EMMN: Emotional Motion Memory Network for Audio-driven Emotional Talking Face Generation

Shuai Tan, Bin Ji , and Ye Pan [*]

Shanghai Jiao Tong University
{tanshuai0219,bin.ji,whitneypanye}@sjtu.edu.cn

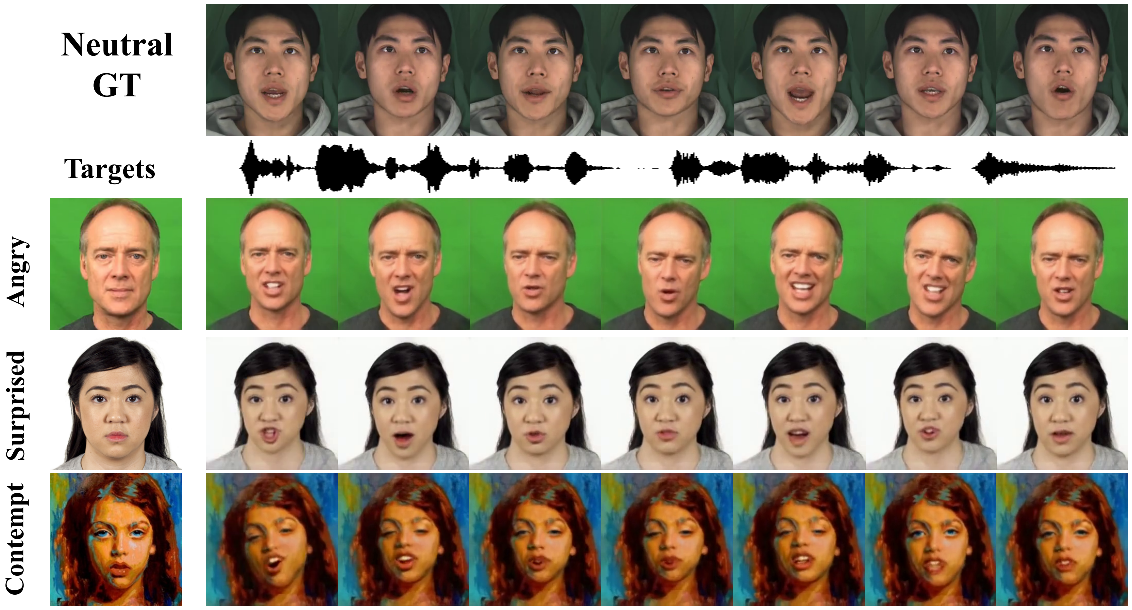


Figure 1: **Example animations generated by our EMMN.** Given a reference image, EMMN is capable of generating emotional talking face driven by an emotional audio clip. The lip motions are synchronized with the contents of the audio, while the facial emotion dynamics are controlled by the emotion embedding in audio.

Abstract

Synthesizing expression is essential to create realistic talking faces. Previous works consider expressions and mouth shapes as a whole and predict them solely from audio inputs. However, the limited information contained in audio, such as phonemes and coarse emotion embedding, may not be suitable as the source of elaborate expressions. Besides, since expressions are tightly coupled to lip motions, generating expression from other sources is tricky and always neglects expression performed on mouth region, leading to inconsistency between them. To tackle the issues, this paper proposes Emotional Motion Memory Net (EMMN) that synthesizes expression overall on the talking face via emotion embedding and lip motion instead of the sole audio. Specifically, we extract emotion embedding from audio and design Motion Reconstruction module to decompose ground truth videos into mouth features and expression features before training, where the latter encode all facial factors about expression. During training, the emotion embedding and mouth features are used as keys, and the corresponding expression features are used as values to create key-value pairs stored in the proposed Motion Memory Net. Hence, once the audio-relevant mouth features

[*]Corresponding author.

and emotion embedding are individually predicted from audio at inference time, we treat them as a query to retrieve the best-matching expression features, performing expression overall on the face and thus avoiding inconsistent results. Extensive experiments demonstrate that our method can generate high-quality talking face videos with accurate lip movements and vivid expressions on unseen subjects.

1. Introduction

High-fidelity audio-driven facial animation has various applications, including education, well-being, and entertainment [25]. Extensive efforts have been devoted to generating not only lip motions [5, 34, 33] synchronized with the audio, but also rhythmic head movements [4, 36, 43, 40] to create more realistic talking heads. However, most approaches rarely consider emotional expression, an essential element for delivering communicative information [10].

Facial emotion dynamics are typically expressed through the coordinated movements of multiple facial muscles in a global manner [12]. For instance, when surprisingly speaking, individuals tend to widen their eyes and open their mouths unconsciously larger than displaying other expressions. It suggests that expression and lip motions are intrinsically interconnected and mouth shape is crucial for conveying emotions [14, 20]. Recently, some works treat expressions and mouth shapes as a whole and predict them from the audio [15, 11]. However, audio, which consists of phoneme-level contents and coarse-grained emotion embeddings, is inadequate for controlling the generation of fine-grained global expressions. Therefore, our goal is to separately predict expression and mouth shape from sources in addition to audio and integrate the predictions globally. Nevertheless, there exist two formidable challenges: (1) Since expression and mouth shape are inherently coupled on the face, it is tough to achieve separate prediction and global integration. (2) Although emotional audio can provide a rough emotion embedding that indicates the emotion category, it struggles to guide detailed facial emotion dynamics effectively. Even though expression can be obtained from other plausible sources, it also requires compromising the consistency between expression and mouth shape to express emotion overall on the face.

To deal with the challenges above, we propose a novel audio-driven emotional talking face generation framework, namely Emotional Motion Memory Network (EMMN). Basically, we extract emotion embedding from audio following EVP [15] and leverage keypoint-based dense motion fields [29, 36] to represent facial dynamics. Our intuition is to completely disentangle the entire facial dynamics into mouth-related space and expression-related space. The former merely contains the content of audio without expression information, while the latter encodes all facial factors about expression to perform expression overall on the face. To this end, we design Motion Reconstruction module that decouples and merges the facial dynamics. To simultaneously train these two processes, the cross-reconstruction training strategy [1] is adopted. However, it brings the desire for paired videos with the exact same expression but different mouth shapes, which is almost impossible to realize in reality [35, 3] due to the strict requirement. To create the training pairs, we present a pseudo label generation strategy via a pre-trained Wav2Lip [26] model with satisfactory lip-sync generation performance. This enables the Motion Reconstruction module to be trained and further employed to decompose ground truth videos into mouth motion features and expression motion features, which serve as pseudo labels for mouth shape and expression, facilitating separate prediction setting. For instance, we design Audio2Mouth module to predict solely the mouth motion features from audio.

To ensure the consistency between expression and lip motion, we hope to recall the expression motion features most relevant to emotion embedding and the mouth motion features. Ideally, we suppose that they are disentangled from the same source. For this purpose, we resort to Memory Network [39] which leverages external memory to store information. Specifically, we construct the Motion Memory Net to store emo-mouth feature (emotion embedding and mouth motion feature) and the corresponding expression motion feature as a key-value pair. During training, the memory aligns emo-mouth features with their corresponding expression motion features in the same address. Hence, at inference time, the best-matching expression motion feature can be retrieved by querying with emo-mouth feature predicted via Audio2Mouth. Using the retrieved expression features and mouth motion features, Motion Reconstruction can globally generate the final representation of the entire emotional facial dynamics. Lastly, a flow estimator and an image generator from FOMM [29] are introduced to synthesize photo-realistic results. Extensive experiments demonstrate the superiority of our method in terms of expression naturalness and emotion accuracy compared with state-of-the-art (SOTA) methods.

Our contributions are summarized as follows: (1) We present a system named Emotional Motion Memory Network (EMMN) to perform one-shot emotional talking face generation solely from a reference image and an emotional audio clip without additional emotion sources. Hence it is flexible in applying the system. (2) We propose Motion Reconstruction that decomposes the face into expression and mouth feature to separately predict them and synthesize expression overall on the face, including the mouth region. (3) We design Motion Memory-Net to store aligned expression and emo-mouth features, which ensures the consistency of the final emotional face motion.

2. Related work

2.1. Audio-Driven Talking Face Generation

There have been several works related to the talking head generation from audio. Generally, the current methods can be divided into intermediate representation based methods and feature based methods. Intermediate representation based methods methods [47, 6, 9, 28, 13, 42, 27] usually leverage the intermediate representation such as landmark to bridge the gap between the input audio and generated video. For instance, Zhou *et al.* [47] first predict landmarks from audio and subsequently utilize them as guidance to animate an image to speak. However, the sparse representations of the landmarks hardly provide delicate information, inevitably leading to error accumulation. Conversely, we look into the dense keypoint based representation [29, 36] that encompasses the entire face motion, including facial dynamics, head motion, and background. Feature based methods [5, 45, 46, 44] encode the inputs into implicit features and leverage them to generate the result images via decoders. Prajwal *et al.* [26] map audio and face image/video into latent space and decode them to reconstruct photo-realistic faces. In addition, they significantly enhance lip-synchronization with the assistance of lip-sync discriminator. Nonetheless, none of these methods perform expression in their results.

2.2. Emotional Talking Face Generation

Compared to lip synchronization synthesis methods, few works [20, 30, 32] focus on consistent emotion in talking face generation, which is crucial in communicating. While Karras *et al.* [16] learn an emotion latent space to control the expression during testing, they fail to cover all emotions owing to the limited data. Recently, Wang *et al.* [35] release a large emotional audio-visual MEAD dataset and utilize one-hot vectors to encode emotion. Similarly, Eskimez *et al.* [11] take a face image and one-hot emotion vectors to generate emotional talking face from audio. However, both approaches integrally predict expression and mouth shape from audio. Ji *et al.* [14] extract expression information from another emotional video with the mouth masked, but the extracted information primarily indicates local emotion displacement, neglecting expression on other facial factors such as the mouth. In contrast, we aim to decouple the whole face into expression and mouth features separately and globally reconstruct dynamic face motion, performing expression with all facial factors on the entire face.

2.3. Memory Network

Memory Network [39, 31, 22] leverages memory components to store scene information for long-term memory. Due to the effectiveness in augmenting features, Memory Network has demonstrated its great power in various fields [18, 17, 23, 19] including talking face generation. For example, Yi *et al.* [41] store spatial features and aligned identity features in memory. To refine unrealistic frames from rough ones, they retrieve the corresponding identity feature by querying with spatial feature. Similarly, Park *et al.* [24] construct Audio-Lip Memory to provide lip features for more precise lip synchronization during inference. However, storing emotion-related features in memory network has yet to be attempted. Sparked by their approaches, we store emo-mouth features and expression motion features as key-value pairs and recall the best-matching expression motion features using the emo-mouth features as a query. In this fashion, dynamic face motion can be produced based on expression and mouth features.

3. Method

3.1. Overview

The proposed network is illustrated in Fig. 2. We first decouple the face representation of ground truth videos into mouth motion features and expression motion features by Motion Reconstruction module (Sec. 3.2) as pseudo labels for the rest of the training. Next, we introduce the Audio2Mouth module (Sec. 3.3) to predict mouth motion features and emotion embeddings from input audios and reference images. Subsequently, both features are utilized to query the corresponding expression motion features stored in the elaborately designed Motion Memory Net (Sec. 3.4). Lastly, we incorporate the Rendering module (Sec. 3.5) to synthesize photo-realistic results from the mouth motion features and retrieved expression motion features. In the following sections, we will explain each part of our algorithm in detail.

3.2. Motion Reconstruction

Pseudo Label Generation. To prepare paired videos that are with the exact same expression but different mouth shapes for cross-reconstruction training [1], we employ the pre-trained Wav2Lip [26] model to process videos in MEAD dataset [35]. Concretely, we feed Wav2Lip the same video with emotion i and different audios with the content a and b, hence videos $v^{i,a}$ and $v^{i,b}$ with the same expression i but with different mouth shapes a and b are obtained. In this fashion, we can generate paired data $v^{i,a}$, $v^{i,b}$ and $v^{j,b}$, where $v^{i,b}$ can be treated as the label for cross-reconstructing $v^{i,a}$ and $v^{j,b}$ during training Motion Reconstruction module. However, the Wav2Lip model is trained on LRS2 [2] and there exists an inevitable gap between LRS2 and MEAD, resulting in slightly poor performance in MEAD. To cope with the issue, we evaluate the synchronization between the generated mouth shapes and the input audio using the confidence score of SyncNet [8], and filter out the generated videos that do not exceed the

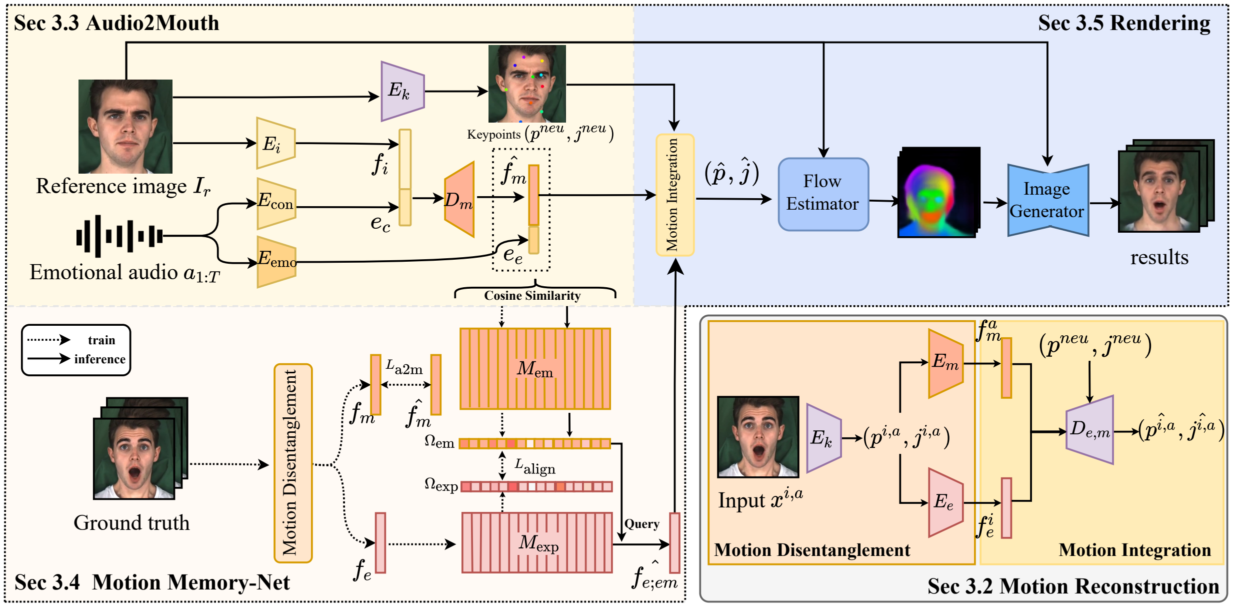


Figure 2: **The overview of proposed Emotional Motion Memory Network.** Before training, we begin by decoupling the ground truth videos into mouth motion features and expression motion features as pseudo labels (Sec. 3.2 Motion Reconstruction). During inference, we employ multiple encoders to extract the corresponding features, which we then utilize to predict mouth motion features and emotion embedding. (Sec. 3.3 Audio2Mouth). By incorporating them, the corresponding expression motion features can be queried from the learned Motion Memory Network (Sec. 3.4 Motion Memory-Net). The mouth motion features and the queried expression motion features are integrated to synthesize photo-realistic results based on the Flow Estimator and Image Generator (Sec. 3.5 Rendering).

preset threshold to obtain adequate high lip-sync paired videos for subsequent training. It is worth noting that since Wav2Lip merely alters the mouth shape of the input video while maintaining the same head pose and expression, the extracted expression features also contain head pose information in addition to facial expression, which aids in generating head motions without requiring an additional pose code input.

Design and Training. Motion Reconstruction consists of Motion Disentanglement and Motion Integration, as depicted in the bottom right of Fig. 2. We employ a pre-trained keypoint Detector E_k [29] to detect input image $x^{i,a}$ for the motion representation ($p^{i,a}$, $j^{i,a}$) pair, where $p^{i,a} \in \mathbb{R}^{N\times 2}$ represents N key-points and $j^{i,a} \in \mathbb{R}^{N\times 2\times 2}$ represents the corresponding Jacobian. The detected ($p^{i,a}$, $j^{i,a}$) is then passed through Expression Encoder E_e and Mouth Encoder E_m to extract expression motion feature f_e^i with emotion i, and mouth motion feature f_m^a with content a, respectively. Subsequently, Motion Decoder $D_{\text{e,m}}$ combines the disentangled features and the face representations (p^{neu}, j^{neu}) detected from a neutral image of the same subject to reconstruct input motion representations ($p^{\hat{i},a}$, $j^{\hat{i},a}$). During training, Motion Disentanglement processes images $x^{i,a}$ and $x^{j,b}$ to obtain disentangled features, which are crossly recombined and fed into Motion Integration to reconstruct the face representations ($p^{\hat{i},b}$, $j^{\hat{i},b}$) and ($p^{\hat{j},a}$, $j^{\hat{j},a}$).

To optimize the network, the cross-reconstruction loss and self-reconstruction loss are established to define the Motion Reconstruction task.

$$\begin{aligned} L_{\text{rec}} &= \left\| D_{e,m}\left(f_e^i, f_m^b\right) - E_k(y^{i,b}) \right\|_2 \\ &+ \left\| D_{e,m}\left(f_e^j, f_m^a\right) - E_k(y^{j,a}) \right\|_2, \end{aligned} \tag{1}$$

$$\begin{aligned} L_{\text{self}} &= \left\| D_{e,m}\left(f_e^i, f_m^a\right) - E_k(x^{i,a}) \right\|_2 \\ &+ \left\| D_{e,m}\left(f_e^j, f_m^b\right) - E_k(x^{j,b}) \right\|_2, \end{aligned} \tag{2}$$

where f_ms and f_es are features disentangled from corresponding input $x^{i,a}$ and $x^{j,b}$. In addition, $y^{i,b}$ and $y^{j,a}$ donate ground truth frames.

In addition, we import two consistency losses, i.e., expression consistency loss $L_{\text{e-con}}$ and mouth consistency loss $L_{\text{m-con}}$, to enforce the matching between the features with the same content and emotion.

$$\begin{aligned} L_{\text{e-con}} &= \left\| E_e\left(E_k(x^{i,a})\right) - E_e\left(E_k(y^{i,b})\right) \right\|_2 \\ &+ \left\| E_e\left(E_k(x^{j,a})\right) - E_e\left(E_k(y^{j,b})\right) \right\|_2. \end{aligned} \tag{3}$$

$$L_{\text{m-con}} = \left\| E_m\left(E_k(x^{i,a})\right) - E_m\left(E_k(y^{j,a})\right) \right\|_2 + \left\| E_m\left(E_k(x^{j,b})\right) - E_m\left(E_k(y^{i,b})\right) \right\|_2 . \tag{4}$$

The total loss function L_{dis} can be represented as:

$$L_{\text{dis}} = L_{\text{rec}} + L_{\text{self}} + \lambda_{\text{e-con}} L_{\text{e-con}} + \lambda_{\text{m-con}} L_{\text{m-con}}, \tag{5}$$

where $\lambda_{\text{e-con}}$ and $\lambda_{\text{m-con}}$ represent the weights for $L_{\text{e-con}}$ and $L_{\text{m-con}}$, respectively.

3.3. Audio2Mouth

We adopt the trained Motion Disentanglement component to decouple training videos into expression motion features and mouth motion features as pseudo labels. Then we design the Audio2Mouth module to map the audio input into the latent mouth motion space, and Fig. 2 shows the pipeline of the Audio2Mouth module. Unlike previous works [36, 37, 14], which take a tuple consisting of reference image, audio and pose code as input and predict N keypoints, our Audio2Mouth omits the pose code input and outputs the mouth motion features in a high dimension. Specifically, we utilize an identity encoder E_i to extract identity feature f_i from reference image I_r, which provides the identity information. Next, we introduce the pre-trained Emotion Disentanglement module in EVP [15] to separate input audio $a_{1:T}$ into emotion embedding e_e and content embedding e_c, where T denotes the length of audio features. The content embedding e_c and identity feature f_i are combined and fed into the Mouth Decoder D_m to predict the mouth motion features $\hat{f_m}$ as described in Sec. 3.2. To predict accurate mouth motion features via the Audio2Mouth module, we apply an L_2 loss, which is calculated as follows:

$$L_{\text{a2m}} = \frac{1}{T} \sum_{t=1}^{T} \left\| {f_m}^{(t)} - \hat{f_m}^{(t)} \right\|_2 , \tag{6}$$

where ${f_m}^{(t)}$ stands for mouth motion feature disentangled from tth frame in ground truth video $v_{1:T}$.

3.4. Motion Memory-Net

To ensure the consistency between mouth shape and expression, we propose the Motion Memory Net, whose purpose is to retrieve the best-matching expression motion features when querying with emo-mouth features. Specifically, we construct the Memory-Net consisting of expression motion memory M_{exp} and emo-mouth memory M_{em} to store the mutually aligned expression motion and emo-mouth features. In particular, the expression motion memory M_{exp} = $\{m^i_{\text{exp}}\}^S_{i=1}$ comprises of S slots and ith slot stores the expression motion feature m^i_{exp}. During training, we take the expression motion feature f_e disentangled from the training videos as guidance for expression motion memory. f_e is first leveraged as a query to calculate the cosine similarity with each slot, which is then processed by the softmax function $\Phi(\cdot)$ as the weight α^i_{exp} of each slot:

$$\alpha^i_{\text{exp}} = \Phi\left(\frac{f_e \cdot m^i_{\text{exp}}}{\|f_e\|_2 \cdot \left\|m^i_{\text{exp}}\right\|_2} \right). \tag{7}$$

In this way, the weight of all slots $\Omega_{\text{exp}} = \left\{\alpha^1_{\text{exp}}, ..., \alpha^i_{\text{exp}}, ..., \alpha^S_{\text{exp}}\right\}$ can be acquired. We regard the Ω_{exp} as the value address for each slot in memory when querying via f_e. Thus we are capable of recalling the feature $\widetilde{f_e}$ most relevant to f_e by following formula:

$$\widetilde{f_e} = \sum_{i=1}^{S} \alpha^i_{\text{exp}} \cdot m^i_{\text{exp}} \tag{8}$$

To update the expression motion memory, we minimize the distance between the retrieved feature $\widetilde{f_e}$ and f_e:

$$L_{\text{exp-mem}} = \left\| f_e - \widetilde{f_e} \right\| \tag{9}$$

However, the f_e is unavailable during inference to acquire the weight for each slot. Thus this brings the need to leverage the available predicted mouth motion feature $\hat{f_m}$ and emotion embedding e_e as the key to query the value in expression motion memory. Concretely, emotion embedding is responsible for the correct emotion category, while the mouth motion feature ensures the consistency between expression and mouth shape. Note that in order to establish the association between audio and memory, we use $\hat{f_m}$ predicted from audio instead of f_m disentangled from the ground truth. We ensure the accuracy of memory by ensuring the accuracy of $\hat{f_m}$ constrained via L_{a2m}. Therefore, we construct emo-mouth memory M_{em} in the same manner as building expression motion memory:

$$\alpha^i_{\text{em}} = \Phi\left(\frac{(\hat{f_m} \bigoplus e_e) \cdot m^i_{\text{em}}}{\left\|(\hat{f_m} \bigoplus e_e)\right\|_2 \cdot \|m^i_{\text{em}}\|_2} \right), \tag{10}$$

where $\bigoplus$ represents the concatenate operation, and m^i_{em} denotes ith slot containing emo-mouth motion information stored in emo-mouth memory M_{em}. Also, we can obtain the weight of each slot $\Omega_{\text{em}} = \left\{\alpha^1_{\text{em}}, ..., \alpha^i_{\text{em}}, ..., \alpha^S_{\text{em}}\right\}$.

To bridge the correlation between expression motion memory and emo-mouth memory, we need to ensure that the matching features in the respective memories are in the equivalent address. Hence, we utilize the KL divergence to align:

$$L_{\text{align}} = KL(\Omega_{\text{em}} || \Omega_{\text{exp}}) \tag{11}$$

By enforcing key address and value address obtained from $\hat{f_m} \bigoplus e_e$ and f_e consistency, both of them point to the same slots in the expression motion memory whether queried with $\hat{f_m} \bigoplus e_e$ or f_e. Therefore during the inference, we retrieve the best-matching expression motion feature by using the available emo-mouth feature as a query:

$$f_{\hat{e;em}} = \sum_{i=1}^{S} \alpha_{\text{em}}^{i} \cdot m_{\text{exp}}^{i}, \tag{12}$$

where $f_{\hat{e;em}}$ denotes the expression motion feature recalled via emo-mouth feature.

3.5. Rendering

We combine the two retrieved features and the keypoints of reference neutral image (p^{neu}, j^{neu}) and feed them into the Motion Integration component in Sec. 3.2 to predict the whole emotional face motion representation ($p^{\hat{(t)}}$, $j^{\hat{(t)}}$) in time step t. To alleviate potential jitter and distortion in results, we apply a one-euro filter as a post-processing step on ($p^{\hat{(t)}}$, $j^{\hat{(t)}}$). Then we adopt the Flow Estimator and Image Generation architectures [29] to estimate the relative dense motion between (p^{neu}, j^{neu}) and ($p^{\hat{(t)}}$, $j^{\hat{(t)}}$), and then to render the synthesized images $y^{\hat{(t)}}$.

To train the whole framework, we adopt the 2-stage train strategy following[36, 14]. Besides the loss functions above, we detect the motion representation ($p^{(t)}$, $j^{(t)}$) from the tth frame $y^{(t)}$ of training video $v_{1:T}$ via a pre-trained Keypoint Detector E_k as guidance in the first stage. The loss function $L_{\text{p,j}}$ can be formulated as:

$$L_{\text{p,j}} = \frac{1}{T} \sum_{t=1}^{T} \left(\left\| p^{(t)} - p^{\hat{(t)}} \right\|_1 + \left\| j^{(t)} - j^{\hat{(t)}} \right\|_1 \right). \tag{13}$$

Therefore, the first stage loss function L_{stage1} is formulated by the weighted sum of the above losses.

$$\begin{aligned} L_{\text{stage1}} = & L_{\text{a2m}} + \lambda_{\text{exp-mem}} L_{\text{exp-mem}} \\ & + \lambda_{\text{align}} L_{\text{align}} + \lambda_{\text{p,j}} L_{\text{p,j}}, \end{aligned} \tag{14}$$

where λs are hyper-parameters to balance these terms.

In the second stage, we import a perceptual loss L_{stage2} to fine-tune the model for a lower disparity between the output $y^{\hat{(t)}}$ and ground truth $y^{(t)}$. Given the ith channel feature of a pre-trained VGG layer with l channels $\text{VGG}_i(\cdot)$, we define L_{stage2} as:

$$L_{\text{stage2}} = \sum_{i=1}^{l} \left\| \text{VGG}_i(y^{(t)}) - \text{VGG}_i(y^{\hat{(t)}}) \right\|_1. \tag{15}$$

4. Experiments

4.1. Experimental Settings

Datasets and Implementation Details. We leverage two public datasets MEAD [35] and LRW [7] for training in our experiment. MEAD is an emotional audio-visual dataset containing videos with 8 emotions and multiple audio clips performed by different actors. We select 32 actors to train our EMMN model. Since MEAD contains limited actors, it is challenging to achieve one-shot setting. Therefore, we introduce LRW, which is collected from various speakers in BBC news, to pre-train our Audio2Mouth and Rendering modules. To evaluate the performance of our method in one-shot setting, we introduce the neutral faces of arbitrary identities from CREMA-D [3] and CFD dataset [21], both of which are unseen in the training sets. We implement our EMMN model with PyTorch. Model training and testing are conducted on 2 NVIDIA GeForce GTX 3090 with 24GB memory. We train the Motion Reconstruction for 200 epochs with Adam optimizer with an initial learning rate of 2×10^{-4}, and the entire framework is trained for 2 days. Further details can be found in the supplementary material.

Comparison Setting. We compare our Emotional Motion Memory Network with several state-of-the-art (SOTA) methods, including audio-driven talking face generation methods: Wav2Lip [26], MakeItTalk [47], PC-AVS [46], Audio2Head [36], as well as emotional talking face generation methods: ETK [11], MEAD [35] and EAMM [14]. We assess the results using evaluation metrics, including PSNR, SSIM, M-LMD, SyncNet and F-LMD. PSNR and SSIM [38] estimate the quality of the generated videos. Landmarks distances on the mouth (M-LMD) [6] and the confidence score of SyncNet [8] measure synchronization between the generated lip motion and the input audio. Compared to M-LMD, F-LMD calculates the disparity of mean distance of all landmarks between predictions and ground truth (GT) to evaluate the generated expression.

4.2. Experimental Results

Quantitative Results. We conduct experiments on MEAD and CREMA-D datasets, and the quantitative comparison is given in Tab. 1. Compared to SOTAs, our method generates expression and head pose without any additional emotion source and head pose source. Besides, it is clear that our method outperforms SOTAs in most metrics on both datasets. On MEAD, Wav2Lip [26] achieves the highest score on $Sync_{\text{conf}}$, mainly due to the help of the SyncNet discriminator when training the model. This also validates the rationality of our pseudo label generation strategy via Wav2Lip. Although ETK [11] is trained on CREMA-D which is unseen in our training set, our method performs better than ETK, only except for F-LMD. We assume that the head movements generated by our method and the expression performance gap between MEAD and CREMA-D datasets lead to disparity with the fixed pose of ground truths in CREMA-D.

Qualitative Results. We first compare our method with the SOTA emotional generation methods. Specifically, we

Method	MEAD [35]					CREMA-D [3]					Input Source		Output	
	PSNR↑	SSIM↑	M-LMD↓	F-LMD↓	$Sync_{\text{conf}}$ ↑	PSNR↑	SSIM↑	M-LMD↓	F-LMD↓	$Sync_{\text{conf}}$ ↑	Pose	Emotion	Pose	Expression
MakeItTalk [47]	28.78	0.58	3.59	3.54	3.07	29.63	0.61	3.36	3.48	2.16	✗	✗	✓	✗
Wav2Lip [26]	28.87	0.57	2.39	3.33	**3.69**	29.58	0.59	3.32	3.41	2.40	✗	✗	✗	✗
Audio2Head [36]	29.07	0.60	3.22	3.26	3.20	29.76	0.59	3.28	3.35	2.19	✗	✗	✓	✗
PC-AVS [46]	28.92	0.59	3.15	3.09	3.12	29.78	0.57	3.11	3.19	2.29	✓	✗	✓	✗
ETK [11]	27.68	0.48	3.73	3.81	2.75	29.83	0.63	3.06	**3.10**	2.39	✗	✓	✗	✗
MEAD [35]	28.88	0.58	3.23	3.35	3.08	-	-	-	-	-	✗	✓	✗	✓
EAMM [14]	29.23	0.62	2.96	3.09	3.43	29.92	0.65	3.12	3.29	2.35	✓	✓	✓	✓
Proposed	**29.38**	**0.66**	**2.78**	**2.87**	3.57	**30.03**	**0.68**	**3.03**	3.16	**2.41**	✗	✗	✓	✓
GT	-	1.00	0.00	0.00	3.77	-	1.00	0.00	0.00	2.69	-	-	-	-

Table 1: **Quantitative comparisons with state-of-the-art methods.** MEAD [35] mainly works on MEAD dataset, and fails to generalize to CREMA-D. For each method, we also list additional input sources and output performance for head pose (Pose) and expression (Emotion).

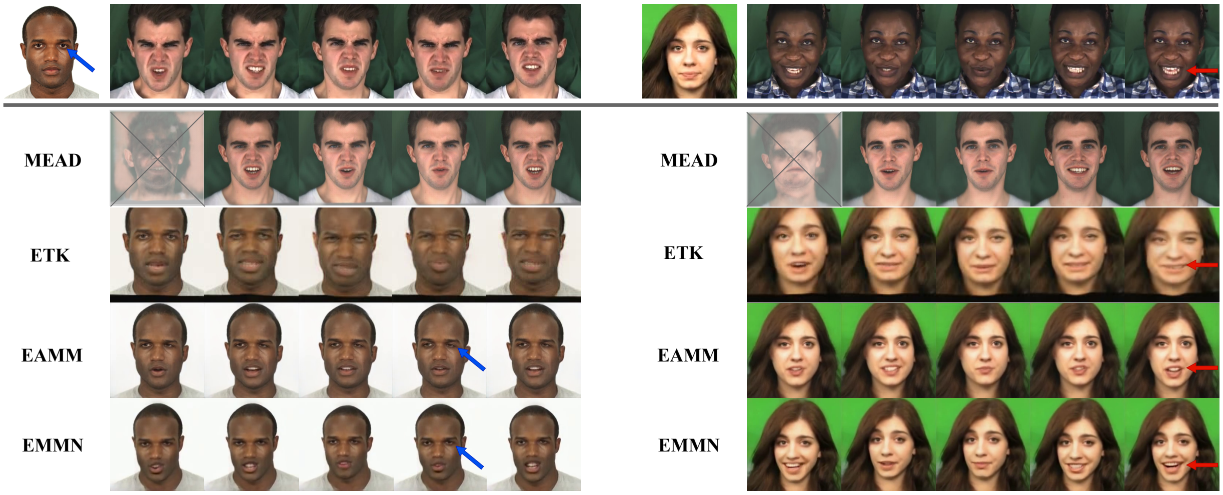


Figure 3: **Qualitative comparisons with state-of-the-art emotional talking face generation methods.** The first row present target images from CFD and CREMA-D datasets and ground-truths with 'disgueted' emotion (left) and 'happy' emotion (right) from MEAD dataset. More results can be found in the supplementary video.

randomly select reference images from CFD and CREMA-D datasets and audios from MEAD as the inputs of each method. Qualitative results are demonstrated in Fig. 3. Our method generates more similar expressions with ground truth. While all methods generate the corresponding expression, target-specific MEAD [35] fails to generalize to unseen subjects as shown in the first image of the second row. Analogously, ETK focuses on synthesizing expressions, resulting in blurred results when inferring reference images from CFD dataset, especially in the mouth region. Although EAMM [38] generates accurate mouth motions and correct expressions, local emotion displacements extracted from different identities neglect expression on mouth region (pointed by red arrows) and introduce inconsistency with the source image, resulting in face deformation in their results, such as different eye sizes (pointed by blued arrows).

Fig. 4 gives the comparison results with SOTA methods without emotion. As observed, the mouth shapes predicted by our method achieve the least disparity from the ground truth. Notably, though both Audio2Head and EAMM are keypoint-based methods in the same way as our method, we only estimate mouth motion feature from the audio while they predict the entire face motion, which conspires towards slightly different upper expression from the reference image (pointed by the red arrows in the 4th column), suggesting the effectiveness of proposed Motion Reconstruction.

User Study. To estimate the quality of our method and SOTAs, we conduct a blind user study with 20 participants (10 males, 10 females). As displayed in Fig. 5a, our method outperforms other methods over all aspects. Moreover, in Fig. 5b, our method achieves 65.20% emotion accuracy, indicating that our method synthesizes more recognizable expressions compared to other methods. Detailed settings and results can be found in the supplementary.

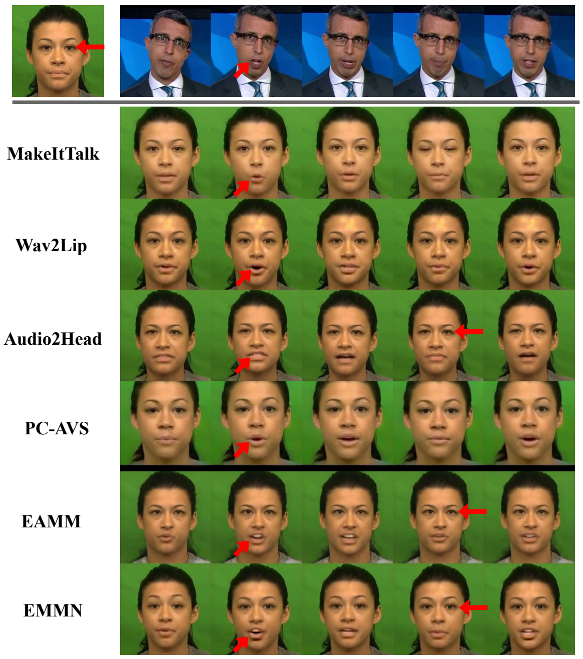
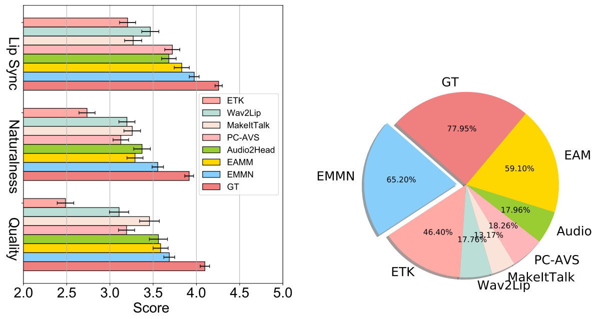


Figure 4: **Qualitative comparisons with state-of-the-art audio-driven talking face generation methods.** Top row shows target subject selected from CREMA-D and ground-truth frames from LRW dataset.



(a) Score

(b) Emotion Accuracry

Figure 5: **User study results.** The score ranges from 1 to 5, and error bars imply the standard deviations.

Method/Score	PSNR ↑	SSIM ↑	M-LMD ↓	F-LMD ↓
w/o L_{rec}	29.55	0.73	2.46	2.57
w/o L_{self}	29.63	0.75	2.49	2.58
w/o $L_{\text{e/m-con}}$	29.57	0.74	2.53	2.71
w/o $filter$	29.61	0.75	2.50	2.62
Ours	**29.89**	**0.79**	**2.36**	**2.45**

Table 2: **Ablation study for Motion Reconstruction.** We decouple features and reconstruct themselves, and then adopt Rendering module to synthesize images for metrics evaluation.

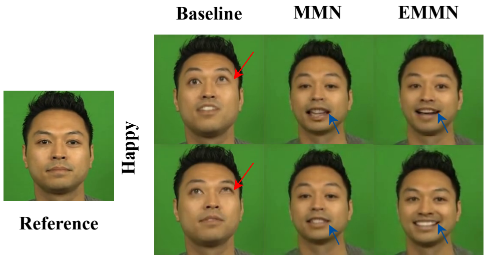


Figure 6: **Visualization Results of ablation study.** The main improvements are pointed by red and blue arrows.

Ablation Study. To investigate the effectiveness of the components of the proposed methods, we conduct the ablation study. Firstly, we evaluate the contributions of the losses and proposed filter process in Sec. 3.2 for motion self-reconstruction. The results in Tab. 2 indicate that the reconstruction loss L_{rec} contributes most to video quality. By incorporating it, PSNR and SSIM scores increase by 0.34 and 0.06, respectively. The improvement on $filter$ also demonstrates the effectiveness of filtering data. Then we verify the contributions of the Motion Reconstruction and Memory Net in a progressive manner. To be specific, the experiments are arranged as: (1) Baseline: we remove both modules from EMMN to directly output keypoints based on f_i, e_c and e_e. (2) Motion Memory Net (MMN): we store the three specific keypoints in the Expression Motion Memory based on the observations in EAMM for expression, instead of expression motion features decomposed by Motion Reconstruction module. (3) EMMN. From the visualization shown in Fig. 6, we can clearly observe that although we provide emotion embeddings for baseline, it fails to generate talking faces with emotion. Besides, the extra emotion embeddings distort normal expression (i.e., eye deformation pointed by red arrows). After adding Memory-Net, it correctly displays emotions on the upper faces, which verifies that Memory Network can effectively store emotion information. However, since only three keypoints cannot globally fuse expression, it neglects the expression affected on mouth shape (pointed by blue arrows). Compared to MMN, EMMN incorporates Motion Reconstruction module, thus generating more realistic expression on the entire face, including the mouth region. The results confirm that the Motion Reconstruction module can significantly disentangle expression from mouth shape and reconstruct it from the global. We also explore the effect of different memory slot numbers on the final performance in the supplementary material.

Limitation Despite the progress achieved, our current work has the following limitations. First of all, some expressions, such as 'disgusted' and 'angry', have similar

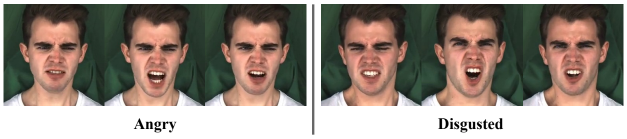


Figure 7: **Cases for limitations.**

characteristics as illustrated in Fig. 7, our method achieves a relatively low emotion accuracy on these expressions, which is consistent with the real videos in MEAD dataset. Secondly, due to the low resolution of the training data, our approach is restricted to 256×256 images. In the future,we will concentrate on overcoming these obstacles.

Ethical Considerations. Our approach has the potential to be misused for harmful purposes, such as producing deepfakes. To prevent this, we will limit the licensing of our model exclusively to research purposes and provide it to the deepfake detection community. Additionally, we will take proactive measures to incorporate watermarks into the deepfake generation process to facilitate their identification.

5. Conclusion

In this paper, we propose a novel Emotional Motion Memory Network (EMMN) to achieve emotional talking face generation. Assuming that expression entangled with lip motion is hardly predicted from audio and that expression typically involves the entire face, we decouple emo-mouth features and expression motion features by Motion Reconstruction and store aligned disentangled features in constructed Memory Network. Consequently, our method produces not only precise lip motions but also generates facial emotion dynamics by considering all facial factors about expression. Extensive experiments conducted on multiple datasets demonstrate the superiority of our method compared to the state-of-the-art methods.

Acknowledgement This work was supported by National Natural Science Foundation of China (NSFC, NO. 62102255), CCF-Tencent Open Research Fund (RAGR20220128), and Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0102).

References

[1] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *ACM Transactions on Graphics*, 2019.

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018.

[3] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[4] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020.

[5] 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)*, pages 520–535, 2018.

[6] 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/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019.

[7] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer, 2017.

[8] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017.

[9] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 408–424. Springer, 2020.

[10] Paul Ekman and Erika L. Rosenberg. What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (facs). 2005.

[11] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 24:3480–3490, 2021.

[12] Gary Faigin. The artist's complete guide to facial expression. 1990.

[13] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[14] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[15] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[16] 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.

[17] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks. *north american chapter of the association for computational linguistics*, 2018.

[18] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. *international conference on machine learning*, 2015.

[19] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. *computer vision and pattern recognition*, 2021.

[20] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022.

[21] Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 2015.

[22] Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *empirical methods in natural language processing*, 2016.

[23] Cesc Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. *computer vision and pattern recognition*, 2022.

[24] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2062–2070, 2022.

[25] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.

[26] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020.

[27] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021.

[28] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021.

[29] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

[30] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2021.

[31] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *neural information processing systems*, 2015.

[32] Zhiyao Sun, Yu-Hui Wen, Tian Lv, Yanan Sun, Ziyang Zhang, Yaoyuan Wang, and Yong-Jin Liu. Continuously controllable facial expression editing in talking face videos. 2022.

[33] Guanzhong Tian, Yi Yuan, and Yong Liu. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *2019 IEEE international conference on Multimedia & Expo Workshops (ICMEW)*, pages 366–371. IEEE, 2019.

[34] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020.

[35] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020.

[36] S Wang, L Li, Y Ding, C Fan, and X Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2021.

[37] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022.

[38] 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.

[39] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv: Artificial Intelligence*, 2014.

[40] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020.

[41] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv: Computer Vision and Pattern Recognition*, 2020.

[42] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF*

international conference on computer vision, pages 9459–9468, 2019.

[43] Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[44] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876, 2021.

[45] 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*, volume 33, pages 9299–9306, 2019.

[46] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021.

[47] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020.