, indicating the extend of movement along that direction.

$$z_{edit} = z_s + \alpha \cdot \boldsymbol{d} \tag{7}$$

where the original latent code representing a face is denoted as z_s, a direction latent code is denoted as $\boldsymbol{d}$ and thescalar α represents the magnitude to which we want to move along a specified direction. The magnitude $\alpha$is multiplied with the direction vector element-wise to have the same dimension of 512. Then the corresponding element-wise values of the original face latent code is added to the magnified direction latent code to give a new latent code z_{edit} which is then passed to pre-trained StyleGAN generator to produce an image of the same identity as z_s but with a substantial difference in its facial attribute along that direction.

B. Training Details

Our model was trained in an unsupervised manner. We used the generated video frames produced by the generator to compare with the ground truth video frames of VoxCeleb2 videos used in our training. The training frames were sampled at 25 fps from VoxCeleb2 training set whereas the mel-spectrogram were obtained from the extracted audio of corresponding VoxCeleb2 videos using 16KHz as sampling rate and FFT length of 1024, hop length of 256 and frequency bin of 80.

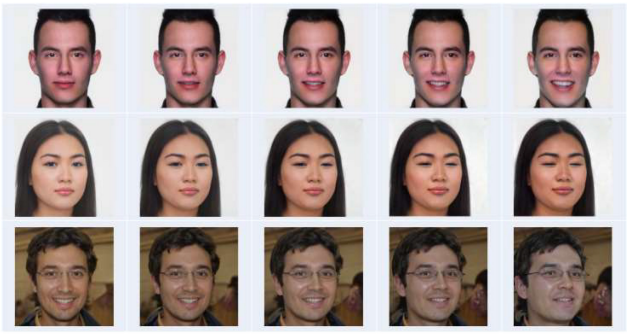

Fig. 6. Experimental result of image manipulation in a StyleGAN latent space. Three different human faces were manipulated along mouth direction (top images), eye direction (middle images) and head direction (bottom images).

We used three loss functions to adjust the reconstruction and lip synchronization errors between generated and ground truth frames.

C. Reconstruction Loss

During the training as indicated in Fig. 1, we learn to reconstruct the same reference video frames by using a combination of perceptual and adversarial losses to guide the generation. For the perceptual loss, we passed the corresponding frames of the generated and ground truth videos to a vgg-16 network to extract their respective feature maps at the last convolutional layer of the network. This layer is a fair balance between high-level and low-level feature representation of the frames. We then used a L2 loss (Mean Square Error) to compare the corresponding values of the feature maps as shown in Equation (8).

$$\mathcal{L}_{perc}(v_t, \widehat{v}_t) = \sum_{t=1}^{T} \|v_t - \widehat{v}_t\|_2 \tag{8}$$

where v_t and $\widehat{v}_t$ are the ground truth and generated video frames respectively at time t. T is total number of frames.

To further guide the image realism of the generated video frames, we used adversarial loss to evaluate the generated frames as either real or fake. We implemented a discriminator network trained on FFHQ dataset, capable of panelising generated frames with blurry image quality. Equation (9) shows the negative log-likelihood loss function used for the discriminator network to correctly classify the generated frames as either real or fake.

$$\mathcal{L}_{adv}(\widehat{v}_t) = \sum_{t=1}^{T} -\log(D(\widehat{v}_t)) \tag{9}$$

where $\widehat{v}_t$ is the generated video frame at time t being passed to the discriminator network D. T is total number of frames.

D. Synchronization Loss

Learning the lip synchronization requires focusing on more concise portion of the face such as the lip, the chin, and parts of the jaw. Instead of focusing on the entire pixels of the generated video frame to learn the lip synchronization, we use a landmark detector, as shown in Fig. 1, to extract just the 68 facial key points from the frames. These points are compared between corresponding generated and ground truth video frames using L1 loss (Mean Absolute Error) to minimize the lip sync error of the generated video frames as shown in Equation (10).

$$\mathcal{L}_{sync}(f_t, \widehat{f_t}) = \sum_{t=1}^{T} \sum_{i=1}^{68} \|f_{i,t} - \widehat{f}_{i,t}\|_1 \tag{10}$$

where $f_{i,t}$ and $\widehat{f}_{i,t}$ represent the facial point i from the 68 points for frame t of the ground truth and generated frames respectively.

E. Contrastive Loss

The contrastive loss function is designed to attract audio embedding and their time-aligned visual context features while repelling the audio embedding from non-corresponding frames. This ensures that the audio embedding are closely aligned with the correct visual features, enhancing the synchronization between lip movements and speech.

Let $\boldsymbol{e}_i$ be the audio embedding for i-th segment, and $\boldsymbol{v}_i$ be the corresponding visual embedding. Similarly, let $\boldsymbol{v}_j$ be a visual embedding that does not correspond to $\boldsymbol{e}_i$. The contrastive loss is constructed as shown in Equation (11):

$$\mathcal{L}_{cont}(e,\ v, m) = \sum_{i=1}^{N} \frac{y_i \|e_i - v_i\|^2 + (1 - y_i)\max\left(0,\ m - \|e_i - v_j\|\right)^2}{2} \tag{11}$$

where y_i is a binary label indicating whether the pair is positive ($y_i = 1$) or negative ($y_i = 0$) and m is a margin parameter that defines the minimum distance between negative pairs, it acts like a threshold encouraging the model to push negative pairs apart, ensuring a clear separation between similar and dissimilar data points.

The combination of these three loss functions Perceptual Loss $\mathcal{L}_{perc}$, Adversarial Loss $\mathcal{L}_{adv}$, Synchronization Loss $\mathcal{L}_{sync}$ and Contrastive Loss $\mathcal{L}_{cont}$ with their respective scales $\lambda_{perc}, \lambda_{adv}, \lambda_{sync}, \lambda_{cont}$ gives the overall loss of our model as shown below:

$$\mathcal{L} = \lambda_{perc} \cdot \mathcal{L}_{perc} + \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{sync} \cdot \mathcal{L}_{sync} + \lambda_{cont} \cdot \mathcal{L}_{cont} \tag{12}$$

Our training implementation was conducted using PyTorch framework with the Adam optimizer, an initial learning rate of $1e^{-4}$ and weight decay of $1e^{-6}$. The training was done using two GPU card of 24GB each with a batch size of 32. It took 7 days to train the pipeline for 200 epochs.

V. Result Discussion

We evaluated our implementation on some quantitative metrics that are commonly used in the field of talking face generation. For the visual qualities of the generated video frames, we used Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Fréchet Inception Distance (FID). SSIM measures the luminance, contrast and structure of the generated frame in comparison with the ground-truth. PSNR measures the quality of a reconstructed image compared to its original image and

TABLE I
COMPARISON WITH SOME STATE-OF-THE-ART MODELS

Models	SSIM↑	PSNR↑	LMD↓	LSE-D↓	FID↓
MakeItTalk	0.53	17.27	1.92	10.60	53.95
PC-AVS	0.64	19.63	2.23	7.03	32.78
Audio2Head	0.62	18.19	2.71	9.04	39.86
Wav2Lip	0.83	24.55	**1.10**	**6.30**	61.58
TalkLip	0.79	25.50	1.18	6.66	-
IP-LAP	0.58	19.70	1.22	6.73	-
VideoReTalking	0.62	24.73	1.17	9.04	-
Ours	**0.85**	**25.70**	1.28	6.89	**28.35**

FID evaluates the quality of generated images in generative models.

For facial landmark-related measure, we used Landmark Distance (LMD), which measures the error distance between the corresponding landmarks of the generated and ground-truth frames. We also used Lip-Sync Error Distance (LSE-D), which evaluates the quality of lip synchronization in generated videos or animations of talking faces.

We compared the performance of our method with some State-of-The-Art models in talking face generation including MakeItTalk[14], PC-AVS [12], Audio2Head [13], Wav2Lip [7], TalkLip [26], IP-LAP [27], and VideoReTalking [28].

As shown in Table I, our approach generates high-quality talking head videos, significantly outperforming existing methods in overall image quality as evidenced by having superior SSIM and PSNR scores and a lower FID score. These indicate better image fidelity, less distortion, and higher diversity. Furthermore, our approach achieves comparable performance to state-of-the-art methods in facial landmark alignment, as measured by LMD and LSE-D scores. These improvements are attributed to our novel approach of leveraging semantic facial attributes within the latent space for accurate facial feature reconstruction and an end-to-end pipeline effectively integrating audio, image, and style information. The resulting videos exhibit clearer quality, frames with more detailed and high structural fidelity.

While our method demonstrates strong lip synchronization, some existing methods, such as Wav2Lip, exhibit slightly better performance on LMD and LSE-D. This difference likely stems from the fact that these methods prioritize lip synchronization above other aspects of video generation, employing architectures and training strategies heavily optimized for this specific task. Our method, in contrast, adopts a more holistic approach, balancing multiple objectives to achieve superior performance across a broader range of metrics. This leads to a potential trade-off where the slightly less optimized lip synchronization allows for superior image quality and expression realism. Additionally, within our approach, the fusion of audio and style embeddings onto the static image's latent code, and potential discrepancies between the video frame rate (25 fps) and the audio processing (mel-spectrogram frame rate), may also hinder optimal lip synchronization. Future work will focus on refining the model to further improve lip synchronization, investigating alternative integration strategies for audio and style information, and aligning the audio and video pre-processing for improved temporal consistency. This will address the identified trade-off and further enhancement of the overall performance of our approach.

TABLE II
RESULTS OF ABLATION STUDIES

Configuration	*SSIM↑*	*PSNR↑*	*LMD↓*	*LSE-D↓*	*FID↓*
Baseline	**0.85**	**25.70**	**1.28**	**6.89**	**28.35**
w/o LipSync	*0.67*	*19.75*	*1.50*	*7.51*	*28.63*
w/o VideoRef	*0.71*	*20.70*	*1.38*	*7.04*	*30.62*

A. Ablation Study

To evaluate the contribution of each learning scheme, we performed ablation studies by systematically removing or altering components of our model and observing the impact on performance. Specifically, we have focused on the following configurations:

1. Baseline: The complete model with both learning schemes (lip synchronization and talking style learning).
2. Without Lip Synchronization (w/o LipSync): The model without the lip synchronization learning scheme, focusing solely on learning the talking style from video reference.
3. Without Video Referencing (w/o VideoRef): The model without the talking style learning scheme, focusing solely on lip synchronization from audio input.

The results of ablation experience are shown in Table II.

The results clearly demonstrate the importance of both learning schemes. The baseline model, which includes both schemes, achieves the best performance across all metrics. Removing the lip synchronization scheme results in a significant increase in LSE-D, indicating poorer lip synchronization. Similarly, removing the talking style learning scheme leads to a decrease in SSIM and PSNR, indicating lower image quality and less realistic facial dynamics. The talking style learning scheme is crucial for capturing the unique talking style of a speaker, including head nods, eye blinks, and other dynamic facial movements. This is achieved by learning the latent directions from motion flow in the video reference, which allows us to generate a realistic talking head that accurately reflects the speaker's individual style.

These ablation studies confirm that both learning schemes are essential for achieving accurate lip synchronization and photo-realistic talking head videos.

Note that an early version (with 2-pages) of this work has been presented and published in IEEE International Conference on Consumer Electronics-Taiwan (ICCE2023-Taiwan) in July 2023, which is cited in this submission as [29]. For convenience to readers some more demonstrations are available with website: https://cis.sfu.edu.hk/2024.TalkingFace/

VI. CONCLUSION

In this research, we have presented a comprehensive approach to generate photo-realistic talking face videos with accurate lip synchronization and realistic visual styles. Our algorithm enables the manipulation of semantic facial

attributes in the latent space, facilitating the generation of open or closed mouth states in a controlled manner. By leveraging the correlation between speech segments and visual facial styles, we capture the talking style of a speaker from a reference video. Additionally, our deep learning pipeline utilizes motion latent variables to navigate a static face image into frames of a talking face video, ensuring smooth and coherent motions.

The combination of these contributions allows for the generation of high-quality and visually convincing talking face videos. The results demonstrate the potential of our approach in producing realistic facial animations that are synchronized with speech, with outstanding SSIM, LMD and FID result compared of the state-of-the-art. The ability to manipulate facial attributes, capture talking styles, and reconstruct videos holds promise for various applications, including entertainment such as forming commentators for automatic football commentary as it was done in Siu et al. [30], virtual avatars, and virtual communication. Future work can explore further improvements in fine-grained attribute manipulation, enhanced synchronization techniques, and the integration of additional cues to generate even more realistic and expressive talking face videos. It appears very promising if some refinement techniques and a suitable image super-resolution network can be designed to enhance video quality. However, this has to be on the condition that the overall complexity should not be significantly increased, since we are working on real-time applications.

REFERENCES

[1] S. S. Sundar, "Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII)," *J. Comput. Mediat. Commun.*, vol. 25, no. 1, pp. 74–88, Mar. 2020, doi: 10.1093/jcmc/zmz026.

[2] X. Sun, M. Wang, R. Lin, Y. Sun, and S. S. Cheng, "Deep-learned perceptual quality control for intelligent video communication," *IEEE Trans. Consum. Electron.*, vol. 68, no. 4, pp. 354–365, Nov. 2022, doi: 10.1109/TCE.2022.3206114.

[3] "Latency vs bandwidth: Unraveling video conferencing quality secrets." Accessed: Jul. 31, 2023. [Online]. Available: https://www.digitalsamba.com/blog/understanding-latency-and-bandwidth-unveiling-the-key-differences

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021, doi: 10.1109/TPAMI.2020.2970919.

[5] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that?: Synthesising talking faces from audio," *Int. J. Comput. Vis.*, vol. 127, nos. 11–12, pp. 1767–1779, 2019, doi: 10.1007/s11263-019-01150-y.

[6] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 919–925, doi: 10.24963/ijcai.2019/129.

[7] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 484–492, doi: 10.1145/3394171.3413532.

[8] 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, 2020, doi: 10.1007/s11263-019-01251-8.

[9] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9240–9249, doi: 10.1109/CVPR42600.2020.00926.

[10] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4431–4440, doi: 10.1109/ICCV.2019.00453.

[11] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for StyleGAN image manipulation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–14, Jul. 2021, doi: 10.1145/3450626.3459838.

[12] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4174–4184, doi: 10.1109/CVPR46437.2021.00416.

[13] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2Head: Audio-driven one-shot talking-head generation with natural head motion," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2021, pp. 1098–1105, doi: 10.24963/ijcai.2021/152.

[14] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-aware talking-head animation," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, 2020, doi: 10.1145/3414685.3417774.

[15] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–11. [Online]. Available: https://arxiv.org/abs/2003.00196v3

[16] S. Tripathy, J. Kannala, and E. Rahtu, "ICface: Interpretable and controllable face reenactment using GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 3374–3383, doi: 10.1109/WACV45572.2020.9093474.

[17] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13648–13657, doi: 10.1109/CVPR46437.2021.01344.

[18] F. T. Hong, L. Zhang, L. Shen, and D. Xu, "Depth-aware generative adversarial network for talking head video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3387–3396, doi: 10.1109/CVPR52688.2022.00339.

[19] T. C. Wang, A. Mallya, and M. Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10034–10044, doi: 10.1109/CVPR46437.2021.00991.

[20] J. Lemley, A. Kar, A. Drimbarean, and P. Corcoran, "Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems," *IEEE Trans. Consum. Electron.*, vol. 65, no. 2, pp. 179–187, May 2019, doi: 10.1109/TCE.2019.2899869.

[21] G. A. Prabhakar, B. Basel, A. Dutta, and C. V. R. Rao, "Multichannel CNN-BLSTM architecture for speech emotion recognition system by fusion of magnitude and phase spectral features using DCCA for consumer applications," *IEEE Trans. Consum. Electron.*, vol. 69, no. 2, pp. 226–235, May 2023, doi: 10.1109/TCE.2023.3236972.

[22] W. Yu and S. Zhao, "Hybrid Contrastive Learning with attention mechanism for unsupervised person re-identification," in *Proc. Int. Conf. Image, Vis. Intell. Syst.*, 2024, pp. 444–454, doi: 10.1007/978-981-97-0855-0_42.

[23] B. Shi, W. N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–24. [Online]. Available: https://arxiv.org/abs/2201.02184v2

[24] Y. Viazovetskyi, V. Ivashkin, and E. Kashin, "StyleGAN2 distillation for feed-forward image manipulation," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 170–186, doi: 10.1007/978-3-030-58542-6_11.

[25] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano, "HyperStyle: StyleGAN inversion with HyperNetworks for real image editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18490–18500, doi: 10.1109/CVPR52688.2022.01796.

[26] J. Wang, X. Qian, M. Zhang, R. T. Tan, and H. Li, "Seeing what you said: Talking face generation guided by a lip reading expert," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14653–14662, doi: 10.1109/CVPR52729.2023.01408.

[27] W. Zhong et al., "Identity-preserving talking face generation with landmark and appearance priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9729–9738, doi: 10.1109/CVPR52729.2023.00938.

[28] K. Cheng et al., "VideoReTalking: Audio-based lip synchronization for talking head video editing in the wild," in *Proc. SIGGRAPH Asia Conf. Papers*, 2022, pp. 1–9, doi: 10.1145/3550469.3555399.

[29] R. Salahudeen, W.-C. Siu, and H. A. Chan, "Activate your face in virtual meeting platform," in *Proc. Int. Conf. Consum. Electron.*, 2023, pp. 791–792, doi: 10.1109/ICCE-TAIWAN58799.2023.10227004.

[30] W.-C. Siu et al., "On the completion of automatic football game commentary system with deep learning," in *Proc. Int. Workshop Adv. Imag. Technol. (IWAIT)*, 2023, Art. no. 125920H, doi: 10.1117/12.2668398.

Ridwan Salahudeen received the B.Sc. and M.Sc. degrees in computer science from Ahmadu Bello University (ABU), Zaria, Nigeria, in 2011 and 2016, respectively, and the Ph.D. degree from the Wuhan University of Technology, China. He serves as a Teaching Staff with ABU in 2020. He is currently a Research Associate with the Artificial Intelligence Laboratory, Saint Francis University, Hong Kong. His research interest includes deep learning, blockchain technology, knowledge representation, and uncertainty resolution.

Wan-Chi Siu (Life Fellow, IEEE) received the M.Phil. degree from The Chinese University of Hong Kong in 1977, and the Ph.D. degree from Imperial College London in 1984. He is currently an Emeritus Professor (formerly the Chair Professor, the HoD(EIE) and the Dean of Engineering Faculty) with The Hong Kong Polytechnic University and Research Professor with St. Francis University, Hong Kong. He was a Vice President, the Chair of Conference Board and core member of Board of Governors of the IEEE SP Society from 2012 to 2014, and the President of APSIPA from 2017 to 2018. He is an Outstanding Scholar with many awards, including the Distinguished Presenter Award, the Best Teacher Award, the Best Faculty Researcher Award (twice) and the IEEE Third Millennium Medal in 2000. He was an APSIPA Distinguished Lecturer from 2021 to 2022, and an Advisor & Distinguished Scientist of the European research project SmartEN (offered by European Commissions). He has been a Keynote Speaker and an Invited Speaker of many conferences, published over 500 research papers (200 appeared in international journals such as IEEE TRANSACTIONS ON IMAGE PROCESSING) in DSP, transforms, fast algorithms, machine learning, deep learning, super-resolution imaging, 2D/3D video coding, object recognition and tracking, and organized IEEE society-sponsored flagship conferences as a TPC Chair of ISCAS1997 and the General Chair of ICASSP2003 and ICIP2010. He was an Independent Non-Executive Director from 2000 to 2015 of a publicly-listed video surveillance company and chaired the First Engineering/IT Panel of the RAE, Hong Kong, from 1992 to 1993. He has been a Guest Editor/Subject Editor/Associate Editor for IEEE Transactions on CAS, IP & CSVT, and Electronics Letters. Recently, he has been a member of the IEEE Educational Activities Board, the IEEE Fourier Award for Signal Processing Committee from 2017 to 2020, the Hong Kong RGC Engineering-JRS Panel from 2020 to 2026, the Hong Kong ASTRI Tech Review Panel from 2006 to 2024, and some other IEEE technical committees.

H. Anthony Chan (Life Fellow, IEEE) received the B.Sc. degree from the University of Hong Kong, the M.Phil. degree from the Chinese University of Hong Kong, and the Ph.D. degree from the University of Maryland in 1982 and then continued basic research there in areas of experimental superconductivity and gravitation. He moved to an industry environment as he joined the Former with AT&T Bell Labs in 1986. As AT&T went through divestiture and shifted emphasis towards business needs, His work moved to practical areas of interconnection and assembly in manufacturing. He focused on hardware reliability from 1991 to 1996, and had become the key reliability expert for numerous design and manufacturing organizations throughout Lucent Technologies, which had separated from AT&T. He then transferred back to AT&T and changed field again to network. While it was another major change from his hardware responsibilities at Lucent Technologies to software and system engineering responsibilities at AT&T, the changes continued from telecommunication to wireless and IP data network. Recently he has joined as the Dean with the School of Computing and Information Sciences, Saint Francis University (formerly call Caritas Institute of Higher Education), Hong Kong. His recent research direction has also focused on artificial intelligent, particularly emphasizing federated learning and deep learning applied to image processing and distributed communications.