JEAN: Joint Expression and Audio-guided NeRF-based Talking Face Generation

Sai Tanmay Reddy Chakkera
schakkera@cs.stonybrook.edu
Aggelina Chatziagapi
aggelina@cs.stonybrook.edu
Dimitris Samaras
samaras@cs.stonybrook.edu

Stony Brook University
NY, USA

Abstract

We introduce a novel method for joint expression and audio-guided talking face generation. Recent approaches either struggle to preserve the speaker identity or fail to produce faithful facial expressions. To address these challenges, we propose a NeRF-based network. Since we train our network on monocular videos without any ground truth, it is essential to learn disentangled representations for audio and expression. We first learn audio features in a self-supervised manner, given utterances from multiple subjects. By incorporating a contrastive learning technique, we ensure that the learned audio features are aligned to the lip motion and disentangled from the muscle motion of the rest of the face. We then devise a transformer-based architecture that learns expression features, capturing long-range facial expressions and disentangling them from the speech-specific mouth movements. Through quantitative and qualitative evaluation, we demonstrate that our method can synthesize high-fidelity talking face videos, achieving state-of-the-art facial expression transfer along with lip synchronization to unseen audio. Project Page: https://starc52.github.io/publications/JEAN

1 Introduction

Talking face generation has increasingly drawn attention due to its wide-ranging applications such as visual dubbing, video content creation and video conferencing. There are two main requirements in synthesizing a photorealistic talking face: (a) accurate lip synchronization to the spoken utterance, and (b) faithful facial expressions to convey a message with the intended affect. In human interaction, facial expressions deliver essential cues while talking [21, 22]. For example, the same sentence spoken with an angry or happy emotion can have a different meaning. Prior work has mostly focused on audio-only [14, 53, 69, 71, 86, 87] or expression-only [2, 3, 20, 38, 41, 56, 68] guidance for face synthesis. A few methods [34, 35, 36, 72] have tried to address the problem of simultaneous control of facial expressions and lips. However, they either struggle to preserve the speaker identity or fail to produce faithful expressions. Recently, neural radiance fields (NeRFs) [48] have demonstrated photorealistic 3D modeling, preserving identity-specific information and

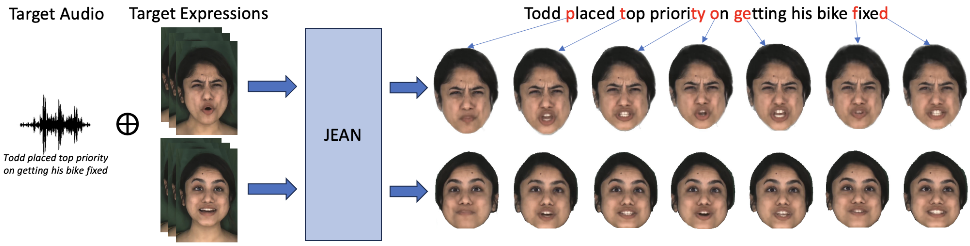

Figure 1: We introduce JEAN, a novel NeRF-based method that simultaneously combines lip-syncing to a target audio with facial expression transfer to generate talking faces.

faithfully reconstructing expressions [24]. However, NeRF-based methods have only addressed the problems of lip-syncing [10, 29, 43, 47, 83] or expression transfer [4, 5, 24, 51] separately.

In this work, we present JEAN, a novel **J**oint **E**xpression and **A**udio-guided **N**eRF for talking face generation. Our network is trained on monocular talking face videos without any ground truth. In these videos, the expression-related facial movements are strongly entangled with the speech-specific mouth movements. Controlling facial expressions and speech-specific lip motion separately requires learning *disentangled* representations for expression and audio correspondingly. To address this, we introduce a self-supervised approach to disentangle facial expressions from lip motion. We observe that mouth motion related to speech and face motion related to expressions in talking faces differ from each other temporally and spatially. Speech-related motion has higher temporal frequency and is spatially localized to the mouth region, while expression-related face motion has a lower temporal frequency and may occur over the entire face region. Moreover, for the same utterance spoken with different expressions, speech-related motion remains consistent. We leverage these observations to disentangle speech-related motion and expression-related motion.

We first learn a powerful audio representation in a self-supervised manner by disentangling the lip motion from the motion of the rest of the face in the feature space of an autoencoder. In general, achieving accurate lip synchronization on unseen audio in NeRFs is hard, as they tend to overfit on the training data [29]. Recently, contrastive learning has shown promise in synchronization in audio-visual tasks [53, 73, 77, 85]. This prompted us to introduce a contrastive learning strategy, in order to align the learned audio features with the lip motion. Next, we introduce a transformer-based architecture that learns expression features, capturing long-range facial expressions and disentangling them from speech-specific lip motion. Finally, we train a dynamic NeRF, conditioned on the learned representations for expression and audio. JEAN can synthesize high-fidelity talking face videos, faithfully following both the input facial expressions and speech signal for a given identity.

In brief, the contributions of our work are as follows:

- We introduce a self-supervised method to extract audio features aligned to lip motion features, achieving accurate lip synchronization on unseen audio.
- We propose a transformer-based module to learn expression features, disentangled from speech-specific mouth motion.
- Conditioning on the disentangled representations for expression and audio, we propose a novel NeRF-based method for simultaneous expression control and lip synchronization, outperforming the current state-of-the-art.

2 Related Work

Audio-driven Talking Face Generation. Audio-driven talking face generation aims to generate portrait images with synchronized lip motion to a given speech. Early attempts in talking face generation with lip synchronization [23, 40, 60] use probabilistic models to map speech phonemes to particular mouth shapes, requiring accurate annotation. More recent methods [11, 14, 55, 65, 69, 71, 85, 86, 87] learn neural networks, such as GANs, using a large amount of video data, containing multiple identities, in order to learn a robust audio-lip space. Our method, on the other hand, is NeRF-based, which enables us to better capture the 3D geometry and appearance of a talking face, and achieve higher output visual quality with just a few minutes of monocular video data.

NeRFs for Human Faces. Implicit neural representations for modeling 3D scenes have recently gained a lot of attention. In particular, neural radiance fields (NeRFs) [48] have shown photorealistic novel view synthesis of complex static [6, 7] and dynamic [44, 45, 57, 79] scenes. They represent a scene using an MLP, where each 3D point is associated with a radiance and density. Various recent works [24, 50, 51, 58, 64] use NeRFs to model the 3D face geometry. AD-NeRF [29] learns a dynamic NeRF conditioned on speech, encoded as DeepSpeech features [1, 30]. Follow-up methods [42, 43, 47, 61, 67, 81, 82, 83] enhance the lip synchronization in case of novel audio. NeRFace [24] conditions the network on 3DMM expression parameters. In contrast to the aforementioned approaches, our proposed NeRF allows for simultaneous control over facial expressions and lip motion to unseen audio.

Representation Learning for Human Faces. The task of representation learning for faces has been widely explored in unsupervised [12, 13, 19, 31, 39, 49, 76] and self-, semi-, or weakly-supervised [18, 27, 56, 59] techniques. In the talking face setting, where supervision is scarce and hard to find, self-supervision has been widely explored in the literature. Some methods have used self-supervision to improve the lip synchronization [53, 73, 81] in talking head generation tasks. Other methods have used self-supervised learning [26, 46, 72] to disentangle pose, expression, eye motion, etc., of a talking face to enable individual control. Other methods have used self-supervision to learn proxies, such as depth [32, 33], latent features [80] or keypoints [36] that improve generated talking faces. Our method disentangles expression from lip motion and aligns audio features to lip motion using self-supervision.

Expression and Audio-driven Talking Face Generation. Expression and audio-driven talking face generation aims to produce portrait images that follow expressions from an expression source and lips synchronized to an audio source. Prior work can be broadly classified into warping-based methods [36, 66] or synthesis-based methods [28, 34, 35, 54, 62, 63, 72]. Warping-based techniques estimate warping flows between source and target images, whereas synthesis-based methods generate images based on intermediate representations. Warping-based techniques, like EAMM [36], frequently produce 3D inconsistencies, since they consider only the 2D space. Synthesis-based methods, like PD-FGC [72], lead to the semantic leakage problem [78], where the output erroneously contains semantic elements of the input or training dataset. Some methods [28, 62, 63], such as EAT [25], learn a categorical emotional space, often based on one-hot encodings. In contrast, our method learns a NeRF controllable by both audio and expression from independent sources, giving 3D accurate and identity-preserved outputs with faithful expressions.

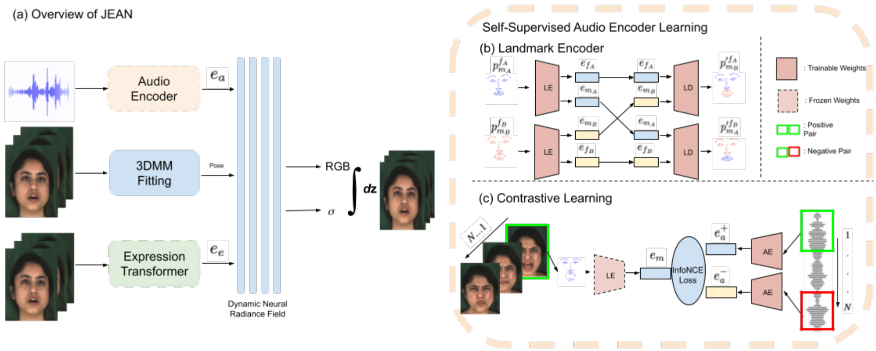

Figure 2: (a) illustrates an overview of JEAN, a novel method for joint expression and audio-guided NeRF-based talking face generation. (b) and (c) illustrate our proposed self-supervised learning of our audio representation. Specifically, (b) demonstrates the self-supervised learning of our landmark autoencoder that disentangles lip motion from the motion of the rest of the face. Then, in (c), our audio encoder *AE* is trained using a contrastive learning regime, in order to align audio features to lip motion.

3 Method

We present JEAN, a novel method for joint expression and audio-guided NeRF-based talking face generation. Fig. 2(a) illustrates an overview of our proposed approach. Given monocular RGB videos of an identity, we learn a NeRF that represents the identity's 4D face geometry and appearance in various expressions and lip positions. We assume three inputs, namely an audio source, the identity's head pose, and an expression source. During training, these inputs come from the same identity. During inference, we can use audio, pose, and expression sources from different videos. Our proposed pipeline consists of three main components: (1) We first learn an audio encoder in a self-supervised manner to align audio features to lip motion features (see Sec. 3.1). (2) We learn an expression transformer to disentangle expression features from lip motion (see Sec. 3.2). (3) Finally, we learn a dynamic NeRF conditioned on our learned representations for both audio and expression (see Sec. 3.3).

3.1 Self-Supervised Audio Encoder

In order to learn a powerful audio representation and achieve a high lip-sync accuracy, we propose a self-supervised contrastive learning method that aligns audio features to lip motion features. Inspired by Yao et al. [81], we first extract lip motion features through a landmark autoencoder. Next, we train our audio encoder using a contrastive learning strategy.

Landmark Autoencoder. We propose a landmark autoencoder that learns to disentangle mouth and eye-nose movements based on 2D landmarks, as illustrated in Fig. 2(b). For a frame A of an identity, we extract face landmarks $\mathbf{p}_{m_A}^{f_A}$, with superscript f_A indicating that the eye-nose landmarks are from frame A and subscript m_A indicating that mouth landmarks are from frame A. Similarly, for a frame B, we extract face landmarks $\mathbf{p}_{m_B}^{f_B}$. A landmark encoder (LE) embeds the input landmarks of each frame into a eye-nose embedding $\mathbf{e}_f$ and a mouth embedding $\mathbf{e}_m$, *i.e.* $\mathbf{e}_{f_A}, \mathbf{e}_{m_A} = LE(\mathbf{p}_{m_A}^{f_A})$ and $\mathbf{e}_{f_B}, \mathbf{e}_{m_B} = LE(\mathbf{p}_{m_B}^{f_B})$. The mouth embeddings $\mathbf{e}_{m_A}$ and $\mathbf{e}_{m_B}$ of the two frames A and B correspondingly are swapped with a probability ε, and passed to a landmark decoder (LD) to predict the corresponding landmarks $\mathbf{p}'^{f_A}_{m_B} =$

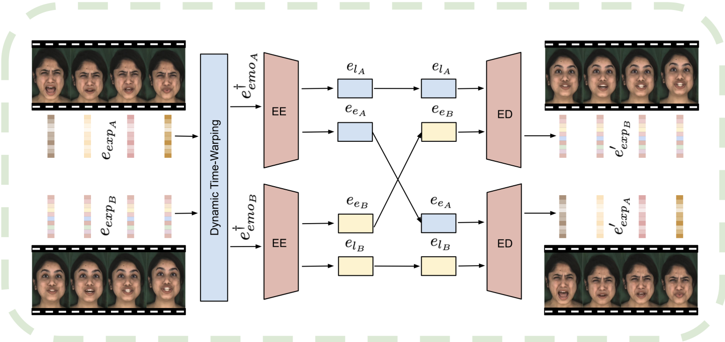

Figure 3: Expression Transformer. We propose an expression transformer encoder that learns to disentangle facial expressions from speech-specific lip motion. We extract emotion features and disentangle them into expression content and speech-specific lip motion content.

$LD(\mathbf{e}_{f_A}, \mathbf{e}_{m_B})$ and $\mathbf{p}'^{f_B}_{m_A} = LD(\mathbf{e}_{f_B}, \mathbf{e}_{m_A})$. We get the ground truth $\mathbf{p}^{f_A}_{m_B}$ by replacing the mouth landmarks of the frame A with the corresponding mouth landmarks of the frame B. The autoencoder is trained using an L1 reconstruction loss:

$$\mathcal{L}_{rec_{lmd}} = \mathbf{E}\left[||\mathbf{p}'^{f_A}_{m_B} - \mathbf{p}^{f_A}_{m_B}||_1 + ||\mathbf{p}'^{f_B}_{m_A} - \mathbf{p}^{f_B}_{m_A}||_1\right]. \tag{1}$$

Using this training regime, the landmark encoder LE learns to represent the lip movements in its latent space, disentangling them from any other face motion. In our implementation, we extracted 68 face landmarks for each video frame. We discarded the first 17 landmarks that correspond to the face contour, in order to pay attention to the eye-nose and mouth movements. The probability ε is set to 0.8. The frames A and B are randomly sampled from the same video of an identity (see also suppl.).

Contrastive Learning. In order to learn audio embeddings aligned to the extracted mouth embeddings $\mathbf{e}_m$, we propose a constrastive training strategy, as illustrated in Fig. 2(c). We learn a CNN-based audio encoder AE that takes DeepSpeech [30] features $\mathbf{a}$ as input and outputs audio embeddings $\mathbf{e}_a$, *i.e.* $\mathbf{e}_a = AE(\mathbf{a})$. For a mouth embedding $\mathbf{e}_m$, we set the corresponding audio feature $\mathbf{e}^+_a$ as the positive key and a randomly picked audio feature $\mathbf{e}^-_a$ as the negative key. We train our audio encoder, using an InfoNCE [20] loss, to ensure that the distance between the positive pair $(\mathbf{e}^+_a, \mathbf{e}_m)$ is smaller than the negative one $(\mathbf{e}^-_a, \mathbf{e}_m)$:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathop{\mathbf{E}}_{x \in \mathcal{X}}\left[\log \frac{\exp(d(\mathbf{e}^+_{a_x}, \mathbf{e}_{m_x}))}{\exp(d(\mathbf{e}^+_{a_x}, \mathbf{e}_{m_x})) + \exp(d(\mathbf{e}^-_{a_x}, \mathbf{e}_{m_x}))}\right], \tag{2}$$

where $\mathcal{X}$ is the set of all $(\mathbf{e}^+_a, \mathbf{e}_m, \mathbf{e}^-_a)$ tuples and $d(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\tau ||\mathbf{x}||_2 ||\mathbf{y}||_2}$ is the temperature-adjusted cosine distance. During training, the negative audio samples are randomly selected from the same identity but different video. The temperature τ is set to 0.1. We use 64-dimensional features for $\mathbf{e}_a$, $\mathbf{e}_f$, and $\mathbf{e}_m$. See suppl. for more details.

3.2 Expression Transformer

We propose to learn an expression transformer that captures long-range facial expressions, disentangling them from the speech-specific lip motion (see Fig. 3). First, we extract emotion features e_{emo} per video frame, using a pre-trained network for emotion recognition [15].

We then learn an expression encoder that disentangles the emotion features into expression content and speech-specific content. The idea is that when a person is speaking, the face movements will have some aspects that are emotion-specific (e.g. cheeks pulled up, flush face, raised eyebrows, etc.) and some speech-specific (e.g. mouth motion for consonant 'b'). Given that we have video pairs of a person saying the same utterance in different emotions, it is possible to capture the facial expressions and successfully disentangle them from the speech-specific motion. More specifically, for an utterance spoken with two different emotions A and B, we extract emotion features $\mathbf{e}_{emo_A[1:m_1]}$ and $\mathbf{e}_{emo_B[1:m_2]}$. We align these sequences, using the dynamic time warping (DTW) algorithm [8]:

$$\mathbf{e}^{\dagger}_{emo_A[1:N]}, \mathbf{e}^{\dagger}_{emo_B[1:N]} = DTW(\mathbf{e}_{emo_A[1:m_1]}, \mathbf{e}_{emo_B[1:m_2]}), \tag{3}$$

where m_1 and m_2 are the initial lengths and N is the output length of DTW. These are then given as input to the expression transformer encoder (EE) in windows of size ω. EE outputs expression features $\mathbf{e}_e$ and speech-specific lip motion features $\mathbf{e}_l$, *i.e.* $\mathbf{e}_l[t:\omega+t-1], \mathbf{e}_e[t:\omega+t-1] = EE(\mathbf{e}^{\dagger}_{emo[t:\omega+t-1]})$ where $t \in \mathbb{N}_{1:N}$. The expression features are randomly swapped with a probability δ. The output features $\mathbf{e}_e$ and $\mathbf{e}_l$ are input to the expression decoder (ED). ED follows an auto-regressive architecture to reconstruct the emotion features, *i.e.* $\mathbf{e}'_{emo[t:\omega+t-1]} = ED(\mathbf{e}_l[t:\omega+t-1], \mathbf{e}_e[t:\omega+t-1])$. The expression transformer is trained using an L1 reconstruction loss:

$$\mathcal{L}_{rec_{emo}} = \mathbf{E}\left[||\mathbf{e}'_{emo_A} - \mathbf{e}^{\dagger}_{emo_A}||_1 + ||\mathbf{e}'_{emo_B} - \mathbf{e}^{\dagger}_{emo_B}||_1\right]. \tag{4}$$

During inference, DTW is skipped and the emotion features are directly input to EE. Our expression encoder is identity-specific, capturing each person's unique way of speaking with a particular emotion. In our experiments, we set $\omega = 8$ and $\delta = 0.8$. EE and ED have 3 layers and 8 attention heads each. The emotion features e_{emo} are of 2048 dimension and mapped to 128 via 2 linear layers. The output of EE is split in half resulting in 64-dimensional $\mathbf{e}_l$, $\mathbf{e}_e$.

3.3 Dynamic NeRF

Our learned audio features $\mathbf{e}_a$ and expression features $\mathbf{e}_e$ are concatenated to an embedding $\mathbf{e}_{in}$, conditioning our dynamic NeRF that models the 4D face dynamics of a subject. For each video frame, we fit a 3DMM [9, 52] and extract the head pose and camera parameters, in order to estimate the viewing direction $\mathbf{d}$. The learned feature $\mathbf{e}_{in}$, the viewing direction $\mathbf{d}$ and a 3D point location $\mathbf{x}$ in canonical space are input to the implicit function F_Θ (MLP), which predicts the corresponding RGB color $\mathbf{c}$ and density σ:

$$F_\Theta : (\mathbf{e}_{in}, \mathbf{d}, \mathbf{x}) \longrightarrow (\mathbf{c}, \sigma) \tag{5}$$

Given the color $\mathbf{c}$ and density σ at each sampled point of every ray, we can reconstruct each video frame using volumetric rendering. For each camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where $\mathbf{o}$ is the camera center, the color C is estimated by accumulating the RGB colors and densities of the points sampled along the ray: $C(\mathbf{r};\Theta) = \int_{t_n}^{t_f} \sigma_\Theta(\mathbf{r}(t))\mathbf{c}_\Theta(\mathbf{r}(t),\mathbf{d})T(t)dt$ [48], where $T(s) = \exp(-\int_{t_n}^{t} \sigma_\Theta(\mathbf{r}(s)))ds$ is the accumulated transmittance from t_n to t, and t_f and t_n are the far and near bounds respectively. We denote the outputs of F_Θ as $\mathbf{c}_\Theta$ and σ_Θ for brevity. Similar to [48], we learn a coarse and a fine model for hierarchical volumetric rendering. We optimize our NeRF using a photo-consistency loss:

$$\mathcal{L}_{photo} = \sum_{\mathbf{r} \in \mathcal{R}} ||\hat{C}(\mathbf{r};\Theta) - C(\mathbf{r};\Theta)||_2^2, \tag{6}$$

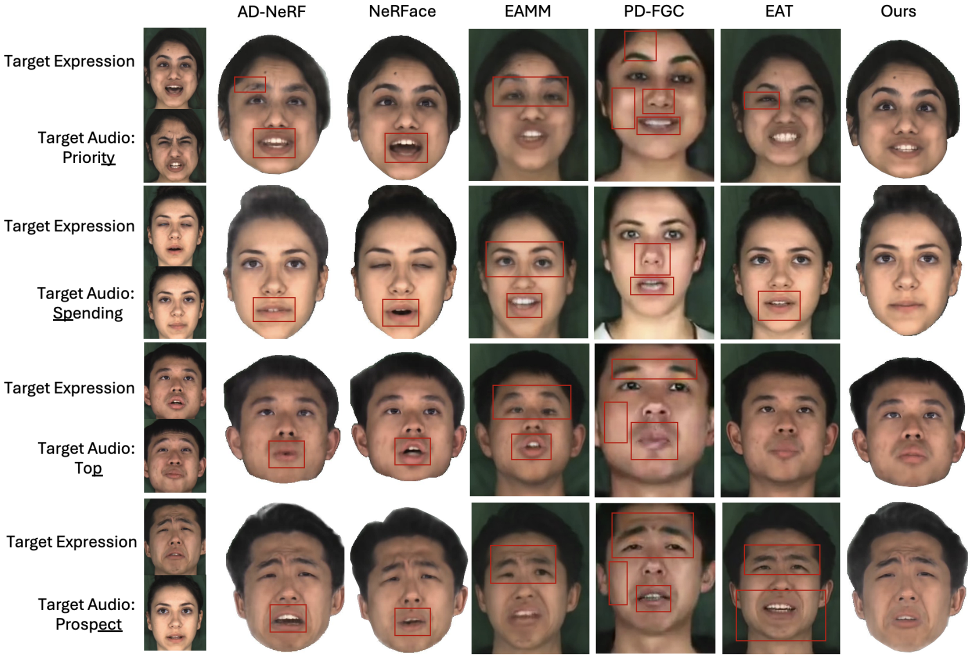

Figure 4: Talking face generation guided by target expression and audio sources (1st column). We compare with state-of-the-art methods for expression and audio-driven talking face generation (EAMM [36], PD-FGC [72]), categorical emotion based talking face generation (EAT [25]), as well as the audio-only AD-NeRF [29], and expression-only NeRFace [24]. Our method outperforms all these methods, transferring the expression and audio inputs with higher fidelity, while preserving the target identity.

which measures the mean squared error between the ground truth color $C(\mathbf{r};\Theta)$ and the predicted color $\hat{C}(\mathbf{r};\Theta)$, and $\mathcal{R}$ is the set of all the rays in each batch (see also suppl.).

4 Experiments

Dataset. In our experiments, we use the MEAD dataset [74]. MEAD includes 48 identities, performing 7 emotions at 3 intensity levels and 1 neutral emotion. The videos are captured by 7 cameras at different viewpoints. Each emotion level contains $\approx 30-40$ videos corresponding to an utterance sampled from a superset of sentences. To train our audio encoder (see Sec. 3.1), we use the complete set of frontal-view videos. For the expression transformer (see Sec. 3.2), we need video pairs, where a person pronounces the same utterance with different emotions. We use the highest level (level 3) of the emotions "angry", "happy" and "sad", and the single level of "neutral", leading to a total of 84 unique pairs of sentences being spoken in 2 emotions. Of those, 60 pairs of videos are used for training, and 24 for validation. Since each person's expressions are unique, we train an expression transformer for each identity. To train our dynamic NeRF, we focus on 4 identities from MEAD, training the network for each identity. Since the videos in MEAD are only 4-8 seconds long, we concatenate videos of the same emotion for each identity. Not all the videos of the same emotion for an identity were captured in the same head pose, so we filter videos based on the

Table 1: Quantitative comparison of our method with the state-of-the-art. Results are highlighted as follows: Best, Second Best and Third Best.

Method	LSE-C ↑	ACD ↓	Exp-Diff ↓	PSNR ↑	SSIM ↑	LPIPS ↓
AD-NeRF [29]	4.380	0.141	0.080	20.467	0.698	0.187
NeRFace [24]	1.966	0.103	0.023	21.335	0.736	0.175
EAMM [36]	1.771	0.284	0.111	18.474	0.592	0.248
PD-FGC [72]	6.220	0.657	0.089	21.094	0.648	0.228
EAT [25]	7.200	0.183	0.078	19.222	0.691	0.205
Ours	4.466	0.095	0.043	21.224	0.720	0.174

estimated focal length and the pose distribution after fitting a 3DMM. After concatenation, we get videos of at least 12 seconds per emotion. See suppl. for more details.

4.1 Results

In this section, we compare our proposed method with the state-of-the-art. We include comparisons with AD-NeRF [29] that takes only audio as input and NeRFace [24] that takes only 3DMM expression parameters as input, in order to illustrate our method's performance against NeRF-based methods with just one of the inputs, audio or expression. Then, we compare with EAMM [36], PD-FGC [72], and EAT [25] that are identity-generic methods for emotion-aware talking face video generation, trained on large datasets. Note that while EAMM and PD-FGC use a video source for expressions, EAT only takes an emotion label as input to produce expressions. Our method achieves the best disentanglement between expression and audio sources, producing high-quality expressive talking faces.

4.1.1 Qualitative Evaluation

Fig. 4 shows our qualitative results. Notice how our method produces accurate lip shapes that follow the target audio (*e.g.* phoneme "t" in row 2), while also synthesizing the input expression (*e.g.* sad) with higher fidelity than the other methods. EAMM generally distorts the input face, adds asymmetrical artifacts, and is unable to produce accurate mouth shapes in all rows. While PD-FGC performs better than EAMM in terms of lip-shape accuracy, it still distorts the input identity and produces artifacts. For example, we observe glossy faces, color distortions and lip artifacts in all rows, a plain white band in place of teeth in rows 1 and 2, and loss of identity-specific characteristics, such as the mole on the face of the woman in row 1. EAT performs best among the other methods, creating accurate lip shapes, synced to the input audio, while also being faithful to the source emotions. However, EAT still struggles with preserving the input identity. For example, it generates artifacts in the eye region and eyebrows in rows 1 and 4 respectively, and wide jaws and crossed eyes in row 4. In general, we observe identity inconsistency problems in all EAMM, PD-FGC, and EAT. In contrast, the NeRF-based methods, *i.e.* AD-NeRF, NeRFace, and our method, learn to preserve the input identity. Our method demonstrates high-quality results, transferring the source facial expression and following the source audio with higher fidelity.

4.1.2 Quantitative Evaluation

Evaluation Metrics. We conduct quantitative evaluation on common metrics used in the talking face generation field. We use LSE-C [55] to measure the lip synchronization of

Table 2: Ablation study on our proposed audio and expression representations. In (a), we train the network without the self-supervised audio encoder learning and without expression transformer (we use features from the pre-trained ResNet-based emotion recognition network from [15] passed through a thin MLP). In (b), we again omit the self-supervised audio encoder learning and use 3DMM expression parameters. In (c), we add the self-supervised audio encoder learning, but we use expression features as in (a). In (d), we train our expression transformer on 3DMM expression parameters. Best results are highlighted in **bold**.

Method					Metrics		
Variant	Self Supervised Audio Encoder	3DMM Expression Parameters	Emotion Recognition Features	Expression Transformer	LSE-C↑	Exp-Diff↓	LPIPS↓
(a)			✓		1.804	0.027	0.170
(b)		✓			1.848	**0.022**	**0.161**
(c)	✓		✓		1.760	0.028	0.168
(d)	✓	✓		✓	2.982	0.071	0.190
Full Net.	✓		✓	✓	**4.466**	0.043	0.174

our method, and Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [75] and Learned Perceptual Image Patch Similarity (LPIPS) [84] to measure the image quality against the expression source images. We also estimate the expression transfer accuracy (Exp-Diff) [72] by using 3D face reconstruction [17] and calculating the Mean Squared Error (MSE) of the extracted expression parameters in the synthesized images with those of the driving expression images. Further, we estimate the identity preservation using the Average Content Distance (ACD) metric, inspired by [71], by calculating the cosine distance between ArcFace [16] face recognition embeddings of synthesized images and driving expression images. Essentially, the idea is that the smaller the distance between those embeddings, the closer are the synthesized images to the driving images in terms of identity.

We show the corresponding quantitative results in Table 1. Since Exp-Diff and the visual quality metrics are computed against expression source frames, we find that the expression-only NeRFace performs best on those metrics. Our method significantly outperforms the state-of-the-art in emotion-aware talking face generation (EAMM, PD-FGC, EAT) in terms of visual quality (PSNR, SSIM, LPIPS), identity preservation (ACD), and expression transfer (Exp-Diff). While PD-FGC and EAT do perform better than our method in terms of lip-syncing, as they are trained on large-scale video data, our method outperforms the rest of the methods. We encourage the readers to watch our suppl. video for additional results demonstrating the efficacy of JEAN.

4.1.3 Ablation Study

Expression Disentanglement. Table 2 shows different variants of our method, demonstrating the efficacy of our self-supervised learning of our audio representation, as well as the disentanglement of our expression representation. More specifically, in variant (a), we omit the self-supervised audio encoder learning and the expression transformer (we directly use features from a pre-trained ResNet-based emotion recognition network from [15] mapped through a thin MLP, and the audio encoder is trained along with the NeRF). In variant (b), we use 3DMM [52] expression parameters and skip the self-supervised audio encoder. In variant (c), we add the self-supervised audio encoder learning, but we use expression features as in (a). Finally, in variant (d), we learn the expression features, using 3DMM expression pa-

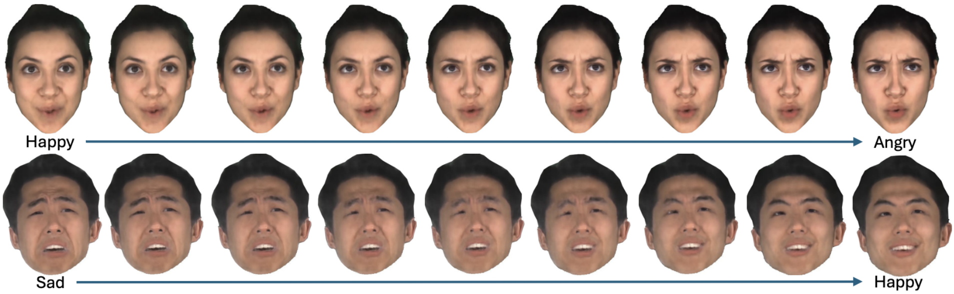

Figure 5: Additional analysis/experiment/visualisation that shows that the expression encoder disentangles features that are semantically grounded and well-behaved. Interpolation of features between different emotional expressions leads to semantically meaningful expressions.

rameters as input to our transformer. We see that not disentangling the emotion recognition features in (a) and (c), and the expression parameters from [9, 52] in (b), cause the NeRF network to only learn expressions from the expression source and ignore the audio source. This leads to a significant decrease in lip-sync metrics on unseen audio and best performance in terms of Exp-Diff and LPIPS. Further, trying to disentangle 3DMM expression parameters in (d) fails to learn meaningful features which leads to poor lip-sync metrics on unseen audio and the worst performance in terms of Exp-Diff and LPIPS. Our proposed expression transformer leads to a successful disentanglement between expression and speech-specific lip motion. Note that Exp-Diff is computed against the driving expression images, which implies that if the network has overfitted to the driving expression the corresponding Exp-Diff would also be lower. Disentangling expressions from lip motion leads to a balanced performance between expression and lip-sync accuracy.
Interpreting Learnt Features. In Fig. 5, we conduct further analysis to investigate the proposed expression disentanglement and the nature of the learned expression features. The interpolation result between two expression features, learned by our expression transformer, shows that our method learns semantically grounded features. In the suppl. material, we show additional interpolation results and show t-SNE plots of the learned expression features, indicating that they are semantically meaningful.

5 Conclusion

In conclusion, we introduce a novel method for joint expression and audio-guided talking face generation. Prior work either struggles to preserve the speaker identity or fails to synthesize faithful facial expressions. We propose a self-supervised method to extract audio features, aligned to lip motion, achieving accurate lip synchronization to unseen audio. In addition, we design a transformer-based module to learn expression features, disentangled from speech-specific mouth motion. By conditioning on the learned representations, our dynamic NeRF synthesizes high-fidelity talking face videos, providing simultaneous control of facial expressions and lip movements, and outperforming the current state-of-the-art. We argue that our proposed representations can be easily extended to other neural rendering pipelines, such as Gaussian Splatting [37], that we plan to explore as future work.
Acknowledgements. This work was supported in part by Amazon Prime Video and a grant from the CDC/NIOSH (U01 OH012476).

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Vaino Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. Deep speech 2: end-to-end speech recognition in english and mandarin. In *Int. Conf. Mach. Learn.*, 2016.

[2] ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Self-supervised deformation modeling for facial expression editing. In *IEEE Conf. Face Gest. Recog. (FG)*, 2019.

[3] ShahRukh Athar, Albert Pumarola, Francesc Moreno-Noguer, and Dimitris Samaras. Facedet3d: Facial expressions with 3d geometric detail prediction. *ArXiv*, abs/2012.07999, 2020.

[4] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[5] ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *IEEE Conf. Face Gest. Recog. (FG)*, 2023.

[6] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021.

[7] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[8] R. Bellman and R. Kalaba. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9, 1959.

[9] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graph.*, 20(3):413–425, 2013.

[10] Aggelina Chatziagapi, ShahRukh Athar, Abhinav Jain, Rohith MV, Vimal Bhat, and Dimitris Samaras. Lipnerf: What is the right feature space to lip-sync a nerf? In *IEEE Conf. Face Gest. Recog. (FG)*, 2023.

[11] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[12] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Adv. Neural Inform. Process. Syst.*, 2018.

[13] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, 2016.

[14] JS Chung, A Jamaludin, A Zisserman, et al. You said that? In *Brit. Mach. Vis. Conf.*, 2017.

[15] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[17] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019.

[18] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[19] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[20] Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou. Head2head++: Deep facial attributes re-targeting. *IEEE Trans. Bio., Beh., Id. Sci.*, 3:31–43, 2020.

[21] Paul Ekman and Wallace V. Friesen. Facial action coding system: a technique for the measurement of facial movement. *Environmental Psychology and Nonverbal Behavior.*, 1978.

[22] Paul Ekman and Erika L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 04 2005.

[23] Shengli Fu, R. Gutierrez-Osuna, A. Esposito, P.K. Kakumanu, and O.N. Garcia. Audio/visual mapping with cross-modal hidden markov models. *IEEE Trans. Multimedia*, 7(2):243–252, 2005.

[24] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[25] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Int. Conf. Comput. Vis.*, 2023.

[26] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[27] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J. Black, and Timo Bolkart. Gif: Generative interpretable faces. In *IEEE Conf. 3D Vis. (3DV)*, 2020.

[28] Sahil Goyal, Sarthak Bhagat, Shagun Uppal, Hitkul Jangra, Yi Yu, Yifang Yin, and Rajiv Ratn Shah. Emotionally enhanced talking face generation. In *Int. Conf. Multimedia Workshop*, 2023.

[29] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Int. Conf. Comput. Vis.*, 2021.

[30] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567*, 2014.

[31] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Int. Conf. Learn. Represent.*, 2017.

[32] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[33] Fa-Ting Hong, , Li Shen, and Dan Xu. Dagan++: Depth-aware generative adversarial network for talking head video generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46 (5):2997–3012, 2023.

[34] Youngjoon Jang, Kyeongha Rho, Jongbin Woo, Hyeongkeun Lee, Jihwan Park, Youshin Lim, Byeong-Yeol Kim, and Joon Son Chung. That's what i said: Fully-controllable talking face generation. In *ACM Int. Conf. Multimedia*, 2023.

[35] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[36] 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 Proc. SIGGRAPH*, 2022.

[37] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023.

[38] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4):163, 2018.

[39] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Int. Conf. Mach. Learn.*, 2018.

[40] Taehwan Kim, Yisong Yue, Sarah Taylor, and Iain Matthews. A decision tree framework for spatiotemporal sequence prediction. In *Proc. ACM SIGKDD Int. Conf. Knowledge Disc. Data Mining*, 2015.

[41] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. In *IEEE Conf. Face Gest. Recog. (FG)*, 2020.

[42] Dongze Li, Kang Zhao, Wei Wang, Bo Peng, Yingya Zhang, Jing Dong, and Tieniu Tan. Ae-nerf: Audio enhanced neural radiance field for few shot talking head synthesis. *Proc. AAAI Conf. on Art. Intel.*, 38(4), 2024.

[43] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Int. Conf. Comput. Vis.*, 2023.

[44] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[45] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[46] Jiyuan Liu, Wenping Wei, Zhendong Li, Guanfeng Li, and Hao Liu. Invariant motion representation learning for 3d talking face synthesis. In *ICASSP*, 2024.

[47] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *Eur. Conf. Comput. Vis.*, 2022.

[48] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.*, 2020.

[49] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[50] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021.

[51] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021.

[52] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009.

[53] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *ACM Int. Conf. Multimedia*, 2023.

[54] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[55] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Int. Conf. Multimedia*, 2020.

[56] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Eur. Conf. Comput. Vis.*, 2018.

[57] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[58] Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pva: Pixel-aligned volumetric avatars. In *arXiv:2101.02697*, 2020.

[59] Yurui Ren, Gezhong Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Int. Conf. Comput. Vis.*, 2021.

[60] Shinji Sako, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Hmm-based text-to-audio-visual speech synthesis. In *"Proc. Int. Speech Comm. Assn., INTERSPEECH"*, 2000.

[61] Shuai Shen, Wanhua Li, Xiaoke Huang, Zheng Zhu, Jie Zhou, and Jiwen Lu. Sd-nerf: Towards lifelike talking head animation via spatially-adaptive dual-driven nerfs. *IEEE Trans. Multimedia*, 26:3221–3234, 2024.

[62] Zhicheng Sheng, Liqiang Nie, Min Zhang, Xiaojun Chang, and Yan Yan. Stochastic latent talking face generation toward emotional expressions and head poses. *IEEE Trans. Cir. Sys. Vid. Tech.*, 34(4):2734–2748, 2024.

[63] Sanjana Sinha, S. Biswas, Ravindra Yadav, and B. Bhowmick. Emotion-controllable generalized talking face generation. In *IJCAI*, 2022.

[64] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. In *Eurographics Symposium on Rendering*, 2021.

[65] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.*, 36(4):1–13, 2017.

[66] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Int. Conf. Comput. Vis.*, 2023.

[67] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022.

[68] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: image synthesis using neural textures. *ACM Trans. Graph.*, 38(4):1–12, 2019.

[69] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *Eur. Conf. Comput. Vis.*, 2020.

[70] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2019.

[71] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019.

[72] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[73] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T. Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[74] 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 *Eur. Conf. Comput. Vis.*, 2020.

[75] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[76] Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *Int. Conf. Comput. Vis.*, 2021.

[77] Xiuzhe Wu, Pengfei Hu, Yang Wu, Xiaoyang Lyu, Yan-Pei Cao, Ying Shan, Wenming Yang, Zhongqian Sun, and Xiaojuan Qi. Speech2lip: High-fidelity speech to lip generation by learning from a short video. In *Int. Conf. Comput. Vis.*, 2023.

[78] Yibo Xia, Lizhen Wang, Xiang Deng, Xiaoyan Luo, and Yebin Liu. Gmtalker: Gaussian mixture based emotional talking video portraits. *arXiv:2312.07669*, 2023.

[79] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[80] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM Proc. SIGGRAPH*, 2023.

[81] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfanerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022.

[82] Zhenhui Ye, Jinzheng He, Ziyue Jiang, Rongjie Huang, Jiawei Huang, Jinglin Liu, Yi Ren, Xiang Yin, Zejun Ma, and Zhou Zhao. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*, 2023.

[83] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In *Int. Conf. Learn. Represent.*, 2023.

[84] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[85] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proc. AAAI Conf. Art. Intel. and Innov. App. Art. Intel. Conf. and AAAI Symp. Edu. Adv. Art. Intel.*, 2019.

[86] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

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