Audio-Driven Emotional Video Portraits

Xinya Ji[1]   Hang Zhou[2]   Kaisiyuan Wang[3]   Wayne Wu[4,5*]   Chen Change Loy[5]
Xun Cao[1*]   Feng Xu[6*]
[1]Nanjing University,   [2]The Chinese University of Hong Kong,
[3]The University of Sydney,   [4]SenseTime Research,
[5]S-Lab, Nanyang Technological University,   [6]BNRist and school of software, Tsinghua University

{xinya@smail., caoxun@}nju.edu.cn, zhouhang@link.cuhk.edu.hk, ccloy@ntu.edu.sg,

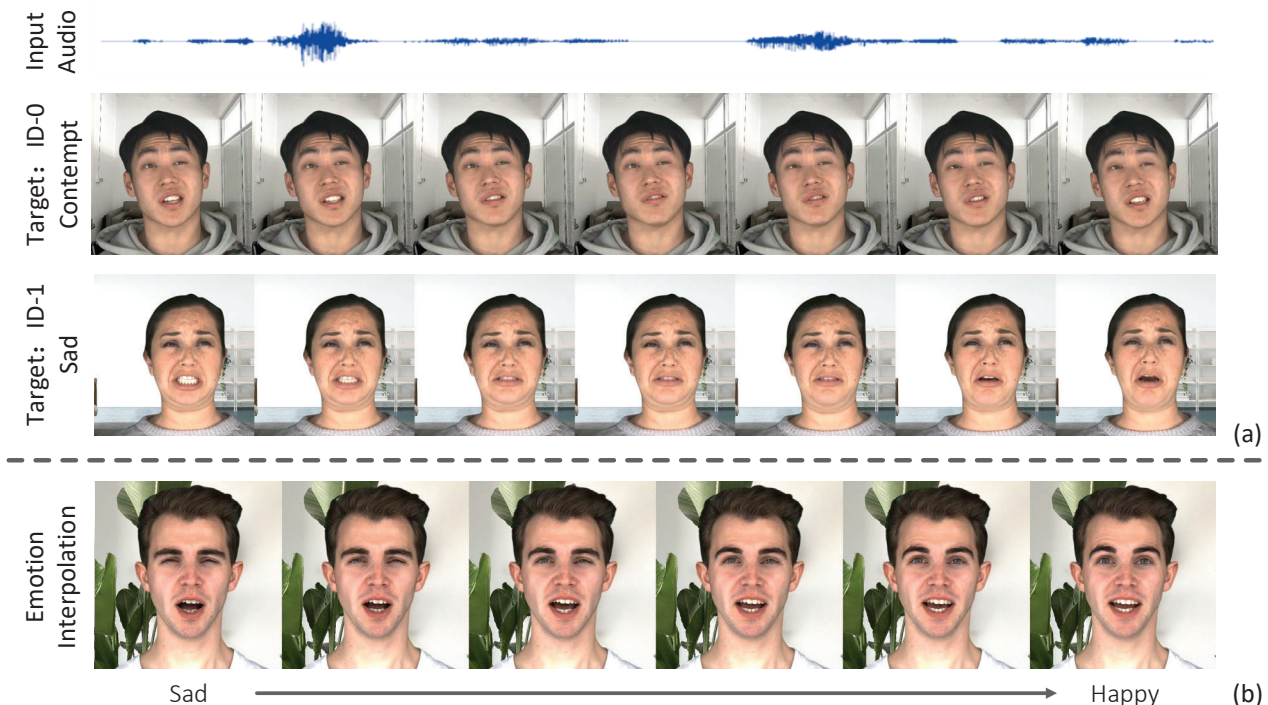kaisiyuan.wang@sydney.edu.au, wuwenyan@sensetime.com, xufeng2003@gmail.com

Figure 1: **Audio-Driven Emotional Video Portraits.** Given an audio clip and a target video, our Emotional Video Portraits (EVP) approach is capable of generating emotion-controllable talking portraits and change the emotion of them smoothly by interpolating at the latent space. (a) Generated video portraits with the same speech content but different emotions (*i.e.*, contempt and sad). (b) Linear interpolation of the learned latent representation of emotions from sad to happy.

Abstract

Despite previous success in generating audio-driven talking heads, most of the previous studies focus on the correlation between speech content and the mouth shape. Facial emotion, which is one of the most important features on natural human faces, is always neglected in their methods.

In this work, we present Emotional Video Portraits (EVP), a system for synthesizing high-quality video portraits with vivid emotional dynamics driven by audios. Specifically, we propose the Cross-Reconstructed Emotion Disentanglement technique to decompose speech into two decoupled spaces, i.e., a duration-independent emotion space and a duration-dependent content space. With the disentangled features, dynamic 2D emotional facial landmarks can be deduced. Then we propose the Target-Adaptive Face Synthesis tech-

---

*Corresponding authors.

nique to generate the final high-quality video portraits, by bridging the gap between the deduced landmarks and the natural head poses of target videos. Extensive experiments demonstrate the effectiveness of our method both qualitatively and quantitatively.[1]

1. Introduction

Generating audio-driven photo-realistic portrait video is of great need to multimedia applications, such as film-making [22], telepresence [2] and digital human animation [25, 14, 47]. Previous works have explored generating talking heads or portraits whose lip movements are synced with the input speech contents. Generally, these techniques can be divided into two categories: 1) *image-based* methods that animate one or few frames of cropped faces [11, 44, 33, 9, 29, 46, 7], and 2) *video-based editing* methods that directly edit target video clips [34, 32, 35]. Nevertheless, most of the previous studies did not model *emotion*, a key factor for the naturalism of portraits.

Only few *image-based* works have discussed emotional information in talking head generation. Due to the lack of appropriate audio-visual datasets with emotional annotations, Vougioukas *et al.* [36] do not model emotions explicitly. Simply encoding emotion and audio content information into a single feature, they produce preliminary results with low quality. Most recently, Wang *et al.* [37] collect the MEAD dataset, which contains high-quality talking head videos with annotations of both emotion category and intensity. Then they set emotion as an one-hot condition to control the generated faces. However, all of these *image-based* methods render only minor head movements with fixed or even no backgrounds, making them impractical in most real-world scenarios.

Whereas, *video-based editing* methods, which are more applicable as discussed in [34, 15, 32, 35], have not considered *emotion* control. Most of them only edit the mouth and keep the upper half of the video portraits unaltered, making free emotion control unaccessible.

In this study, we propose a novel algorithm named **Emotional Video Portraits** (**EVP**), aiming to endow the *video-based editing* talking face generation with the ability of *emotion* control from audio. We animate full portrait with emotion dynamics that better matches the speech intonation, leading to more vivid results. However, it is *non-trivial* to achieve this. There exist several intricate challenges: 1) The extraction of emotion from audio is rather difficult, since the emotion information is stickily entangled with other factors like the speech content. 2) The blending of the edited face and the target video is difficult while synthesizing high fidelity results. Audio does not supply any

cues for head poses and the global movements of a head, thus the edited head inferred from audio may have large head pose and movement variances with the target videos.

To tackle the challenges mentioned above, we manage to achieve audio-based emotion control in the proposed Emotional Video Portraits system with two key components, namely *Cross-Reconstructed Emotion Disentanglement*, and *Target-Adaptive Face Synthesis*. To perform emotion control on the generated portraits, we firstly propose the *Cross-Reconstructed Emotion Disentanglement* technique on audios to extract two separate latent spaces: i) a duration-independent space, which is a content-agnostic encoding of the emotion; ii) a duration-dependent space, which encodes the audio's speech content. Once extracted, features from these latent spaces are recombined to yield a new audio representation, allowing a cross-reconstruction loss to be computed and optimized. However, to enable the cross-reconstructed training, paired sentences with the same content but different emotions at the same length should be provided. This is nearly unreachable in real-world scenarios. To this end, we adopt Dynamic Time Warping (DTW) [3], a classic algorithm in time series analysis, to help form pseudo training pairs with aligned uneven-length speech corpus.

Following previous methods [34, 9], an audio-to-landmark animation module is then introduced with the decomposed features to deduce emotional 2D landmark dynamics. As no pose information is provided in audio, there is a gap to be bridged between the generated landmarks and the large variances of head pose and movement in target video. To this end, we propose the *Target-Adaptive Face Synthesis* technique to bridge the pose gap between the inferred landmarks and the target video portraits in 3D space. With a carefully designed 3D-aware keypoint alignment algorithm, we are able to project 2D landmarks into the target video. Finally, we train an Edge-to-Video translation network to generate the final high-quality emotional video portraits. Extensive experiments demonstrate the superior performance of our method and the effectiveness of several key components.

Our contributions are summarized as follows:

- We propose the Emotional Video Portraits (EVP) system, which is the first attempt to achieve emotional control in video-based editing talking face generation methods.

- We introduce Cross-Reconstructed Emotion Disentanglement technique, to distill content-agnostic emotion features for free control.

- We introduce Target-Adaptive Face Synthesis, to synthesize high quality portrait by making the generated face adapt to the target video with natural head poses and movements.

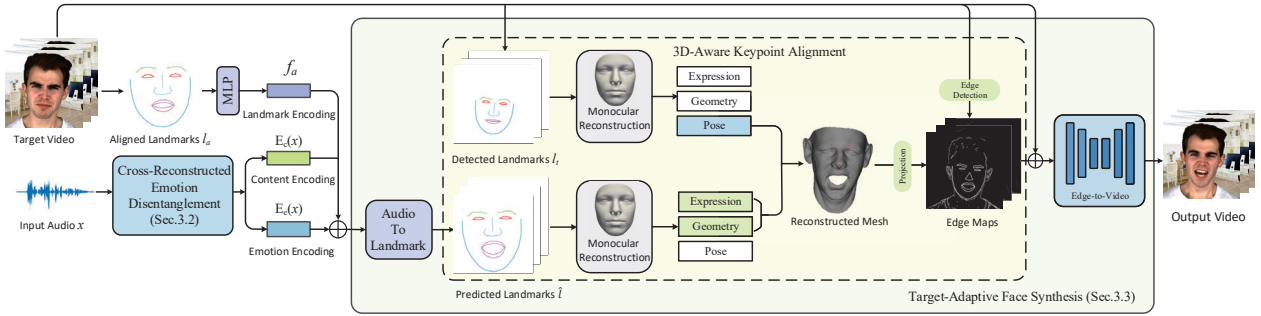[1]All materials are available at https://jixinya.github.io/projects/evp/.

Figure 2: **Overview of our *Emotional Video Portrait* algorithm.** We first extract disentangled content and emotion information from the audio signal. Then we predict landmark motion from audio representations. The predicted motion is transferred to the edge map of the target video via a 3D-aware keypoint alignment module. Finally, the rendering network gives us photo-realistic animations of the target portrait based on the target video and edge maps.

2. Related Work

2.1. Audio-Driven Talking Face Generation

The task of audio-driven talking-head generation aims at synthesizing lip-synced videos of a speaking person driven by audio. It is a topic of high demand in the field of entertainment, thus has long been the research interest in the area of computer vision and computer graphics [4, 21, 11, 38, 34, 44, 33, 48, 8, 9, 35, 29, 46, 37, 6, 45]. We can divide these methods into two categories according to the differences in the visualization of their results.

Image-Based Methods. One type of models focuses on driving the cropped facial areas with one or more frames as the identity reference. Chung *et al.* [11], for the first time, propose to generate lip-synced videos in an image-to-image translation [19] manner. Then Zhou *et al.* [44] and Song *et al.* [33] improve their results using disentangled audio-visual representation and recurrent neural networks. Moreover, Chen *et al.* [9] leverage landmarks as intermediate representation and split the process into two stages. However, these methods can only promise the synchronization between generated mouths and audios. The results have barely any expression or head movements. Zhou *et al.* [46] successfully generate identity-related head movements, but their model also fails to control emotions.

As for emotional talking faces, Vougioukas *et al.* [36] adopt three separated discriminators to enhance synthesis details, synchronization, and realistic expressions, respectively. However, their experiments are carried out on limited scales. Most recently, Wang *et al.* [37] propose the MEAD dataset and propose to generate emotional talking faces by splitting the manipulation for the upper and lower part of the face, respectively. Nevertheless, their results are less realistic and limited to only the facial areas.

Video-Based Editing Methods. Full-frame video portraits contain not only the facial areas but also the neck and the shoulder part of the person, together with the background. It is without doubt that this setting is more realistic, but more difficult to reconstruct. As a result, most methods edit only the mouth areas. Suwajanakorn *et al.*[34] synthesize photo-realistic talking videos of Obama by training an audio to landmark RNN. A re-timing module is proposed for head-poses. Song *et al.*[32] and Thies *et al.*[35] all regress facial expression parameters of 3DMM models, and inpaint the mouth regions. While high-quality results can be rendered through these pipelines using the videos of a target subject, it is difficult for their models to manipulate the upper face, left alone emotions. In this work, we propose to generate emotional manipulable full-frame talking-heads.

2.2. Conditional Emotion Generation

Inspired by the great success of unsupervised image translation[49, 19, 24, 18, 10, 20], several methods focusing on emotion conditioned generation have been proposed in recent years. Ding *et al.* [13] design a novel encoder-decoder architecture to control expression intensity continuously by learning an expressive and compact expression code. Pumarola *et al.* [31] introduce an unsupervised framework named GANimation, which is able to generate continuous facial expressions of a specified emotion category by activating the action units (AU) to various states. However, unsatisfying artifacts have always been a challenging problem for these methods due to the lack of explicit and accurate guidance. Inspired by [9], our method also chooses facial landmarks as a more reliable intermediary to generate talking face sequences with high-fidelity emotions.

3. Method

3.1. Overview

As shown in Fig. 2, our *Emotional Video Portrait (EVP)* algorithm consists of two key components. The first is *Cross-Reconstructed Emotion Disentanglement* that learns the disentangled content and emotion information from the
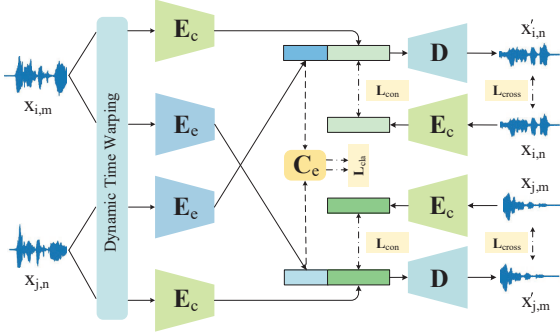
Figure 3: **Cross-reconstruction for disentanglement.** The emotion and content representations extracted from different audio signals are combined to reconstruct corresponding samples. Part of the training losses are also shown on this figure.

audio signals. We use a temporal alignment algorithm, Dynamic Time Warping [3] to produce pseudo training pairs, and then design a Cross-Reconstructed Loss for learning the disentanglement (Sec. 3.2). The second part of our algorithm is *Target-Adaptive Face Synthesis* that adapts the facial landmarks inferred from the audio representations to the target video. We design a 3D-Aware Keypoint Alignment algorithm to rotate landmarks in 3D space, and thus the landmarks can be adaptive to various poses and motions in the target video. Then, we subsequently use the the projected 2D landmarks as the guidance to edit the target video via an Edge-to-Video network(Sec. 3.3). In the following sections, we describe each module of our algorithm in detail.

3.2. Cross-Reconstructed Emotion Disentanglement

To achieve audio-based emotional control for talking face synthesis, the inherently entangled emotion and content components need to be independently extracted from audio signals. Unlike previous methods [37], which learn one single representation from audio signals, we propose to extract two separate latent audio spaces: i) a duration-independent space, which is a content-agnostic encoding of the emotion and ii) a duration-dependent space, which encodes the audio's speech content. While cross-reconstruction [1] technique seems promising on such a task, it can only be enabled through paired audio clips with the same content but different emotions at the same length. Nevertheless, this is nearly unreachable in real-world scenarios. To this end, we firstly build aligned *pseudo training pairs*, and then adopt the *cross-reconstructed training* for emotion disentanglement in audios.

Build Pseudo Training Pairs. An audio-visual dataset [37] with various characters speaking the same corpus under different emotion states is leveraged to train this disentanglement network. Since the speeches with the same content

but different emotions vary in speech rate, we resort to a temporal alignment algorithm to align the uneven-length speeches.

Specifically, we use Mel Frequency Cepstral Coefficients (MFCC) [26] as audio representations and use the Dynamic Timing Warping (DTW) [3] algorithm to warp the MFCC feature vectors by stretching or shrinking them along the time dimension. Given two MFCC sequences S_a and S_b of the same content but different lengths, DTW calculates a set of index coordinate pairs $\{(i, j), ...\}$ by dynamic programming to force $S_a[i]$ and $S_b[j]$ to be similar. The optimal match between the given sequences is achieved by minimizing the sum of a distance cost between the aligned MFCC features:

$$\min \sum_{(i,j) \in P} d(S_a[i], S_b[j]), \tag{1}$$

where d is the distance cost, P is the path for alignment. The path constraint is that, at (i, j), the valid steps are $(i + 1, j)$, $(i, j + 1)$, and $(i + 1, j + 1)$, making sure that the alignment always moves forward each time for at least one of the signals. These aligned audio samples can then be used as the inputs to the disentanglement network for cross-reconstruction.

Cross-Reconstructed Training. The cross-reconstructed training procedure is as shown in Fig. 3. To independently extract the emotion and content information lie in an audio clip $x_{i,m}$ with content i and emotion m, two encoders E_c and E_e are leveraged for embedding the two information respectively. Intuitively, when the two representations are completely disentangled, we can use the information in both the content embedding $E_c(x_{i,m})$ and the emotion embedding $E_c(x_{j,n})$ from audio clips $x_{i,m}$ and $x_{j,n}$ to reconstruct the clip $x_{i,n}$ from a decoder D. By leveraging the pseudo training pairs we build before, we introduce two new samples $x_{i,n}, x_{j,m}$ to serve as supervisions for the reconstruction procedure. Since each sample can only provide one type of information that is beneficial to the cross-reconstruction, the disentanglement can be finally achieved.

We supervise the training process with a loss function including four parts: cross reconstruction loss, self reconstruction loss, classification loss, and content loss. Given four audio samples $x_{i,m}, x_{j,n}, x_{j,m}, x_{i,n}$, we formulate the cross reconstruction loss as :

$$L_{cross} = \|D(E_c(x_{i,m}), E_e(x_{j,n})) - x_{i,n}\|_2 \\ + \|D(E_c(x_{j,n}), E_e(x_{i,m})) - x_{j,m}\|_2. \tag{2}$$

Besides, we are also able to reconstruct the original input by using the encoders and the decoder, namely the self reconstruction loss defined as:

$$L_{self} = \|D(E_c(x_{i,m}), E_e(x_{i,m})) - x_{i,m}\|_2 \\ + \|D(E_c(x_{j,n}), E_e(x_{j,n})) - x_{j,n}\|_2 \tag{3}$$

14083

In order to encourage the E_e to map samples with the same emotion type into clustered groups in the latent space, we add a classifier C_e for the emotion embedding and an additional classification loss defined as:

$$L_{cla} = -\sum_{k=1}^{N}(p_k * \log q_k).\qquad(4)$$

Here, N denotes the number of different emotion types, p_k denotes whether the sample takes emotional label k, and q_k denotes the corresponding network prediction probability. Moreover, we also constrain the samples with the same utterance to share similar content embedding:

$$L_{con} = \sum_{k=i,j} \|E_c(x_{k,m}) - E_c(x_{k,n})\|_1.\qquad(5)$$

Summing these four terms, we obtain the total loss function:

$$L_{dis} = L_{cross} + L_{self} + \lambda_{cla}L_{cla} + \lambda_{con}L_{con},\qquad(6)$$

where λ_{cla} and λ_{con} are weights for the classification and the content loss respectively.

3.3. Target-Adaptive Face Synthesis

To generate photo-realistic facial animations of the input portrait, we first introduce an *audio-to-landmark* network, following [34, 9], that predicts landmark motions from the disentangled audio embeddings. Afterwards, video-based editing methods normally perform facial editing on a target video clip recorded on the target person. However, this would lead to two challenges in our setting which alter the whole face rather than only the mouth: 1) The misalignment of head poses. The head movements of the predicted landmarks may differ from the target video severely, and barely any pose information is provided in the audio. 2) The blending of the edited face and the target video is difficult while synthesizing high fidelity results.

To cope with such challenges, we propose the *3D-Aware Keypoint Alignment* algorithm to align our generated landmarks with guidance landmarks in the 3D space. Then we propose to merge our generated landmarks with the edge map of the target image. The two of them together can serve as the guidance to train an *Edge-to-Video* translation network for the final results.

Audio-to-Landmark Module. Our first goal is to learn landmark displacements from emotional audio clips, thus requiring the facial shape, or identity information from the aligned landmark l_a unchanged. So we extract the landmark identity embedding f_a with a multi-layer perceptron (MLP) as shown in Fig 2. Then f_a is sent into an audio-to-landmark module along with the two disentangled audio embeddings $E_c(x)$ and $E_e(x)$. The audio-to-landmark module predicts the landmark displacements l_d by a long short-term memory (LSTM) network followed by a two-layer MLP.

In terms of the loss function, we minimize the distance between the reference landmarks l and the predicted ones $\hat{l}$ defined below:

$$L_a = \|\hat{l} - l\|_2 = \|l_a + l_d - l\|_2.\qquad(7)$$

3D-Aware Keypoint Alignment. For aligning head poses, we first perform a landmark detection on the target video using an off-the-shelf method [42]. Then we operate in the 3D space, where the pose information is explicitly defined. A parametric 3D face model [5] is utilized to recover the 3D parameters from 2D landmarks by solving a non-linear optimization problem. Utilizing the 3D geometry and expression parameters, we get a set of pose-invariant 3D landmarks $\boldsymbol{L}_p^{3d}$ (See supplementary for details). The pose parameters $\boldsymbol{p}$ contains a 3×3 rotation matrix $\boldsymbol{R}$, 2 translation coefficients $\boldsymbol{t}$, and 1 scaling coefficient s. By replacing the pose parameters of the predicted landmark with the detected ones in the target video ($\boldsymbol{R}_t, \boldsymbol{t}_t, s_t$), we obtain the adapted 3D key points and then project them to the image plane with scale orthographic projection:

$$\boldsymbol{l_p} = s_t * \boldsymbol{Pr} * \boldsymbol{R}_t * \boldsymbol{L}_p^{3d} + \boldsymbol{t}_t,\qquad(8)$$

where $\boldsymbol{l_p}$ is the projected 2D landmark and $\boldsymbol{Pr}$ is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$. Since the geometry and expression parameters remain unchanged, the projected landmarks $\boldsymbol{l_p}$ naturally share consistent identity and facial expressions with the predicted landmarks. While the head pose, scale and position are set the same as the face in the target video frame.

Note we use 3D models only for pose alignment. An alternative way is to directly predict 3D facial parameters for facial reconstruction [23, 43]. However, inaccurate fitted facial expressions are not sufficient for representing the detailed emotions in our work.

Edge-to-Video Translation Network.

Given the adapted landmarks and the target frame, we merge the landmarks and the edge map extracted from this frame into a guidance map for portrait generation. In particular, we extract edges outside the face region using an edge detection algorithm [16], and replace the original landmarks with our aligned ones. Then we connect adjacent facial landmarks to create a face sketch.

Following [39], we adopt a conditional-GAN architecture for our Edge-to-Video translation network. The generator part G is designed as a coarse-to-fine architecture [40], while the discriminator part is designed to guarantee both the quality and the continuity of the generated frames. Please refer to [39] for more details about the network architecture.

Figure 4: **Qualitative comparisons with the state-of-the-art methods.** We show three examples with different speech content and emotions. Note that we choose the same target video with frontal face for Song *et al.*[32] and ours, and we use the first frame of the target video for Chen *et al.*[9] and Wang *et al.*[37] as they edit a target image rather than a target video.

Method/Score	M-LD ↓	M-LVD ↓	F-LD ↓	F-LVD ↓	SSIM ↑	PSNR ↑	FID ↓
Chen *et al.*[9]	3.27	2.09	3.82	1.71	0.60	28.55	67.60
Wang *et al.*[37]	2.52	2.28	3.16	2.01	0.68	28.61	22.52
Song *et al.*[32]	2.54	1.99	3.49	1.76	0.64	29.11	36.33
Ours	**2.45**	**1.78**	**3.01**	**1.56**	**0.71**	**29.53**	**7.99**

Table 1: **Quantitative comparisons with the state-of-the-art methods.** We calculate the landmark accuracies and video qualities of the results of different solutions by comparing them with the ground truth. M- represents mouth and F- stands for face region.

4. Experiment

Implementation Details. We evaluate our method on MEAD, a high-quality emotional audio-visual dataset with 60 actors/actresses and eight emotion categories. The models are trained and tested on the train/test splits of the dataset. All the emotional talking face videos are converted to 25 fps and the audio sample rate is set to be 16kHz. For the video stream, we align all the faces based on the detected facial landmarks. As for the audio stream, we follow the design in [9] to extract a 28×12 dim MFCC feature corresponding to each frame in the video. Before training the disentanglement module, the emotion encoder is pretrained through an emotion classification task [30]. Meanwhile, the content encoder is pretrained on LRW [12], a lip-reading dataset with barely any emotion. Then we discard the decoder and use the two pretrained encoders in our training process. More implementation details can be found in the

supplementary materials.

Comparing Methods. We compare our work with three prior works [9, 37, 32]. The method of Chen *et al.* [9] is an image-based method, which synthesizes facial motions based on landmarks and employs an attention mechanism to improve the generation quality. The method of Song *et al.* [32] is a video-based method which applies 3D face models to realize audio-based video portrait editing. Then we compare our work with Wang *et al.* [37], the most relevant work that proposes the first talking face generation approach with the capacity of manipulating emotions. We believe these three works are the most representative works to compare with.

4.1. Experimental Results

Qualitative Comparisons. We make comparisons with other methods on various sequences as shown in the accompanying video. We also select some frames as shown
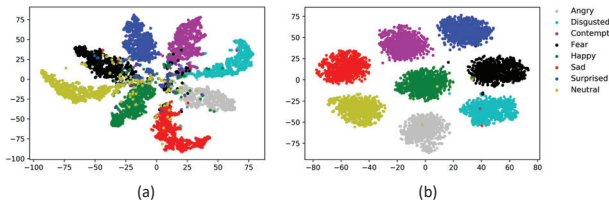
Figure 5: **Emotion latent space clusters with and without the cross-reconstruction part.** (a) Emotion latent codes of the pre-trained emotion encoder. (b) Emotion latent codes of the encoder trained by cross-reconstruction. Different colors indicate different emotion types.

in Fig. 4. Our method is able to generate high-fidelity emotional talking face video which is better than others. Concretely, Chen *et al.*[9] and Song *et al.*[32] do not consider emotions, so they generate plausible mouth shapes but always with the neutral emotion. Wang *et al.*[37] is able to generate desired emotions. However, the emotion of the predicted mouth shape is sometimes inconsistent with the facial expression (left) since it directly learns mouth shapes from audio signals where the emotion and content information are closely entangled. In addition, the algorithm in Wang *et al.*[37] is not robust enough to data with large head movements and background variations, leading to implausible facial expressions (middle) and changes with characteristics like hairstyles (right).

Quantitative Comparisons. To quantitatively evaluate different methods, we extract facial landmarks from the aligned result sequences and the ground truth sequences. The alignment is also for compensating head motions. Then, the metrics of Landmark Distance(LD) and Landmark Velocity Difference(LVD)[9, 46] are utilized to evaluate facial motions. LD represents the average Euclidean distance between generated and recorded landmarks. Velocity means the difference of landmark locations between consecutive frames, so LVD represents the average velocity differences of landmark motions between two sequences. We adopt LD and LVD on the mouth and face area to evaluate how well the synthesized video represents accurate lip movements and facial expressions separately. To further evaluate the quality of the generated images of different methods, we compare the SSIM [41], PSNR, and FID [17] scores. The Results are illustrated in Table 1. Our method obviously outperforms others in audio-visual synchronization (M-LD, M-LVD), facial expressions (F-LD, F-LVD) and video quality (SSIM, PSNR, FID).

User Study. To quantify the quality (including the accuracy of emotion and facial motion) of the synthesized video clips, we design thoughtful user studies to compare real data with generated ones from EVP, Wang *et al.*[37] , Chen *et al.*[9] and Song *et al.* [32] . We generate 3 video clips for each of the 8 emotion categories and each of the 3 speakers,
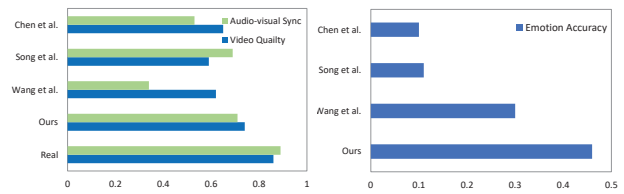

Figure 6: **User study.** User study results of audio-visual synchronization, video quality and emotion accuracy.

hence 72 videos in total. They are evaluated w.r.t three different criteria: whether the synthesized talking face video is realistic, whether the face motion sync with the speech, and the accuracy of the generated facial emotion. The evaluation consists of two stages. First, the attendees are asked to judge the given video upon audio-visual synchronization and video quality and score from 1 (worst) to 5 (best). Then we show them real emotional video clips without background sound. After that they need to choose the emotion category for the generated video without voice. 50 participants finished our questionnaire and the results are shown in Figure 6. As can be seen, our method obtains the highest score on visual quality and audio-visual sync apart from the real data. We also achieve the highest accuracy on emotion classification compared with other methods.

More Results. We show image results of our EVP algorithm in Figure 1 and more results can be found in our supplementary video[2]. Our method can synthesize high-quality emotional video portraits adaptive to various head poses and backgrounds. What's more, during inference different audio signals or even learned features can be taken as content and emotion encoder inputs, leading us to more applications described in Sec. 4.2.

4.2. Disentanglement Analysis

As illustrated in Sec. 3, our network disentangles content and emotion information from audio signals. To validate this, we feed different audio inputs to the content encoder and the emotion encoder. As shown in part (a) of Fig. 1, the mouth motion in the generated video is in accordance with the audio fed into the content encoder, while the generated facial expression matches the emotion of the audio of the emotion encoder. Extensive experiments (shown in the accompanying video) also indicate that the speech content and emotion are successfully decoupled from the audio signals.

Moreover, to quantitatively evaluate the generated emotions in the final video, we adopt an off-the-shelf emotion classification network [28] in our experiments. We trained the classification network on the training set of MEAD and calculated the numerical emotion accuracy of a face video by comparing its emotion classification result with

---

[2]The materials are available at https://jixinya.github.io/projects/evp/.

Method/Score	M-LD	M-LVD	F-LD	F-LVD
Ours w/o L_{cla}	2.72	1.83	3.68	1.63
Ours w/o L_{con}	2.65	1.86	3.03	1.60
Ours w/o L_{self}	2.47	1.83	3.02	1.62
Ours w/o L_{cross}	2.54	1.80	3.19	1.59
Ours	**2.45**	**1.78**	**3.01**	**1.56**

Table 2: **Quantitative ablation study for Cross-Reconstructed Emotion Disentanglement component.** We show quantitative results of landmarks with different losses.

the ground truth label. Since the videos in MEAD have ground truth emotion labels, quantitative evaluation can be performed here. The testing set of MEAD gets 90.2% accuracy, indicating the classification network outputs reasonable emotion labels. Our method gets 83.58% accuracy exceeding the 76.00% accuracy of Wang *et al.*[37], showing that our method better maintains the emotion.

Emotion Editing. As our method encodes emotion features in a continuous latent space, alternating features in the latent space is able to achieve emotion manipulation including emotion category as well as the intensity manipulation . In particular, we perform emotion category manipulation by calculating the mean latent feature of each emotion cluster and interpolating between the mean codes. Results are shown in part (b) of Fig. 1. By tuning the weight α between the source emotion E_s and the target emotion E_t, we get image sequences conditioned on a linear interpolated emotion feature: $\alpha E_s + (1-\alpha)E_t$. We can find that the emotion transformation between frames is consistent and smooth, which means our work is capable of continuously editing emotion via speech features.

4.3. Ablation Study

Cross-Reconstructed Emotion Disentanglement. The cross-reconstruction is the key to our disentanglement. To evaluate the disentanglement, in Fig. 5, we compare the emotion latent spaces obtained by networks with and without the cross-reconstruction training. We use t-SNE [27] to visualize the latent codes. Different colors represent audio with eight different emotion categories. It can be seen that by using the cross-reconstruction, the samples with the same emotion class are more clustered than those obtained without it. This indicates that the cross-reconstruction does contribute to decoupling emotion information from audios. We also evaluate the effectiveness of our cross-reconstruction by comparing the emotion classification accuracy of the final synthesized video clips. Without the reconstruction part, our method gets an accuracy of 69.79%, lower than results with it(83.58%). It demonstrates that the reconstruction module enhances the emotion correctness of the final talking head videos and facilitates the emotion control of our technique. Moreover, we conduct experiments to demonstrate the contributions of the four losses introduced
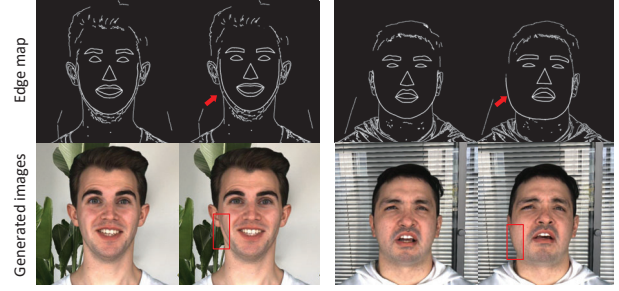

Figure 7: **Ablation study for 3D-aware keypoint alignment module.** We show cases with (left) and without (right) 3D keypoint alignment. The red arrows point out the displacements between landmarks and the face contour, while the red boxes show the artifacts in the synthesized frame.

in Sec. 3.2. Quantitative results are shown in Table 2, which prove that each loss contributes to the component.

Target-Adaptive Face Synthesis. Fig. 7 shows the qualitative results of the 3D alignment in face synthesis. 3D landmark alignment enables us to change head motions to be consistent with the target portrait video, so that the Edge-to-Video Translation network generates smooth and realistic frames. On the other hand, directly using 2D facial landmarks brings displacements between the synthesized face and the face contour of the input video in the edge map, which results in noticeable artifacts in the final synthesized video.

5. Conclusion

In this paper, we present an audio-driven video editing algorithm to synthesize emotional video portraits via effective learning in the decoupled representation space. We propose *Cross-Reconstructed Emotion Disentanglement* to decompose the input audio sample into a pair of disentangled content and emotion embeddings, based on which, 2D facial landmarks with emotion dynamic can be generated. Then, we propose *Target-Adaptive Face Synthesis* to produce the video portraits with high fidelity by aligning the newly generated facial landmarks with the natural head poses of target videos. Qualitative and quantitative experiments have validated the effectiveness of our method.

References

[1] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *arXiv preprint arXiv:1905.01680*, 2019. 4

[2] Sigurdur Orn Adalgeirsson and Cynthia Breazeal. Mebot: A robotic platform for socially embodied telepresence. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010. 2

[3] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, 1994. 2, 4

[4] Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 3

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 2013. 5

[6] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*, 2020. 3

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *European Conference on Computer Vision (ECCV)*, 2020. 2

[8] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[9] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 5, 6, 7

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 3

[11] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*, 2017. 2, 3

[12] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, 2016. 6

[13] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 3

[14] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)*, 2016. 2

[15] Ohad Fried, Maneesh Agrawala, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, and Christian Theobalt. Text-based editing of talking-head video. *ACM TOG*, 2019. 2

[16] Bill Green. Canny edge detection tutorial. *Retrieved: March*, 2002. 5

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems (NeurIPS)*, 2017. 7

[18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 3

[20] Xiaoqi Jia, Jianwei Tai, Hang Zhou, Yakai Li, Weijuan Zhang, Haichao Du, and Qingjia Huang. Et-gan: Cross-language emotion transfer based on cycle-consistent generative adversarial networks. *ECAI*, 2020. 3

[21] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 3

[22] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 2019. 2

[23] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 2018. 5

[24] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 3

[25] Yilong Liu, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo. Video-audio driven real-time facial animation. *ACM Transactions on Graphics (TOG)*, 2015. 2

[26] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, 2000. 4

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 8

[28] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2019. 7

[29] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3290–3298, 2020. 2, 3

[30] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 2014. 6

[31] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[32] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020. 2, 3, 6, 7

[33] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *IJCAI*, 2019. 2, 3

[34] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 2017. 2, 3, 5

[35] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[36] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 2019. 2, 3

[37] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 2, 3, 4, 6, 7, 8

[38] Lijuan Wang, Xiaojun Qian, Wei Han, and Frank K Soong. Synthesizing photo-real talking head via trajectory-guided sample selection. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010. 3

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 5

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 5

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 7

[42] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[43] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[44] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2, 3

[45] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[46] Yang Zhou, Dingzeyu Li, Xintong Han, Evangelos Kalogerakis, Eli Shechtman, and Jose Echevarria. Makeittalk: Speaker-aware talking head animation. *arXiv preprint arXiv:2004.12992*, 2020. 2, 3, 7

[47] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 2018. 2

[48] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. *IJCAI*, 2020. 3

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3